# La conversion génique biaisée : origine, dynamique et intensité de la quatrième force d'évolution des génomes eucaryotes

Yann Lesecque

N° 122-2014 Année 2014

Thèse de l'université de Lyon

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON 1

pour l'obtention

du Diplôme de doctorat
(arrêté du 7 août 2006)

soutenue publiquement le
11 juillet 2014

par

Yann Lesecque

# La conversion génique biaisée : origine, dynamique et intensité de la quatrième force d'évolution des génomes eucaryotes

Directeur de thèse : Laurent Duret

| | | |
|---|---|---|
| Jury : | Bernard de Massy | Rapporteur |
| | Laurent Duret | Directeur de thèse |
| | Nicolas Galtier | Rapporteur |
| | Evelyne Heyer | Examinateur |
| | Cristina Vieira-Heddi | Examinateur |

# UNIVERSITE CLAUDE BERNARD - LYON 1

| | |
|---|---|
| **Président de l'Université** | **M. François-Noël GILLY** |
| Vice-président du Conseil d'Administration | M. le Professeur Hamda BEN HADID |
| Vice-président du Conseil des Etudes et de la Vie Universitaire | M. le Professeur Philippe LALLE |
| Vice-président du Conseil Scientifique | M. le Professeur Germain GILLET |
| Directeur Général des Services | M. Alain HELLEU |

## *COMPOSANTES SANTE*

| | |
|---|---|
| Faculté de Médecine Lyon Est – Claude Bernard | Directeur : M. le Professeur J. ETIENNE |
| Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux | Directeur : Mme la Professeure C. BURILLON |
| Faculté d'Odontologie | Directeur : M. le Professeur D. BOURGEOIS |
| Institut des Sciences Pharmaceutiques et Biologiques | Directeur : Mme la Professeure C. VINCIGUERRA |
| Institut des Sciences et Techniques de la Réadaptation | Directeur : M. le Professeur Y. MATILLON |
| Département de formation et Centre de Recherche en Biologie Humaine | Directeur : Mme. la Professeure A-M. SCHOTT |

## *COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE*

| | |
|---|---|
| Faculté des Sciences et Technologies | Directeur : M. F. DE MARCHI |
| Département Biologie | Directeur : M. le Professeur F. FLEURY |
| Département Chimie Biochimie | Directeur : Mme C. FELIX |
| Département GEP | Directeur : M. H. HAMMOURI |
| Département Informatique | Directeur : M. le Professeur S. AKKOUCHE |
| Département Mathématiques | Directeur : M. G. TOMANOV |
| Département Mécanique | Directeur : M. le Professeur H. BEN HADID |
| Département Physique | Directeur : M. J.C. PLENET |
| UFR Sciences et Techniques des Activités Physiques et Sportives | Directeur : M. Y.VANPOULLE |
| Observatoire des Sciences de l'Univers de Lyon | Directeur : M. B. GUIDERDONI |
| Polytech Lyon | Directeur : M. P. FOURNIER |
| Ecole Supérieure de Chimie Physique Electronique | Directeur : M. G. PIGNAULT |
| Institut Universitaire de Technologie de Lyon 1 | Directeur : M. C. VITON |
| Ecole Supérieure du Professorat et de l'Education | Directeur : M. A. MOUGNIOTTE |
| Institut de Science Financière et d'Assurances | Directeur : M. N. LEBOISNE |

*A Nadine Catry,*
*Emmanuel Pietre, Claude Richard et Philippe Donaire*
*qui m'ont donné le goût des belles lettres : A, C, T et G...*
*... et à Gérard Cachet qui m'a appris à les dénombrer.*

# Table des matières

# Remerciements

Je tiens à remercier, en premier lieu, les membres du jury qui ont accepté d'évaluer mon travail : les professeurs Evelyne Heyer et Cristina Vieira-Heddi ainsi que les docteurs Bernard de Massy et Nicolas Galtier, qui ont également relu et corrigé le manuscrit.

Je remercie tout particulièrement mon directeur de thèse, Laurent Duret. J'aimerais ici te dire à quel point je te suis reconnaissant de m'avoir orienté vers cette voie, de m'avoir montré ce qu'est vraiment la science, comment elle se construit chaque jour et bouleverse nos *a priori*. De m'avoir appris les bases de la programmation avec patience, d'avoir été disponible souvent, j'insiste, très souvent ! Mais pour être un bon tuteur, il ne suffit pas de bien apprendre et de donner envie de connaitre, il faut aussi créer les conditions favorables à l'entente et au respect mutuel. De ce côté encore, Laurent, les choses n'auraient pu mieux se passer. Merci donc pour la qualité de vie que tu as amené dans mon travail chaque jour depuis cinq ans maintenant, on ne peut rêver meilleur directeur.

Merci à ceux qui ont accepté de collaborer avec nous durant ces quelques années. Merci à Nicolas Lartillot, Sylvain Glémin, Sylvain Mousset, Franck Picard et Laurent Gueguen d'avoir pallier mes lacunes en maths, stats et génétique des populations. Merci à Richard Bourgon, Anamaria Necsulea et Henrik Kaessmann de nous avoir autorisé l'accès à leurs données lorsque nous en avions besoin. Merci à nos collègues du CRCL, Sylvie Mazoyer et Chloé Tessereau d'avoir bien voulu nous associer à leur étude.

Il me faudrait ici prendre 5 double-pages pour lister l'ensemble de mes amis et collègues de l'équipe BGE et du PRABI, anciens et nouveaux. Merci à tous pour votre soutien, la facilité avec laquelle nous travaillons tous ensemble

dans une atmosphère chaleureuse, conviviale et détendue est un trésor qu'il ne faut pas laisser s'échapper. Vous allez profondément me manquer lorsque je serai en salle des profs ! Je tiens à remercier tout particulièrement mon collègue et ami Mathieu Groussin pour son soutien fidèle depuis notre entrée commune au laboratoire. Merci à Thomas Bigot, mon couteau-suisse, le trouveur de solutions, une canne blanche dans le monde obscur du doctorat. Merci à Erika Kvikstad et Joanna Parmley for your support and help. Merci à Fanny Pouyet, Florent Lassalle, Rémi Planel, Aline Muyle, Laurent Jacob, Ghislain Durif, Michel Lecoq, Dominique Guyot, Murray Patterson, Simon Penel, Lionel Humblot, Stéphane Delmotte, Adil El-Filali, Nicolas Rochette, Magali Semeria, Héloïse Philippon, Clément Goubert, Laurent Modolo, Gabriel Terraz, Floriane Plard, Christophe Plantamp et Clothilde Deschamps pour vos éclats de rire en salle détente. Merci aux anciens, Jean-François Gout, Eugénie Pessia, Louis-Marie Bobay et Alexandra Popa.

Ma chère Marie Cariou, mon inestimable collègue et amie, comme je suis heureux que nous ayons traversé ensemble ces années intenses en travail, en réunions, en rires et parfois en larmes. Bonne route à toi et à bientôt !

Les Chevaliers du Zodiac avaient Athéna, Frodon avait Gandalf, Bambi avait Pan-pan de mon côté c'est Marie Sémon qui a été présente pour m'épauler dans chacune des étapes qui ont marqué ces années. Merci de m'avoir initié à la génomique, à la bioinfo, à l'évolution. Merci de m'avoir fait confiance pour les enseignements, merci de m'avoir écouté quand ça n'allait pas. Quoique tu en dises et que tu veuilles bien me croire ou non, tu seras toujours le premier de mes mentors. Merci aussi à ta sœur, Pauline Sémon, de m'avoir fourni une délicieuse illustration pour la page 133.

Merci aussi à ceux qui ont beaucoup compté dans mon parcours, professeurs ou tuteurs, vous avez indirectement contribué à ce travail : Adam Eyre-Walker, Emmanuel Douzery, Jean-Nicolas Volff, Pierre Thomas, Lluis Quintana-Murci, Raquel Tavares, Cristina Vieira-Heddi, Manolo Gouy, Gwenaël Piganeau, Kateryna Makova, Anton Nekrutenko, Gaël Yvert, Sylvain Charlat et Ludovic Orlando. Plus particulièrement, j'aimerais ici remercier Dominique Mouchiroud pour le soutien et la confiance qu'elle m'a accordé, merci aussi pour ces longues discussions sur le métier de la recherche et celui de professeur, merci pour ta patience et ta bienveillance, merci de m'avoir donné le goût du Neutralisme et la détermination nécessaire à la réalisation de nos projets scientifiques.

Un merci tout particulier à Tristan Lefebure, mon tuteur d'école docto-

rale, merci pour ton soutien et ta disponibilité.

Je tiens en outre à remercier ceux qui ont été à mes côtés lors des enseignements du monitorat, l'une des expériences les plus enrichissantes de ma thèse. Merci donc à Laurence Mouton, Marie-Claude Venner, Raquel Tavares, Marie Fablet, Clément Goubert, Marie Sémon, Marc Bailly-Béchet, Céline Brochier-Armanet et Cristina Vieira-Heddi, j'aimerais que mes futurs collègues soient aussi agréables que vous. Merci aux étudiants de licence et master qui ont subi mes cours et TD : les L3 de l'ENS de Lyon, les L2 de l'UCBL et les 4BIM de l'INSA de Lyon. Merci aux différents stagiaires de Laurent avec qui j'ai pu interagir : Audrey, Damien, Ivo, Amanda, Du et Marie.

Merci à la dreamteam du pôle administratif de l'UMR 5558, que serions nous sans les sourires et l'efficacité de Nathalie Abrasetti, Aline Maitrias, Laetitia Mangeot et Odile Mulet-Marquis ?

Mais la vie ne s'est pas arrêtée aux portes de la Doua et pendant ce temps, les amis de toujours étaient bien présents pour m'entourer, à Paris, à Grenoble ou ailleurs. Vous êtes mes guides et pas un pas ne pourrait être fait sans vous, Alexander Clevering, Hubert Wolff, Camille Matinal, Clémence Tota, Emilie et Virginie Foray, Nelly Jolmes, Adrien Beck, Vincent Mary et Louis Rambert, je vou(s) aime. Merci aussi à mes fidèles amis de Lyon, Alexandre Zagdoun, Sara Lefevre, Germain Lesoeur, Richard Griffon, Pierre-François Labousse et Carl Daniels, nous partageons tout : GoT, QVEMF, LAEDLP, TaupeCheffent, des Snap-LoLz et bien plus encore ! Un merci tout particulier aux sœurs C. et M. Odieuse de Beauregard.

Merci aussi à mes compères du théâtre, Laura Tatoueix, Gabriella Serban, Théo Tacail, Léa Bello, Lucy Michel et Etienne Besson. Many thanks to my collegues and friends from Sussex Uni especially Eduardo Medina Barcenas, Lynne Robinson and Romain Blanc-Mathieu.

Il est temps maintenant de remercier ceux qui étaient avec moi à l'école, ceux qui ont partagé plus que des bancs d'amphis inconfortables, plus que des cours sans fin sur le cycle cellulaire, la stratigraphie séquentielle ou la matrice extra-cellulaire, plus que le travail : la fête, la joie d'être ensemble, ici et ailleurs. Merci les types, Florent Mazel, Blaise Tymen, Mathieu Groussin, Pierre Levy, Pierre Lemerre, je ris rien qu'en pensant à vous. Merci aussi à Harrisson Hor et Damien Fournier mes colocataires de rêve. Merci à Domitille Chalopin pour ses conseils en arts martiaux et à Anne Juras et Filipe

de Vadder d'avoir partagé avec nous les affres de l'agreg !

Un merci tout particulier à celle qui a été la femme de ma vie pendant près de six ans. Docteur Chloé Tessereau je te dois tellement, scientifiquement et surtout humainement. Merci d'avoir subi ma musique, mes bavardages stériles, mes joies, mes peines pendant ces années. Merci d'avoir été toujours là. Merci de me soutenir, de me faire rire. On ne s'oublie pas.

Un merci très spécial, qui se veut discret mais qui ne le sera pas, à Luc Lauro. Merci de me supporter, merci d'être toi, avec moi.

Enfin, que serait l'éternel étudiant sans sa chère famille ? Je remercie mes parents Chantal et Jean-Luc Lesecque qui ont eux-même commencé à poser les bases de cette thèse, sans le savoir, il y à environ 25 ans et n'ont jamais cessé d'y contribuer. Si je ne devais être fier que d'une seule chose, ce serait de l'éducation que vous m'avez donnée. Je remercie mes oncles et tantes ainsi que mes cousins, de métropole et d'outre-mer, pour avoir toujours cru en moi, particulièrement messieurs Yorick Lesecque, Mathieu et Florent Loir, quelle belle brochette ! Je remercie aussi mes deux grands-mères Janine Lesecque et Odette Loir qui m'inspirent depuis toujours et j'ai une pensée toute particulière pour Oscar Loir qui nous a quitté peu avant que je ne commence ce travail.

# Préambule

*Cependant la lune se lève*
*Et l'esquif en sa course brève*
*File gaîment sur l'eau qui rêve.*

C'est par ces vers que s'achève le poème "En bateau" dans lequel Paul Verlaine utilise la métaphore de la barque ("l'esquif") pour décrire la recherche, malheureusement vouée à l'échec, de l'eldorado amoureux. Aujourd'hui, c'est d'un autre voyage en mer dont nous parlerons : l'évolution. Dans le cadre de cette thèse, je me suis intéressé à l'une des forces qui dirige cette régate biologique : la conversion génique biaisée (BGC). Sous ce nom quelque peu barbare se cache en réalité ce qui n'était d'abord qu'une hypothèse, un modèle à l'origine évolutive floue, permettant d'expliquer les surprises issues du séquençage des génomes. Le BGC est un processus lié à la recombinaison qui crée un biais dans la ségrégation des caractères à l'échelle de la population des gamètes produits lors de la méiose chez les eucaryotes.
Classiquement, la régate de l'évolution admet être gouvernée par trois forces. Premièrement, la mutation est l'armateur qui crée les nouveaux variants (allèles) et les met à flot sur une mer alors occupée principalement par un seul autre navire. A ce moment, c'en est déjà fini du rôle de cet inventeur et le destin de la barque est remis uniquement entre les mains des deux autres protagonistes. Le second acteur de cette course est la sélection naturelle qui, à la manière du pilote impose un cap à la force de ses bras. Cependant, maintenir le cap n'est permis que si les vents de la dérive génétique, aléatoires et changeant continuellement de direction sont assez faibles. Ainsi, deux forces s'opposent ou s'allient dans l'évolution des fréquences alléliques de génération en génération. La sélection est orientée car elle impose une direction précise : le variant est favorisé ou non. Elle est aussi adaptative car le variant n'est

favorisé que si il améliore significativement la survie et/ou le succès reproducteur de individu. A l'inverse, la dérive génétique est stochastique car elle n'impose pas une trajectoire évolutive prédictible d'une génération à l'autre, de ce fait elle est aussi non-adaptative.

*La Porte d'aval avec des bateaux partant à la pêche - Claude Monet - 1885*

Cependant, depuis les années 1980, une quatrième force vient s'ajouter à ce tableau. Dans cette galère, le BGC occupe une position intermédiaire. A la manière d'un courant atlantique, il est orienté car il favorise ou non le variant dans sa course. Malgré cela, il n'est pas adaptatif. En effet, en dehors du cas où la sélection ramerait assidûment dans le sens de ce courant, celui-ci représente une gêne à la navigation surtout si il est très fort. Nous entrevoyons ici ce qui fait l'essence même de la notion de force évolutive : il s'agit d'un processus qui joue sur l'apparition, le maintien ou la disparition d'allèles dans les populations à condition que son intensité soit assez importante pour surpasser l'influence de ses semblables.
Ainsi, afin de montrer que le BGC a une place à part entière aux cotés de la mutation, la sélection et la dérive sur l'océan de l'évolution, il est primordial de caractériser cette force. Pour cela nous tenterons de répondre à trois

questions fondamentales sur le BGC. Tout d'abord, quelle est son origine moléculaire, c'est à dire quels mécanismes, quels enzymes, sous-tendent ce processus. Deuxièmement, il nous faudra savoir quand, à l'échelle des temps évolutifs, et où, à l'échelle des génomes, le BGC agit. Enfin, nous quantifierons son intensité afin de mieux comprendre si il est apte à surpasser, voire à contrecarrer, les autres forces, notamment la sélection et la dérive.
Afin de répondre à ces questions, mon travail de thèse s'est divisé en deux études chacune portée sur une espèce différente : la levure de boulanger *Saccharomyces cerevisiae* et l'homme. Ces deux analyses ont permis la production de deux articles scientifiques qui seront présentés dans deux chapitres de résultats (chapitres II & III), eux-mêmes précédés d'un chapitre introductif faisant l'état des connaissances actuelles sur le BGC (chapitre I). Nous restreindrons notre propos au domaine des eucaryotes. Cependant, l'éventualité de l'existence du BGC chez les bactéries sera évoquée aux cotés d'un bilan et d'autres perspectives dans un dernier chapitre de discussion (chapitre IV).

# Liste des abréviations

| | |
|---|---|
| **A** | Adénine |
| **AA** | Américains d'origine africaine (population *cf.* [Hinch et al., 2011]) |
| **aa** | Acides aminés (unité) |
| **ABC** | Méthode bayésienne approchée (Approximate Bayesian Computation) |
| **ADN** | Acide DésoxyriboNucléique |
| **ARN** | Acide RiboNucléique |
| **BER** | Système de réparation par excision de base (Base Excision Repair) |
| **BGC** | Conversion génique biaisée (Biased Gene Conversion) |
| **C** | Cytosine |
| **CEU** | Résidents de l'Utah originaires de l'Europe du Nord et de l'Ouest (population humaine *cf.* [The International HapMap Consortium, 2007]) |
| **CHB** | Chinois Han de Beijing en Chine (population humaine *cf.* [The International HapMap Consortium, 2007]) |
| **ChIP** | Immunoprécipitation de chromatine (Chromatine ImmunoPrecipitation) |
| **ChIPseq** | Idem suivi d'un séquençage à haut débit |
| **cM** | CentiMorgan (unité) |
| **CMH** | Complexe Majeur d'Histocompatibilité |
| **CO** | Enjambement (Crossing-Over) |

| | |
|---|---|
| **DAF** | Fréquence de l'allèle dérivé (Derived Allele Frequency) |
| **DAPI** | 4',6'-diamidino-2-phénylindole (colorant) |
| **dBGC** | Conversion génique biaisée d'initiation (associée aux cassures double brins) |
| **DL** | Déséquilibre de liaison |
| **dHj** | Double jonction de Holliday |
| **DSB** | Cassure double brin (Double Strand Break) |
| **DSBR** | Modèle de réparation des DSB (Double Strand Break Repair model) |
| **FITC** | Isothiocyanate de fluorescein (colorant) |
| **G** | Guanine |
| **Gb** | Giga paire de bases (unité) |
| **gBGC** | Conversion génique biaisée vers GC |
| **GC*** | Taux de G+C d'équilibre |
| **GC3** | Taux de G+C mesuré à la $3^{ème}$ position des codons des exons |
| **H3K4Me3** | Triméthylation de la lysine 4 de l'histone 3 (marque épigénétique) |
| **HACNS** | Régions non-codantes à évolution accélérée chez l'homme (Human Accelerated Conserved Noncoding Sequences) |
| **HAR** | Régions à évolution accélérée chez l'homme (Human Accelerated Regions) |
| **HR** | Haute résolution |
| **ILS** | Tri de lignée incomplet (Incomplete Lineage Sorting) |
| **IS** | Informations Supplémentaires liées à une publication |
| **JPT** | Japonais de Tokyo au Japon (population humaine *cf.* [The International HapMap Consortium, 2007]) |
| **kb** | Kilo paire de base (unité) |
| **LD** | Déséquilibre de liaison (Linkage Desequilibrium) |
| **Ma** | Million d'années (unité) |
| **Mb** | Méga paire de bases (unité) |

| | |
|---|---|
| **MMR** | Système de réparation des mésappariements (MisMatch Repair) |
| **N** | Taille des populations |
| **NCO** | Non Crossing-Over |
| **NGS** | Méthodes de séquençage de nouvelle génération (Next Generation Sequencing) |
| $N_e$ | Taille efficace des populations |
| **NER** | Système de réparation par excision de nucléotide (Nucleotide Excision Repair) |
| **PAR** | Régions pseudo-autosomales des chromosomes sexuels (Pseudo-Autosomal Regions) |
| **pb** | Paire de bases (unité) |
| **PCR** | Amplification en chaîne par polymérase (Polymerase Chain Reaction) |
| **PMS** | Ségrégation post-méiotique (Post Meiotic Segregation) |
| **RFLP** | polymorphisme de longueur des fragments de restriction (Restriction fragment length polymorphism) |
| **S** | G ou C (Strong) |
| **SDSA** | Modèle de déplacement et réassociation de brin dépendant de la synthèse (Synthesis Dependent Strand Annealing) |
| **SEI** | Invasion de brin (Single End Invasion) |
| **SNP** | Polymorphisme nucléotidique (Single Nucleotide Polymorphism) |
| **SW** | Mutations de GC vers AT |
| **T** | Thymine |
| **W** | A ou T (Weak) |
| **WS** | Mutations de AT vers GC |
| **YRI** | Yoruba d'Ibadan au Nigeria (population humaine *cf.* [The International HapMap Consortium, 2007]) |
| **ZnF** | Domaine à doigt de zinc (Zinc Finger domain) |

Chapitre I

# Introduction

Le BGC, un jeune concept d'évolution moléculaire intimement lié à la recombinaison.

# I.A. Le BGC, conséquence mécanique de la recombinaison

L'histoire de la découverte de la conversion génique et de la recombinaison ressemble à beaucoup d'autres aventures scientifiques : chaque grande avancée sur le sujet est associée à une approche expérimentale particulière, permise par les progrès techniques de l'époque. Elle commence par l'étude des croisements qui a permis la découverte du principe de liaison génétique et fait appel, de nos jours, à des techniques de plus en plus sophistiquées telles que le génotypage de cellules germinales (sperm-typing). Dans cette partie, nous donnerons d'abord une définition historique du processus de recombinaison et verrons que ce concept s'est transformé au fil des découvertes pour devenir celui que nous utilisons aujourd'hui.

## I.A.1. La découverte de la recombinaison

### La liaison génétique

C'est avec la redécouverte et la réinterprétation des lois de Mendel, au début du XXème siècle que le concept de liaison génétique est né. Dans une étude de 1905, Bateson, Saunders et Punnett réalisèrent un croisement de plants de pois homozygotes (on parle alors de « lignées pures ») pour deux caractères : la forme du grain pollen et la couleur de la fleur [Bateson et al., 1905]. Le premier plant de la génération mère (F0) était à fleurs rouges et pollen rond, ces deux caractères sont récessifs sur ceux du second plant : fleurs violettes et pollen allongé (Figure I.1). Ainsi, la génération F1 issue de ce croisement était composée d'individus tous hétérozygotes aux loci concernés qui portent tous des fleurs violettes et des grains de pollen allongés. Selon la loi de ségrégation indépendante des caractères édictée quarante ans plus tôt par Mendel, à la génération suivante F2, issue du croisement d'individus F1 entre eux, on devait obtenir chacune des 4 combinaisons des 2 versions de chacun des 2 caractères dans les proportions 9:3:3:1. Cependant, Bateson et collègues observent un déficit significatif de plants à fleurs violettes et pollen rond ainsi que de fleurs rouges à pollen allongé comparé aux effectifs attendus (Figure I.1). Ceci indique que les caractères qui étaient présents chez les ancêtres F0 ségrègent plus fréquemment ensemble à travers les générations. C'est la liaison génétique.

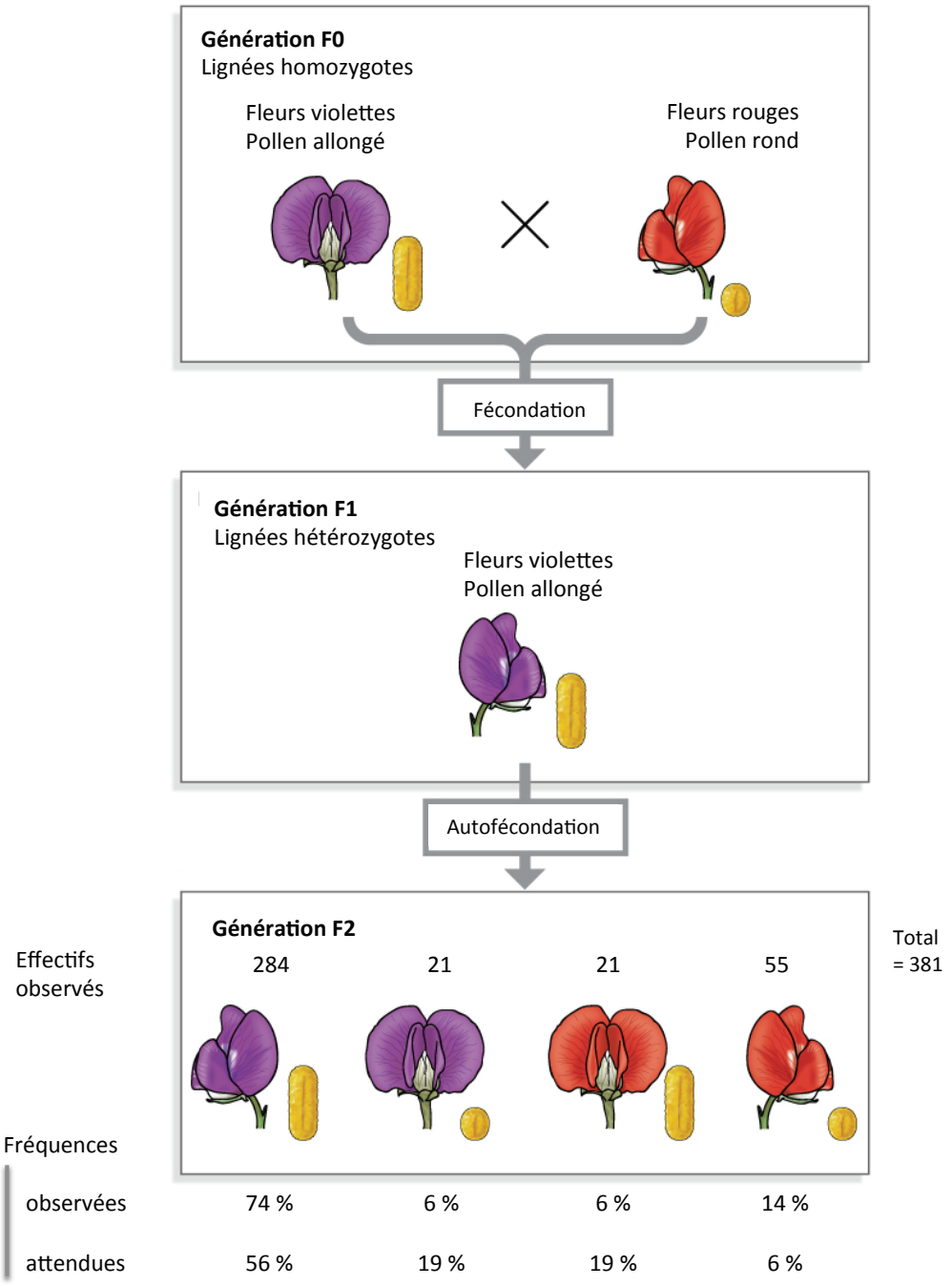



FIGURE I.1 : Example simplifié de croisement effectué par [Bateson et al., 1905] sur des variétés de pois. A la génération F2, sous l'hypothèse que les caractères de couleur de la fleur et de forme du pollen ségrègent indépendamment, les fréquences attendues de chaque phénotype sont en proportions 9:3:3:1. L'expérience montre que cette hypothèse n'est pas vérifiée : il y a un déficit de phénotypes non-parentaux (c'est à dire ceux non représentés à la génération F0). Adapté de [Lobo & Show, 2008].

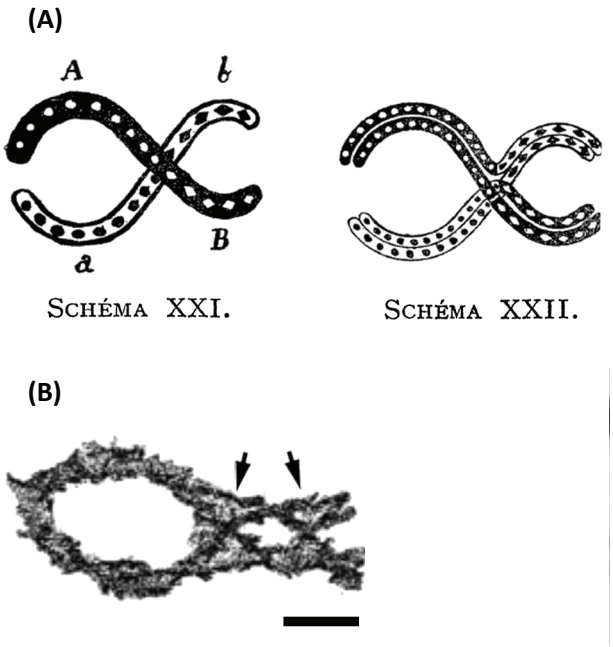



Figure I.2 : (A) Schémas originaux de Janssens représentant l'interprétation moléculaire des figures de chiasma. XXI : chiasma entre deux chromosomes homologues AB (noir) et ab (blanc) XXII : idem avec représentation des deux chromatides de chaque chromosome. (B) Deux chiasmas (flèches) observés dans un meiocyte de salamandre (*Oedipina poelzi*). La barre d'échelle représente environ 5 $\mu$m. Adapté de [Koszul et al., 2012].

### Apport de la théorie chromosomique

C'est dix ans plus tard, en 1915, que Thomas Hunt Morgan fit la démonstration que la transmission des caractères et la ségrégation des chromosomes observés dans des cellules de drosophiles sont liés [Morgan et al., 1915]. La liaison génétique correspond donc à une liaison physique des gènes le long des chromosomes. Cependant, dans l'expérience de Bateson, les caractères fleurs rouges et pollen rond ne sont pas systématiquement liés comme ils le sont chez les parents (générations F0 et F1). Il faut donc imaginer que chez certains plants, ces caractères ont été réassortis, on parle alors d'individus recombinants. Il y a donc eu des échanges entre groupes de caractères liés dont la fréquence peut être estimée directement grâce à la proportion de recombinants. Ainsi, si l'on compte 42 recombinants sur 381 plants comme dans l'exemple de la Figure I.1, on dira que les gènes (définis comme segments de chromosomes correspondant chacun à un caractère) de la couleur de la fleur et de la forme du grain de pollen du pois sont liés et à une distance génétique de (42 / 381) x 100 = 11 centiMorgan (cM). Ainsi, historiquement, la recombinaison a été définie comme la capacité de rompre la liaison génétique entre deux caractères [Morgan, 1911]. Cette capacité est d'autant plus grande que la distance génétique entre les deux caractères est grande.

### Les chiasmas

Parallèlement à cela, grâce aux progrès de la microscopie et à l'avènement de la *camera lucida*, Frans Janssens mit en évidence, en 1909, l'existence de structures chromosomiques "en croix" dans des cellules germinales de tritons et salamandres : les chiasmas (ou chiasmatas) [Koszul et al., 2012] (Figure I.2). Les chiasmas correspondent à l'attachement ponctuel de paires de chromosomes précédant leur séparation lors de la méiose (cf. I.A.2. p. 25). Morgan émit alors l'hypothèse que ces évènements correspondent à l'échange de matériel génétique lors de la recombinaison [Schwartz, 2009]. Il nomme alors "crossing-over" (CO) ces évènements. Cette hypothèse sera vérifiée grâce à l'étude de croisements de plants de riz par le groupe de Barbara McClintock, en 1931 [Creighton & McClintock, 1931]. La recombinaison est donc un processus d'échange de gènes entre chromosomes permis par les CO (ou enjambements en français) dont l'une des manifestations cellulaires est le chiasma. Replaçons maintenant ce processus dans son contexte : la méiose.

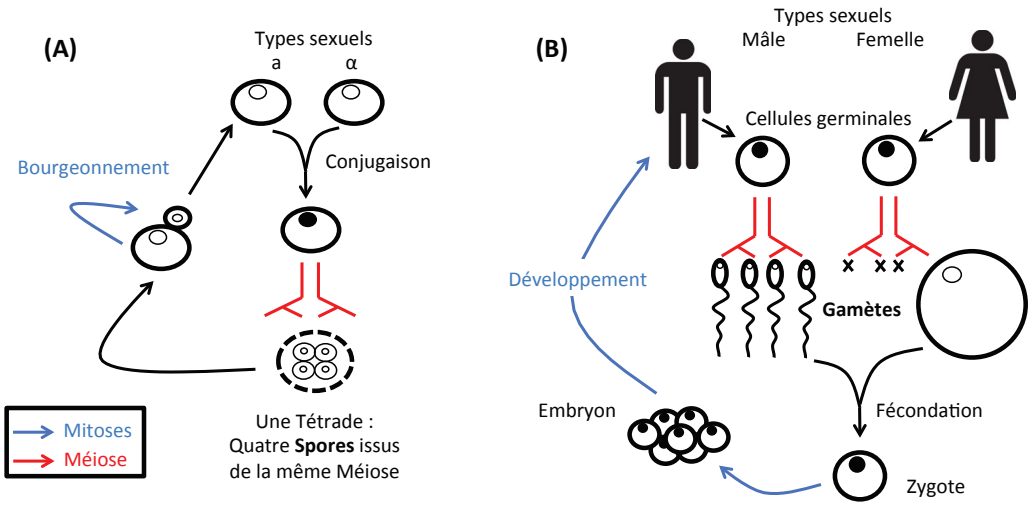



FIGURE I.3 : Cycles de développement des deux espèces eucaryotes : (A) *Saccharomyces cerevisiae* . (B) *Homo sapiens*. Les cellules à noyau plein indiquent un stade diploïde et celles à noyau vide un stade haploïde. Les flèches rouges correspondent à un évènement de méiose et les flèches bleues à plusieurs mitoses. Les produits de méiose (gamètes et spores) sont indiqués en gras. Notez que l'homme passe la grande majorité de son cycle sous forme diploïde, ce qui n'est pas le cas pour la levure. De plus, chez les mammifères la méiose femelle ne produit qu'un ovule car, après chaque division, une cellule sur deux dégénère, laissant la majorité de son cytoplasme à l'autre.

## I.A.2. Aspects cytogénétiques de la recombinaison

### La méiose dans les cycles de développement

La recombinaison homologue est un processus programmé qui se produit, entre autre, lors de la méiose. La méiose est un processus cellulaire ayant lieu chez les espèces sexuées. Elle correspond à la production de quatre cellules haploïdes ($n$ chromosomes) à partir d'une cellule diploïde ($n$ paires de chromosomes homologues). Cette production s'inscrit différemment dans le cycle de développement des organismes à l'échelle des eucaryotes. Chez la plupart des métazoaires sexués, la méiose est une étape de la formation des gamètes (spermatozoïdes et ovules chez les mammifères, par exemple) à partir des cellules de la lignée germinale (Figure I.3 (B)). Chez les Fungi (champignons au sens strict) il existe une grande diversité de cycles. Ainsi, chez la levure *Saccharomyces cerevisiae* , organisme unicellulaire, les individus haploïdes peuvent se reproduire aussi bien par mitose (bourgeonnement) ce qui donne deux clones, que par voie sexuée faisant intervenir une méiose. Pour cela, deux levures haploïdes de types sexuels différents (a/$\alpha$) fusionnent puis l'individu diploïde ainsi formé entre en méiose pour donner quatre spores associées temporairement dans un asque appelé tétrade (Figure I.3 (A)) [Herskowitz, 1988].

### La recombinaison au sein de la méiose

La méiose comprend deux divisions cellulaires successives (méiose I et méiose II) (Figure I.4). La première est dite "réductionnelle", elle assure la séparation des chromosomes homologues en deux lots ce qui permet un premier brassage génétique. La seconde division est dite "équationnelle", elle assure la séparation des deux chromatides soeurs de chaque chromosome et aboutit à un second brassage génétique, analogue à celui observé en mitose sur cellule haploïde. La recombinaison proprement dite a lieu dans les premières étapes de la méiose I : la prophase I. Lors de cette étape, qui marque l'entrée en méiose, des cassures double brins (DSB, Double Strand Break en anglais) se forment le long des chromosomes. Ces cassures sont ensuite réparées grâce à l'intervention du chromosome homologue, utilisé comme matrice (*cf.* I.A.3. p. 32). Il est donc indispensable que les chromosomes homologues se rapprochent et entrent en contact au sein de bivalents afin qu'il y ait CO (Figure I.4). La prophase I de méiose doit donc assurer plusieurs fonctions aboutissant à la recombinaison :

- L'alignement *i.e.* le rapprochement des chromosomes homologues à moins de 400nm l'un de l'autre.

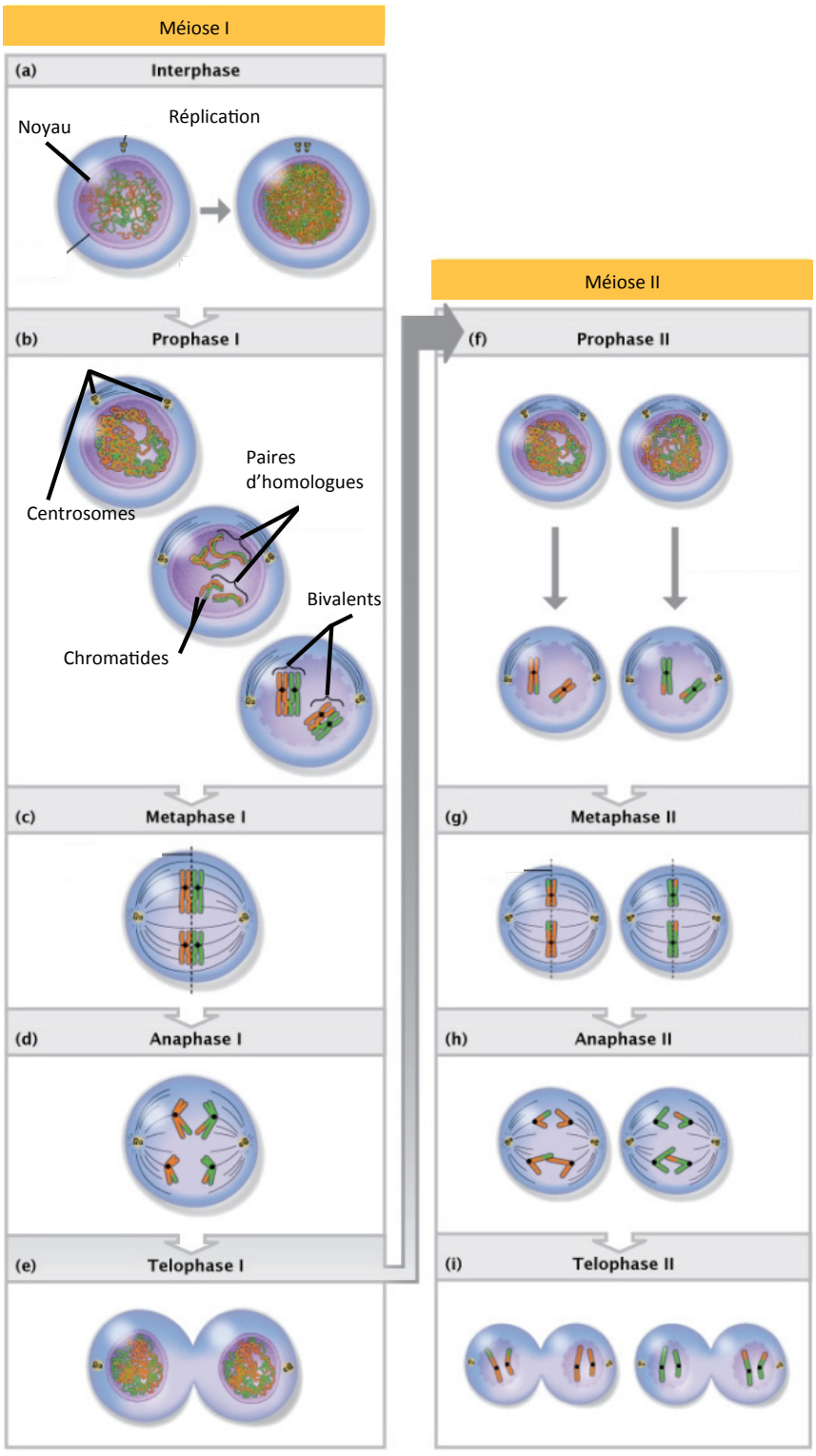



FIGURE I.4 : Schéma simple des évènements génétiques et cellulaires de la méiose dans une cellule à deux paires de chromosomes homologues (même tailles). Les couleurs (vert et orange) permettent de distinguer l'information génétique portée par chacun des homologues d'une paire. La recombinaison a donc lieu en prophase I. Adapté de [O'Connor, 2008].

- Le synapsis *i.e.* l'accolement des deux chromosomes au sein d'une structure nucléoprotéique appelée complexe synaptonémal.

- La formation et la réparation des DSB.

Les modalités de l'enchaînement programmé de ces étapes, parfois chevauchantes, varient selon les espèces. Il est cependant à noter que plusieurs points clés de ces étapes sont fonctionnellement conservés à l'échelle des eucaryotes [Yanowitz, 2010]. De plus, la séquence du gène *spo11*, codant pour l'enzyme qui catalyse la formation programmée des DSB en méiose, est elle aussi conservée [Keeney et al., 1997].

### L'appariement des chromosomes homologues

La prophase I est divisée en plusieurs étapes marquées par l'état d'association des chromosomes homologues et le degré de compaction de leur ADN :

- Leptotène : Les chromosomes sortent d'une phase de réplication et ont chacun deux chromatides sœurs. Des protéines appelées cohésines forment des complexes en anneaux assurant l'attachement de ces chromatides soeurs entre-elles [Klein et al., 1999]. Parallèlement, c'est aussi lors de cette phase que se produit l'alignement des chromosomes homologues. L'alignement commence par l'attachement des télomères à l'enveloppes nucléaire et leur regroupement dans une zone correspondant, sur la face cytoplasmique, à une machinerie cellulaire appelée "spindle body" chez *Saccharomyces cerevisiae* [Gerton & Hawley, 2005]. Des expériences de microscopie à fluorescence sur des mutants *spo11* de *Schizosaccharomyces pombe* ont montré que les étapes précoces de cet alignement sont indépendantes de la formation des DSB, qui commence aussi à ce stade (Figure I.5). L'association des télomères à l'enveloppe nucléaire se fait grâce à des protéines des familles SUN-KASH [Tzur et al., 2006] qui interagissent avec des dynéines du "spindle body". Cette interaction serait à l'origine de mouvements des chromosomes dans le noyau qui permettraient l'alignement en augmentant la probabilité de rencontres des homologues. Ces mouvements favoriseraient spécifiquement l'association des homologues tout en évitant l'association des non-homologues [Koszul & Kleckner, 2009]. A la fin de ce processus, les chromosomes adoptent une forme en "bouquet" dans le noyau (Figure I.5).

- Zygotène : La formation des bivalents passe non seulement par l'alignement des chromosomes homologues, mais aussi par leur accolement

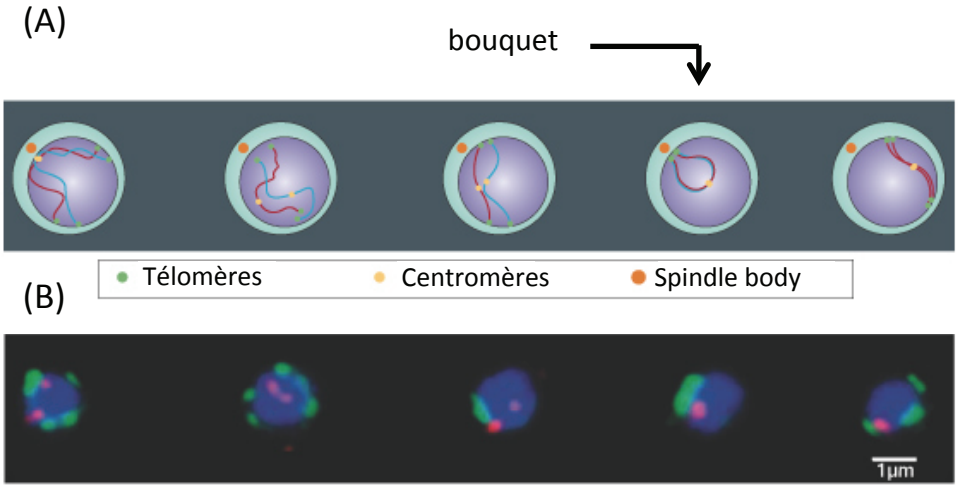



FIGURE I.5 : (A) Modèle de formation du "bouquet" dans des cellules de *Saccharomyces cerevisiae* en méiose. (B) Observations de noyaux de *Saccharomyces cerevisiae* en méiose, au microscope confocal. L'ADN est coloré au DAPI (bleu) et permet de délimiter l'espace nucléaire. Les télomères sont marqués au FITC (vert) et une zone témoin spécifique de chaque chromosome XI est marquée à la rhodamine (rouge). Sur la troisième image on observe la localisation des télomères sur l'enveloppe nucléaire puis, sur l'image suivante, la colocalisation des régions homologues du chromosome XI indiquant l'alignement. Enfin la dernière image montre la dissolution du bouquet et la persistance de l'alignement. Adapté de [Gerton & Hawley, 2005].

intime au sein du complexe synaptonémal. Un des "défis" pour la cellule est ici d'accoler chaque chromosome avec son homologue et non un autre segment d'ADN.
Chez les eucaryotes, il existe deux modalités majeures d'appariement des chromosomes homologues. La première est dépendante des DSB. En effet, des mutants de *spo11* de *Saccharomyces cerevisiae* , *Arabidopsis thaliana* et de souris montrent une baisse forte de nombre de synapsis en prophase I [Romanienko & Camerini-Otero, 2000, Grelon et al., 2001, Ding et al., 2010]. Lors de la formation des DSB, les extrémités 5' sont digérées par une exonucléase laissant des brins 3' libres rapidement associés aux pritéines Dmc1 et Rad51 [Neale & Keeney, 2006]. La protéine Rad51 est ensuite impliquée dans la recherche du brin homologue (*cf.* I.A.3. p 32). Ainsi, chez ces espèces, c'est le couplage du synapsis avec la recherche d'homologie lors de la réparation des DSB qui permettrait un bon appariement des chromosomes homologues.
A l'inverse, chez d'autres espèces comme le ver *Caenorhabditis elegans* ou la mouche *Drosophila melanogaster*, le synapsis se fait indépendamment de la recombinaison. Chez *Caenorhabditis elegans* des centres d'appariements ont été identifiés le long des chromosomes. En méiose, ceux-ci se lient spécifiquement à des protéines à doigt de zinc de la famille ZIM. L'enchainement de ces protéines créé alors un code barre le long du chromosome, permettant la reconnaissance des homologues entre eux [Phillips & Dernburg, 2006]. Chez *Drosophila melanogaster*, on a longtemps cru que le problème de l'appariement méiotique n'était pas posé car les chromosomes étaient observés comme "pré-appariés" de manière somatique [Metz, 1926]. Cependant, des recherches récentes ont remis en question cette affirmation [Christophorou et al., 2013]. Enfin, le rôle des centromères dans l'appariement indépendant des DSB a été mis en évidence chez ces deux invertébrés et *Schizosaccharomyces pombe* [Karpen et al., 1996, Dernburg et al., 1998, Ding et al., 2004].
A la fin du zygotène, le complexe synaptonémal est assemblé en certains points des bivalents. Ce complexe présente une structure tripartite qui connecte de manière serrée (à moins de 100nm) les chromosomes homologues (Figure I.6). Les éléments latéraux sont liés aux deux chromatides de chaque chromosome prématurément, puis autour de cela s'assemblent les éléments centraux et transverses. La propagation du complexe le long des chromosomes se fait à partir de points de nucléation tels que les télomères ou les segments concernés par des DSB. Elle procède par assemblage des éléments transverses et centraux à la manière d'une fermeture-éclair liant les homologues [Page & Hawley, 2004].

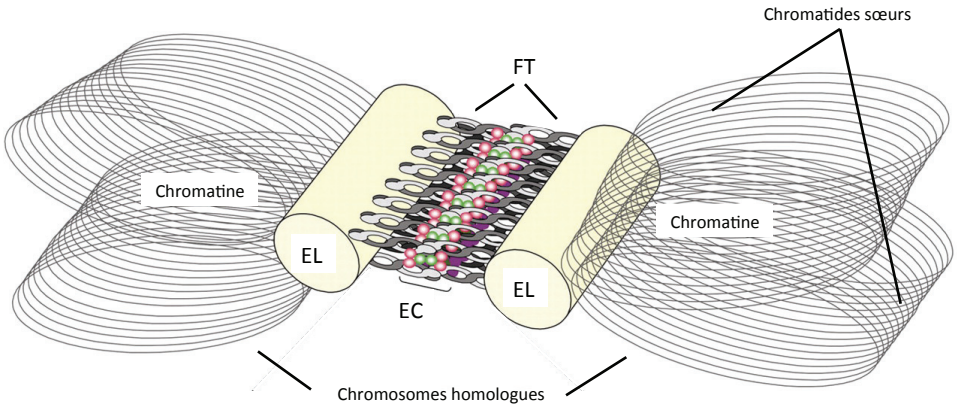



FIGURE I.6 : Modèle moléculaire du complexe synaptonémal. Les deux parties du complexe sont : (EL) les éléments latéraux (dont Rec8, SCP2 et SCP3) et (EC) l'élément central, formé de l'association des filaments transverses (FT ; dont Zip1, SCP1 et SYP1). Adapté de [Costa et al., 2005].

- Pachytène : C'est à cette étape que la réparation des DSB a lieu (voir I.A.3. p 32). Elle peut être longue : jusqu'à deux semaines dans les spermatocytes humains.
- Diplotène : Les bivalents commencent à se séparer par dissolution du complexe synaptonémal. C'est à partir de ce stade que sont observés les chiasmas. Il s'agit aussi d'un stade intense en synthèses, notamment dans l'ovocyte des mammifères formés lors du développement fœtal. Ces ovocytes sont alors bloqués en diplotène jusqu'à la puberté où certains d'entre eux achèveront leur méiose de manière régulière à chaque cycle menstruel.
- Diacinèse : Elle correspond à la disparition de l'enveloppe nucléaire qui précède la séparation des chromosomes homologues en deux lots lors des phases plus tardives de la méiose I : métaphase, anaphase et télophase (Figure I.4).

### Le rôle de la recombinaison dans la ségrégation du matériel génétique en méiose

Les étapes suivantes de la méiose permettent une bonne ségrégation des chromosomes dans les quatre produits (gamètes ou spores) formés. Cependant, il arrive que ces processus soient défaillants et aboutissent à une mauvaise distribution des chromosomes créant des produits (i) nullisomiques : dont un

chromosome manque et (ii) disomiques : à deux chromosomes homologues (erreur en méiose I) ou à deux chromosomes issus des deux chromatides d'un même parent (erreur en méiose II). Chez l'homme, après fécondation les gamètes nullisomiques donnent des œufs monosomiques et les gamètes disomiques donnent des œufs trisomiques. Les individus atteints du syndrome de Down, ou trisomie 21, sont issus de ce type de gamètes : disomiques pour le chromosome 21. Ils sont sujets à des désordres cognitifs et à des malformations congénitales [Källén et al., 1996]. Dans 80% des cas, l'erreur vient de la méiose I mettant en lumière le potentiel rôle de la recombinaison dans la ségrégation correcte des chromosomes [Hassold & Hunt, 2001, Lamb et al., 2005, Székvölgyi & Nicolas, 2010]. En effet, les CO génèrent des tensions qui permettent aux homologues de s'aligner sur la plan équatorial en métaphase I (Figure I.4). Les cohésines liant les chromatides sœurs permettant, à contrario, de maintenir l'édifice autour des chiasma [Petronczki et al., 2003]. De fait, lorsque les DSB sont abolis, chez la majorité des eucaryotes, la ségrégation méiotique se fait au hasard aboutissant à de nombreuses erreurs. De plus, 40% des trisomies 21 dues à une erreur en méiose I sont associées à des bivalents sans CO, les autres cas correspondant à la formation d'un unique CO "mal placé", c'est à dire anormalement proche des télomères ou des centromères [Székvölgyi & Nicolas, 2010]. Des observations similaires ont été faites chez la levure [Rockmill et al., 2006]. Ceci suggère que la formation des CO est déterminante dans la ségrégation méiotique, une erreur pouvant avoir des conséquences pathologiques importantes. L'hypothèse communément acceptée est qu'un CO, au moins, est requis par paire d'homologue par méiose. Cette hypothèse est soutenue chez la levure où le nombre de CO par paire d'homologues est lié par une relation affine à la longueur des chromosomes avec une ordonnée à l'origine de 1,0 marquant cette contrainte mécanique [Mancera et al., 2008].

Ainsi, la recombinaison a lieu dans le contexte cellulaire de la méiose. N. B. : Dans le corps de ce travail nous n'aborderons pas le cas de la recombinaison homologue mitotique qui est encore mal connue et qui utilise, la plupart du temps, l'homologue situé sur la chromatide sœur plutôt que celui du chromosome homologue ce qui n'aboutit généralement pas à de la conversion génique [Pâques & Haber, 1999] (*cf.* article de revue sur le sujet : [LaFave & Sekelsky, 2009] et chapitre IV p. 181). De plus, la recombinaison méiotique répond à la formation de DSB programmée alors que la recombinaison mitotique répond à la formation spontanée de DSB (lésions de l'ADN). Les modalités et degrés d'imbrication de la recombinaison à la méiose varient selon les espèces malgré l'apparente nécessité d'engager un CO

par chromosome par méiose (en complément voir revues : [Gerton & Hawley, 2005, Zickler, 2006, Ding et al., 2010, Székvölgyi & Nicolas, 2010, Yanowitz, 2010]). Intéressons nous maintenant aux mécanismes de réparation des DSB qui forment le cœur même du processus de recombinaison.

## I.A.3. Quand les brins s'emmêlent : origine de la conversion génique

Les éléments présentés dans cette partie résultent principalement de l'étude de mutants et de profils de migration d'intermédiaires nucléiques de la recombinaison sur gel chez la levure, principalement *Saccharomyces cerevisiae* . Cependant, étant donnée la conservation des principales protéines impliquées, la plupart des résultats, sauf mention contraire explicite, est généralisable aux eucaryotes.

### Les étapes d'initiation de la recombinaison

La réparation des cassures double brins (DSBR, Double Strand Break Repair en anglais) est le modèle communément admis de fonctionnement de la recombinaison à l'échelle moléculaire [Szostak et al., 1983]. Celui-ci fait intervenir trois étapes principales concernant la phase initiale de la recombinaison. Premièrement, le DSB est formé grâce à un complexe comportant la topoisomérase de type II Spo11 initialement identifiée chez une archée [Bergerat et al., 1997, Keeney et al., 1997]. Celle-ci coupe l'ADN sans spécificité de motif (à part à l'échelle du nucléotide coupé [Murakami & Nicolas, 2009]) et reste attachée de manière covalente aux brins coupés (Figure I.7 (A)). Puis les brins sont excisés de 5' en 3' grâce à l'intervention du complexe Mre11-Rad50-Xrs2 [Borde & Cobb, 2009] (Figure I.7 (B)). Cela met à nu deux simples brins du chromosome dit "receveur" (en rouge sur la Figure I.7). Le chromosome intact étant dit "donneur" (en bleu sur la Figure I.7). Ces résections laissent donc une lacune dans le chromosome receveur. Les extrémités simple brins de celui-ci sont alors utilisées comme base dans la recherche du chromosome homologue. Cette étape (Figure I.7 (C)), appelée invasion de brin (SEI, Single-End Invasion en anglais), est favorisée par le rapprochement physique des homologues à moins de 100 nm au sein du complexe synaptonémal (*cf.* I.A.2. p. 25) et est permise par l'intervention de de deux recombinases : Rad51 et Dmc1 [Neale & Keeney, 2006]. Le SEI est aussi favorisé par les protéines Hop1 et Red1 qui font partie du complexe synaptonémal [Schwacha & Kleckner, 1997, Smith & Nicolas, 1998].

A ce stade, la recherche d'homologie peut-être infructueuse. En effet, elle peut-être interrompue si les molécules trouvées présentent trop peu de simila-

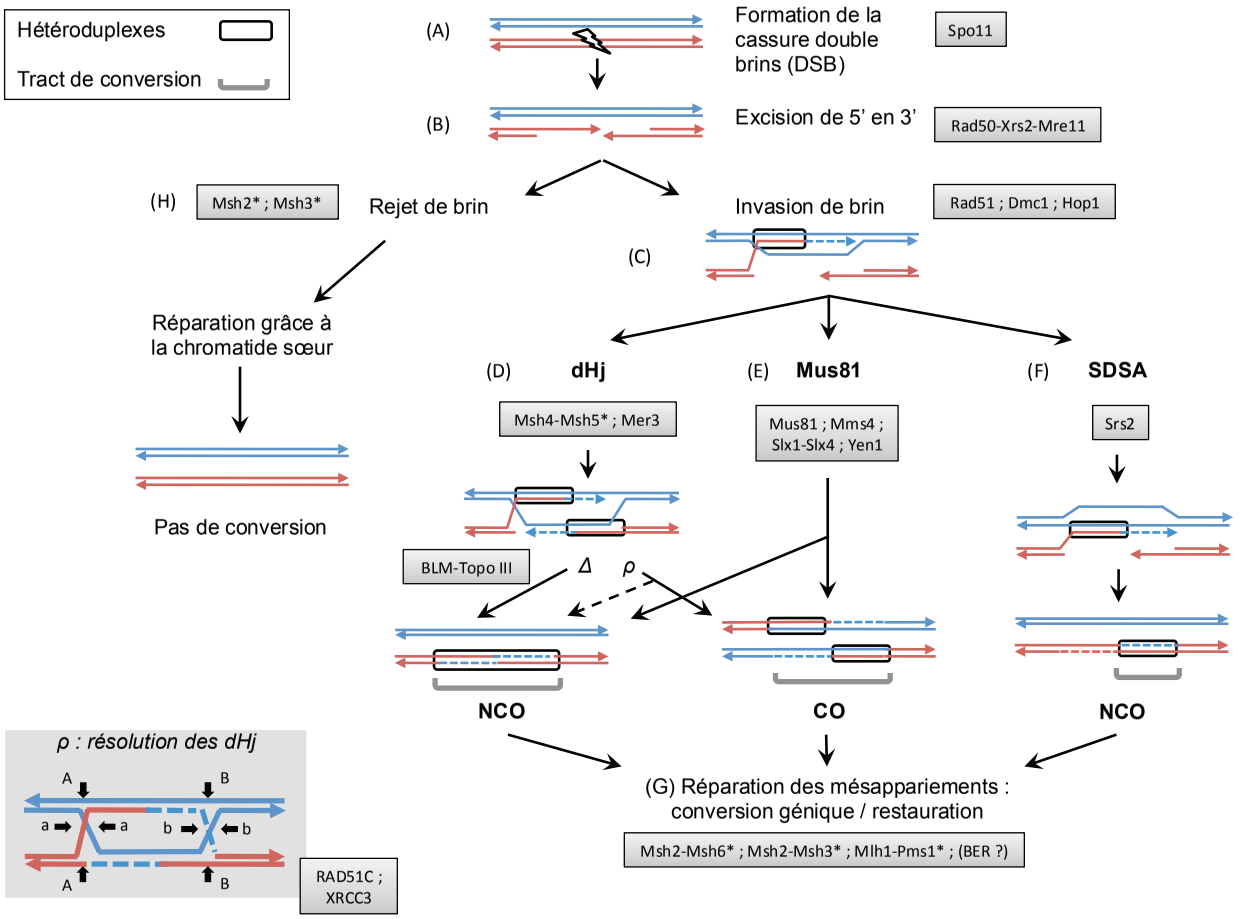



FIGURE I.7 : Modèle moléculaire des voies de recombinaison méiotique. Les flèches représentent les brins de deux chromosomes homologues (bleu : chromosome donneur, rouge : chromosome receveur) dans le sens 5' vers 3'. Seule une chromatide par chromosome est représentée. Les lignes pointillées représentent les segments néo-synthétisés. Les boites grisées mentionnent le nom des principales protéines impliquées dans ces voies chez *Saccharomyces cerevisiae* . L'étoile mentionne les protéines du système MMR. Les trois voies de réparation des DSB par recombinaison homologue sont mentionnées en gras. Les double jonctions de Holliday (dHj) peuvent être résolues ($\rho$) en CO (plus souvent) ou en NCO (plus rarement, *cf.* flèche en pointillés) ou dissoutes ($\Delta$) en NCO. L'encart grisé montre les points de coupure de résolution des dHj : A+b (ou a+B) donne un CO alors que a+b (ou A+B) donne un NCO. Voir texte principal pour plus de précisions.

rités avec l'extrémité libre associée aux recombinases. Les agents responsable de ce rejet de brin ont été identifiés, il s'agit de protéines du système de réparation des mésappariements nommé MMR (MisMatch Repair en anglais) [Datta et al., 1996, Evans & Alani, 2000, Surtees et al., 2004]. Plus les séquences sont divergentes, plus le rejet sera fréquent. Dans ce processus, les protéines Msh2 et Msh3, homologues de MutS (protéine du MMR bactérien), agissent comme des senseurs du degré de similarité des deux molécules. Plus précisément, il a été montré, dans ce contexte, que l'hétérodimère Msh2-Msh3 est capable de reconnaitre différents types de mésappariements : base-base ou petits indels [Nicholson et al., 2000]. D'autres études ont montré q'un segment de 20 pb montrant 100% d'identité) était requis dans une région homologue d'au moins 610 pb pour éviter le rejet de brin [Chen & Jinks-Robertson, 1998]. Ce système permet d'éviter la recombinaison ectopique (*i.e.* avec un segment non-homologue). En effet, la recombinaison ectopique peut-être délétère notamment lorsqu'elle aboutit à des réarrangements chromosomiques comme c'est le cas pour plusieurs pathologies telle que la maladie de Charcot Marie-Tooth [Székvölgyi & Nicolas, 2010]. Après le rejet, le DSB est réparé grâce à la chromatide sœur (Figure I.4 (H)) avortant ainsi le processus de recombinaison [Goldfarb & Lichten, 2010, Martini et al., 2011].

### Les voies de réparation utilisant l'homologue

- La voie dHj. Dans le modèle DSBR, les cassures sont réparées grâce à un intermédiaire faisant intervenir une double jonction de Holliday (dHj pour double Holliday junction) (Figure I.7 (D)). La dHj se forme lorsque les deux brins 3' libres sont associés à leurs homologues (*cf.* revue [Liu & West, 2004]). Les séquences manquantes du chromosome receveur sont alors toutes les deux synthétisées grâce aux brins homologues puis liguées ensemble ce qui forme deux jonctions de Holliday. Cette conformation est stabilisée par plusieurs protéines appelées ZMM [Börner et al., 2004] dont les protéines de la famille Zip et l'hélicase Mer3 capable d'étendre la région d'invasion de brin [Mazina et al., 2004] ainsi que deux protéines du MMR : Msh4 et Msh5 qui agissent en hétérodimère et permettent une stabilisation des Hj naissantes [Snowden et al., 2004].
  Les dHj sont ensuite "résolues" par des enzymes particulières appelées résolvases (Figure I.7 (D.$\rho$)) [Zakharyevich et al., 2012]. Selon le couple de brins coupés par les résolvases, les extrémités de chaque molécule sont échangées ou non donnant lieu respectivement à un CO ou un non crossing-over (NCO) (Figure I.7 (encart)) [Szostak et al., 1983]. Ce modèle prévoit donc un nombre égal de CO et NCO. Cependant,

chez des mutants [Bishop & Zickler, 2004] pour lesquels la formation des CO est bloquée, on n'observe pas de baisse du nombre de NCO. De plus, CO et NCO ne sont pas formés en nombre égaux [Székvölgyi & Nicolas, 2010] comme le prévoit le modèle DSBR. Il existe donc une voie indépendante des dHj qui produit des NCO (voir paragraphe suivant).
Plus récemment, on a découvert que les dHj peuvent aussi être "dissoutes" par l'intervention tardive (après la synthèse du second brin) de deux protéines : BLM et la Topoisomérase de type III [Wu & Hickson, 2003]. Cette voie conduit uniquement à la formation de NCO (Figure I.7 (D.Δ)).

- La voie SDSA. Dans ce modèle, le brin naissant qui participe au SEI est re-déplacé vers son origine : le chromosome receveur (Figure I.7 (F)) [Bishop & Zickler, 2004]. L'autre brin est alors réparé en utilisant son complémentaire comme matrice. Cette seconde réparation est donc dépendante de la synthèse du premier brin, d'où le nom de la voie SDSA (pour Synthesis-Dependant Strand Annealing). Parmi les enzymes qui participent à cette voie, on compte l'hélicase Srs2 qui permet le retour du premier brin sur le chromosome receveur [Ira et al., 2003]. Cette voie est sans coupure du chromosome donneur et ne donne donc que des NCO. Notons qu'elle semble très majoritaire dans la production des NCO chez *Saccharomyces cerevisiae* [Martini et al., 2011].

- Plus récemment, une seconde voie de mise en place des CO, indépendante de la voie ZMM-dHj, a été identifiée chez *Schizosaccharomyces pombe* (Figure I.7 (E)) [Osman et al., 2003]. Elle serait permise par le clivage de la boucle faite par le chromosome donneur lors du SEI, lui même catalysé par deux enzymes : Mus81 et Eme1 [Argueso et al., 2004]. Cette voie semble être la seule à produire des CO chez *Schizosaccharomyces pombe* [Osman et al., 2003] mais cohabite avec la voie ZMM-dHj chez *Saccharomyces cerevisiae* et *Arabidopsis thaliana* [Mercier et al., 2005, Whitby, 2005]. Elle est, par ailleurs, présente chez la souris [Holloway et al., 2008]. Enfin, contrairement à la voie ZMM, la voie Mus81 semble être exempte du phénomène d'interférence [Hollingsworth & Brill, 2004] (voir ci-dessous).

### L'interférence

Nous avons déjà vu comment le nombre de CO par méiose devait être contrôlé afin de garantir une bonne ségrégation des chromosomes homologues. Il

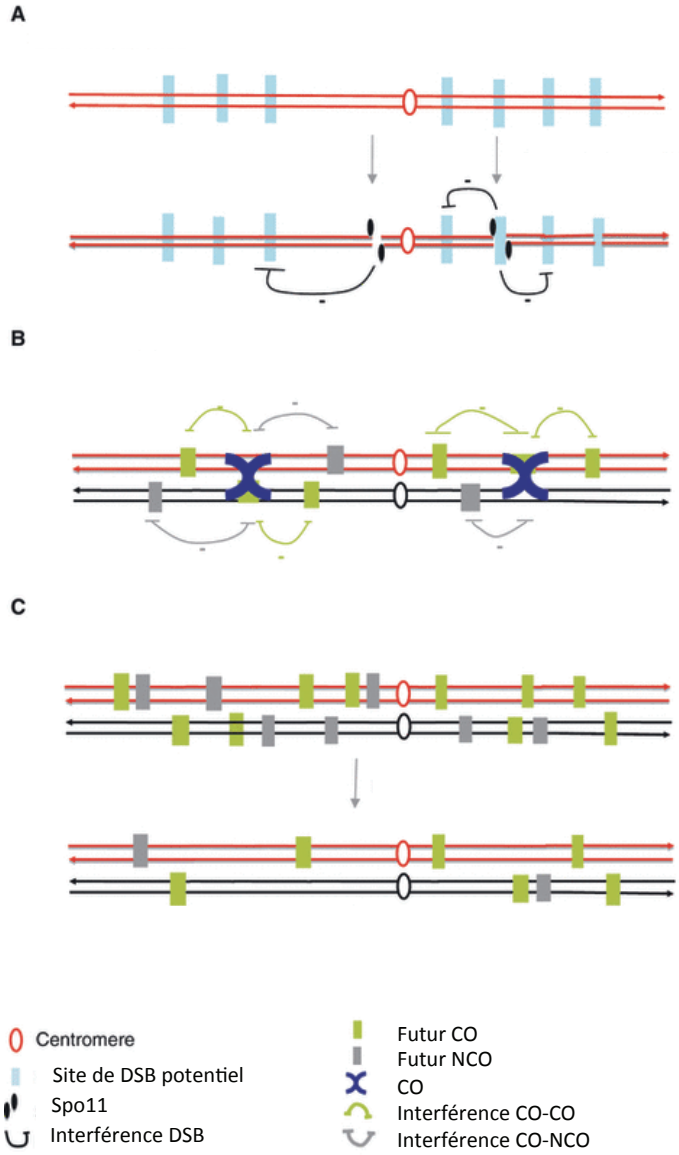



FIGURE I.8 : Les trois modes d'interférence : (A) L'interférence des DSB. (B) L'interférence des CO. (C) L'homéostasie des CO, notez comment les CO sont maintenus aux dépens des NCO. Adapté de [Székvölgyi & Nicolas, 2010].

semble cependant exister un deuxième niveau de contrôle sur le nombre et la répartition des CO lors de la méiose appelé interférence. On distingue trois types d'interférence (Figure I.8) :

(i) L'interférence des DSB : l'induction artificielle d'un DSB semble réduire le taux de DSB aux loci voisins [Robine et al., 2007].

(ii) L'interférence des CO : la formation d'un CO dans une région réduit la probabilité d'apparition d'un autre CO dans un voisinage pouvant excéder 100 Mb chez les Mammifères par exemple [Muller, 1925, de Boer et al., 2007]. Cette interférence est présente chez la plupart des espèces eucaryotes analysées jusqu'à maintenant. [Székvölgyi & Nicolas, 2010].

(iii) L'homéostasie des CO : elle permet un baisse du nombre de DSB en tout en garantissant la présence de CO qui se fait alors aux dépends des NCO dans une région donnée [Martini et al., 2006].

La répartition globale des taux de recombinaison est une question complexe que nous aborderons dans la partie I.B. p. 48.

### La conversion génique

Chez la levure, les spores issues d'une même méiose sont groupées en tétrades qui peuvent être disséquées. Chacune des spores peut ensuite être cultivée sur un milieu favorisant la reproduction clonale, par mitoses. C'est ce genre d'expériences qui a permis de mettre en évidence la conversion génique (Figure I.10) [de Massy et al., 1995, Surtees et al., 2004]. A un locus hétérozygote, lorsque, sur les quatre spores, deux portent un allèle et les deux autres l'allèle alternatif, la ségrégation est dite "2:2" (ou "4:4" si on compte les brins plutôt que les chromosomes) ou mendélienne. La conversion génique se produit à un locus lorsque cette ségrégation est déséquilibreée en faveur d'un haplotype : "3:1" (ou "6:2"). Dans ce dernier cas, un allèle (celui qui apparait en excès) a converti l'autre. La conversion génique correspond donc à un transfert unidirectionnel d'information génétique d'un chromosome dit "donneur" vers son homologue dit "receveur". A l'échelle moléculaire, la conversion génique se fait toujours du chromosome intact (n'ayant pas subi le DSB) vers le chromosome coupé, au niveau des zones concernées par des hétéroduplexes. Un hétéroduplex consiste en l'association de deux brins homologues. C'est typiquement ce qui se produit lors du SEI, mais aussi après synthèse et réassociation des brins du chromosome receveur (*cf.* cadres noirs de la Figure I.7). Les deux séquences de l'hétéroduplex n'étant pas, généralement 100% identiques sur toute leur longueur, il se créé des mésappariements. Dans la majorité des cas,

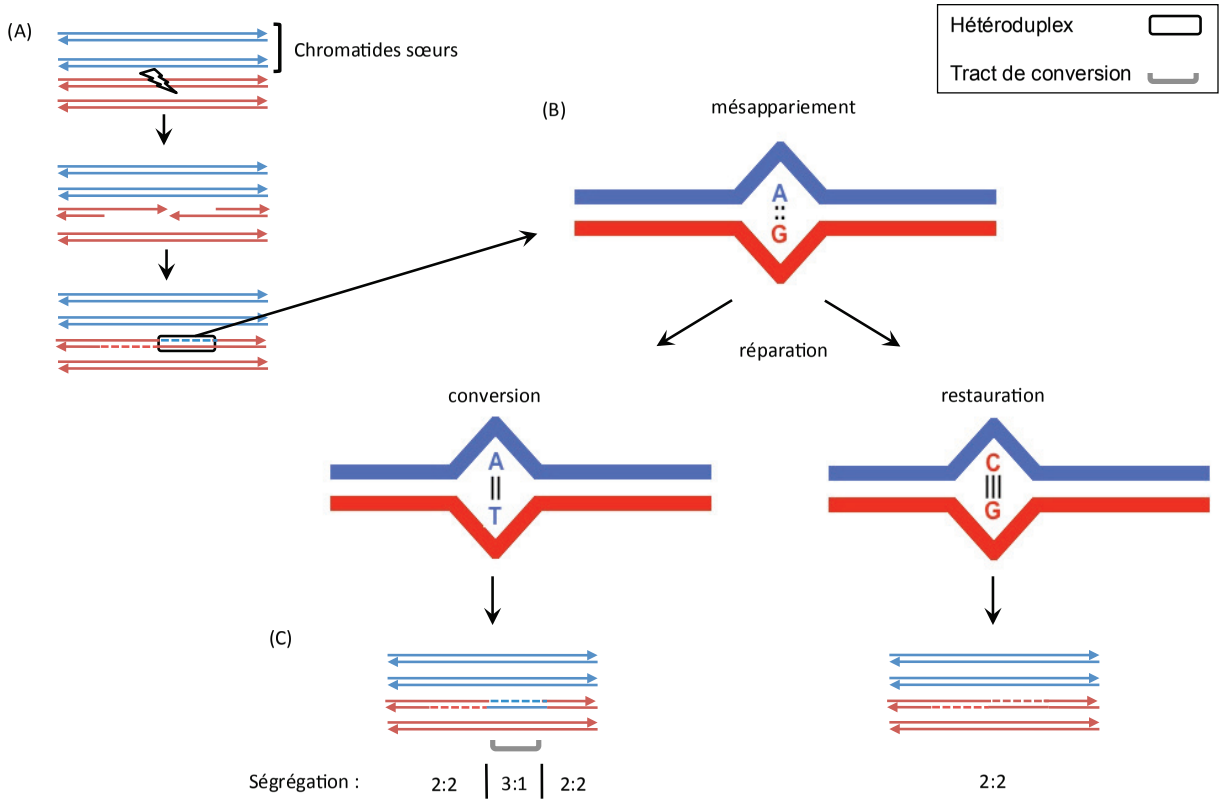



FIGURE I.9 : Conversion génique lors de la réparation des mésappariements d'hétéroduplexes. (A) Un hétéroduplex est créé dans la voie SDSA après formation d'un DSB sur l'une des chromatides du chromosome rouge (l'homologue est repérésenté en bleu). On a représenté les deux chromatides de chaque chromosome. (B) Zoom sur la réparation d'un mésappariement (A::G) de l'hétéroduplex. (C) Les profils de ségrégation au locus concerné dans les produits méiotiques en terme de nombre de chromosomes. La conversion génique est marquée par une ségrégation non-mendelienne (différente de 2:2) de type 3:1. La région concernée par la conversion est le tract de conversion (segment gris).

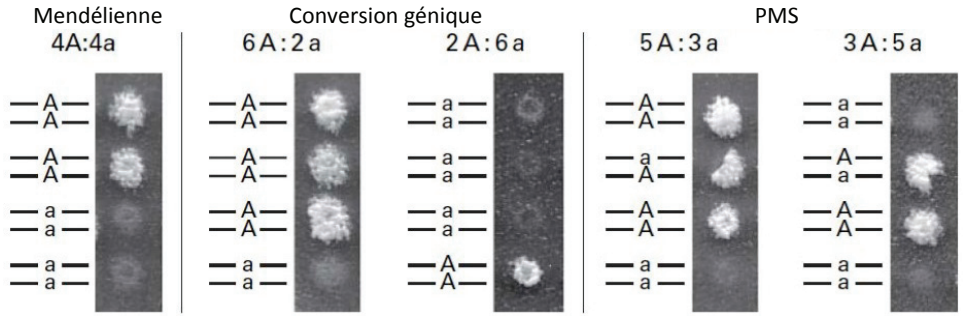



FIGURE I.10 : Example de ségrégations mendéliennes et non-mendéliennes au locus *arg4* chez *Saccharomyces cerevisiae* . Chaque paire de ligne représente les deux brins de la molécule d'ADN de la spore cultivée. "A" représente l'allèle qui confère le phénotype $Arg^+$ (colonies blanches sur les photographies) et "a" l'allèle mutante du phénotype $Arg^-$ (colonies sombres). Les rapports de ségrégation sont donnés en nombre de brins par tétrade. Les spores qui ont hérité d'un mésappariement non réparé a::A donnent des colonies mosaïques qui témoignent d'une ségrégation post-méiotique (PMS). Adapté de [Surtees et al., 2004].

ces mésappariements sont réparés (Figure I.9). Si la réparation se fait vers le génotype du chromosome coupé, on parle de restauration, la ségrégation 2 :2 est maintenue, il n'y a pas de conversion. Si la réparation se fait en faveur du génotype du chromosome intact, il y a conversion. Celle-ci est alors détectable sur un segment génomique donné appelé "tract de conversion" (N. B. : nous utiliserons directement le mot anglais "tract" pour décrire ces structures). Puisque les hétéroduplexes concernent toutes les voies de recombinaison (Figure I.7), les tracts de conversion peuvent être associés aux CO comme aux NCO. Notons, de plus, que le tract est dit "simple" si, pour un même évènement de recombinaison (CO ou NCO) tous les mésappariements des hétéroduplexes sont convertis dans le même sens (*i.e.* pas d'alternance conversion / réparation). Dans le cas contraire on parle de tract de conversion "complexe" [Mancera et al., 2008]. Chez *Saccharomyces cerevisiae* , il arrive que la réparation ne se fasse pas. Les colonies issues des spores présentent alors un phénotype mosaïque et la ségrégation est "5 :3" (Figure I.10). La ségrégation se fait alors après la méiose, lors de la première mitose de la spore, on parle de ségrégation post-méiotique (PMS) [Surtees et al., 2004]. La PMS semble être associée plus spécifiquement à certaines voies de recombinaison car on la retrouve moins souvent associée aux CO qu'aux NCO [Székvölgyi & Nicolas, 2010].

La conversion génique dépend donc de deux processus : le choix de l'ho-

mologue où se produit la coupure détermine le sens de la conversion si elle a lieu et le sens de réparation des hétéroduplexes détermine si il y a conversion ou non. Nous aborderons les déterminants du choix de l'homologue dans l'établissement du DSB dans la partie suivante (*cf.* I.B.2. p. 52). Les mécanismes de réparation des hétéroduplexes en méiose ont été étudiés via l'analyse des taux de conversion et de PMS chez des mutants de *Saccharomyces cerevisiae* . L'idée généralement admise est que c'est le système MMR qui assurerait cette tâche [Surtees et al., 2004] avec des modalités proches de celles mises en œuvre par ce même système lors de la réplication. Les mésappariements seraient reconnus par les hétérodimères Msh2-Msh3 et Msh2-Msh6 homologues de MutS du MMR bactérien [Kirkpatrick et al., 1998]. Le signal de réparation est ensuite transmit par l'intermédiaire du dimère Mlh1-Pms1 (homologues du MutL bactérien) à une exonucléase encore non identifiée [Kolodner & Marsischky, 1999]. Il se forme alors une lacune qui est réparée grâce à une polymérase [Surtees et al., 2004]. Ainsi, les modalités du choix du brin dégradé sont mal connues. Contrairement à son homologue bactérien, le MMR eucaryote ne semble pas posséder d'endonucléase. Le choix du brin pourrait donc découler mécanistiquement de la présence d'une cassure simple brin d'un côté ou de l'autre du mésappariement, sur l'un des deux brins. Cette cassure serait utilisée par une exonucléase telle que EXO1 (identifiée chez *Mus musculus*) pour amorcer la dégradation [Wei et al., 2003]. Ce processus de réparation est au centre du modèle proposé dans nos résultats (*cf.* II p. 101).

Dans les cellules dont le MMR a été aboli, on observe une forte augmentions de PMS : 40 à 90% des évènements [Alani et al., 1994, Hunter & Borts, 1997]. Cependant, les autres mésappariements sont réparés laissant penser que d'autres mécanismes que le MMR peuvent jouer un rôle dans la conversion génique tel que le mécanisme de réparation par excision de base (BER, Base Excision Repair en anglais) [Coïc et al., 2000]. Une des différences entre MMR et BER réside dans la longueur des patches réparés autour du mésappariement. Le MMR est un système de patches longs : plusieurs centaines de paires de bases [Jiricny, 2006] alors que le BER est un système de patches courts : 1-13pb [Memisoglu & Samson, 2000]. Une autre différence est que le BER possède des endonucléases qui lui permettent d'exciser spécifiquement les bases mésappariées avec certaines préférences [Marais, 2003] (*cf.* II p. 101). N.B. : Un autre système de patches courts, le NER (Nucléotide Excision Repair en anglais) semble, lui, ne pas affecter les mésappariements d'hétéroduplex méiotiques [Coïc et al., 2000].

### Les tracts de conversion

Les données sur la longueur et la morphologie des tracts proviennent principalement d'études de cartes de recombinaison de levure à haute résolution. Une carte de recombinaison est un outil qui fait correspondre une mesure de liaison génétique avec une distance physique pour un grand nombre de marqueurs (de nos jours principalement des SNP ou indels courts) à l'échelle du génome entier ou d'un chromosome. Elle permet ainsi de calculer des taux moyens de recombinaison sur des fenêtres de taille variable, le long du génome. Il existe, de nos jours, plusieurs façons d'étudier les taux de recombinaisonque nous détaillerons dans une série d'encarts "Méthode" tout au long de ce premier chapitre.

**Méthode**

**Les cartes à haute résolution**

Depuis la fin des années 2000, les progrès techniques réalisés par le génotypage sur puces ont permis de mettre en évidence directement les évènements de recombinaison, à l'échelle du génome entier, chez la levure. Le principe repose sur le croisement de deux souches haploïdes de levures présentant un certain degré de polymorphisme. Ce polymorphisme doit être assez important pour qu'il y ait suffisamment de marqueurs le long du génome ce qui garantit la haute résolution. Il ne doit pas être trop grand cependant car alors la recombinaison serait inhibée par rejet de brin (voir plus haut). Pour la carte de Mancera et collaborateurs [Mancera et al., 2008], par exemple, les deux souches présentaient environ 52000 marqueurs interrogeables par puce et répartis uniformément le long du génome ce qui aboutit à une résolution de 78pb (distance inter-marqueurs médiane). Cette carte est donc en moyenne 20 fois plus précise que les cartes de recombinaison classiques de levure et 360 fois plus que celles établies chez l'homme. L'autre intérêt majeur de ces cartes est qu'elles sont établies par dissection de tétrades ce qui veut dire que l'on connait le génotype des quatre spores provenant d'une même méiose, ce qui est impossible chez l'homme car les spermatozoïdes ne restent pas associés après leur formation. Ceci permet de détecter précisément les CO mais aussi les NCO ainsi que les tracts de conversion associés. Il s'agit donc aussi de cartes de conversion. Pour la carte de Mancera et collaborateurs, 51 méioses ont été analysées ce qui fournit, à notre connaissance, le premier et le seul outil, à ce jour, permettant une étude statistique de la conversion génique (CO et NCO) à l'échelle du génome entier. *En bilan sur les techniques d'étude de*

*la recombinaison, voir Annexe A p. 219.*

Chez *Saccharomyces cerevisiae* , on compte de 140 à 170 DSB par méiose [Buhler et al., 2007]. Parmi les évènements de recombinaison détectables, 58% sont des CO et 42% des NCO [Mancera et al., 2008]. D'autres études, menées chez la souris indiquent que la proportion de NCO serait beaucoup plus grande que celle des CO ($\approx$ 10 fois plus) chez les mammifères à certains loci [Cole et al., 2012a]. Ceci serait confirmé à l'échelle du génome entier par comparaison du nombre de DSB et de CO [Cole et al., 2012b]. Cependant, il est impossible de détecter directement les NCO à l'échelle du génome entier chez les mammifères car cela demanderait le génotypage d'un nombre trop important de SNP dans un nombre trop important de cellules. Enfin, sur deux loci à haute fréquence de recombinaison étudiés chez *Arabidopsis thaliana*, un présentait des proportions comparables de CO et NCO alors que l'autre présentait 30 fois plus de CO que de NCO [Drouaud et al., 2013]. Chez *Saccharomyces cerevisiae* , globalement, 1% du génome est sujet à la conversion génique à chaque méiose [Mancera et al., 2008]. La taille médiane d'un tract de conversion varie significativement selon le contexte recombinatoire : les tracts associés à des CO ont une longueur d'environ 2kb contre 1,8kb pour les NCO. Ceci semble aussi vrai chez l'homme même si les tracts (CO et NCO) semblent globalement plus courts. Par exemple au locus *DNA3* les tracts associés à des CO mesurent en moyenne 460pb contre 55-290pb pour ceux associés aux NCO [Jeffreys & May, 2004]. De plus, chez *Saccharomyces cerevisiae* , environ 11% des CO sont associés à des tracts complexes contre seulement 3,4% des NCO [Mancera et al., 2008]. Ceci est en accord avec le fait que le SDSA, principale voie de formation des NCO, ne donne que rarement des tracts complexes. On ignore encore l'origine des tracts complexes. Comme évoqué plus haut, ils pourraient venir d'une réparation des hétéroduplexes par patches, alternant entre conversion et restauration. Ils pourraient aussi provenir de la résolution de dHj complexes comportant plusieurs hétéroduplexes comme suggéré par [Mancera et al., 2008].

Ainsi, à ce stade de notre exploration du processus de recombinaison (en complément voir les articles de revue suivants : [Smith & Nicolas, 1998, Krogh & Symington, 2004, Liu & West, 2004, Surtees et al., 2004, Székvölgyi & Nicolas, 2010]), il semble intéressant de lui donner une définition à l'échelle moléculaire, différente de celle vue plus haut. La recombinaison consiste donc à la

réparation des DSB méiotiques à l'aide du chromosome homologue. Cette réparation peut aboutir à la formation d'un CO ou d'un NCO. Ainsi, à l'échelle moléculaire, un évènement de recombinaison n'est pas toujours associé à la production de recombinants (cas des NCO) comme le suggère la définition historique.

## I.A.4. Le "meiotic-drive" de conversion : le BGC

Comme nous venons de le voir, la conversion génique correspond au transfert unidirectionnelle d'information génétique d'un chromosome à son homologue. A l'échelle d'une population de produits de méioses, il est possible qu'à un même locus hétérozygote, un allèle ait plus de chance que l'autre de convertir son homologue. Ceci aboutit donc à un biais en faveur d'un allèle à l'échelle de la population, c'est ce que l'on appelle la conversion génique biaisée (BGC, Biased Gene Conversion en anglais). Ce phénomène correspond donc à une forme particulière de distorsion méiotique (meiotic drive en anglais, pour revue voir [Zimmering et al., 1970]) associée spécifiquement à la conversion génique. Cette sous-partie a pour but de mettre en évidence, puis de présenter comment les mécanismes vus plus hauts peuvent conduire à ce phénomène. D'autres preuves (indirectes cette fois) de l'existence du BGC seront développées plus tard dans la troisième partie de cette introduction (*cf.* I.C. p. 73).

### Mise en évidence du BGC

**Méthode**

**Le sperm-typing**

Les résultats présentés ci-dessous proviennent principalement de l'étude ciblée de zones du génome (humain et murin, principalement) à haute activité recombinatoire appelées points chauds (*cf.* I.B.2. p. 52) par une méthode de génotypage de gamètes mâles : le sperm-typing [Li et al., 1988]. Afin d'utiliser cette technique, les points chauds doivent avoir été découverts *a priori* par d'autres méthodes (*cf.* I.B.1. p. 49). Cette technique consiste à recueillir un échantillon de sperme d'un mâle dont le génotype, au locus du point chaud, est connu. Ce locus doit contenir un maximum de marqueurs pour lesquels l'individu est hétérozygote. Les spermatozoïdes sont ensuite analysés par PCR avec des amorces spécifiques de chacun des marqueurs : un couple pour le premier haplotype et un

couple pour le second. Ainsi, lorsqu'un CO a lieu entre deux marqueurs, il y aura amplification spécifiquement si deux amorces complémentaires des deux haplotypes sont utilisées. Cette technique permet d'analyser rapidement un grand nombre de molécules recombinantes. Depuis son invention, de nombreux progrès ont été réalisés et le sperm-typing permet aujourd'hui de déterminer précisément la fréquence des CO et NCO à un locus donné (pour peu qu'il s'agisse d'un point chaud), la structure des tracts de conversion qui les accompagne ainsi que le taux et le sens de conversion d'un allèle [Kauppi et al., 2009]. De plus, récemment, une variante de cette méthode a vu le jour : le "pollen-typing" qui adapte le sperm-typing au modèle végétal : *Arabidopsis thaliana* [Drouaud et al., 2013]. *En bilan sur les techniques d'étude de la recombinaison, voir Annexe A p. 219.*

Le BGC a été directement observé lors de l'étude de points chauds de recombinaison humains. Grâce au sperm-typing (*cf.* encadré Méthode), le profil de transmission des haplotypes autour de 6 points chauds a été analysé en 2002 par Jeffreys et Neumann [Jeffreys & Neumann, 2002]. Parmi ces 6 points chauds, le locus *DNA2* du complexe majeur d'histocompatibilité (CMH, chromosome 6) montre un profil de transmission particulier : pour deux SNP situés dans le voisinage immédiat du site préférentiel de formation de DSB, l'un des allèles est retrouvé dans les tracts de conversions plus fréquemment que l'autre allèle (Figure I.11). L'intensité de ce biais peut-être mesurée directement : sur les échantillons analysés [Jeffreys & Neumann, 2002], dans un contexte de CO, l'allèle G du SNP FG11 a été transmis favorablement dans 76,4% des cas (proportions 76,4:23,6). Ceci représente un taux de transmission de l'allèle avantagé (notée $\gamma$) de 0.764. Afin de quantifier l'impact du biais à l'échelle de la population, il faut calculer l'avantage moyen conféré à ces allèles dans leur transmission d'une génération à l'autre : il s'agit de la fréquence moyenne de cet allèle dans le pool de gamètes issu d'un hétérozygote (noté $x$). Ceci dépend de l'intensité du biais de conversion (noté $\alpha = \gamma - 0.5$) et du taux de conversion à ce locus ($r = 3.7\ 10^{-5}$ CO par méiose dans le cas de *DNA2*, on néglige ici la contribution, potentiellement importante, des NCO par manque de données). La relation entre ces paramètres est donnée par l'équation (1) [Nagylaki, 1983].

$$x = \frac{1}{2}(1 + r\alpha) = \frac{1}{2}(1 + r(\gamma - \frac{1}{2})) \quad (1)$$

Mentionnons, dès à présent, que le produit $r\alpha$ est généralement noté $b$

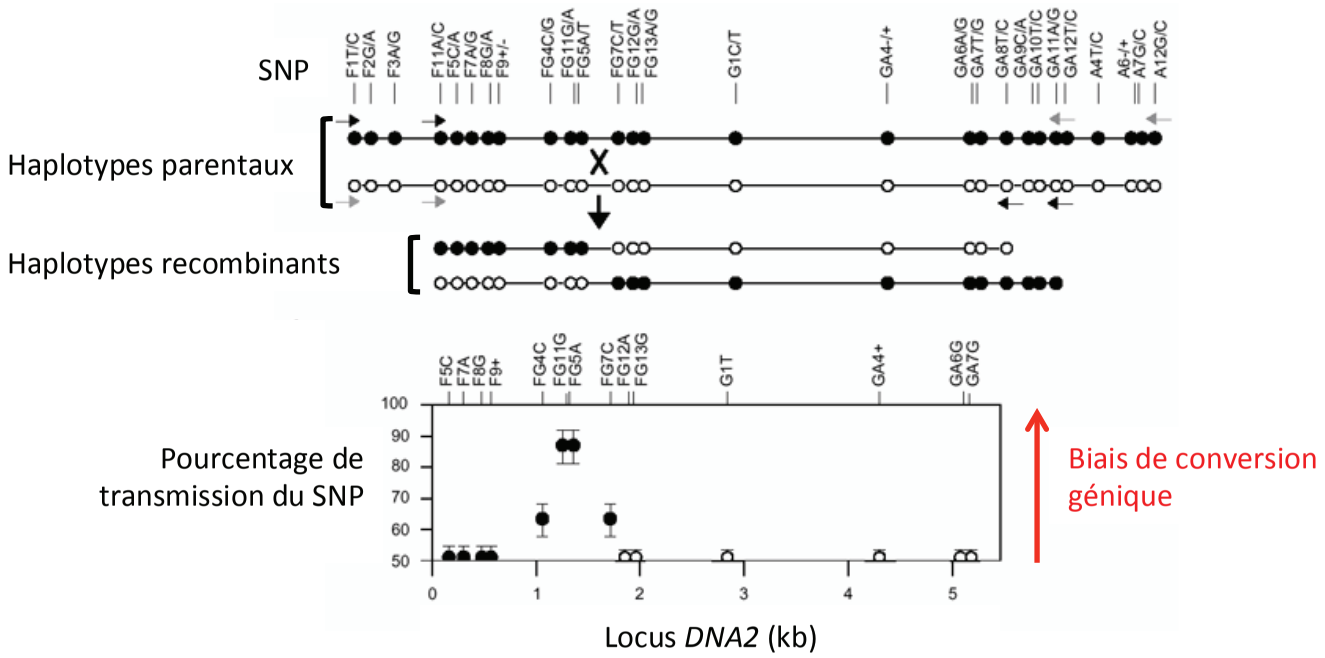



FIGURE I.11 : BGC au locus *DNA2* chez l'homme. La partie supérieure de la figure montre les deux haplotypes de l'homme étudié. Chaque cercle correspond à un polymorphisme (SNP). Le nom et les allèles de ces marqueurs sont indiqués au dessus. La partie inférieure de la figure représente le pourcentage de transmission de chacun des haplotypes, la couleur du cercle renvoie à l'allèle le plus transmit. Notez comment les allèles G et A sont transmises significativement plus souvent qu'attendu (50%) aux loci FG11 et FG5 respectivement. Adapté de [Jeffreys & Neumann, 2002].

(ou $g$) et qu'il correspond au coefficient de BGC (ou coefficient de disparité) utilisé en génétique des populations [Nagylaki, 1983]. Ainsi, $x = 50,00049\%$ dans le cas de *DNA2*. Ceci peut sembler faible mais est suffisant pour multiplier par 1,5 la probabilité de fixation de l'allèle G dans une population humaine de taille réaliste [Jeffreys & Neumann, 2002]. D'autres loci montrent un biais comparable chez l'homme : les point chauds *NID1* (chromosome 1) [Jeffreys & Neumann, 2005], *B* (chromosome 21), *J1* (chromosome 5) [Webb et al., 2008], *S1* et *S2* (chromosome 13) [Jeffreys & Neumann, 2009] ainsi que chez la souris [Wu et al., 2010].

### Le BGC d'initiation : dBGC

A tous les loci cités ci-dessus, l'allèle avantagé par le BGC semble aussi responsable de la détermination en *cis* du point chaud : l'allèle donneur est associé à un taux de recombinaison plus faible que le receveur. De plus, dans le cas de *NID1* par exemple, le biais de conversion semble aussi bien affecter les CO que les NCO. Ces deux observations suggèrent que le biais de conversion serait déterminé en amont de la séparation des voies spécifiques aux CO et NCO, au moment de la cassure. Ce modèle, que nous nommerons "biais d'initiation" ou "BGC d'initiation" (noté dBGC pour DSB-associated BGC), repose sur le principe que l'allèle favorisant le DSB (allèle chaud) initie plus souvent la recombinaison que l'autre allèle (allèle froid). Ceci aboutit à plus de conversions de l'allèle chaud vers l'allèle froid et cela sans que la balance conversion/restauration du MMR ne soit affectée [Jeffreys & Neumann, 2002]. Notons que dans la plupart des exemples ci-dessus, seule la conversion associée aux CO est détectée. Au locus *A3* chez la souris, la recombinaison est fortement déséquilibrée en faveur des NCO et la conversion associée à ces évènements est aussi biaisée que celle associée aux CO. Ainsi, si on ne considère que ces CO, l'avantage conféré à l'allèle froid par conversion biaisée est $x = 50,044\%$ mais atteint la valeur de $x = 50,23\%$ lorsque l'on y ajoute la contribution des NCO ce qui prédit la fixation de l'allèle froid en moins de 1200 générations. Ceci correspond à moins de 200 ans si l'on considère un temps de génération de 8 semaines [Silver, 1995]. Une des conséquences du dBGC est que les séquences qui promeuvent la formation de DSB et donc la recombinaison, sont susceptibles de disparaitre rapidement des populations sous l'influence de ce dBGC. C'est le "paradoxe des points chauds de recombinaison" [Boulton et al., 1997] que nous développerons dans la partie suivante (*cf.* I.B.2. p. 52).

| Locus | A1 | A2 | NCO:CO | $\gamma_{NCO}$ | $\gamma_{CO}$ | $\gamma_{NCO}$ vs. $\gamma_{CO}$ | $x_{NCO}$ | $T_{fix}$ |
|---|---|---|---|---|---|---|---|---|
| F | G | A | 1 :11 | 71%* | 51% | $6\ 10^{-4}$ | 50,0167% | 300 |
| K | C | T | 1 :1,3 | 68%* | 48% | $ND$ | 51,4400% | 95 |

TABLEAU I.1 : BGC spécifique des NCO à deux points chauds humains selon [Odenthal-Hesse et al., 2014]. Le taux de transmission de l'allèle "A1" ($\gamma$) par rapport à l'allèle "A2" est mesuré séparément pour les NCO ($\gamma_{NCO}$) et les CO ($\gamma_{CO}$). Lorsque ce biais est significativement différent de l'attendu (50%) cela est mentionné par une étoile. On donne aussi le rapport du nombre de NCO sur le nombre de CO (NCO:CO). La colonne "$\gamma_{NCO}$ *vs.* $\gamma_{CO}$" donne la p-value du test de $\chi^2$ comparant le BGC dans les deux types de contextes (ND : donnée non disponible). "$x_{NCO}$" donne la fréquence moyenne de l'allèle avantagé par le BGC associé aux NCO compte tenu du taux de NCO au locus concerné. Une estimation du temps de fixation de l'allèle avantagé est donnée ($T_{fix}$) en milliers d'années.

### Le BGC biaisé vers G et C : gBGC

La carte haute résolution (HR) de Mancera et collaborateurs est, pour le moment, le seul outil qui permet de mesurer le biais de conversion à l'échelle d'un génome entier sur un grand nombre de méiose [Mancera et al., 2008]. (Chez l'homme et *Arabidopsis thaliana* l'étude de produits méiotiques à l'échelle du génome entier est possible [Lu et al., 2012, Hou et al., 2013] mais pas à grande échelle ce qui ne permet pas une estimation précise du biais) Grâce à cette carte, les auteurs ont montré qu'il existait un bias global dans la conversion des bases W (A ou T) vers les bases S (G ou C), chez la levure. Ceci correspond à une fréquence $x = 50.065\%$ des allèles S dans le pool des gamètes. Ce résultat corrobore de nombreuses observations indirectes faites chez différentes espèces indiquant qu'il existerait un biais global dans la conversion des allèles le long du génome. Afin d'expliquer ces observations, un concept, nommé gBGC (BGC biaisé vers GC) a émergé. Nous détaillerons ces observations, ainsi que la genèse du concept de gBGC dans la partie I.C. p. 73. Retenons pour le moment que le gBGC est une forme de BGC qui favorise spécifiquement les allèles riches en GC par rapport aux allèles riches en AT et qu'il n'a pu être démontré directement que chez la levure jusqu'à maintenant. Les méthodes d'études actuelles ne permettent, en effet, pas d'étudier les tracts de conversion sur l'ensemble du génome chez les mammifères.

Récemment, Jeffreys et collaborateurs ont mis en évidence, chez l'homme,

deux points chauds de recombinaison, $F$ et $K$, soumis au BGC [Odenthal-Hesse et al., 2014] (Tableau I.1). Dans les deux cas le biais est important et favorise l'allèle S sur l'allèle W. Il pourrait donc s'agir de gBGC. De manière inattendue, le biais est observé uniquement dans les tracts associés à des NCO et non à des CO contrairement à ce qui est observé aux loci soumis au dBGC. De plus, les allèles désavantagées ici n'ont pas d'influence particulière sur le taux de DSB local. Il ne s'agirait donc pas d'un biais d'initiation (dBGC) mais potentiellement d'un biais dans la réparation de l'hétéroduplex [Odenthal-Hesse et al., 2014]. Il est donc possible que gBGC et dBGC aient des origines moléculaires différentes. Nous étudierons les mécanismes moléculaires du gBGC dans la première partie des résultats de cette thèse (*cf.* II p. 101) afin de mieux caractériser son origine évolutive et les conséquences potentielles que ce processus peut avoir sur les génomes.

Ainsi, dBGC et gBGC sont des processus intimement liés à la recombinaison méiotique. Afin de mieux caractériser l'impact de ces phénomènes sur l'évolution des génomes, il est donc utile d'étudier la dynamique de la recombinaison elle-même, c'est ce qui est proposé dans la partie suivante.

## I.B. La dynamique spatiale et temporelle de la recombinaison et du BGC

Afin de mieux décrire le BGC, il est important de décrire en détails deux aspects centraux de la recombinaison. Premièrement, sa distribution spatiale, c'est à dire comment les DSB, les CO et les NCO sont répartis le long du génome. Ceci permet d'identifier les "compartiments génomiques" dans lesquels le BGC est susceptible de jouer un rôle majeur. Par "compartiments génomique" nous entendons ici à la fois la position physique sur les chromosomes (bras courts, longs, centromères, télomères...) mais aussi le type de séquence touché (région codante, intergénique, promoteurs, éléments transposables, ...). Deuxièmement, il est primordial d'analyser les changements de cette distribution dans le temps afin de mieux appréhender la capacité du BGC à interférer avec la sélection dont la force varie elle aussi dans le temps-dans les différentes régions du génome. Ceci se fait grâce à la comparaison des profils de recombinaison entre espèces ou entre populations.

## I.B.1. Mesurer la recombinaison

Dans cette sous-partie, nous verrons trois techniques différentes permettant de reconstruire les taux de recombinaison le long des génomes.

**Méthode**

**Les cartes de DSB**

Cette technique s'effectue directement sur des cellules en méiose : a/$\alpha$ chez la levure ou spermatocytes chez la souris. Elle consiste à (i) isoler spécifiquement l'ADN ayant subi un DSB puis (ii) à déterminer sa position sur la génome. La mise en œuvre de ces deux étapes varie selon les plan d'expériences :

(i) La sélection de l'ADN enrichi en sites de DSB peut se faire par immunoprécipitation de chromatine (ChIP) grâce à des anticorps anti-Spo11 [Gerton et al., 2000] ou anti-DMC1 [Smagulova et al., 2011] ou par des méthodes chimiques d'isolement spécifique de l'ADN simple brin [Buhler et al., 2007, Khil et al., 2012].

(ii) La détermination des loci concernés se fait sur puces (ChIP-chip) [Gerton et al., 2000] ou, plus récemment, par séquençage haut-débit (ChIP-seq) [Smagulova et al., 2011].

Cette méthode a une bonne résolution (200pb à 1kb) et n'est pas conditionnée par la recherche de polymorphisme. L'inconvénient majeur est que l'on ne peut pas déterminer si les DSB vont être réparés en CO, NCO ou grâce à la chromatide sœur. De plus, elle semble difficilement adaptable à l'homme car elle nécessite l'extraction de spermatocytes. *En bilan sur les techniques d'étude de la recombinaison, voir Annexe A p. 219.*

**Méthode**

**Les cartes basées sur le déséquilibre de liaison (DL)**

Cette méthode permet de quantifier les taux de recombinaison par analyse du déséquilibre de liaison (DL) à partir de données de polymor-

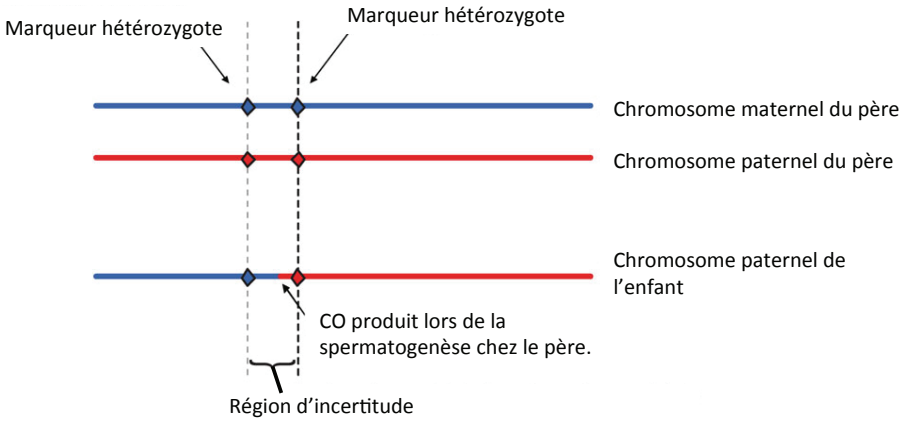



FIGURE I.12 : Etablissement d'une carte de recombinaison à partir de pedigrees. Les barres représentent une portion de deux chromosomes homologues du père (en haut) et du chromosome paternel de l'enfant (en bas). Les couleurs permettent de distinguer l'origine des portions recombinantes. Le site de formation d'un CO se trouve toujours entre deux marqueurs, à l'intérieur de cette "Région d'incertitude", on ne peut déterminer où le CO s'est exactement formé. Ainsi, plus la région est dense en marqueurs, meilleure sera la résolution. Adapté de [Kong et al., 2010].

phisme. L'histoire des haplotypes est alors reconstruite grâce à un modèle de coalescent dont la version la plus courante est distribuée sour le nom de "LDhat" [McVean et al., 2004, Auton & McVean, 2007]. Si deux allèles de deux marqueurs adjacents (SNP) sont systématiquement associés dans les données, le taux de recombinaison entre eux est prédit comme faible et inversement. Les données sont obtenues par séquençage [The 1000 Genomes Project Consortium, 2010] ou génotypage [The International HapMap Consortium, 2007] massifs ce qui permet l'analyse d'un grand nombre de marqueurs chez un grand nombre d'individus et donne donc une bonne résolution. Cependant, la précision atteinte avec le séquençage de "seulement" 10 individus non apparentés permet une étude fine de la répartition de la recombinaison le long du génome de la population considérée chez les primates [Auton et al., 2012]. Un inconvénient majeur est que lors de la reconstruction du coalescent, la sélection est un facteur confondant qui peut conduire à une estimation erronée des taux de CO [O'Reilly et al., 2008]. De plus, cette méthode ne donne accès qu'à un taux moyen de recombinaison (sur les deux sexes) résumant l'histoire recombinatoire des dernières générations correspondant à la diversification des populations étudiées. On parle alors de taux de recombinaison historique. Cette méthode a été appliquée entre autres chez l'homme (à partir de différents jeux de données) [The International HapMap Consortium, 2007, The 1000 Genomes Project Consortium, 2010], le chimpanzé [Auton et al., 2012] et la souris [Brunschwig et al., 2012], la résolution pouvant atteindre 2kb. *En bilan sur les techniques d'étude de la recombinaison, voir Annexe A p. 219.*

**Méthode**

**Les cartes basées sur l'étude de pedigree**

Cette méthode consiste à génotyper des parents ainsi que leur progéniture à l'échelle du génome entier et à déterminer les points, dans le génome de l'enfant, où l'on passe de l'haplotype du grand-père paternel (respectivement maternel) à l'haplotype de la grand-mère paternelle (respectivement maternelle), ou l'inverse. Ces points marquent la présence d'un CO (Figure I.12). Elle est communément réalisée sur des familles humaines et des hybrides de souris de laboratoire [Cox et al., 2009, Kong et al., 2010]. La

résolution de la carte est limitée par la densité en marqueurs polymorphes, d'où l'intérêt d'utiliser des souris hybrides. Parmi ces cartes, mentionnons celle du projet deCODE [Kong et al., 2010] pour laquelle plus de 38000 islandais ont été génotypés pour 300000 SNP. Ceci représente l'analyse de plus de 25000 méioses et aboutit à une résolution de 10kb. Malgré cette faible résolution relative, les cartes de pedigree présentent l'avantage d'être les seules qui permettent de mesurer séparément la contribution des recombinaisons mâle et femelle. *En bilan sur les techniques d'étude de la recombinaison, voir Annexe A p. 219.*

## I.B.2. Les taux de recombinaison le long des génomes

### Les points chauds

La description de l'intensité de la recombinaison le long du génome est très dépendante de l'échelle spatiale (*i.e.* en terme de nombre de paires de bases) à laquelle on se place (Figure I.13 (A)). A grande échelle (plusieurs Mb), il y a peu de variations, les taux de recombinaison sont distribués relativement normalement. Lorsque l'échelle se réduit, en dessous de 100kb, la distribution est déséquilibrée vers les faibles taux. Ainsi, il existe un grand nombre de longs segments à faibles taux de recombinaison et un petit nombre de courts segments à forts taux de recombinaison. La Figure I.13 (B) montre qu'une petite part du génome, environ 20% de la séquence totale, est responsable de 80% de l'activité recombinatoire du génome. Ceci est vérifié sur l'ensemble des chromosomes humains et met en évidence l'existence des points chauds de recombinaison [Myers et al., 2005] (Figure I.14). Les points chauds ont d'abord été identifiés dans le génome de *Saccharomyces cerevisiae* [Lichten & Goldman, 1995] et semblent être un trait commun à un grand nombre d'espèces eucaryotes : souris [Smagulova et al., 2011], chimpanzé [Auton et al., 2012] et plusieurs plantes [Mézard, 2006]. Cependant, chez la drosophile, et le ver *Caenorhabditis elegans* aucune structure comparable n'a été trouvée [Rockman & Kruglyak, 2009, Comeron et al., 2012]. Les points chauds sont donc des segments courts du génome qui concentrent une forte probabilité de former un DSB en méiose. Cependant, chez l'homme, aucune technique ne permet à ce jour d'identifier directement les DSB à l'échelle du génome entier. Des cartes de DL sont alors utilisées, elle permettent d'identifier des points chauds de CO uniquement. D'autre part, quelque soit la technique utilisée, la détection des points chauds dépend de seuils empiriques, il n'y a

donc pas de définition pratique et universelle d'un point chaud.
La caractérisation des points chauds dépend donc de l'approche utilisée. Ainsi, une carte de DSB a mis en évidence un peu plus de 1300 points chauds chez la levure[Buhler et al., 2007] soit 1 point chaud pour 10kb. Chez l'homme, les cartes de DL les plus récentes mettent en évidence plus de 30000 points chauds chez l'homme [The International HapMap Consortium, 2007] avec environ 1 point chaud tous les 50kb. Ces 30000 points chauds représentent environ 60% de la recombinaison humaine totale et 6% du génome avec une taille médiane de 5,5kb, proche d'une estimation faite chez la souris : 5kb (sur plus de 47000 points chauds) [Brunschwig et al., 2012]. Cependant, la précision de ces estimations est limitée par la résolution de la carte. Ainsi, les approches par sperm-typing prédisent que l'influence d'un point chaud de CO s'étend sur 1,2 à 1,9 kb environ [Webb et al., 2008]. Enfin, il existe aussi une diversité d'intensité des points chauds à l'échelle du génome. Les cartes de DL montrent que la distance génétique médiane induite par les points chauds humains est de 0,043 cM (1 CO pour 2300 méioses) mais atteint 1,2 cM (1 CO pour 80 méioses) pour le plus chaud d'entre eux [The International HapMap Consortium, 2007]. Ces estimations sont complétées par l'étude de plusieurs de ces points chauds par sperm-typing : a ces loci, les taux de CO extrêmes sont séparés par un facteur $\approx 900$ [Jeffreys et al., 2001, Jeffreys & Neumann, 2005, Berg et al., 2010]. Chez la souris, le point le plus chaud des points chauds de DSB pourrait atteindre 6 cM [Smagulova et al., 2011] ce qui montre que les NCO contribuent aussi grandement à l'activité de ces loci.

Les déterminants exacts de positionnement des points chauds le long du génome sont encore mal connus. Cependant, plusieurs corrélations ont été observées entre la présence de ces points chauds et différents paramètres génomiques locaux. Chez la levure, l'homme et la souris, la présence de points chauds est corrélée au contenu en bases G+C (ou GC génomique) calculé sur de petites fenêtres (environ 5kb), cette corrélation s'atténue lorsque le GC est calculé sur de plus grandes fenêtres [Gerton et al., 2000, Myers et al., 2006, Smagulova et al., 2011]. Ceci semble être directement imputable au gBGC comme nous le verrons dans la partie I.C. p. 73. D'autre part, il semble y avoir une association entre l'état de la chromatine et la présence de points chauds. Chez la levure, les DSB sont majoritairement trouvés dans les régions déplétées en nucléosomes des promoteurs de gènes [Pan et al., 2011] là où la chromatine est ouverte [Lichten, 2008]. Ils peuvent aussi être associés à l'absence de certains sites de fixation de facteurs de transcription [Pan et al., 2011]. Chez la souris, cela semble également être le cas [Baker et al., 2014]. Dans les deux cas, les points chauds sont favorablement associés à la marque épigénétique H3K4Me3 (tri-méthylation de la lysine 4 de l'histone 3

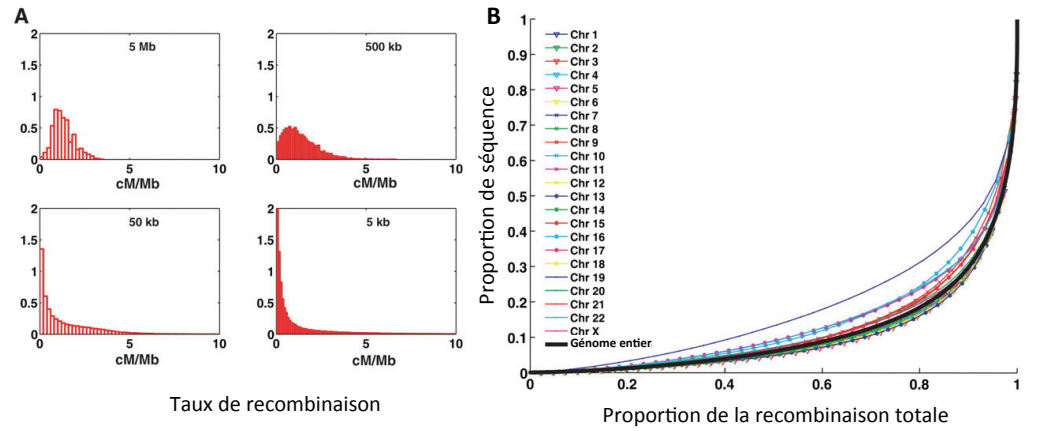



FIGURE I.13 : Hétérogénéité des taux de recombinaison dans le génome humain. (A) : Histogrammes des taux de recombinaisons en cM/Mb mesurés à différentes échelles. (B) La proportion de la recombinaison totale correspondant à la proportion de séquence concernée pour les différents chromosomes. Adapté de [Myers et al., 2005].

des nucléosomes) [Borde & Cobb, 2009, Buard et al., 2009]. Cette marque est communément associée, sur les promoteurs, avec la présence d'une activité transcriptionnelle chez la levure et la souris. Cependant, chez la souris, les points chauds sont associés à d'autres marques H3K4Me3 présentes à d'autres loci et spécifiques d es tissus testiculaires [Smagulova et al., 2011]. Ainsi, la marque H3K4Me3 ne semble pas suffisante pour expliquer la position des points chauds. De plus, une inversion de "skew" a été enregistrée autour des points chauds de souris [Smagulova et al., 2011]. Le skew correspond à un enrichissement en purines (A et G) en 5' du centre du point chaud et à un enrichissement en pyrimidines (C et T) en 3'. Ce skew pourrait être dû à un effet mutationnel de la recombinaison (*cf.* I.C.2. p. 82).

Les résultats les plus informatifs concernant les déterminants de positionnement des points chauds proviennent de la recherche de motifs ou d'éléments répétés associés à ces points chauds. Chez la levure, aucun contexte de la sorte n'a été clairement identifié si ce n'est que les éléments Ty (la famille d'éléments répétés la plus représentée chez la levure) semblent évités [Pan et al., 2011]. Chez l'homme, deux motifs de 7 et 9 pb (*CCTCCCT* et *CCCCACCCC*) ont d'abord été identifiés grâce aux points chauds mis en évidence par analyse du DL [Myers et al., 2005]. Cette découverte a été confirmée fonctionnellement par le fait que ces motifs forment le cœur des allèles qui déterminent le caractère chaud des loci *DNA2* et *NID1* respecti-

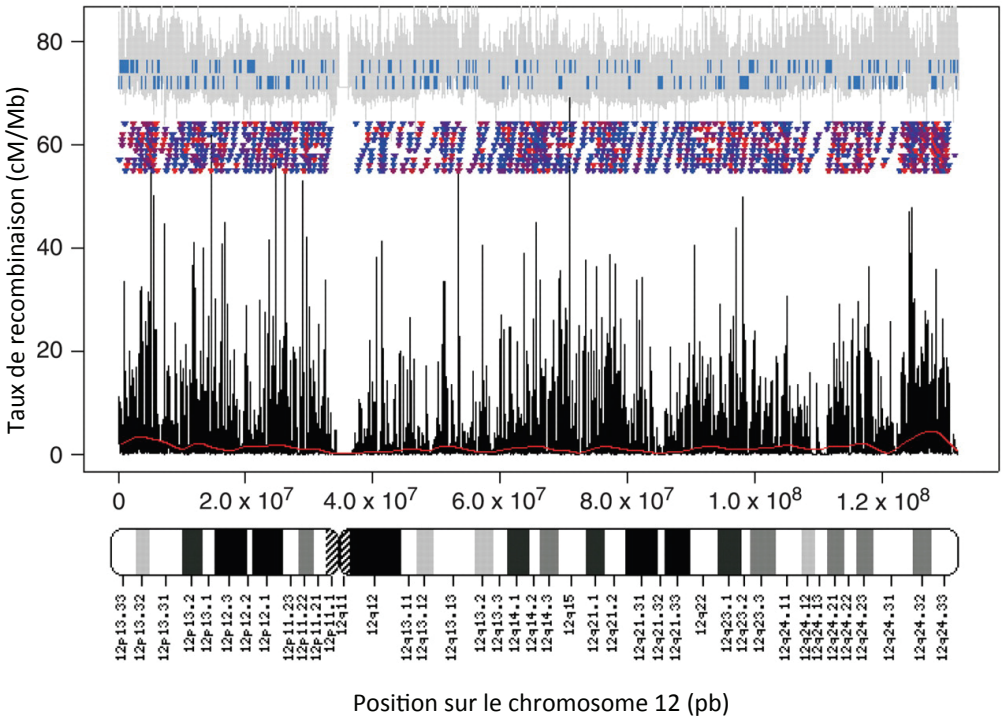



FIGURE I.14 : Taux de recombinaison déterminés par DL le long du chromosome 12 humain (en noir). La comparaison avec la carte basée sur les pedigrees de deCODE [Kong et al., 2002] est donnée par la ligne rouge. Observez la différence de résolution entre les deux cartes. Les triangles indiquent les points chauds détectés, la couleur donne leur intensité : de faible (bleu) à fort (rouge). Les barres bleues donnent la position des gènes de la base ENSEMBL. Les fluctuations locales de contenu en G+C génomique (fenêtres de 1kb) sont données en gris. Adapté de [Myers et al., 2005].

vement, caractérisés par sperm-typing [Jeffreys & Neumann, 2002, Jeffreys & Neumann, 2005]. Plus tard, le premier motif a été précisé : le 13-mer 5 fois dégénéré *CCNCCNTNNCCNC* semble impliqué dans 40% des points chauds [Myers et al., 2008] (Figure I.15). Le pouvoir prédictif de ce motif est plus marqué si il est lui-même inclus dans la séquence d'un des éléments de la famille de rétrotransposons THE1 (THE1A et THE1B). Par exemple, la présence du motif *CCTCCCTNNCCAC* dans son contexte THE1A correspond à un point chaud dans 73% des cas contre 10% hors de l'élément. Nous verrons dans la sous-partie suivante que ce motif correspond au site de fixation de la protéine à doigt de zinc PRDM9 qui a une grande influence sur la répartition des points chauds le long du génome. Sur la base de l'étude des homologues

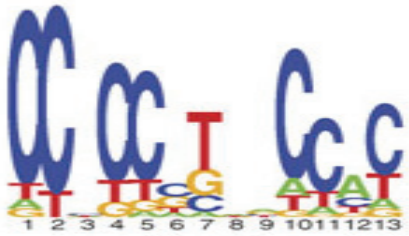


FIGURE I.15 : Séquence du motif de 13 pb associé aux points chauds humains. La hauteur des lettres montre la représentation relative de chacune d'elle dans les motifs associés aux points chauds. Certains sites (3, 6, 8, 9 et 12) semblent avoir un pouvoir prédictif inférieur aux autres. Adapté de [Myers et al., 2008].

de cette protéine, d'autres motifs, plus ou moins dégénérés ont été trouvés chez la souris [Smagulova et al., 2011] et le chimpanzé [Auton et al., 2012]. Chez le chimpanzé, outre les courts motifs *CGCG* et *CCCGGC*, l'élément répété MER92B et le microsatellite $(GGAA)_n$ sont significativement associés à la présence de points chauds. Ainsi aucun de ces éléments ne semble être conservé à l'échelle des eucaryotes et même des hominidés.

### Le paradoxe des points chauds

L'idée que les points chauds puissent être déterminés en *cis* par un motif, ou plus généralement une séquence particulière, pose un problème théorique quant à l'existence même de ces points chauds. En effet, si une séquence promeut la recombinaison en *cis*, elle sera plus fréquemment soumise à des DSB en méiose qui seront réparés grâce à la séquence homologue (Figure I.16). Ceci aboutit plus souvent à la conversion de la séquence chaude par la séquence froide que l'inverse. A l'échelle de plusieurs générations, l'allèle chaud est donc victime de dBGC qui tend à le faire disparaitre. Ainsi, l'existence même des points chauds conditionne leur disparition : c'est la paradoxe des points chauds [Boulton et al., 1997]. Un indice suggérant que ce processus est à l'œuvre chez l'homme a été observée dans l'évolution des motifs associés aux points chauds. En effet, il a été montré que le motif de 13pb associé aux points chauds humains dans le contexte THE1 était significativement plus perdu dans la lignée humaine que dans la lignée chimpanzé depuis la divergence entre les deux espèces [Myers et al., 2010]. Le motif n'étant pas associé aux points chauds chez le chimpanzé, cette observation suggère que c'est le dBGC agissant spécifiquement aux points chauds humains qui a entrainé l'excès de pertes du motif dans cette lignée. Le paradoxe des points

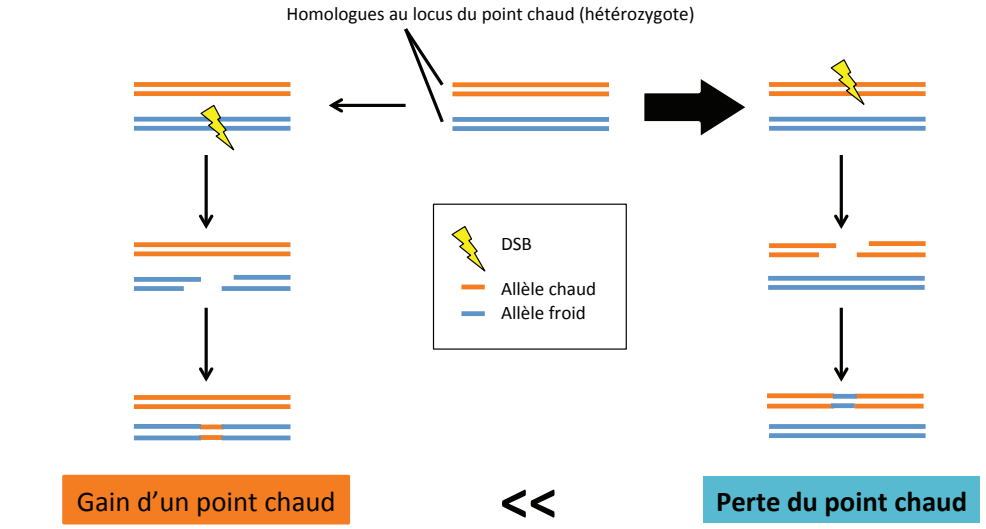



FIGURE I.16 : La paradoxe des points chauds. Les segments représentent les brins de deux chromosomes homologues au locus d'un point chaud de recombinaison. La séquence orange contient un déterminant du caractère chaud du point chaud (un motif par exemple). Son homologue, en bleu, ne le contient pas (motif muté par exemple). Le chromosome possédant l'allèle chaud initie donc plus souvent la recombinaison par formation de DSB (flèche épaisse) que son homologue. Cela aboutit à du dBGC en faveur de l'allèle froid. A l'échelle de la population, il y a donc perte du point chaud.

chauds est au cœur de la dynamique temporelle de ces structures que nous développerons dans la sous-partie suivante.

**La répartition de la recombinaison dans les compartiments génomiques**

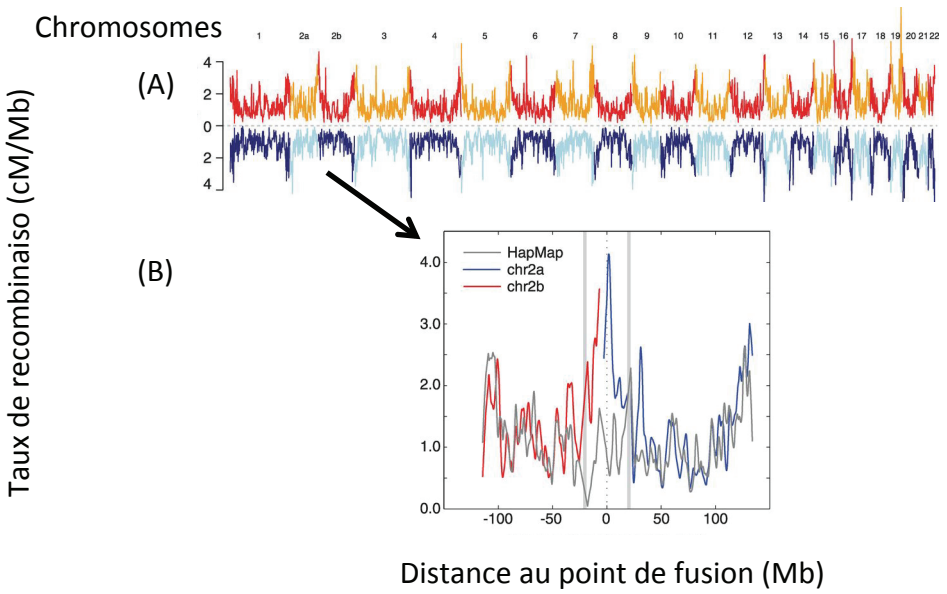



FIGURE I.17 : (A) Comparaison des taux de recombinaisons calculés par DL à grande échelle (1Mb) chez l'homme (bleus) et le chimpanzé (rouge/orange). Notez les forts taux de recombinaison aux télomères. (B) Taux de recombinaison calculés sur des fenêtres de 2Mb au point de fusion des chromosomes 2a et 2b. La ligne grise donne les taux estimés chez l'homme par DL grâce aux données HapMap. Les lignes verticales grises délimitent les zones sous-télomériques des chromosomes 2a et 2b qui correspondent à la zone de fusion chez l'homme. Adapté de [Auton et al., 2012].

A l'échelle des eucaryotes, il semble y avoir une tendance générale à la suppression de la recombinaison autour du centromère [Myers et al., 2005, Pan et al., 2011]. Les centromères étant formés de régions répétées, ceci pourrait permettre de limiter la recombinaison ectopique responsable de réarrangements potentiellement délétères. Chez la levure, cette suppression n'est plus perceptible 8 à 10kb autour du centromère [Buhler et al., 2007]. Les télomères de levures semblent aussi exempt de recombinaison [Buhler et al., 2007] alors que les régions sous-télomériques présentent de forts taux chez les primates [The International HapMap Consortium, 2007, Auton et al., 2012] (Figure

I.17 (A)). L'influence des télomères sur les taux de recombinaison s'observe particulièrement bien lorsque l'on compare le chromosome 2 chez l'homme et les chromosomes 2a et 2b du chimpanzé (Figure I.17 (B)). Le chromosome 2 humain résulte, en effet, de la fusion acrocentrique de deux chromosomes ancestraux restés intacts chez le chimpanzé (2a et 2b). A l'échelle du génome entier, les taux calculés sur une fenêtre de 1Mb sont bien conservés entre les deux espèces (Figure I.17). Cependant, on observe que la relocalisation intrachromosomique (*i.e.* hors des télomères), chez l'homme, de la région homologue aux télomères 2a et 2b du chimpanzé a conduit à une réduction du taux de recombinaison local à un niveau comparable à celui du reste du génome [Auton et al., 2012].

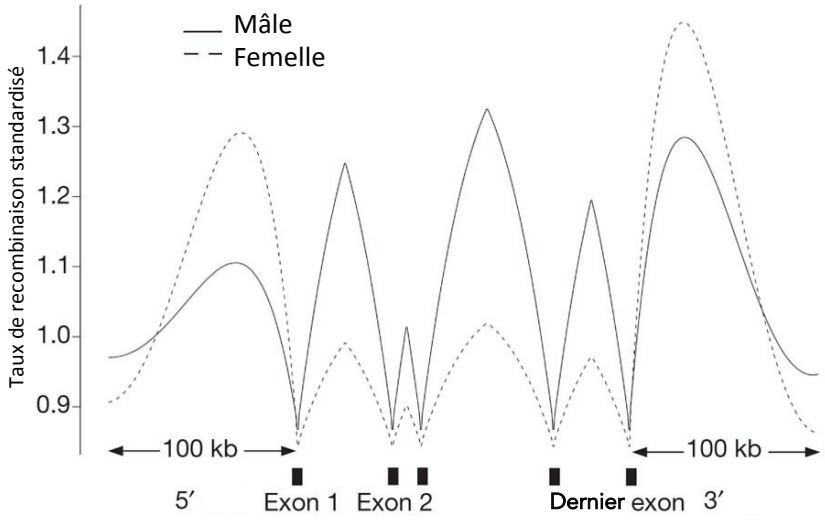



FIGURE I.18 : Représentation schématique des taux de recombinaison par sexe le long d'un gène. Le taux de recombinaison standardisé est obtenu grâce à la carte de pedigree de [Kong et al., 2010]. Adapté de [Kong et al., 2010].

Nous avons déjà vu que les points chauds sont concentrés dans les promoteurs chez la levure [Pan et al., 2011]. Il semble en être de même chez le chien [Auton et al., 2013] et *Arhabidopsis thaliana* [Choi et al., 2013]. Chez l'homme et la souris, la relation entre la recombinaison et les gènes semble plus complexe. Chez l'homme, le taux de CO est globalement réduit dans les régions transcrites mais on observe un pic dans la région 5' des promoteurs [The International HapMap Consortium, 2007] ainsi qu'en 3' des gènes [Kong et al., 2010]. Plus particulièrement, les régions codantes (exons) pré-

sentent des taux de recombinaison très faibles [Kong et al., 2010]. Ces résultats sont résumés sur la Figure I.18. On observe que la répartition moyenne de la recombinaison le long des gènes n'est pas la même selon les sexes (*cf.* paragraphe suivant). Chez la souris, à l'inverse, les points chauds semblent répartis plus favorablement dans les régions géniques qu'intergéniques pour une raison encore inconnue [Smagulova et al., 2011]. Notons aussi qu'il existe des différences significatives de taux de recombinaison entre loci situés dans le voisinage de gènes ayant des fonctions biochimiques différentes [The International HapMap Consortium, 2007].

#### Les variations des taux de recombinaison entre mâles et femelles

Les cartes basées sur les pedigree permettent de distinguer les contributions relatives des méioses mâles par rapport aux méioses femelles dans la répartition de la recombinaison le long du génome. Ainsi, chez de nombreuses espèces (homme, souris, chien, porc et *Arabidopsis thaliana*) les cartes génétiques femelles sont plus longues que les cartes mâles [Paigen & Petkov, 2010]. D'autre part, chez l'homme, Kong et collaborateurs ont identifié 4762 points chauds mâles et 4129 points chauds femelle dont 15% de chaque set semblent être spécifiques à la recombinaison mâle ou femelle [Kong et al., 2010]. Enfin, chez l'homme et d'autres mammifères, les régions sous-télomériques semblent affectées par plus de CO chez les mâles que chez les femelles [Popa et al., 2012]. Les hypothèses évoquées pour expliquer les différences de taux de recombinaison entre les sexes font le plus souvent intervenir le contexte cellulaire. Chez l'homme et la souris, on pense que le fait que les chromatides femelles soient moins compactées, et donc plus grandes, que les chromatides mâles favorise la formation de CO au sein du complexe synaptonémal [Paigen & Petkov, 2010]. D'autres hypothèses invoquent le fait que chez les mammifères, la recombinaison femelle est plus proche dans le temps de la fécondation que la recombinaison mâle. La phase haploïde mâle est donc plus longue que la phase haploïde femelle. La sélection naturelle étant plus forte lors de phases haploïdes, elle restreindrait d'avantage la recombinaison chez le mâle que chez la femelle et cela pour minimiser les effets délétères associés à ce processus (réarrangements, délétions...) [Lenormand & Dutheil, 2005].

### I.B.3. La dynamique temporelle des points chauds de recombinaison

Nous avons vu que la position des points chauds de recombinaison pouvait varier entre les sexes, cependant, des variations entre espèces ou populations

mettent en évidence le caractère dynamique de ces structures, à l'échelle des temps évolutifs, comme nous allons le voir dans cette sous-partie.

### Les variations interspécifiques des taux de recombinaison

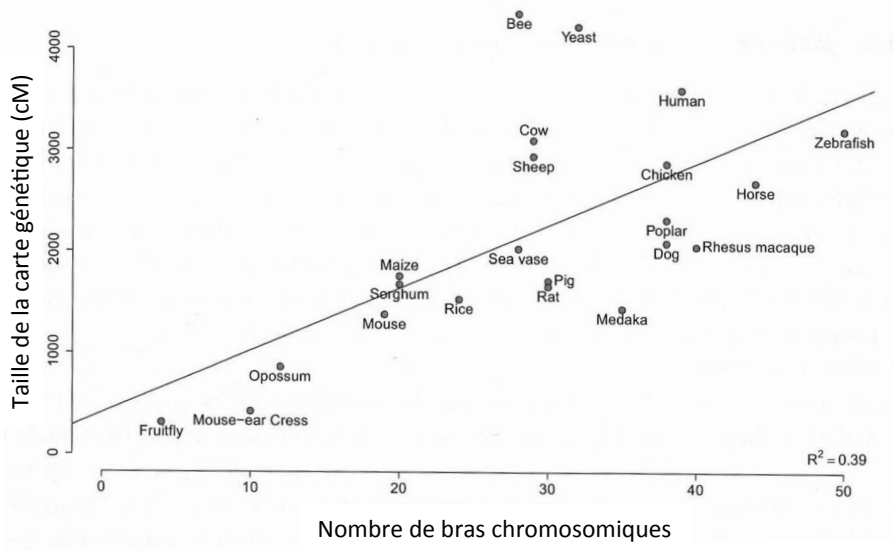



FIGURE I.19 : Corrélation entre la taille de la carte génétique et le nombre de bras chromosomiques chez différents eucaryotes. Le nom des espèces est indiqué à côté de chaque point, en anglais. Tiré de [Popa, 2011] et initialement adapté de [Coop & Przeworski, 2007] .

La multiplication du nombre de cartes génétiques disponibles pour différentes espèces a permis de montrer la diversité d'intensité de la recombinaison chez les eucaryotes. Ainsi, le taux moyen de CO à l'échelle du génome entier s'étend de 300 cM/Mb chez la levure à 0.05 cM/Mb chez certaines plantes [Awadalla, 2003]. Cette diversité est en grande partie expliquée la diversité des karyotypes. En effet, on observe une corrélation positive entre la longueur des cartes génétiques et le nombre de bras chromosomiques du génome [Coop & Przeworski, 2007] (Figure I.19). Ceci reflète la contrainte sur le nombre minimal de CO devant se former à chaque méiose : il faut, en effet, au moins 1 CO par chromosome pour assurer une bonne ségrégation du matériel génétique.

A plus petite échelle génomique, les variations de taux de recombinaison dépendent des groupes taxinomiques considérés. Ainsi, au sein du genre *Saccharomyces*, les taux de recombinaisons semblent être relativement conservés.

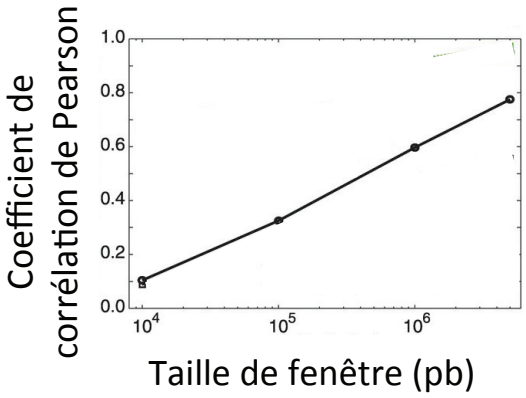



FIGURE I.20 : Coefficient de corrélation de Pearson ($r$) des taux de recombinaisons estimés entre l'homme et le chimpanzé pour différentes tailles de fenêtres. Adapté de [Auton et al., 2012].

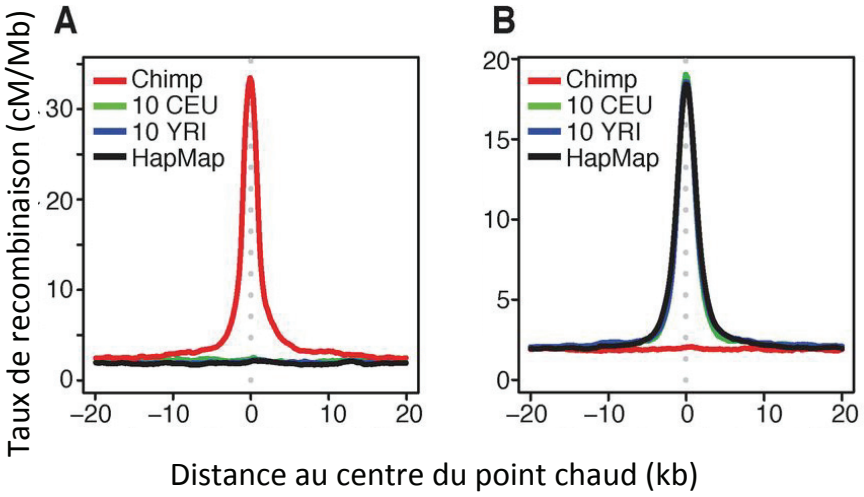



FIGURE I.21 : Profil des taux moyens de recombinaison mesurés au voisinage des points chauds du chimpanzé (A) et de l'homme (B). Les taux de recombinaisons sont calculés sur des fenêtres de 7,5kb à partir de cartes de DL établies chez le chimpanzé (rouge), l'homme (noir) et deux échantillons de deux populations humaines différentes : CEU (vert) et YRI (bleu). Adapté de [Auton et al., 2012].

En effet, la comparaison de la carte de DL du chromosome III de *Saccharomyces paradoxus* avec la carte de DSB et la carte haute résolution de ce même chromosome chez *Saccharomyces cerevisiae* montre que la localisation des points chauds chez ces deux espèces, séparées par 14% de divergence nucléotidique, est significativement plus recouvrante qu'attendu sous un modèle de distribution uniforme de ces sites [Tsai et al., 2010]. A l'inverse, chez la souris, il semble y avoir des différences significatives à petite et à plus grande échelle (de l'ordre du Mb) entre sous-espèces proches telles que *Mus musculus musculus* et *Mus musculus castaneaus* dont la divergence a eu lieue il y a moins de 450000 ans [Goios et al., 2007]. Chez les primates, une situation intermédiaire semble prévaloir entre l'homme et le chimpanzé séparés par 1% de divergence nucléotidique. En effet, comme suggéré plus haut (Figure I.17), on observe une corrélation importante des taux de recombinaison à l'échelle du mégabase [Auton et al., 2012] (Figure I.20). Cependant, cette corrélation baisse significativement lorsque l'on s'intéresse aux taux de recombinaison locaux (10kb). Ceci reflète le fait que la position des points chauds n'est pas conservée entre ces deux espèces malgré leur importante similarité de séquence (99%) [Ptak et al., 2005, Winckler et al., 2005, Auton et al., 2012] (Figure I.21). Les points chauds ont donc une certaine dynamique temporelle dont les causes ne sont pas encore totalement connues. Un élément de réponse se trouve dans le rôle de la protéine PRDM9 que nous développerons plus bas.

### Les variations inter et intra populations des taux de recombinaison

**Méthode**

#### Les cartes basées sur l'étude de populations métisses

Cette méthode n'a, à notre connaissance, été réalisée qu'une fois à partir de l'étude du polymorphisme détecté dans la population humaine métisse d'américains d'origine africaine [Hinch et al., 2011]. Elle tire profit de l'organisation particulière des haplotypes dans ce type de populations. Dans ces génomes, il y a une alternance de segments haplotypiques issus d'une des deux populations mères (population de l'Afrique de l'Ouest et population CEU dans notre exemple) entrecoupés de segments issus de la seconde population mère. A la jonction de deux segments d'origine différente, on prédit donc un évènement de CO ayant eu lieu depuis le métissage des deux populations. Il s'agit ici, comme pour les cartes basées sur le DL, d'une carte de CO uniquement qui reflète la moyenne des taux

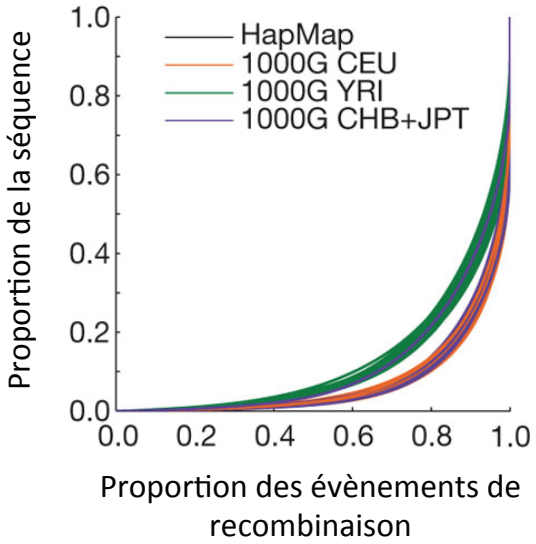



FIGURE I.22 : Concentration des CO le long du génome pour différentes populations humaines : CEU (orange), YRI (vert) et CHB+JPT (violet). La ligne noire correspond à la carte de DL basée sur les données de polymorphisme du projet HapMap. Adapté de [The 1000 Genomes Project Consortium, 2010].

de recombinaison mâle et femelle. Cependant, elle possède une bonne résolution : <3kb [Hinch et al., 2011]. *En bilan sur les techniques d'étude de la recombinaison, voir Annexe A p. 219.*

A l'échelle des populations humaines, la répartition de la recombinaison varie aussi sensiblement. Ainsi, les cartes de DL basées sur les données de polymorphisme du projet 1000 Génomes [The 1000 Genomes Project Consortium, 2010] montrent que la recombinaison est plus concentrée dans les populations ayant subi la sortie d'Afrique (CEU, CHB et JPT) que dans les populations africaines (YRI). En effet, si 80% des CO surviennent dans 10% du génome des populations CEU et CHB+JPT, seulement 70% ont lieu dans la même portion du génome de la population YRI (Figure I.22). Ceci a été confirmé par la construction d'une carte de recombinaison réalisée à partir de polymorphismes trouvés dans la population américaine d'origine africaine. Cette population étant le résultat d'un métissage d'une population d'origine européenne avec une population d'Afrique de l'ouest, le long du génome, le passage de l'haplotype d'une de ces populations à l'autre marque

un évènement de CO [Hinch et al., 2011]. La comparaison de cette carte AA (africain-américains) avec la carte de DL de la population européenne CEU [The International HapMap Consortium, 2007] met en évidence environ 2500 points chauds ($\approx 10\%$) spécifiques de la population d'Afrique de l'ouest [Hinch et al., 2011]. Cependant, la grande majorité des points chauds de la population CEU est présente dans la population AA.
De plus, il a été montré, grâce aux études de sperm-typing, que certains loci présentent un caractère de point chaud uniquement si ils portent un allèle particulier. Nous l'avons vu avec l'exemple des points chauds humains *DNA2* et *NID1* [Jeffreys & Neumann, 2002, Jeffreys & Neumann, 2005]. Ceci met en évidence une diversité intra population des patrons de recombinaison chez l'homme. Les causes de ces variations inter et intra populations des patrons de recombinaison seront détaillées dans le paragraphe suivant.

### Rôle de PRDM9 dans la dynamique des points chauds

Les différences de localisation des points chauds de recombinaison aux différentes échelles (espèces, populations, individus) montrent que ces structures sont très dynamiques chez l'homme et la souris. Chez l'homme, des études de génétique quantitative associées à la détermination de cartes génétiques personnelles dans des familles à larges cohortes issues de populations huttérites ont montré que les individus n'utilisent qu'une partie des points chauds "historiques" détectés par DL. De plus, la variabilité dans l'utilisation de ces points chauds est héritable ($h = 22\%$) [Coop et al., 2008]. Ceci suggère qu'il existe un ou des loci associés à l'utilisation de certains points chauds plus que d'autres. Ainsi, après la découverte du motif de 13pb associé en *cis* aux points chauds humains, plusieurs équipes ont tenté de trouvé un facteur *trans* gouvernant leur utilisation. Chez la souris, ceci a été fait par analyse de l'association entre des régions génomiques candidates et la mesure de l'activité des points chauds dans des lignées de laboratoires. Deux analyses de ce type ont montré, de manière convergente, que l'utilisation différentielles des points chauds est associée à une région génomique située sur le chromosome 17 [Grey et al., 2009, Parvanov et al., 2009]. Le génotype d'un locus de cette région semble influencer l'utilisation d'un point chaud situé sur le même chromosome : *Psmb9* [Grey et al., 2009] et d'un autre situé sur un chromosome différent : *Hlx1* [Parvanov et al., 2009]. Ainsi, il existe bien un facteur agissant en *trans* sur l'utilisation des points chauds. Le locus correspondant à ce facteur a été appelé *Dsbc1* [Grey et al., 2009] ou *Rcr1* [Parvanov et al., 2009].

Baudat et collaborateurs ont ensuite cherché un candidat parmi les loci annotés de la région génomique mise en évidence sur le chromosome 17 [Bau-

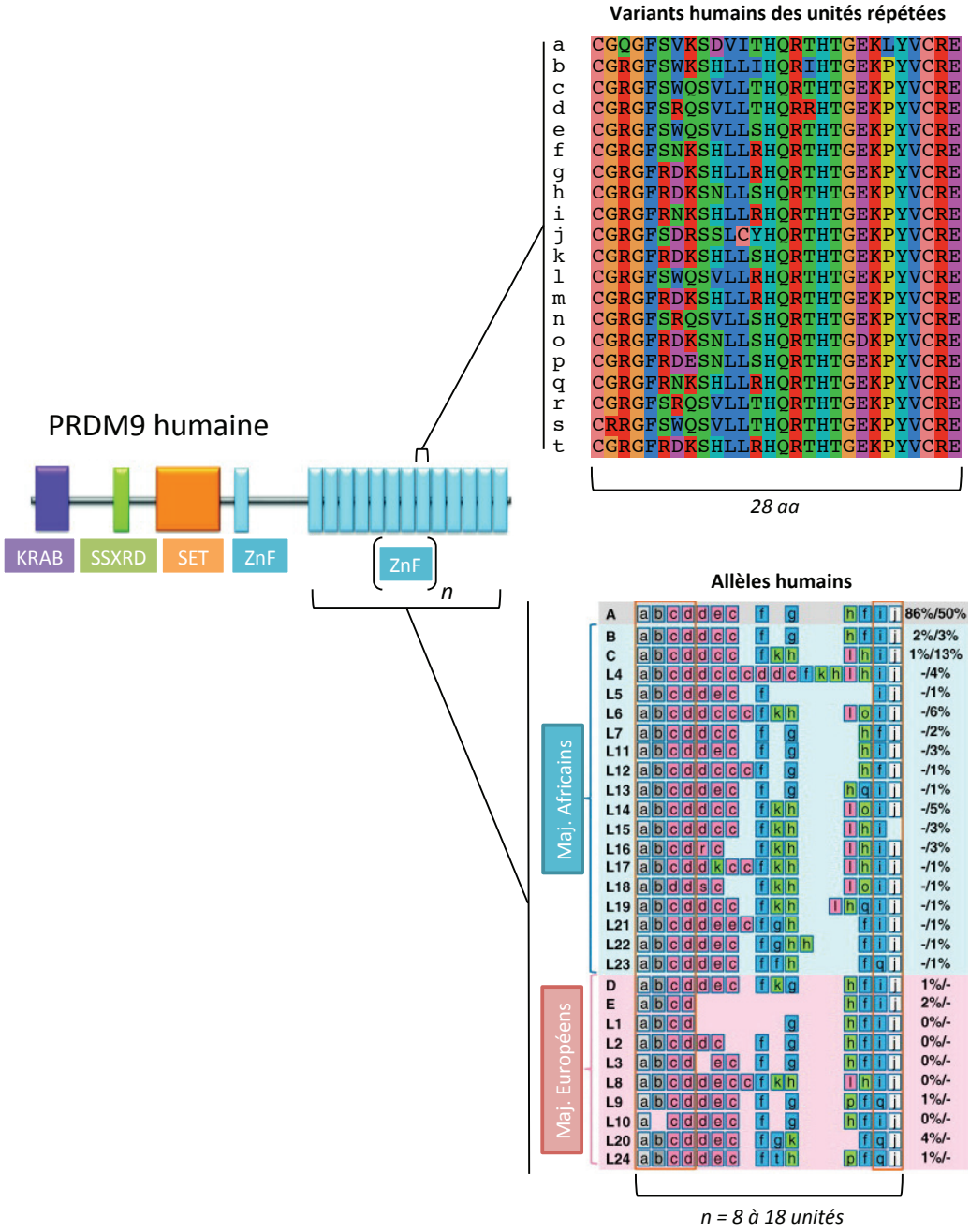



FIGURE I.23 : Structure du gène *PRDM9* et de ses variants humains. La partie gauche montre les principaux domaines protéiques codés par *PRDM9*. La partie supérieure droite montre les séquences protéiques (codes IUPAC) des variants connus de l'unité de 84pb (28 acides aminés) codant pour les ZnF [Berg et al., 2010]. La partie inférieure droite montre les différents allèles identifiés chez l'homme (lettres majuscules). Ils se différentient par le nombre ($n$) et l'enchaînement des différents variants du ZnF (lettres minuscules) : le "ZnF array". Les cadres oranges montrent la conservation des extrémités du ZnF array. La colonne de droite donne les fréquences alléliques dans la population européenne/africaine. Les allèles sur fond bleu (rose) sont majoritairement africaines (européennes). Adapté de [Ponting, 2011].

dat et al., 2010]. Le gène codant pour la protéine PRDM9 présente plusieurs caractéristiques compatibles avec une fonction liée au patron de recombinaison (Figure I.23) :

- Il est exprimé en prophase I de méiose spécifiquement.
- Il code pour un un domaine PRSET qui catalyse la triméthylation des histones H3 sur la lysine 4 aux points chauds *Hlx1* et *Psmb9*.
- Il code pour les domaines KRAB et SSRXD caractéristiques d'une activité nucléaire.
- Il possède une région exonique codée par un minisatellite dont l'unité de base est une séquence de 84pb codant pour un domaine à doigt de zinc (ZnF pour Zinc Finger en anglais). La séquence protéique résultant de cet enchaînement d'unités répétées est appelée "ZnF array". Elle est capable de se fixer spécifiquement sur certains motifs de l'ADN.

Le rôle de PRDM9 a ensuite été confirmé chez l'homme grâce à trois arguments principaux [Baudat et al., 2010] :

- L'allèle A étant le plus répandu dans l'ensemble des populations humaines : 90% des européens et 50% des africains (Figure I.23), on s'attend en effet à ce que la carte de DL définie à partir de ces populations reflète majoritairement l'effet de l'allèle A. Or les porteur homozygotes de cet allèle utilisent plus préférentiellement ces points chauds que les hétérozygotes A/I. L'allèle rare I (2% chez les Hutterites) possède un ZnF array différent de l'allèle A.
- Il est possible de prédire la séquence d'un motif d'ADN reconnu par un ZnF array à partir de sa séquence protéique. Ainsi, l'allèle A, à l'inverse de l'allèle I, est capable de reconnaitre le motif de 13pb trouvé dans 40% des points chauds définis pas DL [Myers et al., 2008] (Figure I.15).
- La fixation de l'allèle A sur le motif de 13pb est possible *in vitro*, à l'inverse de l'allèle I.

Au même moment, Myers et collaborateurs ont recherché des protéines codant pour des ZnF array capable de reconnaitre le motif de 13pb. Parmi les candidats, seul *PRDM9* présentait une séquence différente chez l'homme et le chimpanzé [Myers et al., 2010]. PRDM9 peut donc expliquer pourquoi les points chauds ne sont pas conservés entre ces deux espèces. Ceci a été confirmé par l'étude des allèles de *PRDM9* chez des chimpanzés d'Afrique

de l'ouest. Les ZnF array correspondant à ces allèles sont incapables de reconnaitre le motif associé aux points chauds chez l'homme [Auton et al., 2012].

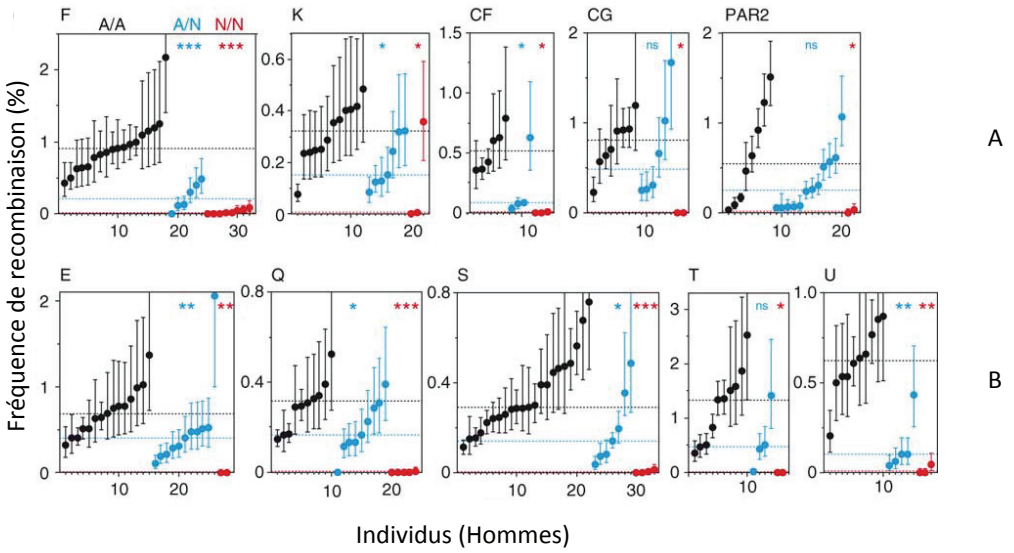



FIGURE I.24 : (A) Fréquence de recombinaison observée dans le sperme d'hommes homozygotes pour l'allèle A de *PRDM9* (noir), hétérozygotes A/N (bleu) où N est un allèle différent de A et homozygotes N/N (rouge) pour 5 points chauds identifiés par DL et contenant le motif de 13pb (Figure I.15). Les étoiles indiquent que la fréquence de recombinaison moyenne est différente à un locus donné entre A/A et A/N (bleu) ou A/A et N/N (rouge). (B) Idem pour 5 points chauds ne contenant pas de motif reconnaissable. Adapté de [Berg et al., 2010].

L'influence de PRDM9 a ensuite était testée par sperm-typing sur différents points chauds : avec ou sans motif de 13pb [Berg et al., 2010] (Figure I.24). Comme attendu, les homozygotes A/A activent les points chauds contenant le motif (Figure I.24 (A)) significativement plus que les hétérozygotes A/N et les homozygotes N/N (où N est un allèle différent de A). En outre, pour la plupart des loci un simple modèle additif permet d'expliquer la fréquence de recombinaison intermédiaire observée chez les hommes A/N. Notons que l'allèle A active aussi les points chauds ne contenant pas de motif reconnaissable (Figure I.24 (B)) ce qui suggère que PRDM9 peut influer sur l'activités de plus de points chauds qu'attendu, peut-être une majorité d'entre eux. De plus, il semble que le contrôle de l'activité des points chauds soit très sensible à la variation du ZnF array. Ainsi, les allèles L20 et L9, qui

ne diffèrent de A que par une substitution non-synonyme, n'activent pas respectivement le point chaud F contenant le motif et le point chaud MSTM1a ne contenant pas le motif, alors qu'ils sont tous deux activés par A. Ainsi, les variations de localisation des points chauds intra populations peuvent être expliquées par les variations d'allèles de *PRDM9*. Il en est de même pour les variations inter populations : le second allèle majoritaire chez les africains : C (Figures I.23) qui ne reconnait pas le motif de 13pb (il reconnait un motif plus long et plus dégénéré : *CCNCNNTNNNCNTNNC*) active les points chauds spécifiques à cette population [Berg et al., 2011]. Enfin, des études menées chez d'autres organismes permettent de différencier trois catégories de taxons, à l'échelle des métazoaires, pour lesquels l'activité ou l'état du locus *PRDM9* est en accord avec la dynamique des points chauds observée [Ponting, 2011] :

(i) Les espèces possédant des points chauds à courte durée de vie associés à un (ou plusieurs) locus *PRDM9* évoluant rapidement comme l'homme, le chimpanzé, la souris. On s'attend à un patron de recombinaison similaire chez le bœuf, le saumon atlantique et l'anémone étoilée chez lesquels *PRDM9* évolue rapidement aussi.

(ii) Les espèces possédant des points chauds à longue durée de vie associés à un locus *PRDM9* absent ou à ZnF array tronqué comme le chien (chez qui les centres des points chauds sont associés à de forts taux de G+C : 67% en moyenne en relation avec l'action prolongée du gBGC [Axelsson et al., 2012]). On s'attend au même patron de recombinaison chez le marsupial *Monodelphis domestica* chez qui le ZnF array de PRDM9 est absent.

(iii) Les espèces ne possédant pas de points chauds et un locus *PRDM9* absent ou à ZnF array tronqué comme le nématode *Caenorhabditis elegans*.

Il est aussi à noter que *Prdm9* a été identifié comme gène provoquant la stérilité hybride lors de croisements de *Mus musculus musculus* et *Mus musculus domesticus*. Cette stérilité est due à un dysfonctionnement de la spermatogenèse chez les hybrides [Mihola et al., 2009].

PRDM9 permet aussi d'apporter un élément de réponse au paradoxe des points chauds (*cf.* I.B.2. p. 56). En effet, on a montré que le ZnF array évolue rapidement chez l'homme et la souris sous l'influence de deux forces. Premièrement, le ZnF array est hypermutable, ceci est dû à sa structure en ministatellite qui favorise la réassociation des unités par recombinaison

inégale intra allélique en mitose et méiose [Jeffreys et al., 2013]. Ceci aboutit à une grande diversité d'allèles observée chez l'homme [Berg et al., 2010, Berg et al., 2011] (Figure I.23). Deuxièmement, le ZnF array semble évoluer sous sélection positive chez l'homme et la souris, plus particulièrement à trois positions impliquées dans l'interaction avec l'ADN : -1, 3 et 6 par rapport à l'hélice alpha de chaque ZnF [Oliver et al., 2009]. Ainsi, même si les cibles de PRDM9 (les motifs) sont perdues par dBGC, d'autres loci sont rapidement recrutés par de nouveaux variants de la protéine. Ce modèle d'évolution des points chauds par l'intervention de facteurs *trans* (PRDM9) et *cis* (les motifs) qui coévoluent est une variante du modèle de "Reine Rouge" [van Valen, 1973]. En effet, à un moment donné, les motifs ciblés par PRDM9 "s'échappent" des populations par dBGC, la sélection pousse alors PRDM9 à changer de cible. PRDM9 "rattrape" ainsi de nouveaux motifs qui, à leur tour, vont commencer à disparaitre des génomes (Figure I.25). Ce modèle d'évolution des points chauds a été initialement proposé par Myers et collaborateurs [Myers et al., 2010] puis développé formellement par Ubeda et collaborateurs [Ubeda & Wilkins, 2011]. Cependant, cette hypothèse a été mise en doute [Ponting, 2011] car le nombre de points chauds du génome chez l'homme (plus de 30000) surpasse largement le nombre de CO se produisant à chaque méiose (environ 60) et a donc peu de chance d'être limitant.

Ainsi, l'origine de cette pression de sélection qui pousse PRDM9 à changer de motif cible et aboutit au "déplacement" des points chauds est encore inconnue. Plusieurs hypothèses ont été avancées et rapportées dans [Ponting, 2011] :

- A l'échelle du génome entier, le nombre de points chauds ne semble pas limitant pour assurer la présence d'un CO par bras chromosomique à chaque méiose sauf pour les étroites régions pseudo-autosomales des chromosomes X et Y (PAR, moins de 3Mb chacune), qui sont les seuls segments de ces chromosomes où des CO peuvent se former. Ainsi, PRDM9 serait poussé à changer de cible pour permettre une bonne ségrégation des chromosomes sexuels en méiose. Cette hypothèse nous semble être remise en question par le fait que les points chauds des régions PAR ne sont pas abolis chez des mutants de souris $PRDM9^{-/-}$ [Brick et al., 2012]. La ségrégation des chromosomes sexuels serait donc indépendante de PRDM9.

- La recombinaison permet la disjonction d'allèles délétères et facilite donc leur élimination par la sélection purificatrice. De même elle permet l'association des allèles avantageux et facilite ainsi la sélection adaptative, ce sont les effets Hill-Robertson [Hill & Robertson, 1966]. Ainsi,

FIGURE I.25 : La Reine Rouge et Alice. Le nom de "Reine Rouge" appliqué à un modèle d'évolution vient de la nouvelle de Lewis Carroll "De l'autre côté du miroir", suite de "Alice au pays des merveilles", dans laquelle la Reine Rouge et Alice sont lancées dans une course sans fin qui ne mène nulle part ailleurs que là où elles se trouvaient initialement, ce qui vaut la remarque suivante à la Reine Rouge : "*Nous courons pour rester à la même place.*"

l'accumulation de mutations délétères dans une région génomique pourrait pousser à la sélection de variants de PRDM9 permettant l'établissement d'un point chaud dans cette zone. Il s'agit donc d'une sélection de second ordre (*i.e.* une sélection qui favorise la sélection). Afin de tester cette hypothèse, il faudrait déterminer si ce type de sélection peut être assez fort pour agir à de courtes échelles de temps (typiquement de l'ordre de la divergence homme - chimpanzé : 6 Ma). Jusqu'à maintenant, les modèles décrivant la sélection sur des modificateurs de la recombinaison [Otto & Barton, 1997] ne se sont intéressés qu'à des modificateurs affectant le taux de CO sur un locus. Dans le cas de PRDM9, le taux de recombinaison est modifié sur plusieurs milliers de loci à la fois [Brick et al., 2012] (vois ci-dessous) mettant en jeu des effets pléiotropes importants qui semblent incompatibles avec la selection de second ordre.

- La recombinaison favorise aussi certains réarrangements chromosomiques délétères. Par exemple, on a montré récemment que l'allèle A de *PRDM9* est associé significativement avec la duplication du gène *PMP22* par

recombinaison asymétrique qui entraine la maladie de Charcot-Marie-Tooth [Berg et al., 2010]. Ainsi, l'évitement de ce genre "d'incidents génétiques" pourrait pousser *PRDM9* à changer de cible régulièrement. Cette hypothèse nous parait peu vraisemblable car il semble plus parcimonieux de muter le locus concerné en *cis* afin d'en éloigner PRDM9 plutôt que de muter ce régulateur ce qui modifierait, par là-même, la recombinaison à d'autres loci de manière potentiellement délétère.

Concernant cette question, nous soutenons une hypothèse différente sur la base de découvertes récentes. Brick et collaborateurs ont mené la première étude associant la localisation des points chauds et les variants de *PRDM9* à l'échelle du génome entier chez la souris [Brick et al., 2012]. Ils ont ainsi montré qu'un changement d'allèle de *PRDM9* provoquait un changement radical dans le patron de formation des DSB : les deux souches aux allèles distincts ne partagent, en effet, que 1,1% de leurs points chauds (majoritairement trouvés dans les PAR). Chez ces deux souches, les points chauds se forment à des loci marqués par des modifications H3K4Me3 spécifiques de l'ADN des cellules testiculaires. D'autre part, les points chauds du mutant $PRDM9^{-/-}$ sont redirigés vers les promoteurs porteurs eux aussi de marques H3K4Me3 retrouvées, de manière non spécifique, dans plusieurs tissus somatiques. Le même phénomène est observé chez le chien dont la copie de *PRDM9* est naturellement pseudogénisée [Auton et al., 2013]. Il est possible que ce repositionnement soit délétère chez la souris car les mutants $PRDM9^{-/-}$ sont stériles [Hayashi et al., 2005]. Par analogie, il est possible qu'un génome contenant peu de cibles pour son variant de PRDM9 voit ses points chauds redirigés de la même sorte entraînant les mêmes effets délétères ce qui créerait une pression de sélection favorable à la montée en fréquence de nouveaux variants. Afin de confronter cette hypothèse aux données, il est au préalable important de quantifier si l'érosion des cibles de PRDM9 par BGC d'initiation est assez forte pour abolir un grand nombre de points chauds rapidement à l'échelle du génome. Ceci est l'objet de la seconde partie des résultats de cette thèse présentée au chapitre III (p. 131).

En complément sur cette partie, voir les articles de revue suivants : [Coop & Przeworski, 2007, Paigen & Petkov, 2010, Lichten & de Massy, 2011, Ponting, 2011].

# I.C. Le BGC, quatrième force de l'évolution des génomes ?

Après avoir replacé le BGC dans son contexte cellulaire et avoir décrit sa dynamique, nous allons maintenant nous pencher sur les conséquences de ce phénomène sur l'évolution des génomes.

## I.C.1. Le modèle d'évolution sous BGC

### Cadre théorique

A l'échelle moléculaire, la variabilité est créée par les mutations qui se produisent principalement lors de la réplication du matériel génétique ou sous l'actions d'agents chimiques (radicaux), physiques (UV) ou biotiques (virus). Les nouveaux variants sont ensuite transmis au générations suivantes après avoir subi deux filtres qui interfèrent entre eux : la sélection naturelle et la dérive génétique. La sélection naturelle favorise les mutations avantageuses (sélection positive, directionnelle ou adaptative) ce qui a pour conséquence de faire augmenter la fréquence de ces mutations dans la population. De même, elle tend à supprimer les variants délétères (sélection négative ou purificatrice) en diminuant leur fréquence. Au final, on définit la fitness d'un allèle, d'une combinaison d'allèle ou d'un individu soumis à la sélection naturelle, par la capacité intrinsèque de cette structure à transmettre son génotype aux générations suivantes. Cependant, ce modèle de sélection pure est trop simple pour décrire fidèlement l'évolution des fréquences alléliques dans les populations naturelles. En effet, les populations naturelles ont une taille limitée par les ressources qu'elles utilisent. De ce fait, le passage d'une génération à une autre se fait par échantillonnage aléatoire des allèles dans "l'urne gamétique". Cet échantillonnage est dépendant de la sélection mais pas seulement. Ainsi, même si l'urne gamétique représente la distribution des valeurs de fitness dans la population, le tirage de ces gamètes dans l'urne est un processus aléatoire simplement modélisable par un tirage binomial, c'est la dérive génétique. L'intensité de la dérive dépend de la taille de l'urne. En effet, la variance dans l'échantillonnage est plus importante dans les petites populations que dans les grandes. Ainsi, la sélection naturelle est plus efficace dans les grandes populations car les fréquences alléliques après échantillonnage reflètent plus fidèlement les fréquences observées dans l'urne et donc la distribution des valeurs de fitness. Dans les populations naturelles, seule une partie des individus participe réellement à la reproduction (à la formation de l'urne gamétique). Par exemple dans une population de souris à sex-ratio très déséquilibré en faveur des mâles, seule une partie de ces mâles accèdent

à la reproduction, par manque de femelles. L'effectif de la population $N$ ne reflète donc pas fidèlement le nombre de variants qui participent à la formation de l'urne de génération en génération. Ainsi, on introduit la taille efficace des populations notée $N_e$ ($\leq N$), qui correspond à la taille d'une population théorique pour laquelle l'intensité de la dérive est aussi forte que pour la population considérée [Wright, 1931].

Afin de quantifier le réel impact de la sélection naturelle sur l'évolution des génomes, il est donc nécessaire de mesurer l'importance des effets stochastiques qui peuvent interférer avec elle. C'est dans ce but que, dans les années 1960, Kimura a introduit la théorie neutre (ou neutraliste) de l'évolution moléculaire [Kimura, 1968]. Cette théorie est basée sur le fait que les taux de substitutions (changements fixés) observés par comparaison de séquences homologues entre espèces sont incompatibles avec des changements de fitness sensibles. Cependant, si ces changements impactaient massivement la fitness, étant donnée leur fort taux, ils conduiraient à une extinction des populations due à la forte présence de mutations délétères (c'est le fardeau de substitution). De ce fait, la majorité des substitutions doit être neutre, c'est à dire qu'elle ne change pas la fitness. Ainsi, l'evolution des fréquences alléliques évolue principalement par dérive à l'échelle du génome. S'appuyant sur cette théorie et sur des modèles mathématiques empruntés à la physique de la diffusion, Kimura a décrit l'évolution d'allèles soumis à la balance sélection/dérive [Kimura, 1962]. Ces équations de diffusion mettent en évidence l'importance du paramètre $4N_es$ ($2N_es$ chez les haploïdes) dans la balance entre sélection et dérive. $s$ est la mesure (entre -1 et 1) du "surplus" de fitness accordé à un génotype par une mutation : lorsque $s$ est négatif, la mutation est contre-sélectionnée et l'inverse quand $s$ est positif. Lorsque $|4N_es| < 1$ l'intensité de la sélection n'est pas assez forte pour surpasser celle de la dérive et le locus évolue de manière neutre. C'est dans ce carde théorique que nous décriront l'action du BGC dans les génomes.

### Le biais de conversion génique dans une population finie

En 1983, Nagylaki a décrit le phénomène de BGC en termes mathématiques [Nagylaki, 1983]. Le point de départ de cette analyse consiste à introduire un coefficient de disparité $x$ qui mesure la fréquence de transmission de l'allèle avantagé par le BGC. Il apparait alors que le BGC agit de manière similaire à la sélection positive en favorisant les allèles vers lesquels la conversion est biaisée. Par analogie avec le coefficient de sélection $s$, l'intensité du BGC sur un allèle est mesurée grâce à un coefficient $b$ (ou $g$). L'équation (2), montre la relation directe entre $x$ et $b$. D'autre part, comme vu plus haut dans l'équation

(1) (p. 44), $b$ est directement proportionnel au taux de recombinaison total (CO + NCO) et à l'intensité du biais au locus considéré. Dans le cas du dBGC, l'allèle favorisé est l'allèle qui initie moins souvent la recombinaison que son homologue. Tout se passe ainsi comme si l'allèle "chaud" était contre sélectionné. De même, dans le cadre du gBGC, les séquences se trouvant dans des régions à fort taux de recombinaison évoluent comme si les allèles riches en G et en C étaient sélectionnés aux dépens des allèles riches en A et T.

$$x = \frac{1}{2}(1 + b) \tag{2}$$

Ainsi, pour que le BGC ait un impact sur l'évolution de la fréquence allélique d'un variant il faut qu'il soit plus fort que la dérive à ce locus. Par analogie avec la sélection, ceci est mesuré grâce au produit $|4N_e b|$, noté $B$ (ou $G$) [Nagylaki, 1983]. Si celui-ci est supérieur à 1, le BGC a un effet plus fort que celui de la dérive et est capable de faire augmenter la fréquence de l'allèle avantagé. Le BGC est donc plus fort dans les grandes populations.

Compte tenu de la proximité théorique de fonctionnement de la sélection et du BGC à l'échelle moléculaire, nous pouvons d'ores et déjà faire deux remarques. Premièrement, si le BGC est assez fort, il peut mimer la sélection dans les données génomiques. Deuxièmement, ces deux forces peuvent fonctionner en synergie ou s'opposer, c'est la cas par exemple si l'allèle avantagé par le BGC est délétère. Ainsi, le BGC pose deux problèmes, le premier est méthodologique alors que le second est fondamental. Avant de traiter ces deux sujets, nous passerons en revue les indices de présence du gBGC dans les génomes.

## I.C.2. Les signatures du gBGC dans les génomes

Nous avons déjà vu, (*cf.* I.A.4. p. 43) les principales preuves de l'existence du dBGC dans les génomes eucaryotes. Dans les deux parties suivantes, nous nous focaliserons sur le gBGC.

### Les patrons de substitution et la recombinaison

L'hypothèse d'une relation entre le contenu en GC et la recombinaison *via* la conversion génique biaisée (*i.e.* le gBGC) a été proposée pour la première fois au début des années 1990 [Holmquist, 1992, Eyre-Walker, 1993]. Cette hypothèse a donc été formulée bien avant la construction de la première (et unique à ce jour) carte de recombinaison à haute résolution chez la levure qui a permis de mettre en évidence de manière directe ce phénomène [Mancera

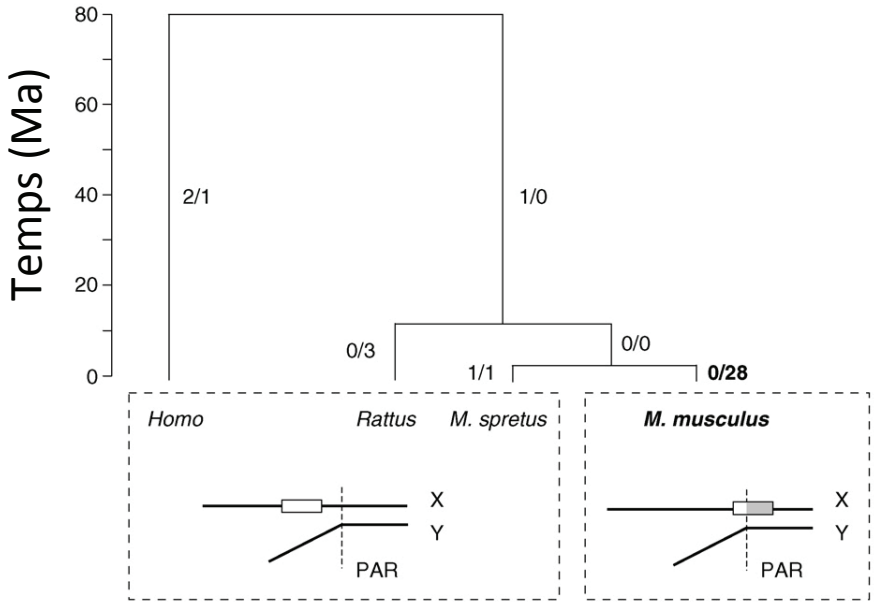



FIGURE I.26 : Evolution du locus *Fxy* chez les mammifères. Les boites encadrées montrent la position du gène (rectangle) par rapport à la PAR sur le chromosome X des espèces mentionnées aux feuilles. Le nombre de changements d'acides aminés du gène est indiqué le long des branches respectivement pour les sites se trouvant hors / dans la PAR chez *Mus musculus*. Ma : millions d'années. Adapté de [Galtier & Duret, 2007].

et al., 2008] (*cf.* I.A.3. p. 41). Ainsi, les preuves de l'existence du gBGC sont très majoritairement indirectes. Rappelons que le modèle du gBGC prédit que, lors de la conversion génique associée à la recombinaison, l'allèle le plus riche en GC a plus de chance de convertir l'allèle plus riche en AT que l'inverse. Un argument majeur de l'existence du gBGC se trouve donc dans les corrélations entre taux de GC et taux de recombinaison. Ainsi, chez l'homme, les gènes soumis à la recombinaison ectopique de manière fréquente, tels que les gènes à grand nombre de copies (ARNt, ARNr...) présentent, un taux de GC bien supérieur (55-83%) à celui du reste du génome (42%) [Galtier et al., 2001]. Deuxièmement, on sait que les régions pseudo-autosomales (PAR) du X et du Y sont riches en points chauds de recombinaison car ce sont les seules qui peuvent recombiner chez les mâles à l'échelle du chromosome entier. Le taux de recombinaison des PAR est donc supérieur à celui des autosomes qui est lui plus fort que celui du chromosome X hors PAR. Or, le GC3 (taux de GC des troisièmes positions des codons) des régions codantes des PAR atteint plus de 70% alors qu'il est de 56% pour le reste du chromosome X

et prend une valeur intermédiaire de 62% sur les autosomes [Galtier et al., 2001]. La troisième position des codons étant souvent synonyme, il est peu probable que les changements ayant conduits à ces taux de GC soient dus à la sélection. A l'inverse, le gBGC favorise indifféremment la fixation de mutations de AT vers GC sur toutes les positions des codons ainsi que dans les régions non codantes. Le cas du locus *Fxy*, chez la souris, permet d'étudier de manière comparative l'influence de la recombinaison sur l'évolution de la composition en base [Perry & Ashworth, 1999] (Figure I.26). Chez l'homme, le rat et la souris *Mus spretus*, ce gène est conservé et se trouve sur le chromosome X hors de la PAR. Cependant, chez *Mus musculus, Fxy* a subi une translocation qui place ses exons 4 à 10 dans la PAR entraînant une augmentation importante de la recombinaison sur ces segments. Ceci a aboutit à une augmentation du taux de GC dans les introns correspondants ainsi que du GC3 des exons 4 à 10 uniquement [Montoya-Burgos et al., 2003]. De plus, dans la branche longue de 1 à 3 Ma séparant *Mus musculus* de son ancêtre commun avec *Mus spretus*, on compte 28 changements d'acides aminés tous dus à des substitutions de AT vers GC localisées dans les exons 4 à 10 (Figure I.26). A titre de comparaison, depuis la séparation entre l'homme et les rongeurs, il y a environ 80 Ma, un seul changement de la sorte s'est produit dans la branche humaine et dans la branche de *Mus spretus* ce qui suggère fortement que la recombinaison est responsable du patron de substitution particulier observé chez *Mus musculus*.

A l'échelle du génome entier, chez l'homme, le taux de GC et le taux de recombinaison, calculés sur des fenêtres de 1Mb, sont corrélés positivement [Meunier & Duret, 2004, Duret & Arndt, 2008] (Figure I.27 (C)). Cette corrélation est assez faible : son coefficient de détermination $R^2$ est de 0.15. Ce lien entre recombinaison et contenu en GC tend à soutenir le modèle gBGC. Sous l'hypothèse que le taux de GC est déterminé par le taux de recombinaison, la faiblesse de la corrélation peut être imputée à la discordance entre les échelles de temps des deux grandeurs analysées. En effet, le taux de GC observé est, en fait, un résumé des patrons de substitution qui se sont succédés sur plusieurs dizaines de millions d'années. Le taux de recombinaison est, lui, historique : il ne reflète que les derniers milliers d'années correspondant à l'histoire des populations utilisées pour reconstruire les cartes de DL [The International HapMap Consortium, 2007]. Donc, pour tester si la recombinaison influe sur les patrons de substitution, il faut considérer des grandeurs qui résument l'activité des deux paramètres sur des échelles de temps proches. En outre, le taux de GC du génome humain n'est pas à l'équilibre. En effet, si l'on analyse les patrons de substitution sur la branche humaine, depuis la divergence avec le chimpanzé, on observe que le GC d'équilibre (GC*) et le GC

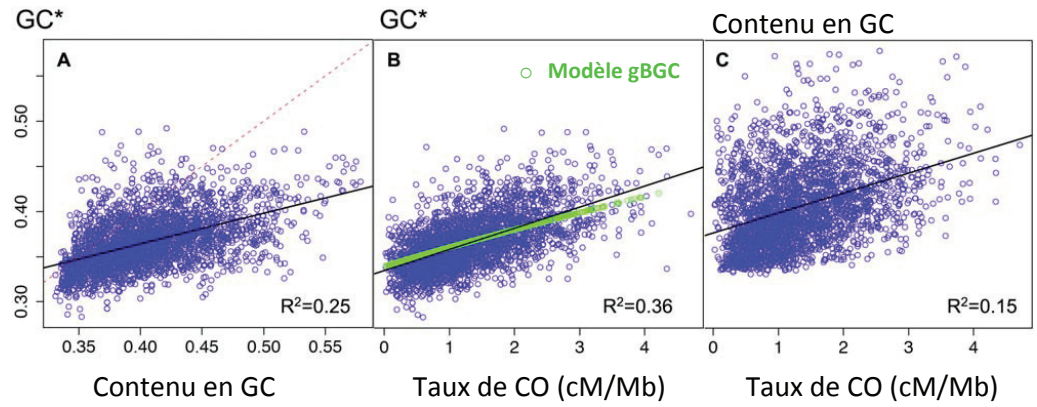



FIGURE I.27 : Corrélations du taux de GC, du GC* et du taux de CO calculés sur des fenêtres de 1Mb chez l'homme. (A) GC* vs. GC. La ligne pointillé représente la première bissectrice. (B) GC* vs. taux de CO. Les points verts montrent les prédictions d'un modèle d'évolution de séquence sous gBGC avec des points chauds de recombinaison dynamiques. (C) GC vs. taux de CO. Adapté de [Duret & Arndt, 2008].

observé sont différents et faiblement corrélés à l'échelle du mégabase [Meunier & Duret, 2004, Duret, 2006, Duret & Arndt, 2008] (Figure I.27 (A)). Le GC* est calculé à partir du nombre de substitutions WS ($\#AT \rightarrow GC$) et SW ($\#GC \rightarrow AT$) se produisant dans une branche ainsi que du nombre de sites AT et GC (#AT et #GC) comme décrit dans l'équation (3). Il donne le taux de GC que doit atteindre une séquence qui évolue sous les mêmes taux de substitution que ceux qui agissent sur la branche considérée. Il ne s'agit donc pas d'une valeur instantanée, mais d'une statistique décrivant l'évolution des séquences sur une période de temps donnée.

$$GC^* = \frac{{}^{\#AT\rightarrow GC}/_{\#AT}}{{}^{\#AT\rightarrow GC}/_{\#AT} + {}^{\#GC\rightarrow AT}/_{\#GC}} \tag{3}$$

Le GC* est un paramètre dont l'estimation est très sensible à la méthode utilisée. La reconstruction de séquences ancestrales par parcimonie peut biaiser cette estimation car elle est très sensible aux forts taux mutationnels tels que ceux des sites CpG (dinucléotide 5'-CG-3'). Pour cela, l'utilisation du maximum de vraisemblance pour estimer les paramètres d'un modèle d'évolution moléculaire [Felsenstein, 1981] qui prend en compte ces sites particuliers améliore l'estimation du GC* [Duret, 2006, Duret & Arndt, 2008]. Grâce à cette méthode, Duret & Arndt ont montré que le GC* mesuré sur la branche humaine depuis la divergence avec le chimpanzé corrèle

mieux avec la recombinaison ($R^2 = 0.36$) que le taux de GC observé [Duret & Arndt, 2008] (Figure I.27 (B)). Ceci suggère que c'est bien la recombinaison qui influe sur le taux de GC et non l'inverse. En effet, si c'était le taux de GC qui déterminait l'intensité de la recombinaison, alors le GC observé serait un meilleur prédicteur de la recombinaison que le GC d'équilibre. La recombinaison agit donc directement sur les patrons de substitution en augmentant le taux de WS. Ceci est en adéquation avec un modèle simulant l'évolution de séquences sous gBGC avec une répartition de la recombinaison en points chauds dont la localisation varie dans le temps [Duret & Arndt, 2008] (Figure I.27 (B)). Ainsi, chez l'homme, 35% de la variance sur le GC* peut être expliquée par la recombinaison. Récemment, Munch et collaborateurs ont développé une méthode basée sur l'étude des patrons d'ILS (tri de ligné incomplet, Incomplete Lineage Sorting en anglais) permettant d'inférer des taux de recombinaison ancestraux (voir encadré "méthode" ci-dessous) [Munch et al., 2014]. En utilisant cette méthode les auteurs ont reconstruit l'activité recombinatoire le long de la branche qui sépare l'ancêtre commun de l'homme et du chimpanzé avec celui qui les sépare du gorille. On parlera alors de taux de recombinaison "ancestral". Les auteurs ont ainsi pu montrer que l'on peut expliquer 64% de la variance sur le GC* grâce à la recombinaison ($R^2$=0.64). Ceci démontre que la recombinaison est un déterminant majeur (voire unique) de l'évolution du GC. De plus, il s'agit d'un argument indirect fort soutenant le modèle du gBGC.
Remarque : l'étude de cartes de pedigree montre que la corrélation recombinaison-GC* est plus forte si l'on ne considère que la recombinaison mâle [Webster et al., 2005, Duret & Arndt, 2008]. Comme montré par Popa et collaborateurs, cette apparente différence entre mâles et femelles est entièrement due aux télomères [Popa et al., 2012]. En effet, la recombinaison étant bien plus forte au niveau des télomères lors de la spermatogenèse que lors de l'ovogenèse, les points du génome dont le GC* et la recombinaison sont calculés sur ces régions soutiennent d'avantage la corrélation chez les mâles que chez les femelles.

**Méthode**

**Les cartes basées sur l'ILS**

Cette technique a été développée par Munch et collaborateurs très récemment [Munch et al., 2014] et n'a été utilisée que chez l'homme et le chimpanzé pour le moment. Elle s'appuie sur le concept d'ILS (tri de ligné incomplet, Incomplete Lineage Sorting en anglais). L'ILS survient au

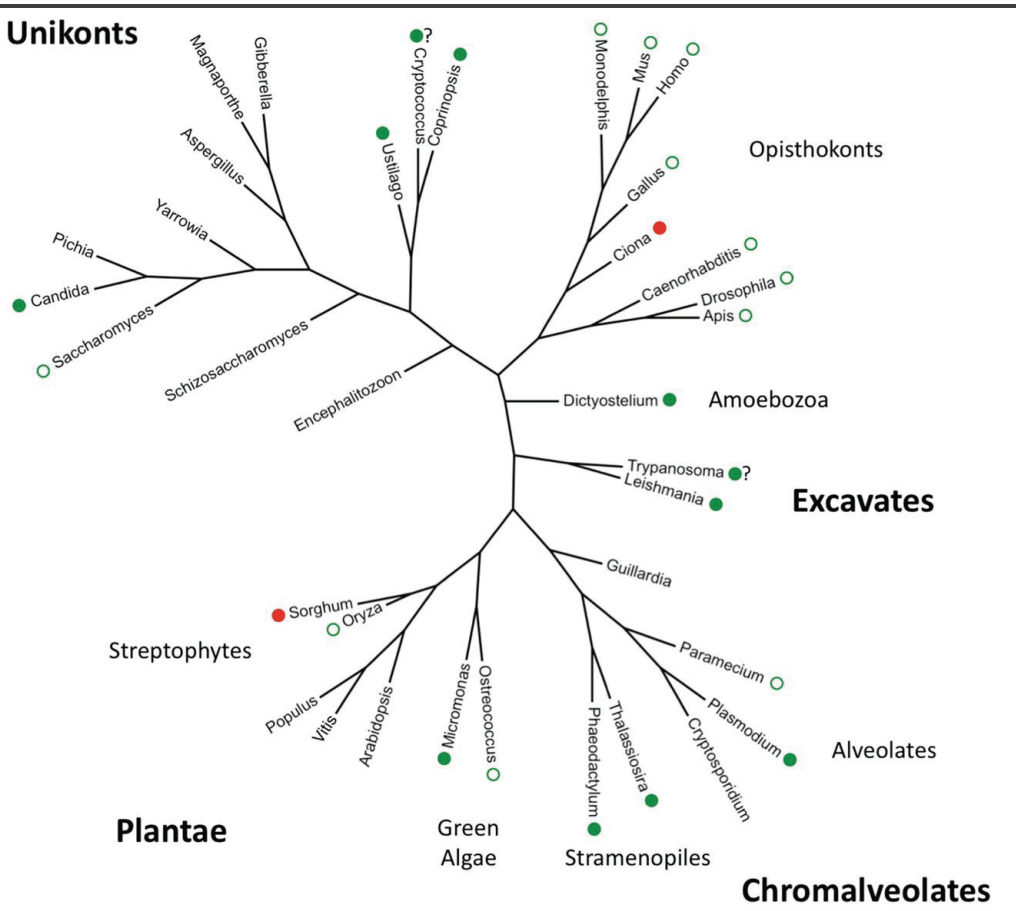



FIGURE I.28 : Phylogénie des 36 espèces étudiées par Pessia et collaborateurs [Pessia et al., 2012]. Les cercles verts indiquent les corrélations significatives positives entre GC (ou GC3) et taux de recombinaison (ou taille des chromosomes) qui suggèrent la présence de gBGC. Les cercles rouges montrent les corrélations significatives négatives entre ces paramètres. Elles ne sont pas compatibles avec le gBGC. Les cercles pleins indiquent que l'observation a été faite pour la première fois dans [Pessia et al., 2012]. Le " ?" indique que les résultats obtenus avec le taux de recombinaison et la longueur des chromosomes ne sont pas en accord. Adapté de [Pessia et al., 2012].

moment de la séparation de deux populations aboutissant à une spéciation. Elle est courante entre l'homme, le chimpanzé et le gorille étant donnée la petite taille des branches terminales conduisant à ces espèces. Au moment de la séparation entre le chimpanzé et l'homme, certains loci sont polymorphes et les deux allèles sont transférés à la fois dans la branche du chimpanzé et la branche humaine. Par dérive, l'allèle ancestral (porté par le gorille) peut se fixer chez l'homme alors que l'allèle dérivé se fixe chez le chimpanzé. Une phylogénie basée sur ce locus infèrera une proximité plus grande entre le gorille et l'homme qu'entre le chimpanzé et l'homme, c'est ce que l'on appelle l'ILS. Ainsi, le long de l'alignement des génomes du gorille, du chimpanzé et de l'homme, on passe successivement d'un bloc groupant l'homme et le gorille à un bloc groupant l'homme et le chimpanzé ou à un bloc groupant le chimpanzé et le gorille et ainsi de suite. La méthode de Munch et collaborateurs repose sur le fait que pour passer d'un patron d'ILS à un autre, il faut nécessairement qu'il y ait eu un évènement de CO le long de la branche qui sépare l'ancêtre de l'homme et du chimpanzé avec celui qu'ils partagent avec le gorille. Pour le moment, cette technique est la seule qui permette une étude des patrons de recombinaison ancestraux. *En bilan sur les techniques d'étude de la recombinaison, voir Annexe A p. 219.*

En se basant sur des corrélations entre GC ou GC3 et taux de recombinaison ou longueur des chromosomes (qui est un prédicteur du taux de recombinaison comme vu précédemment), Pessia et collaborateurs ont montré que des indices suggérant la présence du gBGC sont présents chez un grand nombre d'espèces balayant les principaux groupes d'eucaryotes [Pessia et al., 2012] (Figure I.28). Ceci confirme les résultats d'Escobar et collaborateurs qui ont montré une augmentation du taux de GC compatible avec le gBGC, dans les séquences d'ARNr 18S (soumis à un fort taux de recombinaison ectopique) de nombreux eucaryotes dont plusieurs vertébrés, angiospermes, un échinoderme et chez l'eucaryote basal *Giardia* [Escobar et al., 2011]. De façon analogue, on a montré que les points chauds de substitution propres au génome de l'homme [Dreszer et al., 2007, Capra & Pollard, 2011], de la souris, du chien, de l'épinoche et du nématode *C. elegans* [Capra & Pollard, 2011] sont biaisés vers GC. Ces points chauds sont estimés par comparaison des taux d'évolution de ces régions entre une espèce et des espèces proches. Dans les clades mentionnés ci-dessus, ces points chauds affectent aussi bien les régions codantes que non codantes. De plus, ils sont significativement re-

groupés dans les régions à fort taux de recombinaison. Tous ces éléments sont compatibles avec le modèle du gBGC. Chez la levure, ces points chauds de substitution semblent ne pas être biaisés vers GC [Capra & Pollard, 2011]. Cependant, outre la preuve directe déjà évoquée plus haut (*cf.* I.A.4. p. 47 et [Mancera et al., 2008]), il existe une corrélation significative entre GC3 des gènes et taux de recombinaison qui suggère la présence du gBGC chez cet organisme [Birdsell, 2002]. L'absence de point chaud de substitution biaisé vers GC chez cette espèce peut s'expliquer par le fait que les points chauds de recombinaison sont conservés à l'échelle des espèces soeurs proches [Tsai et al., 2010]. On ne s'attend donc pas à trouver des régions spécifiquement "accélérées" (en terme de taux d'évolution) par le gBGC chez *Saccharomyces cerevisiae* . Chez le poulet, Capra et collaborateurs n'enregistrent pas non plus de signatures de gBGC dans le points chauds de substitution [Capra & Pollard, 2011]. Ceci semble être dû à la grande dépendance statistique entre de multiples facteurs génomiques et le taux de GC spécifiquement dans les génomes à karyotypes stables comme ceux des oiseaux [Mugal et al., 2013]. Selon Mugal et collaborateurs, lorsque l'effet de ces facteurs confondants est pris en compte, on observe une signature claire de gBGC chez le poulet [Mugal et al., 2013]. Enfin, chez la drosophile, aucune trace de gBGC n'a été, enregistrée à ce jour [Comeron et al., 2012, Robinson et al., 2013].

### Apport des données de polymorphisme

Les corrélations entre la recombinaison et les patrons de substitution tendent à montrer que la recombinaison permet l'augmentation locale du taux de GC. Ceci est en accord avec le modèle du gBGC. Cependant, il est aussi possible que la recombinaison soit mutagène et que le patron de mutation associé soit biaisé vers GC. La capacité de la recombinaison à générer des mutations est encore peu documentée mais a été invoquée à plusieurs reprises pour expliquer certains patrons de polymorphisme chez l'homme [Lercher & Hurst, 2002, Hellmann et al., 2003]. Cependant, des études des patrons de substitution des zones de jonctions des PAR et du reste du chromosome X suggèrent que la recombinaison ne serait pas mutagène, chez l'homme [Yi et al., 2004] et la souris [Huang et al., 2005].
Malgré tout, il est possible de distinguer d'éventuels effets de la mutation des effets du gBGC. En effet, alors que le processus mutationnel donne naissance à de nouveaux variants (qui sont alors à faible fréquence dans la population, typiquement 1/N) le gBGC, lui, augmente la probabilité de transmission de ces variants. Cela aboutit à l'augmentation de la fréquence allélique des variants riches en GC dans la population. Le processus mutationnel n'a aucun effet sur les fréquences alléliques contrairement aux autres forces : dérive,

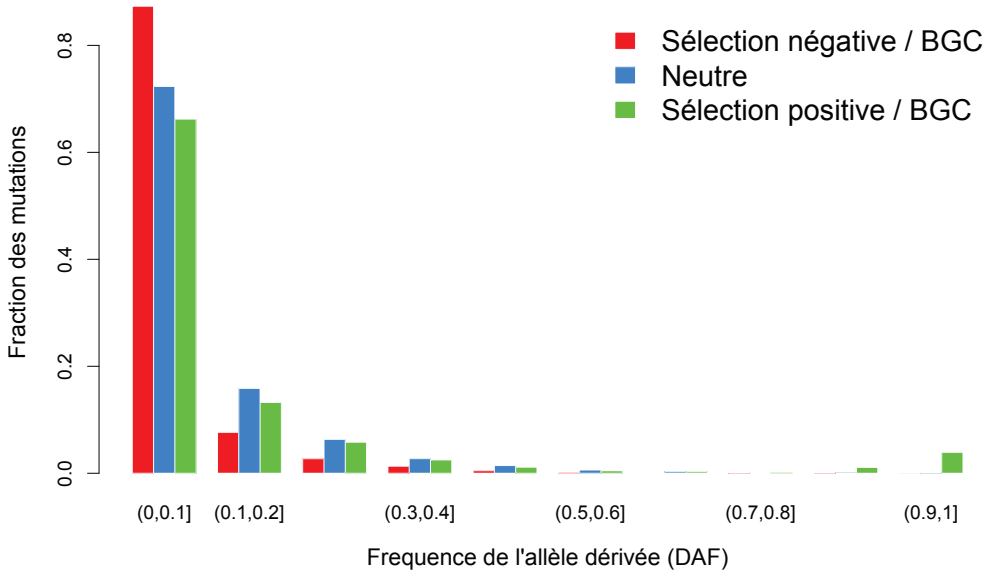



FIGURE I.29 : Représentation schématique des spectres de DAF attendus pour différents types de sites.

BGC et sélection.
Il est possible d'analyser le comportement des fréquences alléliques grâce aux spectres de fréquence d'allèles dérivés (DAF pour Derived Allele Frequency en anglais). La DAF est une mesure de la fréquence allélique d'un variant dans une population. Elle est calculée sur la base du génotypage d'un échantillon d'individus pour le locus considéré. Si, à ce locus, un polymorphisme segrège, il faut déterminer : (i) la fréquence de chaque allèle estimée directement grâce à sa proportion dans l'échantillon, (ii) lequel des deux variants représente l'état dérivé. Pour cela, un ou plusieurs groupes externe(s) est/-sont utilisé(s). Ils donnent l'état ancestral de l'allèle à ce locus, l'autre étant alors inféré comme dérivé. Le spectre de DAF représente la distribution des DAF d'un ensemble de mutations dans une population. Le spectre de DAF pour des mutations affectant des sites neutres a une forme exponentielle décroissante car la majorité des mutations qui ségrègent viennent d'apparaitre dans la population (elles ont donc une faible DAF) et une faible fraction (probabilité : $1/2N_e$) d'entre elles augmente en fréquence car la plupart est perdue par dérive (Figure I.29 (barres bleues)). Le spectre de DAF pour des mutations dont l'état dérivé est délétère (ou défavorisées par le BGC) est

décalé vers les basses fréquences par rapport au spectre neutre car les nouvelles mutations sont rapidement éliminées ou gardées à une faible fréquence (Figure I.29 (barres rouges)). A l'inverse, les mutations dont l'état dérivé est adaptatif ou avantagé par le BGC (les allèles riches en GC dans le cas du gBGC) ont un spectre décalé vers les hautes fréquences avec un pic autour de 1 montrant la forte proportion des mutations fixées (Figure I.29 (barres vertes)).
Sous le modèle du gBGC on attend donc que les mutations WS ségrègent à plus haute fréquence que les mutations SW, ce qui n'est pas le cas sous un modèle mutationnel. La construction des spectres de DAF affectant les mutations SW d'une part et WS d'autre part a été rendue possible par le génotypage, puis le séquençage, massif d'individus de différentes populations humaines [The International HapMap Consortium, 2007, The 1000 Genomes Project Consortium, 2010]. Dans toutes les populations étudiées (CEU, YRI, CHB+JPT) les spectres de DAF des mutations WS sont décalés vers les hautes fréquences par rapport à l'attendu neutre et au spectre des mutations SW [Katzman et al., 2011] (Figure I.30). Ceci confirme les résultats d'études plus anciennes basées sur de plus petits échantillons [Eyre-Walker, 1999, Duret et al., 2002, Webster & Smith, 2004, Spencer et al., 2006]. De plus, le "décalage" des spectres des mutations WS est plus fort dans les régions à fort taux de recombinaison comme attendu sous le modèle du gBGC [Katzman et al., 2011]. L'étude des allèles rares (fréquences $\leq 10^{-4}$) permet d'étudier les processus mutationnels en excluant l'influence des facteurs qui agissent sur les fréquences alléliques : sélection, dérive et BGC. Une analyse de la sorte, réalisée chez l'homme, permet, elle aussi, de rejeter l'hypothèse d'un effet mutationnel de la recombinaison en remarquant, entre autres, qu'il n'y a pas plus de variants rares GC que AT au voisinage des points chauds [Schaibley et al., 2013].

Enfin, la comparaison des spectres de DAF WS et SW permet de quantifier le l'intensité du gBGC ($B$). A l'échelle du génome entier, une analyse datant de 2004 donne une valeur de $B = 0.4$ ce qui est trop faible pour avoir une influence significative sur toutes les mutations WS [Webster & Smith, 2004]. Cependant, ce paramètre est une moyenne de l'effet du gBGC sur le génome, on s'attend donc à ce que $B$ soit bien plus fort dans les points chauds. C'est ce qui est observé dans les données de polymorphisme humain : $B \approx 1.3$ dans les régions du génome qui recombinent le plus [Spencer et al., 2006]. Comme nous le verrons dans notre étude, ceci s'applique aussi, dans d'autres proportions, au dBGC (*cf.* III p. 131).

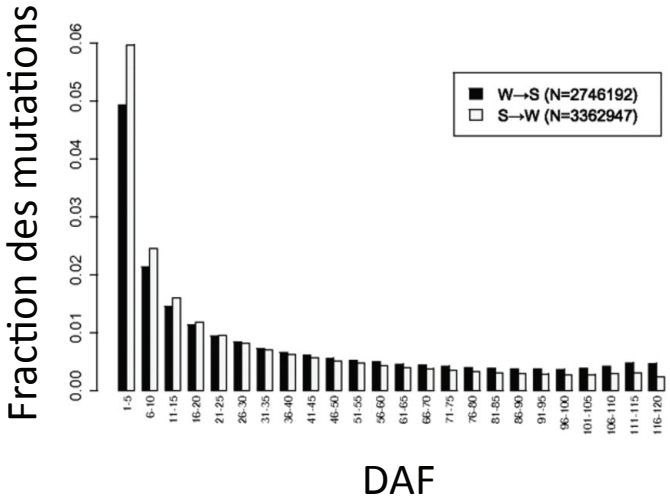



FIGURE I.30 : Spectre de DAF pour les mutations WS (AT vers GC) et SW (GC vers AT) qui ségrègent dans la population CEU. La DAF est donnée en nombre d'individus (sur 120 au total) partageant l'allèle dérivé pour chaque SNP. Les effectifs de chaque catégorie de mutations sont indiqués en légende. Le décalage des deux spectres est statistiquement significatif ($p \approx 0$). Adapté de [Katzman et al., 2011].

### Apports des modèles d'évolution de séquence

Récemment, plusieurs études ont permis de mieux caractériser le phénomène de gBGC grâce à sa modélisation. Mentionnons à ce titre trois études apportant des éclairages sur trois aspects de ce processus : sa répartition le long du génome, son évolution au sein des mammifères et sa relation avec d'autres caractéristiques génomiques et écologiques.

- Un modèle d'évolution de séquence sous sélection et gBGC, basé sur un processus de Markov caché proposé par Capra et collaborateurs [Capra et al., 2013] permet d'identifier les zones du génome évoluant sous gBGC. Cette méthode utilise un alignement de deux espèces qui peuvent être prises indifféremment comme cible (target) de l'étude et de deux groupes externes. La présence ou l'absence de sélection est inférée à partir de l'état de conservation des séquences. A cela se superpose la présence ou l'absence de gBGC d'intensité $B$ définie *a priori* (paramètre d'entrée). Les résultats obtenus sur l'alignement homme, chimpanzé, orang-outan et macaque, avec l'homme ou le chimpanzé comme cible, montrent des patrons de gBGC en accord avec l'attendu. En effet, les tracts de gBGC ainsi identifiés (pour $B = 3$ : longueur

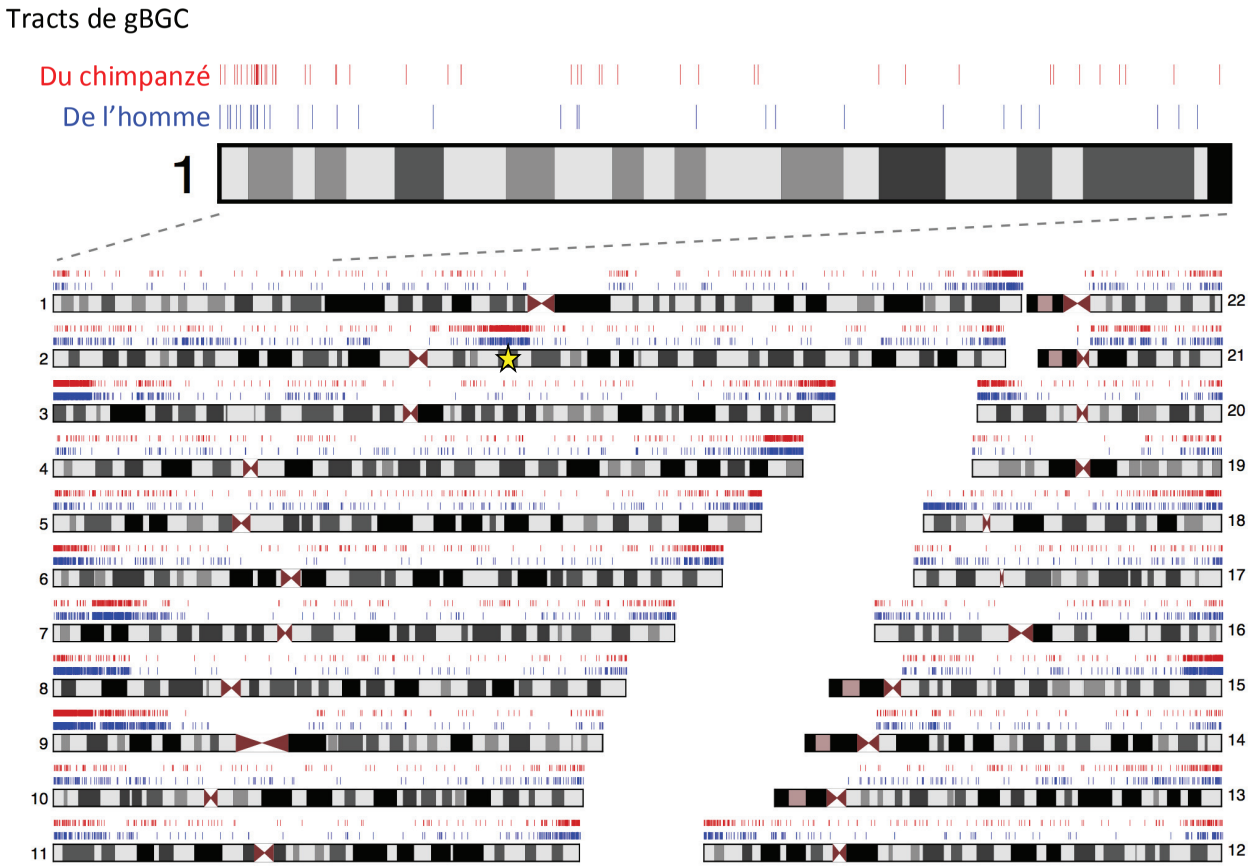



FIGURE I.31 : Tracts de gBGC détectés par Capra et collaborateurs chez l'homme (bleu) et le chimpanzé (rouge). Le numéro des chromosomes humains est reporté. L'étoile jaune montre le point de fusion des chromosomes ancestraux 2a et 2b chez l'homme, il correspond à un tract majeur de gBGC. La partie supérieure montre le détail d'un bras du chromosome 1. Observez la concentration des tracts de gBGC dans les régions sous-télomériques et leur déplétion au niveau des centromères. Adapté de [Capra et al., 2013].

moyenne de 1kb) couvrent 0,3% de chacun des génomes environ et se concentrent dans les régions sous-télomériques et les points chauds de recombinaison [Capra et al., 2013] (Figure I.31). De plus, l'analyse des DAF des mutations trouvées dans les tracts de gBGC détectés chez l'homme montre que le biais vers GC est dû à un processus de fixation. Enfin, les tracts de gBGC sont peu chevauchants entre l'homme et le chimpanzé à petite échelle (10kb) alors qu'à grande échelle (1Mb), la densité en tract est bien corrélée entre les deux espèces. Ceci reflète la conservation des patrons de recombinaison et la dynamique des points chauds. La carte des tracts de gBGC est disponible via le serveur de l'UCSC [Rhead et al., 2010].

- L'histoire de $B$ a été reconstruite par par inférence bayésienne des paramètres d'un modèle d'évolution de séquences incluant le gBGC à l'échelle des mammifères [Lartillot, 2013]. Dans cette étude, l'intensité du gBGC agissant sur chaque branche de la phylogénie est inférée pour chaque exon de l'alignement regroupant une large gamme de mammifères de tous ordres. Il apparait que l'intensité du gBGC estimée chez les espèces modèles (homme et souris) n'est pas représentatif du patron observé chez le reste des mammifères [Lartillot, 2013]. Ainsi, le gBGC estimé sur les exons humains ($B = 0.1$ en moyenne) serait l'un des plus faible des mammifères (Figure I.32). En effet, la valeur moyenne est d'environ 1 sur l'ensemble des mammifères étudiés et plusieurs taxons présentent des valeurs de gBGC particulièrement fortes comme le lapin ou la chauve-souris *Myotis* ($B \approx 3 - 5$). Ces résultats confirment la distribution du gBGC estimée sur un échantillon taxinomique similaire grâce à la reconstruction des valeurs de GC3 ancestrales par maximum de vraisemblance [Romiguier et al., 2010].

- Cette dernière étude a aussi permis d'analyser la relation entre l'intensité du gBGC et d'autres caractéristiques propres à chaque espèce [Romiguier et al., 2010]. Les auteurs ont montré que le gBGC est plus fort dans les petits que dans les grands génomes ($\rho = -0.48; p = 0.01$) car le taux de recombinaison par mégabase y est plus fort, comme vu précédemment [Romiguier et al., 2010]. De plus, l'intensité du gBGC est corrélée négativement à la taille corporelle et à d'autres traits d'histoire de vie connus pour être des prédicteurs forts de la taille efficace des populations. Puisque $B$ est directement proportionnel à $N_e$, on s'attend donc à ce qu'il soit plus fort chez le petit hérisson-tenrec (*Echinops*) que chez l'éléphant (*Loxodonta*) malgré leur proximité phylogénétique à l'échelle des mammifères. Ceci correspond à ce qui est inféré par les

deux études : [Romiguier et al., 2010, Lartillot, 2013] (Figure I.32).

## I.C.3. Le gBGC et la sélection naturelle

Pour que le BGC soit reconnu comme une force majeure de l'évolution des génomes, il ne suffit pas d'apporter les preuves de son existence, il faut aussi s'intéresser à sa relation avec les autres forces afin de comprendre si il peut s'y opposer ou les accompagner de manière significative. Nous avons déjà vu que le BGC est fort lorsque la dérive est faible (*i.e.* dans les grandes populations). Dans cette partie, nous analyserons deux aspects de la relation entre gBGC et sélection. Tout d'abord, nous verrons deux cas où le gBGC mime la sélection à travers deux "débats" scientifiques sur l'origine évolutive de deux structures génomiques : les isochores et les gènes à évolution accélérée. Ensuite nous verrons comment le gBGC et la sélection peuvent s'opposer ou agir en synergie ce qui nous amènera finalement à nous poser la question de l'origine évolutive de ce phénomène.

### Le débat sur les paysages génomiques : les isochores

Les génomes de mammifères et d'oiseaux montrent des variations de contenu en GC à grande échelle appelées isochores. Les isochores ont été découverts au milieu des années 1980 grâce à l'étude biochimique de fragments d'ADN de grande taille ($\approx 100kb$) [Bernardi et al., 1985, Mouchiroud et al., 1987]. Ces variations de GC affectent indifféremment les parties codantes et non-codantes (Figure I.33 (b)) et semblent dessiner les contours de "l'anatomie" [Duret et al., 2006] ou de "l'architecture" [Duret & Galtier, 2009a] du génome. En effet, le taux de GC est corrélé à la densité en gènes et à la "compaction" des gènes (*i.e.* le rapport entre la longueur de la partie non-codante et de la partie codante d'un gène) suggérant une réelle organisation des éléments fonctionnels au sein de compartiments dont la nature est définie par le taux de GC (Figure I.33 (a)). De plus, il semble au premier abord que ces structures sont assez stables à l'échelle des mammifères, sauf chez les muridés [Mouchiroud et al., 1987]. Cependant, des analyses plus récentes montrent qu'elles s'érodent : on observe, en effet, chez les primates et les muridés notamment, des taux de substitutions enclins à réduire le taux global de GC ce qui aboutit à son homogénéisation le long du génome [Duret et al., 2002, Belle et al., 2004]. Les particularités de l'organisation des isochores le long du génome et de leur variabilité à l'échelle des vertébrés posent la question de leur origine évolutive. A la fin des années 1990, trois hypothèses majeures se détachent concernant cette question.

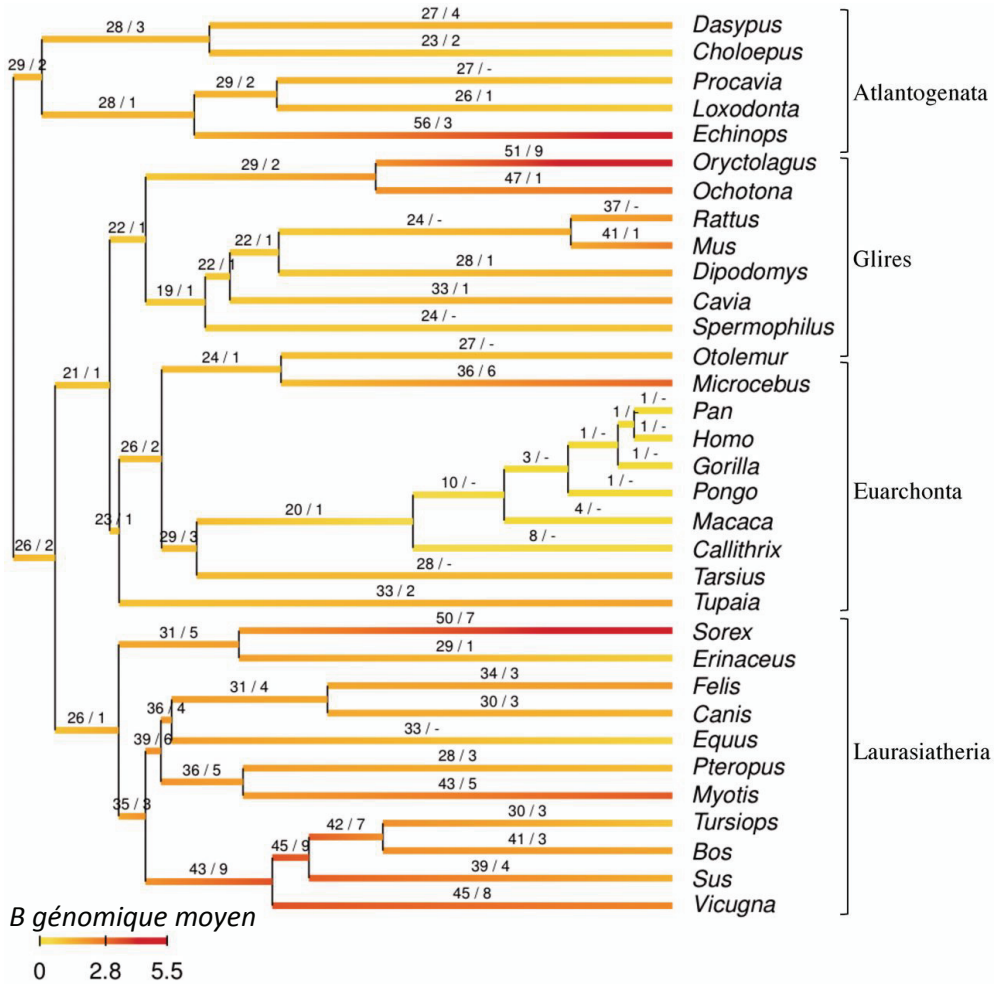



FIGURE I.32 : Reconstruction de l'histoire évolutive de $B$ chez les mammifères. Les noms d'ordres sont donnés dans la colonne de droite. Les deux nombres au dessus de chaque branche donnent le pourcentage d'exons évoluant sous $B > 1/B > 10$. Adapté de [Lartillot, 2013].

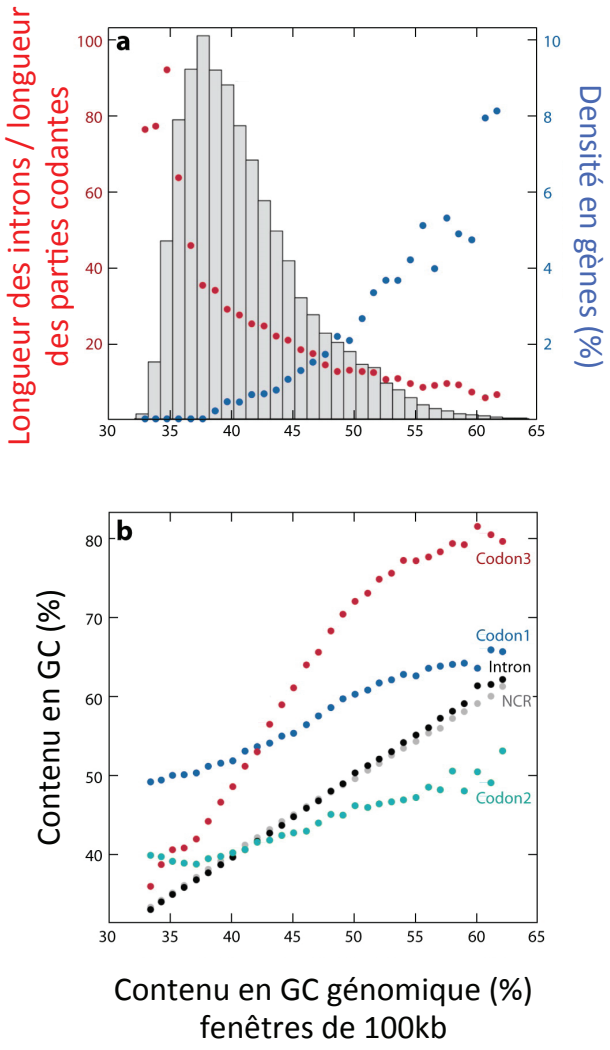



FIGURE I.33 : Contenu en GC et architecture du génome humain. (a) Distribution des taux de GC calculés sur des fenêtres de 100kb et relation avec la compaction des gènes (longueur des introns/longueur des parties codantes, en rouge, échelle de gauche) et la densité en séquences codantes (en bleu, échelle de droite). (b) Relations entre le taux de GC calculé sur des fenêtres de 100kb et le GC des introns (noir), des autres régions non codantes (gris) et des régions codantes différenciées selon la position dans le codon : 1 (bleu), 2 (vert) et 3 (rouge). Adapté de [Duret & Galtier, 2009a].

Premièrement, plusieurs auteurs ont proposé que le patrons de mutation soient responsable de l'apparition des isochores. Ceci se ferait *via* les variations dans le timing d'initiation de la réplication le long du génome [Wolfe et al., 1989] ou les variations de taux de désamination des cytosines qui aboutit à un plus grand nombre de mutations de C vers T dans les régions déjà riches en AT [Fryxell & Zuckerkandl, 2000]. Cependant, la première hypothèse a été remise en cause car il ne semble pas y avoir de relation entre le taux de GC et le timing de réplication dans les cellules de la lignée germinale. La seconde hypothèse est séduisante mais impose que les isochores soient en croissance continue ce qui est en désaccord avec leur érosion apparente [Eyre-Walker & Hurst, 2001]. De plus, comme nous l'avons vu dans la partie précédente grâce à la comparaison des spectres de DAF WS et SW, les variations de taux de GC à l'échelle du génome entier, sont conduites par un processus de fixation (tel que la sélection, le BGC ou la dérive) et non pas un processus mutationnel [Eyre-Walker, 1999, Duret et al., 2002, Webster & Smith, 2004, Spencer et al., 2006].
Deuxièmement, Bernardi, l'un des principaux instigateurs de la découverte des isochores dans les années 1980, défend une hypothèse sélectionniste qui stipule que les isochores sont une adaptation à la température corporelle élevée des homéothermes (mammifères et oiseaux), les paires G-C étant plus résistantes à la dénaturation thermique que les paires A-T [Bernardi, 2000]. Cette hypothèse était soutenue par le fait que lors des premières analyses effectuées, les vertébrés ectothermes ne présentaient pas de trace d'isochores [Bernardi & Bernardi, 1991]. Cependant, lors d'études ultérieures, des structures semblables ont été détectées chez plusieurs animaux à sang froids comme le crocodile du Nil ou la tortue de Floride [Hughes et al., 1999]. De plus, le taux de GC n'est pas corrélé à la température optimale de croissance chez les bactéries malgré une empreinte forte de la sélection naturelle dans l'évolution de leur génome due à leur forte taille de population [Galtier & Lobry, 1997]. Au delà de l'hypothèse de Bernardi, il parait difficile d'imaginer une pression de selection agissant à la fois sur les sites codants et non-codants à l'échelle du génome entier et de manière différentielle sur des segments de plusieurs centaines de kb. En effet, si tous les sites des isochores étaient sous selection alors, étant donnés les taux de mutations chez les mammifères, les génomes accumuleraient un fardeau de mutations délétères bien trop important pour être toléré [Eyre-Walker & Hurst, 2001].
Troisièmement, le gBGC a été proposé comme modèle pouvant expliquer l'apparition des isochores [Eyre-Walker, 1993, Galtier et al., 2001]. Nous avons déjà vu, dans la sous-partie précédente, de nombreuses preuves de l'existence du gBGC. En plus de ces signatures, le gBGC a plusieurs propriétés qui permettent d'expliquer l'évolution des isochores et qui en font le meilleur

modèle, à ce jours, pour expliquer l'origine de ces structures [Duret et al., 2006, Duret & Galtier, 2009a]. A l'échelle des eucaryotes, on sait depuis plusieurs années que le patron de mutation est biaisé vers AT [Lynch, 2010]. Ainsi, à l'équilibre mutationnel (*i.e.* en l'absence de sélection et gBGC), le taux de GC est bas (32% chez l'homme). La dynamique particulière du gBGC associé aux points chauds de recombinaison permet donc d'expliquer la structure des isochores et le fait que les isochores riches en GC sont hétérogènes alors que les isochores pauvres en GC ont des compositions en base similaires [Clay & Bernardi, 2005]. En effet, considérons une région du génome exempte de points chauds de recombinaison sur une longue période évolutive, on attend à ce qu'elle soit à l'équilibre mutationnel, elle constitue donc un isochore pauvre en GC. Lorsqu'un points chaud est recruté en son sein, le gBGC agit rapidement et intensément pendant plusieurs générations. Cette région voit alors son taux de GC augmenter, cependant, il atteint rarement le GC* car le point chaud disparait rapidement. A ce moment, le taux de GC est ramené vers l'équilibre mutationnel. Cette décroissance du GC est lente car elle se fait sous l'effet de la dérive uniquement. Ainsi, à un moment donné, à l'échelle du mégabase, une région riche en points chauds présents et passés (éteints) aura un taux de GC fort (isochore riche en GC). La majorité des points chauds étant éteints, le taux de GC décroit : c'est l'érosion des isochores. De plus, sous ce modèle, le fait que la richesse en GC dépende à la fois de la force, du nombre et de la durée de vie des points chauds explique pourquoi le isochores GC riches sont hétérogènes [Duret & Arndt, 2008]. Enfin, sous ce modèle, la population homogène des isochores pauvres en GC est le résultat de l'absence de points chauds dans ces régions qui restent donc à l'équilibre mutationnel. Encore une fois, pour mieux valider ce modèle, il est utile de pouvoir quantifier l'intensité des BGC : gBGC et dBGC qui est responsable de l'extinction des points chauds. C'est ce que nous proposons au chapitre III p. 131. Chez le poulet, plusieurs observations suggèrent que les isochores ne s'érodent pas [Webster et al., 2006, Capra & Pollard, 2011]. Des auteurs ont suggéré que ceci était en lien direct avec le fait que le génome des oiseaux est peu réarrangé et ne possède pas de protéine PRDM9 ce qui conduirait à une faible dynamique de la recombinaison dans ces génomes. La faible mobilité des points chauds entraînerait à son tour des hétérogénéités de gBGC à l'échelle du génome et sur de longs temps évolutifs aboutissant à un renforcement et non une érosion des isochores [Capra & Pollard, 2011, Mugal et al., 2013].

Le gBGC semble donc être le scénario le plus vraisemblable pour expliquer l'émergence et l'érosion (ou le maintien) des isochores chez les amniotes ce qui démontre l'importance de son empreinte dans l'évolution des génomes de ces espèces.

### Le débat sur les séquences à évolution accélérées

Grâce au développement des techniques de séquençage de masse, de nombreux génomes d'espèces proches de l'homme et d'autres vertébrés sont disponibles. Ceci a permis de mettre en évidence le fait que ces génomes sont constitués d'une part importante ($> 4\%$) de séquences non codantes fonctionnelles : régulateurs, ARN etc... Ces éléments, sont conservés entre espèces distantes, sur plusieurs centaines de millions d'année, ce qui traduit une pression de sélection négative qui tend à "protéger" les régions fonctionnelles du génome des mutations délétères. La conservation est donc une signature du caractère fonctionnel de ces régions. Au milieu des années 2000, plusieurs groupes ont recherché indépendamment des régions conservées à l'échelle des vertébrés à l'exception de l'homme. Le but premier de ces études était de trouver, parmi ces régions, celles qui font l'unicité de l'homme à l'échelle moléculaire. Ces régions fonctionnelles ont été appelées HAR (régions à évolution accélérées chez l'homme, Human Accelerated Regions en anglais) ou HACNS (idem mais spécifiquement dans les régions non-codantes, Human Accelerated Conserved Noncoding Sequences) [Pollard et al., 2006a, Prabhakar et al., 2006, Bird et al., 2007] (voir [Bird et al., 2007] pour une comparaison des résultats de ces trois études). Dans la suite, nous utiliserons le terme de "HAR" pour faire référence à l'ensemble des régions mises en évidence dans ces différentes études. Le fait que les HAR ne soient pas conservés (*i.e.* accélérés) chez l'homme a été interprété comme une signature de sélection adaptative spécifique de la lignée humaine.
Cependant, il est possible que ces accélérations soient le résultat de l'action épisodique du gBGC dans la lignée humaine [Galtier & Duret, 2007, Duret & Galtier, 2009b]. Prenons ici l'exemple de *HAR1* et *HAR2* (*HACNS1*) qui sont les deux loci de ce type qui présentent les plus forts taux de substitution chez l'homme [Pollard et al., 2006b, Prabhakar et al., 2006]. *HAR1* contient la séquence d'un ARN non traduit exprimé lors du développement cortical du cerveau et capable de former une structure secondaire après transcription [Pollard et al., 2006b]. *HAR2* contient, lui, une séquence régulatrice capable de cibler l'expression d'un gène rapporteur dans les futurs membres antérieurs d'embryons de souris avec des patrons différents selon que la copie humaine ou celle présente chez les autres amniotes est utilisée [Prabhakar et al., 2008]. Ces deux HAR semblent donc être deux bons candidats dans la quête des régions qui ont façonné les caractères propres à l'homme. Cependant, une analyse plus fine des substitutions affectant ces régions dans la lignée humaine montre que le gBGC est un meilleur candidat pour expliquer les patrons de substitution que la sélection positive. En effet, sur les 18 substitutions détectées dans *HAR1* (16 dans *HAR2*), 18 (14 pour *HAR2*) sont

WS (on compte aussi 2 substitutions de GC vers CG dans *HAR2*). Ce patron hautement biaisé vers GC est compatible avec le gBGC et inattendu sous un modèle de sélection positive [Galtier & Duret, 2007, Duret & Galtier, 2009b]. De plus, les HAR ont tendance à être localisés dans des régions à fort taux de recombinaison [Pollard et al., 2006a], par exemple, *HAR2* se trouve dans une région sous-télomérique du chromosome 2 où le taux de CO est 2.8 fois supérieur au taux moyen des autosomes le plaçant dans les 7% du génome qui recombinent le plus [Duret & Galtier, 2009b]. Ajoutons à cela que la région de *HAR1* ne présente pas de dépression significative de diversité compatible avec un balayage sélectif mais montre, au contraire, plusieurs substitutions biaisées vers GC sur plus d'un kb au delà de ses limites, ce qui est compatible avec le gBGC [Galtier & Duret, 2007]. De plus, grâce à la fusion de la séquence de *HAR2* à un gène rapporteur, un groupe a montré que la copie humaine entrainait le même patron d'expression qu'une copie dont le site d'interaction avec le facteur de transcription était délété [Sumiyama & Saitou, 2011]. Les substitutions inférées dans *HAR2* chez l'homme correspondent donc à une perte de fonction qui, si elle était adaptative, aurait plus vraisemblablement émergé *via* une délétion que *via* l'accumulation de 16 substitutions. Enfin, le fait que ces deux loci montrent une activité fonctionnelle particulière n'en fait pas des produits de la sélection naturelle (voir le récent débat autour du projet ENCODE à ce sujet [Bernstein et al., 2012, Doolittle, 2013, Graur et al., 2013, Kellis et al., 2014]). Il est possible que le gBGC ait fixé des changement d'acides aminés neutres, voire délétères (*cf.* paragraphe suivant) conduisant à des changement fonctionnels [Duret & Galtier, 2009b] sans augmentation de fitness. Plusieurs substitutions enregistrées dans *HAR1* permettent d'ailleurs de compenser l'effet d'autres substitutions au même locus garantissant ainsi la même structure secondaire à l'ARN produit et donc peu d'effets sur la fitness [Galtier & Duret, 2007]. A l'échelle de tous les HAR, un excès de substitution WS et de forts taux de recombinaison, montrent que le gBGC est la force majeure à l'origine de ces structures [Pollard et al., 2006a, Galtier & Duret, 2007]. Ceci a été confirmé, plus tard, grâce à la comparaison des spectres de DAF WS et SW des sites impliqués dans les substitutions observées chez l'homme [Katzman et al., 2010].

L'inférence de la sélection adaptative à partir de données de séquences est donc une entreprise périlleuse chez les organismes où le gBGC est présent. En effet, à cause du gBGC, l'hypothèse nulle d'évolution neutre ne peut plus simplement être rejetée lorsque l'on observe un fort taux de substitution (non-synonyme) dans une branche. Ainsi, Galtier et collaborateurs ont proposé d'étendre cette hypothèse nulle en énonçant quatre observations qui indiquent que l'empreinte du gBGC ne peut pas être rejetée dans l'évolution d'une séquence [Galtier & Duret, 2007, Duret & Galtier, 2009a] :

(i) Un excès de substitutions WS par rapport aux substitutions SW associé, éventuellement, à une ségrégation à plus haute fréquence des mutations WS par rapport aux mutations SW.

(ii) Le locus est dans une région à fort taux de recombinaison.

(iii) Les substitutions concernent des sites fonctionnels (exons, ARN et éléments régulateurs) et des sites non fonctionnels (introns, régions intergéniques) du voisinage génomique.

(iv) Le patron de polymorphisme entourant la région n'est pas réduit indiquant une absence de balayage sélectif.

Le gBGC peut donc mimer la sélection positive dans de nombreux tests classiques basés sur l'étude des patrons de substitution [Berglund et al., 2009, Ratnakumar et al., 2010]. De plus, à l'échelle moléculaire, le gBGC a dirigé une partie majeure de l'évolution humaine de manière neutre. Examinons maintenant comment le gBGC peut avoir des conséquences sur la fitness des génomes.

### Le talon d'Achille des génomes

Nous avons déjà vu que le gBGC est capable d'augmenter la probabilité de fixation de mutations WS dans un contexte où la recombinaison est forte (*cf.* I.C.1. p. 73). Puisque le gBGC affecte indifféremment les régions codantes et non-codantes, il est donc possible qu'il promeuve la fixation de mutations délétères entrant ainsi directement en compétition avec la sélection naturelle. Sous un modèle théorique simple, une mutation WS (respectivement SW) affectée d'un coefficient de sélection $s$, qui se trouve dans une région soumise au gBGC d'intensité $b$ se comporte comme si elle était sélectionnée avec un coefficient $s+b$ (respectivement $s-b$). Ainsi, si une mutations WS est délétère, elle peut tout de même se répandre dans la population si $b$ surpasse $s$ et que la dérive est relativement faible. En 2009, Galtier et collaborateurs ont utilisé ce modèle pour mesurer l'impact du gBGC sur une population théorique aux caractéristiques proches de celles mesurées chez l'homme : $N_e = 10000$ et taux de mutation $\mu = 10^{-9}$ [Duret & Galtier, 2009a]. La simulation donne les proportions de substitutions délétères, neutres et adaptatives attendues après $10^7$ générations sur un gène d'un exon de 1000 bp affecté ou non par un gBGC d'intensité $N_e b = 1$ (Figure I.34). Avec le gBGC, la proportion de substitutions délétères observée est quadruplée montrant ainsi que le gBGC peut-être mal adaptatif.
Plusieurs études confirment ces attendus théoriques grâce à l'étude des patrons de substitution. Tout d'abord, revenons sur le cas du gène *Fxy* (*cf.*

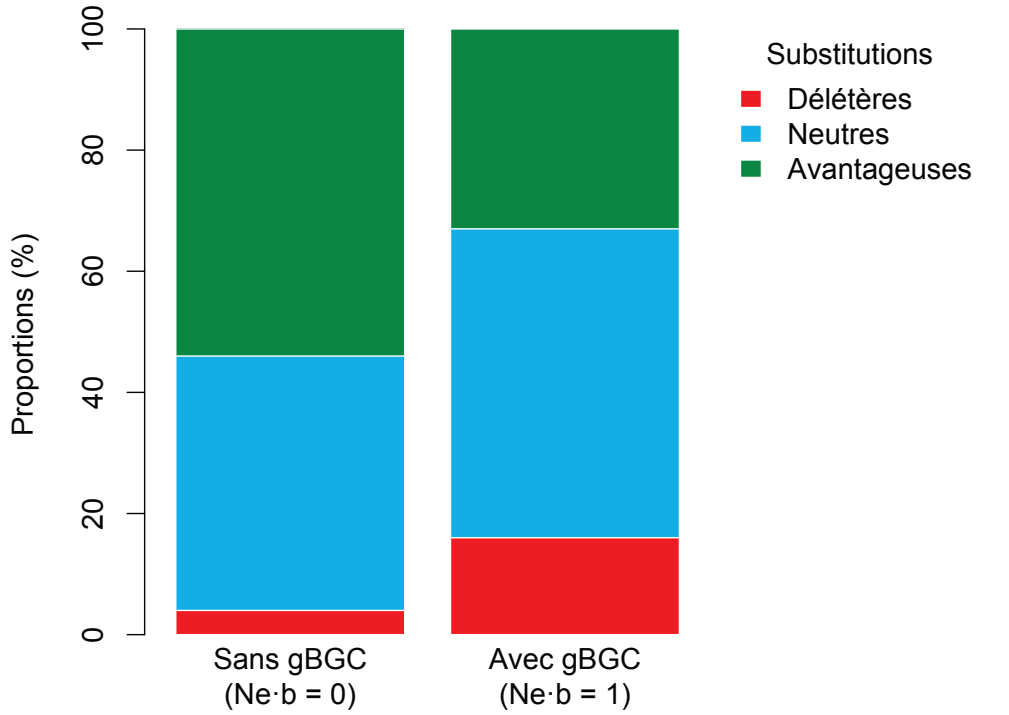



FIGURE I.34 : Proportions de substitutions délétères, neutres et adaptatives attendues avec ou sans gBGC dans une population de taille efficace 10000. D'après [Duret & Galtier, 2009a].

I.C.2. p. 75). Ce locus a subi 28 substitutions non-synonymes chez la souris depuis sa translocation dans la PAR (*i.e.* il y a 1 à 3 Ma) (Figure I.26 p. 76). Toutes ces substitutions sont WS et sont donc probablement dues au gBGC. Le fait que chez l'homme ou chez le rat on ne compte que 3 substitutions non-synonymes à ce locus en 80 Ma montre qu'il est d'autre part très conservé et évolue principalement sous sélection purificatrice. Ainsi, le gBGC agissant chez la souris sur *Fxy* a surpassé cette pression de sélection en favorisant la fixation de mutations délétères [Galtier & Duret, 2007]. Sur le même modèle et pour les mêmes raisons, les loci *HAR1* et *HAR2* ont aussi accumulé des mutations délétères à cause du fort gBGC agissant dans leur voisinage [Duret & Galtier, 2009a].

Afin de démontrer le rôle du gbGC dans la fixation de mutations délétères, Galtier et collaborateurs se sont intéressés au rapport du taux de substitutions non-synonymes sur le taux de substitutions synonymes ($^{d_N}/_{d_S}$ ) des exons accélérés (en terme de changements d'acides aminés) chez les primates [Galtier et al., 2009]. L'analyse du $^{d_N}/_{d_S}$ est couramment utilisée pour inférer

l'action de la sélection positive dans une lignée [Yang, 2007]. Cependant, dans une zone à fort taux de recombinaison et au patron de substitution biaisé vers GC, un fort $^{d_N}/_{d_S}$ a plus de chance de résulter de l'action du gBGC. Or, sur les 9 exons accélérés par le gBGC de plus de 600pb mis en évidence dans [Galtier et al., 2009], 4 montrent un $^{d_N}/_{d_S}$ significativement plus grand dans la branche où a eu lieu l'accélération que dans le reste de l'arbre. Ceci indique que, dans cette lignée, le gBGC est assez fort pour surpasser la sélection purificatrice et donc favoriser la fixation de mutations faiblement délétères. D'autres analyses du même type ont conduit à des conclusions similaires [Haudry et al., 2008, Berglund et al., 2009, Ratnakumar et al., 2010].
Une autre manière de montrer le caractère mal-adaptatif du gBGC est d'exploiter les données de polymorphismes et les spectres de DAF (voir plus haut). En 2011, Necşulea et collaborateurs ont montré l'implication du gBGC dans la fixation de mutations associées à des phénotypes pathologiques chez l'homme [Necşulea et al., 2011]. Pour cela, des données populationnelles (SNP) et des bases de données recensant des variants à risque ou responsables de maladies génétiques ont été croisées. Les spectres de DAF des mutations WS et SW touchant spécifiquement des zones à forts taux de recombinaison ont été comparés. Lorsque les sites intergéniques ou les mutations synonymes (et non-synonymes) sont analysées, le décalage des spectres est en accord avec le modèle du gBGC : les mutations WS ségrègent à plus forte fréquence que les mutations SW (Figure I.35 A & B). De plus, cet effet est aussi présent (voire plus fort) lorsque l'on considère uniquement les mutations à risque (Figure I.35 C) ou responsables de maladies génétiques (Figure I.35 D). Le gBGC favorise donc la fixation d'allèles délétères chez l'homme.

Ainsi, le gBGC peut être mal-adaptatif. A ce titre il a été pointé comme le "talon d'Achille des génomes" par Galtier et Duret [Galtier & Duret, 2007]. Cependant, il est difficile d'expliquer le fait que ce processus soit observé chez plusieurs organismes à l'échelle des eucaryotes alors qu'il joue contre la sélection naturelle. Ceci pose la question de l'origine évolutive du gBGC.

En complément sur cette partie, voir les articles de revue suivants : [Eyre-Walker & Hurst, 2001, Marais, 2003, Duret et al., 2006, Galtier & Duret, 2007, Galtier et al., 2009].

### L'origine du gBGC

La question du *pourquoi* du gBGC est encore peu documentée. Deux pistes semblent cependant se dégager. Les patrons de mutations étant biaisés vers AT à l'échelle des eucaryotes [Lynch, 2010], le gBGC aurait pu évoluer de manière à compenser ce biais. Les deux hypothèses suivantes se basent sur

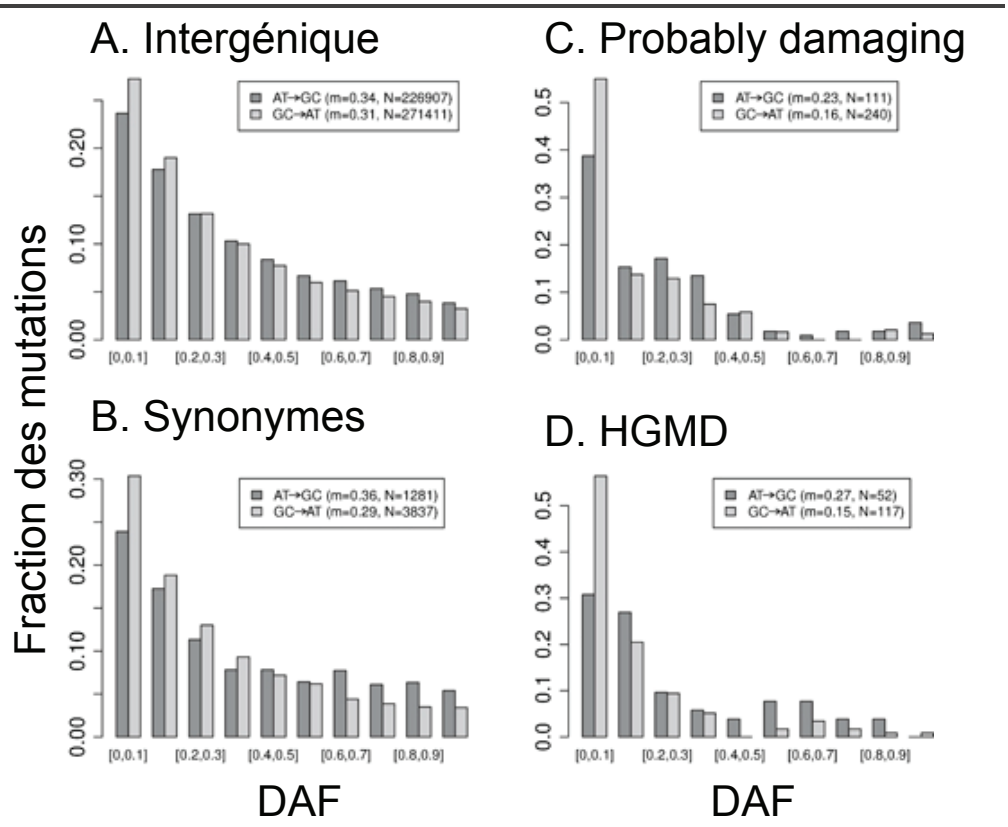



FIGURE I.35 : Spectres de DAF pour des mutations WS (AT vers GC) et SW (GC vers AT) dans des régions à forts taux de recombinaison chez l'homme. Pour chaque subset la médiane de la DAF est donnée (m), ainsi que le nombre de mutations analysées (N). (A) Mutations détectées dans les régions intergéniques du génome. (B) Mutations synonymes. (C) Mutations détectées comme "Probably damaging" par la base PolyPhen [Sunyaev et al., 2001]. (D) Mutations associées à des maladies génétiques dans la base HGMD (Human Gene Mutation Database) [Stenson et al., 2009]. Adapté de [Necşulea et al., 2011].

cette idée. Sous la première hypothèse, le gBGC serait bénéfique à l'individu car les mécanismes qui sont à son origine seraient aussi capable de limiter les mutations somatiques en favorisant la réparation des erreurs de réplication vers GC. Sous la seconde hypothèse, le gBGC serait bénéfique à la population car il permettrait une baisse du fardeau mutationnel en limitant la probabilité de fixation des allèles mutants AT. Cette hypothèse est soutenue par des résultats théoriques, dans une gamme de paramètres restreints cependant [Glémin, 2010]. Afin de confirmer ou d'invalider ces deux modèles, il est nécessaire de se concentrer sur les mécanismes moléculaires, encore peu connus, qui donnent naissance au gBGC. C'est ce que nous proposons dans la première partie de nos résultats (chapitre II p. 101).

Chapitre *II*

# Etude des mécanismes moléculaires sous-tendant le gBGC chez la levure

# II.A. Présentation de l'étude

Afin de déterminer l'origine évolutive du gBGC, il est utile de connaitre son origine moléculaire. En effet, il faut tout d'abord délimiter les bornes d'action du processus conduisant au gBGC pour déterminer le contexte dans lequel il a pu émerger. Comme vu plus haut (*cf.* I.C.3. p. 97) les principales hypothèses avancées jusque là font intervenir le fait que le gBGC pourrait contrebalancer le biais mutationnel, universellement biaisé vers AT à cause de la désamination des cytosines [Lynch, 2010]. Cependant, le biais mutationnel touche l'ensemble du génome, de manière continue. Comment donc le gBGC, qui est lui localisé et dynamique, permettrait de contrecarrer efficacement cette force au point qu'il soit sélectionné malgré les effets mal-adaptatifs qu'il peut engendrer ? Dans ce chapitre, nous présentons la première partie des résultats de ce travail de thèse qui vise à déterminer le mécanisme moléculaire sous-jacent au phénomène de gBGC.

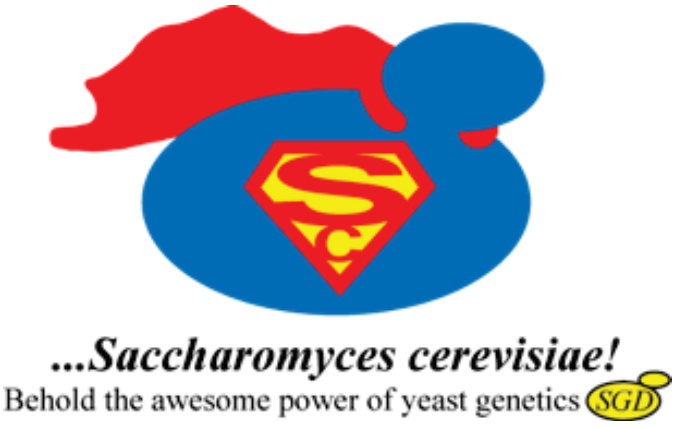
Le dBGC est lié à l'initiation de la recombinaison : le DSB. Ainsi, il se place en amont de la différenciation des voies donnant lieu aux CO et NCO. Il est possible que le gBGC soit un sous-produit du dBGC si certains sites de cassure sont riches en AT, même si cela ne se retrouve pas dans les motifs mis en évidence dans les points chauds de recombinaison (*cf.* I.B.2. p. 52). Il est donc possible que dBGC et gBGC aient une origine moléculaire commune. Récemment Odenthal-Hesse et collaborateurs ont mis en évidence qu'à deux points chauds de recombinaison chez l'homme, la transmission d'allèles GC par rapport à leur homologue AT était significativement favorisée uniquement dans les tracts de conversion associés à des NCO [Odenthal-Hesse et al., 2014] (Tableau I.1). Bien que ce comportement ne soit observé qu'à deux loci, il montre qu'il est possible qu'un biais de conversion vers GC soit introduit après le DSB. D'autres résultats de ce type, étendus à l'échelle du génome, permettraient de déterminer si il s'agit là d'une signature NCO-spécifique du gBGC. Les auteurs proposent, comme d'autres avant eux [Galtier et al., 2001, Birdsell, 2002, Marais, 2003], l'hypothèse selon laquelle le gBGC serait dû à un biais dans la réparation des mésappariements d'hétéroduplexes formés en méiose [Odenthal-Hesse et al., 2014]. Marais, propose le système BER (Base Excision Repair) comme meilleur candidat car il montre un bias de réparation des mésappariements en faveur de GC dans différents systèmes cellulaires [Marais, 2003]. Cependant, comme vu en introduction, l'action du BER n'a encore jamais été avérée en méiose.

La carte de recombinaison à haute résolution de Mancera est une opportunité unique d'étudier le biais de conversion et sa répartition dans les différents types de tracts de conversion (simples/complexes, CO/NCO, longs/courts, riches en GC/AT...) chez la levure [Mancera et al., 2008] (*cf.* I.A.3. p. 41

et Annexe A p. 219). L'étude qui suit consiste donc à pister le signal de gBGC (déjà enregistré par Mancera grâce à cette même carte) à travers les différents types de tracts de conversion afin de déterminer si il peut-être associé à une/des étape(s) particulière(s) du processus de recombinaison. Ceci permettra de proposer une origine évolutive au gBGC.

## II.B. Le gBGC est associé spécifiquement aux CO chez la levure : mécanismes et enjeux évolutifs

*Voir article joint page suivante.*
Référence : [Lesecque et al., 2013]
Les informations supplémentaires sont disponibles à la section II.C. p. 115.




*La levure, un organisme modèle super-puissant en génétique ! - par La Saccharomyces Genome Database (www.yeastgenome.org ).*

# GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance

Yann Lesecque,[1] Dominique Mouchiroud,[1] and Laurent Duret*[,1]
[1]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université de Lyon, Université Lyon 1, Villeurbanne, France
***Corresponding author:** E-mail: laurent.duret@univ-lyon1.fr.
**Associate editor:** Kenneth Wolfe

## Abstract

**GC-biased gene conversion (gBGC) is a process associated with recombination that favors the transmission of GC alleles over AT alleles during meiosis. gBGC plays a major role in genome evolution in many eukaryotes. However, the molecular mechanisms of gBGC are still unknown. Different steps of the recombination process could potentially cause gBGC: the formation of double-strand breaks (DSBs), the invasion of the homologous or sister chromatid, and the repair of mismatches in heteroduplexes. To investigate these models, we analyzed a genome-wide data set of crossovers (COs) and noncrossovers (NCOs) in *Saccharomyces cerevisiae*. We demonstrate that the overtransmission of GC alleles is specific to COs and that it occurs among conversion tracts in which all alleles are converted from the same donor haplotype. Thus, gBGC results from a process that leads to long-patch repair. We show that gBGC is associated with longer tracts and that it is driven by the nature (GC or AT) of the alleles located at the extremities of the tract. These observations invalidate the hypotheses that gBGC is due to the base excision repair machinery or to a bias in DSB formation and suggest that in *S. cerevisiae*, gBGC is caused by the mismatch repair (MMR) system. We propose that the presence of nicks on both DNA strands during CO resolution could be the cause of the bias in MMR activity. Our observations are consistent with the hypothesis that gBGC is a nonadaptive consequence of a selective pressure to limit the mutation rate in mitotic cells.**

***Key words:*** **recombination, biased gene conversion, crossover, meiotic drive.**

## Introduction

In many eukaryotes, recombination is required for the proper segregation of chromosomes during meiosis. This process involves the programmed formation of double-strand breaks (DSBs), which are subsequently repaired by using homologous sequences as a template. It is generally accepted that the profound *raison d'être* of meiosis is to enhance the efficacy of natural selection by allowing the formation of new combinations of alleles via this process of recombination. Thus, asexual taxa (which cannot create new haplotypes by recombination) are expected to be evolutionary dead ends, because of their reduced potential for adaptation (for review, see Coop and Przeworski [2007]).

Recently, many studies have shown that besides its fundamental impact on selection efficacy, recombination also strongly contributes to genome evolution via the nonadaptive process of biased gene conversion (BGC) (for review, see Duret and Galtier [2009] and Webster and Hurst [2012]). Gene conversion is a process intrinsically associated with recombination that results in the nonreciprocal transfer of genetic information between the two recombining sequences. This process is said to be biased if one of the two alleles has a higher probability to be the donor than its homolog. BGC tends to raise the frequency of the donor allele in the pool of gametes and therefore leads to increase its probability of fixation in the population. It is a nonadaptive process, because the spread of one allele through BGC is independent of its effect on fitness. However, its impact on the dynamics of allele frequency within populations is very similar to that of directional selection (Nagylaki 1983). Different lines of evidence indicate that in many eukaryotes, BGC tends to favor the transmission of GC alleles in AT/GC heterozygotes (for review, see Duret and Galtier [2009] and Webster and Hurst [2012]). In mammals, it has been shown that gBGC (i.e., GC-favoring BGC) is the main determinant of the evolution of genomic base composition (Meunier and Duret 2004; Duret and Arndt 2008; Katzman et al. 2011; Auton et al. 2012), and there is indirect evidence that this process is widespread in eukaryotes (Capra and Pollard 2011; Escobar et al. 2011; Pessia et al. 2012). Moreover, it has been shown that gBGC can interfere with natural selection and lead to the fixation of deleterious alleles (Galtier and Duret 2007; Berglund et al. 2009; Galtier et al. 2009; Glémin 2010, 2011; Ratnakumar et al. 2010; Necşulea et al. 2011). However, despite its major impact on genome evolution, the molecular mechanisms leading to gBGC are still unknown.

Much of our knowledge of the molecular mechanisms of meiotic recombination in eukaryotes has come from the study of yeasts (for review, see de Massy [2003]). Recombination is initiated by the formation of DSBs followed by 5′- to 3′-end resection (Smith and Nicolas 1998; Krogh and Symington 2004). DSBs are then repaired, using homologous sequences as a template, either from the sister chromatid or, more frequently, from the nonsister chromatid (the

homolog). Recombination events between homologs can lead to the exchange of flanking regions (i.e., crossovers [COs]) or not (i.e., noncrossover [NCO] recombination events). The two types of events result from different recombination pathways (fig. 1). In budding yeast (*Saccharomyces cerevisiae*), NCOs result principally from the synthesis-dependant strand annealing pathway, and secondarily from double Holliday junction (dHj) dissolution, whereas COs result from dHj resolution (class I COs) and from the Mus81 pathway (class II COs) (McMahill et al. 2007; Martini et al. 2011). In all cases, the repair of DSBs by the homolog involves the formation of heteroduplex DNA, with one DNA strand coming from the broken chromosome and the other from the intact template. When homologs are not identical, mismatches are formed in this heteroduplex, and their repair leads to the conversion of one allele by the other. The segment of the chromosome affected by a conversion event is called the conversion tract. Mancera et al. (2008) recently published a high-resolution recombination map that allowed a very detailed genome-wide analysis of conversion tracts in *S. cerevisiae*. The median length of conversion tracts is 2 kb for COs and 1.8 kb for NCOs. They found that the majority of conversion tracts (89% for COs and 97% for NCOs) are "simple," that is, with one single-donor haplotype along the whole tract (Mancera et al. 2008). They notably demonstrated that conversion events overlapping AT/GC heterozygous sites lead to a significant overtransmission of the GC allele (1.3% greater than expected under the null hypothesis of Mendelian transmission), thus providing the first direct evidence of gBGC in a eukaryote (Mancera et al. 2008).

Several hypotheses can be proposed concerning the molecular mechanisms responsible for gBGC. First, the analysis of gene conversion tracts, in yeasts or in mammals, indicates that in most cases, gene conversion occurs from the intact chromosome toward the broken one (Nicolas et al. 1989; Mancera et al. 2008; Webb et al. 2008). Thus, if in an AT/GC heterozygote, DSBs occur more frequently on the AT-richer haplotype, this could lead to the overtransmission of the GC allele. This model is hereafter referred to as the "initiation bias" hypothesis. An alternative model is that gBGC could result from the activity of the mismatch repair (MMR) machinery. MMR plays a major role during recombination, not only for the repair of mismatches in heteroduplex DNA but also for the choice of the DNA template to be used to repair the DSB. Indeed, during the process of invasion of the homologous chromosome by the single-stranded 3′-overhang, MMR is able to sense the mismatches present in the heteroduplex and to reject the invading strand if the level of sequence divergence is too high (Hunter et al. 1996; Chen and Jinks-Robertson 1999). This activity is crucial to avoid recombination between nonallelic loci (ectopic recombination) (Surtees et al. 2004). Current models suggest that in the cases where MMR prevents the invasion of the homolog, DSBs get subsequently repaired by using the sister chromatid (Martini et al. 2011), which leads to Mendelian transmission, without any conversion (fig. 1). In theory, it is possible that the decision to reject the invading strand or to repair the mismatch depends on the nature of the allele present on the invading strand: If strands carrying AT alleles were less prone to be rejected than those carrying GC alleles, then the

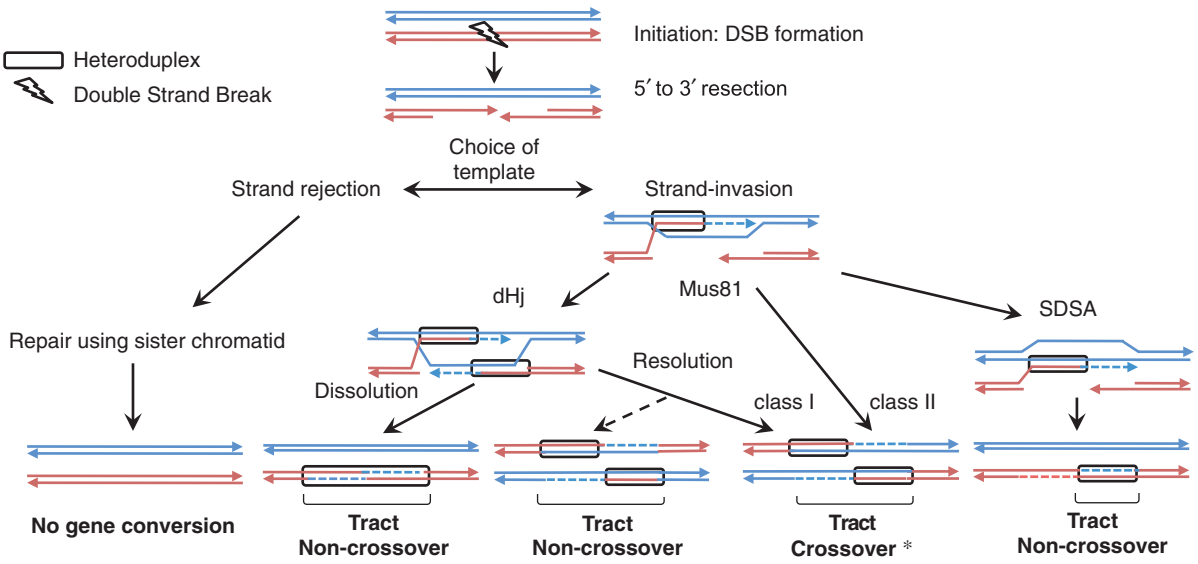



**Fig. 1.** Canonical model of meiotic recombination in *Saccharomyces cerevisiae*. For simplicity, only two homologous dsDNA molecules are represented (one red and one blue). Meiotic recombination is initiated by the formation of a DSB (here represented by a flash on the red haplotype), followed by 5′- to 3′-end resection. The DSB is subsequently repaired using as a template either the sister chromatid (not shown here; left part) or the homolog (here represented in blue; right part). There exist several DSB repair pathways, which, when the homolog is used as a template, can lead to COs or NCOs. Current models indicate that NCOs result principally from the synthesis-dependant strand annealing (SDSA) pathway and secondarily from double Holliday junction (dHj) dissolution, whereas COs result from dHj resolution (class I) and from the Mus81 pathway (class II) (Martini et al. 2011). The resolution of dHj into NCOs is represented by a dashed harrow, to indicate that this is a minor pathway. Dashed lines (blue and red) represent newly synthesized DNA and boxes show heteroduplex associations during the whole process. The * symbol next to the CO product refers to figure 4.

former would have more opportunities to get converted, which would lead to a conversion bias in favor of GC alleles. An alternative hypothesis is that MMR could cause gBGC via its activity in the repair of mismatches in heteroduplex DNA. The directionality of the repair by MMR depends on the presence of nicks flanking the mismatch (Jiricny 2006) and is not known to be biased toward the GC allele. It is, however, possible that a weak bias, such as the one causing gBGC in *S. cerevisiae*, might have remained unnoticed. Finally, we and others proposed that gBGC could be caused by the base excision repair (BER) machinery (Memisoglu and Samson 2000). Indeed, although MMR is the prominent repair system active during recombination (Evans and Alani 2000; Hoffmann and Borts 2004; Surtees et al. 2004; Jiricny 2006), there is evidence that other systems contribute to the repair of mismatches in heteroduplex DNA (Coïc et al. 2000). Given that BER is intrinsically biased toward GC, it is a priori expected that if this repair machinery is active on heteroduplex DNA during meiotic recombination, then it should induce gBGC (Brown and Jiricny 1989; Galtier et al. 2001; Birdsell 2002; Marais 2003). One clear difference between BER and MMR is the length of the region affected by the repair: Although MMR involves DNA resynthesis over hundreds of base pairs (i.e., about the size of conversion tracts) (Holmes and Clark 1990; Thomas et al. 1991), BER leads only to short-patch repair (1–13 bp) (Memisoglu and Samson 2000). Given the length of gene conversion tracts (~2 kb on average), if some single-nucleotide polymorphism (SNP) conversion events are driven by BER, then the conversion of these SNPs should occur independently of the conversion of neighboring SNPs. Thus, although MMR is expected to produce predominantly simple conversion tracts, BER—if active during recombination—is expected to lead frequently to complex conversion tracts (i.e., tracts involving conversion events from both parental haplotypes). Hence, if BER is responsible for the conversion bias, then gBGC should be much stronger among complex conversion tracts compared with simple conversion tracts.

To try to distinguish between the different processes possibly responsible for gBGC (initiation bias, MMR, or BER), we decided to analyze the high-resolution recombination data published by Mancera et al. (2008). We demonstrate that in *S. cerevisiae*, gBGC is associated with long-patch DNA repair and is specific of CO events. We further show that gBGC is associated with longer conversion tracts and that the conversion bias depends on the nature of mismatches at the boundaries of the tract. These observations are not consistent with the initiation bias and BER models and suggest that gBGC is caused by MMR.

## Results

To analyze gene conversion tracts in yeast, we used the high-resolution recombination data published by Mancera et al. (2008). These data were obtained by genotyping tetrads resulting from 46 meioses, in a diploid hybrid of two wild-type *S. cerevisiae* strains (S96 and YJM789). Several other similar data sets have been published (Winzeler et al. 1998; Chen et al. 2008; Qi et al. 2009). However, the Mancera data set is currently the only one to provide exhaustive genotyping data (i.e., almost all the sites that differ between the two strains have been genotyped) for such a large number of meioses. The median distance between two consecutive markers is 78 bp. We analyzed all recombination events associated with detectable conversion tracts (2,884 COs and 2,090 NCOs). On average, conversion tracts overlap nine SNPs. Each of these SNP sites was genotyped in the two resulting spores. Thus, in total, 89,538 SNP sites involved in a conversion event have been genotyped. To test whether gene conversion shows a bias in favor of GC or AT allele, we focused on the subset of sites that correspond to AT/GC heterozygotes in the parental hybrid (87% of the total set of SNPs involved in conversion events). For this set of sites, we counted the proportion of GC alleles in the offspring ($x$). The existence of a conversion bias was tested by comparing $x$ to the Mendelian expectation (50%), with a one-sample proportion test (see Materials and Methods). The intensity of the conversion bias in favor of GC alleles was measured by the coefficient $b = 2x - 1$. (NB: We chose this expression because it is equivalent to the definition of the selection coefficient of a semidominant mutation, see Nagylaki [1983].) In agreement with previous results (Mancera et al. 2008), we observed a significant conversion bias toward GC alleles ($b = 0.013$, $P < 10^{-3}$; table 1; NB: The properties of the conversion tracts that we studied are summarized in supplementary table S1, Supplementary Material online).

**Table 1.** Conversion Bias Toward GC Bases for AT/GC SNPs Involved in a Recombination Event.

| Conversion Tract Type | Number of Genotyped SNP Sites with AT/GC Polymorphism | Conversion Bias Toward GC Bases ($b$)[a] | $P$[a] |
|---|---|---|---|
| All | 77,901 | 0.013 | <0.001 |
| Simple | 64,898 | 0.014 | <0.001 |
| Complex | 13,003 | 0.008 | 0.36 (NS) |

NOTE.—NS, nonsignificant.
[a]One-sample proportion test.

### Transmission Biases in Simple and Complex Conversion Tracts

If BER is the unique cause of gBGC, it is expected that the conversion bias in favor of GC alleles should be much stronger among complex conversion tracts than among simple tracts. To test this prediction, we measured the conversion bias in favor of GC alleles separately for SNPs located in simple and complex conversion tracts. Interestingly, we observed that the conversion bias is not reduced among simple conversion tracts compared with complex ones (table 1). On the contrary, $b$ tends to be higher for SNPs located in simple conversion tracts (although the difference is not significant; two-sample proportion test). It should be noted that complex conversion tracts tend to be longer than simple ones (because, by definition, complex tracts must contain at least two SNPs, whereas simple tracts may contain just one SNP).

To test whether this ascertainment bias might have affected our conclusions, we repeated the analysis on SNP sites located in tracts overlapping at least five SNPs. The results remained unchanged (supplementary table S2, Supplementary Material online). Note that a large majority (83%) of SNPs involved in a recombination event are located in simple conversion tracts. Hence, quantitatively, the conversion bias in favor of GC alleles is essentially due to recombination events associated with simple conversion tracts. This observation is, therefore, not consistent with the predictions of the BER model.

## Conversion Biases Operate on Multiple Adjacent SNPs

In the above analysis, as in Mancera et al. (2008), the statistical significance of the conversion bias in favor of GC alleles was assessed under the assumption that each SNP conversion was an independent event. However, given that the observed conversion bias is essentially associated with simple conversion tracts, this assumption is clearly incorrect: All SNPs in a simple tract are converted together from the same donor haplotype. This nonindependence might lead to overestimate the statistical significance of conversion biases. To avoid this potential artifact, we reanalyzed conversion biases at the scale of the conversion event (i.e., a set of SNPs involved in a common conversion tract), focusing exclusively on simple conversion tracts ($N = 4{,}428$ recombination events). For each tract, we measured the difference in GC content between the two haplotypes involved in the conversion event ($\Delta$GC, supplementary fig. S1, Supplementary Material online). We selected all cases where one of the two haplotypes had a higher GC content than the other (i.e., $\Delta GC \neq 0$, $N = 3{,}676$ recombination events). These conversion tracts were said to have a "AT/GC-richer" polymorphism. Among the 7,352 corresponding haplotypes in the pool of spores, we observed a clear and statistically significant conversion bias in favor of the GC-richer haplotype (fig. 2; $b = 0.030$, $P = 0.01$), which confirms the existence of gBGC. Note that this conclusion remains when using a more stringent threshold to categorize AT/GC-richer haplotypes (supplementary text S2 and fig. S2, Supplementary Material online). In the rest of this article, to avoid any statistical artifact due to the nonindependence of SNPs located in a same tract, we analyzed conversion biases at the scale of conversion tracts (and not individual SNPs), excluding complex conversion tracts.

## gBGC Is CO Specific

Among CO recombination events, we observed a strong conversion bias in favor of the GC-richer haplotype (fig. 2; $b = 0.057$, $P = 3.6 \times 10^{-4}$). Interestingly, NCOs did not exhibit any conversion bias. This difference between COs and NCOs conversion biases was significant (fig. 2; $P = 0.014$). This indicates that gBGC observed in the whole data set is essentially driven by COs. This observation is not consistent with the initiation bias model, which predicts that gBGC should affect both COs and NCOs (see Discussion).

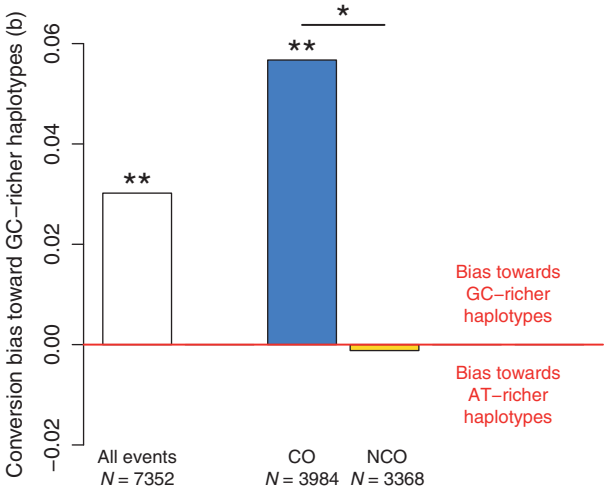



**FIG. 2.** Conversion bias toward GC-richer haplotypes. The conversion bias toward GC-richer haplotypes ($b$) was computed for simple conversion tracts, taken all together (white bar) or separating tracts associated with COs (blue bar) and NCOs (yellow bar). "$N$" is the number of genotyped haplotypes in each category. The red horizontal line indicates the Mendelian expectation ($b = 0$). Significant conversion biases are indicated by "*" for a $P$ value $\leq 0.05$ and "**" for a $P$ value $\leq 0.01$ (one-sample proportion test). The "*" between "CO" and "NCO" bars denotes the fact that the conversion bias toward GC-richer haplotypes is significantly different between CO and NCO events (two-sample proportion test).

## gBGC Is Driven by Mismatches Located at the Extremities of Conversion Tracts and Is Associated with Longer Tracts

The previous observations are inconsistent with the BER and initiation bias models. We, therefore, investigated further the hypothesis of a mismatch repair bias driven by MMR. The fact that gBGC is observed in simple conversion tracts is compatible with a role of MMR in gBGC. However, this hypothesis raises the question of how the MMR machinery would be able to distinguish AT-richer versus GC-richer haplotypes. It seems a priori unlikely that the MMR machinery could sense the global difference in GC content along the region, typically 2-kb long, affected by the conversion. Given that the directionality of the repair by MMR depends on the presence of flanking nicks (Jiricny 2006), we hypothesized that the bias could depend specifically on the nature of the mismatches found at the boundaries of the conversion tract, that is, those that are closest to the nicks flanking the heteroduplexes (fig. 4).

To test this prediction, we classified conversion tracts according to the nature of the first and the last SNPs of the tract. When one particular strain had a G or a C for first and last SNPs in the region corresponding to the conversion tract, whereas the other strain had a A or T at those positions, the first haplotype was called "$GC_f$" (which stands for GC-flanked haplotype) and the second "$AT_f$" (AT-flanked haplotype) (supplementary fig. S1, Supplementary Material online). These conversion tracts were said to have a "$GC_f/AT_f$ polymorphism." Similarly, conversion tracts with a GC/AT SNP at

one end and an AT/TA or GC/CG SNP at the other end were classified as "one-side $GC_f/AT_f$ polymorphisms." All other cases were excluded: When haplotypes are flanked by G or C at one extremity and A or T at the other one, it is impossible to define a conversion bias in this fashion because the two parental haplotypes are indistinguishable in term of GC/AT flanking SNPs.

Among CO-associated simple conversion events with $GC_f/AT_f$ polymorphism ($N = 1{,}104$ events, i.e., 38% of the set of CO-associated simple conversion tracts), we observed a strong conversion bias toward the $GC_f$ haplotype (fig. 3). For CO-associated simple conversion events with one-side $GC_f/AT_f$ polymorphism, the conversion bias toward the $GC_f$ haplotype was slightly weaker and only marginally significant

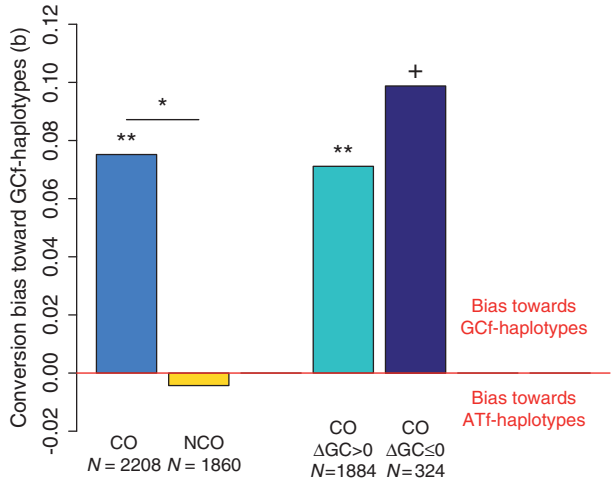



**Fig. 3.** Conversion bias toward GCf haplotypes. The conversion bias toward GCf haplotypes ($b$) was computed for simple conversion tracts, associated with COs (sky blue bar), NCOs (yellow bar), COs with $\Delta GC > 0$ (light blue bar), and COs with $\Delta GC \leq 0$ (dark blue bar). $\Delta GC$ is positive ($\Delta GC > 0$) when the GCf haplotype is globally richer in G + C than the ATf haplotype, it is negative or null otherwise ($\Delta GC \leq 0$). The red horizontal line indicates the Mendelian expectation ($b = 0$). Significant conversion biases are indicated by "**" for two-tailed one-sample proportion test with a $P$ value $< 0.01$ and "+" for one-tailed one-sample proportion test with alternative hypothesis "$b > 0$" and $P$ value $< 0.05$. The "*" between "CO" and "NCO" bars denotes the fact that the conversion bias toward GCf haplotypes is significantly different between CO and NCO events (two-sample proportion test, $P$ value $< 0.05$).

($b = 0.06$, $P = 0.07$, one-sample proportion test), probably because of limited sample size ($N = 435$ events). Note that for NCO recombination events, the conversion of $GC_f/AT_f$ haplotypes was unbiased (fig. 3), Thus, as noticed previously, the conversion bias appears to be CO specific. Interestingly, we noticed that for CO-associated simple conversion tracts, the length of tracts varies according to the direction of conversion: The median tract length (computed as the distance between the two most distal SNPs within the tract, for all conversion tracts overlapping at least two SNPs) is 1,322 bp for $GC_f$ conversion tracts, compared with 1,146 bp for other conversion tracts (Wilcoxon test, $P = 0.0017$). This difference is not observed for NCO recombination events (median tract length: 1,046 bp for $GC_f$ conversion tracts, compared with 1,044 bp for other conversion tracts).

The fact that we observed a conversion bias toward the $GC_f$ haplotype is consistent with the hypothesis that the bias depends on the nature of mismatches located at the extremities of the conversion tracts. However, given the way they are defined, $GC_f$ haplotypes also tend to be GC rich. Thus, the observed conversion bias toward the $GC_f$ haplotype might in fact be driven by a conversion bias toward the GC-richer haplotype (i.e., it might depend on the GC richness of the whole haplotype and not specifically on the SNPs located at the extremities). To test this hypothesis, we considered the subset of CO-associated simple conversion tracts with $GC_f/AT_f$ polymorphism for which the $GC_f$ haplotype is not richer in GC than the $AT_f$ haplotype ($\Delta GC \leq 0$ in fig. 3). If the bias toward $GC_f$ haplotypes was driven by the bias toward GC-richer haplotypes, one would expect the $GC_f$-conversion bias to be negative for these 162 events. In contradiction with this prediction, we observed a strong and positive bias in favor of $GC_f$ haplotypes ($b = 0.099$, fig. 3). This indicates that the conversion bias toward $GC_f$ haplotypes exists regardless of the difference in GC content between homologous haplotypes and that this conversion bias is predominant over the conversion bias toward the GC-richer haplotype. And indeed, when we categorized conversion tracts into AT/GC-richer haplotypes based on internal SNPs (i.e., ignoring the two SNPs at the extremities of the tract), then the conversion bias in favor of the GC-richer haplotype becomes much weaker and nonsignificant (table 2, supplementary table S3, Supplementary Material online). Thus, gBGC in yeast is essentially driven by a conversion bias in favor of $GC_f$ haplotypes. Given that in 85% of the cases, $GC_f$ haplotypes are GC richer

**Table 2.** Conversion Bias Toward GC-Richer Haplotypes among $AT_f/GC_f$ Polymorphism, Considering All SNPs, or Only Flanking or Internal SNPs.

| SNPs Considered to Classify Haplotypes as AT- or GC Richer | Number of Genotyped Haplotypes with AT/GC-Richer Polymorphism[a] | Conversion Bias Toward GC-Richer Haplotypes ($b$)[b] | $P$[b] |
|---|---|---|---|
| All | 1,114 | 0.070 | 0.02 |
| Flanking SNPs only | 1,246 | 0.101 | <0.001 |
| Flanking SNPs excluded | 1,072 | 0.034 | 0.28 (NS) |

NOTE.—NS, nonsignificant.

[a]Haplotypes were categorized in AT- or GC richer according to their difference in GC content, considering all SNPs in the tract or only the two flanking SNPs or only the SNPs that are not the two flanking SNPs.

[b]One-sample proportion test.

than $AT_f$ haplotypes, this $GC_f$ bias leads to an overall bias in favor of GC-richer haplotypes.

## Discussion

Our analyses confirm that in yeast, when a GC/AT heterozygote site is involved in the conversion tract of a recombination event, the GC allele has a higher probability to be transmitted than the AT allele (Mancera et al. 2008). We show that this pattern of non-Mendelian segregation is specific of CO recombination events. Furthermore, we found that gBGC is essentially associated with simple conversion tracts (i.e., where all SNPs within the tract are converted from the same donor haplotype) and that the conversion bias depends on the nature of mismatches located at the extremities of the conversion tract. Thus, it appears that the decision to repair distal mismatches in one direction or the other affects all other mismatches in the heteroduplex, independently of their base composition. This phenomenon of "conversion sweep" (by analogy to selective sweeps) therefore tends to decrease the strength of gBGC. Indeed, the bias observed at the scale of conversion events in favor of GC-flanked haplotypes ($b = 0.075$, fig. 3) is much stronger than gBGC observed among the whole set of SNPs ($b = 0.013$, table 1). The departure from Mendelian expectation observed in the whole set of SNPs (50.6% instead of 50%, $b = 0.013$) might seem relatively weak. However, similar to natural selection, the impact of gBGC on the probability of allele fixation depends on the effective population size ($N_e$) and becomes strong when $N_e b \gg 1$ (Nagylaki 1983). Given that 1% of the yeast genome is affected by gene conversion during each meiosis (Mancera et al. 2008), the genome-wide gBGC coefficient is $b = 1.3 \times 10^{-4}$. Thus, in an obligate outcrossing species, such a gBGC drive would have a very strong impact, even for relatively small effective population sizes ($N_e \geq 10^5$). Yeast show a very low level of sexual reproduction and outcrossing, which reduces the population genetic effect of gBGC (Tsai et al. 2010). Nonetheless, there is evidence that gBGC affects the long-term evolution of yeast genomes (Birdsell 2002; Lynch et al. 2010; Tsai et al. 2010; Cutter and Moses 2011; Harrison and Charlesworth 2011).

### Invalidation of the BER Hypothesis

To better understand the proximal causes of gBGC and the selective pressure that might operate on this genetic system, it is essential to identify the molecular mechanisms responsible for this conversion bias. In mammals, experiments in somatic cells demonstrated that the repair of DNA mismatches is strongly GC biased (Brown and Jiricny 1988, 1989; Bill et al. 1998). This GC bias results, at least in part, from the activity of the BER pathway, which involves DNA glycosylases that specifically excise thymines (and/or uracils) in DNA mismatches. Given that BER is intrinsically GC biased, it has been previously proposed that this repair mechanism, if active during meiosis, could be the cause of gBGC (Brown and Jiricny 1989; Galtier et al. 2001; Birdsell 2002; Marais 2003). BER leads to short patch repair and should therefore be frequently associated to complex conversion tracts. In the Mancera data set, the majority (≥89%) of conversion tracts are simple, as expected given the prominent role of MMR during recombination. However, a minor contribution of BER to the repair of mismatches in heteroduplex DNA cannot be a priori excluded. Calculations show that if a fraction of SNP conversion events result from the action of BER, then such cases must be at least 10 times more frequent among complex conversion tracts compared with simple conversion tracts (for details, see supplementary text S1, Supplementary Material online). Hence, if BER is the unique cause of gBGC, it is expected that the conversion bias in favor of GC alleles should be much stronger among complex conversion tracts than among simple tracts. However, in contradiction with this prediction, our analyses show that the largest source of gBGC corresponds to recombination events associated with simple conversion tracts. We, therefore, conclude that in *S. cerevisiae*, gBGC occurs in conversion events associated with a long-patch repair machinery and that the contribution of BER to the gBGC process, if any, is at most very minor.

### Invalidation of the Initiation Bias Hypothesis

An alternative hypothesis is that gBGC could be the consequence of a bias in the initiation of recombination. It has been shown that the rate of DSB formation at a given locus may vary strongly between different haplotypes (Webb et al. 2008), and there is clear evidence that this initiation bias leads to a strong conversion bias in favor of the haplotype that is less prone to initiate recombination (Myers et al. 2010). Thus, if DSBs tend to occur more frequently on the AT-richest haplotype, this initiation bias might lead to gBGC. The analysis of DSB maps in *S. cerevisiae* did not reveal any clear association with AT-rich motifs (Murakami and Nicolas 2009; Pan et al. 2011), but a weak sequence preference, sufficient to cause the observed gBGC, cannot be a priori excluded. However, this initiation bias hypothesis is not consistent with our observation that gBGC is exclusively associated with CO recombination events. In yeast, CO hotspots and NCO hotspots generally coincide: Some recombination hotspots with biased CO/NCO ratios have been detected, but they represent only a tiny fraction (1.4%) of the regions involved in recombination events (Mancera et al. 2008). This indicates that generally, the same initiating regions can lead to both COs and NCOs. Hence, if the distribution of DSBs was the cause of gBGC, one would expect the same conversion bias in CO and NCO recombination events. The fact that gBGC is CO specific is therefore a strong argument indicating that the conversion bias is the consequence of a process that is posterior to the formation of DSBs. Note that in humans, the location of recombination hotspots is determined by a DNA-binding protein (PRDM9), which recognizes a specific sequence motif (Baudat et al. 2010). As predicted by the initiation bias model, the 13-bp genomic sequence motif targeted by PRDM9 has been subject to a rapid accumulation of substitutions in the human lineage (Myers et al. 2010). However, given that this motif is GC rich, this initiation bias tends to favor the fixation of G:C to A:T mutations. Hence, this

initiation bias cannot account for the gBGC process observed in the human genome.

### MMR Model 1: Strand Rejection

Given that the BER and initiation bias models are rejected, an alternative hypothesis is that gBGC could be due to MMR. MMR plays a major role during recombination as a sensor of sequence homology during the process of strand invasion (Hunter et al. 1996; Chen and Jinks-Robertson 1999; Surtees et al. 2004). As mentioned in the introduction, it is in principle possible that the decision to reject the invading strand depends on the nature of mismatches present in the heteroduplex DNA. It is also possible that even in cases where the invading strand is not rejected, the extent of the heteroduplex is influenced by the presence of SNPs: When an SNP is encountered during the process of strand invasion, then either it is included in the heteroduplex (resulting in an additional mismatch) or the process of strand invasion is interrupted. Let us suppose that when the SNP that is encountered corresponds to an AT allele on the single-stranded 3′-overhang (and a GC allele on the intact homolog), the probability of interruption is lower than in the opposite configuration. Under this assumption, one expects an excess of cases where the mismatches at the extremities of heteroduplex DNA correspond to an AT on the broken chromatid and to a GC allele on the intact homolog. Thus, given that gene conversion occurs from the intact homolog toward the broken chromatid, this model predicts an excess of GC-flanked conversion tracts. Moreover, this model also predicts that GC-flanked conversion tracts should, on average, be longer than other conversion tracts. Both predictions, therefore, fit with our observations. However, one difficulty with this model is to explain why gBGC is CO specific. Yeast mutants lacking MSH2 show an increase both in the number of COs and NCOs (Martini et al. 2011), which suggests that MMR affects strand invasion for both categories of recombination events. Thus, if gBGC was due to the sensing of mismatches by MMR during the process of strand invasion, one would a priori expect to detect gBGC both in COs and NCOs.

### MMR Model 2: Biased Mismatch Repair

An alternative (but non exclusive) hypothesis is that gBGC could result from the repair activity of MMR. MMR is composed of two main protein classes (Evans and Alani 2000; Jiricny 2006): MSH (MutS Homologs) proteins act as heterodimers to recognize mismatches along the sequence and recruit MLH (MutL Homologs) heterodimer proteins to form complexes. These complexes then migrate, in both directions from the mismatch, up to encountering a nick, where they will recruit an exonuclease. The degradation of the nick-containing strand is then followed by DNA resynthesis. It has been shown, both in vivo and in vitro, that the efficiency of mismatch repair by MMR depends on the nature of mismatches (Bishop et al. 1989; Mazurek et al. 2009; Martini et al. 2011). However, to our knowledge, it has not been investigated whether the "direction" of repair by MMR is affected by the nature of mismatches. In principle, it is only when nicks are present on both strands that there is a possibility of choice in the direction of repair. In the context of heteroduplex DNA formed during recombination, nicks are always present on the strand coming from the broken chromosome, but nicks can also be formed on the other strand during the resolution of recombination intermediates (Martini et al. 2011). The choice of one strand or the other will lead either to the conversion of the broken haplotype toward the nonbroken haplotype or to the restoration of Mendelian segregation. We propose a model, wherein the direction of the repair by MMR (toward conversion or restoration) depends on the nature of the mismatched bases that are close to the nicks flanking the heteroduplex (fig. 4). According to that model, when nicks are present on both strands, MMR would preferentially initiate DNA degradation from the nick closest to a mismatched A or T base. Thus, when the strand coming from the broken chromosome carries the GC allele, MMR would more frequently lead to the restoration of this haplotype, when compared with when it carries the AT allele. Hence, AT alleles would be converted more frequently than GC alleles, which would lead to an overall conversion bias in favor of GC alleles (fig. 4; supplementary text S3, Supplementary Material online). Note that the extent of the detected conversion tract (and hence the nature of the SNPs at its boundaries) depends on whether mismatch repair is directed toward conversion or restoration. As shown in figure 4, if AT alleles are less frequently restored than GC alleles, then GC-flanked conversion tracts are expected to be on average larger than other conversion tracts (supplementary text S3, Supplementary Material online, for details). Thus, this model would explain both the fact that gBGC is directed by the nature of the alleles located at the extremities of the conversion tract and the fact that GC-flanked conversion tracts are larger than other conversion tracts.

Again, one difficulty with this hypothesis is to explain why gBGC is CO specific. Current models indicate that COs result from dHj resolution (class I COs) and from the Mus81 pathway (class II COs) (Martini et al. 2011) (fig. 1). The class I CO pathway requires several meiosis-specific homologs of the MMR system (Hunter and Borts 1997; Argueso et al. 2004): The MLH1–MLH3 complex is involved in dHj resolution, whereas the MSH4–MSH5 complex is required in earlier steps of this pathway (Zakharyevich et al. 2012). However, MSH4 and MSH5 lack mismatch recognition domain and activity (Ross-Macdonald and Roeder 1994; Hollingsworth et al. 1995) and hence cannot be directly responsible for the biased mismatch repair. In fact, both in meiotic and mitotic cells, the recognition of base-base mismatches relies on the MSH2–MSH6 complex (for review, see Jiricny [2006]). In MSH2 mutants, meiotic recombination proceeds normally, but mismatches in heteroduplex DNA are left unrepaired, both for COs and NCOs (Martini et al. 2011). Given that both COs and NCOs rely on the same machinery for mismatch recognition, then how can gBGC be CO specific? One possible explanation is that the resolution of COs requires the formation of nicks of both DNA strands, in close vicinity (the average distance between Hj is approximately 260 bp [Cromie

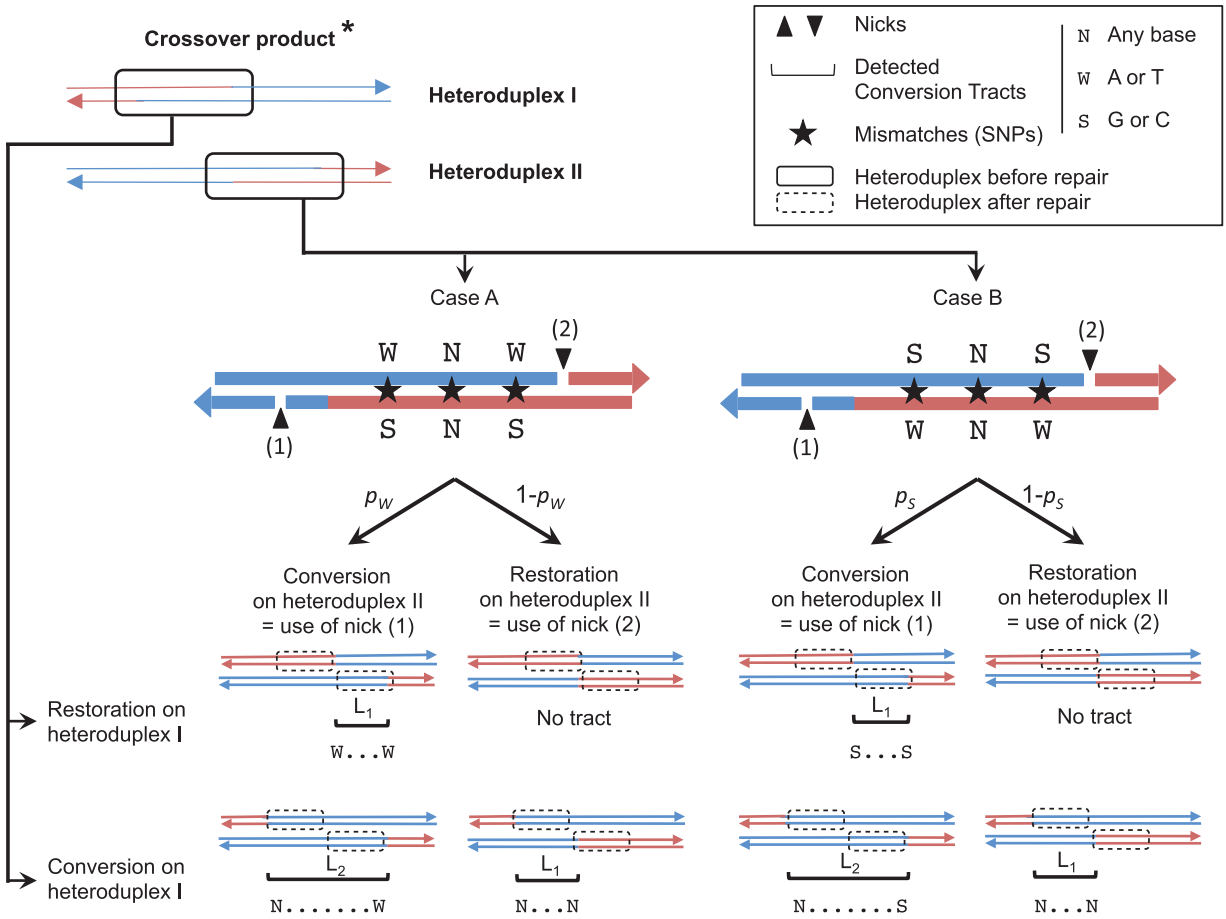



**Fig. 4.** Model of gBGC driven by GC-biased MMR repair. According to our model, gBGC results form a bias in the repair of mismatches by MMR, when nicks are present on both strands of the heteroduplex. This configuration potentially occurs during CO pathways (indicated by an * in fig. 1). COs involve the formation of two heteroduplexes. In the example shown, heteroduplex II consists of one GC-flanked haplotype (S = G or C) and one AT-flanked haplotype (W = A or T; *N* represents any mismatched base within the heteroduplex). MMR repair from nick (1) leads to the conversion of the red haplotype (the one that encountered the DSB initiating recombination), whereas the use of nick (2) leads to restoration. According to our model, the probability to use nick (2), instead of nick (1), is higher when the red strand carries the GC-flanked haplotype (case A) compared with when it carries the AT-flanked haplotype (case B). Thus, the probability of conversion is higher in case A than in case B (i.e., $P_S > P_W$). The detected conversion tract depends on the repair of both heteroduplexes: If both are restored, no tract can be detected. If only one heteroduplex is converted, the size of the tract ($L_1$) is expected to be smaller than if both are converted ($L_2$, with $L_1 < L_2$). In the case where the heteroduplex I is converted, the nature of the donor allele detected at the 5′-end of the tract (represented by an *N*) is independent of the haplotypes present in heteroduplex II. Given that $P_S > P_W$, this model predicts that among detected tracts with $GC_f/AT_f$ polymorphism, there should be a transmission bias in favor of GC-flanked haplotype. Moreover, the model predicts that GC-flanked conversion tracts should on average be longer than AT-flanked ones (see details in supplementary text S3, Supplementary Material online). For simplicity, failures of mismatch repair (leading to postmeiotic segregation) are not considered here.

et al. 2006]). Thus, the presence of nicks on both DNA strands provides an opportunity for a bias in the direction of repair by MMR according to the nature of mismatches. Molecular pathways leading to NCOs also involve nicks on both strands (fig. 1). However, if these nicks are not in close proximity, or not present at the same time in NCO intermediates, then there would be no possible choice in the direction of repair. Thus, the fact that gBGC is CO specific could be due to differences in the spatiotemporal configuration of nicks in CO and NCO recombination intermediates.

## Conclusion

In conclusion, our observations reject the BER and initiation bias models and are consistent with the hypotheses that gBGC is caused by MMR (via its role in strand invasion, in mismatch repair, or both). At this stage, the models of MMR-induced gBGC presented here remain speculative, and more data will be needed to test them. The hypothesis that gBGC is due to the repair activity of MMR makes several predictions that could be tested experimentally. First, this model predicts that the repair of AT:GC mismatches by MMR should be biased toward GC when nicks are present on both DNA strands. Second, this model implies that MSH2 should be active, not only during the early steps of recombination but also at the final step of CO pathway(s), during the resolution of joint molecules. Furthermore, it would be interesting to test whether gBGC is associated with both class I and class II CO pathways. In their analyses of recombination in yeast, Mancera et al. (2008) included five meioses from a mutant

of the class I CO pathway. Unfortunately, the limited number of recombination events detected was not sufficient to test whether gBGC occurs or not in this mutant (data not shown).

Given that the components of the recombination machinery are conserved across eukaryotes (Kolodner and Marsischky 1999), it seems likely that the same processes may be responsible for gBGC in other eukaryotes. One should note, however, that the relative contribution of the different CO recombination pathways differs among taxa. For example, fission yeast appears to rely exclusively on the class II pathway (Cromie et al. 2006), whereas most COs in mice result from the class I pathway (Holloway et al. 2008). If gBGC is specific of one of the two CO pathways, then one may expect differences in gBGC intensity among taxa.

One important issue is to understand the primary cause of the evolution of gBGC. In all taxa where some evidence of BGC has been reported, the conversion bias tends to favor GC alleles over AT alleles (Capra and Pollard 2011; Escobar et al. 2011; Pessia et al. 2012). This probably results from the fact that in most taxa, the pattern of mutation is biased toward AT (Lynch 2010), and hence any selective pressure to reduce the mutation rate is expected to favor the evolution of GC-biased mismatch repair. It should be noted that meiosis represents only a small fraction of the life cycle of eukaryotes. For example, in humans, germline cells are on average subject to 33 (in females) to approximately 200 (in males) mitotic cell divisions before meiosis (Chang et al. 1994). In nature, budding yeasts divide by meiosis only once every 1,000 generations (Tsai et al. 2010). Hence, most mutations occur in mitotic cells, where MMR plays a major role in the repair of DNA replication errors (Jiricny 2006). Thus, if the GC bias of MMR results from a selective pressure to reduce the mutation rate, then the strongest selective pressure should come from mutations that occur during mitosis. We therefore propose that the evolution of GC-biased MMR is driven by a selective pressure to reduce the rate of mutation in mitotic cells (including somatic cells, in the case of multicellular eukaryotes) and that gBGC simply results from the activity of this repair system during meiosis. Thus, under this hypothesis, gBGC would be a nonadaptive (and possibly maladaptive) indirect consequence of a selective pressure to limit the mutation rate in mitotic cells.

## Materials and Methods

### Data

We used recombination data, obtained by genotyping meiosis products of wild-type strains of *S. cerevisiae*, that were produced by Mancera et al. (2008). The list of conversion events associated with COs and NCOs, with details about parental and transmitted alleles, was kindly provided by Richard Bourgon. We filtered SNPs for which the base found in the spore was not called with enough confidence (labeled "NA" in the data). This led to a final list of 89,538 genotyped SNPs involved in conversion events, corresponding to 2,884 COs and 2,090 NCOs.

### Measure of Gene Conversion Biases

We measured gene conversion biases at different scales: individual SNPs or haplotypes (see main text). In all cases, we considered sets of sites that are heterozygous in the parental hybrid and that were involved in conversion events (for the sake of generality, the two alleles are hereafter denoted Z and Y). For this set of sites, we counted the proportion of the allele Z in the offspring ($x$). We tested the existence of a conversion bias in favor of the allele Z by comparing $x$ to the Mendelian expectation (50%), with a one-sample proportion test (see later). The intensity of the conversion bias in favor of the allele Z was measured by the coefficient $b = 2x - 1$.

### Statistical Testing

Two types of tests were used on the proportion $x$ as defined earlier. We used normal approximate two-tailed Z test with continuity correction to compare $x$ to the Mendelian expectation of 50%. This is referred as "one-sample proportion test" in the text and legends. Additionally, we used normal approximate Z test with continuity correction to compare two different observed $x$ proportions. This is referred as "two-sample proportion test" in the text and legends. Two-sample proportion tests are all two-tailed except when specified differently.

## Supplementary Material

Supplementary texts S1–S3, figure S1 and S2, and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

The authors thank Richard Bourgon for providing the data set and for fruitful discussions about the observed transmission biases. They thank Bernard de Massy, Erika Kvikstad, Nicolas Lartillot, and Gaël Yvert for their very useful comments on an early version of the manuscript. This work was supported by the Centre National de la Recherche Scientifique.

## References

Argueso JL, Wanat J, Gemici Z, Alani E. 2004. Competing crossover pathways act during meiosis in *Saccharomyces cerevisiae*. *Genetics* 168:1805–1816.

Auton A, Fledel-Alon A, Pfeifer S, et al. (23 co-authors). 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* 336: 193–198.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327:836–840.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol.* 7:45–62.

Bill CA, Duran WA, Miselis NR, Nickoloff JA. 1998. Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells: competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. *Genetics* 149: 1935–1943.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol.* 19:1181–1197.

Bishop DK, Andersen J, Kolodner RD. 1989. Specificity of mismatch repair following transformation of *Saccharomyces cerevisiae* with heteroduplex plasmid DNA. *Proc Natl Acad Sci U S A.* 86:3713–3717.

Brown TC, Jiricny J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* 54: 705–711.
Brown TC, Jiricny J. 1989. Repair of base-base mismatches in simian and human cells. *Genome* 31:578–583.
Capra JA, Pollard KS. 2011. Substitution patterns are GC-biased in divergent sequences across the metazoans. *Genome Biol Evol.* 3: 516–527.
Chang B, Shimmin L, Shyue S-K, Hewett-Emmett D, Li W-H. 1994. Weak male driven molecular evolution in rodents. *Proc Natl Acad Sci U S A.* 91:827–831.
Chen SY, Tsubouchi T, Rockmill B, Sandler JS, Richards DR, Vader G, Hochwagen A, Roeder GS, Fung JC. 2008. Global analysis of the meiotic crossover landscape. *Dev Cell.* 15:401–415.
Chen W, Jinks-Robertson S. 1999. The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. *Genetics* 151:1299–1313.
Coïc E, Gluck L, Fabre F. 2000. Evidence for short-patch mismatch repair in *Saccharomyces cerevisiae*. *EMBO J.* 19:3408–3417.
Coop G, Przeworski M. 2007. An evolutionary view of human recombination. *Nat Rev Genet.* 8:23–34.
Cromie GA, Hyppa RW, Taylor AF, Zakharyevich K, Hunter N, Smith GR. 2006. Single Holliday junctions are intermediates of meiotic recombination. *Cell* 127:1167–1178.
Cutter AD, Moses AM. 2011. Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. *Mol Biol Evol.* 28:1745–1754.
de Massy B. 2003. Distribution of meiotic recombination sites. *Trends Genet.* 19:514–522.
Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4:1–19.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10:285–311.
Escobar JS, Glémin S, Galtier N. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Mol Biol Evol.* 28:2561–2575.
Evans E, Alani E. 2000. Roles for mismatch repair factors in regulating genetic recombination. *Mol Cell Biol.* 20:7839–7844.
Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23:273–277.
Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25:1–5.
Galtier N, Piganeau G, Mouchiroud D. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159:907–911.
Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. *Genetics* 185: 939–959.
Glémin S. 2011. Surprising fitness consequences of GC-biased gene conversion. II. Heterosis. *Genetics* 187:217–227.
Harrison RJ, Charlesworth B. 2011. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces* sensu stricto group of yeasts. *Mol Biol Evol.* 28:117–129.
Hoffmann ER, Borts RH. 2004. Meiotic recombination intermediates and mismatch repair proteins. *Cytogenet Genome Res.* 107: 232–248.
Hollingsworth N, Ponte L, Halsey C. 1995. MSH5, a novel MutS homolog, facilitates meiotic reciprocal recombination between homologs in *Saccharomyces cerevisiae* but not mismatch repair. *Genes Dev.* 9: 1728–1739.
Holloway JK, Booth J, Edelmann W, McGowan CH, Cohen PE. 2008. MUS81 generates a subset of MLH1-MLH3-independent crossovers in mammalian meiosis. *PLoS Genet.* 4:e1000186.
Holmes J, Clark S. 1990. Strand-specific mismatch correction in nuclear extracts of human and *Drosophila melanogaster* cell lines mismatch. *Proc Natl Acad Sci U S A.* 87:5837–5841.
Hunter N, Borts RH. 1997. MLH1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. *Genes Dev.* 11:1573–1582.
Hunter N, Chambers SR, Louis EJ, Borts RH. 1996. The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. *EMBO J.* 15:1726–1733.
Jiricny J. 2006. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol.* 7:335–346.
Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol.* 3:614–626.
Kolodner RD, Marsischky GT. 1999. Eukaryotic DNA mismatch repair. *Curr Opin Genet Dev.* 9:89–96.
Krogh BO, Symington LS. 2004. Recombination proteins in yeast. *Annu Rev Genet.* 38:233–271.
Lynch DB, Logue ME, Butler G, Wolfe KH. 2010. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biol Evol.* 2: 572–583.
Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A.* 107:961–968.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19:330–338.
Martini E, Borde V, Legendre M, Audic S, Regnault B, Soubigou G, Dujon B, Llorente B. 2011. Genome-wide analysis of heteroduplex DNA in mismatch repair-deficient yeast cells reveals novel properties of meiotic recombination pathways. *PLoS Genet.* 7:1–18.
Mazurek A, Johnson CN, Germann MW, Fishel R. 2009. Sequence context effect for hMSH2-hMSH6 mismatch-dependent activation. *Proc Natl Acad Sci U S A.* 106:1–6.
McMahill MS, Sham CW, Bishop DK. 2007. Synthesis-dependent strand annealing in meiosis. *PLoS Biol.* 5:2589–2601.
Memisoglu A, Samson L. 2000. Base excision repair in yeast and mammals. *Mutat Res.* 451:39–51.
Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol.* 21:984–990.
Murakami H, Nicolas A. 2009. Locally, meiotic double-strand breaks targeted by Gal4BD-Spo11 occur at discrete sites with a sequence preference. *Mol Cell Biol.* 29:3500–3516.
Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* 327:876–879.
Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A.* 80:6278–6281.
Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat.* 32: 198–206.
Nicolas A, Treco D, Schultes NP, Szostak JW. 1989. An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. *Nature* 338:35–39.
Pan J, Sasaki M, Kniewel R, et al. (12 co-authors). 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144:719–731.
Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol.* 4:675–682.
Qi J, Wijeratne A, Tomsho L, Hu Y, Schuster S, Ma H. 2009. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in *Saccharomyces cerevisiae*. *BMC Genomics* 10:475.
Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci.* 365:2571–2580.

Ross-Macdonald P, Roeder GS. 1994. Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. *Cell* 79:1069–1080.

Smith KN, Nicolas A. 1998. Recombination at work for meiosis. *Curr Opin Genet Dev.* 8:200–211.

Surtees JA, Argueso JL, Alani E. 2004. Mismatch repair proteins: key regulators of genetic recombination. *Cytogenet Genome Res.* 107: 146–159.

Thomas DC, Roberts JD, Kunkel TA. 1991. Heteroduplex repair in extracts of human HeLa cells. *J Biol Chem.* 266:3744–3751.

Tsai IJ, Burt A, Koufopanou V. 2010. Conservation of recombination hotspots in yeast. *Proc Natl Acad Sci U S A.* 107:7847–7852.

Webb AJ, Berg IL, Jeffreys A. 2008. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc Natl Acad Sci U S A.* 105:10471–10476.

Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet.* 28:101–109.

Winzeler EA, Richards DR, Conway AR, et al. (11 co-authors). 1998. Direct allelic variation scanning of the yeast genome. *Science* 281: 1194–1197.

Zakharyevich K, Tang S, Ma Y, Hunter N. 2012. Delineation of joint molecule resolution pathways in meiosis identifies a crossover-specific resolvase. *Cell* 149:334–347.

# II.C. Informations supplémentaires de l'article

Informations supplémentaires de l'article : *GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers : Molecular Mechanisms and Evolutionary Significance.*

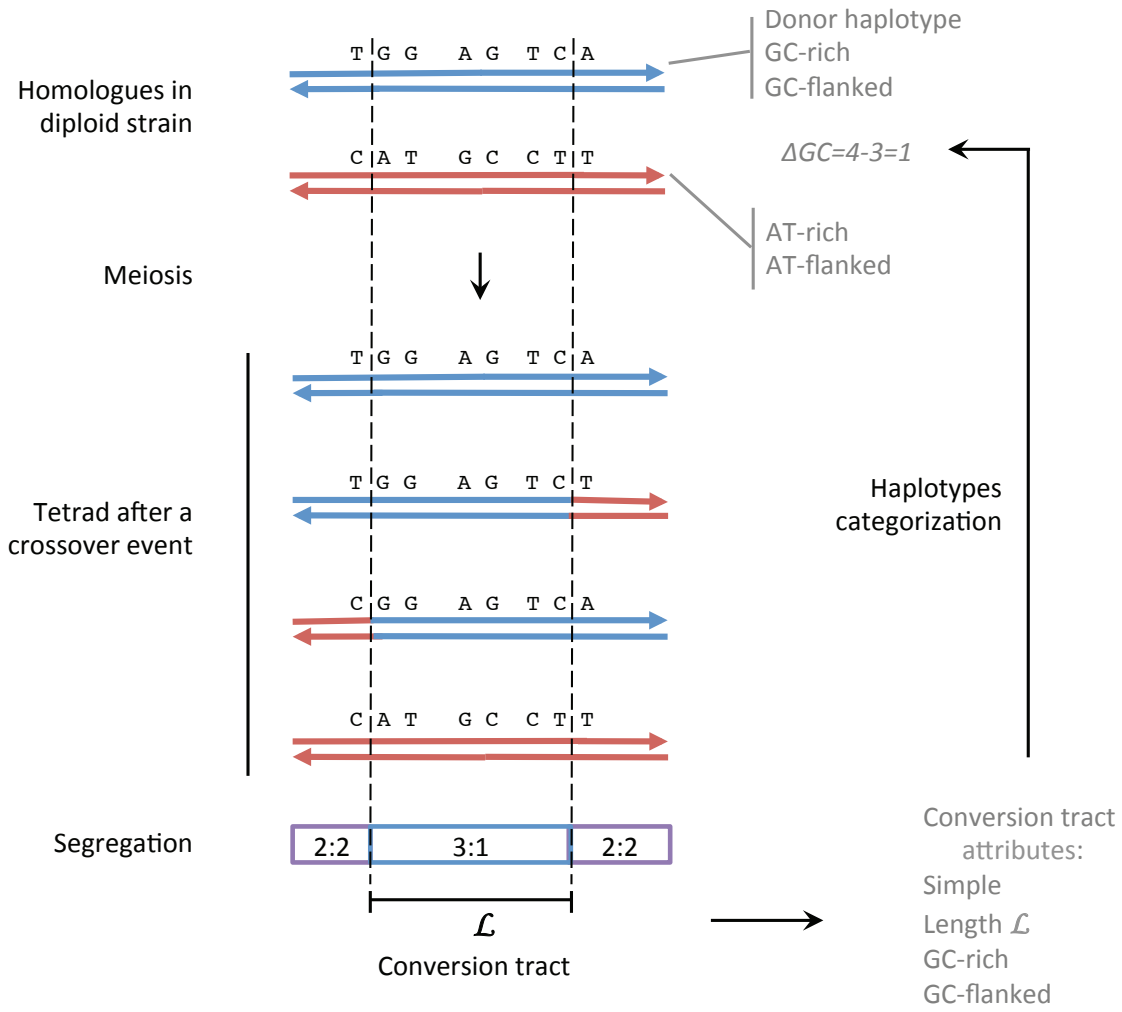



**Figure S1. Conversion tract definition and haplotypes categorization**

Double stranded DNA is shown by red and blue arrows. Only varying positions between homologues (SNPs) are shown. In this example the product of a crossover is shown. The conversion tract is defined as the interval (between dashed lines) in which we observe non-Mendelian segregation (3:1) as opposed to Mendelian segregation (2:2). The length ($\mathcal{L}$) of this tract is the physical distance between the two SNPs at the boundaries of the conversion tract interval. Those two SNPs are used to assign parental haplotypes to GC/AT-flanked categories. The ΔGC measure corresponds to the difference in the number of G+C nucleotides between the two parental haplotypes within the conversion tract. The haplotype categories corresponding to this specific example are indicated in grey.

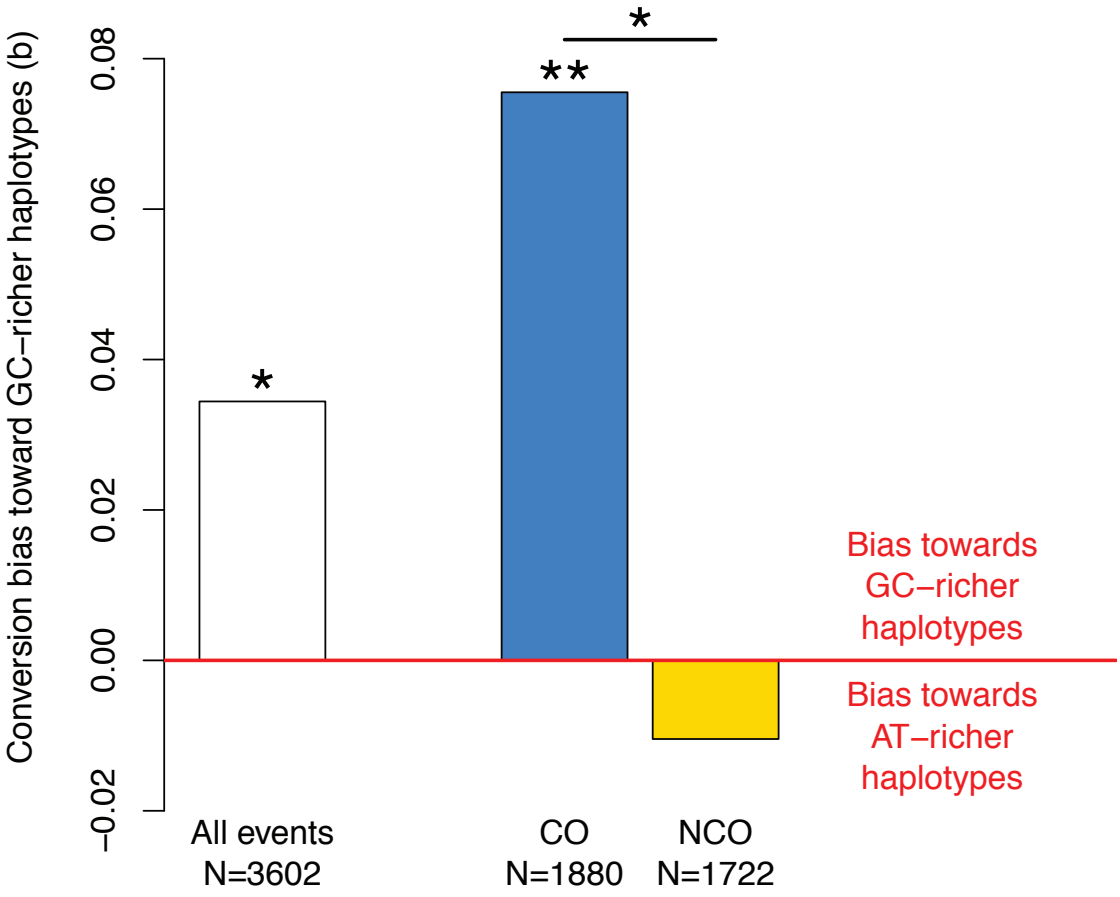



**Figure S2. Conversion bias towards GC-richer haplotypes (defined with a more stringent**

AT/GC-richer polymorphisms are defined here by $|\Delta GC| > 1/3$ (Text S2). The conversion bias towards GC-richer haplotypes ($b$) was computed for simple conversion tracts, taken all together (white bar) or separating tracts associated with COs (blue bar) and NCOs (yellow bar). "N" is the number of genotyped haplotypes in each category. The red horizontal line indicates Mendelian expectation ($b$=0). Significant conversion biases are indicated by "*" for a p-value≤0.05 and "**" for a p-value≤0.01, one-sample proportion test (see methods). The "*" between "CO" and "NCO" bars denotes the fact that the conversion bias towards GC-richer haplotypes is significantly different between CO and NCO events (two-sample proportion test, see methods).

**Table S1. Tract lengths and number of SNPs for each conversion event subset studied.**

| Figure/Table reference | Subset | Median tract length (bp) | Mean (Median) number of SNPs in the tract |
|---|---|---|---|
| Table 1 | All | 906 | 9.5 (6) |
| | Simple | 752.5 | 14.2 (5) |
| | Complex | 1,948 | 8.8 (9.5) |
| Table S2 | All | 1,813 | 15.1 (12) |
| | Simple | 1,740 | 16.8 (12) |
| | Complex | 2,249 | 14.8 (12) |
| Figure 2 | All | 869.5 | 9.5 (6) |
| | CO | 926 | 9.7 (6) |
| | NCO | 790.5 | 9.3 (5) |
| Figure S1 | All | 145 | 4.2 (2) |
| | CO | 134 | 4.3 (2) |
| | NCO | 158 | 4.1 (2) |
| Figure 3 | CO | 1,017 | 9.8 (6) |
| | NCO | 726.5 | 8.9 (5) |
| | CO; ΔGC>0 | 204 | 6.5 (2) |
| | CO; ΔGC≤0 | 1,878 | 15.0 (12) |
| Table 2 | All | 1,471 | 12.9 (10) |
| | Flanking SNPs only | 1,497 | 12.8 (10) |
| | Flanking SNPs excluded | 1,564 | 13.2 (10) |
| Table S3 | All | 1,825 | 15.3 (12) |
| | Flanking SNPs only | 1,834 | 15.0 (12) |
| | Flanking SNPs excluded | 1,864 | 15.2 (12) |

This table indicates the median length of conversion tracts (see below) and the mean (and median) number of SNPs in the conversion tract for each subset analysed in the study. The tract length is defined as the physical distance in base pairs (bp) between the two most distal SNPs within the tract.

**Table S2. Conversion bias towards GC-bases for AT/GC SNPs involved in a recombination event (conversion tracts with at least 5 SNPs).**

| Conversion tract type | Number of genotyped SNP sites with AT/GC polymorphism[a] | Conversion bias towards GC-bases (*b*) | P-value[b] |
|---|---|---|---|
| All | 70,374 | 0.013 | < 0.001 |
| Simple | 57,897 | 0.013 | 0.0014 |
| Complex | 12,477 | 0.011 | *N.S.* |

[a]The conversion bias towards GC alleles (*b*) was computed for SNP sites located in conversion tracts containing at least 5 SNPs, taken all together or separating sites found in simple and complex conversion tracts.

[b]one-sample proportion test, see methods, N.S.: non-significant.

**Table S3. Conversion bias towards GC-richer haplotypes among $AT_f/GC_f$-polymorphism, considering all SNPs, or only flanking or internal SNPs (conversion tracts with at least 5 SNPs).**

| SNPs considered to classify haplotypes as AT- or GC-richer | Number of genotyped haplotypes with AT/GC-richer polymorphism[a] | Conversion bias towards GC-richer haplotypes (*b*) | P-value[b] |
|---|---|---|---|
| All | 890 | 0.052 | 0.13 |
| Flanking SNPs only | 1,006 | 0.101 | 0.001 |
| Flanking SNPs excluded | 888 | 0.018 | 0.61 |

[a]The conversion bias towards GC-richer haplotypes (*b*) was computed for $AT_f/GC_f$ polymorphisms located in CO-associated simple conversion tracts overlapping at least 5 SNPs. Haplotypes were categorized in AT- or GC-richer according to their difference in GC-content, considering all SNPs in the tract or only the two flanking SNPs or only the SNPs that are not the two flanking SNPs.

[b]one-sample proportion test, see methods.

**Table S4. Relative abundance of conversion tracts types, as predicted by the model of GC-biased MMR.**

| Case # | Parental haplotypes | Conversion tract | Relative abundance | Tract length |
|---|---|---|---|---|
| 1 | N…W<br>N…S | N…W | $p_w\, x$ | $L_2$ |
| 2 | W…W<br>S…S | W…W | $p_w\, (1\text{-}x)$ | $L_1$ |
| 3 | N…W<br>N…S | N…S | $p_S\, x$ | $L_2$ |
| 4 | S…S<br>W…W | S…S | $p_S\, (1\text{-}x)$ | $L_1$ |
| 5 | N…N<br>N…N | N…N | $(2\text{-}\, p_S\, \text{-}\, p_w)\, x$ | $L_1$ |

For each of the five different types of conversion tracts produced after the repair of mismatches in heteroduplexes in crossover products (see Figure 4), we show the identity of the flanking bases observed in parental haplotypes (second column; S = G or C, W = A or T, N = Any base) and in the conversion tract (third column). The relative abundance in the gamete pool (fourth column) and the tract length (fifth column) are shown using parameters defined in Supplementary Text S3.

**Table S5. Relative abundance of conversion tracts types among $AT_f/GC_f$-polymorphism, as predicted by the model of GC-biased MMR.**

| Case # | Parental haplotypes | Conversion tract | Relative tract abundance | Tract length |
|---|---|---|---|---|
| 1 | W…W<br>S…S | W…W | $r\, p_w\, x$ | $L_2$ |
| 2 | W…W<br>S…S | W…W | $p_w\,(1\text{-}x)$ | $L_1$ |
| 3 | W…W<br>S…S | S…S | $r\, p_S\, x$ | $L_2$ |
| 4 | S…S<br>W…W | S…S | $p_S\,(1\text{-}x)$ | $L_1$ |
| 5 | W…W<br>S…S | W…W | $r^2\,(2\text{-}\,p_S\,\text{-}\,p_w)\,x$ | $L_1$ |
| 5 | S…S<br>W…W | S…S | $r^2\,(2\text{-}\,p_S\,\text{-}\,p_w)\,x$ | $L_1$ |

For each of the five different types of conversion tracts produced after the repair of mismatches in heteroduplexes in crossover products among the subset of $AT_f/GC_f$-polymorphism, we show the identity of the flanking bases observed in parental haplotypes (second column; S = G or C, W = A or T, N = Any base) and in the conversion tract (third column). The relative abundance in the gamete pool (fourth column) and the tract length (fifth column) are shown using parameters described in Supplementary Text S3.

**Text S1. Expected relative contribution of BER to simple and complex conversion tracts.**

To distinguish between the BER and MMR models, we compared the strength of gBGC in simple and complex conversions tracts (a complex conversion tracts is a tract involving conversion events from both parental haplotypes, whereas simple tracts involve only one donor haplotype). Indeed, one clear difference between BER and MMR is the length of the region affected by the repair: whereas MMR involves DNA re-synthesis over hundreds of base pairs (i.e. about the size of conversion tracts) (Winzeler et al. 1998; Holmes and Clark 1990), BER leads only to short-patch repair (1-13bp) (Evans and Alani 2000). Given that the median distance between contiguous markers is 78 bp, if some SNP conversion events are driven by BER, then the conversion of these SNPs should occur independently of the conversion of neighboring SNPs. Thus, if BER is active during recombination, it is expected to lead frequently to complex conversion tracts.

To justify this assertion more formally, let us consider a simple model, in which conversion tracts can be affected either by MMR only (probability $p$), or by both MMR and BER (probability 1-$p$), with BER acting independently on $n$ SNPs in the tact ($n$>0). When both MMR and BER are acting on the tract, then for each SNP repaired by BER, the probability to be repaired using the same template strand as MMR is 0.5. Thus, when BER is active, the expected proportions of simple and complex conversion tracts are respectively ($0.5^n$) and ($1 - 0.5^n$). BER is probably not the unique cause of complex conversion tracts. Let us assume that in absence of BER, MMR can lead to complex tracts with probability $q$. We can now express the proportion of complex ($f_{cpx}$) and simple ($f_{smp}$) tracts:

$$\begin{cases} f_{cpx} = p\,(1-0.5^n) + (1-p)\,q \\ f_{smp} = p\,0.5^n + (1-p)(1-q) \end{cases} \qquad (1)$$

Thus, the proportion of conversion tracts in which BER was active, among complex tracts ($BER_{cpx}$) and among simple tracts ($BER_{smp}$), can be expressed as follows:

$$\begin{cases} BER_{cpx} = p\,(1-0.5^n) / f_{cpx} \\ BER_{smp} = p\,0.5^n / f_{smp} \end{cases} \qquad (2)$$

In the Mancera dataset [23], the proportion of complex events is $f_{cpx}$ = 8.8%. Thus, the ratio $f_{smp}/f_{cpx}$ is about 10. Hence, we can deduce that:

$$\frac{BER_{cpx}}{BER_{smp}} \approx 10\,(2^n - 1) \qquad (3)$$

Given that n≥1, the value of this ratio is at least 10. Thus, under the assumption that a fraction of SNP conversions results from the activity of BER, then the proportion of tracts in which BER was involved should be at least 10 times higher among complex tracts than among simple tracts. Hence, if gBGC was due to BER, then the signal of gBGC should be much stronger among complex conversion tracts as compared to simple ones.

**Text S2. Conversion bias towards GC-richer haplotypes (defined with a more stringent threshold).**

In the Figure 2, we considered all conversion events for which there was a difference in GC-content between the two haplotypes, even if this difference was weak. In order to test if this categorization does not introduce any bias in our study, we repeated the same analysis using only conversion events for which there is a strong difference in GC-content between the two haplotypes. To measure this difference, we introduced the $\Delta GC_h$ parameter, which is defined as follow:

$$\Delta GC_h = \frac{|GC_Y - GC_S|}{n} \quad (4)$$

Where $n$ is the number of AT/GC SNP sites located in the conversion tract, and $GC_Y$ and $GC_S$, are the numbers of GC-alleles at those sites respectively for the YJM789 and S96 strains. In this section, we only used events for which $\Delta GC_h > 1/3$ (N=1,801 recombination events). Using this threshold, the intensity of the bias ($b$) remains the same as in previous analyses (compare Figure S2 with Figure 2) for all three categories (All, CO and NCO). This shows that our results are not affected by the threshold used to define GC-richer/AT-richer haplotypes.

**Text S3. Model of gBGC driven by GC-biased MMR-repair: expected relative abundance and lengths of conversion tracts with $GC_f$ or $AT_f$ haplotype.**

Recombination is initiated by the formation of a DSB on one chromatid, followed by 5' to 3' resection, which produces two single-stranded 3' overhangs. The DSB is subsequently repaired using an intact template (sister chromatid or homolog) (Figure 1). This DSB repair involves the formation of heteroduplex DNA, with one strand coming from the template chromatid and the other corresponding to one single-stranded 3' overhang of the broken chromatid. If the template chromatid is the homolog, then the two haplotypes in the heteroduplex may be different. The two corresponding haplotypes will hereafter be referred to as the "template haplotype" and the "broken haplotype". Mismatches in the heteroduplex can be repaired and, depending to the direction of repair, this process leads either to conversion of the broken haplotype by the template haplotype, or to restoration of Mendelian segregation.

Let us consider a heteroduplex with $AT_f/GC_f$ polymorphism, i.e. in one haplotype the alleles located at the extremities of the heteroduplex are A or T, whereas in the other they are G or C (Figure 4). Let us note $p_s$ the probability of conversion when the template haplotype is $GC_f$ and $p_w$ the probability of conversion when the template haplotype is $AT_f$. The repair of mismatches by MMR is nick-directed. By definition, one nick is always present on the strand corresponding to the broken haplotype. But the resolution of recombination intermediates may also involve the formation of nicks on the other strand. According to our model, in that situation, the direction of repair is biased towards the $GC_f$ haplotype. Hence, $p_s > p_w$.

Many of the recombination pathways involve the invasion of the template chromatid by both single-stranded 3' overhangs of the broken chromatid, and hence involve the formation of two heteroduplex DNA (Figure 1). The extent of the detected conversion tract (and hence the nature of the SNPs at its boundaries) depends on the direction of repair (conversion or restoration) on both heteroduplex. We assume that for a given haplotype, the nature of the alleles (GC or AT) at the polymorphic sites located in one heteroduplex is independent of the alleles present in the other heteroduplex (i.e. we assume that there is no correlation between the nature of neighboring alleles on a same haplotype). Under this assumption, the direction of repair (restoration or conversion) in one heteroduplex is independent of the direction of repair in the other one.

In Figure 4, we present the different conversion tracts that can be obtained, depending on whether each of the two heteroduplex is subject to conversion or restoration. When both heteroduplex are subject to restoration, no conversion tract can be observed. The other cases can lead to 5 different configurations, with relatively long (length = $L_2$) or short (length = $L_1$, with $L_2 > L_1$) conversion tracts. In Figure 4, we present in detail the different possible outcomes of mismatch repair by MMR for one heteroduplex (heteroduplex II). Here we focus on cases where all mismatches in the heteroduplex are repaired in the same direction (i.e. we focus on cases that lead to simple conversion tracts, which represent the vast majority of recombination events; see Main Text). For the sake of simplicity, we consider that in all cases, the second heteroduplex (heteroduplex I) is subject to conversion with probability $x$ or restoration with probability $(1 - x)$. We assume that there is no initiation bias, i.e. the frequency of cases where the broken haplotype is $AT_f$ (case A in Figure 4) is equal to the frequency of cases

where the broken haplotype is $GC_f$ (case B). Under those assumptions, it is possible to compute the relative abundance of each of the five configurations of detectable conversion tracts (Table S3).

In our study, we measured the conversion bias among conversion tracts for which the two parental haplotypes present $AT_f/GC_f$-polymorphisms, i.e. only a subset of all observed conversion tracts (see Main Text). As we assume that there is no correlation between the nature (GC or AT) of neighboring alleles on a same haplotype, the nature of N::N mismatches present in chromatid I heteroduplex is independent of the identity of mismatches in chromatid II heteroduplex. Hence, we can assume that in Table S3, each N::N parental mismatch has a probability *r* to be W::S or S::W and 1-*r* to be W::W or S::S. Thus, we can derive the relative abundance of $GC_f$ and $AT_f$ haplotypes expected among the subset of conversion tracts with $AT_f/GC_f$-polymorphism, for each of the five cases (Table S4). From this, we can compute the expected relative abundance of $GC_f$ and $AT_f$ haplotypes, among the subset of conversion tracts with $AT_f/GC_f$-polymorphism:

$$\begin{aligned} n_{GCf} &= r^2(2-p_W-p_S)x + rp_S x + p_S(1-x) \\ n_{ATf} &= r^2(2-p_W-p_S)x + rp_W x + p_W(1-x) \end{aligned} \qquad (5)$$

Thus,

$$n_{GCf} - n_{ATf} = (p_S - p_W)(1 + x(r-1)) \qquad (6)$$

According to our model $p_s > p_w$. Given that *x* and *r* are probabilities, $(1+x(r-1))$ is positive. Hence, our model predicts that $n_{GCf} - n_{ATf}$ is positive, in agreement with the observation that conversion tracts with $AT_f/GC_f$-polymorphism lead to an over-transmission of $GC_f$ haplotypes (Table 2).

Let us now consider the average length of $GC_f$ and $AT_f$ conversion tracts: $L_{GCf}$ and $L_{ATf}$. Using the relative abundance of the different tracts and their length (Table S4), we can derive the following results:

$$\begin{aligned} L_{GCf} &= \frac{r^2(2-p_W-p_S)xL_1 + rp_S xL_2 + p_S(1-x)L_1}{n_{GCf}} \\ L_{ATf} &= \frac{r^2(2-p_W-p_S)xL_1 + rp_W xL_2 + p_W(1-x)L_1}{n_{ATf}} \end{aligned} \qquad (7)$$

Thus

$$L_{GCf} - L_{ATf} = \frac{r^3 x^2 (2 - p_W - p_S)(p_S - p_W)(L_2 - L_1)}{n_{GCf} n_{ATf}} \qquad (8)$$

Given that $L_2 > L_1$, and $p_s > p_w$ this value is expected to be positive. Hence, in agreement with our observations (see Main Text), our model predicts that, on average, $GC_f$ conversion tracts should be longer than $AT_f$ conversion tracts.

## II.D. Bilan de l'étude et perspectives

Cette étude a amené plusieurs résultats importants.

- Tout d'abord, elle a permis de rejeter l'hypothèse du gBGC dû au BER [Marais, 2003] car le signal de gBGC est concentré dans les tracts de conversion simples qui ne peuvent pas être issus de la réparation des hétéroduplexes méiotiques par ce système.

- Le fait que le gBGC soit retrouvé uniquement dans les tracts associés à des CO montre que le biais est instigué après la différenciation des voies donnant lieu à des CO ou NCO. Ceci permet de rejeter la possibilité d'un biais causé lors de l'initiation ou lors du rejet de brin en méiose. Ainsi, dBGC et gBGC sont deux processus qui trouvent leur origine moléculaire dans deux mécanismes différents malgré une dynamique spatiale et temporelle commune associée à la recombinaison.

- Le gBGC semble être dû à un biais dans la réparation des mésappariements. Le modèle que nous proposons fait intervenir le système MMR qui est capable de convertir l'ensemble des marqueurs d'un même tract de conversion dans le même sens. Ainsi, si ce système est biaisé vers les allèles riches en GC, il conduit à un signal de gBGC majoritairement retrouvé dans les tracts simples, ce qui est conforme aux observations. Un biais peut alors être instigué dans ce processus si on imagine que le choix de brin effectué par le MMR lors de la réparation des hétéroduplexes est lui-même biaisé. Sur la base de l'étude fine de la transmission des allèles retrouvés aux extrémités des tracts de conversion, nous proposons un modèle dans lequel ce biais est en faveur du brin le plus riche en GC précisément à ces sites. Rappelons que ces sites sont, avant réparation, ceux des mésappariements les plus proches des cassures simple brins (nick) utile à l'amorce de la réparation par le MMR.

Récemment, Odenthal-Hesse et collaborateurs ont montré, au niveau deux points-chauds humains, l'association d'un biais de conversion vers GC uniquement dans des tracts de NCO [Odenthal-Hesse et al., 2014]. Ce BGC n'est pas du dBGC car il est associé à un seul type de voie de recombinaison (NCO) et que les allèles AT (moins fréquemment transmis) ne semblent pas spécifiquement associés au fort taux de recombinaison de ces points chauds. Il pourrait s'agir de gBGC. Cependant, pour affirmer que le gBGC est associé aux NCO chez l'homme, il faudrait montrer cette association à l'échelle du génome entier, d'autres études doivent donc venir confirmer ou infirmer cette

hypothèse. Ainsi, si cette association est confirmée, elle serait en opposition avec ce que nous avons observé dans notre étude : le gBGC est spécifique des CO chez la levure . Il faudra donc imaginer que le gBGC trouve des origines moléculaires différentes à l'échelle des eucaryotes. Sous ce modèle, le gBGC serait donc apparu plusieurs fois de manière convergente dans différentes lignées de ce domaine. Il est aussi possible, *a contrario* que le signal enregistré par Odenthal-Hesse et collaborateurs ne soit pas du gBGC. La paire de nucléotides impliquée dans chacun de ces points chauds montrerait donc un biais de conversion déterminé par un élément autre que leur nature (G ou C).

Notre modèle propose, en outre, une origine évolutive au gBGC. Les mutations spontanées étant plus fréquentes de GC vers AT, le MMR permettrait un rééquilibrage en favorisant les réparations donnant lieu à des changement de AT vers GC en mitose, là où il est majoritairement actif dans la lignée germinale (200 mitoses pour 1 méiose chez l'homme, mâle, par exemple). Ainsi, en méiose, le biais observé de conversion vers GC ne serait qu'une conséquence de ce biais sélectionné en mitose. Sous ce modèle, l'intensité du biais introduit par le MMR est donc soumise à un trade-off. En effet, même si le MMR permet une baisse du taux de mutation en favorisant les réparations vers GC en mitose, il ne faut pas que ce biais soit trop fort car il aboutirait à un gBGC trop important synonyme de fixation de mutations délétères. Ceci est illustré à la Figure II.1.

Cependant, le modèle proposé reste spéculatif, aussi bien sur le plan moléculaire que sur le plan évolutif. Afin d'améliorer notre connaissance de ces aspects, il serait utile de créer des cartes de recombinaison à haute résolution sur des mutants du système MMR. Ceci a déjà été effectué sur un petit nombre de méioses pour Msh2 [Martini et al., 2011] et Msh4 [Mancera et al., 2008]. Malheureusement, pour chaque étude, le faible échantillon statistique représenté ne permet pas d'analyser le gBGC avec assez de puissance. D'autre part, les protéines du MMR n'agissant pas uniquement sur la réparation des mésappariements en méiose (Figure I.7 p. 33), d'autres étapes du processus de recombinaison sont perturbées chez ces mutants ce qui conduit à des changements dans l'utilisation relative des différentes voies de réparation des DSB [Mancera et al., 2008]. Ceci pourrait constituer un facteur confondant important pour notre approche.

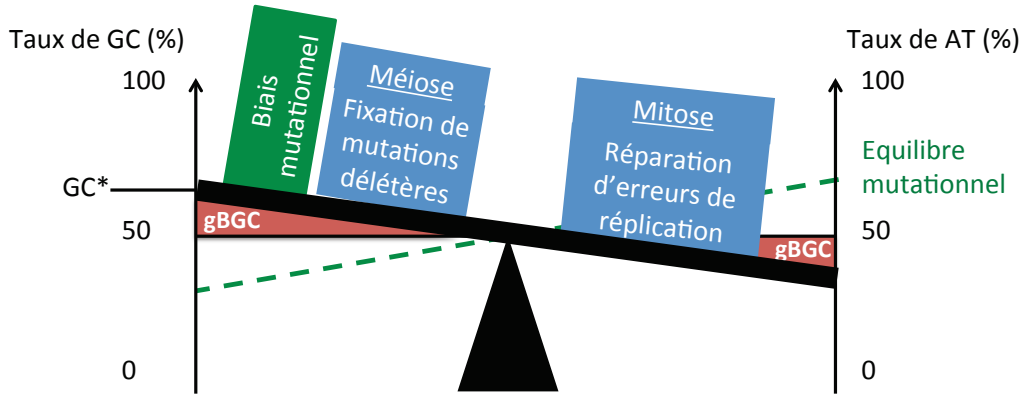



FIGURE II.1 : Modèle proposé de gBGC dû à une réparation biaisée par le MMR. L'inclinaison de la balance montre l'intensité du biais de réparation occasionné par le MMR. Cette intensité est tempérée par deux pressions de sélection qui s'opposent (boites bleues). La première est favorable : elle maintien ce biais permettant la réparation d'erreurs dues aux mutations elles-même biaisées vers AT (boite verte) qui apparaissent en mitose. La seconde est défavorable : elle tend à diminuer ce biais qui conduit au gBGC (en méiose) dont les conséquences peuvent être délétères. Lorsque le biais est assez fort, il surpasse le biais mutationnel et favorise les bases GC ce qui permet l'observation du gBGC (aires rouges). Les axes représentent les taux de GC et AT globaux attendus sous les deux contraintes : gBGC et biais mutationnel, à l'échelle du génome.

Chapitre *III*

# Etude de la dynamique et de l'intensité du BGC associé aux points chauds de recombinaison humains

# III.A. Présentation de l'étude

Après s'être penché sur l'origine moléculaire et évolutive du gBGC, nous allons nous intéresser à deux autres aspects majeurs de ce phénomène.
Tout d'abord, nous allons étudier la dynamique temporelle du gBGC et du dBGC chez l'homme. Comme vu en introduction, cette dynamique est conditionnée par celle des points chauds de recombinaison, elle même principalement orchestrée par l'apparition et la disparition des allèles de PRDM9 dans les populations. Cette évolution rapide semble, en effet, être à l'origine du fait qu'à petite échelle génomique, les patrons de recombinaison ne sont pas conservés entre l'homme et le chimpanzé [Auton et al., 2012]. Ceci indique que le patron d'action actuel du BGC à l'échelle du génome humain est vieux de 6 à 12 Ma au maximum (date de divergence entre l'homme et le chimpanzé [Langergraber et al., 2012]). Cependant, la rapidité de l'évolution de PRDM9 suggère que cette limite supérieure pourrait être beaucoup plus récente. C'est pourquoi, nous avons voulu préciser cette date en utilisant le génome de l'hominidé de Denisova [Meyer et al., 2012] (groupe frère de Neanderthal) dont la séparation avec l'homme date de 0,4 à 0,8 Ma [Langergraber et al., 2012]. Ce génome, diploïde, séquencé avec une bonne couverture (30x), offre l'opportunité d'explorer les processus moléculaires ayant agit dans les 10% les plus récents de la branche qui sépare l'homme du dernier ancêtre partagé avec le chimpanzé. Afin de déterminer si l'hominidé de Denisova et l'homme partagent les mêmes points chauds, nous allons analyser l'évolution des motifs ciblés par l'allèle majoritaire de *PRDM9* chez l'homme : l'allèle A. Comme vu dans [Myers et al., 2010] la perte de ces motifs témoigne de l'activité locale du dBGC et donc de la recombinaison. Nous analyserons de même les patrons de gBGC chez l'hominidé de Denisova et l'homme afin de déterminer si leur établissement précède ou non la séparation des deux lignées.
D'autre part, nous allons quantifier l'intensité du dBGC agissant sur les motifs ciblés par PRMD9 chez l'homme grâce aux données des 1000 Génomes [The 1000 Genomes Project Consortium, 2012]. Ceci permettra de déterminer si la distribution de $B$ estimée à l'échelle du génome est compatible avec l'hypothèse d'évolution des points chauds de recombinaison sous un modèle de Reine Rouge (*cf.* I.B.3.). En effet, le modèle de Reine Rouge prédit que c'est la destruction du nombre de sites reconnus par PRDM9 *via* le dBGC qui est à l'origine de la montée en fréquence de nouveaux variants de *PRDM9*. Pour que ce modèle soit réaliste, il faut donc que le dBGC soit assez fort pour que cela occasionne une perte sensible du nombre de points chauds à l'échelle du génome entier.

# III.B. Des séquences humaines modernes et archaïques soutiennent le modèle de Reine Rouge des points chauds de recombinaison

*Voir article joint page suivante.*
En review chez *PLoS Genetics*.
Les informations supplémentaires sont disponibles à la section III.C. p. 158.

*L'analyse du génome de Denisova et le modèle de Reine Rouge d'évolution des points chauds de recombinaison - par Pauline Sémon (www.p-sem.com).*

# The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes

Yann Lesecque[1], Sylvain Glémin[2], Nicolas Lartillot[1], Dominique Mouchiroud[1] and Laurent Duret[1].

[1] Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, Villeurbanne, France

[2]Institut des Sciences de l'Evolution, UMR CNRS 5554, Université Montpellier 2, Montpellier, France

Corresponding author: Laurent Duret, laurent.duret@univ-lyon1.fr

Recombination is a major molecular process, which increases diversity by disrupting genetic linkage between loci and ensures the proper segregation of chromosomes during meiosis. In the human genome, recombination events are clustered in hotspots, whose location is determined by the PRDM9 protein. There is evidence that the location of hotspots evolves rapidly, as a consequence of changes in PRDM9 DNA-binding domain. However, the reasons for these changes and the rate at which they occur are not known. In this study, we investigated the evolution of human hotspot loci and of PRDM9 target motifs, both in modern and archaic human lineages (Denisovan) to quantify the dynamics of hotspots turnover during the recent period of human evolution. We show that human hotspots started to be active shortly before the split between Denisovans and modern humans, about 0.7 to 1.3 MYR ago. Surprisingly, however, our analyses indicate that Denisovan recombination hotspots did not overlap with modern human ones, despite sharing similar PRDM9 target motifs. We further show that high-affinity PRDM9 target motifs are subject to a strong self-destructive drive, known as biased gene conversion (BGC), which should lead to the loss of the majority of them in the next 3 MYR. This depletion of PRDM9 genomic targets is expected to decrease fitness, and thereby to favor new PRDM9 alleles binding different motifs. Our refined estimates of the age and life expectancy of human hotspots provide empirical evidence in support the Red Queen hypothesis of recombination hotspots evolution.

## Introduction

Meiotic recombination is a highly regulated process, initiated by the programmed formation of double-strand breaks (DSBs). These DSBs are subsequently repaired, using homologous chromosomes as a template, thus leading to crossover (CO) or non-crossover (NCO) recombination events. In mammals, as in many other eukaryotes, the formation of at least one CO on each chromosome is required for the proper disjunction of chromosomes during meiosis (for review see [1]). Hence, the recombination machinery must be tightly controlled to promote a sufficient number of COs on each chromosome, while ensuring that all DSBs can be efficiently repaired to produce viable gametes.

Recombination events are not randomly distributed across the genome, but cluster in hotspots, typically 1 to 2 kb long [2–8]. About 33,000 recombination hotspots have been identified in the human genome, which account for 60% of COs and 6% of the sequence [2]. Many independent observations have clearly demonstrated that in human and mouse, the location of hotspots is primarily determined by the zinc finger protein PRDM9, through its sequence-specific DNA-binding domain [9–12]. PRDM9 contains a SET domain, which catalyzes histone H3 Lys4 trimethylation (H3K4me3) at hotspot loci [4,12–14]. PRDM9 is highly polymorphic, specifically in its DNA binding domain, and the location of recombination hotspots differs among individuals carrying different alleles [9,11,12,15]. At the population scale, the set of recombination hotspots that are the most frequently used can be inferred from patterns of linkage disequilibrium [3] or of genetic admixture [16]. These analyses revealed that more than 90% of recombination hotspots are shared between European and African populations [16]. This strong overlap is due to the fact that the same major allele of PRDM9 (allele A) is present at high frequency both in European and African populations [11]. Interestingly, this A allele presents affinity for the 13-bp motif *CCTCCCTNNCCAC*, which was initially identified on the basis of its enrichment within human recombination hotspots [17] (we will hereafter refer to this sequence motif as HM – for human hotspot motif).

It has been shown that the location of recombination hotspots is not conserved between human and chimpanzee [18–20]. This rapid shift is due to the fact that the major PRDM9 alleles present in each species have different DNA binding specificities [10,19]. There is clear evidence that PRDM9 has evolved under strong positive selection, in primates as well as in many other animal lineages, specifically at those sites involved in DNA sequence recognition [21,22]. This indicates that PRDM9 has been under selective pressure to switch to new targets [21,22]. However, the reasons for this selective pressure remain mysterious.

One interesting hypothesis, proposed by Myers and colleagues [10], is that the turnover of PRDM9 alleles might be a consequence of the self-destruction of recombination hotspots by the process of biased gene conversion (BGC) [23–25]. Indeed, the repair of DSBs is expected to lead to the conversion of recombination-prone alleles by hotspot-disrupting alleles [23–25] (we will hereafter refer to this form of BGC as 'dBGC', for DSB-driven BGC). In agreement with the dBGC model, it was shown that the HM motif was subject to accelerated evolution in the human lineage [10]. The authors suggested that the progressive degradation of recombination hotspots through dBGC might lead to a loss of fitness. Indeed, there is evidence that lower CO rates are associated with lower fertility, possibly due to improper chromosome disjunction [26]. Hence, the loss of PRDM9 target motifs might favor the increase in frequency of new PRDM9 alleles, targeting different motifs [10].

Simulation studies have shown that this model, termed the 'red queen theory of recombination hotspots', might explain the rapid turnover of recombination hotspots [27]. It is however not established whether this model is quantitatively realistic. Notably it has been argued that the number of human recombination hotspots (~30,000) largely exceeds the number of COs per meiosis (~60) and hence is unlikely to be limiting [22]. Thus, one key issue is to determine whether, during the lifespan of a given PRDM9 allele, the loss of its target motifs by dBGC is fast enough to have a significant impact on genome-wide recombination patterns. To address this issue, we first determined when the HM motif started to be the target of PRDM9 allele A in the human lineage. For this, we analyzed the genome sequence of Denisovan [28], an archaic human that diverged from the modern humans about 400,000-800,000 years ago [29]. We then used polymorphism data to quantify the strength of dBGC on HM motifs in extant human populations. This combined analysis of polymorphism and divergence, made possible thanks to Denisovan genomic data, demonstrates that the life expectancy of human recombination hotspots is very short, and brings support for the red queen theory of recombination hotspots.

Results

**The HM motif started to be targeted by PRMD9 shortly before the Human-Denisovan split**

The major human allele of PRDM9 (allele A, present at a frequency of 86% in European populations and 50% in African populations [11]) recognizes a specific sequence motif, whose core consensus is *CCTCCCTNNCCAC* [9,10]. This motif promotes recombination specifically in humans, not in chimpanzee, and is particularly active in the context of THE1 transposable elements [10,19]. As predicted by the self-destructive dBGC drive model, it was previously shown that this motif has accumulated an excess of substitutions specifically in the human lineage, after its divergence from chimpanzee, and that the HM loss rate was particularly strong within THE1 elements [10]. Based on the dBGC model [25], the authors proposed that the erosion of HM motifs might have started about 1 to 2 million years ago [10]. This estimate was however based on poorly known parameters, and was therefore provided as a conservative upper bound [10]. To obtain a more direct dating of the onset of the HM motif activity, we used the Denisovan genome so as to determine when the HM motifs started to be subject to dBGC during the evolution of modern and archaic humans (Figure 1). We analyzed the evolution of HM motifs both within and outside human recombination hotspots. For this, we used recombination maps inferred by HapMap from patterns of linkage disequilibrium in human populations [2]. These maps reflect the average crossover rates across human populations over many generations. We will hereafter refer to these data as human "historical" recombination rates. Given that the list of human historical hotspots is currently available only for autosomes, we excluded sex chromosomes from our analyses.

We first identified HM motifs (N=5,704) in the reconstructed autosomal sequences of the human-chimpanzee ancestor (HC), and then counted base replacement changes along the four branches of the phylogeny (hereafter termed modern human, Denisovan, Hominini and chimpanzee branches, Figure 1), by comparing sequences of reference genomes to the ancestral one (see methods). It should be noted that the detected base changes include both fixed and polymorphic mutations. To quantify the

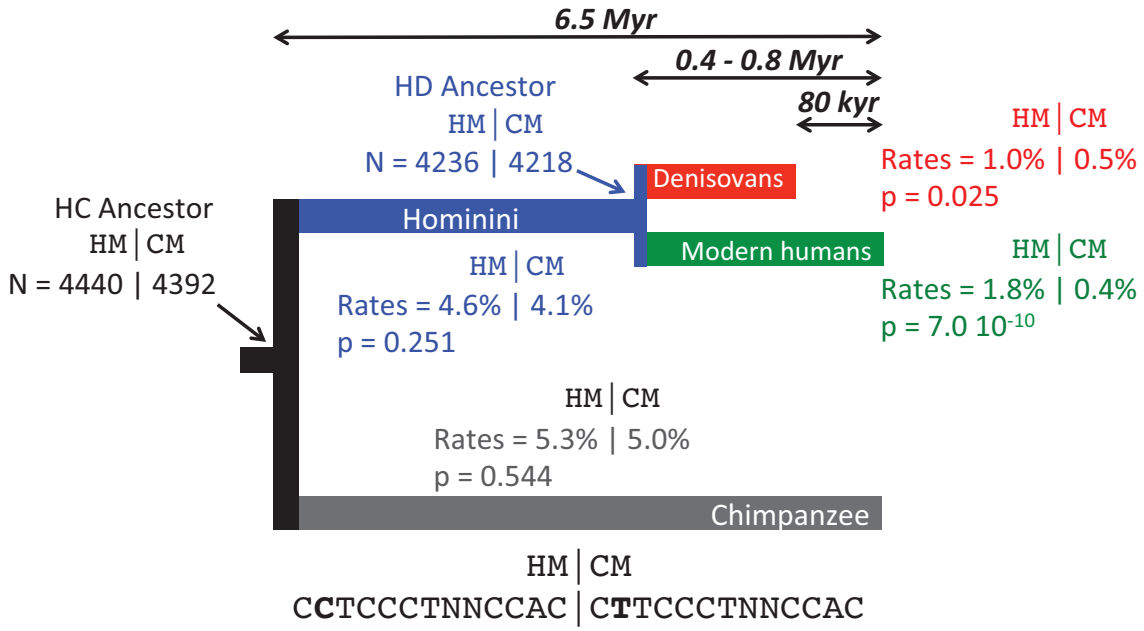



**Figure 1. Differential loss of HM motifs across recent human history.**

The number of intact HM and CM motifs found in the reconstructed sequence (F2 subset) of human and chimpanzee last common ancestor (HC) and in the last human-Denisovan common ancestor (HD) is indicated with a simple arrow. Figures on the left (respectively right) of the bar refer to HM (respectively CM) motifs. On each branch, *Rates* indicate the motif loss rate by point mutation for HM and CM motifs and *p* is the p-value of a proportion test comparing HM and CM loss rates. Sequences of both motifs are shown below the tree. Double arrows represent average sequence divergence times and difference in branch length as estimated in [28,29].

excess of base changes (if any) on HM motifs along each branch of the phylogeny, we used as a reference the rate of base change within a control motif (CM: *CTTCCCTNNCCAC, N=5,483*), which differs from HM by the second position and does not show any effect on the recombination pattern [30] (Figure 2).

We counted base changes only at informative sites (i.e. we ignored the two N positions) and excluded the second position, which differs between HM and CM motifs. Thus, we only examined sites that are *a priori* expected to have the same rate of mutation (and possibly sequencing errors) in HM and CM motifs. We considered a motif to be lost as soon as it was subject to one mutation in one informative site. To minimize errors in the inference of motif losses, it is necessary to avoid regions with low sequencing quality or erroneous alignment. Thus, we created three levels of filters (F1, F2 and F3) successively applied to our data so as to keep three subsets of motifs. A motif is discarded from a subset if at least one informative site does not pass the filter. Filter F1 retains all aligned sites common to human, chimpanzee and Denisovan, while filter F2 favors a more accurate HC ancestral sequence reconstruction. Finally, the most stringent filter F3 accounts for sequence errors specific to ancient DNA in Denisovan (see methods). Unless explicitly mentioned, results presented below correspond to the F2 dataset, totalizing 4,440 HM and 4,393 CM motifs present in the human-chimpanzee ancestor.

In the modern human branch, we observed that the HM loss rate (1.8%) is more than four times higher than the CM loss rate (0.4%; green branch in Figure 1). As expected, the HM loss rate is much higher within THE1 elements (6.7% vs. 1.7%; proportion test: $p = 8.2\ 10^{-5}$) (Table 1). However, the excess of HM losses is not limited to THE1 elements: at non-THE1 loci, the HM loss rate is significantly higher than the CM loss rate (1.7% vs. 0.4%, $p = 2\ 10^{-8}$). Conversely, we observed no significant difference in HM and CM loss rates along the Chimpanzee branch (in grey in Figure 1), as expected given that the HM motif is not a target of PRDM9 alleles in chimpanzees [10,19]. This negative control confirms that there is no intrinsic difference in mutation rate between the two motifs, and hence that the CM motif is a good reference to detect accelerated evolution of the HM motif.

These observations are consistent with the self-destructive dBGC drive model. However, they could also be explained by a possible mutagenic effect of recombination. To distinguish between these two possibilities, we analyzed the derived allele frequency (DAF) spectra of mutations in HM and CM motifs: under the hypothesis that the increased HM loss rate is simply due to a higher mutation rate (and not a fixation bias, like BGC), the two DAF spectra are expected to be identical. We included in these analyses all modern-human mutations detected as polymorphic by the 1000 genomes project [31], as well as fixed ones. We observed that the DAF spectrum of HM mutations is shifted towards higher frequencies compared to CM mutations (Figure 3), with an average mean DAF almost three times higher (13% vs. 5%; Wilcoxon test $p = 1.9\ 10^{-6}$). Overall 3.7% of HM mutations detected in the modern human branch are fixed, compared to 0.2% for CM mutations (Proportion test $p = 1.6\ 10^{-4}$). This demonstrates that the accumulation of HM losses in the human branch is a consequence of a fixation bias, as predicted by the dBGC model [32].

In the Hominini branch, ancestral to Denisovans and modern humans (in blue in Figure 1), the HM loss rate appears slightly higher than the CM loss rate, but this difference is not statistically significant. In the Denisovan branch (in red in Figure 1), the HM loss rate (1%) is two times higher than the CM loss rate (0.5%). This excess is weaker than that observed in the modern human branch, but it is still significant

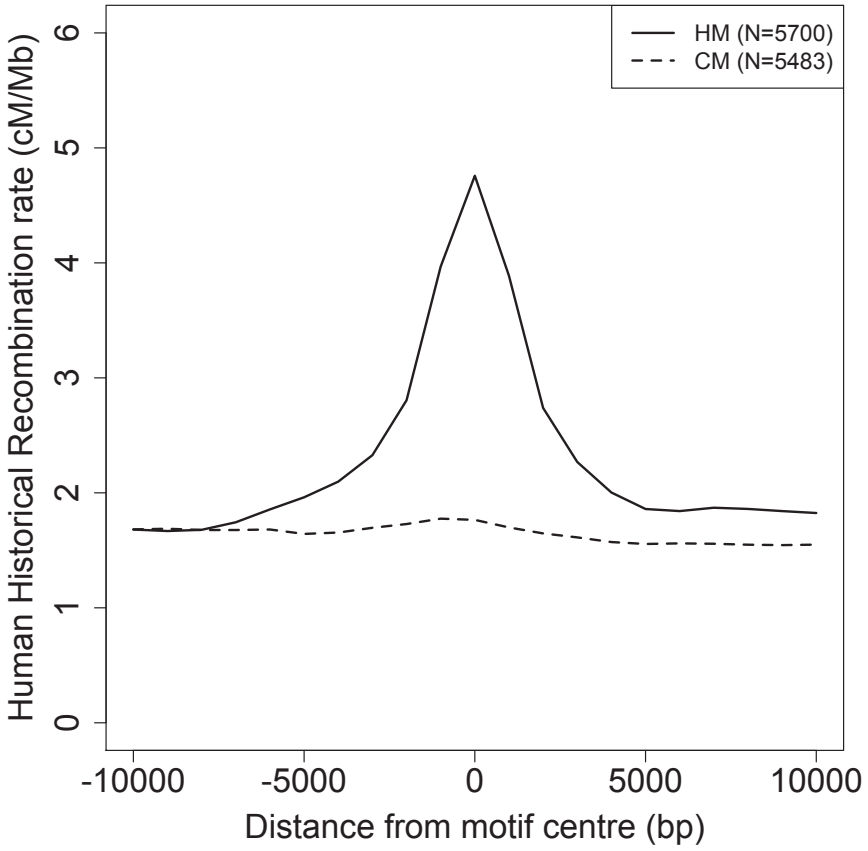



**Figure 2. Modern human recombination profiles around HM and CM motifs found in the HC-ancestral sequence.**

Human historical recombination rates (cM/Mb) around CM (dotted line) and HM (solid line) motifs found in the human-chimpanzee reconstructed ancestral sequence (F1 subset). Recombination rates are averaged on 2kb overlapping windows (overlap = 1kb).

**Table 1. HM motifs loss rates within *versus* outside THE1 elements**

| Branch | Within THE1 | | Outside THE1 | | p[c] |
|---|---|---|---|---|---|
| | N[a] | Rate[b] | N[a] | Rate[b] | |
| Chimpanzee | 145 | 5.5% | 4295 | 5.3% | 0.999 |
| Hominini | 145 | 7.6% | 4295 | 4.5% | 0.122 |
| Denisovan | 134 | 0.0% | 4102 | 1.0% | 0.463 |
| Human | 134 | 6.7% | 4102 | 1.7% | $8.2\ 10^{-5}$ |

[a] Intact motif count at ancestral edge of the branch (cf. Figure 1)
[b] Motif loss rate along the branch
[c] P-value of proportion test comparing HM loss rates within *vs.* outside THE1 elements along the branch

(p=0.025). Additionally, the rate of homozygosity of these mutations (computed using the diploid sequence of the Denisovan individual) is higher for HM than for CM (0.80 *vs.* 0.69). This trend is consistent with the hypothesis that in Denisovans, as in modern humans, HM mutations segregated on average at higher frequency than CM mutations. The fact that the signature of dBGC on HM motifs is weaker in Denisovan compared to human might be explained by slightly different sequence affinities of their PRDM9 alleles, or by a lower population frequency of HM-targeting PRDM9 alleles in Denisovans. This weaker signature of dBGC might also be due to the fact that the effective population size was smaller in Denisovans compared to modern humans [33], which is expected to enhance the effects of random genetic drift, and hence to decrease the strength of dBGC [32].

Given that ancient DNA is prone to sequencing errors, we repeated our analyses with more stringent criteria to keep only data with the highest sequence quality (filter F3). In that F3 subset, we found the same two-fold excess of HM losses compared to CM losses in the Denisovan branch (Table S1). The above results are robust to data filtering criteria (F1, F2 or F3; Table S1, S2 and Figure 1). The only notable difference is that in the F1 data set, due to the larger sample size, the slight excess of HM losses in the Hominini branch is detected as statistically significant (Table S2). All these observations indicate that HM has been subject to dBGC both in Denisovans and modern humans lineages, and suggest that HM started to be a target of PRDM9 shortly before the Denisovan/modern human split.

**Quantifying the intensity of dBGC on HM motifs in the human branch**

To estimate the intensity of dBGC against HM motifs, we fitted a population genetic model to the DAF spectra of CM and HM mutations (Figure 3), using a maximum likelihood framework. Since dBGC behaves like selection on semi-dominant mutations [32], we used the model of Eyre-Walker et al. [34] to quantify it. Under the simplifying assumption that all HM informative sites are subject to the same dBGC strength, the population scaled dBGC coefficient ($G = 4N_e g$) estimated on all HM motifs is 8.55 (95% confidence interval = 2.76-2655). This result is robust to the number of categories used to describe DAF spectra (Table S3). It should be noticed that large values of G are difficult to estimate accurately because above a given threshold (G>20), all values of this parameter are expected to give very similar DAF spectra (Figure S1). This explains why the upper bound of the confidence interval of this estimate of G is very high.

The strength of dBGC at a given locus is proportional to the absolute difference in recombination rate between the original (hot) allele and the mutant (colder) allele [25]. This difference can be large only if the recombination rate at this locus is high. Hence HM motifs that are located at lowly recombining loci are not expected to undergo dBGC. It is important to note that the recombination rate at HM motifs is highly variable across the genome: 8% of HM motifs concentrate 60% of all crossover events located in the vicinity of HM motifs (± 2 kb) (Figure 4). It is therefore expected that the intensity of dBGC should be stronger for HM motifs located in a genomic context prone to recombination. To test that prediction, we re-estimated $G$ from DAF spectra, in three subsets of equal sample size, binned according to the local historical recombination rate (measured on a 2 kb window centered on motif position). As expected, $G$ increases with increasing historical recombination rates, from $G = 0.96$ in the first tercile to $G = 14.64$ in the third tercile (Table S4).

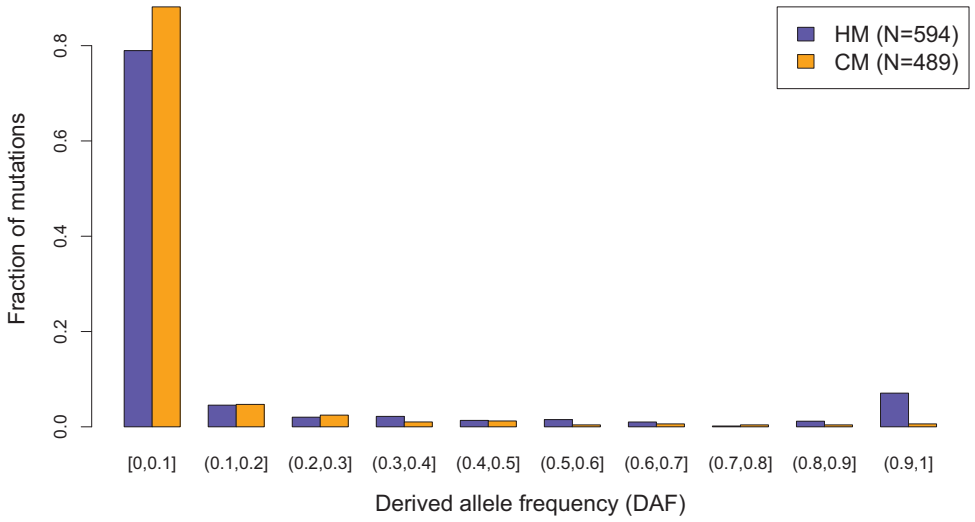



**Figure 3. Derived allele frequency (DAF) spectra of mutations leading to motifs loss in the human branch.**

DAF of mutations affecting HM (purple bars) and CM (orange bars) along the human branch (green branch in Figure 1). Allele frequencies are extracted from 1000 genomes phase I, using all available populations [31]. Mutations count for each motif (F2 subset) is indicated (N).

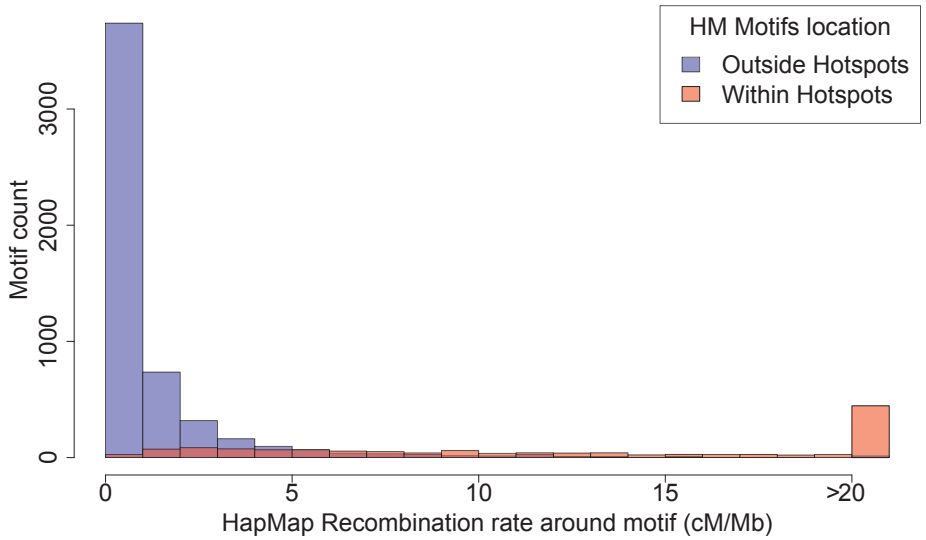



**Figure 4. Distribution of recombination rates around HM motifs of the human genome.**

Human historical recombination rates (cM/Mb) are measured over a 2kb window centered on HM motifs from human autosomes (hg19 assembly; no filter). Red: motifs located within HapMap recombination hotspots. Blue: motifs located outside hotspots.

To get a better picture of the distribution of *G* across all HM motifs, we fitted a simple model where the dBGC coefficient at a given locus is directly proportional to the local crossover rate at this locus (Supplementary Text S1). Given the observed distribution of recombination rates around HM motifs (Figure 4), this model suggests that the 8% most highly recombining motifs should be subject to very strong dBGC (on average, G=174, CI: 29-291).

**Mutations of HM motifs do not immediately silence recombination hotspots**

The dBGC model predicts that the small subset of HM motifs located in a highly recombining context should accumulate substitutions extremely rapidly. In agreement with that prediction, we observed that, along the modern human branch, the loss rate is almost 3 times higher for HM motifs located within historical hotspots compared to other HM motifs (3.5% *vs.* 1.2%; $p = 4.6\ 10^{-7}$). Overall, 55% of the HM motifs detected as being mutated along the modern human branch are located within historical recombination hotspots (compared to 28% for motifs that have remained intact) (Table 2). Thus, on average, the historical recombination rate at HM motifs mutated in the modern human branch is more than two times higher than that at intact HM motifs (11.2 cM/Mb vs. 4.9 cM/Mb; Figure 5). Noteworthy, we observed the same pattern with present-day recombination rates, inferred from pedigree-based genetic maps [35] (Figure S2). Moreover this pattern is observed even for the subset of HM mutations that are fixed in human populations (Figure S3). These observations show that mutations of HM motifs that were fixed in modern humans are generally located in loci that still have a high recombination activity in present-day populations. Hence, although mutations of HM motifs diminish the local recombination rate, they generally do not directly convert a hotspot into a coldspot.

Interestingly, HM motifs that are located outside of historical recombination hotspots also show a signature of dBGC. This signature is weaker than for HM located within hotspots, but still clearly significant: there is a 3-fold excess of HM losses compared to CM losses in the modern human branch (Table 2), and HM mutations segregate at higher frequencies than CM mutations (10% *vs.* 4%; p = 0.0014). This suggests that these HM losses occurred in ancient recombination hotspots that are not active anymore. Overall, we detected 78 HM losses along the modern human branch, whereas only 19 would have been expected if the loss rate were the same as that of CM motifs. Among these 59 extra losses that can be attributed to dBGC, 23 occurred at loci that are not detected as recombination hotspots (Table 2). Thus, among all loci that used to be recombination hotspots in the human lineage and that have lost the HM motif by dBGC, 39% are no longer active.

**No overlap between human and Denisovan recombination hotspots**

With only one single individual sequenced, it is not possible to establish recombination maps in Denisovans. However, different analyses can be performed to test whether recombination hotspots identified in modern human populations correspond to hotspots in Denisovans.

A first approach to detect past recombination activity consists in analyzing substitution patterns, so as to infer the equilibrium GC-content (denoted GC*) along different branches of the phylogeny (see methods). Many lines of evidence indicate that

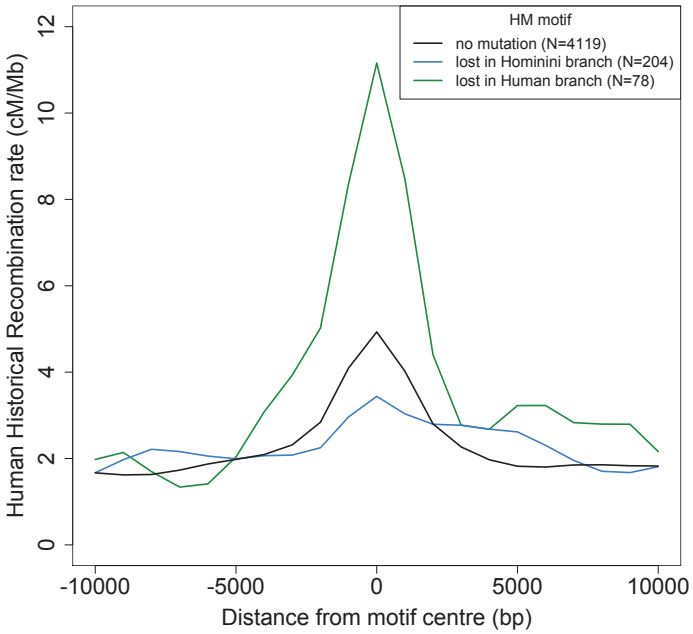



**Figure 5. Modern human recombination profiles around lost and conserved HM motifs.**

Human historical recombination rates (cM/Mb) around HM motifs found in the human-Chimpanzee reconstructed ancestral sequence (F2 subset) and conserved in the human genome (black) or lost in the Hominini (blue) or human (green) branch. Recombination rates are averaged on 2kb overlapping windows (overlap = 1kb).

**Table 2. Motifs loss rates within and outside modern human recombination hotspots (HapMap)**

| Motifs | Branch | N[a] | | Rate[b] | | p[c] | # obs. loss[d] | # exp. loss[e] | # dBGC loss[f] |
|---|---|---|---|---|---|---|---|---|---|
| | | HM | CM | HM | CM | | | | |
| Within hotspots | Chimpanzee | 1272 | 724 | 5.5% | 5.4% | 0.994 | | | |
| | Hominini | 1272 | 724 | 3.9% | 4.3% | 0.792 | | | |
| | Denisovan | 1222 | 693 | 0.8% | 0.4% | 0.485 | | | |
| | Human | 1222 | 693 | 3.5% | 0.6% | $1.2\ 10^{-4}$ | 43 | 7 | 36 |
| Outside hotspots | Chimpanzee | 3167 | 3669 | 5.2% | 4.9% | 0.605 | | | |
| | Hominini | 3167 | 3669 | 4.9% | 4.0% | 0.109 | | | |
| | Denisovan | 3013 | 3521 | 1.1% | 0.5% | 0.024 | 32 | 15 | 17 |
| | Human | 3013 | 3521 | 1.2% | 0.4% | $3.3\ 10^{-4}$ | 35 | 12 | 23 |

[a] Intact motif count at ancestral node of the branch (cf. Figure 1)
[b] Motif loss rate along the branch
[c] P-value of proportion test comparing HM *vs.* CM loss rates along the branch
[d] Number of observed HM losses
[e] Number of HM losses expected in absence of dBGC (i.e. based on CM loss rate)
[f] Estimated number of HM losses caused by dBGC (obs. - exp.). NB: these values are reported only for lineages showing evidence of dBGC ($p<0.05$).

in primates, recombination is driving the evolution of GC-content via the process of GC-biased gene conversion (gBGC), which results from a bias in the repair of AT:GC mismatches in heteroduplex DNA during meiotic recombination [36,37]. Notably, it has been shown that GC* strongly correlates with present or past recombination rates [38–40]. We therefore measured GC* separately for each branch of the phylogeny at loci corresponding to the 32,981 human historical recombination hotspots [2]. As expected, we observed a strong peak of GC* centered on the middle of historical recombination hotspots, in the modern human branch (Figure 6D). In agreement with previous results [19], this peak is absent in the chimpanzee branch (Figure 6A), consistent with the fact that human and chimpanzee recombination hotspots do not overlap. Interestingly, we observed only a very limited bump of GC* in the Hominini branch (Figure 6B). This indicates that, up to a recent time, shortly before the Denisovan/modern human split, loci corresponding to human historical recombination hotspots were not subject to gBGC.

Surprisingly, we observed no peak of GC* in the Denisovan branch (Figure 6C). This result was unexpected: given our observations indicating that the HM motif started to be a target of PRDM9 before the split between modern humans and Denisovans, we presumed, *a priori*, that the two populations should share the same recombination hotspots. We first hypothesized that the absence of peak of GC* could be due to the fact that, owing to the relatively low effective population size in Denisovan, gBGC was too weak to leave any detectable signature. To test this hypothesis, we investigated whether we could detect the hallmarks of gBGC in Denisovan, by analysing correlations between GC* (inferred along different branches of the phylogeny) and recombination rates, measured in 1 Mb-windows. At this genomic scale, recombination rates are well conserved between human and chimpanzee [19] and hence are expected to be also conserved in Denisovan. As predicted by the gBGC model, and in agreement with previous results [38–40], we observed a significant correlation between human historical recombination rates and GC* along the modern human branch ($R^2$ = 13%; $p < 10^{-74}$). Noteworthy, this correlation is as strong for GC* computed in the Denisovan branch ($R^2$ = 14%; $p < 10^{-80}$; Figure S4). This indicates that, genome-wide, the signature of gBGC is as visible in Denisovan as it is in human. Thus, the absence of peak of GC* in the Denisovan branch at human recombination hotspots loci cannot be attributed to a possibly weaker gBGC effect in Denisovan. Instead, it indicates that recombination hotspots were not shared between humans and Denisovans.

To further test this conclusion, we used an independent approach. The self-destructive drive model predicts that HM motifs located in recombination hotspots should be subject to stronger dBGC than other HM motifs. Thus, if the location of recombination hotspots was conserved, then HM motifs located in loci corresponding to human recombination hotspots should show an enhanced signature of dBGC not only in human (as shown previously), but also in the Denisovan branch. As already mentioned, we observed an excess of HM losses compared to CM losses in Denisovan (Figure 1), which indicates that there is a detectable signature of dBGC on HM in Denisovan. However, the HM loss rate is not different between HM loci that correspond to human historical hotspots and other HM loci (respectively 0.8% and 1.1%, p = 0.579; Table 2). Thus, we see no evidence for stronger dBGC in Denisovan at the location of human historical hotspots.

Finally, it has been shown that HM motifs located within THE1 transposable elements are particularly prone to recombination in humans [10,17]. As expected, we observed a markedly elevated HM loss rate within THE1 elements in human (6.7%). In

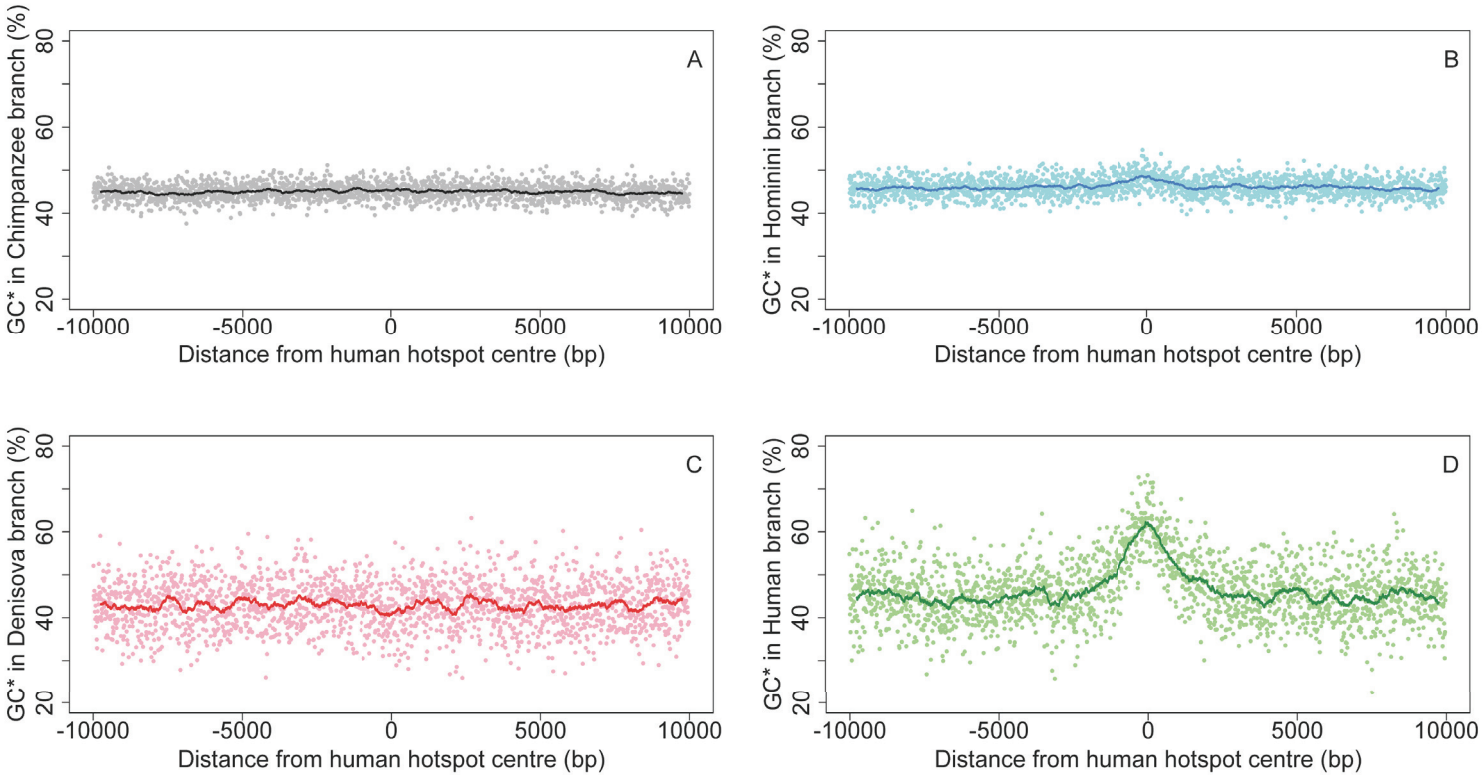



**Figure 6. Equilibrium GC-content (GC*) around human recombination hotspots in different branches of the phylogeny.**

GC* is computed on each branch of the phylogeny (Figure 1): (A) Chimpanzee branch; (B) Hominini branch; (C) Denisovan branch; (D) Modern human branch. Profiles show the mean GC* computed on 32,987 human historical hotspots, over a 20kb region centered on the middle of hotspots. Each dot is the average GC* over a 10bp window. The line shows average GC* over 500bp window.

contrast, we did not detect any mutation in the Denisovan branch among HM motifs located in THE1 (loss rate = 0%; proportion test: p = 0.0067; Table 1). This suggests that contrarily to human, HM motifs located within THE1 elements were not associated to elevated recombination rates in Denisovan.

All these observations concur to the conclusion that fine-scale recombination rates were not conserved between Denisovans and humans. This therefore suggests that the major PRDM9 alleles present in the two populations, despite sharing similar binding sites, targeted different recombination hotspots.

**Expected lifespan of human recombination hotspots**

The lifetime of HM motifs can be predicted using standard population genetic approximation [25]. Under the simplifying assumption that motif mutations are immediately either lost or fixed in the population, the probability that a hotspot motif accumulates at least one disrupting substitution after $T$ generations, can be approximated by:

$$p = 1 - e^{\left(-\mu k T \frac{G}{1-e^{-G}}\right)}$$

where $\mu$ is the mutation rate ($1.2\ 10^{-8}$ mutations/bp/generation in humans [41]), $k$ the length of the motif (here $k$=11 for the HM motif) and $G$ the population-scaled dBGC coefficient ($G=4N_e g$) [25].

Given the distribution of G estimated previously, this model predicts that after 100,000 generations (i.e. about 3 MYR [29]), overall, 18% of HM motifs should be lost (CI: 5%-23%). But importantly, for the subset of most highly-recombining HM motifs (top 8% of HM motifs, which concentrate 60% of HM-associated recombination events), the predicted loss rate is extremely high (87%; CI: 32%-96%). Thus the model indicates that if the dBGC drive against HM remains as strong as it is in extant human populations, then the subset of highly-recombining HM motifs should be rapidly lost. We observed that in many cases, the loss of HM does not totally abolish the hotspot activity. This is most probably due to the fact that the affinity of PRDM9 depends not only on the HM motif, but also on interactions with other sites in its vicinity. However, our observations indicate that losses of a HM motif by dBGC in the human branch were associated with hotspot extinction in 39% of cases. Given that the human branch is relatively short (14,000-28,000 generations), this suggests that within the next 100,000 generations, the loss of HM motifs should be accompanied by the loss of recombination hotspots activity.

Discussion

PRDM9 is the major determinant of the location of recombination hotspots in humans and mice [9–12,15,16,35,42]. At the population scale, the chromosomal distribution of recombination events is therefore expected to depend on the allelic composition at the PRDM9 locus. The location of human historical recombination hotspots reflects the DNA binding specificity of the A allele [9]. This allele is present at high frequency both in African and European populations, and as expected, most historical recombination hotspots are shared between these populations [16]. This implies that the majority of human historical recombination hotspots are older than 50,000 years. To determine more precisely the age of historical hotspots (i.e. to determine when the A allele started to reach substantial frequency within populations),

we searched for signatures of recombination hotspot activity by analyzing patterns of sequence evolution across different branches of the phylogeny, before and after the divergence between modern humans and Denisovans. We used the fact that when a locus is recombining at a high rate (at the population scale), it then becomes subject to two forms of BGC: BGC in favor of mutations disrupting PRDM9 target motifs (dBGC), and BGC in favor of GC-alleles (gBGC).

Along the modern human branch, we observed clear signatures of dBGC against HM motifs: these motifs accumulated an excess of mutations, which tend to segregate at higher allelic frequencies. Moreover, we showed that the strength of this fixation bias in favor of HM-disrupting mutations increases with increasing local recombination rate. All these observations are perfectly consistent with the fact that HM is targeted by the major allele of PRDM9 in human populations (allele A). Interestingly, we also observed an excess of HM losses along the Denisovan branch, which suggests that HM started to be a target of PRDM9 before the population split between Denisovans and modern humans. However, several independent lines of evidence indicate that recombination hotspots were not shared between Denisovans and modern humans: in Denisovan, at loci corresponding to human recombination hotspots, we observed no signature of gBGC and no evidence of stronger dBGC against HM motifs. Moreover, in Denisovan, contrarily to human, HM motifs located within THE1 elements are not subject to accelerated loss.

The fact that fine-scale recombination rates were not conserved between humans and Denisovans might *a priori* seem in contradiction with the observation that the same motif (HM) was subject to accelerated loss in both lineages. However, the affinity of PRDM9 to its targets is not determined by this 13-bp motif alone, but also depends on interactions with surrounding sites [17,43]. For example, in human, the HM motif is much more prone to recombination when located within the context of THE1 [17]. Overall, among the 6,671 HM motifs found in human autosomes, only 1,358 (20%) overlap with one of the 32,987 recombination hotspots identified by HapMap [2]. Thus, only a subset of HM motifs in the human genome are in a context for which the A allele of PRDM9 presents a high affinity. It is therefore possible that the major PRDM9 allele(s) present in Denisovan populations had affinity to HM, but within a different context.

In summary, the fact that Denisovans and humans had different hotspots but similar target motifs suggests that they had slightly different PRDM9 alleles, with distinct context specificity. This conclusion is compatible with two scenario: i) the A allele of PRDM9 was already present at a substantial frequency in the ancestral population (before the population split between Denisovans and modern humans), but was lost (or present at very low frequency) in the Denisovan lineage or ii) the A allele increased in frequency specifically in the modern human branch. The genotype of the Denisovan individual who was sequenced does not correspond to any known human PRDM9 allele (Supplementary Text S2). However data from more individuals would be needed to determine whether or not the A allele was present in Denisovan populations. Noteworthy, at loci corresponding to human recombination hotspots, we observed a small bump of GC* in the Hominini branch (Figure 6B), which shows that some hotspots were already active in the ancestor. Under the simplifying assumptions that all human recombination hotspots started to be active at the same date and that the intensity of gBGC has been constant since then, the onset of human hotspots activity can be estimated to have occurred at about 95% of the length of the Hominini branch (i.e. 0.7 to 1.3 MYR ago, depending on the estimate of the Denisovans/modern human population split date; see Supplementary Text S3). Noteworthy, the small excess of HM losses observed along the Hominini branch also indicates that the acceleration of HM loss rate

started shortly before the population split (see Supplementary Text S3). Thus, the onset of activity of human historical hotspots coincides with the onset of dBGC on HM motifs. The most parsimonious explanation for these observations is that the A allele increased in frequency shortly before the human/Denisovan split.

To understand the dynamics of recombination hotspots it is necessary to establish not only when they were born, but also when they will die. The analysis of DAF spectra indicates that the subset of most highly-recombining HM motifs (top 8% of HM motifs, which concentrate 60% of HM-associated recombination events) is subject to very strong dBGC in extant human populations (G>90). If the intensity of dBGC remains stable over time, then after 100,000 generations (i.e. about 3 MYR), 87% (CI: 32%-96%) of these motifs are predicted to be lost. Up to now, the erosion of HM motifs in the modern human lineage has been quite limited: since the human/Denisovan split (14,000-28,000 generations), only 0.6% of HM motifs have accumulated fixed mutations (1.1% for motifs located within recombination hotspots). This relatively limited erosion can be explained by the fact that initially, when the A allele appeared and progressively increased in frequency in ancestral populations, the intensity of dBGC against HM motifs was certainly much weaker than it is in extant human populations (where the frequency of allele A reaches up to 90%). Moreover, many mutations did not have time to reach fixation. For instance, standard population genetics models [44] indicate that the fixation of a HM mutation subject to strong dBGC (G=90) should take about 9,000 generations on average (for an effective population size in humans of 10,000). Thus, we are just observing the beginning of the erosion of HM motifs. However, in the long term (3 MYR), if the frequency of the A allele remains as high as in extant populations, the vast majority of the most active HM motifs are expected to be lost.

What might be the consequences of the genomic depletion of high affinity PRDM9 target sites? In mice, the knockout of *Prdm9* does not lead to a decrease the number of recombination hotspots [12]. However the location of hotspots in *Prdm9*$^{-/-}$ mice is totally different from that of wild-type mice, with a strong enrichment towards promoters and other sites of PRDM9-independent H3K4 trimethylation [12]. It is therefore plausible that the loss of high affinity PRDM9 target sites would also lead to relocate recombination hotspots to these regions. *Prdm9*$^{-/-}$ mice are sterile, which suggests that this re-patterning of recombination hotspot location is deleterious [12]. Hence, it is expected that the loss of high affinity PRDM9 target sites should provide a strong selective pressure for new PRDM9 alleles, with different DNA binding affinities, to rise in frequency in the population. This constraint is expected to appear much before all high affinity PRDM9 target sites have been lost. Thus, the next turnover of PRDM9 alleles (and hence of hotspot locations) is expected to occur before 3 MYR.

Our observations are therefore consistent with the Red Queen model of hotspot turnover [10,27]. This does not imply that all cases of hotspot turnover are due to this evolutionary scenario. For example, the shift in hotspot location in Denisovans might simply be due to changes in PRDM9 allelic frequencies driven by random genetic drift in this small population. However, this model provides a plausible and simple explanation for the recurrent selective pressure on PRDM9 to switch to new targets, as observed in many animal taxa.

Methods

## Data

We used genomic alignments of Chimpanzee (PanTro2 assembly), Denisovan and modern human (hg19 assembly), published by Meyer and colleagues [28] and available at http://cdna.eva.mpg.de/denisova/VCF/hg19_1000g/. Those files contain different information among which we used the following:

- human-chimpanzee ancestral sequence inferred by Ensembl Compara EPO 6 primate whole genome alignments (Ensembl release 64) [45]
- Denisovan sequence coverage.
- A « TS » string indicating the number of different sequences available for each species of the original Ensembl Compara EPO 6 primates whole genome alignment blocks. This field is used to discard paralogous segments.
- 1000 genomes polymorphism and corresponding averaged allele frequencies (AF) from the 1000 genomes 20101123 intermediate release which contains samples from 1,094 individuals of 15 populations [31].
- Duke mappability scores of 20-mers (Map20), which allows the filtering of regions with low mappability quality.
- Systematic errors (SysErr), which allows the filtering of regions with low Illumina sequencing quality.
- Low Quality (LowQual), which allows to filter regions with uncertain genotype call in Denisovan.

For more information on those annotations, see note 6 in supplementary material of [28].

We used LiftOver software to convert HapMap [2] and DeCODE [35] recombination data from hg18 into hg19 coordinates [46]. This discards 4 HM motifs from experiments using recombination data because their loci are not present in the hg18 assembly.

## Filters

We found 5,704 HM (*CCTCCCTNNCCAC*) and 5,483 CM (*CTTCCCTNNCCAC*) motifs in the human-chimpanzee (HC) reconstructed ancestral sequence described above. Those data were filtered using three increasingly stringent methods named F1, F2 and F3. Each degree of filtering results in a subset of HM and CM motifs (respectively named F1, F2 and F3) used for subsequent analysis. In both motifs the two "N" sites as well as the second position, which is different in CM and HM are classified as non-informative. One motif is used in one given subset if all of its informative sites pass the corresponding level of filtering.

Filter F1 excludes sites that have no genotype call in Denisovan along with those experiencing indels in one of the three species: human, Denisovan or chimpanzee. We excluded indels because the reconstruction of the HC ancestral sequence is particularly difficult at those sites (this specifically excluded 146 HM and 105 CM). This led to the F1 subset composed of 5,474 HM and 5,314 CM motifs.

Filter F2 includes filter F1 and aims at conserving sites for which the HC ancestral sequence is the most reliable. Thus, we only used sites from a filtered subset of the EPO alignment of human, chimpanzee, gorilla and orangutan produced by the Gorilla

Sequencing Consortium [47]. This subset has been previously used and filtered as described below and kindly provided by Kasper Munch [40]: "*To increase data quality the alignment is filtered to remove regions of low sequencing quality and regions with a large proportion of gaps or uncalled bases. All alignment blocks that do not contain one and only one sequence for each of the four species are discarded. Then all alignment columns with a gap in both human, chimpanzee and gorilla sequence are removed. To take base call uncertainty into account we then slide a 10nt window by 1nt. If the mean quality score is below 7 the window is removed and the alignment block is split accordingly. To further filter for gap content we slide a window of size 50 by 1nt. If a window contains 49 gaps or more it is removed and the alignment block is split accordingly. Blocks smaller than 300 are removed. The resulting alignment blocks are joined if less than 100 bases apart (and padded accordingly with 'N'), or split where they contain runs of more than 100 alignment columns of all 'N'*" [40]. Based on the resulting alignment, we retained only positions for which at least 3 out of the 4 primate species were concordant. This filtered alignment (FA) was used to create the F2 motifs subset: 4,440 HM and 4,393 CM and compute GC* estimates (see below).

Filter F3 includes filter F2 and aims at eliminating potential sequence errors occurring in ancient DNA. This filtering was used in note 9 and following in supplementary material of [28] to estimate substitution rates in Denisovan and *Homo sapiens*. One particular site is excluded if:

- it is in a LowQual or a SysErr region as described in the data section above.
- it has a Map20 score different from 1 which indicates potential mapping error.
- it has a Denisovan sequence coverage below 16 which avoids regions with unreliable Denisovan genotype call.
- it has a Denisovan sequence coverage higher than 46 which avoids repeated and duplicated regions.
- it is in an EPO alignment block with more than one human sequence or more than one Chimpanzee sequence, which avoids paralogies.

This stringent filtering left only 2,019 HM and 2,274 CM motifs.

**HM and CM motif loss rates estimation**

To estimate motif loss rates, the number of motifs that are mutated at informative sites along one branch was counted and then divided by the number of intact motifs found at the ancestral node of this branch. For a given motif, if mutations at informative sites occurred both in the Hominini branch and in one terminal branch (modern human or Denisovan), the motif loss is attributed to the Hominini branch. A mutation is inferred in the chimpanzee branch if the human-chimpanzee ancestral sequence differs from the Chimpanzee sequence. Similarly, a mutation is inferred in the Hominini branch if the human-chimpanzee ancestral sequence differs from the human and Denisovan sequences, with human and Denisovan having the same genotype. Finally, a mutation is inferred in the human (respectively Denisovan) branch if the human-chimpanzee ancestral sequence differs from the human (respectively Denisovan) but not from the Denisovan (respectively human) sequence. If a site is different in the 3 species, it is excluded from the analysis (this concerns only 2 sites excluding 2 distinct HM motifs in the F1 dataset). As the Denisovan genome is diploid, we randomly selected one genotype at heterozygous sites. We repeated all motif loss counts 100 times and provided averaged motif loss rates in Figure 1, Table S1 and S2.

## DAF of mutations affecting motifs in the human branch

To obtain the Derived Allele Frequency (DAF) spectrum of mutations affecting motifs in the human branch, we used the F2 subset of ancestral motifs. We inferred, as previously described, motifs that were present in the human-Denisovan ancestor. For each informative site of those 4,236 HM and 4,218 CM motifs, we computed the DAF using the allele frequency (AF) of the polymorphic sites found in 1000 genomes data if any [31]. If there is no SNP at a particular position (most of them), no DAF is computed except if the reference genome (hg19) is different from the intact motif, in this case this change is considered as fixed (DAF = 1).

## Detection of the THE1 transposable elements

We used the hg19 RepeatMasker 3.3.0 (with repeat library 20120124) list of repeats in the human genome [48] from which we extracted all intervals corresponding to "THE1A", "THE1A-int", "THE1B" or "THE1B-int" LTR elements.

## Estimation of dBGC intensity

We fitted a population genetic model to the derived allele frequency (DAF) spectra of CM and HM to estimate the intensity of gene conversion against *PRDM9* motifs, $G = 4N_e g$, using a maximum likelihood framework. We used the model of Eyre-Walker et al. [34] except that we fitted constant positive selection (as gene conversion is equivalent to selection, see [32]) instead of a distribution of deleterious effects. CM sites (resp. HM) play the role of synonymous (resp. non-synonymous) sites. The probability of observing $k_i$ SNPs having $i$ derived alleles out of $n$ follows a Poisson distribution, $P(\mu,k_i)$, with mean:

$$\mu_{CM}(i) = \frac{4N_e u L_{CM} r_i}{i} \quad (1a)$$

and

$$\mu_{HM}(i) = 2N_e u L_{HM} r_i \int_0^1 C_n^i x^i (1-x)^{n-i} H(x) dx \quad (1b)$$

where $H(x)$ is the time a converted allele spends between frequency $x$ and $x + dx$. $N_e$ is the effective population size, $u$ the mutation rate, $L_{CM}$ and $L_{HM}$ the number of CM and HM motifs, respectively. The $r_i$ have been introduced by Eyre-Walker et al. [34] to take demography and/or population structure (and sampling) into account. There is one $r_i$ for each SNP class, corresponding to the deviation from the standard equilibrium model relative to the singleton class for which $r_1$ is set to one (for the discussion of the robustness of this kind of model see [34,49]). The first term within the integral corresponds to the binomial sampling of $i$ alleles over $n$ given their frequency $x$. Because $n$ is very large in the 1000 genomes dataset ($n$ = 2,184), we used the continuous approximation that gives very similar results and facilitates numerical computations:

$$\int_0^1 C_n^i x^i (1-x)^{n-i} H(x) dx \approx \frac{1}{n} H(i/n) \quad (2)$$

We used the two following nested models:
M0: no conversion

$$H(x) = \frac{2}{x} \qquad (3)$$

M1: constant gene conversion of intensity $G = 4N_e g$:

$$H(x) = 2\frac{1-e^{-G(1-x)}}{x(1-x)(1-e^{-G})} \qquad (4)$$

Because the number of SNPs is much lower than the number of chromosomes sampled, we grouped the SNPs by categories of frequencies. The expectations of these groups of SNPs simply becomes:

$$\mu(i_1, i_2) = \sum_{i_1}^{i_2} \mu(i) \qquad (5)$$

Assuming independence between motifs, the likelihood of the model can thus be written down as:

$$\Gamma = \prod_{c=1}^{n_{cat}} P(\mu_{CM}(i_c, j_c), \sum_{z=i_c}^{j_c} k_z^{CM}) P(\mu_{HM}(i_c, j_c), \sum_{z=i_c}^{j_c} k_z^{HM}) \quad (6)$$

Parameters estimates were obtained by maximization of the log-likelihood function. The significance of the model with gene conversion is tested by a LRT with 1 degree of freedom. The goodness of fit (*Gof*) of model M1 is assessed by comparing its likelihood with the saturated model for which all $\mu$ are free. Confidence intervals on $G$ are computed by fixing all other parameters at optimum and searching for $G$ such that the log-likelihood is two points lower than the maximum likelihood (*lnLmax*).

We tested several values of $m$ to assess the robustness of estimations. To do so, we used the following categories of DAF: $f < 0.01$ plus $m$ other categories defined as $0.01 \le f < \frac{1}{m}$ and $\frac{1}{m} \le f < \frac{i+1}{m}$ for $1 \le i \le m - 1$.

Assuming a constant dBGC intensity does not allow to capture the possible very high $G$ values in highly recombining regions. To get a better determination of $G$, we assumed that $G$ is proportional to recombination rates, $G = c\,X$, where $X$ is the crossover rate (in cM/Mb) in a 2-kb window centered on the motif. We used the observed distribution of $X$ in HM motifs. We first fit a gamma distribution to the observed distribution of $X$, which gave a mean of $M_X = 5.02$ and a shape of $\beta = 0.28$. Then we fitted the following model with:

$$H(x) = \int_0^{\infty} 2\frac{1-e^{-G(1-x)}}{x(1-x)(1-e^{-G})}\Phi(G)\,dG \qquad (7)$$

where $\Phi(G)$ is a gamma distribution with mean $M_G = c\,M_X$, and shape $\beta$, fixed to $\beta = 0.28$. In this model, $c$ is optimized.

**Equilibrium GC content (GC*) estimates**

To compute GC* we used the filtered alignment FA corresponding to F2 filter (see above). As GC* estimates are strongly biased by CpG hypermutable sites [50], we discarded all potential CpG sites in the alignment by excluding all G (respectively C) sites for which the previous (respectively following) site is a C (respectively G) in at least one of the four species. We inferred separately AT to GC and GC to AT mutations in each branch of our phylogeny as described above for the count of motif losses. GC* is then computed as follow:

$$GC^* = \frac{\#AT \to GC / \#AT}{\#AT \to GC / \#AT + \#GC \to AT / \#GC} \qquad (8)$$

#X->Y is the number of mutations from X to Y in the branch and #X is the number of X bases in the ancestral node of the branch. "AT" refers either to an "A" or a "T" and "GC" refers either to a "G" or a "C".

**Statistics**

We used normal approximate Z-test with continuity correction to compare motif loss rates. This test is referred as "proportion test" in the text. To compare mean DAFs we used the Wilcoxon test, as DAFs are not distributed normally (Figure 3). All tests and regression computations were made using R software (2.15.0) [51].

Acknowledgments

Authors would like to thank Sylvain Mousset for fruitful discussions on their work and Kasper Munch for providing alignment data.

References

1. Coop G, Przeworski M (2007) An evolutionary view of human recombination. Nat Rev Genet 8: 23–34. doi:10.1038/nrg1947.

2. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861. doi:10.1038/nature06258.

3. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324. doi:10.1126/science.1117196.

4. Smagulova F, Gregoretti I V, Brick K, Khil P, Camerini-Otero RD, et al. (2011) Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature 472: 375–378. doi:10.1038/nature09869.

5. Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K (2012) Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Res 22: 51–63. doi:10.1101/gr.124123.111.

6. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, et al. (2013) Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genet 9: e1003984. doi:10.1371/journal.pgen.1003984.

7. Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, et al. (2013) Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat Genet 45: 1327–1336. doi:10.1038/ng.2766.

8. Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, et al. (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. Cell 144: 719–731. doi:10.1016/j.cell.2011.02.009.

9. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840. doi:10.1126/science.1183439.

10. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876–879. doi:10.1126/science.1182363.

11. Berg IL, Neumann R, Lam K-WG, Sarbajna S, Odenthal-Hesse L, et al. (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42: 859–863. doi:10.1038/ng.658.

12. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova G V. (2012) Genetic recombination is directed away from functional genomic elements in mice. Nature 485: 642–645. doi:10.1038/nature11089.

13. Hayashi K, Yoshida K, Matsui Y (2005) A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature 438: 374–378. doi:10.1038/nature04112.

14. Grey C, Barthès P, Chauveau-Le Friec G, Langa F, Baudat F, et al. (2011) Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. PLoS Biol 9: e1001176. doi:10.1371/journal.pbio.1001176.

15. Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, et al. (2011) Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc Natl Acad Sci U S A 108: 12378–12383. doi:10.1073/pnas.1109531108.

16. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, et al. (2011) The landscape of recombination in African Americans. Nature 476: 170–175. doi:10.1038/nature10336.

17. Myers S, Freeman C, Auton A, Donnelly P, McVean G (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet 40: 1124–1129. doi:10.1038/ng.213.

18. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, et al. (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37: 429–434. doi:10.1038/ng1529.

19. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, et al. (2012) A fine-scale chimpanzee genetic map from population sequencing. Science 336: 193–198. doi:10.1126/science.1216872.

20. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al. (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111. doi:10.1126/science.1105322.

21. Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, et al. (2009) Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet 5: e1000753. doi:10.1371/journal.pgen.1000753.

22. Ponting CP (2011) What are the genomic drivers of the rapid evolution of PRDM9? Trends Genet 27: 165–171. doi:10.1016/j.tig.2011.02.001.

23. Boulton A, Myers RS, Redfield RJ (1997) The hotspot conversion paradox and the evolution of meiotic recombination. Proc Natl Acad Sci U S A 94: 8058–8063.

24. Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat Genet 31: 267–271. doi:10.1038/ng910.

25. Coop G, Myers SR (2007) Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet 3: e35. doi:10.1371/journal.pgen.0030035.

26. Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, et al. (2004) Recombination rate and reproductive success in humans. Nat Genet 36: 1203–1206. doi:10.1038/ng1445.

27. Ubeda F, Wilkins JF (2011) The Red Queen theory of recombination hotspots. J Evol Biol 24: 541–553. doi:10.1111/j.1420-9101.2010.02187.x.

28. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226. doi:10.1126/science.1224344.

29. Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, et al. (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci U S A 109: 15716–15721. doi:10.1073/pnas.1211740109.

30. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi:10.1038/nature09534.

31. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi:10.1038/nature11632.

32. Nagylaki T (1983) Evolution of a finite population under gene conversion. Proc Natl Acad Sci U S A 80: 6278–6281.

33. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, et al. (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science (80- ) in press: 1–10. doi:10.1126/science.1224344.

34. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900. doi:10.1534/genetics.106.057570.

35. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467: 1099–1103. doi:10.1038/nature09525.

36. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10: 285–311. doi:10.1146/annurev-genom-082908-150001.

37. Lesecque Y, Mouchiroud D, Duret L (2013) GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Mol Biol Evol 30: 1409–1419. doi:10.1093/molbev/mst056.

38. Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 4: 1–19. doi:10.1371/journal.pgen.1000071.

39. Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol 21: 984–990. doi:10.1093/molbev/msh070.

40. Munch K, Mailund T, Dutheil JY, Schierup MH (2014) A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. Genome Res 24: 467–474. doi:10.1101/gr.158469.113.

41. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475. doi:10.1038/nature11396.

42. Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, et al. (2011) Variation in human recombination rates and its genetic determinants. PLoS One 6: e20321. doi:10.1371/journal.pone.0020321.

43. Billings T, Parvanov ED, Baker CL, Walker M, Paigen K, et al. (2013) DNA binding specificities of the long zinc-finger recombination protein PRDM9. Genome Biol 14: R35. doi:10.1186/gb-2013-14-4-r35.

44. Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. Genetics 14: 24–32.

45. Paten B, Herrero J, Fitzgerald S (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 4: 1829–1843. doi:10.1101/gr.076521.108.

46. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39: D876–82. doi:10.1093/nar/gkq963.

47. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169–175. doi:10.1038/nature10842.

48. Smit A, Hubly R, Green P (2010) RepeatMasker Open-3.0. http://www.repeatmasker.org.

49. Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S (2011) GC-biased gene conversion and selection affect GC content in the Oryza genus (rice). Mol Biol Evol 28: 2695–2706. doi:10.1093/molbev/msr104.

50. Duret L (2006) The GC content of primates and rodents genomes is not at equilibrium: a reply to Antezana. J Mol Evol 62: 803–806. doi:10.1007/s00239-005-0228-7.

51. R Development Core Team (2012) R: A language and environment for statistical computing. Vienna: R foundation for statistical computing.

# III.C. Informations supplémentaires de l'article

Informations supplémentaires de l'article : *The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes.*

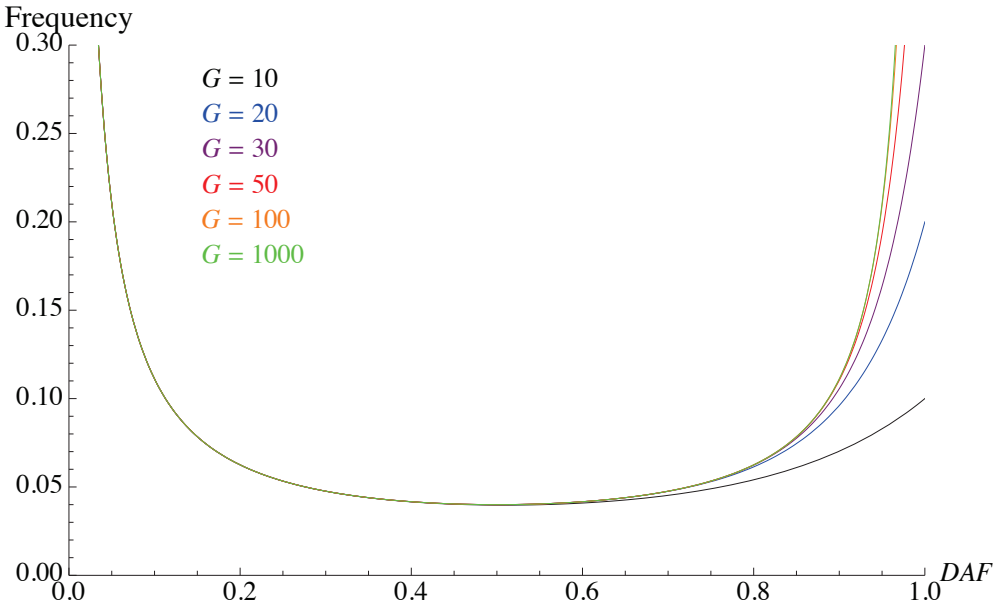



**Figure S1. Expected DAF distribution of mutations affecting HM motifs for different dBGC intensities.**

Derived Allele Frequency (DAF) distribution expected on HM motifs under different dBGC coefficients (G). Equation (4) is plotted for dBGC coefficients ranging from 10 to 1000, as indicated.

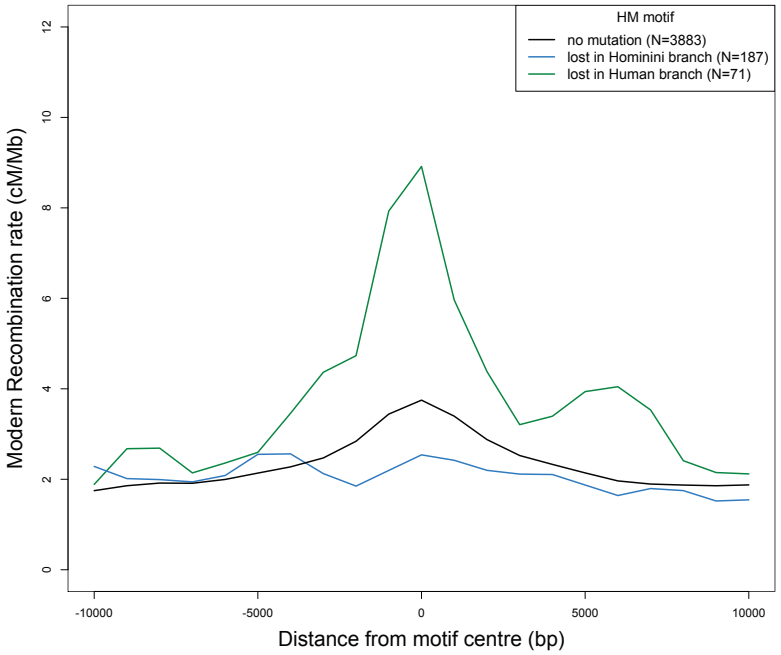



**Figure S2. Present-day human recombination profiles around HM motifs.**

DeCODE recombination rates around HM motifs found in the human-chimpanzee reconstructed ancestral sequence (Filter F2) and conserved in the human genome (black) or lost in the Hominini (blue) or human (green) branch. Recombination rates are averaged on 2kb overlapping windows (overlap = 1kb).

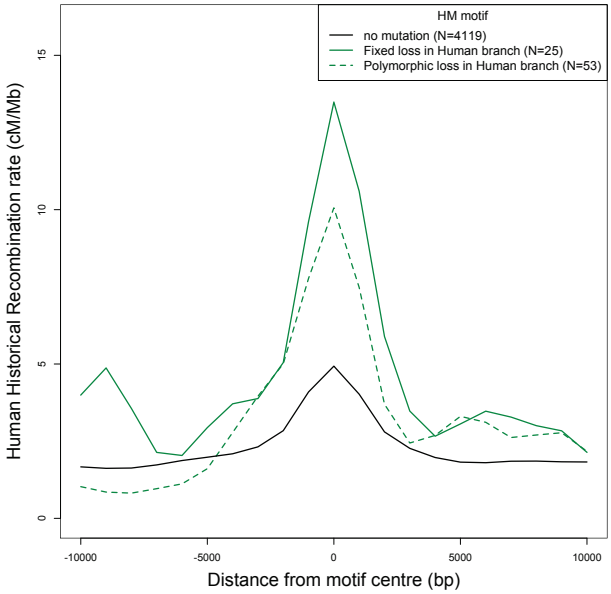



**Figure S3. Historical human recombination profiles around HM motifs differentiating between fixed and non-fixed losses.**

Historical human recombination rates (cM/Mb) around HM motifs found in the human-chimpanzee reconstructed ancestral sequence (Filter F2) and conserved in the human genome (black) or lost in the human branch (green). If the ancestral allele is present in the 1000 genomes data set [31], the motif loss is considered as not-fixed (dotted line). In the opposite case it is considered as fixed (solid line). Recombination rates are averaged on 2kb overlapping windows (overlap = 1kb).

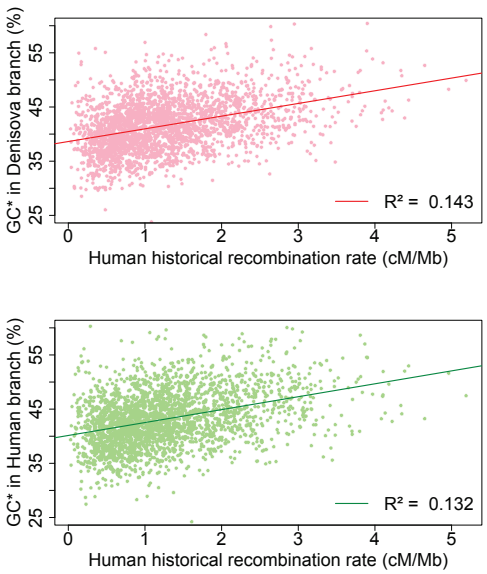



**Figure S4. Genome-wide correlations between equilibrium GC content and recombination rate.**

Each dot represents historical recombination rate (cM/Mb) and equilibrium GC-content (GC*) estimated on the Denisovan branch (red) and human branch (green) over a 1 Mb genomic window.

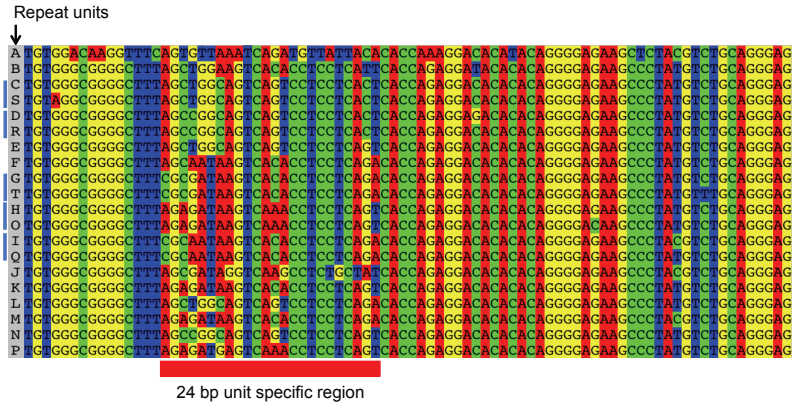



**Figure S5. PRDM9 repeat unit sequences.**

Zinc finger coding repeat sequences are extracted from [11]. The red horizontal box indicates the 24 bp region used to characterize units in Denisovan sequence data. This region is unique for 10 units out of 20. Blue vertical lines on the left show the 5 pairs of units for which the 24 bp region is identical.

**Table S1. Motifs loss rates computed on F3 motif subset.**

| Branch | N[a] | | Rate[b] | | p[c] |
|---|---|---|---|---|---|
| | HM | CM | HM | CM | |
| Chimpanzee | 2019 | 2274 | 5.2% | 5.1% | 0.938 |
| Hominini | 2019 | 2274 | 5.1% | 4.3% | 0.071 |
| Denisovan | 1908 | 2177 | 0.8% | 0.4% | 0.078 |
| Human | 1908 | 2177 | 1.9% | 0.4% | $5.6\ 10^{-6}$ |

[a] Intact motif count at ancestral node of the branch (cf. Figure 1)
[b] Motif loss rate along the branch
[c] P-value of proportion test comparing HM *vs.* CM loss rates along the branch

**Table S2. Motifs loss rates computed on F1 motif subset.**

| Branch | N[a] | | Rate[b] | | p[c] |
|---|---|---|---|---|---|
| | HM | CM | HM | CM | |
| Chimpanzee | 5474 | 5314 | 6.4% | 5.8% | 0.209 |
| Hominini | 5474 | 5314 | 5.3% | 4.3% | 0.024 |
| Denisovan | 5185 | 5084 | 1.1% | 0.7% | 0.036 |
| Human | 5185 | 5084 | 1.8% | 0.5% | $5.7\ 10^{-10}$ |

[a] Intact motif count at ancestral node of the branch (cf. Figure 1)
[b] Motif loss rate along the branch
[c] P-value of proportion test comparing HM *vs.* CM loss rates along the branch

**Table S3. Estimates of the BGC intensity *G* on HM motifs in the human branch.**

| Ncat[a] | G[b] | lnL0[c] | lnL1[c] | lnLmax[d] | p-value[e] | Gof[f] |
|---|---|---|---|---|---|---|
| 6 | 9.30 | -43.20 | -28.82 | -27.91 | 8.23E-08 | 0.87 |
| 7 | 7.61 | -46.54 | -34.26 | -31.69 | 7.14E-07 | 0.53 |
| 10 | 9.55 | -60.29 | -46.67 | -41.47 | 1.79E-07 | 0.32 |
| 12 | 8.55 | -67.37 | -54.17 | -47.35 | 2.75E-07 | 0.25 |
| 15 | 6.93 | -69.64 | -59.81 | -55.25 | 9.27E-06 | 0.82 |
| 22 | 8.55 | -96.06 | -82.57 | -72.01 | 2.06E-07 | 0.45 |
| 42 | 8.26 | -137.38 | -124.23 | -100.97 | 2.91E-07 | 0.26 |

[a] number of DAF categories
[b] Population scaled BGC coefficient ($G = 4N_e g$)
[c] Log-likelihood of neutral model (L0) and BGC model (L1)
[d] see methods
[e] p-value of LRT test comparing M0 and M1
[f] Goodness of fit (see methods)

**Table S4. Estimates of the BGC intensity *G* on HM motifs according to local recombination rate.**

| Rate[a] | CM[b] | HM[b] | G[c] | lnL0[d] | lnL1[d] | lnLmax[e] | p-value[f] | Gof[g] |
|---|---|---|---|---|---|---|---|---|
| 0.098 | 1736 | 1177 | 0.96 | -38.11 | -37.94 | -28.85 | 0.565 | 0.052 |
| 0.565 | 1562 | 1407 | 4.24 | -40.47 | -38.45 | -34.90 | 0.045 | 0.716 |
| 9.412 | 1092 | 1852 | 14.64 | -49.10 | -40.66 | -34.97 | 3.96E-05 | 0.330 |

[a] Mean human historical recombination rate (cM/Mb) over 2 kb window around motifs
[b] Number of CM and HM motifs used in this recombination category
[c] Population scaled BGC coefficient ($G = 4N_e g$)
[d] Log-likelihood of neutral model (L0) and BGC model (L1)
[e] see methods
[f] p-value of LRT test comparing M0 and M1
[g] Goodness of fit (see methods)

**Table S5. Number of reads matching 24 bp PRDM9 Zn-finger unit specific regions and estimated number of unit copies per genotype in Denisova.**

| Unit region | A | B* | C/S | D/R | E | F | G/T | H/O | I*/Q | J | K | L | M | N | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **R[a]** | 16 | 20 | 62 | 62 | 0 | 42 | 0 | 60 | 20 | 17 | 0 | 0 | 0 | 0 | 0 |
| **N[b]** | 2 | 2\|3 | 7\|8 | 7\|8 | 0 | 5\|6 | 0 | 7\|8 | 2\|3 | 2\|3 | 0 | 0 | 0 | 0 | 0 |

[a] Number of reads matching the 24 bp region
[b] Estimated number of copies of the unit per genotype. "2|3" stands for "2 or 3 copies per genotype".
* reads are from $B^{den}$ and $I^{den}$ units

**Table S6. Number of W to S substitutions in the Hominini and modern human branches used in Text S3.**

| | Focal sites (F) | Background sites (B) |
|---|---|---|
| Number of W sites in the human-chimpanzee ancestor | $F_a$ = 1699374 | $B_a$ = 1726440 |
| Number of W to S substitutions in the Hominini branch | $F_{12}$ = 7097 | $B_{12}$ = 6742 |
| Number of W to S substitutions in the modern human branch | $F_3$ = 207 | $B_3$ = 83 |

**Table S7. Number of HM and CM motif losses in the Hominini and modern human branches used in Text S3.**

| | HM motifs (F) | CM motifs (B) |
|---|---|---|
| Number of motifs in the human-chimpanzee ancestor | $F_a$ = 4440 | $B_a$ = 4392 |
| Number of motif losses in the Hominini branch | $F_{12}$ = 204 | $B_{12}$ = 179 |
| Number of motif losses in the modern human branch | $F_3$ = 25 | $B_3$ = 1 |

**Supplementary Text S1: Estimating the strength of BGC in favor of hotspot-disrupting alleles**

To estimate the intensity of BGC against HM motifs, we analyzed the DAF spectra of CM and HM mutations (see main text). To take into account the variability in local recombination rates, we considered a simple model, where the population-scale BGC parameter ($G = 4N_e g$) at a given HM motif is directly proportional to the local crossover rate at this locus ($G = c\,X$; where $X$ is the crossover rate in a 2-kb window centered on the motif, in cM/Mb). We estimated the value of $c$ by fitting a population genetic model to the DAF spectra of CM and HM mutations. This procedure gives a ML estimate of $c$ = 4.8 (CI: 0.8 - 125000). The upper bound of this confidence interval is extremely high. This is due to the fact that DAF spectra do not allow one to differentiate between strong values of $G$.

To obtain a more realistic upper bound for the estimate of $c$, we used the model developed by [1], which describes BGC in favor of hotspot-disrupting alleles. Let us consider a window of 2-kb centered on a HM motif, with two alleles: A corresponding to the intact HM motif, and B to the mutated motif. In an AB heterozygote, a chromosome with the A allele initiates a DSB within that window with probability $r_A$ and a chromosome with the B allele initiates a DSB in this window with probability $r_B$ (with $r_A > r_B$). When a DSB is initiated then with probability $p$, the allele that initiated the DSB is transmitted. Biologically possible values of $p$ range from $p$=0 (the DSB-carrying allele is systematically converted by the other one) to $p$=1/2 (no transmission bias). According to this model, the population-scale BGC parameter ($G = 4N_e g$, where $N_e$ is the effective population size and $g$ the BGC coefficient) is given by:

$$G = 8N_e(r_A - r_B)(\frac{1}{2} - p) \quad (1)$$

It is important to note that the repair of DSB can lead either to crossover (CO) or non-crossover (NCO) recombination events. Thus, the frequency of DSB formation in the window is distinct from the frequency of crossovers. $G$ can be expressed according to the local crossover rate ($X$, in cM/Mb), with the formula:

$$G = 8N_e\left(\frac{X_A - X_B}{f}\right)\left(\frac{1}{2} - p\right)\frac{2000}{100\times10^6} = 1.6\ 10^{-4}N_e\left(\frac{X_A - X_B}{f}\right)\left(\frac{1}{2} - p\right) \quad (2)$$

where $f$ is the fraction of DSBs that are repaired as crossover events.

In theory, the maximum possible value of G could be obtained for $p$=0 (if all DSBs initiated within the window occurred very close to the HM motif) and $r_B$=0 (if the mutation of the motif totally prevented the formation of DSBs in the window):

$$G_{max} = 8\ 10^{-5}N_e\left(\frac{X_A}{f}\right) \quad (3)$$

Empirical data indicate that mutations of the HM motif generally do not abolish the activity of hotspots (see main text). Thus, in reality $r_B$>0, and hence $G_{max}$ corresponds to an upper bound for the true value of $G$.

The level of polymorphism in human populations corresponds to an effective population size of about 10,000. The total number of DSBs per genome is about 10 times the number of crossover [2]. With these parameters ($f$ = 0.1 and $N_e$ =10,000), this would give:

$$G_{max} = 8\, X_A \qquad (4)$$

Hence, this leads to a more realistic upper bound for estimate of $c$ = 8.

References


1. Coop G, Myers SR (2007) Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet 3: e35. doi:10.1371/journal.pgen.0030035.

2. Baudat F, Imai Y, de Massy B (2013) Meiotic recombination in mammals: localization and regulation. Nat Rev Genet 14: 794–806. doi:10.1038/nrg3573.

**Supplementary Text S2: The Denisovan individual PRDM9 does not correspond to any known human allele**

In humans, 29 different alleles of PRDM9 have been described (named from A to E and L1 to L24) [1]. These alleles differ by the number (8 to 18), the identity and the order of minisatellite repeat units (84 bp long) that encode the Zn-finger DNA-binding domain (named from A to T). It is impossible to assemble properly the sequence of this minisatellite repeat in Denisovan, because ancient DNA is highly fragmented (the median size of sequence reads in the Denisovan data set is 56 bp). However, we can use two types of information from reads mapped at the PRDM9 locus to get insight on the alleles present in the Denisovan individual.

First, reads mapped on the human PRDM9 locus display two synonymous SNPs. Those are respectively located in the B and I repeat units of the reference genome. At those sites (chr5: 23,526,925 and 23,527,717 respectively) all mapped Denisovan reads are mutated (G->A and C->T respectively) meaning that units B and I, as known in present human populations, are absent from the ancient genome. Additionally, both of those mutations do not correspond to any other PRDM9 repeat unit known so far in humans. Hereafter, we call the new corresponding repeat units $B^{den}$ and $I^{den}$ (see supplementary data). Thus, PRDM9 Denisovan individual genotype does not correspond to any human PRDM9 genotype described so far. However, as those changes are synonymous, $B^{den}$ and $I^{den}$ are likely to be functionally identical to units B and I respectively. Thus this analysis does not provide information on the PRDM9 target motif recognized by the corresponding alleles (i.e. the PRDM9 phenotype).

Secondly, the analysis of sequence coverage allowed us to estimate the copy number of each of the different units known to constitute the Zn-finger domain of the Denisovan individual. To characterize repeat units, we used only the 24 bp variable region of those units (positions 16 to 39; Figure S5). For 10 characterised units out of 20 (A, B, E, F, J, K, L, M, N and P) this region is unique among all described units. For the 10 others, the sequence of this region is shared by one and only one other unit among the 20: C/S, D/R, G/T, H/O and I/Q (Figure S5). NB: $B^{den}$ and $I^{den}$ units have the same 24 bp region than units B and I/Q respectively. Then we extracted a wide set of mapped Denisovan reads potentially resulting from the sequencing of the PRDM9 Zn-finger locus. This includes the PRDM9 Zn-finger region ±100 bp (chr5: 23,526,242 - 23,528,806; 1,344 reads) and the homologous region of the PRDM7 locus (chr16: 90,123,447 - 90,125,042; 869 reads), which is a close paralog to PRDM9. For each of the 15 specific 24 bp regions (A, B, E, F, J, K, L, M, N, P, C/S, D/R, G/T, H/O and I/Q), we counted the number of reads containing a segment with 100% identity with the region (Table S5). Noting that all PRDM9 alleles known so far have one and only one A unit, we can consider that the number of reads matching the 24 bp region of unit A reflects the coverage of a unit represented once by allele (twice per diploid genotype). Assuming that the number of copies is proportional to the corresponding coverage of the 24 bp specific region of this unit, we estimated copy number of each repeat unit per genotype (Table S5). The pattern of estimated copy number found in Denisovan is not compatible with any combination (homozygous and heterozygous) of alleles known so far in humans.

Reference

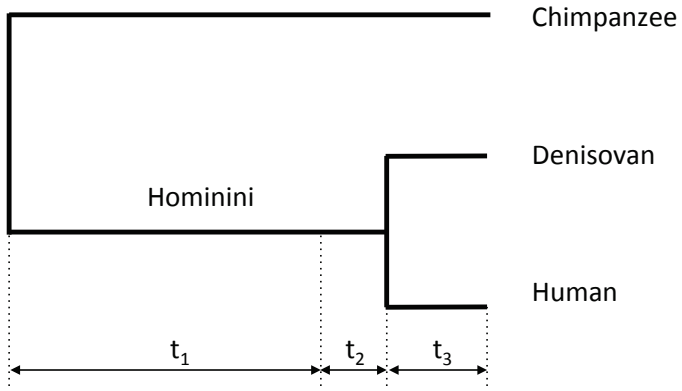
1. Berg IL, Neumann R, Lam K (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature 42: 859–863. doi:10.1038/ng.658.PRDM9.

**Supplementary Text S3: Estimation of the onset of human historical hotspots activity along the Hominini branch.**

## 1) Principle of the approach

To date when human historical hotspots started to be active, we searched for two signatures of recombination activity: BGC against PRDM9 target motifs, and BGC in favor of GC-alleles (gBGC). The first type of BGC can be detected by comparing the loss rate of HM motifs to the loss rate of CM motifs (see main text). The process of gBGC can be detected by analyzing substitution patterns, to infer the equilibrium GC-content (denoted GC*). Along the human branch, as expected, there is a strong signal for these two signatures (see main text). Interestingly, along the Hominini branch, we observed a small bump of GC* in the center of loci corresponding to human historical recombination hotspots (Figure 6B). This indicates that these historical hotspots (or at least a fraction of them) were already active before the population split between Denisovans and modern humans. Moreover, we also observed a slight excess of HM losses over CM losses along the Hominini branch. This suggests that HM started to be a target of PRDM9 before this split.

To try to date the onset of activity of human historical recombination hotspots, we considered a simple model, assuming that all human recombination hotspots started to be active at the same date and that their intensity has been constant since then. Thus, according to this model, loci corresponding to human historical recombination hotspots have been subject to constant BGC during the time period $t_2$+$t_3$, but were not affected by BGC during the time period $t_1$:




Our goal here is to estimate $t_2$. Given the uncertainties about the human-chimpanzee divergence time [1], we will express the date of the onset of recombination hotspots as a fraction ($f$) of the Hominini branch:

$f = \frac{t_2}{t_1+t_2}$ : fraction of the Hominini branch during which hotspots have been active.

Let us consider two categories of sites: focal sites (F), subject to BGC during the time period $t_2$+$t_3$, and background sites (B) not affected by BGC. We will use the following notation:

$u$ : substitution rate (per site per unit of time) in absence of BGC

$v$ : substitution rate (per site per unit of time) at sites affected by BGC

$F_a$ : number of focal sites at the root of the tree

$F_{12}$ : total number of substitutions at focal sites during time $t_1$+$t_2$

$F_3$ : total number of substitutions in the human branch at focal sites during time $t_3$

$B_a$ : number of background sites at the root of the tree

$B_{12}$ : total number of substitutions at background sites during time $t_1$+$t_2$

$B_3$ : total number of substitutions in the human branch at background sites during time $t_3$

Given the very short evolutionary distances considered here, the number of substitutions along the Hominini branch can be approximated by:

$$B_{12} = B_a\, u\, (t_1 + t_2) \quad (1)$$

$$F_{12} = F_a\, (u\, t_1 + v\, t_2) \quad (2)$$

From this, one can obtain:

$$\frac{F_{12}}{B_{12}} = \frac{F_a}{B_a}\left(1 + f\left(\frac{v}{u} - 1\right)\right) \quad (3)$$

and hence:

$$f = \left(\frac{F_{12}B_a}{B_{12}F_a} - 1\right)\left(\frac{v}{u} - 1\right)^{-1} \quad (4)$$

The number of substitutions along the human branch can be expressed as:

$$B_3 = (B_a - B_{12})\, u\, t_3 \quad (5)$$

$$F_3 = (F_a - F_{12})\, v\, t_3 \quad (6)$$

From this, one can obtain the ratio $v/u$:

$$\frac{v}{u} = \left(\frac{F_3}{F_a - F_{12}}\right)\left(\frac{B_a - B_{12}}{B_3}\right) \quad (7)$$

And hence, from equation (4):

$$f = \left(\frac{F_{12}B_a}{B_{12}F_a} - 1\right)\frac{(F_a - F_{12})B_3}{(B_a - B_{12})F_3 - (F_a - F_{12})B_3} \quad (8)$$

**2) Dating the onset of gBGC activity at human historical hotspots**

To date the onset of gBGC activity, we computed the number of W (A or T) to S (G or C) substitutions (using the F2 sequence alignment data set) along the different branches of the phylogeny, in a window of 100 bp centered on the middle of each of the 32,981 human historical hotspots (focal sites). As a reference of sites not affected by gBGC (background sites), we

considered two 50 bp-long windows, located respectively 10 kb upstream and downstream of the center of human historical hotspots. To determine whether mutations along the modern human branch were fixed (i.e. substitutions) or not, we used polymorphism data from the 1000 Genomes project (dataset 20101123 intermediate release).

Taking substitution counts from Table S6, and using equation (8) we inferred f = 4.5%.

The human-chimpanzee divergence time is estimated to 7-13 MYR ago (Langergraber et al. 2012). The human-Denisovan population split is estimated to 0.4-0.8 MYR ago (Langergraber et al. 2012). Thus, the onset of human historical hotspot activity is dated to 0.7 MYR ago (if we consider the lower estimates of divergence times) or 1.3 MYR ago (if we consider the upper estimates of divergence times).

**3) Dating the onset of BGC activity on HM motifs**

To date the onset of BGC on PRDM9 motifs, we computed the number of HM motif losses (focal sites) fixed along the different branches of the phylogeny (using the F2 sequence alignment data set). We used CM motifs as a reference of sites not affected by BGC (background sites). To determine whether mutations along the modern human branch were fixed (i.e. substitutions) or not, we used polymorphism data from the 1000 Genomes project (dataset 20101123 intermediate release).

Taking substitution counts from Table S7, and using equation (8) we inferred f = 0.5%.

Thus, this analysis suggests that the HM motif started to be the target of PRDM9 0.44 MYR ago (if we consider the lower estimates of divergence times) or 0.87 MYR ago (if we consider the upper estimates of divergence times).

Given the limited number of observed motif losses (notably in the modern human branch), these values have to be considered as very rough estimates. However, this result is in agreement with the analyses of gBGC signatures, which indicate that human historical hotspots started to be active shortly before the human-Denisovan population split.

Reference

1. Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, et al. (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci U S A 109: 15716–15721. doi:10.1073/pnas.1211740109.

# III.D. Complément : particularité des patrons de substitutions autours des points chauds humains

Au cours de l'étude de la dynamique récente de la recombinaison chez l'homme, nous avons analysé plus en détail les patrons de substitution sur la branche humaine autour des points chauds. Pour cela nous avons utilisé les changements inférés dans la branche humaine aux sites non CpG comme décrit dans la partie "No overlap between human and Denisovan recombination hotspots" de l'article ci-dessus. Nous avons différencié ces changements selon deux critères. Tout d'abord leur nature : SW ou WS, les autres types de changements on été écartés car il ne sont pas affectés par le gBGC. Deuxièmement, nous avons séparé les changements fixés (substitutions) des changements polymorphes grâce aux fréquences alléliques inférées aux sites correspondants par le projet 1000 Génomes [The 1000 Genomes Project Consortium, 2012]. Un changement est considéré comme fixé si la fréquence de l'allèle dérivée est 1, c'est à dire si l'allèle ancestral n'est pas détecté dans les données 1000 Génomes.

Comme attendu sous le modèle de gBGC, on observe une augmentation des taux de changements WS fixés et non fixés au centre des points chauds (Figure III.1, courbes vertes). En effet, à ces positions, la transmission des allèles GC est augmentée ce qui favorise leur montée en fréquence dans la population. Ceci joue à la fois sur le patron de substitution en augmentant la probabilité de fixation des allèles GC et sur le taux de changements non fixés en favorisant localement ces mêmes allèles. On infère donc plus de changements WS (fixés et non fixés) au centre des points chauds que dans le reste du génome. En ce qui concerne les mutations SW (Figure III.1, courbe rouge partie basse), la dépression relative en changements non fixés au voisinage des points chauds est en accord avec le modèle gBGC : les allèles AT étant moins transmis, ils ségrègent à plus faible fréquence dans ces régions que dans le reste du génome. A l'inverse, le profil des substitutions SW semble, au premier abord, inattendu. En effet, on ne détecte pas de baisse dans le taux de changements SW fixés au voisinage des points chauds (Figure III.1, courbe rouge partie haute). Pour tenter d'expliquer cette observation, nous avons fait l'hypothèse suivante : sur la période de temps considérée, l'impact du gBGC n'est pas le même entre WS et SW. En effet, en dehors des points chauds, lorsque l'effet du BGC est plus faible que celui de la dérive, la probabilité de fixation des allèles GC et AT est faible ($1/2N_e$). Comme vu plus haut, au sein d'un point chaud, $B$ atteint généralement des valeurs

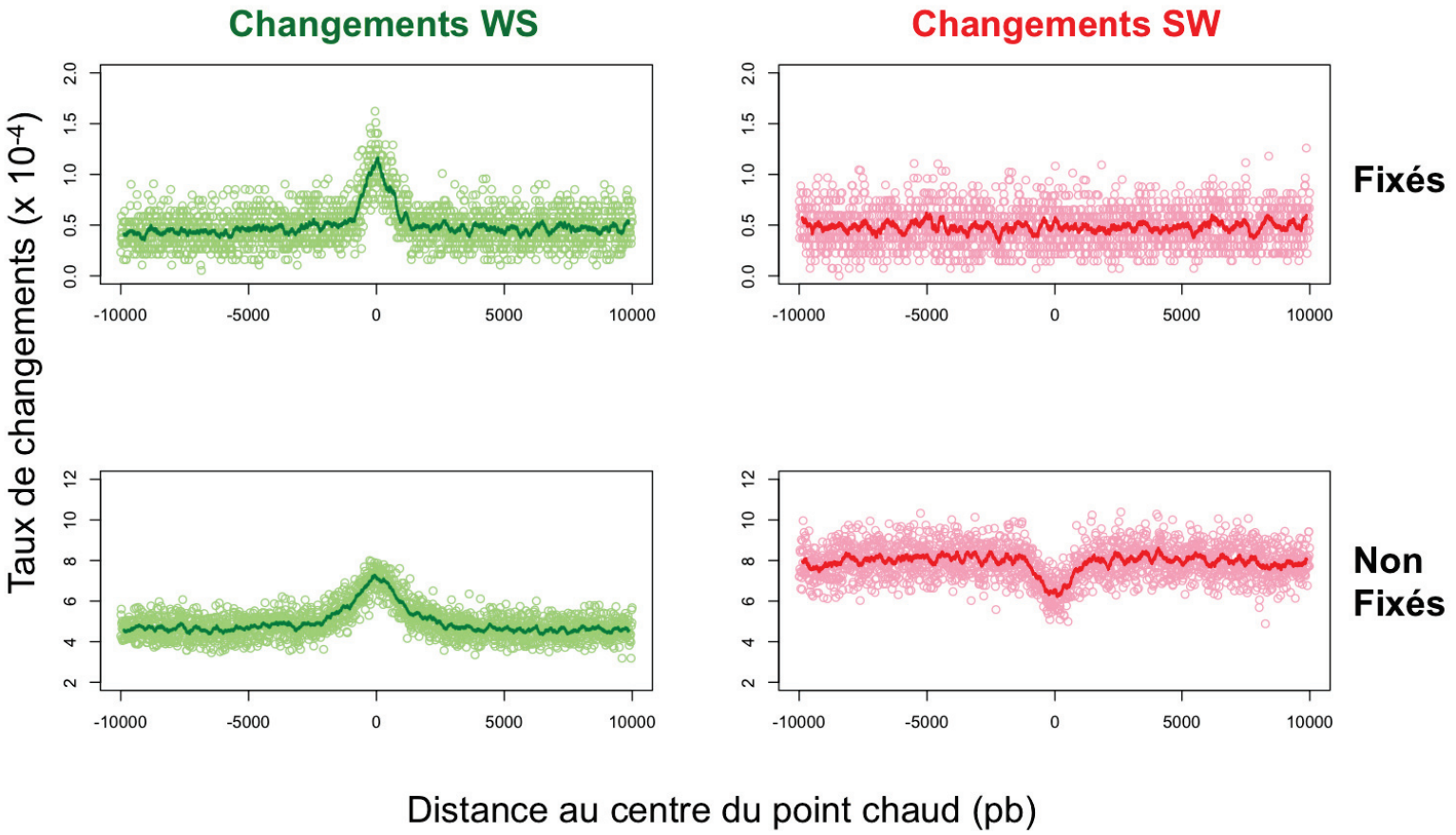



FIGURE III.1 : Taux moyens des changements de AT vers GC (WS) et de GC vers AT (SW) fixés et non fixés autour des 32987 points chauds historiques (HapMap). Chaque cercle représente la moyenne du taux de substitution sur 10 pb. La ligne représente le lissage des taux de substitution sur des fenêtres non chevauchantes de 250 pb.

supérieures à 5 (*cf.* Table S4 dans IS p. 162). Dans ce contexte, le gBGC augmente bien plus la probabilité de fixation des allèles GC qu'elle n'abaisse celle des allèles AT, déjà basse hors des points chauds (Figure III.2). Un épisode de gBGC intense et récent aura donc un effet potentiellement fort et facilement détectable sur les changements WS fixés tout en restant sous le seuil de détection pour les changements SW fixés.

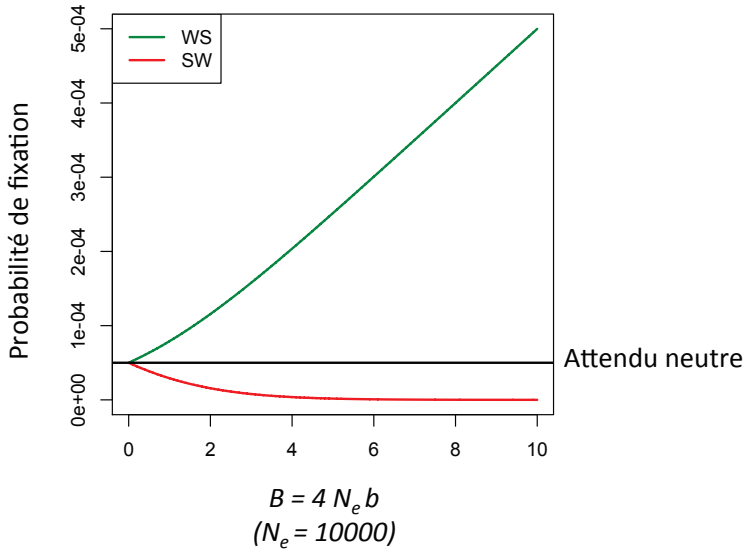



FIGURE III.2 : Probabilité de fixation des mutations AT vers GC (WS, en vert) et GC vers AT (SW, en rouge) soumises au gBGC d'intensité $B$. Ces valeurs sont calculées pour une population de taille efficace $N_e = 10000$ (proche de celle estimée chez l'homme) grâce aux équations de diffusion proposées dans [Nagylaki, 1983] et rapportées dans [Duret & Galtier, 2009a]. L'attendu neutre correspond à la probabilité de fixation d'un allèle uniquement soumis à la dérive : $1/2N_e$. Remarquez l'asymétrie dans l'influence du gBGC sur la fixation des deux types de mutations par rapport à l'attendu neutre pour les valeurs de $B$ supérieures à 3.

Ainsi, si notre interprétation est correcte, ces profils, particulièrement ceux des changements SW, contiennent une information sur l'âge d'apparition des points chauds. Nous avons donc tenté d'estimer plus précisément cet âge à partir de ces données grâce à une méthode bayésienne approchée (ABC pour Approximate Bayesian Computing en anglais). Pour cela, nous avons simulé l'évolution de séquences le long de la lignée humaine depuis la divergence d'avec le chimpanzé. Dans un premier temps, les séquences évoluent de manière neutre, puis à partir d'un temps donné ($t$), elles sont soumises à un

gBGC d'intensité $B$. Le temps total de simulation, $T = 320000$ générations, a été choisi pour correspondre au temps séparant l'homme de son dernier ancêtre commun partagé avec le chimpanzé [Langergraber et al., 2012]. A un temps donné correspondant à la date de divergence homme-denisovien (il y a environ 32000 générations), le modèle simule indépendamment l'évolution sur la branche de l'hominidé de Denisova et sur celle de l'homme (Figure III.3). Ce modèle simule donc l'impact de l'apparition des points chauds humains sur les patrons de mutation et substitution dans cette branche. Pour des raisons pratiques d'optimisation du temps de calcul, nous avons simulé une population de taille efficace réduite d'un facteur 10 par rapport à ce qui est généralement attendu chez l'homme ($N_e = 10000$). L'ensemble des paramètres d'entrée, comme les taux de mutation, a été remis à l'échelle en conséquence. N.B. : Ce modèle a été développé et exploité grâce à la précieuse collaboration de Nicolas Lartillot.

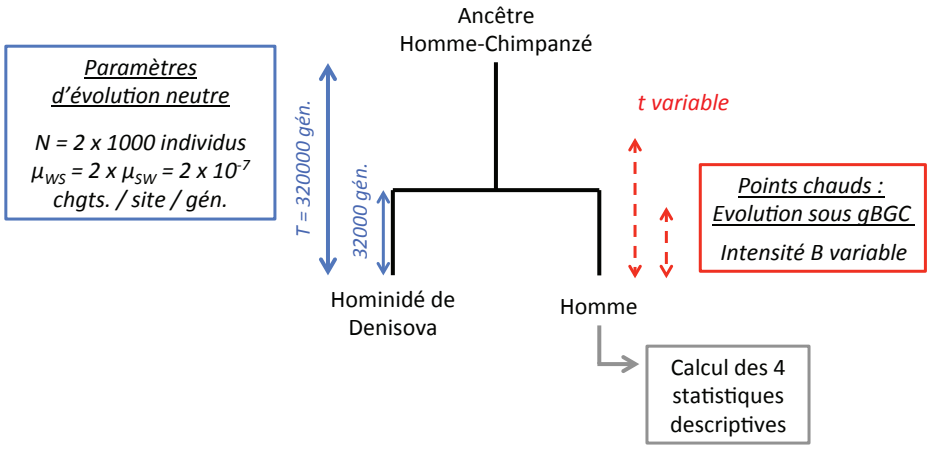



FIGURE III.3 : Modèle utilisé dans l'approche ABC d'estimation de l'âge des points chauds humains. Ce modèle simule l'évolution de séquences sur $T$ générations dans une population de taille constante $N$ et soumise à des taux de mutation WS et SW $\mu_{WS}$ et $\mu_{SW}$. A $t$ générations, les points chauds s'allument et les sites évoluent sous un gBGC d'intensité $B$. A la fin de la simulation, 4 rapports de nombre de changements sont calculés afin d'être confrontés aux données observées, comme décrit dans le texte. N. B. : "chgts." = changements, "gén." = générations.

Dans une approche de type ABC, un grand nombre de simulations sont menées sous différents paramètres d'entrée (ici $B$ et $t$). A la suite de ces simulations, des statistiques descriptives sont calculées afin de pouvoir être confrontées aux observations. Les estimations retenues seront celles qui four-

nissent les statistiques les plus proches de ces observations. Dans le cadre de notre étude, les statistiques choisies sont les rapports du nombre de changements inférés sous le modèle décrit ci-dessus sur le nombre de changements sous un modèle sans points chauds ($B = 0$ et $t = T$) pour chacune des 4 catégories de changements : WS / SW, fixés / non-fixés. Sous le modèle, un changement est inféré comme fixé si il est détecté par parcimonie sur la branche humaine (par comparaison de la séquence homologue ancestrale et denisovienne) et que sa fréquence est de 1 dans un échantillon, de $n = 100$ individus tirés aléatoirement dans la population de taille $N$. Ceci simule notre méthode d'inférence de l'état de fixation grâce aux données des 1000 Génomes. N.B. : Les changements inférés sous le modèle sans points chauds simulent ceux que nous enregistrons hors des points chauds dans nos données.

Sous un premier modèle simple à deux paramètres ($B$ et $t$) simulant l'apparition simultanée des points chauds, nous ne sommes pas parvenus à reproduire les quatre statistiques descriptives observées. Nous avons donc complexifié le modèle pour prendre en compte l'hétérogénéité des points chauds. Pour cela, nous avons considéré l'évolution d'une fraction $\pi$ de sites soumis à un type de points chauds : $B_1$, $t_1$ et d'une fraction $1 - \pi$ de sites soumis à un autre type de points chauds : $B_2$, $t_2$. Avec ce modèle, nous sommes parvenus à reproduire les statistiques descriptives. Ceci montre que le fait que l'on ne détecte pas de gBGC dans le profil des changements SW fixés peut être expliqué par la dynamique particulière de ce processus. Malheureusement, l'estimation des paramètres sous ce modèle ($\pi$, $B_1$, $t_1$, $B_2$ et $t_2$) n'est pas informative car elle recouvre une large gamme de valeurs possibles pour l'âge des points chauds, répartie quasi-uniformément le long de la branche séparant l'homme de son dernier ancêtre commun partagé avec le chimpanzé. Cette approche n'a donc pas permis de préciser l'âge des points chauds. Afin de pouvoir mieux exploiter ces données, il pourrait être intéressant de déterminer des statistiques descriptives plus riches permettant une estimation plus fine des paramètres d'entrée.

## III.E. Bilan de l'étude et perspectives

Cette étude associant génomique des populations humaines et génomique comparative à l'échelle des hominidés, actuels et éteints, a permis de préciser différents aspects du BGC.
Tout d'abord, elle montre que la dynamique des points chauds, et donc celle du BGC est, chez l'homme, encore plus rapide que précédemment estimé par comparaison avec le chimpanzé [Auton et al., 2012]. Cela a été montré *via*

différentes observations. Premièrement, l'âge des points chauds humains estimée à partir des patrons de dBGC (pertes des motifs HM) et gBGC montre que ces structures sont jeunes : au maximum 1,3 Ma. De plus, l'estimation de l'intensité du dBGC sur les motifs HM prédit que leur demi-vie vie est courte : 3 Ma au maximum. La mise en perspective de ces deux dates montre que les points chauds humains actuels ne sont qu'à l'aube de leur existence, ainsi, il est probable que l'empreinte du BGC telle qu'elle est observée actuellement dans le génome continue à s'accentuer durant encore plusieurs dizaines de milliers de générations.
Nos résultats montrent que l'hominidé de Denisova et l'homme ne partagent pas les mêmes points chauds. Le patron de BGC chez ces deux groupes, très proches génétiquement, est donc radicalement différent. Notre hypothèse est que ce changement s'est opéré sur la branche des denisoviens *via* la perte de l'allèle A de *PRDM9* par dérive. Afin d'affiner cette hypothèse, il pourrait être intéressant d'analyser les patrons de dBGC (sur le motif HM) et gBGC chez le groupe frère de l'hominidé de Denisova : Neanderthal. Ceci, a été rendu possible très récemment grâce à la publication du génome d'un néandertalien à haute couverture (52x) : Neanderthal Altaï [Prüfer et al., 2014]. Jusqu'à maintenant nous ne disposions, en effet, que d'un génome à faible couverture pour les néandertaliens (1,3x). L'utilisation des données de Neanderthal Altaï permettrait de déterminer si la perte de l'allèle A a eu lieu récemment ou si elle est partagée avec ces néandertaliens. Il est attendu que cette perte soit assez ancienne car, dans le cas contraire, on enregistrerait un faible signal de gBGC autour des points chauds humains sur la branche Denisova, ce qui n'est pas le cas (Figure 6 de l'article précédent).
Notre étude fournit également la première estimation de la distribution de $B$ à l'échelle du génome entier (motifs HM). Cette distribution correspond au dBGC et montre que son intensité peut atteindre des valeurs très élevées : jusqu'à plusieurs centaines dans certains points chauds (Figure III.4). Jusqu'à présent, notre estimation ne peut être confrontée qu'à une seule étude de sperm-typing pour laquelle la recombinaison totale (CO et NCO) a été étudiée au locus *5A* (chromosome 5 humain) affecté par un fort dBGC [Odenthal-Hesse et al., 2014]. Grâce aux données fournies par cette analyse, nous avons déterminé que la valeur de $B$ à ce point chaud est de 56,6 (voir Tableau III.1). Cette valeur correspond à la médiane de la distribution de $B$ que nous avons estimée autour des motifs HM (57,5) se trouvant dans des points chauds. Notre estimation est donc en accord avec cette mesure "directe" de $B$. N.B. : Le point chaud *5A* n'est pas déterminé par le motif reconnu par l'allèle A de *PRDM9* mais par l'allèle C. Cependant, nous n'avons aucune raison de penser que les distributions de l'intensité du dBGC autour de ces deux motifs soient différentes.

| | **CO** | **NCO** | **Total** |
|---|---|---|---|
| #A transmis | 39 | 6 | 45 |
| #G transmis | 78 | 28 | 106 |
| Total | 117 | 34 | $151^{(1)}$ |
| Fréquence par cellule | 0,54% | - | $^{151}/_{\frac{117}{0,54\%}}{}^{(2)}$ |

TABLEAU III.1 : Calcul de $B$ au locus 7.2 (polymorphisme A/G) du point chaud humain *5A* à partir des données de sperm-typing de [Odenthal-Hesse et al., 2014]. $^{(1)}$ : Nombre de recombinants. $^{(2)}$ : Fréquence de recombinaison totale (CO + NCO) $r$ calculée comme le nombre de recombinants sur le nombre de cellules analysées. Le nombre de cellules analysées est donné par le nombre de CO sur la fréquence des CO par cellule. $B$ est estimé grâce à la formule : $B = 4N_e r(\gamma - 0.5)$ où $\gamma$ est la distortion méiotique totale (CO + NCO) en faveur de l'allèle G : $\gamma = \frac{106}{151}$. Voir Equation 1 p. 44. Ainsi, avec $N_e = 10000$, on obtient $B \approx 56,6$.

Enfin, les quantifications conjointes de l'intensité et de la dynamique du dBGC autour des points chauds humains montre que ce processus est assez fort pour aboutir à une érosion importante et rapide de ces structures à l'échelle du génome entier. Il est donc attendu que ce processus d'auto-destruction des points chauds aboutisse rapidement à une réorganisation des taux de recombinaison à travers le génome. On connait encore mal les modalités de cette réorganisation. Chez la souris $Prdm9^{-/-}$ et chez le chien dont la copie de *Prdm9* est pseudogénisée, les points chauds sont préférentiellement localisés dans les promoteurs des gènes. Cependant, il n'est pas certain que ceci rende compte de ce qui se passe lorsqu'une copie fonctionnelle de PRDM9 ne rencontre plus de motif cible intact le long du génome. Afin de préciser cela, il serait peut-être possible de construire, par transgenèse, un modèle murin possédant un variant de PRDM9 dont le ZnF array reconnaisse un motif fortement sous-représenté dans le génome de la souris. Cette lignée serait un modèle d'étude de la dynamique de la recombinaison (et donc du BGC) dans un génome dont les points chauds sont fin de vie. L'analyse de la carte de DSB et de la fertilité (prédicteur de la fitness) de ce mutant permettraient de déterminer quelles sont les conséquences du processus d'auto-destruction rapide des points chauds qui est au coeur du modèle de Reine Rouge d'évolution de ces structures (*cf.* I.B.3. p. 65).

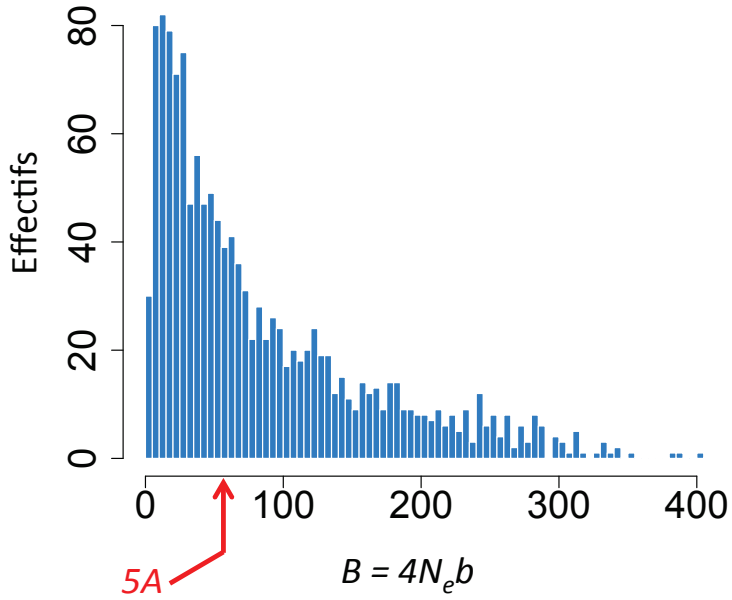



FIGURE III.4 : Distribution des valeurs de $B$ d'intensité du dBGC sur 1358 motifs HM présents dans les points chauds de recombinaison (HapMap). Méthode : voir article (*cf.* III p. 131) et Texte supplémentaire S1 associé (p. 165) La flèche rouge montre la valeur de $B$ au locus $5A$ estimée grâce aux données de sperm-typing tirées de [Odenthal-Hesse et al., 2014] (avec $N_e = 10000$).

Chapitre *IV*

# Discussion

# IV.A. Bilan : caractéristiques de la quatrième force d'évolution des génomes

Dans cette partie, nous ferons le bilan des résultats obtenus dans le cadre de ce travail et montrerons qu'ils contribuent à faire du BGC une force d'évolution moléculaire à part entière.

Tout d'abord, nous avons étudié, chez la levure, la carte de recombinaison haute résolution de [Mancera et al., 2008]. Cette carte montre que les évènements de conversion sont biaisés vers GC à l'échelle du génome entier. Nous avons donc analysé les tracts associés à ces évènements afin de mieux appréhender le déterminisme moléculaire du gBGC. Nous avons montré qu'à un locus pris dans un hétéroduplex lors de la recombinaison, un allèle bordé par des sites G ou C hétérozygotes avait plus de chance de convertir son homologue lorsque celui-ci était bordé par des sites A ou T. Ce biais vers GC, trouvé dans la réparation des hétéroduplexes donnant des tracts de conversion simples, en grande majorité, est compatible avec l'hypothèse d'un gBGC causé par le système MMR. Ce résultat apporte ainsi la confirmation que dBGC et gBGC sont deux processus aux origines moléculaires distinctes. En effet, le dBGC, lui, découle directement du biais dans l'initiation de la recombinaison et non pas dans la réparation des hétéroduplexes. Si l'un des homologues initie plus fréquemment la recombinaison que l'autre, alors il sera plus fréquemment convertit par l'autre. Ainsi, la formation du DSB étant une étape commune à toutes les voies de recombinaison, on s'attend à observer du dBGC associé à la fois aux CO et aux NCO. Ceci n'est pas forcément attendu avec gBGC. En effet, il est possible que le biais de réparation du MMR ne puisse s'exprimer que dans certains contextes comme la formation des CO lors de laquelle la présence d'une cassure sur chaque brin de l'hétéroduplex permet "un choix" dans la conversion ou la restauration, propice à l'instigation d'un biais. Ceci est en accord avec l'observation de l'association spécifique du gBGC avec les CO et non les NCO, que nous avons faite chez la levure. Chez l'homme et la souris, l'absence de carte de conversion haute résolution ne permet pas de déterminer si le gBGC est associé à certains évènements ou d'autres. Cependant, une étude de sperm-typing récente a mis en évidence un BGC spécifiquement associé aux NCO, différent du dBGC, sur deux SNP de deux points chauds chez l'homme [Odenthal-Hesse et al., 2014]. Le fait que le biais soit en faveur des polymorphismes GC face aux polymorphismes AT suggère qu'il s'agit peut-être de gBGC. Dans ce cas l'association spécifique de ce BGC avec les NCO serait en contradiction avec ce qui a été observé chez la levure. Cependant, cette étude ne reposant sur l'analyse

que de deux loci, d'autres résultats de la sorte sont nécessaires pour tester si l'association gBGC-NCO est généralisable à l'échelle du génome entier.
D'autre part, nous avons proposé une origine évolutive au gBGC sur la base de nos résultats concernant ses mécanismes moléculaires. Le MMR étant aussi recruté pour la réparation des mésappariements formés lors de la réplication de l'ADN, en mitose, il est possible que le biais vers GC, observé en méiose (conversion génique), ne soit qu'un sous-produit d'un biais sélectionné pour son action en mitose. Cette hypothèse est compatible avec le fait qu'il existe un biais vers AT dans le processus de mutation, à l'échelle du vivant [Lynch, 2010]. Un biais de réparation vers GC pourrait donc contrebalancer cette force. Cette hypothèse semble difficile à tester *via* une approche comparative car le biais mutationnel vers AT semble partagé par tous les eucaryotes. Nous ne pouvons donc pas vérifier directement si l'absence de ce biais est associée à une absence de gBGC chez un organisme. Cependant, il semble possible de réaliser une expérience permettant de tester le potentiel biais de réparation du MMR sur des constructions nucléiques contenant différents arrangements de cassures simple brin et de mésappariements AT : :GC. Ceci permettrait éventuellement d'appuyer notre modèle de choix de brin déterminé par les mésappariements proches des cassures. D'autre part, comme mentionné plus haut, l'étude de cartes de recombinaison à haute résolution chez des mutants de composants du MMR pourrait permettre de mieux caractériser le biais vers GC imposé par ce système lors de la conversion génique. Malheureusement, cette approche est limitée par le fait que le MMR agit dans plusieurs autres étapes clés du processus de recombinaison dont le rejet de brin et la formation des dHj menant aux CO. Ceci rend particulièrement difficile l'étude précise du rôle de chaque composant du MMR dans l'apparition du gBGC.

Comme vu en introduction, le BGC ne peut avoir un réel impact sur l'évolution que si il est capable de surpasser la sélection naturelle et la dérive génétique à de nombreux loci du génome. Le dBGC et le gBGC influencent l'évolution des génomes avec des modalités différentes.
Le gBGC modifie les patrons de substitution en augmentant la probabilité de fixation des allèles riches en GC. Des résultats récents obtenus par analyse des spectres de DAF WS et SW calculés à partir des données 1000 Génomes [The 1000 Genomes Project Consortium, 2012] montrent que $B$ atteint des valeurs supérieures à 10 dans les 2% du génomes qui recombinent le plus [Sylvain Glémin, communication personnelle]. Ceci prouve que le gBGC est capable de surpasser la dérive et la sélection au niveau des points chauds. De son côté, le dBGC conduit à la disparition des séquences recombinogènes, comme les motifs reconnus par PRDM9, chez l'homme et la souris. L'étude

que nous avons menée sur la dynamique récente des points chauds de recombinaison chez l'homme a permis de quantifier cet aspect, à l'échelle du génome entier (dans les motifs HM, dans et hors des points chauds). Ainsi, dans les points chauds, $B$ peut atteindre plusieurs centaines (Figure III.4), montrant que le dBGC est capable de surpasser la dérive génétique et d'interférer avec la sélection à ces loci.
D'autre part, nous avons montré que l'allèle A de *PRDM9*, allèle majoritaire à l'échelle des populations humaines, déterminait la position des points chauds actuels depuis 1,3 Ma, au maximum. De plus, ces points chauds ne sont pas partagés avec l'hominidé de Denisova dont la séparation avec la lignée humaine a eu lieu il y a moins de 800000 ans. Ceci démontre que la dynamique du BGC est très rapide. Enfin, l'étude du GC* calculé sur la branche terminale humaine (depuis la séparation avec Denisova) au niveau des points chauds montre que, bien que le patron de BGC soit soumis à des changements rapides, il est capable d'impacter profondément l'évolution des régions à fort taux de recombinaison. Le patron de BGC est donc une caractéristique propre du génome humain, c'est à dire, non partagée avec ses plus proches cousins : chimpanzé et denisoviens. Ceci renforce l'idée selon laquelle le BGC peut être un facteur confondant important dans la détection de la sélection naturelle affectant spécifiquement l'homme, à l'échelle moléculaire. Notre travail pointe donc, encore une fois, la nécessité d'adapter les méthodes de détection de la sélection. Une voie de recherche se dessine actuellement dans ce sens avec le développement de modèles d'évolution de codons permettant à la fois une estimation du $^{d_N}/_{d_S}$ et du $B$ associé au gBGC. Ceci permettra de quantifier la contribution du gBGC dans les variations du $^{d_N}/_{d_S}$ et ainsi de mettre en évidence la part imputable à la sélection dans ces variations.
Au final, à un moment donné, le BGC a donc un impact fort sur l'évolution d'un petit nombre de loci (points chauds). Cependant, la dynamique de ce processus est telle qu'à l'échelle des temps évolutifs, une fraction beaucoup plus importante du génome y est soumise. Ainsi, la conjonction des deux facteurs étudiés ici : intensité et dynamique, montre que le BGC est une force importante de l'évolution des génomes.

Grâce à la caractérisation du dBGC agissant sur les motifs HM, nous avons, en outre, apporté de nouveaux résultats soutenant le modèle de Reine Rouge d'évolution des points chauds de recombinaison. Ce modèle, déjà décrit en introduction et dans notre article, est résumé à la Figure IV.1, notre contribution y est indiquée. Il est basé sur des observations faites chez les primates et les rongeurs. A ce jour, nous ne savons pas si il pourrait s'appliquer à d'autres espèces car la dynamique des points chauds associée aux allèles de *PRDM9* n'a été montrée que dans ces deux groupes. Cependant,

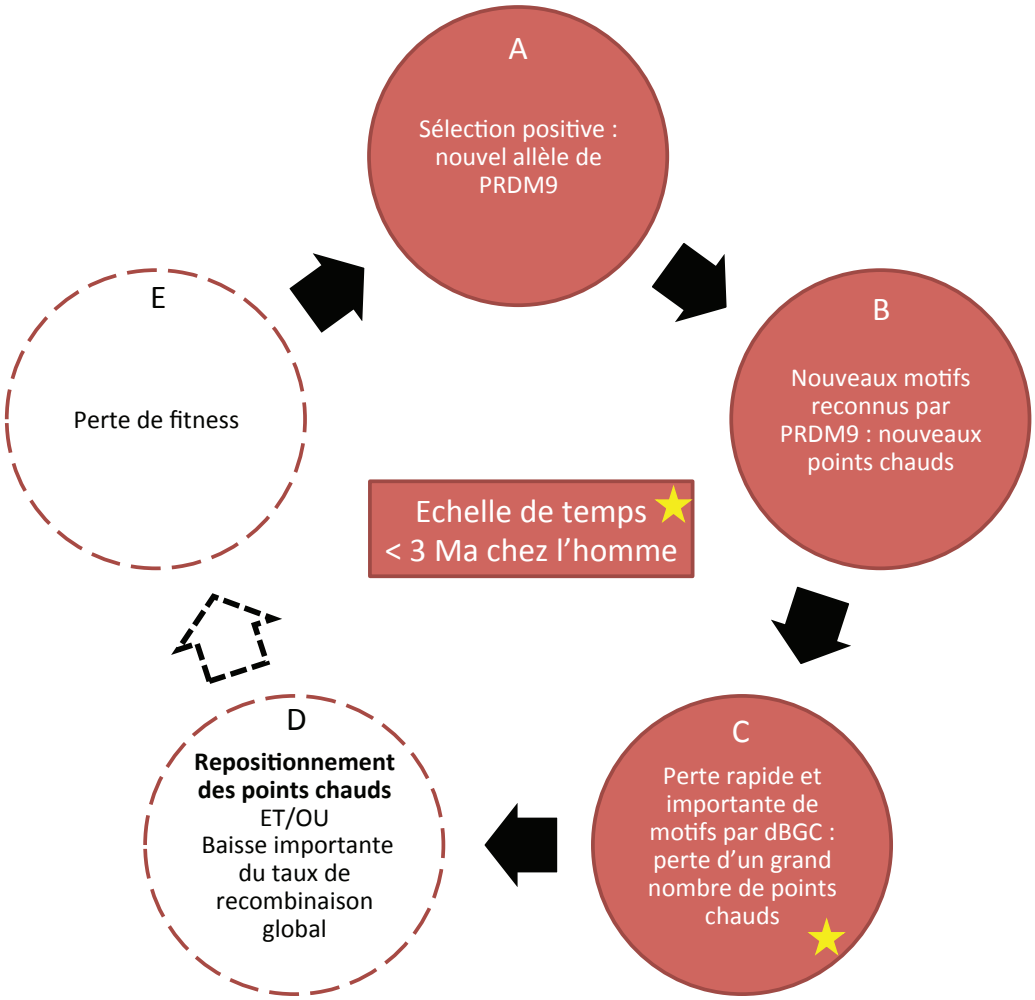



FIGURE IV.1 : Modèle de Reine Rouge d'évolution des points chauds de recombinaison. Les cercles pleins montrent les étapes confirmées par des observations chez l'homme et/ou la souris. Références pour l'étapes A : [Oliver et al., 2009] et B : [Baudat et al., 2010, Myers et al., 2010, Berg et al., 2010, Brick et al., 2012]. Les contributions apportées par nos résultats sont indiquées par une étoile. Les cercles en pointillés indiquent les éléments du modèle pour lesquels aucune preuve n'a encore été avancée. Etape D : nous privilégions l'hypothèse d'un repositionnement des points chauds vers les régions promotrices (en gras) comme origine de la perte de fitness (E) poussant *PRDM9* à se diversifier, comme mentionné en introduction.

chez la levure et le chien, PRDM9 est absente ou non fonctionnelle ce qui est compatible avec l'absence de dynamique de leurs points chauds [Tsai et al., 2010, Axelsson et al., 2012]. Le modèle de Reine Rouge ne s'applique donc pas dans leur cas. De plus, chez ces deux espèces, les points chauds sont localisés préférentiellement dans les promoteurs des gènes [Mancera et al., 2008, Auton et al., 2013], comme chez les souris $PRDM9^{-/-}$ [Brick et al., 2012]. Chez le chien, cette absence de renouvellement est accompagnée par une très forte signature de gBGC : certains centres de points chauds ont un contenu en GC supérieur à 60% [Axelsson et al., 2012]. A l'inverse, chez la levure, le gBGC laisse une trace comparable à celles observées chez l'homme malgré la conservation de l'emplacement génomique des points chauds sur plusieurs millions d'années. Ceci est dû au fait que la levure effectue beaucoup moins de méioses ($\approx 10^{-5}$ par génération) que les mammifères (1 par génération) car elle peut aussi se reproduire par mitose. Ainsi, la signature de gBGC observée dans son génome est atténuée d'un facteur de l'ordre de $10^5$ [Tsai et al., 2010]. Une étape déterminante du modèle de Reine Rouge (Figure IV.1) n'a pas encore été élucidée : quelle est l'origine précise de la perte de fitness qui impose à *PRDM9* de se renouveler permettant ainsi un "déplacement" des points chauds ? De notre côté nous privilégions l'hypothèse selon laquelle lorsque PRDM9 arrive à court de motifs cibles (détruits par dBGC), tout se passe comme si elle était absente du génome et, comme nous l'enseigne le mutant de souris $PRDM9^{-/-}$, ses points chauds sont redirigés vers les promoteurs des gènes [Brick et al., 2012]. Par la suite, il est possible que ce repositionnement impacte négativement la fitness. Ce modèle reste encore spéculatif mais est soutenu par le fait que les souris $PRDM9^{-/-}$ sont stériles. Comme proposé plus haut (*cf.* III.E. p. 176), ce modèle pourra être affiné par l'étude d'autres mutants *PRDM9* de souris. Il sera ainsi possible de déterminer le lien causal, si il existe, entre le repositionnement des points chauds dans les promoteurs et la baisse de fitness des mutants. Cependant, sous ce modèle, restera à expliquer comment cette réduction de fitness est évitée chez le chien.

Au final, nos résultats viennent confirmer l'hypothèse selon laquelle les BGC (dBGC et gBGC), même si ils ont des origines moléculaires différentes, sont des forces évolutives au même titre que la mutation, la sélection et la dérive car ils sont capables de façonner les patrons de substitution et cela à l'échelle du génome grâce à leur dynamique rapide chez la plupart des mammifères.

# IV.B. Perspectives : le BGC au delà de la recombinaison méiotique

Comme vu en introduction, le BGC est un concept biologique encore jeune au sein d'un domaine de recherche lui même récent. Dans cette partie, nous nous éloignons de l'étude des mécanismes, de l'intensité et de la dynamique du BGC pour nous pencher sur les limites qui nuancent ou étendent l'idée que nous nous faisons de l'influence de ce processus sur l'évolution des génomes. Ces limites sont encore incertaines car leur étude est encore peu documentée. Cependant, le développement rapide des nouvelles méthodes de séquençage, comme la technologie nanopore [Schneider & Dekker, 2012], permettant l'analyse à haut débit de longs fragments d'ADN, contrairement aux NGS [Metzker, 2010], ouvre la voie vers une étude plus précise de la recombinaison chez les mammifères dont le génome est plus de 250 fois plus étendu que celui de la levure. Récemment, plusieurs résultats ont été obtenus par séquençage du génome entier d'un spermatozoïde humain [Wang et al., 2012, Kirkness et al., 2013]. Ceci est la première étape vers la construction d'une carte de haute résolution chez un mammifère. Reste maintenant à pouvoir répliquer ce type d'expérience afin d'avoir une puissance statistique permettant d'inférer des biais de conversion de l'ordre de ceux observés sur les points chauds par sperm-typing. Cette tâche semble encore difficile à mettre en œuvre car, par exemple, pour détecter un biais de transmission significatif au point chaud humain *DNA2* ($x = 50,00049\%$, *cf.* I.A.4. p. 43) il faudrait analyser au moins $4\ 10^{10}$ cellules, ce qui semble impossible à réaliser à l'échelle du génome entier, avec les techniques actuelles. De plus, comme nous l'avons vu (*cf.* Introduction et Annexe A p. 219), la majorité des méthodes d'étude de la recombinaison passe par l'analyse d'un grand nombre de marqueurs polymorphes le long du génome. La relative faible densité en SNP des génomes humains, comparé à celle qui peut exister entre deux souches de *Saccharomyces cerevisiae* par exemple (voir par exemple [Mancera et al., 2008] et [Martini et al., 2011]), est donc une limite biologique à l'étude de la recombinaison chez l'homme.

Dans l'ensemble de ce travail, nous nous sommes focalisés sur le BGC associé à la recombinaison méiotique allélique eucaryote. Nous allons maintenant voir que le BGC n'est potentiellement pas limité à ce contexte précis. Tout d'abord, il existe plusieurs indices montrant que la conversion génique ectopique (*i.e.* entre paralogues) pourrait aussi donner lieu aux deux types de BGC. En effet, l'étude du GC3 des gènes présents en grand nombre de copies dans le génome montre qu'ils sont plus riches en GC que les autres. C'est le

cas par exemple des ARNr, des ARNt [Galtier et al., 2001] et des histones [Galtier, 2003]. Or, on sait que ces gènes évoluent de manière concertée par conversion ectopique. Cependant, ces résultats ne suffisent pas à prouver que la recombinaison ectopique provoque du gBGC. En effet, les gènes représentés par plusieurs copies dans le génome ont un $N_e$ plus fort que les autres, il est donc possible que leur fort taux de GC traduise simplement un gBGC allélique plus fort (rappelons que l'intensité du BGC, $B$, est proportionnelle à $N_e$). En outre, une forme de dBGC a pu être mise en évidence à plusieurs loci où certaines copies (de loci répétés) ont plus de chance de convertir leur paralogue que l'inverse. Par exemple, chez l'homme, en ce qui concerne les éléments répétés de la famille HERV, les conversions d'une copie proximale vers une copie distale (par rapport au centromère) sont 20 fois plus fréquentes que l'inverse [Bosch et al., 2004]. Ceci est en accord avec le fait qu'en méiose, comme nous l'avons déjà vu, la recombinaison est plus fréquemment initiée sur les parties télomériques. De plus, il est possible que ce dBGC ectopique soit directement lié aux propriétés fonctionnelles de certains loci comme ceux des $\beta$-globines humaines pour lesquels les copies donneuses ont un taux d'expression significativement moins fort que les copies reçeveuses [Papadakis & Patrinos, 1999].

Deuxièmement, on a récemment mis en évidence un biais dans la conversion allélique mitotique chez le pétoncle (mollusque bivalve) [Wang et al., 2010]. Dans cette étude, les auteurs ont d'abord créé un hybride à partir de deux pétoncles (un mâle et une femelle) de deux espèces différentes ($\approx 20\%$ de divergence). Ils ont ensuite suivi le devenir d'une partie d'un locus codant pour un ARNr par étude du patron de restriction (RFLP) au cours des premières divisions de la cellule œuf de l'hybride. Ce patron est, en effet, différent entre la copie mâle et la copie femelle. Il apparait qu'au cours du développement de l'hybride, la copie mâle disparait au profit de la copie femelle. Les auteurs invoquent le BGC pour expliquer ces résultats. Ils proposent que la copie mâle soit soumise à des DSB fréquents imputables à des enzymes de restriction propres au génome femelle. La réparation de ces DSB par recombinaison homologue pourrait donc aboutir à la conversion fréquente du génotype mâle en génotype femelle. Il s'agirait donc ici d'une forme de dBGC mitotique. D'autre part, la recombinaison homologue mitotique a été observée chez la levure, où elle est réalisée sur de très longs patches, pouvant atteindre la longueur d'un bras chromosomique entier. Ce type de conversion est appelé BIR (Break Induced Repair, en anglais) [Pâques & Haber, 1999]. Cependant, aucun parallèle direct n'a été mis en évidence entre ce mécanisme et le BGC mitotique observé chez le pétoncle. Ainsi, le BGC ne semble limité ni à la conversion allélique, ni à la conversion méiotique, mais semble pouvoir être potentiellement associé à toutes les formes de recombinaison homologue.

Enfin, la recombinaison homologue n'est pas restreinte au domaine des eucaryotes. En effet, les procaryotes sont aussi capable, pour certaines espèces, de parasexualité permise par l'échange de matériel génétique *via* les mécanismes de transduction, transformation et conjugaison. Dans de nombreux cas, l'intégration de séquences exogènes est permise par un système de recombinaison homologue dont de nombreuses protéines sont homologues à celles du système de recombinaison eucaryote [Eisenstark, 1977, Krogh & Symington, 2004]. Des résultats récents montrent que le gBGC ne serait pas limité au domaine des eucaryotes. Lassalle et collaborateurs se sont intéressés aux déterminants évolutifs des disparités de taux de GC à l'échelle des procaryotes [Florent Lassalle, communication personnelle]. Sous les modèles classiques, les variations de GC chez les bactéries (taux de GC génomique moyen allant de 13% à 75% à l'échelle des bactéries) seraient le résultat du biais mutationnel [Sueoka, 1962]. Cependant, ces modèles ont été récemment remis en question par le fait que le GC observé chez les bactéries résulte, en réalité, d'un biais de fixation en faveur des allèles GC. Ceci a été interprété comme une signature de sélection sur la composition en bases [Hershberg & Petrov, 2010, Hildebrand et al., 2010]. Cependant, l'origine de cette pression de sélection reste un mystère. Dans ce contexte, Lassalle et collaborateurs ont analysé les patrons de recombinaison dans les génomes bactériens. Ils ont montré que pour 11 espèces sur 14 étudiées, le GC3 des gènes qui recombinent est significativement plus haut que celui des gènes qui ne recombinent pas. Par ailleurs, ces variations de GC3 ne peuvent pas s'expliquer par une sélection sur l'usage des codons. Ces résultats semblent donc compatibles avec le modèle du gBGC. Ils montrent ainsi que le gBGC pourrait affecter l'évolution des bactéries et pourrait donc être commun à l'ensemble des organismes cellulaires capable de recombinaison homologue. De nôtre côté, nous avons montré que le gBGC était probablement associé à l'activité du système MMR chez les eucaryotes. Le fait que plusieurs protéines de ce système soient conservées et actives lors de la recombinaison chez les procaryotes montre donc qu'il est vraisemblable que les observations faites par Lassalle et collaborateurs soient dues au gBGC.

Notre étude montre que les courants dynamiques et impétueux du BGC peuvent mener les allèles vers des rivages inattendus au même titre que la sélection et la dérive, mettant ainsi en lumière la quatrième force d'évolution moléculaire. Beaucoup de travail reste encore à réaliser afin de mieux analyser et comprendre à la fois l'origine évolutive de cette force et son étendue, à l'échelle des génomes et à l'échelle du vivant. Cependant, la démonstration de l'importance de ce processus impose qu'à l'avenir, il soit systématiquement pris en compte dans les études de génomique comparative.

# Bibliographie


[Alani et al., 1994] Alani, E., Reenan, R. and Kolodner, R. (1994). Interaction Between Mismatch Repair and Genetic Recombination in Saccharomyces cerevisiae. Genetics *137*, 19–39.

[Argueso et al., 2004] Argueso, J. L., Wanat, J., Gemici, Z. and Alani, E. (2004). Competing crossover pathways act during meiosis in Saccharomyces cerevisiae. Genetics *168*, 1805–16.

[Auton et al., 2012] Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. a., Bontrop, R., Myers, S., Donnelly, P., Przeworski, M. and McVean, G. (2012). A fine-scale chimpanzee genetic map from population sequencing. Science (New York, N.Y.) *336*, 193–8.

[Auton & McVean, 2007] Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome research *17*, 1219–27.

[Auton et al., 2013] Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., Hayward, J. J., Cohen, P. E., Greally, J. M., Wang, J., Bustamante, C. D. and Boyko, A. R. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS genetics *9*, e1003984.

[Awadalla, 2003] Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. Nature reviews. Genetics *4*, 50–60.

[Axelsson et al., 2012] Axelsson, E., Webster, M. T., Ratnakumar, A., Ponting, C. P. and Lindblad-Toh, K. (2012). Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome research *22*, 51–63.

[Baker et al., 2014] Baker, C. L., Walker, M., Kajita, S., Petkov, P. M. and Paigen, K. (2014). PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome research *24*, 724–32.

[Bateson et al., 1905] Bateson, W., Saunders, E. and Punnett, R. (1905). Experimental studies in the physiology of heredity. Reports to the Evolution Committee of the Royal Society *2*, 1–55.

[Baudat et al., 2010] Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G. and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science (New York, N.Y.) *327*, 836–40.

[Belle et al., 2004] Belle, E. M. S., Duret, L., Galtier, N. and Eyre-Walker, A. (2004). The decline of isochores in mammals : an assessment of the GC content variation along the mammalian phylogeny. Journal of molecular evolution *58*, 653–60.

[Berg et al., 2010] Berg, I. L., Neumann, R., Lam, K.-W. G., Sarbajna, S., Odenthal-Hesse, L., May, C. a. and Jeffreys, A. J. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature genetics *42*, 859–63.

[Berg et al., 2011] Berg, I. L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N. J. and Jeffreys, A. J. (2011). Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proceedings of the National Academy of Sciences of the United States of America *108*, 12378–83.

[Bergerat et al., 1997] Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.-C., Nicolas, A. and Forterre, P. (1997). An atypical topoisomerase II from archea with implications for meiotic recombinaison. Nature *386*, 414–417.

[Berglund et al., 2009] Berglund, J., Pollard, K. S. and Webster, M. T. (2009). Hotspots of biased nucleotide substitutions in human genes. PLoS Biology *7*, 45–62.

[Bernardi, 2000] Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. Gene *241*, 3–17.

[Bernardi & Bernardi, 1991] Bernardi, G. and Bernardi, G. (1991). Compositional Properties of Nuclear Genes from. Compositional Properties of Nuclear Genes from Cold-Blooded Vertebrates *33*, 57–67.

[Bernardi et al., 1985] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science (New York, N.Y.) *228*, 953–8.

[Bernstein et al., 2012] Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

[Bird et al., 2007] Bird, C. P., Stranger, B. E., Liu, M., Thomas, D. J., Ingle, C. E., Beazley, C., Miller, W., Hurles, M. E. and Dermitzakis, E. T. (2007). Fast-evolving noncoding sequences in the human genome. Genome biology *8*, R118.

[Birdsell, 2002] Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Molecular Biology and Evolution *19*, 1181–97.

[Bishop & Zickler, 2004] Bishop, D. K. and Zickler, D. (2004). Early Decision. Cell *117*, 9–15.

[Borde & Cobb, 2009] Borde, V. and Cobb, J. (2009). Double functions for the Mre11 complex during DNA double-strand break repair and replication. The international journal of biochemistry & cell biology *41*, 1249–53.

[Börner et al., 2004] Börner, G. V., Kleckner, N. and Hunter, N. (2004). Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. Cell *117*, 29–45.

[Bosch et al., 2004] Bosch, E., Hurles, M., Navarro, A. and Jobling, M. (2004). Dynamics of a human interparalog gene conversion hotspot. Genome research *14*, 835–844.

[Boulton et al., 1997] Boulton, A., Myers, R. S. and Redfield, R. J. (1997). The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences of the United States of America *94*, 8058–63.

[Brick et al., 2012] Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D. and Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. Nature *485*, 642–645.

[Brunschwig et al., 2012] Brunschwig, H., Levi, L., Ben-David, E., Williams, R. W., Yakir, B. and Shifman, S. (2012). Fine-scale maps of recombination rates and hotspots in the mouse genome. Genetics *191*, 757–64.

[Buard et al., 2009] Buard, J., Barthès, P., Grey, C. and de Massy, B. (2009). Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. The EMBO journal *28*, 2616–24.

[Buhler et al., 2007] Buhler, C., Borde, V. and Lichten, M. (2007). Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae. PLoS biology *5*, e324.

[Capra et al., 2013] Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S. and Siepel, A. (2013). A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. PLoS genetics *9*, e1003684.

[Capra & Pollard, 2011] Capra, J. A. and Pollard, K. S. (2011). Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biology and Evolution *3*, 516–27.

[Chen & Jinks-Robertson, 1998] Chen, W. and Jinks-Robertson, S. (1998). Mismatch Repair Proteins Regulate Heteroduplex Formation during Mitotic Recombination in Yeast. Molecular and cellular biology *18*, 6525–6537.

[Choi et al., 2013] Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. a., Copenhaver, G. P., Franklin, F. C. H., McVean, G. and Henderson, I. R. (2013). Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nature genetics *45*, 1327–36.

[Christophorou et al., 2013] Christophorou, N., Rubin, T. and Huynh, J. (2013). Synaptonemal complex components promote centromere pairing in pre-meiotic germ cells. PLoS genetics *9*, 1–9.

[Clay & Bernardi, 2005] Clay, O. and Bernardi, G. (2005). How not to search for isochores : a reply to Cohen et Al. Molecular biology and evolution *22*, 2315–7.

[Coïc et al., 2000] Coïc, E., Gluck, L. and Fabre, F. (2000). Evidence for short-patch mismatch repair in Saccharomyces cerevisiae. EMBO Journal *19*, 3408–17.

[Cole et al., 2012a] Cole, F., Kauppi, L., Lange, J., Roig, I., Wang, R., Keeney, S. and Jasin, M. (2012a). Homeostatic control of recombination is

implemented progressively in mouse meiosis. Nature cell biology *14*, 424–30.

[Cole et al., 2012b] Cole, F., Keeney, S. and Jasin, M. (2012b). Preaching about the converted : how meiotic gene conversion influences genomic diversity. Annals of the New York Academy of Sciences *1267*, 95–102.

[Comeron et al., 2012] Comeron, J. M., Ratnappan, R. and Bailin, S. (2012). The many landscapes of recombination in Drosophila melanogaster. PLoS genetics *8*, e1002905.

[Coop & Przeworski, 2007] Coop, G. and Przeworski, M. (2007). An evolutionary view of human recombination. Nature Reviews Genetics *8*, 23–34.

[Coop et al., 2008] Coop, G., Wen, X., Ober, C., Pritchard, J. K. and Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science (New York, N.Y.) *319*, 1395–8.

[Costa et al., 2005] Costa, Y., Speed, R., Ollinger, R., Alsheimer, M., Semple, C. a., Gautier, P., Maratou, K., Novak, I., Höög, C., Benavente, R. and Cooke, H. J. (2005). Two novel proteins recruited by synaptonemal complex protein 1 (SYCP1) are at the centre of meiosis. Journal of cell science *118*, 2755–62.

[Cox et al., 2009] Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. a., Wergedal, J. E., Bult, C., Paigen, B., Flint, J., Tsaih, S.-W., Churchill, G. a. and Broman, K. W. (2009). A new standard genetic map for the laboratory mouse. Genetics *182*, 1335–44.

[Creighton & McClintock, 1931] Creighton, H. and McClintock, B. (1931). A correlation of cytological and genetical crossing-over in zea mays. ... of the National Academy of Sciences ... *17*, 492–497.

[Datta et al., 1996] Datta, A., Adjiri, A. and New, L. (1996). Mitotic crossovers between diverged sequences are regulated by mismatch repair proteins in Saccaromyces cerevisiae. Molecular and cellular ... *16*.

[de Boer et al., 2007] de Boer, E., Dietrich, A. J. J., Höög, C., Stam, P. and Heyting, C. (2007). Meiotic interference among MLH1 foci requires neither an intact axial element structure nor full synapsis. Journal of cell science *120*, 731–6.

[de Massy et al., 1995] de Massy, B., Rocco, V. and Nicolas, a. (1995). The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in Saccharomyces cerevisiae. The EMBO journal *14*, 4589–98.

[Dernburg et al., 1998] Dernburg, a. F., McDonald, K., Moulder, G., Barstead, R., Dresser, M. and Villeneuve, a. M. (1998). Meiotic recombination in C. elegans initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. Cell *94*, 387–98.

[Ding et al., 2010] Ding, D.-Q., Haraguchi, T. and Hiraoka, Y. (2010). From meiosis to postmeiotic events : alignment and recognition of homologous chromosomes in meiosis. The FEBS journal *277*, 565–70.

[Ding et al., 2004] Ding, D.-Q., Yamamoto, A., Haraguchi, T. and Hiraoka, Y. (2004). Dynamics of homologous chromosome pairing during meiotic prophase in fission yeast. Developmental cell *6*, 329–41.

[Doolittle, 2013] Doolittle, W. F. (2013). Is junk DNA bunk ? A critique of ENCODE. Proceedings of the National Academy of Sciences of the United States of America *110*, 5294–300.

[Dreszer et al., 2007] Dreszer, T. R., Wall, G. D., Haussler, D. and Pollard, K. S. (2007). Biased clustered substitutions in the human genome : the footprints of male-driven biased gene conversion. Genome research *17*, 1420–30.

[Drouaud et al., 2013] Drouaud, J., Khademian, H., Giraut, L., Zanni, V., Bellalou, S., Henderson, I. R., Falque, M. and Mézard, C. (2013). Contrasted patterns of crossover and non-crossover at Arabidopsis thaliana meiotic recombination hotspots. PLoS genetics *9*, e1003922.

[Duret, 2006] Duret, L. (2006). The GC content of primates and rodents genomes is not at equilibrium : a reply to Antezana. Journal of molecular evolution *62*, 803–6.

[Duret & Arndt, 2008] Duret, L. and Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genetics *4*, 1–19.

[Duret et al., 2006] Duret, L., Eyre-Walker, A. and Galtier, N. (2006). A new perspective on isochore evolution. Gene *385*, 71–4.

[Duret & Galtier, 2009a] Duret, L. and Galtier, N. (2009a). Biased gene conversion and the evolution of mammalian genomic landscapes. Annual Review of Genomics and Human Genetics *10*, 285–311.

[Duret & Galtier, 2009b] Duret, L. and Galtier, N. (2009b). Comment on "Human-specific gain of function in a developmental enhancer". Science (New York, N.Y.) *323*, 714.

[Duret et al., 2002] Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. and Galtier, N. (2002). Vanishing GC-rich isochores in mammalian genomes. Genetics *162*, 1837–47.

[Eisenstark, 1977] Eisenstark, A. (1977). Genetic Recombination in Bacteria. Annual review of genetics *11*, 369–96.

[Escobar et al., 2011] Escobar, J. S., Glémin, S. and Galtier, N. (2011). GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. Molecular Biology and Evolution *28*, 2561–75.

[Evans & Alani, 2000] Evans, E. and Alani, E. (2000). Roles for Mismatch Repair Factors in Regulating Genetic Recombination. Molecular and Cellular Biology *20*, 7839.

[Eyre-Walker, 1993] Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. Proceedings of the Royal Society of London B *252*, 237–243.

[Eyre-Walker, 1999] Eyre-Walker, a. (1999). Evidence of selection on silent site base composition in mammals : potential implications for the evolution of isochores and junk DNA. Genetics *152*, 675–83.

[Eyre-Walker & Hurst, 2001] Eyre-Walker, a. and Hurst, L. D. (2001). The evolution of isochores. Nature reviews. Genetics *2*, 549–55.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences : a maximum likelihood approach. Journal of molecular evolution *17*, 368–76.

[Fryxell & Zuckerkandl, 2000] Fryxell, K. J. and Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. Molecular biology and evolution *17*, 1371–83.

[Galtier, 2003] Galtier, N. (2003). Gene conversion drives GC content evolution in mammalian histones. Trends in genetics : TIG *19*, 65–8.

[Galtier & Duret, 2007] Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion ? Extending the null hypothesis of molecular evolution. Trends in Genetics *23*, 273–7.

[Galtier et al., 2009] Galtier, N., Duret, L., Glémin, S. and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends in Genetics : TIG *25*, 1–5.

[Galtier & Lobry, 1997] Galtier, N. and Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. Journal of molecular evolution *44*, 632–6.

[Galtier et al., 2001] Galtier, N., Piganeau, G. and Mouchiroud, D. (2001). GC-Content Evolution in Mammalian Genomes : The Biased Gene Conversion Hypothesis. Genetics *159*, 907–11.

[Gerton et al., 2000] Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O. and Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America *97*, 11383–90.

[Gerton & Hawley, 2005] Gerton, J. L. and Hawley, R. S. (2005). Homologous chromosome interactions in meiosis : diversity amidst conservation. Nature Reviews Genetics *6*, 477–87.

[Glémin, 2010] Glémin, S. (2010). Surprising fitness consequences of GC-biased gene conversion : I. Mutation load and inbreeding depression. Genetics *185*, 939–59.

[Goios et al., 2007] Goios, A., Pereira, L., Bogue, M., Macaulay, V. and Amorim, A. (2007). mtDNA phylogeny and evolution of laboratory mouse strains. Genome research *17*, 293–8.

[Goldfarb & Lichten, 2010] Goldfarb, T. and Lichten, M. (2010). Frequent and efficient use of the sister chromatid for DNA double-strand break repair during budding yeast meiosis. PLoS biology *8*, e1000520.

[Graur et al., 2013] Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. a. and 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*, 578–90.

[Grelon et al., 2001] Grelon, M., Vezon, D., Gendrot, G. and Pelletier, G. (2001). AtSPO11-1 is necessary for efficient meiotic recombination in plants. The EMBO journal *20*, 589–600.

[Grey et al., 2009] Grey, C., Baudat, F. and de Massy, B. (2009). Genome-wide control of the distribution of meiotic recombination. PLoS biology *7*, e35.

[Hassold & Hunt, 2001] Hassold, T. and Hunt, P. (2001). To err (meiotically) is human : the genesis of human aneuploidy. Nature reviews. Genetics *2*, 280–91.

[Haudry et al., 2008] Haudry, a., Cenci, a., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S., David, J. and Glémin, S. (2008). Mating system and recombination affect molecular evolution in four Triticeae species. Genetics research *90*, 97–109.

[Hayashi et al., 2005] Hayashi, K., Yoshida, K. and Matsui, Y. (2005). A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature *438*, 374–8.

[Hellmann et al., 2003] Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S. and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. American journal of human genetics *72*, 1527–35.

[Hershberg & Petrov, 2010] Hershberg, R. and Petrov, D. a. (2010). Evidence that mutation is universally biased towards AT in bacteria. PLoS genetics *6*, e1001115.

[Herskowitz, 1988] Herskowitz, I. (1988). Life cycle of the budding yeast Saccharomyces cerevisiae. Microbiological reviews *52*, 536–53.

[Hildebrand et al., 2010] Hildebrand, F., Meyer, A. and Eyre-walker, A. (2010). Evidence of Selection upon Genomic GC-Content in Bacteria. PLoS Genetics *6*.

[Hill & Robertson, 1966] Hill, W. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. Genetical Research *8*, 269–294.

[Hinch et al., 2011] Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., Berndt, S. I., Bernstein, L., Blot, W. J., Bock, C. H., Boerwinkle, E., Cai,

Q., Caporaso, N., Casey, G., Cupples, L. A., Deming, S. L., Diver, W. R., Divers, J., Fornage, M., Gillanders, E. M., Glessner, J., Harris, C. C., Hu, J. J., Ingles, S. a., Isaacs, W., John, E. M., Kao, W. H. L., Keating, B., Kittles, R. a., Kolonel, L. N., Larkin, E., Le Marchand, L., McNeill, L. H., Millikan, R. C., Murphy, A., Musani, S., Neslund-Dudas, C., Nyante, S., Papanicolaou, G. J., Press, M. F., Psaty, B. M., Reiner, A. P., Rich, S. S., Rodriguez-Gil, J. L., Rotter, J. I., Rybicki, B. a., Schwartz, A. G., Signorello, L. B., Spitz, M., Strom, S. S., Thun, M. J., Tucker, M. a., Wang, Z., Wiencke, J. K., Witte, J. S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K. a., Zheng, W., Ziegler, R. G., Zhu, X., Redline, S., Hirschhorn, J. N., Henderson, B. E., Taylor, H. a., Price, A. L., Hakonarson, H., Chanock, S. J., Haiman, C. a., Wilson, J. G., Reich, D. and Myers, S. R. (2011). The landscape of recombination in African Americans. Nature *476*, 170–5.

[Hollingsworth & Brill, 2004] Hollingsworth, N. M. and Brill, S. J. (2004). The Mus81 solution to resolution : generating meiotic crossovers without Holliday junctions. Genes & development *18*, 117–25.

[Holloway et al., 2008] Holloway, J. K., Booth, J., Edelmann, W., McGowan, C. H. and Cohen, P. E. (2008). MUS81 generates a subset of MLH1-MLH3-independent crossovers in mammalian meiosis. PLoS Genetics *4*, e1000186.

[Holmquist, 1992] Holmquist, G. (1992). Chromosome Bands, Their Chromatin Flavors, and Their Functional Features. American journal of human genetics *51*, 17–37.

[Hou et al., 2013] Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X. S. and Qiao, J. (2013). Genome analyses of single human oocytes. Cell *155*, 1492–506.

[Huang et al., 2005] Huang, S.-W., Friedman, R., Yu, N., Yu, A. and Li, W.-H. (2005). How strong is the mutagenicity of recombination in mammals ? Molecular biology and evolution *22*, 426–31.

[Hughes et al., 1999] Hughes, S., Zelus, D. and Mouchiroud, D. (1999). Warm-blooded isochore structure in Nile crocodile and turtle. Molecular biology and evolution *16*, 1521–7.

[Hunter & Borts, 1997] Hunter, N. and Borts, R. H. (1997). Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. Genes & Development *11*, 1573–1582.

[Ira et al., 2003] Ira, G., Malkova, A., Liberi, G., Foiani, M. and Haber, J. E. (2003). Srs2 and Sgs1–Top3 Suppress Crossovers during Double-Strand Break Repair in Yeast. Cell *115*, 401–411.

[Jeffreys et al., 2013] Jeffreys, A. J., Cotton, V. E., Neumann, R. and Lam, K.-W. G. (2013). Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. Proceedings of the National Academy of Sciences of the United States of America *110*, 600–5.

[Jeffreys et al., 2001] Jeffreys, a. J., Kauppi, L. and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature genetics *29*, 217–22.

[Jeffreys & May, 2004] Jeffreys, A. J. and May, C. a. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nature genetics *36*, 151–6.

[Jeffreys & Neumann, 2002] Jeffreys, A. J. and Neumann, R. (2002). Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nature Genetics *31*, 267–71.

[Jeffreys & Neumann, 2005] Jeffreys, A. J. and Neumann, R. (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Human molecular genetics *14*, 2277–87.

[Jeffreys & Neumann, 2009] Jeffreys, A. J. and Neumann, R. (2009). The rise and fall of a human recombination hot spot. Nature genetics *41*, 625–9.

[Jiricny, 2006] Jiricny, J. (2006). The multifaceted mismatch-repair system. Nature Reviews Molecular Cell Biology *7*, 335–46.

[Källén et al., 1996] Källén, B., Mastroiacovo, P. and Robert, E. (1996). Major congenital malformations in Down syndrome. American journal of Medical Genetics *65*, 160–166.

[Karpen et al., 1996] Karpen, G. H., Le, M. H. and Le, H. (1996). Centric heterochromatin and the efficiency of achiasmate disjunction in Drosophila female meiosis. Science (New York, N.Y.) *273*, 118–22.

[Katzman et al., 2011] Katzman, S., Capra, J. A., Haussler, D. and Pollard, K. S. (2011). Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. Genome Biology and Evolution *3*, 614–26.

[Katzman et al., 2010] Katzman, S., Kern, A. D., Pollard, K. S., Salama, S. R. and Haussler, D. (2010). GC-biased evolution near human accelerated regions. PLoS genetics *6*, e1000960.

[Kauppi et al., 2009] Kauppi, L., May, C. A. and Jeffreys, A. J. (2009). Analysis of Meiotic Recombination Products from Human Sperm. Methods in Molecular Biology *557*, 323–355.

[Keeney et al., 1997] Keeney, S., Giroux, C. N. and Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell *88*, 375–84.

[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. and Hardison, R. C. (2014). Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences *Early Edit*, 1–8.

[Khil et al., 2012] Khil, P. P., Smagulova, F., Brick, K. M., Camerini-Otero, R. D. and Petukhova, G. V. (2012). Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. Genome research *22*, 957–65.

[Kimura, 1962] Kimura, M. (1962). On the Probability of Fixation of Mutant Genes in a Population. Genetics *47*, 713–719.

[Kimura, 1968] Kimura, M. (1968). Evolutionary Rate at the Molecular Level. Nature *217*, 624–626.

[Kirkness et al., 2013] Kirkness, E. F., Grindberg, R. V., Yee-Greenbaum, J., Marshall, C. R., Scherer, S. W., Lasken, R. S. and Venter, J. C. (2013). Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome research *23*, 826–32.

[Kirkpatrick et al., 1998] Kirkpatrick, D., Dominska, M. and Petes, T. (1998). Conversion-Type and Restoration-Type Repair of DNA Mismatches Formed During Meiotic Recombination in Saccharomyces cerevisiae. Genetics *149*, 1693–1705.

[Klein et al., 1999] Klein, F., Mahr, P., Galova, M., Buonomo, S. B., Michaelis, C., Nairz, K. and Nasmyth, K. (1999). A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. Cell *98*, 91–103.

[Kolodner & Marsischky, 1999] Kolodner, R. D. and Marsischky, G. T. (1999). Eukaryotic DNA mismatch repair. Current Opinion in Genetics and Development *9*, 89–96.

[Kong et al., 2002] Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R. and Stefansson, K. (2002). A high-resolution recombination map of the human genome. Nature genetics *31*, 241–7.

[Kong et al., 2010] Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., Gudjonsson, S. a., Frigge, M. L., Helgason, A., Thorsteinsdottir, U. and Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–103.

[Koszul & Kleckner, 2009] Koszul, R. and Kleckner, N. (2009). Dynamic chromosome movements during meiosis : a way to eliminate unwanted connections ? Trends in cell biology *19*, 716–24.

[Koszul et al., 2012] Koszul, R., Meselson, M., Van Doninck, K., Vandenhaute, J. and Zickler, D. (2012). The centenary of Janssens's chiasmatype theory. Genetics *191*, 309–17.

[Krogh & Symington, 2004] Krogh, B. O. and Symington, L. S. (2004). Recombination proteins in yeast. Annual Review of Genetics *38*, 233–71.

[LaFave & Sekelsky, 2009] LaFave, M. C. and Sekelsky, J. (2009). Mitotic recombination : why ? when ? how ? where ? PLoS genetics *5*, e1000411.

[Lamb et al., 2005] Lamb, N. E., Yu, K., Shaffer, J., Feingold, E. and Sherman, S. L. (2005). Association between maternal age and meiotic recombination for trisomy 21. American journal of human genetics *76*, 91–9.

[Langergraber et al., 2012] Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., Robbins, M. M., Schubert, G., Stoinski, T. S.,

Viola, B., Watts, D., Wittig, R. M., Wrangham, R. W., Zuberbühler, K., Pääbo, S. and Vigilant, L. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proceedings of the National Academy of Sciences of the United States of America *109*, 15716–21.

[Lartillot, 2013] Lartillot, N. (2013). Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. Molecular biology and evolution *30*, 489–502.

[Lenormand & Dutheil, 2005] Lenormand, T. and Dutheil, J. (2005). Recombination difference between sexes : a role for haploid selection. PLoS biology *3*, e63.

[Lercher & Hurst, 2002] Lercher, M. J. and Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. Trends in Genetics *18*, 337–40.

[Lesecque et al., 2012] Lesecque, Y., Keightley, P. D. and Eyre-Walker, A. (2012). A Resolution of the Mutation Load Paradox in Humans. Genetics *191*, 1321–1330.

[Lesecque et al., 2013] Lesecque, Y., Mouchiroud, D. and Duret, L. (2013). GC-biased gene conversion in yeast is specifically associated with crossovers : molecular mechanisms and evolutionary significance. Molecular biology and evolution *30*, 1409–19.

[Li et al., 1988] Li, H., Gyllensten, U., Cui, X., Saiki, R., Erlich, H. and Arnheim, N. (1988). Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature *335*, 414–417.

[Lichten, 2008] Lichten, M. (2008). Meiotic Chromatin : The Substrate for Recombination Initiation. Genome Dynamics and Stability *3*, 165–193.

[Lichten & de Massy, 2011] Lichten, M. and de Massy, B. (2011). The impressionistic landscape of meiotic recombination. Cell *147*, 267–70.

[Lichten & Goldman, 1995] Lichten, M. and Goldman, a. S. (1995). Meiotic recombination hotspots. Annual review of genetics *29*, 423–44.

[Liu & West, 2004] Liu, Y. and West, S. C. (2004). Happy Hollidays : 40th anniversary of the Holliday junction. Nature reviews. Molecular cell biology *5*, 937–44.

[Lobo & Show, 2008] Lobo, I. and Show, K. (2008). Discovery and Types of Genetic Linkage. Nature education *1*, 139.

[Lu et al., 2012] Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A. J., Li, T. and Ma, H. (2012). Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome research *22*, 508–18.

[Lynch, 2010] Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America *107*, 961–8.

[Mancera et al., 2008] Mancera, E., Bourgon, R., Brozzi, A., Huber, W. and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature *454*, 479–85.

[Marais, 2003] Marais, G. (2003). Biased gene conversion : implications for genome and sex evolution. Trends in Genetics *19*, 330–338.

[Martini et al., 2011] Martini, E., Borde, V., Legendre, M., Audic, S., Regnault, B., Soubigou, G., Dujon, B. and Llorente, B. (2011). Genome-Wide Analysis of Heteroduplex DNA in Mismatch Repair–Deficient Yeast Cells Reveals Novel Properties of Meiotic Recombination Pathways. PLoS Genetics *7*, 1–18.

[Martini et al., 2006] Martini, E., Diaz, R. L., Hunter, N. and Keeney, S. (2006). Crossover homeostasis in yeast meiosis. Cell *126*, 285–95.

[Mazina et al., 2004] Mazina, O. M., Mazin, A. V., Nakagawa, T., Kolodner, R. D. and Kowalczykowski, S. C. (2004). Saccharomyces cerevisiae Mer3 Helicase Stimulates 3'–5' Heteroduplex Extension by Rad51. Cell *117*, 47–56.

[McVean et al., 2004] McVean, G. a. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science (New York, N.Y.) *304*, 581–4.

[Memisoglu & Samson, 2000] Memisoglu, A. and Samson, L. (2000). Base excision repair in yeast and mammals. Mutation Research *451*, 39–51.

[Mercier et al., 2005] Mercier, R., Jolivet, S., Vezon, D., Huppe, E., Chelysheva, L., Giovanni, M., Nogué, F., Doutriaux, M.-P., Horlow, C., Grelon,

M. and Mézard, C. (2005). Two meiotic crossover classes cohabit in Arabidopsis : one is dependent on MER3, whereas the other one is not. Current biology : CB *15*, 692–701.

[Metz, 1926] Metz, C. (1926). Observations on Spermatogenesis in Drosophila. Zeitschrift für Zellforschung und Mikroskopische ... *4*, 1–28.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature reviews. Genetics *11*, 31–46.

[Meunier & Duret, 2004] Meunier, J. and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. Molecular Biology and Evolution *21*, 984–90.

[Meyer et al., 2012] Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F., Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J. and Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science (New York, N.Y.) *338*, 222–6.

[Mézard, 2006] Mézard, C. (2006). Meiotic recombination hotspots in plants. Biochemical Society transactions *34*, 531–4.

[Mihola et al., 2009] Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. and Forejt, J. (2009). A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science *323*, 373–375.

[Montoya-Burgos et al., 2003] Montoya-Burgos, J. I., Boursot, P. and Galtier, N. (2003). Recombination explains isochores in mammalian genomes. Trends in genetics : TIG *19*, 128–30.

[Morgan, 1911] Morgan, T. H. (1911). Random Segregation versus Coupling in Mendelian Inheritance. Science (New York, N.Y.) *34*, 384.

[Morgan et al., 1915] Morgan, T. H., Sturtevant, A. H., Muller, H. J. and Bridges, C. B. (1915). The Mechanism of Mendelian Heredity. New-York.

[Mouchiroud et al., 1987] Mouchiroud, D., Fichant, G. and Bernardi, G. (1987). Compositional compartmentalization and gene composition in the genome of vertebrates. Journal of Molecular Evolution *26*, 198–204.

[Mugal et al., 2013] Mugal, C. F., Arndt, P. F. and Ellegren, H. (2013). Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. Molecular biology and evolution *30*, 1700–12.

[Muller, 1925] Muller, H. (1925). The Regionally Differential Effect Of X Rays On Crossing Over In Autosomes Of Drosophila. Genetics *10*, 470–507.

[Munch et al., 2014] Munch, K., Mailund, T., Dutheil, J. Y. and Schierup, M. H. (2014). A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. Genome research *24*, 467–74.

[Murakami & Nicolas, 2009] Murakami, H. and Nicolas, A. (2009). Locally, Meiotic Double-Strand Breaks Targeted by Gal4BD-Spo11 Occur at Discrete Sites with a Sequence Preference. Molecular and Cellular Biology *29*, 3500–3516.

[Myers et al., 2005] Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science (New York, N.Y.) *310*, 321–4.

[Myers et al., 2010] Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G. and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science (New York, N.Y.) *327*, 876–9.

[Myers et al., 2008] Myers, S., Freeman, C., Auton, A., Donnelly, P. and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nature genetics *40*, 1124–9.

[Myers et al., 2006] Myers, S., Spencer, C. C. a., Auton, a., Bottolo, L., Freeman, C., Donnelly, P. and McVean, G. (2006). The distribution and causes of meiotic recombination in the human genome. Biochemical Society transactions *34*, 526–30.

[Nagylaki, 1983] Nagylaki, T. (1983). Evolution of a finite population under gene conversion. Proceedings of the National Academy of Sciences of the United States of America *80*, 6278–81.

[Neale & Keeney, 2006] Neale, M. J. and Keeney, S. (2006). Clarifying the mechanics of DNA strand exchange in meiotic recombination. Nature *442*, 153–8.

[Necşulea et al., 2011] Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C. and Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. Human Mutation *32*, 198–206.

[Nicholson et al., 2000] Nicholson, a., Hendrix, M., Jinks-Robertson, S. and Crouse, G. F. (2000). Regulation of mitotic homeologous recombination in yeast. Functions of mismatch repair and nucleotide excision repair genes. Genetics *154*, 133–46.

[O'Connor, 2008] O'Connor, C. (2008). Meiosis, Genetic Recombination, and Sexual Reproduction. Nature education *1*, 174.

[Odenthal-Hesse et al., 2014] Odenthal-Hesse, L., Berg, I. L., Veselis, A., Jeffreys, A. J. and May, C. a. (2014). Transmission distortion affecting human noncrossover but not crossover recombination : a hidden source of meiotic drive. PLoS genetics *10*, e1004106.

[Oliver et al., 2009] Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., Beatson, S. a., Lunter, G., Malik, H. S. and Ponting, C. P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics *5*, e1000753.

[O'Reilly et al., 2008] O'Reilly, P. F., Birney, E. and Balding, D. J. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome research *18*, 1304–13.

[Osman et al., 2003] Osman, F., Dixon, J., Doe, C. L. and Whitby, M. C. (2003). Generating crossovers by resolution of nicked Holliday junctions : a role for Mus81-Eme1 in meiosis. Molecular cell *12*, 761–74.

[Otto & Barton, 1997] Otto, S. and Barton, N. (1997). The Evolution of Recombination : Removing the limits to natural selection. Genetics *147*, 879–906.

[Page & Hawley, 2004] Page, S. L. and Hawley, R. S. (2004). The genetics and molecular biology of the synaptonemal complex. Annual review of cell and developmental biology *20*, 525–58.

[Paigen & Petkov, 2010] Paigen, K. and Petkov, P. (2010). Mammalian recombination hot spots : properties, control and evolution. Nature reviews. Genetics *11*, 221–33.

[Pan et al., 2011] Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H. G., Tischfield, S. E., Zhu, X., Neale, M. J., Jasin, M., Socci, N. D., Hochwagen, A. and Keeney, S. (2011). A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. Cell *144*, 719–31.

[Papadakis & Patrinos, 1999] Papadakis, M. and Patrinos, G. (1999). Contribution of gene conversion in the evolution of the human $\beta$-like globin gene family. Human genetics *104*, 117–125.

[Pâques & Haber, 1999] Pâques, F. and Haber, J. (1999). Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae. Microbiology and molecular biology reviews *63*, 349–404.

[Parvanov et al., 2009] Parvanov, E. D., Ng, S. H. S., Petkov, P. M. and Paigen, K. (2009). Trans-regulation of mouse meiotic recombination hotspots by Rcr1. PLoS biology *7*, e36.

[Perry & Ashworth, 1999] Perry, J. and Ashworth, a. (1999). Evolutionary rate of a gene affected by chromosomal position. Current biology : CB *9*, 987–9.

[Pessia et al., 2012] Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L. and Marais, G. A. B. (2012). Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. Genome Biology and Evolution *4*, 675–82.

[Petronczki et al., 2003] Petronczki, M., Siomos, M. F., Nasmyth, K., Correns, C., Vries, H. D. and Tscher, E. (2003). Un Ménage à Quatre : The Molecular Biology of Chromosome Segregation in Meiosis. Cell *112*, 423–440.

[Phillips & Dernburg, 2006] Phillips, C. M. and Dernburg, A. F. (2006). A family of zinc-finger proteins is required for chromosome-specific pairing and synapsis during meiosis in C. elegans. Developmental cell *11*, 817–29.

[Pollard et al., 2006a] Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J. and Haussler, D. (2006a). Forces shaping the fastest evolving regions in the human genome. PLoS genetics *2*, e168.

[Pollard et al., 2006b] Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, M., Vanderhaeghen,

P. and Haussler, D. (2006b). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–72.

[Ponting, 2011] Ponting, C. P. (2011). What are the genomic drivers of the rapid evolution of PRDM9 ? Trends in genetics : TIG *27*, 165–71.

[Popa et al., 2012] Popa, A., Samollow, P., Gautier, C. and Mouchiroud, D. (2012). The sex-specific impact of meiotic recombination on nucleotide composition. Genome biology and evolution *4*, 412–22.

[Popa, 2011] Popa, A. M. (2011). The Evolution of Recombination and Genomic Structures : a Modeling Approach. PhD thesis, Université Claude-Bernard Lyon 1.

[Prabhakar et al., 2006] Prabhakar, S., Noonan, J. P., Pääbo, S. and Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. Science (New York, N.Y.) *314*, 786.

[Prabhakar et al., 2008] Prabhakar, S., Visel, A., Akiyama, J. a., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D. R., Afzal, V., Pennacchio, L. a., Rubin, E. M. and Noonan, J. P. (2008). Human-specific gain of function in a developmental enhancer. Science (New York, N.Y.) *321*, 1346–50.

[Prüfer et al., 2014] Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L. F., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J. and Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–9.

[Ptak et al., 2005] Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. a. and Pääbo, S. (2005). Fine-scale recombination patterns differ between chimpanzees and humans. Nature genetics *37*, 429–34.

[Ratnakumar et al., 2010] Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L. and Webster, M. T. (2010). Detecting

positive selection within genomes : the problem of biased gene conversion. Philosophical Transactions of the Royal Society of London Series B Biological Sciences *365*, 2571–80.

[Rhead et al., 2010] Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. a., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. a., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D. and Kent, W. J. (2010). The UCSC Genome Browser database : update 2010. Nucleic acids research *38*, D613–9.

[Robine et al., 2007] Robine, N., Uematsu, N., Amiot, F., Gidrol, X., Barillot, E., Nicolas, A. and Borde, V. (2007). Genome-wide redistribution of meiotic double-strand breaks in Saccharomyces cerevisiae. Molecular and cellular biology *27*, 1868–80.

[Robinson et al., 2013] Robinson, M. C., Stone, E. a. and Singh, N. D. (2013). Population genomic analysis reveals no evidence for GC-biased gene conversion in Drosophila melanogaster. Molecular biology and evolution *31*, 425–33.

[Rockman & Kruglyak, 2009] Rockman, M. V. and Kruglyak, L. (2009). Recombinational landscape and population genomics of Caenorhabditis elegans. PLoS genetics *5*, e1000419.

[Rockmill et al., 2006] Rockmill, B., Voelkel-Meiman, K. and Roeder, G. S. (2006). Centromere-proximal crossovers are associated with precocious separation of sister chromatids during meiosis in Saccharomyces cerevisiae. Genetics *174*, 1745–54.

[Romanienko & Camerini-Otero, 2000] Romanienko, P. J. and Camerini-Otero, R. D. (2000). The mouse Spo11 gene is required for meiotic chromosome synapsis. Molecular cell *6*, 975–87.

[Romiguier et al., 2010] Romiguier, J., Ranwez, V., Douzery, E. J. P. and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes : relationship with life-history traits and chromosome sizes. Genome research *20*, 1001–9.

[Schaibley et al., 2013] Schaibley, V. M., Zawistowski, M., Wegmann, D., Ehm, M. G., Nelson, M. R., St Jean, P. L., Abecasis, G. R., Novembre, J., Zöllner, S. and Li, J. Z. (2013). The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome research *23*, 1974–84.

[Schneider & Dekker, 2012] Schneider, G. F. and Dekker, C. (2012). DNA sequencing with nanopores. Nature biotechnology *30*, 326–8.

[Schwacha & Kleckner, 1997] Schwacha, a. and Kleckner, N. (1997). Interhomolog bias during meiotic recombination : meiotic functions promote a highly differentiated interhomolog-only pathway. Cell *90*, 1123–35.

[Schwartz, 2009] Schwartz, J. (2009). In pursuit of the gene : from Darwin to DNA. Harvard University Press, Cambridge.

[Silver, 1995] Silver, L. M. (1995). Mouse Genetics. Concepts and Applications. Oxford University Press, Oxford.

[Smagulova et al., 2011] Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D. and Petukhova, G. V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature *472*, 375–8.

[Smith & Nicolas, 1998] Smith, K. and Nicolas, A. (1998). Recombination at work for meiosis. Current Opinion in Genetics and Development *8*, 200–211.

[Snowden et al., 2004] Snowden, T., Acharya, S., Butz, C., Berardini, M. and Fishel, R. (2004). hMSH4-hMSH5 recognizes Holliday Junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. Molecular cell *15*, 437–51.

[Spencer et al., 2006] Spencer, C. C. a., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D. and McVean, G. (2006). The influence of recombination on human genetic diversity. PLoS genetics *2*, e148.

[Stenson et al., 2009] Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S. and Cooper, D. N. (2009). The Human Gene Mutation Database : 2008 update. Genome medicine *1*, 13.

[Sueoka, 1962] Sueoka, N. (1962). On The Genetic Basis Of Variation And Heterogeneity Of Dna Base Composition. Proceedings of the National Academy of Sciences of ... *48*, 582–592.

[Sumiyama & Saitou, 2011] Sumiyama, K. and Saitou, N. (2011). Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. Molecular biology and evolution *28*, 3005–7.

[Sunyaev et al., 2001] Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, a. S. and Bork, P. (2001). Prediction of deleterious human alleles. Human molecular genetics *10*, 591–7.

[Surtees et al., 2004] Surtees, J. A., Argueso, J. L. and Alani, E. (2004). Mismatch repair proteins : key regulators of genetic recombination. Cytogenetic and Genome Research *107*, 146–59.

[Székvölgyi & Nicolas, 2010] Székvölgyi, L. and Nicolas, A. (2010). From meiosis to postmeiotic events : homologous recombination is obligatory but flexible. The FEBS journal *277*, 571–89.

[Szostak et al., 1983] Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. and Stahl, F. W. (1983). The double-strand-break repair model for recombination. Cell *33*, 25–35.

[The 1000 Genomes Project Consortium, 2010] The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–73.

[The 1000 Genomes Project Consortium, 2012] The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

[The International HapMap Consortium, 2007] The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–61.

[Tsai et al., 2010] Tsai, I. J., Burt, A. and Koufopanou, V. (2010). Conservation of recombination hotspots in yeast. Proceedings of the National Academy of Sciences of the United States of America *107*, 7847–52.

[Tzur et al., 2006] Tzur, Y. B., Wilson, K. L. and Gruenbaum, Y. (2006). SUN-domain proteins : 'Velcro' that links the nucleoskeleton to the cytoskeleton. Nature reviews Molecular cell biology *7*, 782–8.

[Ubeda & Wilkins, 2011] Ubeda, F. and Wilkins, J. F. (2011). The Red Queen theory of recombination hotspots. Journal of evolutionary biology *24*, 541–53.

[Vallot et al., 2013] Vallot, C., Huret, C., Lesecque, Y., Resch, A., Oudrhiri, N., Bennaceur-Griscelli, A., Duret, L. and Rougeulle, C. (2013). XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. Nature genetics *45*, 239–41.

[van Valen, 1973] van Valen, L. (1973). A new evolutionary law. Evolutionary theory *30*, 1–30.

[Wang et al., 2012] Wang, J., Fan, H. C., Behr, B. and Quake, S. R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell *150*, 402–12.

[Wang et al., 2010] Wang, S., Zhang, L., Hu, J., Bao, Z. and Liu, Z. (2010). Molecular and cellular evidence for biased mitotic gene conversion in hybrid scallop. BMC evolutionary biology *10*, 6.

[Webb et al., 2008] Webb, A. J., Berg, I. L. and Jeffreys, A. (2008). Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proceedings of the National Academy of Sciences of the United States of America *105*, 10471–6.

[Webster et al., 2006] Webster, M. T., Axelsson, E. and Ellegren, H. (2006). Strong regional biases in nucleotide substitution in the chicken genome. Molecular biology and evolution *23*, 1203–16.

[Webster & Smith, 2004] Webster, M. T. and Smith, N. G. C. (2004). Fixation biases affecting human SNPs. Trends in genetics : TIG *20*, 122–6.

[Webster et al., 2005] Webster, M. T., Smith, N. G. C., Hultin-Rosenberg, L., Arndt, P. F. and Ellegren, H. (2005). Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. Molecular biology and evolution *22*, 1468–74.

[Wei et al., 2003] Wei, K., Clark, A. B., Wong, E., Kane, M. F., Mazur, D. J., Parris, T., Kolas, N. K., Russell, R., Hou, H., Kneitz, B., Yang, G., Kunkel, T. a., Kolodner, R. D., Cohen, P. E. and Edelmann, W. (2003). Inactivation of Exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. Genes & development *17*, 603–14.

[Whitby, 2005] Whitby, M. C. (2005). Making crossovers during meiosis. Biochemical Society transactions *33*, 1451–5.

[Winckler et al., 2005] Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G. a. T., Gabriel, S. B., Reich, D., Donnelly, P. and Altshuler, D. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. Science (New York, N.Y.) *308*, 107–11.

[Wolfe et al., 1989] Wolfe, K., Sharp, P. and Li, W. (1989). Mutation rates differ among regions of the mammalian genome. Nature *337*, 283–5.

[Wright, 1931] Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97–159.

[Wu & Hickson, 2003] Wu, L. and Hickson, I. (2003). The Bloom's syndrome helicase suppresses crossing over during homologous recombination. Nature *426*, 15–19.

[Wu et al., 2010] Wu, Z. K., Getun, I. V. and Bois, P. R. J. (2010). Anatomy of mouse recombination hot spots. Nucleic acids research *38*, 2346–54.

[Yang, 2007] Yang, Z. (2007). PAML 4 : phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution *24*, 1586–91.

[Yanowitz, 2010] Yanowitz, J. (2010). Meiosis : making a break for it. Current opinion in cell biology *22*, 744–51.

[Yi et al., 2004] Yi, S., Summers, T. J., Pearson, N. M. and Li, W.-H. (2004). Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome research *14*, 37–43.

[Zakharyevich et al., 2012] Zakharyevich, K., Tang, S., Ma, Y. and Hunter, N. (2012). Delineation of joint molecule resolution pathways in meiosis identifies a crossover-specific resolvase. Cell *149*, 334–47.

[Zickler, 2006] Zickler, D. (2006). From early homologue recognition to synaptonemal complex formation. Chromosoma *115*, 158–74.

[Zimmering et al., 1970] Zimmering, S., Sandler, L. and Nicoletti, B. (1970). Mechanisms of meiotic drive. Annual review of genetics *4*, 409–36.

# Annexes

Annexe A

# Les méthodes d'étude de la recombinaison : un bilan

Dans cette annexe nous présentons un tableau permettant de comparer les méthodes les plus couramment utilisées pour étudier et mesurer la recombinaison dans les génomes (*cf.* page suivante). NB : La colonne "Tract de conversion" indique les techniques qui permettent d'analyser, avec une bonne résolution, les tracts de conversion (simples et complexes). La colonne "Espèce(s)" donne la liste des espèces chez qui cette technique a été utilisée, de manière non exhaustive (H = Homme, C = Chimpanzé, L = Levure, S = Souris, A = *Arabidopsis thaliana*, Anc = Ancêtre de l'homme et du chimpanzé). La colonne "Résolution" donne le(s) facteur(s) limitants de l'approche et la meilleure résolution atteinte avec cette technique à notre connaissance, entre parenthèses. La colonne "Age" donne l'âge des évènements détectés (*cf.* introduction).

| Technique | Génome entier | Evénements détectés | Tract de conversion | Résolution | Age | Espèce(s) | Références |
|---|---|---|---|---|---|---|---|
| Sperm-typing | Non | CO & NCO | Oui | # de SNP ($\sim 50pb$) | Présents | H, S, A | [Li et al., 1988, Kauppi et al., 2009] |
| Génotypage massif de produits méiotiques | Oui | CO & NCO | Oui | # de SNP ($\sim 100pb$) | Présents | L | [Mancera et al., 2008] |
| Cartes de DSB | Oui | DSB | Non | Profondeur de séquençage (200pb) | Présents | L, S | [Gerton et al., 2000, Smagulova et al., 2011, Khil et al., 2012] |
| Cartes de DL | Oui | CO | Non | # de SNP (2kb) | Historiques | H, C, S | [McVean et al., 2004, Auton et al., 2012] |
| Cartes basées sur l'étude du métissage | Oui | CO | Non | # de SNP ($<$ 3kb) | Historiques | H | [Hinch et al., 2011] |
| Cartes basées sur les pedigree | Oui | CO | Non | Profondeur de séquençage et # de SNP (10kb) | Présents | H | [Kong et al., 2002, Kong et al., 2010] |
| Cartes basées sur l'ILS | Oui | CO | Non | ($\sim 10kb$) | Ancestraux | Anc | [Munch et al., 2014] |



TABLEAU A.1 : Bilan sur les méthodes d'étude de la recombinaison et de la conversion génique méiotique. *Légende : cf. page précédente.*

Annexe B

# Articles publiés en dehors du sujet de thèse

## B.1. A Resolution of the Mutation Load Paradox in Humans

*Voir article joint page suivante.*
Référence : [Lesecque et al., 2012]

# A Resolution of the Mutation Load Paradox in Humans

**Yann Lesecque,*,†,1 Peter D. Keightley,‡ and Adam Eyre-Walker*,2**

*School of Life Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom, †Ecole Normale Superieure, Lyon, BP 7000 69342 Lyon, Cedex 07, France, and ‡Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

**ABSTRACT** Current information on the rate of mutation and the fraction of sites in the genome that are subject to selection suggests that each human has received, on average, at least two new harmful mutations from its parents. These mutations were subsequently removed by natural selection through reduced survival or fertility. It has been argued that the mutation load, the proportional reduction in population mean fitness relative to the fitness of an idealized mutation-free individual, allows a theoretical prediction of the proportion of individuals in the population that fail to reproduce as a consequence of these harmful mutations. Application of this theory to humans implies that at least 88% of individuals should fail to reproduce and that each female would need to have more than 16 offspring to maintain population size. This prediction is clearly at odds with the low reproductive excess of human populations. Here, we derive expressions for the fraction of individuals that fail to reproduce as a consequence of recurrent deleterious mutation ($\varphi$) for a model in which selection occurs via differences in relative fitness, such as would occur through competition between individuals. We show that $\varphi$ is much smaller than the value predicted by comparing fitness to that of a mutation-free genotype. Under the relative fitness model, we show that $\varphi$ depends jointly on $U$ and the selective effects of new deleterious mutations and that a species could tolerate 10's or even 100's of new deleterious mutations per genome each generation.

ALL organisms are subject to recurrent deleterious mutation, which cause some individuals to die or fail to reproduce. Deleterious mutations therefore impose a cost or load on the population. The evolutionary consequences of deleterious mutations were first studied by J. B. S. Haldane, who showed that the reduction in mean fitness in a diploid organism caused by recurrent semidominant deleterious mutation at a single locus is equal to twice the mutation rate (Haldane 1937). This led H. J. Muller to suggest that each new deleterious mutation ultimately leads to one genetic death, irrespective of the mutation's fitness effect (Muller 1950). Subsequently, the mutation load was more formally defined as the proportional reduction in mean fitness of a population relative to that of a mutation-free genotype, brought about by deleterious mutations (Crow 1970):

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}}, \tag{1}$$

where $\bar{w}$ is the mean fitness of the population at equilibrium and $w_{\max}$ is the mean fitness of a deleterious mutation-free individual.

Under viability selection, the mutation load is equivalent to the proportion of individuals that fail to survive and hence leave no descendants in the next generation. For example, if an individual carries 10 mutations, each reducing the chance of surviving to reproductive age by 10%, then this individual is expected to survive with probability $(1–0.1)^{10} = 0.35$. If all individuals in the population have this genotype, then 65% of them would fail to have any descendants in the next generation. The load does not have such a simple interpretation under fertility selection, as we discuss below.

If the fitness effects of deleterious mutations are independent from one another, the mutation load across all loci subject to recurrent mutation is approximately


doi: 10.1534/genetics.112.140343
Manuscript received March 8, 2012; accepted for publication May 24, 2012
[1]Present address: UMR CNRS 5558, Biométrie et Biologie évolutive, UCB Lyon 1, Bât. Grégor Mendel, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France.
[2]Corresponding author: University of Sussex, School of Life Sciences, Brighton, BN1 9QG, United Kingdom. E-mail: a.c.eyre-walker@sussex.ac.uk

$$L \approx 1 - e^{-U} \quad (2)$$

(Kimura and Maruyama 1966), where $U$ is the overall rate of deleterious mutation per diploid genome per generation. This simple formula is a classic result of evolutionary genetics and appears in almost every textbook on the subject.

It has previously been estimated that $U$ is considerably greater than one in humans (Kondrashov and Crow 1993; Eyre-Walker and Keightley 1999; Nachman and Crowell 2000) and may be as high as 10 (Reed *et al.* 2005). Under Crow's (1970) definition of the mutation load and a viability selection model, the fraction of individuals that fail to reproduce, $\varphi$, is predicted to be considerable; for example, if $U$ is as high as 3, $\varphi = 1 - e^{-3} = 95\%$. However, previous estimates of $U$ have relied on indirect estimates of the mutation rate, based on the neutral divergence between human and chimpanzee, and inaccurate estimates of the proportion of sites in the genome that are subject to natural selection (Kondrashov and Crow 1993; Eyre-Walker and Keightley 1999; Nachman and Crowell 2000). The mutation rate per nucleotide site in humans ($\mu$) has recently been directly estimated by comparing the genome sequences of offspring and their parents. Three studies (Awadalla *et al.* 2010; Durbin *et al.* 2010; Roach *et al.* 2010) have yielded consistent estimates, with a mean of $\mu = 1.1 \times 10^{-8}$. Assuming a diploid genome of $6 \times 10^9$ nucleotides, each newborn therefore receives ~66 new single nucleotide mutations from its parents. To estimate $U$, we need to multiply this figure by the fraction of mutations that are deleterious ($\zeta$) (Kondrashov and Crow 1993). Comparisons of the human and mouse genomes and the human and macaque genomes suggest that 5–6.5% of sites are subject to some degree of purifying selection (Meader *et al.* 2010; Lindblad-Toh *et al.* 2011; Mouse Genome sequencing Consortium 2002). However, the level of conservation, and hence $\zeta$, was not explicitly estimated in these analyses, making it difficult to estimate $U$. A more formal analysis has estimated $\zeta$ by comparing the human–chimp nucleotide divergence for transposable element (TE) remnants, which appear to evolve largely neutrally (Lunter *et al.* 2006; Meader *et al.* 2010), with the divergence for the remainder of the genome (Eory *et al.* 2010). The non-TE fraction evolves at 94.3% the rate of the TE fraction, suggesting that 5.7% of non-TE mutations are deleterious and removed by natural selection. The non-TE fraction constitutes ~55% of the genome (Cordaux and Batzer 2009), so an estimate of $U = 66 \times 0.55 \times 0.057 = 2.1$. This is an underestimate, because some TEs are subject to selection (Brosius 2003) and we have disregarded insertion and deletion mutations, which occur at 0.05–0.1 the rate of single nucleotide mutations (Nachman and Crowell 2000; Kondrashov 2003) and are more likely to be deleterious. Furthermore, we have ignored adaptive mutations, which leads to an underestimate of the proportion of sites in genome that are subject to negative selection.

Our estimate of $U$ is similar to previous estimates, but this is largely coincidental, since those analyses generally considered only the rate of deleterious mutation in protein coding genes (Eyre-Walker and Keightley 1999; Nachman and Crowell 2000). If we calculate the deleterious mutation rate for protein-coding sequences using a recent estimate for the number of genes and the mutation rate we obtain a much smaller estimate. There are estimated to be ~20,000 genes in the human genome of average length 1500 bp; ~70% of mutations in protein coding genes are nonsynonymous and the mean level of constraint (*i.e.*, the proportion of the mutations that are deleterious) is estimated to be ~0.75 at nonsynonymous sites (Eory *et al.* 2010). This yields an estimate of 0.35 deleterious nonsynonymous mutations per diploid genome, which is substantially smaller than previous estimates (Eyre-Walker and Keightley 1999; Nachman and Crowell 2000), principally because recent estimates of the mutation rate and the number of protein coding loci are lower than previous estimates.

Our conservative estimate of $U = 2.1$, which includes mutations in coding and noncoding DNA, predicts that $\phi = 1 - e^{-2.1} = 88\%$ if mutations act independently; *i.e.*, 88% of the population is predicted to fail to reproduce as a consequence of recurrent deleterious mutation under a viability selection model (Equation 2). Furthermore, each individual would have to have an average of $1/(1 - 0.88) = 8.3$ offspring, and since there are two sexes in humans, each female would have to have at least 16 children to maintain the population size. Such a high frequency of genetic death is implausible in humans, particularly if many individuals fail to reproduce for nongenetic reasons. This is the mutation load paradox (Kondrashov and Crow 1993; Eyre-Walker and Keightley 1999; Nachman and Crowell 2000; Reed and Aquadro 2006; Barton *et al.* 2007; Charlesworth and Charlesworth 2010).

A number of factors can lead to a reduction in the mutation load (Agrawal and Whitlock 2012), two of which have been discussed in relation to the problem in humans. First, it has been suggested that many genetic deaths occur in the cell lineages leading to the gametes (Reed and Aquadro 2006) and prior to birth, since many pregnancies spontaneously abort at an early stage (Wang *et al.* 2004). However, this can explain only a small proportion of the mutation load, because the fraction of sites in the genome effectively selected in germ-line cell lineages is likely to be small, and most spontaneous abortions occur for nongenetic reasons or because of major genetic defects (Nagaishi *et al.* 2004), which are not included in our calculation of $\varphi$. Second, the mutation load can be reduced by synergistic epistasis, such that the combined effects of deleterious mutations are more severe than their independent effects (Kimura and Maruyama 1966; Crow and Kimura 1979). However, there is little empirical evidence that synergistic epistasis is more frequent than diminishing returns epistasis (Kouyos *et al.* 2007; Halligan and Keightley 2009), which has the opposite effect on the load.

It might also be argued that a load problem is unlikely to exist in humans if most selection acts on differences in

fertility. Under fertility selection, in which fitnesses are absolute, the load is the reduction in fertility relative to that of a mutation-free individual, not the proportion of individuals that fail to have descendants in the next generation. Defining $x$ as the number of offspring that a deleterious mutation-free individual can produce, the average number of offspring per individual is $z = x\,e^{-U}$, since the mean fitness of the population is $e^{-U}$. Each offspring has two parents, so $z = 2$ when the population size is stationary. Hence, if $x > 2/e^{-U}$, the population is expanding and potentially at a rate such that $\varphi \approx 0$. On the other hand if $x < 2/e^{-U}$ the population is contracting and $\varphi$ may approach 1 for small $x$. Therefore, the proportion of individuals that fail to have offspring depends both on the deleterious mutation rate and $x$. The rate of deleterious mutation that can be tolerated is therefore limited by $x$, but unfortunately, the value of $x$ is not known with any degree of certainty. To prevent population decline $x$ must be greater than 16 in humans if selection acts solely on absolute fertility differences. Agrawal and Whitlock (2012) have argued that $x$ may substantially exceed 16, since human family sizes can be large in modern societies and males can potentially have many offspring by mating with multiple females. However, reproductive potential may have been much more limited in ancestral human populations. In hunter–gatherer societies, which may have reproductive patterns similar to ancient hominid populations, females breastfeed their offspring for several years; this suppresses ovulation and leads to an average interbirth interval of approximately 3 years (Eaton *et al.* 1994). Since hunter–gatherer females typically reach menarche at ~16 years and menopause at ~47 years (Eaton *et al.* 1994), they have the potential to produce only ~11 children, and actual average family size is ~6 live births per female (Eaton *et al.* 1994). Since the ages of menarche, menopause, and weaning are probably under stabilizing selection in such populations, and close to their optima, it is difficult to envisage how $x$ could be much greater than 11 offspring for hunter–gatherer females. Fertility selection could potentially be stronger in males, if males can mate with several females. However, humans seem to have been largely monogamous, at least over the last million years (Labuda *et al.* 2010), and this trait is also likely to be under stabilizing selection. It therefore seems difficult to explain the mutation-load paradox in humans by assuming that selection acts largely on fertility, given what is known about human reproductive biology in extant populations.

Here we examine an alternative explanation for how humans can tolerate their high rate of harmful mutation. Wallace (1970) noted that the classic formulation of the load implicitly assumes that selection acts upon absolute fitness differences, such that the effect of a mutation in one individual is independent of the genotypes of other individuals in the population. Examples of mutations falling into this category are those that reduce cold tolerance or completely penetrant lethal mutations that knock out developmental pathways. However, Wallace argued that if selection occurs via competition between individuals within a species, then the proportion of individuals that fail to survive or reproduce depends on the variation in fitness between individuals, not the difference in fitness between the population mean and a deleterious mutation-free individual, as in Equation 1 (Crow 1970). The consequences of recurrent deleterious mutation for the magnitude of $\varphi$ might therefore be much smaller than suggested by Equation 2 under a relative fitness model. Similar arguments have been made by Sved *et al.* (1967) in relation to the number of balanced polymorphisms that can be maintained in a population, and by Ewens (1970) in relation to the substitution load.

Wallace (1970) argued that the proportion of individuals failing to reproduce is significantly reduced under a relative fitness model compared to the prediction from the classic calculation of the load, but he did not demonstrate this either theoretically or empirically. Here, we calculate the proportion of individuals that fail to have an adult descendant in the next generation under a relative fitness model. We refer to the fraction of nonreproducing individuals under this model as $\varphi_r$, and under the old definition of load under a viability selection model (Equation 2), as $\varphi_a$.

## Models

Consider a diploid organism with a genome containing $M$ loci, each subject to deleterious mutation at rate $u$. Assume that mutations are not completely recessive and that their fitness effects in heterozygous individuals ($s$) are sufficiently strong that mutant alleles segregate only in heterozygous form. Assuming free recombination between loci, the average frequency of a deleterious mutation is expected to be $u/s$. An individual will therefore carry $2Mu/s = U/s$ deleterious mutations, on average. Information on the rate and distribution of fitness effects of deleterious mutations suggests that $U/s > 20$ (Lohmueller *et al.* 2008; Charlesworth and Charlesworth 2010), so the number of deleterious mutations per individual is expected to be approximately normally distributed with a variance equal to its mean. Assuming that the fitness effects of mutations are multiplicative, then the fitness of an individual carrying $k$ mutations is $w(k) = (1 - s)^k$, and fitness is approximately lognormally distributed with a location parameter $\mu = U/s\ \mathrm{Log}(1 - s)$ and a squared scale parameter $\sigma^2 = U/s\ (\mathrm{Log}(1 - s))^2$.

### *Viability selection*

To calculate $\varphi_r$ under viability selection (*i.e.*, survival to reproductive age), assume that the population is censused at the zygote stage. Since, there is no selection on fertility, the proportion of individuals that survive viability selection is also the proportion of individuals that have descendants in the next generation, whether reproduction is monogamous or not. To enforce direct competition between individuals, we assume that the population size is stationary, as would be

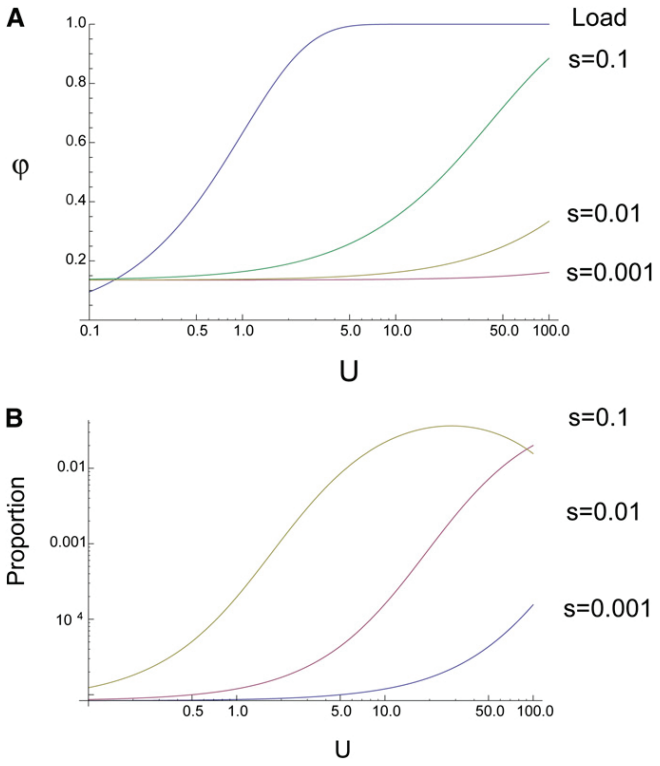



**Figure 1** The fraction of nonreproducing individuals, $\varphi_r$, (A) and the proportion of couples that have more than 10 offspring (B) plotted as a function of the deleterious mutation rate ($U$) and the strength of selection against a deleterious mutation ($s$).

the case, for example, if individuals compete for some finite resource that limits population size. An individual's fitness is then determined both by its own genotype and the genotypes of other individuals. If all individuals have the same fitness then individuals can fail to reproduce by chance. However, $\varphi_r$ increases if there is variation in fitness, because some individuals will have greater reproductive success than average, and others will have few or no surviving offspring. The fraction of nonreproducing individuals in this model can be calculated as follows. The proportion of offspring in the next generation contributed by a zygote with $k$ mutations is $w'(k) = w(k)/\bar{w}$, where $\bar{w}$ is the mean of $w(k)$, and the distribution of $w$' is lognormal with $\sigma^2 = U/\mathrm{s}\ (\mathrm{Log}(1-s))^2$. We assume that the population size is stationary, so the number of offspring to which an individual contributes is Poisson distributed with a mean of 2. The proportion of individuals leaving $x$ descendants is therefore

$$Q(x) = \int_0^\infty D(w')P(2w', x)dw, \qquad (3)$$

where $P(m, x)$ is the Poisson distribution with a mean of $m$ and $D(w')$ is the distribution of $w'$. The probability of a couple producing no offspring is $\varphi_r = Q(0)$.

Evaluation of Equation 3 shows that $\varphi_r$ increases as both the genomic deleterious mutation rate and the strength of selection on a new mutation increase (Figure 1A). However,

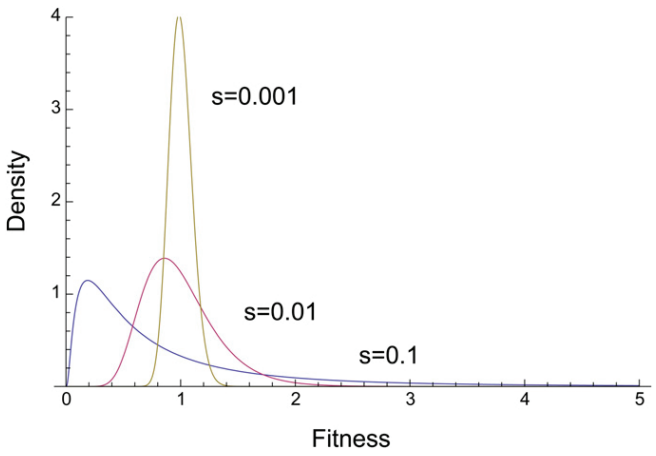



**Figure 2** The density of fitness for $U = 10$. Fitness has been normalized such that the mean is 1.

$\varphi_r$ is generally substantially lower than $\varphi_a$ (calculated using Equation 2), and approaches $\varphi_a$ only if selection is very strong and the deleterious mutation rate very high. For example, if we assume, unrealistically, that every new mutation in the human genome is deleterious (*i.e.*, $U = 66$) and $s = 1\%$ $\varphi_r$ is only 28% whereas $\varphi_a$ is close to 100%.

The fraction of nonreproducing individuals has a minimum value, which represents the probability that an individual has no offspring by chance alone. We assume a stationary population size, so the mean number of offspring per individual is 2, and therefore the chance of an individual having no offspring is $e^{-2} = 0.14$. This component is not included in $\varphi_a$. The proportion of nonreproducing individuals explained by selection alone is $(\varphi_r - e^{-2})/(1 - e^{-2})$, which is lower than $\varphi_r$ and hence even lower than $\varphi_a$ (see below).

Although the predicted proportion of nonreproducing individuals is small under a relative fitness model, it is important to check that the model does not predict the existence of super-fit individuals, since even the most successful individuals have limited reproductive potential. We investigated this by estimating the proportion of individuals that have >10 offspring by evaluating Equation 3 for a range of $U$ and $s$ values, summing the result for $x > 10$. It is evident that the proportion is generally small and consistent with levels of reproduction seen in humans (Figure 1B).

The proportion of individuals having no descendants in the next generation is smaller under a relative than absolute fitness model because $\varphi_r$ is determined by the variance in fitness among individuals, and this is generally small. Unless the deleterious mutation rate is very high and the selection strength against each deleterious mutation very strong, the model predicts that the fittest individuals (or couples) are not much fitter than the least fit individuals (Figure 2). For example, if $U = 10$ and $s = 0.01$ and we scale fitness to a mean of 1, then 97% of individuals have relative fitnesses between 0.5 and 2.

### *Other models of selection*

In the model described above, we calculated $\varphi_r$ assuming viability selection since $\varphi_a$ is equal to the mutation load under this model, and the consequences of recurrent deleterious mutation are therefore comparable under relative and absolute fitness models. However, it is also of interest to calculate $\varphi_r$ under a fertility selection model. If individuals are free to interbreed, rather than forming monogamous relationships, then the proportion of offspring produced by an individual with $k$ mutations is $w'(k) = w(k)/\bar{w}$ and relative fitness is lognormally distributed with a mean of 1 and a squared scale parameter of $\sigma^2 = U/s$ $(\mathrm{Log}(1 - s))^2$. Since we assume that the population size is stationary each individual will contribute to an average of two offspring in the next generation, so an individual with $k$ mutations will contribute to a Poisson distributed number of offspring with a mean of $2w'(k)$. The proportion of the population leaving $x$ offspring is therefore as given by Equation 3.

To investigate the consequences of monogamy let us assume that there is random mating and that the fertility of a couple is a function of the total number of deleterious mutations carried by the couple. In this case the proportion of offspring contributed by a couple to the next generation is $w'(k) = w(k)/\bar{w}$, which is lognormally distributed with a squared scale parameter $\sigma^2 = 2U/s$ $(\mathrm{Log}(1 - s))^2$. Since we assume that the population is stationary each couple is expected to have two offspring. The proportion of couples leaving $x$ offspring is therefore given by Equation 3, but $D(w')$ has a squared scale parameter of $\sigma^2 = 2U/s$ $(\mathrm{Log}(1 - s))^2$ rather than $\sigma^2 = U/s$ $(\mathrm{Log}(1 - s))^2$; *i.e.*, the mutation rate is effectively doubled by considering couples rather individuals.

For completeness, let us consider an asexual organism with discrete generations. Each generation, an individual can have several offspring, but the carrying capacity of the environment is such that the population is reduced to its former size before the next round of reproduction. Although asexual, we ignore the complication of Hill–Robertson interference, so the average frequency of a deleterious mutation is expected to be $u/s$ as above. As before, the contribution of an individual with $k$ mutations to the next generation is $w'(k) = w(k)/\bar{w}$. Hence relative fitness, $w'$, is lognormally distributed with a mean of one and a squared scale parameter $\sigma^2 = U/s$ $(\mathrm{Log}(1 - s))^2$. However, because there is no sex, an individual is expected to have only one adult descendent, on average, in the next generation rather than two if population size is stationary. The proportion of the population leaving $x$ offspring is therefore:

$$Q(x) = \int_0^\infty D(w')P(w', x)dw. \qquad (4)$$

Let us refer to the three models above as MF (monogamy with fertility selection), SEX (all models involving sex except MF) and ASEX (asexual). The proportion of nonreproducing individuals, $\varphi_r = Q(0)$, under these three models is shown in Figure 3. As expected, given the difference in the effective mutation rate between the two models, $\varphi_r$ under SEX is always lower than that under MF. The value of $\varphi_r$ under the ASEX model is generally higher than under either MF or SEX; this is largely due to the greater proportion of individuals having no offspring at low mutation rates under the ASEX model, since each individual is expected to have only one descendant, not two as under the SEX models. If we remove the effect of chance by calculating $\varphi_r$ attributable to selection alone as $(\varphi_r - e^{-2})/(1 - e^{-2})$ for the MF and SEX models and $(\varphi_r - e^{-1})/(1 - e^{-1})$ for the ASEX model, then we find $\varphi_r$ explained by selection is identical for the SEX and ASEX models and consistently lower than for the MF model (Figure 4).

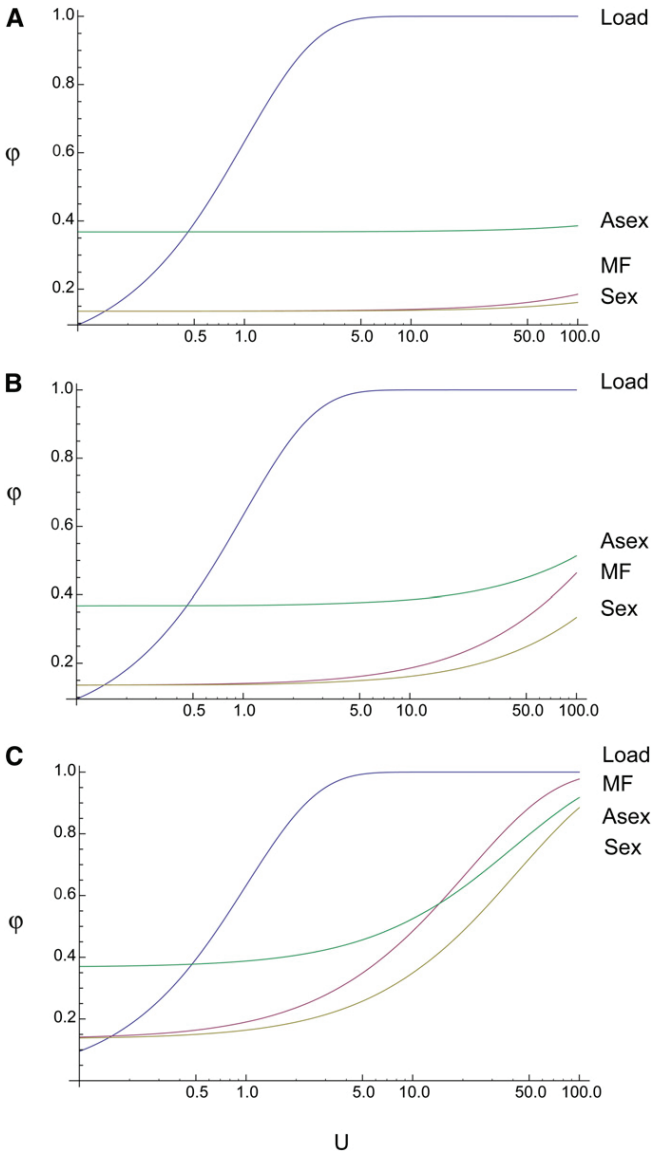



**Figure 3** The fraction of nonreproducing individuals, $\varphi_r$, under various models assuming strengths of selection: (A) $s = 0.001$, (B) $s = 0.01$, (C) $s = 0.1$.

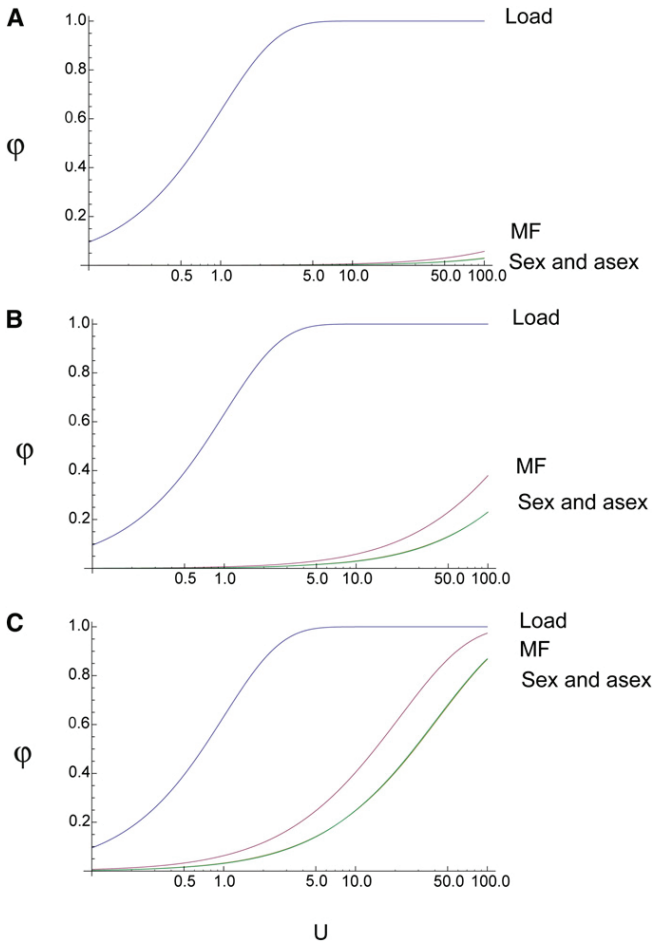



**Figure 4** The fraction of nonreproducing individuals, $\varphi_r$, with the effect of chance removed: (A) $s = 0.001$, (B) $s = 0.01$, (C) $s = 0.1$.

### *Recessive mutations*

The analysis above has assumed that mutations are not completely recessive and sufficiently strongly selected that we need consider only the selection against them when they are heterozygous. Let us now consider the value of $\varphi_r$ predicted under a model of recessive mutations. We consider a model in which selection is due to viability, or equivalently, fertility with free interbreeding. If the fitness of the three genotypes are 1, $1 - 2hs$, and $1 - 2s$ then Kimura (1964) has shown that the time that a new mutation spends at frequency $x$ is

$$f(x;S,h) = \frac{e^{-2Shx - S(1-2h)x^2}}{x(1-x)} \frac{\int_x^1 e^{2Shq + S(1-2h)q^2} dq}{\int_0^1 e^{2Shq + S(1-2h)q^2} dq}, \qquad (5)$$

where $S = 4N_e s$ and $N_e$ is the effective population size. To estimate $\varphi_r$, we need to know the expected number of loci for which an individual is homozygous for the recessive allele. This is

$$R_r(S,\Theta) = \Theta \int_{x=0}^{1} f(x;S,0)\, x^2 dx, \qquad (6)$$

where $\Theta = 4MN_e u$ and $M$ is the number of loci.

The expected number of loci that are expected to be heterozygous for a semidominant mutation is

$$R_s(S,\Theta) = \Theta \int_{x=0}^{1} f(x;S,1/2) 2x(1-x)\, dx, \qquad (7)$$

which is approximately $2U/s = 2\Theta/S$. Evaluation of Equations 6 and 7 suggests that the average number of homozygous recessive loci is between 25% and 50% of the number of heterozygous semidominant loci (assuming equal numbers of loci and mutation rates) (Table 1). This can also be seen by an analytical approximation. The average frequency of a deleterious recessive mutation is approximately $u\sqrt{\pi N_e/s} = \theta\sqrt{\pi}/2\sqrt{S}$, where $\theta = 4N_e u$ and $S = 4N_e s$, if $\theta \ll 1$ (Nei 1968) (note the classic formula $\sqrt{u/s}$ applies only in infinite populations). Hence the average frequency of each deleterious mutation introduced into the population is $\sqrt{\pi}/2\sqrt{S}$, so the expected frequency of the homozygous genotype for each of these mutations is approximately $\pi/4S$ and the expected number of loci that are homozygous is

$$R_r'(S,\Theta) = \Theta\pi/4S. \qquad (8)$$

Comparing this against the expected number of semidominant loci that are heterozygous suggests that we expect approximately $8/\pi = 2.5$ times more sites to be heterozygous for semidominant mutations than homozygous for recessive mutations (Table 1).

The contribution of an individual with $z$ loci that are homozygous for a deleterious recessive to the next generation is $w'(z) = w(z)/\bar{w}$. Hence the relative fitness, $w'$, is lognormally distributed with a mean of one and a squared scale parameter $\sigma^2 = R_r(\Theta,S)\,(\mathrm{Log}(1 - 2s))^2$. Since we assume that the population size is stationary, each individual will

**Table 1 The expected number of homozygous and heterozygous loci, when $\Theta = 1$, for recessive and semidominant mutations respectively**

| $S$ | No. of homozygous recessive loci (Equation 6) | Approximate no. of homozygous recessive loci (Equation 8) | No. of heterozygous semidominant loci (Equation 7) | Ratio (column 2/4) |
|---|---|---|---|---|
| 0.01 | 0.50 | 79.0 | 1.0 | 0.50 |
| 0.1 | 0.49 | 7.9 | 0.98 | 0.50 |
| 1 | 0.44 | 0.78 | 0.84 | 0.52 |
| 10 | 0.073 | 0.078 | 0.20 | 0.37 |
| 100 | 0.0055 | 0.0079 | 0.020 | 0.28 |
| 1000 | 0.00051 | 0.00078 | 0.0020 | 0.26 |

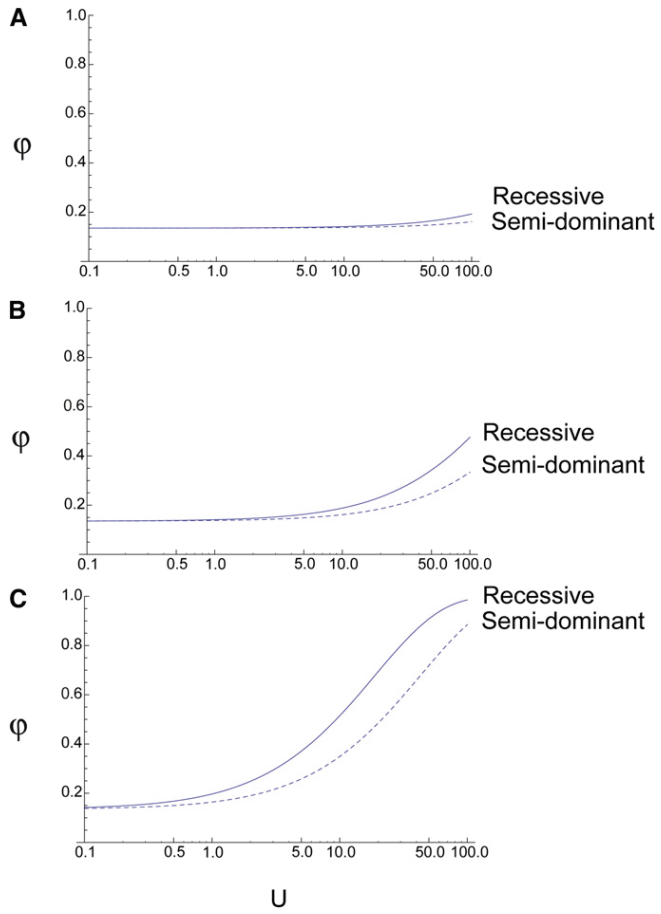



**Figure 5** The fraction of nonreproducing individuals, $\varphi_r$, under models with semidominant and completely recessive mutations for $N_e$ = 10,000. (A) $s$ = 0.001, (B) $s$ = 0.01, (C) $s$ = 0.1.

contribute to two offspring in the next generation and so the proportion of the population leaving $x$ offspring is

$$Q(x) = \int_0^\infty D(w')P(2w', x)\,dw, \qquad (9)$$

where $P(m, x)$ is the Poisson distribution with a mean of $m$ and $D(w')$ is the distribution of $w'$.

The value of $\varphi_r$, predicted under a model of recessive mutations (Equation 9), is compared to $\varphi_D$ under a model of semidominant mutations (Equation 3), in Figure 5. From this it can seen that with recessive mutations $\varphi_r$ is somewhat higher than under a model with semidominant mutations for the same rate of mutation for $N_e$ = 10,000. The situation can be reversed if $N_e$ is much smaller, but the $\varphi_r$ is always quite similar.

### *Simulations*

We ran a series of simulations to check our analytical approach. A population of $N$ diploid individuals with $M$ independent loci was subject to recurrent deleterious mutation at a rate $u$ per locus, such that $2Mu = U$. The fitness of each individual with $k$ mutations was calculated as $(1 - s)^k$. In each generation we randomly selected pairs of individuals in proportion to their relative fitnesses (*e.g.*, if we had four individuals with absolute fitnesses of 0.1, 0.2, 0.3, and 0.2 we would select the first individual on average 0.1/(0.1 + 0.2 + 0.3 + 0.2) = 0.125 of the time to mate). Each mating produced one offspring, with alleles drawn at random from the parental genomes (*i.e.*, assuming free recombination). This process of selecting individuals to form pairs was repeated until $N$ offspring had been produced; individuals could contribute to multiple matings. The value of $\varphi_r$ in the simulations was very close to that expected from Equation 3 suggesting that our analytical derivation of $\varphi_r$ was satisfactory (Table 2).

## Discussion

We have shown that the proportion of individuals that fail to have descendants in the next generation under a relative fitness model is substantially lower than that predicted under an absolute viability fitness model and that species could potentially survive a mutation rate of 10's if not 100's of deleterious mutations per genome per generation if selection was largely mediated through competition.

The fraction of nonreproducing individuals ($\varphi$) depends on both the rate of deleterious mutation and the strength of

**Table 2 Simulations under a relative fitness model**

| $S$ | $U$ | $\phi_r$ (theory) | $\phi_r$ (simulated)(SE) | Observed average frequency over expected (SE) |
|---|---|---|---|---|
| 0.01 | 0.1 | 0.136 | 0.135 (0.000) | 1.03 (0.00) |
| | 1 | 0.138 | 0.138 (0.000) | 1.03 (0.00) |
| | 2 | 0.141 | 0.141 (0.000) | 1.03 (0.00) |
| | 5 | 0.149 | 0.148 (0.000) | 1.03 (0.00) |
| | 10 | 0.161 | 0.161 (0.000) | 1.04 (0.00) |
| 0.1 | 0.1 | 0.138 | 0.138 (0.000) | 1.00 (0.00) |
| | 1 | 0.164 | 0.163 (0.000) | 1.00 (0.00) |
| | 2 | 0.190 | 0.188 (0.000) | 1.01 (0.00) |
| | 5 | 0.258 | 0.254 (0.000) | 1.02 (0.00) |
| | 10 | 0.349 | 0.342 (0.000) | 1.06 (0.00) |

The table gives the fraction of nonreproducing individuals, $\varphi_r$ , along with its theoretical prediction, and the average frequency of a deleterious mutations divided by its expected value under an absolute fitness model (*i.e.*, $u/s$). The simulations were run with a population size of 1000 and 100,000 sites.

selection acting on deleterious mutations. The mean fitness effect of a deleterious mutation in humans is unknown, but mutation accumulation experiments in other animals and plants suggest that nonlethal mutations have fitness effects of at most 1–20% (Keightley and Halligan 2009). However, such estimates are upwardly biased because they are generally made under the unrealistic assumption that mutations have equal selective effects, implying that the true mean value of $s$ is likely to be substantially lower. Even assuming $s$ as high as 20%, $\varphi_r$ would be only 51% if $U = 5$. The load from lethal mutation is expected to be much lower than that for nonlethals, since lethal mutations have been estimated to occur at about one-hundredth the rate of nonlethals (Crow and Simmons 1983).

The extent to which selection is mediated through competition between conspecifics is unknown. If individuals compete for resources or mates, and competitive ability is genetically determined, then the success of an individual will depend both on its own genotype and the genotypes of its competitors. This might suggest that there is epistasis generated in a relative fitness model, and it has been shown that the mutation load can be substantially reduced if there is synergistic epistasis (Kimura and Maruyama 1966). However, synergistic epistasis is not expected to be a feature of our model, since the contribution of a genotype with $k$ mutations to the next generation is $w'(k) = w(k)/\bar{w}$, so $\log(w'(k))$ is linear with respect to $k$. To check that epistasis is not an emergent property of our model we tabulated the number of offspring produced in our simulation (see above). As expected, the log of the mean number of offspring produced by individuals with $k$ mutations is linearly related to $k$ (Figure 6A), demonstrating that epistasis does not emerge within this model. We also kept track of the mean absolute fitness of the population within the simulation. As expected, the mean absolute fitness is $e^{-U}$. If synergistic epistasis had been present then we would expect the mean absolute fitness to be higher than this expected value (Figure 6B).

It has been suggested that sexual reproduction might be maintained because sexual species can have substantially lower mutation loads than asexual species if there is synergistic epistasis (Kimura and Maruyama 1966). If $U > 1$ this can be sufficient to offset the twofold cost of sex (Kondrashov 1982; Kondrashov 1988). This is known as the deterministic mutation hypothesis. However, since the overall effect of recurrent deleterious mutation on population fitness is considerably reduced, if selection is mediated by competition, it is likely that the conditions under which sexual species have an advantage will also be greatly reduced.

The consequences of recurrent deleterious mutation for the proportion of the population that fails to reproduce is less extreme under a relative compared to an absolute fitness model. One might therefore expect natural selection to be weaker under a relative fitness model and that deleterious mutations would accumulate in the population. However, this is not the case: in our simulation the average

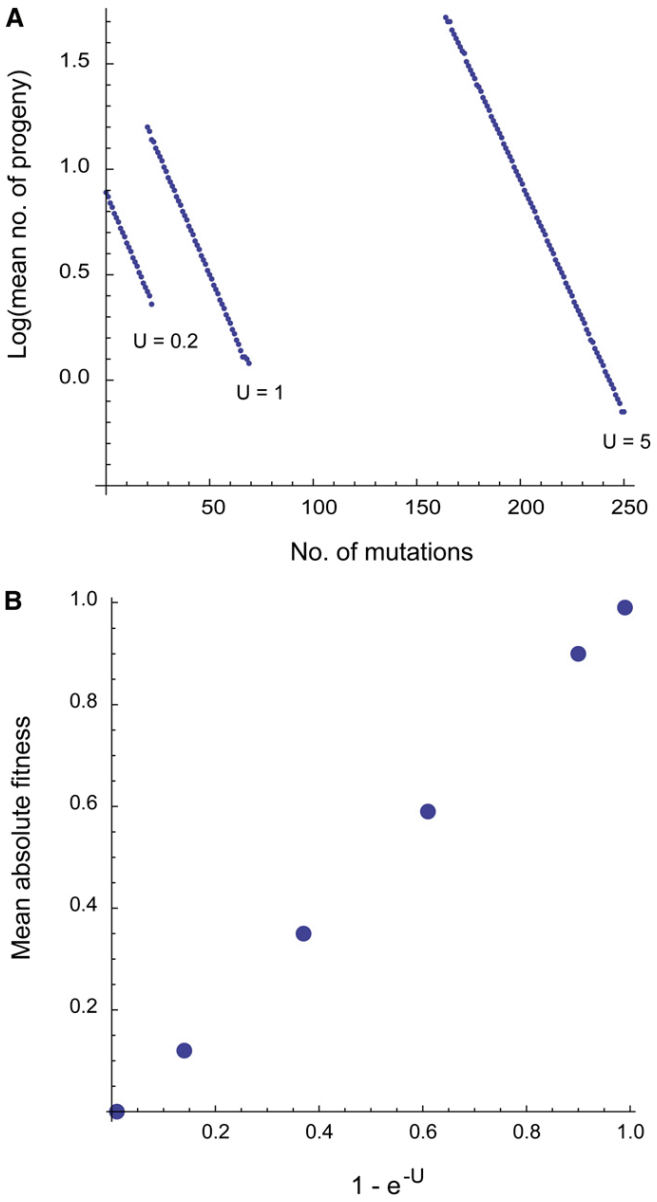



**Figure 6** Individual and population fitness in a simulated population under a relative fitness model. (A) The relationship between the log mean progeny number and the number of mutations for three populations subject to genomic deleterious mutation rates of 1, 2, and 5. (B) The relationship between the mean absolute fitness of a population and the expected absolute fitness, $e^{-U}$ for different values of $U$.

frequency of a deleterious mutation is close to the value expected under an absolute fitness model (Table 2). The average frequency of a deleterious mutation is very slightly higher than we expect, but this is likely to be due to Hill–Robertson interference.

Following Wallace (1970), we have shown that the fraction of individuals that fail to reproduce as a consequence of recurrent deleterious mutation depends on whether selection is mediated via absolute or relative differences between individuals. The classic definition of the mutation load equates

to the fraction of nonreproducing individuals if selection acts on absolute differences under a viability selection model, as would be the case if the fitness of a genotype were independent of the genotypes of conspecifics. If fitness depends on the genotypes of conspecifics, then the proportion of nonreproducing individuals depends on the distribution of fitness among individuals and tends to be much lower than predicted by the absolute mutation load. Evaluation of our model, assuming plausible values for the genomic deleterious mutation rate and strength of selection against a new mutation, suggests that the proportion of individuals that fail to reproduce is much lower than predicted by the classic formula for the absolute load, and there is no requirement for some individuals to be unrealistically fecund. Our analytical results and simulations suggest a resolution of the mutation load paradox by showing that a very high number of deleterious mutations can be eliminated from the population each generation and that the population can still be viable. Our results also demonstrate that one mutation does not necessarily result in one genetic death.

## Acknowledgments

The authors are grateful to Austin Burt, Michael Whitlock, Aneil Agrawal, Alex Kondrashov, Liz Somerville, and anonymous referees for helpful discussion and comments.

## Literature Cited

Agrawal, A., and M. C. Whitlock, 2012 Mutation load: the fitness of individuals in populations where deleterious mutations are abundant. Annu. Rev. Ecol. Evol. Syst. 43: (in press).

Awadalla, P., J. Gauthier, R. A. Myers, F. Casals, F. F. Hamdan *et al.*, 2010 Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am. J. Hum. Genet. 87: 316–324.

Barton, N. H., D. E. G. Briggs, J. A. Eisen, D. B. Goldstein, and N. H. Patel, 2007 *Evolution*. Cold Spring Harbor Laboratory Press, New York.

Brosius, J., 2003 The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118: 99–116.

Charlesworth, B., and D. Charlesworth, 2010 *Elements of Evolutionary Genetics*. Ben Roberts, Greenwood Village, CO.

Cordaux, R., and M. A. Batzer, 2009 The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10: 691–703.

Crow, J. F., 1970 Genetic loads and the cost of natutal selection. *Mathematical Topics in Population Genetics*, pp. 128–177. Springer-Verlag, New-York.

Crow, J. F., and M. Kimura, 1979 Efficiency of truncation selection. Proc. Natl. Acad. Sci. USA 76: 396–399.

Crow, J. F., and M. J. Simmons, 1983 The mutation load in Drosophila, pp. 1–26 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner, H. L. Carson, and J. L. Thompson. Academic Press, London.

Durbin, R. M., G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks *et al.*, 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

Eaton, S., M. Pike, R. Short, N. Lee, J. Trussell *et al.*, 1994 Women's reproductive cancers in evolutionary context. Q. Rev. Biol. 69: 353–367.

Eory, L., D. L. Halligan, and P. D. Keightley, 2010 Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol. Biol. Evol. 27: 177–192.

Ewens, W. J., 1970 Remarks on the substitutional load. Theor. Popul. Biol. 1: 129–139.

Eyre-Walker, A., and P. D. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature 397: 344–347.

Haldane, J. B. S., 1937 The effect of variation in fitness. Am. Nat. 71: 337–349.

Halligan, D. L., and P. D. Keightley, 2009 Spontaneous mutaion accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. 40: 151–172.

Keightley, P. D., and D. L. Halligan, 2009 Analysis and implications of mutational variation. Genetica 136: 359–369.

Kimura, M., 1964 Diffusion models in population genetics. J. Appl. Probab. 1: 177–232.

Kimura, M., and T. Maruyama, 1966 The mutational load with epistatic gene interactions in fitness. Genetics 54: 1337–1351.

Kondrashov, A. S., 1982 Selection against harmful mutations in large sexual and asexual populations. Genet. Res. 40: 325–332.

Kondrashov, A. S., 1988 Deleterious mutations and the evolution of sexual reproduction. Nature 336: 435–440.

Kondrashov, A. S., 2003 Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum. Mutat. 21: 12–27.

Kondrashov, A. S., and J. F. Crow, 1993 A molecular approach to estimating the human deleterious mutation rate. Hum. Mutat. 2: 229–234.

Kouyos, R. D., O. K. Silander, and S. Bonhoeffer, 2007 Epistasis between deleterious mutations and the evolution of recombination. Trends Ecol. Evol. 22: 308–315.

Labuda, D., J. F. Lefebvre, P. Nadeau, and M. H. Roy-Gagnon, 2010 Female-to-male breeding ratio in modern humans-an analysis based on historical recombinations. Am. J. Hum. Genet. 86: 353–363.

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker *et al.*, 2011 A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997.

Lunter, G., C. P. Ponting, and J. Hein, 2006 Genome-wide identification of human functional DNA using a neutral indel model. PLOS Comput. Biol. 2: e5.

Meader, S., C. P. Ponting, and G. Lunter, 2010 Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 20: 1335–1343.

Mouse Genome Sequencing Consortium, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Muller, H. J., 1950 Our load of mutations. Am. J. Hum. Genet. 2: 111–176.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.

Nagaishi, M., T. Yamamoto, K. Iinuma, K. Shimomura, S. A. Berend *et al.*, 2004 Chromosome abnormalities identified in 347 spontaneous abortions collected in Japan. J. Obstet. Gynaecol. Res. 30: 237–241.

Nei, M., 1968 The frequency distribution of lethal chromosomes in finite populations. Proc. Natl. Acad. Sci. USA 60: 517–524.

Reed, F. A., and C. F. Aquadro, 2006 Mutation, selection and the future of human evolution. Trends Genet. 22: 479–484.

Reed, F. A., J. M. Akey, and C. F. Aquadro, 2005 Fitting background-selection predictions to levels of nucleotide variation

and divergence along the human autosomes. Genome Res. 15: 1211–1221.
Roach, J. C., G. Glusman, A. F. Smit, C. D. Huff, R. Hubley *et al.*, 2010 Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–639.
Sved, J., T. E. Reed, and W. F. Bodmer, 1967 The number of balanced polymrophisms that can be maintained in a natural population. Genetics 55: 469–471.
Wallace, B., 1970 *Genetic Load: Its Biological and Conceptual Aspects*. Prentice-Hall, Englewood Cliffs, NJ.
Wang, J. X., R. J. Norman, and A. J. Wilcox, 2004 Incidence of spontaneous abortion among pregnancies produced by assisted reproductive technology. Hum. Reprod. 19: 272–277.

*Communicating editor: L. M. Wahl*

## B.2. XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells

*Voir article joint page suivante.*
Référence : [Vallot et al., 2013]

# *XACT*, a long noncoding transcript coating the active X chromosome in human pluripotent cells

Céline Vallot[1,2], Christophe Huret[1,2], Yann Lesecque[3], Alissa Resch[4], Noufissa Oudrhiri[5], Annelise Bennaceur-Griscelli[5], Laurent Duret[3] & Claire Rougeulle[1,2]

**X-chromosome inactivation (XCI) in mammals relies on *XIST*, a long noncoding transcript that coats and silences the X chromosome in *cis*. Here we report the discovery of a long noncoding RNA, *XACT*, that is expressed from and coats the active X chromosome specifically in human pluripotent cells. In the absence of *XIST*, *XACT* is expressed from both X chromosomes in humans but not in mice, suggesting a unique role for *XACT* in the control of human XCI initiation.**

A major recent breakthrough in the conception of eukaryotic gene regulation has been the identification of a multitude of long noncoding RNAs (lncRNAs) in mammals that are scattered throughout the genome and are located, in particular, in intergenic regions (lincRNAs)[1]. Although the number of lncRNAs is estimated to be more than 1,000, only 12 have been functionally characterized. These lncRNAs were shown to function in diverse cellular processes, most predominantly in the regulation of gene expression. In this context, many lncRNAs participate in gene silencing pathways, some of which involve the recruitment of repressive chromatin complexes[2].

XCI is one of the most studied processes involving lncRNA-mediated repression. *XIST* is a noteworthy transcript in that in addition to silencing an (approximate) entire chromosome, it is the only lncRNA described thus far to widely coat the chromosome from which it is expressed. This coating by *XIST* induces substantial nuclear reorganization and recruitment of histone-modifying complexes that are important for the initiation and maintenance, respectively, of X-chromosome silencing[3].

XCI is established early during embryonic development, and embryonic stem cells can be used to decipher the kinetics and molecular actors of the process. In female human embryonic stem cells (hESCs), one X chromosome is, in most cases, already inactivated[4]. Using an RNA sequencing (RNA-seq) analysis of female H9 hESCs (C.V., Ouimette, J.-F., Makhlouf, M., Féraud, O., N.O., A.B.-G., Côme, J., Martinat, C., A.R., Lalande, M. & C.R., unpublished data), we identified a large unannotated region (~252 kb) on the X chromosome for which we could detect a substantial amount of expression originating from the minus strand (**Fig. 1a**). We called this transcript *XACT*. The *XACT* region is located on chromosome Xq23 between the protein-coding genes *AMOT* and *HTR2C* in an unusually large intergenic domain of 1.7 Mb (only 1% of intergenic regions in humans are >1.5 Mb). This domain is GC poor (37%) and rich in repeat sequences (68% compared to 62% for the X chromosome and 45% for the overall human genome), notably long interspersed nuclear elements and long terminal repeats (LTRs) (**Supplementary Fig. 1**). All RNA-seq reads were co-linear with the genomic sequence, suggesting that the *XACT* transcript is unspliced (**Supplementary Methods**). We detected similar transcription in male H1 hESCs, and alignment with published chromatin immunoprecipitation sequencing data[5] revealed histone H3 Lys36 trimethylation (H3K36me3) and histone H3 Lys79 dimethylation (H3K79me2) blocks along the region, as well as peaks of H3K4me3 and RNA polymerase II (PolII) at the telomeric extremity of the transcribed region (**Supplementary Fig. 2a**). These peaks correspond to LTR elements, which might serve as promoters for this transcribed region[6]. We further characterized the 5′ end of *XACT* by RT-PCR and 5′ RACE (**Supplementary Fig. 2b**). This analysis revealed multiple transcription start sites (TSSs) clustered in a 126-bp region characteristic of broad-type promoters[7], which coincide with peaks of H3K4me3 and RNA PolII (**Supplementary Fig. 2**). Together these data suggest that *XACT* corresponds to a single 251.8-kb transcription unit (112,983,323–113,235,148 bp). The *XACT* transcript is polyadenylated and mostly nuclear (**Fig. 1b**), further suggesting that it acts, at least in part, as a noncoding RNA. We also identified a distinct transcript 5′ of *XACT* on the plus strand called *T113.3* (**Fig. 1a** and **Supplementary Fig. 3**). Unlike *XACT*, *T113.3* is spliced and mostly cytoplasmic, and its TSS, mapped by 5′ RACE, lies within a peak of H3K4me3 located 48 kb upstream of the TSS of *XACT*.

We next investigated transcription of *XACT* at the cellular level by RNA FISH. Remarkably, a BAC probe covering 151 kb of the transcribed region detected a unique, large signal in female H9 hESCs that was reminiscent of the *XIST* RNA cloud. However, this signal corresponded to the active X chromosome, as determined by combined RNA FISH with a *XIST* probe, which labels the inactive X chromosome (**Fig. 1c**), and simultaneous RNA and DNA FISH with an X-paint probe (**Fig. 1d**). The *XACT* RNA cloud signal was partially resistant to the stringent denaturation steps used in the DNA FISH experiments (**Fig. 1d**; note that the *XACT* signal appears smaller than it does in classical RNA FISH experiments) but not to RNase treatment (**Fig. 1e**), indicating a strong association of the RNA with the active X chromatin. The BAC probe detected two pinpoints in the DNA FISH

[1]Université Paris Diderot, Sorbonne Paris Cité, Epigenetics and Cell Fate, Paris, France. [2]Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 7216 Epigenetics and Cell Fate, Paris, France. [3]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université de Lyon, Université Lyon 1, Villeurbanne, France. [4]Stem Cell Institute, University of Connecticut Health Center, Farmington, Connecticut, USA. [5]ESTeam Paris Sud, Institut National de la Santé et de la Recherche Médicale (INSERM) U935, Université Paris Sud 11, Assistance Publique–Hôpitaux de Paris (AP-HP), Villejuif, France. Correspondence should be addressed to C.R. (claire.rougeulle@univ-paris-diderot.fr).

Received 18 July 2012; accepted 20 December 2012; published online 20 January 2013; doi:10.1038/ng.2530

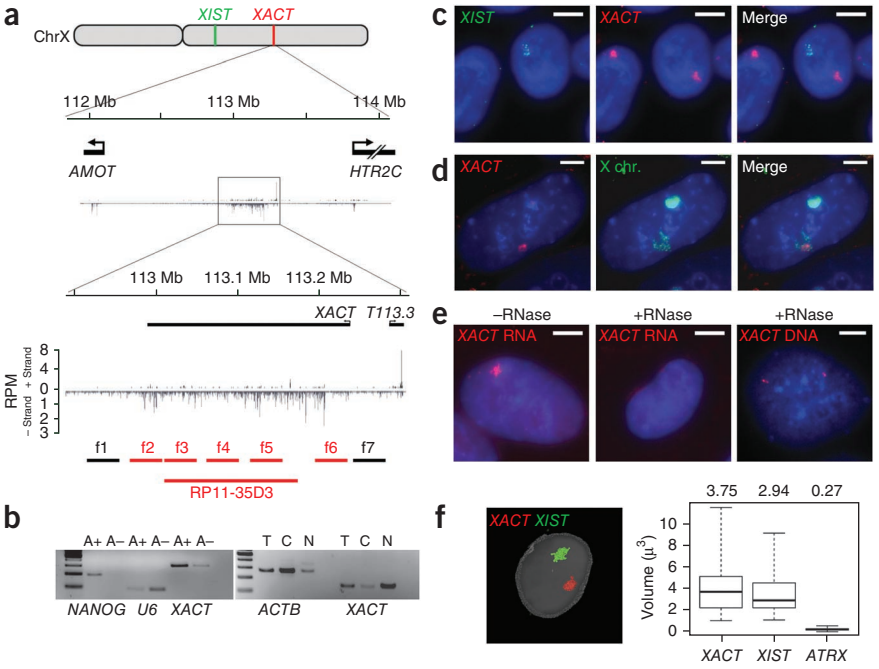



**Figure 1** Characterization of *XACT*, a long intergenic RNA that coats the active X chromosome in hESCs. (**a**) Schematic map showing the localization of *XACT* and *XIST* on the human X chromosome. Bottom, strand-specific RNA-seq showing >250 kb of continuous transcription originating from the minus strand in female H9 cells. The boxed area is a magnification of the locus. The locations of the BAC probe (RP11-35D3) and fosmid probes (f1–f7) used for further FISH analysis are indicated below (**Supplementary Fig. 4**). Probes giving an RNA cloud signal for *XACT* are in red, and probes not giving a signal are in black. RPM, reads per million mapped reads. (**b**) Left, semiquantitative RT-PCR analysis of poly(A)-positive (A+) and poly(A)-negative (A–) enriched RNA fractions of H9 hESCs. *NANOG* and *U6* were used as positive controls for the poly(A)-positive and poly(A)-negative fractions, respectively. Right, RT-PCR analysis of total (T), cytoplasmic (C) and nuclear (N) RNAs. (**c**) RNA FISH analysis of H9 cells with a *XIST* probe (green) and a *XACT* probe (red). All *XACT* FISH experiments were performed using the BAC RP11-35D3 probe, except where otherwise indicated. (**d**) Simultaneous RNA and DNA FISH of H9 cells with the *XACT* probe (red) detecting *XACT* RNA and DNA and an X-paint probe (green) detecting X chromosomes. (**e**) Left and middle, *XACT* RNA FISH with and without initial treatment with RNase. Right, *XACT* DNA FISH with initial treatment with RNase. Scale bars (**c**–**e**), 5 μm. (**f**) Three-dimensional model of the nuclear volumes occupied by *XACT* and *XIST* RNAs. The boxplot represents the distribution of the volumes ($n = 50$ nuclei). The distribution of the volumes corresponding to the transcription signal of an X-linked gene (*ATRX*) is shown for comparison. The median value of each distribution is indicated above the corresponding boxplot.

experiments (after RNase treatment of the slides), which confirms that the RNA cloud signal corresponds to actual coating of the chromosome by *XACT* RNA and not to the genomic organization of the locus. We called this RNA *XACT* (X active coating transcript). The nuclear volume occupied by *XACT* in the nucleus was similar to that occupied by *XIST* (median volumes of 3.75 $\mu m^3$ and 2.94 $\mu m^3$, respectively; **Fig. 1f**). RNA FISH analysis of *XACT* expression using a series of fosmid probes covering the region (**Fig. 1a**) further confirmed the extent and expression profile of *XACT* (**Supplementary Fig. 4**).

X inactivation is highly unstable in female hESCs, and *XIST* downregulation tends to occur spontaneously in culture[8]. Loss of *XIST* expression leads to substantial but incomplete gene reactivation from the originally inactive X chromosome (Xi*)[9] (C.V., Ouimette, J.-F., Makhlouf, M., Féraud, O., N.O., A.B.-G., Côme, J., Martinat, C., A.R., Lalande, M. & C.R., unpublished data). RNA-seq performed in H9 hESCs not expressing *XIST* revealed a slight increase in *XACT* expression compared to H9 cells that did express *XIST* (**Fig. 2a**; reads per kilobase per million reads (RPKM) = 1.78 in cells expressing *XIST*, and RPKM = 2.37 in cells not expressing *XIST*). We took advantage of a SNP (rs5929175) within *XACT* and the clonal X-inactivation pattern shown by H9 cells[10,11] to address the allelic expression of *XACT* in cells not expressing *XIST* compared to *XIST*-expressing cells. Whereas *XACT* was monoallelically transcribed in cells expressing *XIST*, it became biallelically expressed in cells not expressing *XIST* (**Fig. 2b**). RNA FISH analysis further revealed that *XACT* is not only re-expressed from but also coats the Xi* in cells not expressing *XIST*, leading to two *XACT* clouds in these cells (**Fig. 2b**). Whether *XACT* re-expression from and coating of the Xi* is a cause or a consequence of *XIST* repression and subsequent partial reactivation of the chromosome remains to be established.

*XACT* is expressed from active X chromosome(s) in hESCs. In contrast, we were not able to detect *XACT* reads from female fibroblast RNA-seq libraries. Similarly, RT-PCR analysis revealed no or weak expression of *XACT* in various tissues, including brain, muscle and placenta (**Supplementary Fig. 5**), suggesting that *XACT* is downregulated after differentiation. To investigate the kinetics of *XACT* silencing, we induced H9 hESCs to differentiate and found substantial downregulation of *XACT* at day 5 and full silencing at day 10 (**Fig. 2c**). To further probe the link between *XACT* and pluripotency, we used a model system in which H9 hESCs not expressing *XIST* were subjected to several rounds of differentiation and reprogramming[12] (**Fig. 2d**). As detected with undirected differentiation, the differentiation of these H9 hESCs into mesenchymal stem cells (MSCs) was accompanied by biallelic silencing of *XACT*. Remarkably, we found strong re-expression of *XACT* from both X chromosomes in MSC-derived induced pluripotent stem (iPS) cells, whereas silencing of *XACT* was re-established when these iPS cells were differentiated into MSCs (iPS-MSCs). Together these results indicate that, in this context, *XACT* expression and coating of the X chromosomes is restricted to pluripotent and early differentiating cells in humans. However, we cannot exclude the possibility that *XACT* is expressed in some differentiated cell types.

We then investigated the conservation of *XACT* by combining *in silico* analyses and experimental studies (**Supplementary Fig. 6**). The organization of the genomic region encompassing *AMOT* and *HTR2C* is well conserved in placental mammals and marsupials. In contrast, the sequence between these two genes shows moderate conservation, with several conserved blocks present in placental mammals but not marsupials, which are interspaced with divergent regions. The LTR that corresponds to the 5′ end of *XACT* is conserved in chimpanzees but not macaques or more distally related species (**Supplementary Fig. 6a**). This suggests that the insertion of these LTR elements is a very recent event. To further probe *XACT* conservation in mice, we undertook systematic DNA and RNA FISH studies in mouse ESCs (mESCs) using a series of six BAC probes spanning the region between *Amot* and *Htr2c*. Although all these probes detected

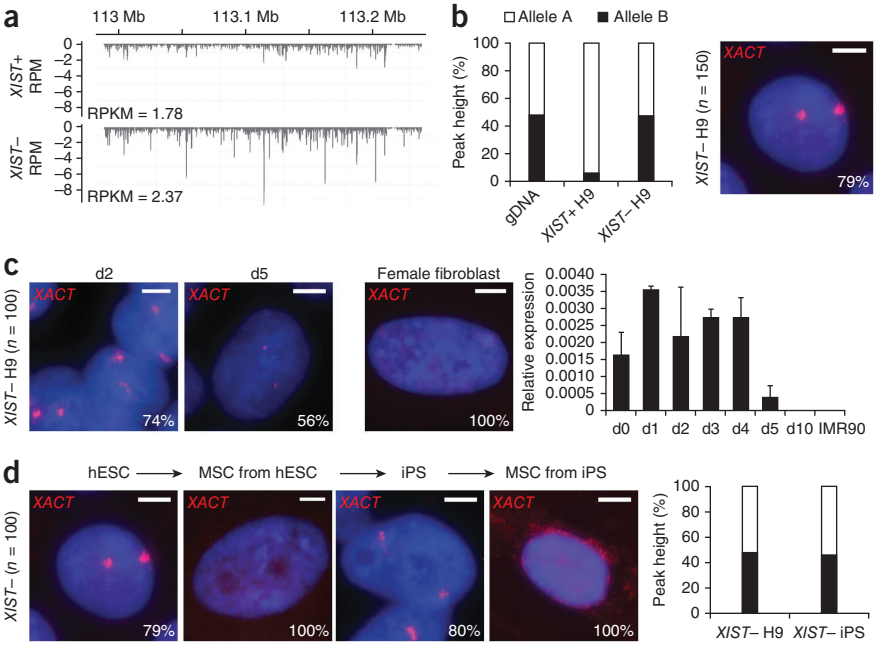



**Figure 2** *XACT* coats both X chromosomes in hESCs not expressing *XIST*, and the expression of *XACT* is restricted to pluripotent cells. (**a**) RNA-seq data for H9 hESCs expressing (*XIST*+) or not expressing (*XIST*–) *XIST* with RPKM values shown. (**b**) Left, pyrosequencing analysis of rs5929175 for H9 genomic DNA (gDNA) and *XIST*+ and *XIST*– complementary DNA. The bar chart indicates the percentage of the peak height corresponding to each allele. Right, RNA FISH using the *XACT* probe in *XIST*– H9 hESCs revealing that 79% of the nuclei ($n = 150$) have two *XACT* RNA clouds. (**c**) RNA FISH and quantitative RT-PCR analysis of *XACT* expression during differentiation of *XIST*– hESCs and in female fetal (IMR90) and adult (CoriellI, AG09603) fibroblasts. d0–d10, days 0–10. The images are representative of the major population. The bar charts correspond to the average of two independent differentiation experiments. Errors bars indicate the s.d. calculated for two samples. (**d**) RNA FISH and pyrosequencing analysis of *XACT* expression during successive rounds of differentiation and reprogramming of *XIST*– H9 cells into MSCs. Scale bars (**b**–**d**), 5 µm. The percentages shown in **b**–**d** indicate the number of nuclei showing similar expression patterns of *XACT*.

the two X chromosomes by DNA FISH, none of them generated a signal by RNA FISH (**Supplementary Fig. 6b**), suggesting that the region syntenic to *XACT* is not expressed in mESCs. In agreement with this result, analysis of mESC RNA-seq data did not reveal broad, '*XACT*-like' transcription in the syntenic region. However, we were able to identify a few discrete peaks in mESCs and other cell types (**Supplementary Fig. 6b**). Whether these peaks correspond to real transcripts remains to be investigated. Together our data suggest that there is no *XACT*-like transcript expressed in mESCs.

We have identified *XACT* as the first lncRNA that coats the active X chromosome in humans. The identification of a lncRNA that coats an active chromosome (whereas most lncRNAs studied so far are involved in gene silencing) underlies the multifaceted nature of lncRNAs. *XACT* might not be conserved in mice, and its function might be related to the specific kinetics of XCI that were recently described in the human. Indeed, in human preimplantation embryos, *XIST* is expressed from the paternal and maternal X chromosomes, but this does not lead to chromosome-wide silencing[13]. In contrast, paternal *Xist* expression and XCI characterize mouse preimplantation development[14]. Given its expression profile, it is tempting to speculate that *XACT* is involved in the control of XCI initiation in humans. More generally, rapidly evolving lncRNAs such as *XACT* in the human and *Tsix* in the mouse[15] could be involved in species-specific regulation of XCI, thus underlying the important plasticity that characterizes this process among mammals.

**Accession codes.** Sequence data and alignments have been submitted to the Gene Expression Omnibus (GEO) database under accession code GSE39757.

*Note: Supplementary information is available in the online version of the paper.*

ACKNOWLEDGMENTS

We thank members of our laboratory and M. Lalande for stimulating discussion and critical reading of the manuscript. The research leading to these results has received funding from the European Research Council (ERC) under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 206875 and the INSERM (Avenir Program R0721HS).

AUTHOR CONTRIBUTIONS

C.V. and C.H. performed the experiments. Y.L. and L.D. did the bioinformatic analysis of *XACT* conservation. N.O. and A.B.-G. provided the H9-MSC-iPS system. A.R. contributed to the bioinformatics analysis of the RNA-seq data. C.V. and C.R. conceived and designed the experiments, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2530.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Guttman, M. *et al. Nature* **458**, 223–227 (2009).
2. Khalil, A.M. *et al. Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
3. Augui, S., Nora, E.P. & Heard, E. *Nat. Rev. Genet.* **12**, 429–442 (2011).
4. Dvash, T. & Fan, G. *Epigenetics* **4**, 19–22 (2009).
5. The ENCODE Project Consortium. *PLoS Biol.* **9**, e1001046 (2011).
6. Romanish, M.T., Lock, W.M., van de Lagemaat, L.N., Dunn, C.A. & Mager, D.L. *PLoS Genet.* **3**, e10 (2007).
7. Sandelin, A. *et al. Nat. Rev. Genet.* **8**, 424–436 (2007).
8. Silva, S.S., Rowntree, R.K., Mekhoubad, S. & Lee, J.T. *Proc. Natl. Acad. Sci. USA* **105**, 4820–4825 (2008).
9. Mekhoubad, S. *et al. Cell Stem Cell* **10**, 595–609 (2012).
10. Mitjavila-Garcia, M.T. *et al. J. Mol. Cell Biol.* **2**, 291–298 (2010).
11. Shen, Y. *et al. Proc. Natl. Acad. Sci. USA* **105**, 4709–4714 (2008).
12. Giuliani, M. *et al. Blood* **118**, 3254–3262 (2011).
13. Okamoto, I. *et al. Nature* **472**, 370–374 (2011).
14. Okamoto, I., Otte, A.P., Allis, C.D., Reinberg, D. & Heard, E. *Science* **303**, 644–649 (2004).
15. Rougeulle, C. & Avner, P. *Semin. Cell Dev. Biol.* **14**, 331–340 (2003).

**Résumé**

En génomique comparative, on considère classiquement trois forces déterminant l'évolution des séquences : la mutation, la sélection et la dérive génétique. Récemment, lors de l'étude de l'origine évolutive des variations de la composition en base des génomes, un quatrième agent a été identifié : la conversion génique biaisée (BGC). Le BGC est intimement lié à la recombinaison méiotique et semble présent chez la plupart des eucaryotes. Ce phénomène introduit une surreprésentation de certains allèles dans les produits méiotiques aboutissant à une augmentation de la fréquence de ces variants dans la population. Ce processus est capable de mimer et d'interférer avec la sélection naturelle. Il est donc important de le caractériser afin de pouvoir le distinguer efficacement de la sélection dans l'étude de l'adaptation à l'échelle moléculaire. C'est ce que nous nous attachons à faire dans le cadre de ce travail. Pour cela nous utilisons deux espèces modèles. Premièrement la levure *Saccharomyces cerevisiae* pour laquelle une carte de recombinaison haute résolution permettant l'analyse du processus de conversion, est disponible. L'étude approfondie de cette carte nous a permis de lever le voile sur les mécanismes moléculaires qui sous-tendent le BGC. Deuxièmement, grâce à des découvertes récentes sur la détermination des patrons de recombinaison *via* la protéine PRDM9 chez les mammifères, nous avons quantifié la dynamique et l'intensité de ce processus dans l'histoire évolutive récente de l'homme. Ces résultats nous ont permis de confirmer la place du BGC comme quatrième force d'évolution moléculaire, mais aussi de discuter de l'origine évolutive de ce phénomène.

**Mots-clés :** Conversion génique biaisée, Crossing-overs, Evolution moléculaire, Génome, Mismatch repair, Points chauds, PRDM9, Recombinaison.

**Abstract**

Usually, three main forces are considered when studying sequences evolution in comparative genomics : mutation, selection and genetic drift. Recently, a fourth process has been identified during the study of base composition landscapes in genomes : biased gene conversion (BGC). This phenomenon introduces an overrepresentation of certain alleles in meiosis products (gametes or spores) leading to an increase of the frequency of those variants in the population. Thus, it is able to mimic and interfere with natural selection. Hence, it is important to describe this phenomenon in order to be able to trustfully distinguish BGC and selection in the study of adaptation at the molecular scale. So, the main goal of this work is to analyze the molecular origin, the intensity and the dynamics of BGC. To do so, we use two model species. First, we use the yeast *Saccharomyces cerevisiae* because, for this specie, a high-resolution recombination map is available which allows a fine study of the conversion process. Analyzing this map led us to shed the light on the molecular mechanisms of BGC. Secondly, recent discoveries on the role of the PRDM9 protein in the determination of recombination landscapes in mammals allowed us to quantify the dynamics and intensity of BGC in the recent human history. Thanks to those two studies, we first confirmed that BGC is the fourth force of molecular evolution and we also provided hypotheses about the evolutionary origin of this process.

**Key words:** Biased gene conversion, Crossing-overs, Genome, Hotspots, Mismatch repair, Molecular evolution, PRDM9, Recombination.