Modélisation théorique de l'évolution de la recombinaison méiotique chez les mammifères : Des mécanismes moléculaires à la dynamique évolutive

Alice Genestier

Université Claude Bernard Lyon 1

THESE de DOCTORAT DE L'UNIVERSITÉ CLAUDE BERNARD LYON 1

Ecole Doctorale 341
Évolution, Écosystèmes, Microbiologie, Modélisation

Discipline : Génomique évolutive

Soutenue publiquement le 05/12/2023, par :

Alice Genestier

Modélisation théorique de l'évolution de la recombinaison méiotique chez les mammifères. Des mécanismes moléculaires à la dynamique évolutive

---

Devant le jury composé de :

Céline BROCHIER-ARMANET Professeure, Université Claude Bernard Lyon 1	**Présidente**
Valérie BORDE Directrice de Recherches, CNRS/Institut Curie (France)	**Rapporteure**
Tom DRUET Professeur, Université de Liège (Belgique)	**Rapporteur**
Bertrand LLORENTE Directeur de Recherche, CNRS/CRCM (France)	**Rapporteur**
Nicolas LARTILLOT Directeur de recherche, CNRS/LBBE	**Directeur de thèse**
Laurent DURET Directeur de recherche, CNRS/LBBE	**Co-Directeur de thèse**

"N'oublie jamais, celui qui croit savoir n'apprend plus."
Pierre Bottero - Le Pacte des Marchombres - Ellana (2006)

À la mémoire de mon grand père Yves Paillard et de mon grand oncle Jacques Picard

Remerciements

Le manuscrit que vous tenez entre vos mains est le fruit de plusieurs années de travail. Rien n'aurait été possible sans le soutien et la participation de nombreuses personnes que je vais remercier ici.

Je voudrais tout d'abord remercier ceux qui m'ont encadrée pendant ces trois années de thèse, Nicolas Lartillot et Laurent Duret. Merci de m'avoir accompagnée et guidée pour mes premiers pas, parfois difficiles, dans le monde de la recherche. J'ai beaucoup appris grâce à vous; en terme de connaissances et de démarche scientifiques et d'écriture. Merci pour votre écoute et votre encadrement pendant cette période plus qu'inhabituelle de pandémie mondiale.

J'aimerais ensuite remercier mon jury de thèse, Céline Brochier-Armanet, Valérie Borde, Tom Druet et Bertrand Llorente. Je suis heureuse que vous ayez accepté d'évaluer ce travail de thèse.

Un grand merci à mon comité de suivi de thèse, Sylvain Glemin, Corrine Grey, Hélène Badouin et Tristan Lefébure pour m'avoir guidée et proposé des pistes de réflexions très intéressantes à explorer, ainsi que pour leur qualité d'écoute.

Je remercie le laboratoire du LBBE, tous ses "habitants et habitantes", et l'Université Claude Bernard Lyon 1 de m'avoir accueillie pendant ces trois ans. Merci à l'école doctorale, qui m'a permis, grâce au concours, d'obtenir une bourse sans laquelle cette thèse n'aurait pas été possible. Je remercie donc le ministère de l'enseignement supérieur et de la recherche ainsi que le CNRS pour m'avoir financée pendant toute la durée de ma thèse.

Je remercie tous les membres de l'équipe BPGE, Céline, Laurent, Nicolas, Hélène, Laure Segurel, Emmanuelle Lerat, Anamaria Necsulea, Carina Mugal, Mark Stoneking, Bénédicte Lafay, Sylvain Mousset, Guy Perrière, Jean-Pierre Flandrois, Dominique Mouchiroud et Manolo Gouy pour leur apport en terme de connaissances scientifiques et pistes à explorer.

Je remercie aussi tous les membres de l'ANR HotRec pour leurs avis précieux concernant mon travail et les discussions scientifiques très intéressantes sur la recombinaison.

Je remercie les différentes personnes avec qui j'ai pu partager mon bureau : Mélodie, Alexia Nguyen Trung (décoratrice officielle), Marie Verneret ainsi que toutes et tous les stagiaires qui ont fait escale dans notre merveilleux bureau.

J'aimerais beaucoup remercier, tous les doctorants, post-doctorants et stagiaires avec qui j'ai eu le plaisir de partager de bons moments, des discussions passionnantes sur des thèmes scientifiques et personnels. Merci à Lisa Nicvert, Florian Lecorvaisier, Rémi Tuffet, Blandine Charrat, Mélodie, Julien Joseph, Florian Bénitiere, Léa Keurinck, Gaspard Dussert, Rémi-Vinh Coudert, Lucas Lalande, Solène Cambreling et Mary Varoux.

Merci à tout ceux, déjà partis, avec qui j'ai partagé des moments sympatiques, au laboratoire comme en conférence. Merci à Thibault Latrille, Alexandre Laverré, Théo Tricout, Djivan Prentout, Émilie Fleurot, Chloé Haberkorn et tant d'autres.

Merci au pôle informatique Bruno Spataro, Stéphane Delmotte, Lionel Humblot,

Simon Penel, Philippe Veber, Adil El Filali, Vincent Miele et Aurélie Siberchicot pour avoir fourni le matériel et m'avoir aidé pendant mes détresses informatiques.

Je remercie aussi le pôle administratif Nathalie Arbasetti, Aurélie Zerfass, Sarah Ferkhous, pour m'avoir aidée dans toutes les démarches à effectuer, que ce soit pour des conférences ou d'autres formalités.

Merci à tous les responsables d'UE et membres de l'équipe pédagogique qui ont accepté de travailler avec moi, jeune doctorante en demande d'heures d'enseignement. Merci à Christelle Lopes, Isabelle Amat, Anne-Béatrice Dufour, Marie-Claude Venner, Vincent Lacroix, Arnaud Mary et Sandrine Charles.

Merci à mes colocs qui m'ont supportée pendant tout le long de cette expérience. Merci à Lisa, Cyprien et Typhaine pour les supers moments passés ensemble, les discussions sur tout et n'importe quoi (surtout n'importe quoi) et les fous rires.

Merci à tous mes amis et amies pour leur soutien inconditionnel, pour ces moments passés avec eux, en présentiel ou en distanciel, à jouer, rire et oublier le travail : Judith, Ambre, François, Benjamin, Alan, Éléanor et tant d'autres ! Merci à celles qui ont réalisé ou réalisent une thèse en ce moment même Marie, Louise (courage les filles) et Hanâ. Un merci tout particulier à Judith, pour m'avoir soutenue (presque) quotidiennement de l'autre bout de la France. Merci beaucoup à tous !

Merci à la famille étendue, grands-parents, oncles, tantes, cousines et tous les autres pour votre soutien et votre intérêt pour mes études. Mon plus grand merci revient au cocon familial, mes parents et mes trois soeurs, pour leur soutien inconditionnel pendant toutes ces années. Merci à mes parents pour m'avoir encouragée en toute circonstance, écoutée, consolée quand j'en avais besoin et hébergée pendant le confinement et pour l'écriture de ma thèse. Merci et félicitations à Lucie et Ninon pour leur épanouissement dans leur vie personnelle et leur réussite professionnelle. Merci à ma petite soeur Mina pour son soutien permanent même si le principe de faire une thèse lui échappe complètement. Et enfin je remercie mes chats, qui ne jugent jamais mon travail, et qui se fichent éperdument d'être cités ici.

Alice Genestier

Résumé

La méiose est une étape fondamentale du cycle de vie eucaryote. Elle est la contrepartie du sexe dans l'alternance entre phases haploïde et diploïde. Elle implémente à la fois l'appariement des chromosomes homologues, étape clé de la ségrégation dans les quatre produits haploïdes de la division, mais aussi, grâce à la recombinaison, le brassage génétique à l'échelle de la population en générant de nouvelles combinaisons d'allèles. Cependant, malgré son importance, la dynamique évolutive de la recombinaison méiotique reste encore une des grandes questions de la biologie. Afin d'aborder ces questions, il est crucial d'incorporer les connaissances sur les mécanismes moléculaires de la méiose dans les modèles théoriques d'évolution de la recombinaison. Chez les mammifères, la recombinaison est régulée par le gène PRDM9, initialement découvert pour son rôle dans la stérilité hybride et la spéciation chez la souris. Ce gène, qui code pour une protéine à doigts de zinc se liant à l'ADN, possède une double fonction. D'une part, la protéine PRDM9 induit l'initiation de la recombinaison en se fixant à l'ADN et en recrutant la machinerie de cassures double brins. D'autre part, elle facilite l'appariement des chromosomes homologues par sa liaison symétrique. Le fonctionnement moléculaire de PRDM9 résulte en un processus cyclique de Reine Rouge intra-génomique. Cette dynamique est la conséquence de l'opposition entre deux forces antagonistes : l'érosion du paysage de recombinaison par destruction des cibles de PRDM9 par conversion génique, compensée par une sélection positive faisant émerger de nouveaux allèles restaurant la recombinaison par la reconnaissance de nouvelles cibles. Cette Reine Rouge aboutit à une très forte instabilité des paysages de recombinaison, dont les modalités et les conséquences ne sont pas encore bien comprises. Dans ce contexte, mon travail de thèse consistait à modéliser, au moyen de simulations informatiques et de développements théoriques, la dynamique évolutive de la recombinaison méiotique et du gène PRDM9 chez les mammifères, en intégrant l'ensemble des connaissances actuelles au sujet des mécanismes moléculaires de la méiose. Dans un premier temps, une version du modèle en population unique a permis de montrer que la double fonction de PRDM9 durant la méiose est suffisante à induire un mécanisme de Reine Rouge. Nous avons pu établir que la cause de la sélection positive agissant sur le gène PRDM9 réside dans l'érosion préférentielle des sites les plus fortement liés par la protéine. On a mis en évidence que le dosage génétique, en conférant une meilleure fertilité aux homozygotes pour PRDM9, aboutit à un phénomène d'éviction des allèles jeunes, et joue de ce fait contre la diversité génétique en ce locus. Enfin, les calibrations empiriques tentées dans cet axe suggèrent qu'il est difficile d'expliquer simultanément la forte diversité génétique et la sélection forte observée sur le gène PRDM9, un dilemme lié au dosage génétique. Dans un deuxième temps, le modèle adapté en contexte bi-populationnel a permis d'observer de la stérilité hybride causée par l'asymétrie de liaison de PRDM9, confirmant les observations empiriques de la littérature. Cette stérilité hybride est cependant marginale et transitoire, ce qui ne plaide par pour un rôle majeur de PRDM9 dans la spéciation. Toutefois, la difficulté du modèle à simultanément expliquer l'ensemble des données empiriques incite à une certaine pru-

dence quant aux conclusions. Les problèmes de calibrations empiriques et la difficulté à prédire de la stérilité hybride suggèrent la nécessité de mieux comprendre certains aspects du mécanisme d'action de PRDM9, comme la distribution d'affinité des cibles, la concentration de PRDM9 ou les effets de dominance. Ultimement, une version plus complexe du modèle pourrait permettre l'étude du rôle de PRDM9 dans la dissipation des effets Hill-Robertson, ou des effets Dobzhansky-Muller pouvant induire de la stérilité hybride.

Résumé étendu

Chez les eucaryotes, la méiose est une étape essentielle du cycle de vie permettant le passage de cellules diploïde à haploïde. Ce processus cellulaire comprend une étape critique d'appariement des chromosomes homologues afin que ceux ci soient ensuite redistribués dans les cellules filles. Durant cette phase d'appariement se déroule un mécanisme moléculaire d'échange de matériel génétique entre chromosomes homologues appelé recombinaison méiotique. Chez la plupart des eucaryotes, la recombinaison est initiée par la formation de cassures double-brin, qui sont ensuite resectées, aboutissant à la formation d'extrémités simple-brin qui vont ensuite rechercher une séquence complémentaire avec laquelle s'apparier. L'ensemble de ce processus permet in fine l'appariement des chromosomes homologues, sur la base d'une reconnaissance de leur homologie de séquence. Chez les mammifères, les cassures double-brin, et par conséquent la recombinaison, ne sont pas uniformément distribuées dans le génome. Au contraire, il existe des petites régions du génome, appelées points chauds de recombinaison, qui concentrent l'essentiel des DSBs. La localisation de ces points chauds est déterminée par une protéine : PRDM9. Cette protéine possède deux fonctions essentielles pendant la méiose. Premièrement, de par son domaine à doigts de zinc, elle reconnaît des séquences ADN spécifiques et peut s'y lier puis tri-méthyle les histones environnant, ce qui recrute la machinerie de cassure double brins. La deuxième fonction de PRDM9 consiste en la facilitation de l'identification des chromosomes homologues, et ce, par sa liaison symétrique (liaison simultanée sur les deux chromosomes, sur des cibles homologues). Cette symétrie semble essentielle au bon déroulé de la méiose. En effet, chez des hybrides de souris, un fort taux d'asymétrie de liaison de PRDM9 a été identifié comme une cause majeure de stérilité, ce qui a mené certains à proposer que *PRDM9* serait un potentiel gène de spéciation.

De plus, lors de la réparation de site cassés par un site érodé, on observe de la conversion génique biaisée en faveur du site érodé ce qui, sur le long terme, cause une disparition progressive des points chauds de recombinaison. Cette auto-destruction des paysages de recombinaison a donné lieu au paradoxe des points chauds stipulant que ce mécanisme d'érosion de la recombinaison aurait pour conséquence la disparition de la recombinaison ce qui ne correspond pas aux taux observés empiriquement. Or, en parallèle, il a été découvert d'une part que les paysages de recombinaison étaient fortement différents entre espèces soeurs faisant penser à une évolution rapide de la localisation des points chauds et, d'autre part, que le gène *PRDM9* était sous forte sélection positive ce qui fait que ce gène évolue très rapidement. Ces deux forces antagonistes ont mené à la proposition d'un modèle dit de Reine Rouge par Ubeda et Wilkins en 2011 qui permettrait de résoudre le paradoxe des points chauds. Cette dynamique fonctionnerait comme un cycle perpétuel d'érosion des cibles par un allèle *PRDM9* faisant baisser la fertilité puis la sélection positive de nouveaux allèles reconnaissants de nouvelles cibles restaurant la fertilité. Cette théorie a été étudiée par des simulateurs informatiques et des développements mathématiques dans des configurations simples de population unique et sans prendre en compte tous les mécanismes moléculaires de PRDM9. Ces modèles ont

montré que la Reine Rouge était un phénomène pouvant contrecarrer le paradoxe des points chauds et que, suivant les paramètres du modèle, cette dynamiques pouvait avoir différents régimes et différents niveaux d'équilibre pour la diversité *PRDM9* ou le niveau d'érosion. Cependant ces modèles restent naïf sur certains points. Premièrement, ils ne prennent pas en compte la deuxième fonction moléculaire de PRDM9 (son rôle dans la facilitation de l'appariement par sa liaison symétrique) qui pourrait expliquer la sélection positive également dans le contexte intra-populationnel. Deuxièmement, les modèles sont réalisés en population unique ne permettant pas d'expliquer toutes les caractéristiques observées empiriquement comme la cause de la sélection positive ou la stérilité hybride.

Dans ce contexte, mon travail de thèse a consisté en l'implémentation d'un simulateur informatique prenant en compte les principes de la génétique des populations et les mécanismes moléculaires de PRDM9 dans un contexte d'une seule population, d'une part, puis de deux populations évoluant indépendemment et générant des hybrides, d'autre part. Sur la base de ces deux versions du modèle, mon travail de thèse est par conséquent divisé en deux parties. Le premier axe traite de l'impact intra-populationnel du rôle dual de PRDM9 pendant la méiose. Dans cet axe sont par ailleurs explorées les conséquences du dosage génétique de PRDM9. Enfin, une calibration empirique est tentée, prenant en compte un certain nombre de connaissances empiriques disponibles dans la littérature et les utilisant pour calibrer les paramètres du modèle et tester ensuite si le modèle peut prédire des valeurs raisonnables pour quelques statistiques descriptives d'intérêt (comme la diversité génétique de *PRDM9* ou les niveaux d'érosion des cibles). Le deuxième axe, de son coté, traite du rôle de *PRDM9* dans la stérilité hybride et permet de tester les conditions d'observation de cette stérilité et de son intensité, que ce soit en conditions empiriques ou en faisant varier certains paramètres.

Dans le contexte du premier axe, nous avons fait tourner notre simulateur dans un contexte uni-populationnel. Le modèle prédit une dynamique de Reine Rouge qui, comme dans les modèles théoriques précédemment publiés, est caractérisée par la compétition entre deux forces : érosion des cibles des allèles de *PRDM9* d'un côté, et sélection positive pour de nouveaux allèles de l'autre. Toutefois, les aspects mécanistes de notre modèle permettent de caractériser plus précisément ces deux aspects, et tout particulièrement l'origine de la sélection positive sur les nouveaux allèles. En effet, l'érosion par conversion génique biaisée attaque préférentiellement les cibles de haute affinité (qui sont plus souvent liés et donc plus souvent cibles de DSBs). La perte des sites de haute affinité a pour conséquence de faire baisser le taux de liaison symétrique à l'échelle du génome, impactant ainsi négativement la fertilité des individus possédant de vieux allèles *PRDM9*. À l'inverse, les nouveaux allèles *PRDM9*, par leur reconnaissance de nouvelles cibles non érodées de haute affinité, restaurent un niveau élevé de fertilité. Par ailleurs, notre modèle implémente le dosage génétique de PRDM9 (en supposant que la concentration de la protéine corrrespondant à un allèle donné dépend du nombre de copies, respectivement une ou deux, de cet allèle, chez les hétérozygotes ou chez les homozygotes). De façon cruciale, le dosage génétique confère un avantage aux homozygotes, du fait que la concentration plus élevée implique un plus fort taux de liaison, et donc de liaison symétrique, et

par conséquent, induit une meilleure fertilité. Ce phénomène induit un régime d'éviction contre les jeunes allèles qui, par nature, apparaissent en contexte hétérozygote. Cette éviction due au dosage induit une forte baisse de diversité génétique dans des régimes où *PRDM9* ne mute pas assez rapidement et où le taux d'érosion est trop faible. Une première tentative de calibration empirique prenant en compte les effets de dosage a donc prédit une diversité *PRDM9* trop faible par rapport aux observations empiriques. Une plus grande diversité peut être restaurée par un découplage entre fertilité des individus et succès de la méiose. Cependant, ce nouveau mécanisme, tout en diminuant les effets dosage, réduit également les différences de fertilité entre jeunes et vieux allèles, et donc la force de la sélection positive, faisant au bout du compte tourner la Reine Rouge dans un régime quasi-neutre. Alternativement, augmenter le nombre de cassures double brins réalisables en une méiose (et ce, afin de moédliser indirectement la régulation du nombre de cassures) a pour impact de restaurer la diversité de *PRDM9* tout en gardant une sélection positive raisonnable, mais cette fois-ci en prédisant une haplo-insuffisance des allèles vieux plus faible que celle observée empiriquement. Ces calibrations démontrent par conséquent que le modèle est fortement contraint. En perspective, faire varier la distribution d'affinité pourrait permettre d'obtenir de nouveaux régimes sans éviction. De plus, la prise en compte de la concentration absolue de PRDM9 dans la cellule pourrait aussi moduler l'intensité de l'effet du dosage. Enfin, d'autre mécanismes empiriques tels que la dominance entre allèles ou la polymérisation des protéines pourraient impacter la dynamique évolutive de la recombinaison.

La baisse de fertilité due à une baisse de taux de symétrie de PRDM9 prédite par notre modèle a, à l'origine, été découverte en contexte hybride. C'est pourquoi, dans un deuxième temps, nous avons adapté notre modèle en contexte bi-populationnel, et ce, afin d'étudier et de mieux quantifier l'impact de l'asymétrie de liaison de PRDM9 sur la fertilité des hybrides. Un des objectifs de ce travail est d'évaluer dans quelle mesure la stérilité hybride induite par *PRDM9* est suffisamment prévalente et forte pour pouvoir considérer *PRDM9* comme un gène de spéciation plausible. Ainsi, dans ce deuxième axe, le simulateur modélise une population ancestrale se scindant en deux populations, qui évoluent alors indépendemment l'une de l'autre et forment régulièrement des hybrides par croisement entre individus des deux populations. Une comparaison des fertilités moyennes des hybrides à celles obtenues en intra-population montre qu'en moyenne, de la stérilité hybride est effectivement observée. L'asymétrie de liaison de PRDM9 est bien une des causes de cette stérilité hybride, comme initialement suggéré à partir de résultats empiriques obtenus par croisement de deux sous espèces de souris – mais non la seule, car les effets de dosage génétique, mis en évidence dans l'axe 1, jouent ici également un rôle. Cependant, notre modèle prédit de la stérilité plutôt marginale et transitoire, sauf en de rares occasions où les deux allèles hérités par l'hybride sont simultanément caractérisés par un fort taux d'érosion dans les populations parentales. Dans ce cas précis, l'asymétrie est assez importante pour générer de la stérilité chez l'hybride. Cette prédiction globale du modèle, d'une stérilité hybride faible ou très occasionnellement forte, ne correspond pas aux observations empiriques. Notons que, dans notre modèle, la

migration entre populations n'a pas été implémentée ne permettant pas de tester l'impact du flux de gène. La migration pourrait pourtant avoir un effet positif sur la diversité intra-populationnelle par la potentielle introgression adaptative de certains allèles d'une population dans l'autre. Enfin les autres paramètres pouvant impacter la dynamique intra-populationnelle, présentés dans le paragraphe précédent, pourraient aussi impacter la stérilité hybride et seraient donc à tester en contexte bi-populationnel.

En conclusion, notre modèle donne de nouveaux résultats en faveur de la dynamique de Reine Rouge et permet de mieux comprendre les déterminations exactes de l'évolution des paysages de recombinaison chez les mammifères possédant *PRDM9*. Cependant, le modèle a permis d'identifier un phénomène d'éviction dû au dosage génétique de PRDM9, agissant contre la diversité en ce locus d'une manière qui rend difficile de faire sens des observations empiriques. Cependant, il existe certains paramètres comme la forme exacte de la distribution d'affinité des cibles, la compétition entre cibles de PRDM9, compétition qui elle-même dépend de la concentration absolue de PRDM9 ainsi que des affinités absolues des cibles, qui sont encore mal connus et dont les valeurs exactes pourraient *in fine* permettre de résoudre ces problèmes. Il devient donc important de démêler les effets de l'affinité, du dosage et de la concentration de PRDM9, que ce soit empiriquement ou théoriquement, pour pouvoir aboutir à un modèle global du mode d'action biochimique et moléculaire de PRDM9. Par ailleurs, lors de nos tentatives de mieux cerner les effets du dosage et de caractériser les conditions qui en limitaient l'impact, nous avons identifié un régime quasi-neutre de la Reine Rouge. Ce dernier pourrait finalement ne pas être totalement incompatible avec les observations empiriques de sélection forte agissant sur *PRDM9*, car il pourrait en effet être intercalé entre des périodes de forte sélection positive agissant sur des allèles ayant assez érodé leurs cibles pour subir une perte significative de fertilité. Par ailleurs, le modèle ne prend pas en compte les effets Dobzhansky-Muller générés par des incompatibilités entre *PRDM9* et un autre locus, comme par exemple *Hstx2*, qui semble jouer un rôle dans la stérilité hybride. La prise en compte de la migration pourrait aussi fortement jouer sur les niveau de diversité de *PRDM9* intra-populationnel. Enfin, de manière générale, notre modèle, qui ne considère que les enjeux de la méiose, et tout particulièrement celui de l'appariement des chromosomes, ne prend pas en compte à ce stade le rôle de la recombinaison dans la dissipation de la liaison génétique entre différents loci sous sélection. Cet autre effet de la recombinaison pourrait représenter une nouvelle cause de sélection agissant sur *PRDM9*. En particulier, elle pourrait jouer en faveur de certains allèles *PRDM9* qui casseraient des liaison génétiques entre deux allèles sous sélection opposée. Enfin, tout ce travail mène à se poser des questions sur les raisons de l'existence et du maintien de *PRDM9* chez certaines espèces, alors même que ce gène a été perdu à de multiples reprises à travers les métazoaires, sans que cela ait des répercussions grave sur la recombinaison. Des perspectives sur la raison d'être du système *PRDM9*, et des modalités possibles de sa perte, sont proposées.

Table des matières

Liste des Figures

Liste des Tableaux

Acronymes

ADN (DNA) : Acide DésoxyriboNucléique (DeoxyriboNucleic Acid)

CO : Crossing-Over

Chip : Immunoprécipitation de chromatine

DSB : Cassure Double Brin (Double Strand Break)

H3K4me3 : Tri-méthylation au 4ème résidu (qui est une lysine) de la protéine d'histone H3

H3K36me3 : Tri-méthylation au 36ème résidu (qui est une lysine) de la protéine d'histone H3

kb : Kilobase

LD : Déséquilibre de liaison (Linkage Desequilibrium)

NCO : Non Crossing-Over

ssDNA : Simple brin d'ADN (Single Stranded DNA)

TSS : Site d'initiation de la transcription (Transcription Start Site)

Préambule

La thèse présentée dans ce manuscrit est à la croisée des chemins entre plusieurs domaines scientifiques. Tout d'abord, elle crée un pont entre biologie moléculaire et biologie évolutive, en se focalisant plus spécifiquement sur les mécanismes et la dynamique évolutive de la méiose et la recombinaison méiotique. Elle fait ensuite le lien avec les mathématiques par l'élaboration de développements analytiques. Enfin, par l'implémentation d'un programme de simulation, elle s'intègre aussi dans le domaine de la modélisation informatique. Cette pluridisciplinarité est à la fois une occasion unique de contribuer à construire de nouvelles passerelles entre domaines distincts de la recherche scientifique mais pose aussi de nombreux défis. En effet, vu l'étendue des domaines concernés, il est impossible de tout traiter, et encore moins de tout maîtriser, dans les moindres détails. J'aurai toutefois fait de mon mieux afin que l'essentiel des idées et des connaissances propres à chacun des domaines, et qui sont par ailleurs nécessaires au traitement de la question scientifique au centre de mes travaux, soient tout de même présentés ici. Concernant la personne qui voudra lire ce manuscrit, suivant son domaine d'étude ou de prédilection, celle-ci devra peut-être s'adapter à de nouvelles notions ou de nouveaux formalismes, parfois difficiles d'accès, en tout cas au premier abord. En vue de faciliter la lecture, j'ai tenté d'expliquer au mieux les idées et les intuitions qui motivent ces formalismes, et ce, afin de garantir au mieux une bonne compréhension de la part de toutes et de tous, quelle que soit la spécialité.

Cette thèse est présentée en vue de satisfaire partiellement aux conditions requises pour l'obtention du titre de Docteur à l'Université de Lyon. Le travail de recherche présenté ici a été réalisé au Laboratoire de Biométrie et Biologie Evolutive (LBBE), sous la supervision des directeurs de recherche Nicolas Lartillot and Laurent Duret. Ce travail a été mené à partir d'octobre 2020 grâce à une subvention de 3 ans par le Ministère de l'enseignement supérieur et de la recherche (Concours de l'école doctorale E2M2) puis par le CNRS pendant 3 mois.

Ce manuscrit de thèse est constitué d'une introduction générale rédigée en français fournissant le contexte et les connaissances nécessaires à la compréhension des travaux de recherches. Ces dernier sont détaillés dans deux manuscrit rédigés en anglais, chacun précédé de résumés court et étendu écrits en français. Le premier article est disponible sur BioRxiv et a été soumis au journal PLOS Genetics. Le deuxième article n'a pas encore été soumis. Enfin le manuscrit de thèse se termine sur une discussion rédigée en français.

Partie I

Introduction

1 Introduction

Contents

1.1 La méiose

1.1.1 Principe général

La méiose est une étape centrale dans le cycle de vie des eucaryotes. Elle est le pendant du sexe et de la reproduction sexuée. La combinaison de la méiose et du sexe permet, dans un premier temps, l'alternance des générations diploïdes et haploïdes du cycle eucaryote. En effet, dans la nature, il existe plusieurs types d'alternance entre phase haploïde et diploïde (initialement découvert chez l'algue rouge par Svedelius [1]). Les organismes diplobiontiques (animaux, certains protistes et algues) ont leur cellules somatiques sous forme diploïde, et seuls leurs gamètes sont haploïdes. À l'inverse, les organismes haplobiontiques (certains champignons, algues et beaucoup de protistes tels que *Chlamydomonas* ou *Plasmodium*) sont sous forme haploïde la majorité du temps, et seul le zygote est diploïde pour une courte durée. Il existe d'autres alternatives comme chez la levure *Saccharomyces cerevisiae*, l'ensemble des plantes et la majorité des algues, capables de rester au stade végétatif ou de se répliquer sous forme aussi bien haploïde que diploïde. Dans tous les cas, la méiose permet la transition de diploïde à haploïde, et inversement pour le sexe.

D'autre part, la méiose implémente un mécanisme moléculaire fondamental : la recombinaison génétique. Ce processus est caractérisé dans un premier temps par l'échange de segments chromosomiques, ce qui génère du brassage génétique participant au maintien de la diversité génétique au sein des populations. Dans un second temps, la séparation des chromosomes homologues dans des cellules filles distinctes génère une seconde phase de réarrangement des génomes, permettant encore plus de brassage et donc de diversité.

Ces deux fonctions de la méiose, sa place dans le cycle de vie pour passer d'une cellule diploïde à des cellules haploïdes d'une part, et l'implémentation de la recombinaison participant au maintien de la diversité génétique d'autre part, posent la question plus générale de la raison d'être et de l'évolution de la méiose. Cette question est très complexe et est intimement liée à d'autres grandes questions de la biologie concernant le maintien du sexe et de la recombinaison (revues [2; 3]).

Dans ce contexte général, mon travail de thèse se focalise sur la question du maintien de la recombinaison au cours du temps, et ce, parmi les métazoaires chez qui la méiose joue le rôle de la création de gamètes. Je me focaliserai donc avant tout sur le rôle de la recombinaison dans le bon accomplissement du processus de la méiose, mettant de côté la question du brassage génétique.

1.1.2 Première fonction : création de gamètes

Le passage de cellules germinales diploïdes à un set de 4 cellules haploïdes est réalisé par la succession de deux divisions cellulaires différentes. La première division cellulaire permet la séparation des paires de chromosomes homologues, formant ainsi des cellules haploïdes et l'implémentation de la recombinaison génétique. La deuxième division cellulaire, à l'instar de la mitose, sert à séparer les chromatides soeurs dans des cellules

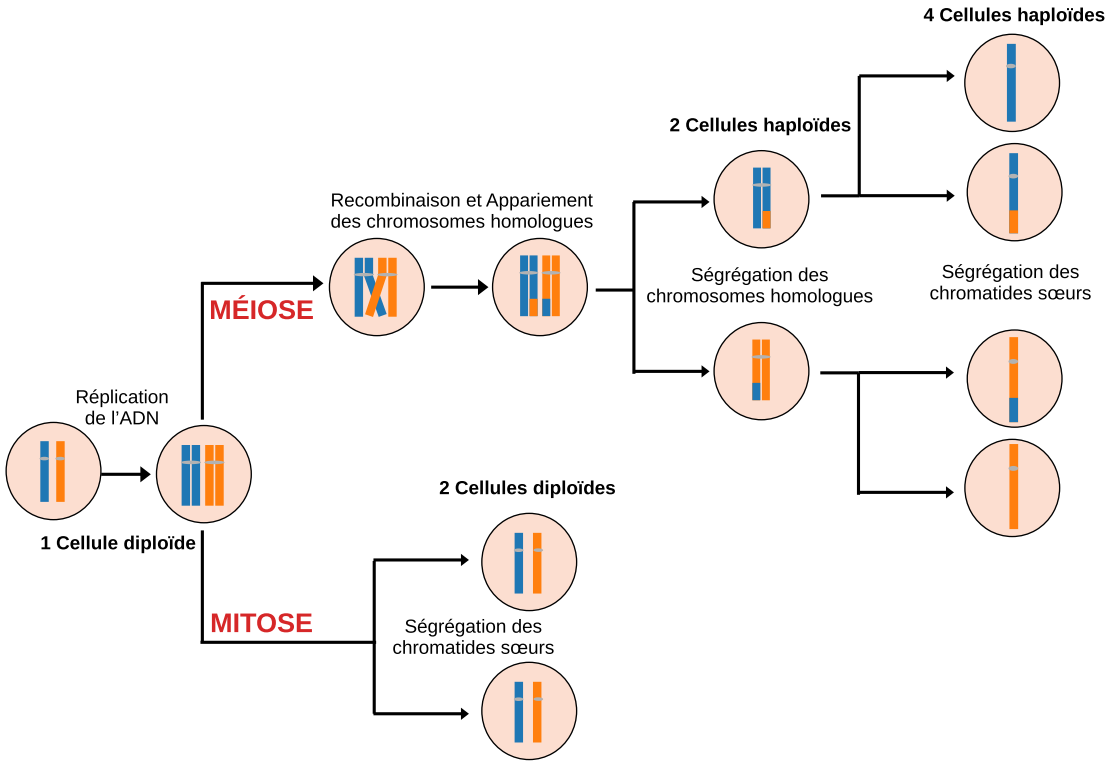

Figure 1.1: Schéma du déroulé des divisions cellulaires de la méiose et de la mitose. *Les deux divisions cellulaires commencent par une cellule diploïde qui subit une étape de réplication de l'ADN. La mitose (chemin vers le bas) est constituée d'une seule division cellulaire durant laquelle les chromatides soeurs de chaque chromosome se séparent dans des cellules filles distinctes diploïdes. La méiose (chemin vers le haut) est constituée de 2 divisions cellulaires successives. La première division permet de séparer les chromosomes homologues, préalablement appariés. Cette division donne naissance à deux cellules haploïdes. Il s'en suit une deuxième division, similaire à la mitose, ou les chromatides soeurs sont ségrégées à leur tour, pour former un ensemble de 4 cellules haploïdes.*

distinctes (cf figure 1.1). Ces deux divisions successives sont toutes deux composées de 4 sous-étapes ayant des mécanismes plus ou moins similaires : prophase, métaphase, anaphase et télophase.

1.1.3 Défi : séparer les chromosomes homologues

La méiose se différencie de la mitose (revue de Ohkura (2015) [4]) par sa première division cellulaire. Celle-ci représente un nouveau défi car, contrairement à la mitose qui sépare des chromatides soeurs d'un même chromosome (donc déjà liées entre elles), la méiose I (première division) doit séparer des chromosomes homologues, sans au préalable connaître quels chromosomes sont associés en paire. Le défi réside donc dans l'identification des paires de chromosomes homologues afin de les ségréger correctement dans les cellules filles. Cette étape est cruciale. En effet, une ségrégation incorrecte génère de l'aneuploïdie, ce qui crée des problèmes graves au niveau phénotypique, tel que le syndrome de Down

(aussi appelé trisomie 21, revue de Hassold *et al.* (2007) [5]).

1.1.4 Mécanismes d'appariement

La phase de reconnaissance des homologues, autrement appelée phase d'appariement, varie grandement selon les eucaryotes. Pour commencer, au tout début de la prophase I (dont le déroulé est détaillé dans la figure 1.2), pendant la phase de leptotène, les chromosomes sont à l'état décondensé, sous forme de filaments très fins dans le noyau. Ils vont alors commencer à se condenser et à s'aligner le long d'un axe, appelé axe chromosomique (revue de Zickler & Kleckner (1998) [6]). Lors de la transition vers la phase de zygotène, chez presque toutes les espèces étudiées, sauf le nématode *Caenorhabditis elegans* et la drosophile, les télomères des chromosomes se fixent sur la paroie interne du noyau et se rassemblent, formant une structure appelée "bouquet" (revues [7; 6]). À la base du bouquet, les chromosomes sont très proches les uns des autres et peuvent ainsi potentiellement commencer à s'apparier.

En complément des bouquets de chromosomes, il existe un autre mécanisme réalisé pendant le début de la prophase I favorisant l'appariement chez beaucoup d'espèces : la recombinaison. À noter que quelques espèces semblent réaliser l'appariement en se passant de la recombinaison, comme *C. elegans* [8]. Chez la drosophile, les études ne sont pas totalement claires. Il semblerait que les chromosomes homologues soient déjà alignés avant le début de la méiose [9] et que les voies d'appariement recombinatoires ne soient présentes que chez les femelles, les mâles ne possédant pas de chiasma [10; 11; 12]. Les phases de pré-appariement peuvent parfois être observées chez les espèces recombinantes telles que la souris ou la levure [13].

1.2 La recombinaison

1.2.1 Introduction

Bien que largement partagée à travers les espèces eucaryotes, la recombinaison diffère entre espèces sur différents points tels que ses mécanismes d'action (revues [14; 15; 16]), ou encore sa distribution à travers le génome. Dans un premier temps je détaillerai ce que l'on sait de la recombinaison méiotique, puis je présenterai les mécanismes de localisation des points de recombinaison à travers le génome (revues [7; 16]).

1.2.2 Mécanismes

Lors des deux premières étapes de la prophase I, des cassures double-brin (DSBs) irréversibles sont induites le long des chromosomes par l'enzyme SPO11 [17; 18]. Au niveau de chaque cassure, une résection de l'extrémité simple brin d'ADN est générée. Cette extrémité simple brin va alors rechercher sa séquence complémentaire afin de réparer la DSB (revue de Hunter (2015) [19]). Ce faisant, elle va représenter le mécanisme clé permettant la bonne identification du chromosome homologue, le stabiliser et faciliter

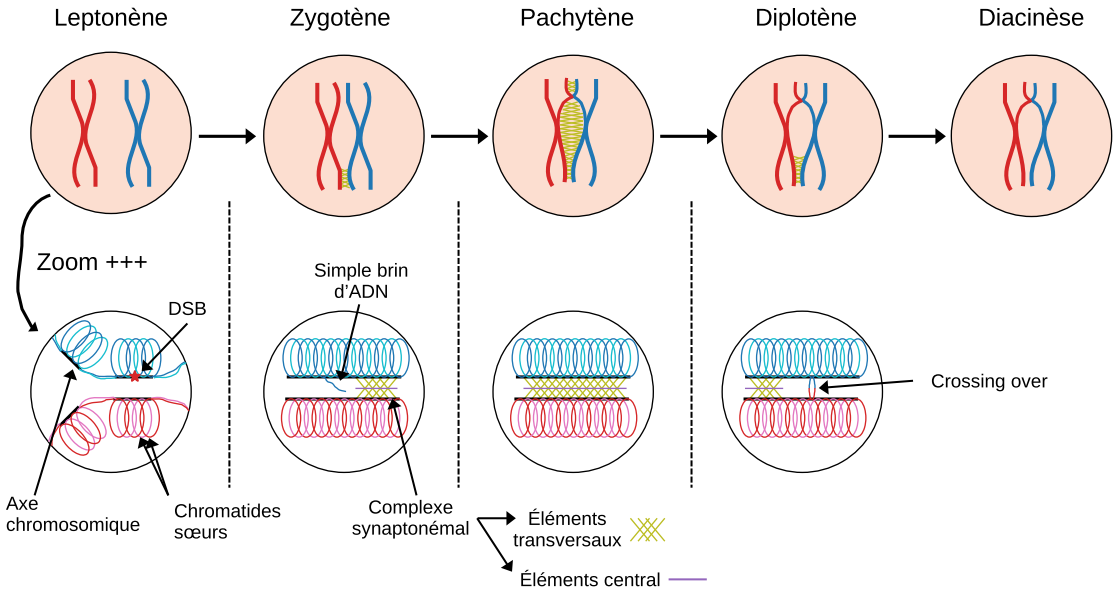

Figure 1.2: Schéma des étapes de la prophase I de la méiose. *La Prophase I de la méiose se découpe en 5 étapes. Durant le leptotène, les chromosomes homologues (rouge et bleu) possèdent déjà chacun deux chromatides soeurs et sont sous la forme décondensée. La chromatine se condense progressivement et s'agence sous forme de boucles. Les chromatides soeurs sont attachées entre elles (non montré pour plus de clarté). Les cassures double-brin (étoile rouge) commencent à se créer le long du génome au niveau de l'axe chromosomique (axe noir). Puis, lors de la phase de zygotène, les chromosomes commencent à s'apparier. Le complexe synaptonémal se forme au niveau des télomères (attachés à la membrane nucléaire pour former un "bouquet", non montré ici). Le complexe synaptonémal est constitué d'un élément central (axe violet) autour duquel s'articulent des éléments transversaux (croix vertes). En même temps, l'extrémité simple brin générée par la cassure recherche sa séquence homologue. Lors de la phase suivante, le pachytène, les chromosomes finissent d'être totalement apparié grâce au complexe synaptonémal qui se propage comme une fermeture éclair le long des chromosomes à partir des points de recombinaison et des télomères. Les évènements de crossing-over et de non crossing-over sont réalisés. Lors du diplotène, le complexe synaptonémal disparaît et les chromosomes homologues restent liés au niveau des points de crossing-over, appelés chiasmata. Lors de la phase finale, les télomères des chromosomes se détachent de l'enveloppe nucléaire lorsque cette dernière disparaît.*

l'invasion du chromosome homologue. Durant cette phase, les recombinases DMC1 et RAD51 se fixent sur le simple brin [20; 21; 22] . L'invasion de l'homologue par le simple brin peut se résoudre de deux manières différentes lors des étapes suivantes. Lors du zygotène et du pachytène, les sites ayant subi des DSB ainsi que les environs du site (petites régions allant de quelques centaines à un millier de paires de bases) sont réparés avec, pour matrice, la séquence du même site sur le chromosome homologue. Cela donne alors naissance à differents types de réparations distincts. J'en présenterai seulement deux (cf figure 1.3). Premièrement, on peut observer une réparation par la formation d'une double jonction de Holliday, résultant le plus souvent en un échange de séquences ADN caractérisant un évènement de type crossing-over (CO). Chez la souris et chez *S. cerevisiae*, les évènements de CO impliquent dans la majorité des cas les protéines MLH1 et MLH3 (chez la souris, cela représente environ 90% des CO [23]). On peut noter que, dans certains cas rares, les jonctions de Holliday sont réparées de manière non crossing-over (NCO). Chez la souris, les évènements de CO représentent environ 10% de réparation de DSB à travers le génome [24]. Deuxièmement, la cassure peut être réparée par "synthesis-dependent strand annealing", résultant en une simple copie de la séquence homologue sur la partie cassée, évènement de type NCO. Ces étapes de réparation permettent non seulement de générer de nouvelles combinaisons d'allèles, mais aussi de créer des points de contacts entre les chromosomes homologues. De manière moins fréquente, on observe aussi des réparations du site cassé par le site situé à la même position sur la chromatide soeur [25]. Ce processus limite donc l'échec de la méiose par non réparation de DSB mais ne permet pas la formation de points de contact entre chromosomes homologues pour leur appariement. En parallèle, ces points de cassures ainsi que les télomères deviennent des foyers de la synapse. Ce processus, qui peut être pensé comme une fermeture éclair, permet, grâce au complexe synaptonémal, de rapprocher les chromosomes d'une même paire sur toute leur longueur [26]. En effet, le complexe synaptonémal est composé des deux axes chromosomiques des homologues, appelés maintenant "éléments latéraux", et d'un élément central (revue de Zickler & Kleckner (1998) [6]). Des éléments transverses s'articulant autour de cet élément central vont rapprocher les éléments latéraux afin de former une seule et même entité. Durant les deux dernières étapes de la prophase, le diplotène et la dacinèse, le complexe synaptonémal se dissout, et les chromosomes se désassemblent, sauf aux centromères et aux emplacements de crossing-over (maintenant nommés chiasmata). Les chromosomes commencent alors à se préparer pour la ségrégation.

1.2.3 Les deux rôles de la recombinaison dans la méiose

Il émerge donc que la recombinaison est l'étape charnière de la réussite de la méiose et de la bonne réalisation de ses deux fonctions essentielles. Elle permet tout d'abord d'obtenir des points d'ancrage entre les chromosomes homologues, permettant ainsi à la cellule de distinguer les paires de chromosomes à séparer, ce qui assure le bon fonctionnement de la méiose. Par ailleurs, la recombinaison génère du brassage génétique. La question de

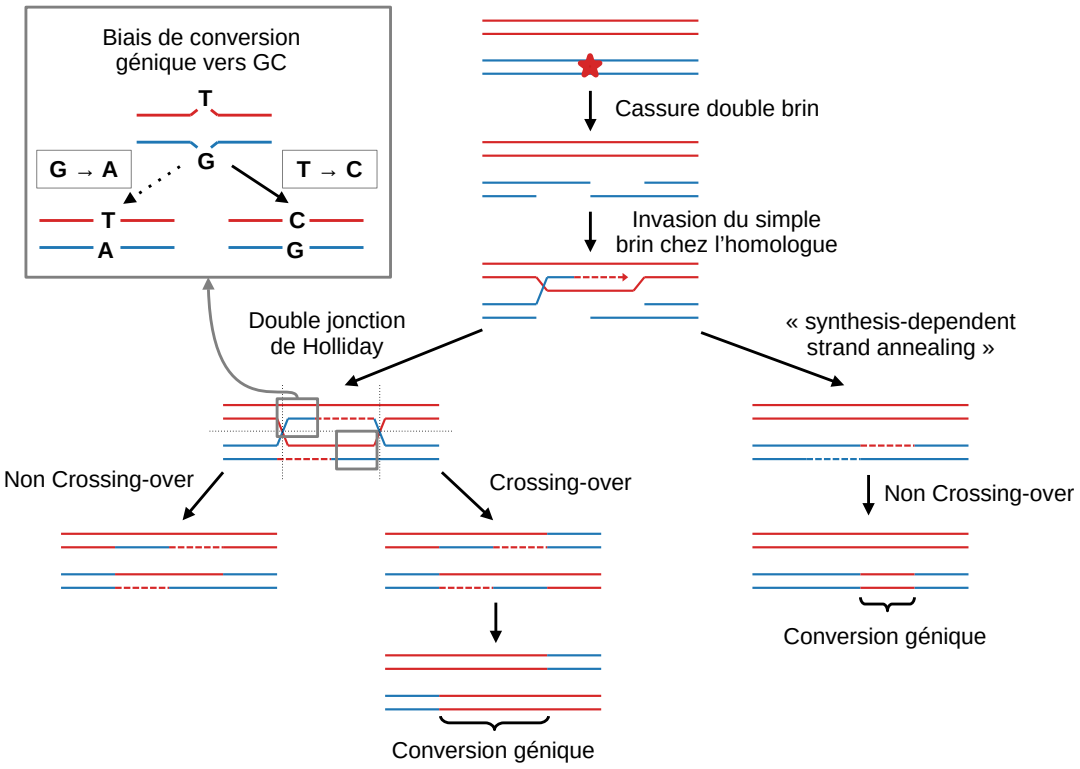

Figure 1.3: Différents moyen de réparation des cassures doubles-brins. *Après la formation d'une cassure double-brin, l'extrémité simple brin d'ADN alors formée va envahir le chromosome homologue. Il existe plusieurs manières de continuer cette réparation. Dans cette figure, deux chemins sont présentés. Soit il y a formation de double jonction de Holliday menant dans la plupart du temps à un évènement de crossing-over, soit on observe une réparation de la forme "synthesis-dependent strand annealing" qui mène à la formation d'un évènement de type non crossing-over. Dans les régions de confrontation entre deux brins de chromosomes différents (carrés gris), il peut parfois y avoir des mismatch entre les séquences ADN menant à de la conversion en faveur de la paire GC au détriment de AT.*

savoir pour quelle(s) raison(s) la recombinaison méiotique a évolué est encore ouverte. En effet, la recombinaison méiotique a-t-elle avant tout évolué dans le but de réussir la méiose ou dans le but de générer du brassage génétique ? Savoir lequel de ces deux aspects est la raison d'être de la recombinaison est une grande énigme. Dans la suite du manuscrit, je me focaliserai sur les mammifères et donc sur la question de l'appariement, laissant ainsi de côté les questions de brassage génétique.

Une question soulevée par la présence de ce mécanisme de recombinaison était de connaître sa distribution à travers le génome, et inversement, connaître cette distribution a joué un rôle clé dans la compréhension des mécanismes impliqués. Pour cela, il est important de décrire au préalable les approches expérimentales utilisées aujourd'hui en vue de détecter des points de recombinaison, et de connaître les mécanismes moléculaires sous-jacents à ce processus.

1.2.4 Méthodes de détection

Il existe plusieurs méthodes pour détecter et cartographier la recombinaison à travers le génome (revue Baudat *et al.* (2013) [7]). On peut tout d'abord utiliser des analyses de pédigrées, qui sont des méthodes utilisant la généalogie. La résolution de la carte est limitée par le nombre d'individus et donc par le nombre de méioses impliquées dans le pédrigrée. Par ailleurs, avant le développement des méthodes de séquençage de nouvelles génération, l'utilisation de puces à ADN limitait la résolution par la densité de polymorphisme (SNP) utilisée sur la puce. Ainsi, ces cartes sont souvent de l'ordre du mégabase. Elles ont par exemple permis d'identifier des variations du taux de recombinaison à l'échelle du génome, avec l'identification de "jungle" (télomères) et de "désert" (centromères) de recombinaison chez l'humain [27; 28; 29]. Chez l'humain, au fur et à mesure des développements technologiques de ces dernières années, l'approche par pedigrées est devenue de plus en plus puissante grâce à des études à grandes échelle [30]. Ces cartes sont par ailleurs sexe spécifique, c'est-à-dire qu'elle ont l'avantage de déterminer dans quelle lignée, mâle ou femelle, ont lieux les CO .

L'approche concurrente à l'analyse de pédigrée est celle consistant à mesurer le déséquilibre de liaison à l'échelle de la population, produisant des estimations de cartes de recombinaison historiques dans la population appelées carte de LD ("Linkage Desequilibrium"). La carte de recombinaison est obtenue par inférence du nombre de CO réalisés entre chaque paire d'allèles. Cette analyse est moins coûteuse que la précédente et permet d'obtenir des cartes à grande résolution avec seulement quelques dizaines d'individus. Cette méthode peut être utilisée sur des espèces modèles, comme l'humain, chez qui cette analyse a permis de générer la première carte de recombinaison à l'échelle du génome entier, ou chez des espèces non modèles, comme le chimpanzé [31]. Elle a aussi été utilisée chez beaucoup d'autres espèces (cf sous-section 1.2.5). Cependant, cette méthode présente quelques limitations. D'une part, elle peut être affectée par d'autres phénomènes que la recombinaison, comme par exemple la sélection. D'autre part, elle ne fournit qu'une estimation moyenne, à travers la population et sur l'echelle de la généalogie

(de l'ordre de 500 000 ans chez l'humain). En particulier, elle ne peut pas démêler les variations entre sexes, entre individus, ou au cours du temps.

Ils existe d'autres méthodes permettant d'obtenir des cartes de recombinaison très précises à très haute résolution et prenant en compte la variabilité entre individus. Chez la levure par exemple, des études de séquençage à grande échelle des spores ont été réalisées [32]. Chez d'autres espèces comme l'humain, des analyses de sperm-typing ont été utilisées, qui ont d'ailleurs permis la première caractérisation directe de régions à fort taux de recombinaison [33]. Bien que cette méthode soit assez précise pour détecter et différencier les évènements de CO et de NCO, ainsi que d'autres processus comme les mutations et l'aneuploïdie, elle reste cependant limitée par le nombre de loci analysables. Par ailleurs, elle ne se focalise que sur la méiose mâle.

Enfin, les méthodes de séquençage dites de "nouvelle génération" ont permis de développer des études de ChIP-seq (immunoprécipitation de chromatine), technique permettant d'identifier les emplacements d'interaction protéine/ADN. En ciblant des marqueurs sur l'ADN associés à des DSB comme H3K4me3 ou H3K36me3 [34; 35; 36], ou des protéines impliquées dans la formation de CO comme DMC1 qui se fixe sur l'extrémité simple brin [34; 35], on a pu identifier des points précis de recombinaison à travers le génome chez différentes espèces. Cette méthode a aussi été utilisée pour cibler la protéine PRDM9, présentée en détail dans la section suivante. En 2016, Lange *et al.* [17] adaptent le protocole Chip-seq pour séquencer les oligonucleotides générés au moment de la DSB et liés de manière covalente à la protéine SPO11, pour l'appliquer chez la souris. Toutefois, cette méthode est très coûteuse. Ces nouvelles méthodes ne permettent pas la différenciation des CO et NCO mais ont permis de démêler les différentes étapes moléculaires détaillées ci-après.

1.2.5 Distribution de la recombinaison

À partir de ces analyses expérimentales, on a découvert que certaines espèces possédaient une recombinaison plutôt uniforme à travers le génome (e.g. *C. elegans* ou la drosophile), tandis que d'autres espèces avaient des points de recombinaison distribués de manière non uniforme le long des chromosomes (revue de Baudat *et al.* (2013) [7]). Cette hétérogénéité de la recombinaison est visible à la fois à l'échelle globale (de l'ordre de la mégabase (mb) ou de la centaine de kilobases (kb)), et de manière locale (à l'échelle du kb). Dans le cas où de très fortes variations locales sont présentes, l'essentiel des évènements, concentrés dans de petites régions de l'ordre de 1kb, représentent une petite proportion du génome. Ces régions sont appelées points chauds ou hotspots de recombinaison car ils caractérisent des régions chromosomiques où les fréquences de CO sont plus élevées que la moyenne à travers le génome ou dans des régions proches [33] (revues [37; 38]). Il n'existe cependant pas de consensus sur la notion de point chaud, car celle-ci dépend du seuil utilisé pour la fréquence de recombinaison en ces points.

Ces points chauds de recombinaison ont d'abord été détectés de manière occasionnelle et individuelle, comme dans la région humaine du HLA [39; 33], ou encore la région

pseudoautosomale [40]. Dans un second temps, l'utilisation de cartes génétiques par déséquilibre de liaison à l'échelle de la population (LD map) a permis de découvrir que les hotspots sont la norme chez l'humain [41], mais aussi chez d'autres espèces eucaryotes (revue de de Massy (2013) [16]) comme les mammifères (le chimpanzé [31], la souris [42] ou le chien [43; 44]), mais également les oiseaux [45], ou encore les plantes telles que *Arabidopsis* [46] et le blé [47].

Parmi les espèces qui ont des points chauds, il existe différents sous-types. En effet, il existe des clades dans lesquels les paysages de recombinaison à fine échelle sont très différents entre espèces soeurs, comme entre le chimpanzé et l'humain par exemple [31], ou même entre sous-espèces, comme observé entre l'humain moderne et Denisova [48]. On pourrait qualifier ces paysages de recombinaison d'instables. Chez ces espèces, les cartes de recombinaison à grande échelle sont assez similaires avec une augmentation de la recombinaison près des télomères, et une réduction près des centromères. Cependant, il existe aussi des clades ayant des paysages de recombinaison à fine échelle, similaires entre espèces proches comme chez les oiseaux ou le chien, qu'on pourrait qualifier de paysages stables [45; 43].

Des études réalisées chez la souris ou l'humain ont permis d'identifier que ces points chauds de recombinaison sont des régions génomiques de 1 à 2 kb de long [33; 49; 41; 42], souvent situées en dehors des gènes. Les études ayant caractérisé des points chauds chez l'humain [49; 41] ou la souris [42] en ont dénombré de l'ordre de plusieurs dizaines de milliers. Il est cependant difficile de donner un nombre absolu de points chauds dans le génome de ces espèces, étant donné que cette quantité dépend fortement du seuil de fréquence de CO utilisé pour définir les points chauds. À noter que les approches par oligos SPO11 (voir plus haut [17]) suggèrent que pas moins de 40% des DSBs semblent avoir lieu en dehors des points chauds, identifiés par ailleurs grâce aux autres méthodes de type Chip-Seq telle que DMC1.

1.2.6 Mechanismes de localisation

Comme montré précédemment, il existe un certain nombre d'espèces chez qui on a détecté la présence de points chauds de recombinaison, mais leur localisation et les mécanismes associés diffèrent suivant la stabilité du paysage de recombnaison. Ainsi, parmi les espèces à paysage stable comme les levures [50] , les plantes, les oiseaux [45] ou le chien [43; 44], les points chauds semblent correspondre à des régions semblables aux régions promotrices des gènes où la chromatine est supposée ouverte, telles que les sites d'initiation de la transcription (TSS) ou les îlots CpG, enrichies en marqueur épigénétique H3K4me3 (tri-méthylation au 4ème résidu de lysine de la protéine d'histone H3) [51; 52; 53]. Ces tri-méthylations d'histones ont aussi été retrouvées aux emplacements de points chauds des espèces à paysage de recombinaison instable [54], comme la majorité des mammifères. Dans ce cas, les points chauds semblent situés en dehors des promoteurs de gènes, et sont par ailleurs caractérisés par un double marquage H3K4me3 et H3K36me3 [36]. Par la suite, je vais me concentrer uniquement sur les paysages de recombinaison instables.

Chez tous les mammifères étudiés jusqu'à aujourd'hui, comprenant la souris, les bovins et les primates dont l'humain, mais pas les canidés, les hotspots se localisent en des positions du génomes caractérisées par la présence de séquences ADN précises. En particulier, chez l'humain, des comparaisons à l'échelle du génome des points chauds détectés par déséquilibre de liaison (LD hotspots) ont détecté la présence d'un certain motif dégénéré (13-mer, CCNCCNTNNCCNC) déterminant jusqu'à 40% des CO [55; 56; 57]. Ceci a permis l'identification du gène responsable de la localisation des hotspots, *PRDM9*. Par la suite, je vais surtout me focaliser sur les espèces possédant *PRDM9*, les autres espèces ne rentrant pas dans le cadre de mon étude. Il devient donc nécessaire de faire une description un peu plus précise de ce gène et de son rôle dans la recombinaison.

1.3 *PRDM9*

PRDM9 code pour une protéine à doigts de zinc désignée en lettre capitales : PRDM9 pour PR domain zinc finger protein 9 [57; 58]. Dans cette partie, je vais tenter de résumer ce que l'on sait sur *PRDM9* (pour aller plus loin : voir la revue de Grey *et al.* 2018 [59])

1.3.1 Les 4 domaines fonctionnels

La protéine PRDM9 possède 4 domaines fonctionnels distincts (cf figure 1.4A). Depuis l'extrémité N-terminale à l'extrémité C-terminale, on trouve dans l'ordre : un domaine KRAB-related (Kruppel-associated box) et un domaine SSXRD (synovial sacroma, X breakpoint repression domain). Ces deux domaines semblent jouer respectivement un rôle d'interaction avec d'autres protéines [60] et de régulation de la transcription [61]. Ensuite, vient un domaine PR/SET possédant une fonction de type histone lysine méthyltransférase [62; 63]. Enfin, vient un domaine de liaison à l'ADN caractérisé par une série de doigts de zinc de type Cystéine(2)-Histidine(2). Ce domaine est codé par un mini-satellite.

Chaque allèle comporte plusieurs doigts de zinc, avec parfois plusieurs copies du même doigt de zinc. Chaque allèle de *PRDM9* est donc caractérisé par sa combinaison spécifique de doigts de zinc, susceptible de reconnaître chacun un motif spécifique [64]. En effet, chaque doigt de zinc possède 3 sites dédiés à la fixation à l'ADN (aux positions -1, 3 et 6 de l'hélice α [57; 65; 66]). Si l'un de ces sites mute, le motif reconnu peut potentiellement changer. Chez l'humain, par exemple, l'allèle *PRDM9* reconnaissant le motif 13-mer attribué à environ 40% des points de recombinaison, possède 13 doigts de zinc [57]. De plus, il existe une forte diversité de combinaisons de doigts de zinc, que ce soit en nombre ou en type (et donc d'allèles *PRDM9*) chez différentes espèces (humain, souris, équidés, bovins et baleine) [57; 64; 67; 68; 69; 70; 71; 72; 73] (revue de Baudat *et al.* (2013) [7]). En effet, chez *Mus musculus*, plusieurs dizaines d'allèles ont été détectés [68; 69]. En 2021, Alleva *et al.* [74] ont détecté 69 allèles (dont 32 nouveaux par rapport aux anciennes études) chez l'humain, chez qui l'allèle A ségrège à une fréquence supérieure à 80% dans les populations non africaines. Les autres allèles ne dépassent jamais les 5% [57].

Ainsi, ces quatre domaines fonctionnels ont chacun un rôle à jouer pendant la méiose,

comme décrit ci-dessous, et semblent être tous requis pour la recombinaison [53].

1.3.2 Mecanisme d'action

Pendant le leptotène, les chromosomes sont alignés le long de l'axe chromosomique, et la chromatine est organisée spatialement sous forme de boucles ancrées à cet axe (cf figure 1.4D) [6]. Une fois le gène *PRDM9* exprimé, les protéines vont se lier à travers le génome sur des motifs séquentiels de 8 à 15 paires de bases, dont la séquence est déterminée par la séquence en acides aminés des doigts de zinc (cf figure 1.4B). Chaque doigt de zinc reconnaît une séquence de trois nucléotides successifs. Ainsi, si le motif ADN est exactement le même que celui qui est reconnu par la protéine, l'affinité de liaison est maximale. À l'inverse, plus il y a de mismatch avec la séquence optimalement reconnue par le domaine à doigts de zinc, plus l'affinité de liaison entre la protéine et l'ADN est faible. Cette affinité protéine-motif a donc un impact assez important sur la capacité qu'a PRDM9 à reconnaître le motif et à s'y lier. D'un point de vue quantitatif, il s'agit du facteur essentiel qui détermine la probabilité d'occupation de PRDM9 à ses cibles. Une fois lié, le domaine PR/SET de PRDM9 triméthyle les histones environnants (H3K4me3 et H3K36me3 [62; 75; 76; 36]), et le site est ramené sur l'axe chromosomique (cf figure 1.4D) [52]. Les mécanismes impliqués lors de cette étape ne sont pas très clairs. Il semblerait cependant que les marqueurs H3K4me3 [35] et les domaines SSXRD et KRAB de PRDM9 y jouent un rôle [77; 78]. Chez l'espèces *S. cerevisiae*, les sites ayant des marques H3K4me3 sont ramenés sur l'axe chromosomique par la protéine Spp1 [79; 52] (revue de de Massy [16]). La méthylation des histones a pour effet de recruter la machinerie de cassure double-brin, constituée en particulier de la protéine SPO11 qui génère alors la cassure [17] (cf figure 1.4C). On dénombre entre 200 et 300 DSB par méiocyte chez la souris [18; 80; 35; 81].

Comme expliqué précédemment, la cassure et le site environnant sont réparés avec pour matrice le site homologue, menant à deux types d'évènements : soit un crossing-over, caractérisé par la réparation du site cassé et par l'échange de matériel génétique entre deux bras chromosomiques homologues (réparation et échange à l'échelle globale, souvent de l'ordre du mégabase), soit un non crossing-over, caractérisé par la simple réparation du site cassé et des ses environs par l'homologue (à l'échelle locale dépassant rarement le kilobase) (revue de Baudat *et al.* (2013) [7]). Dans les deux cas, on observe de la conversion génique, c'est-à-dire un transfert unilatéral de matériel génétique afin de réparer le site cassé. Cette conversion génique est donc biaisée en faveur du site non cassé, qui sert de matrice pour réparer le site ayant subit la cassure. La conversion génique peut également être biaisée d'une seconde manière, lors de la réparation des mésappariement entre la séquence cassée et la séquence matrice. En effet, lors de mismatch entre deux nucléotides, la conversion va être en faveur de la paire de base GC au détriment de la paire AT. On observe alors une conversion génique biaisée vers GC (revue de Duret & Galtier (2009) [82]). Cette dernière conversion génique biaisée ne sera pas prise en compte par la suite.

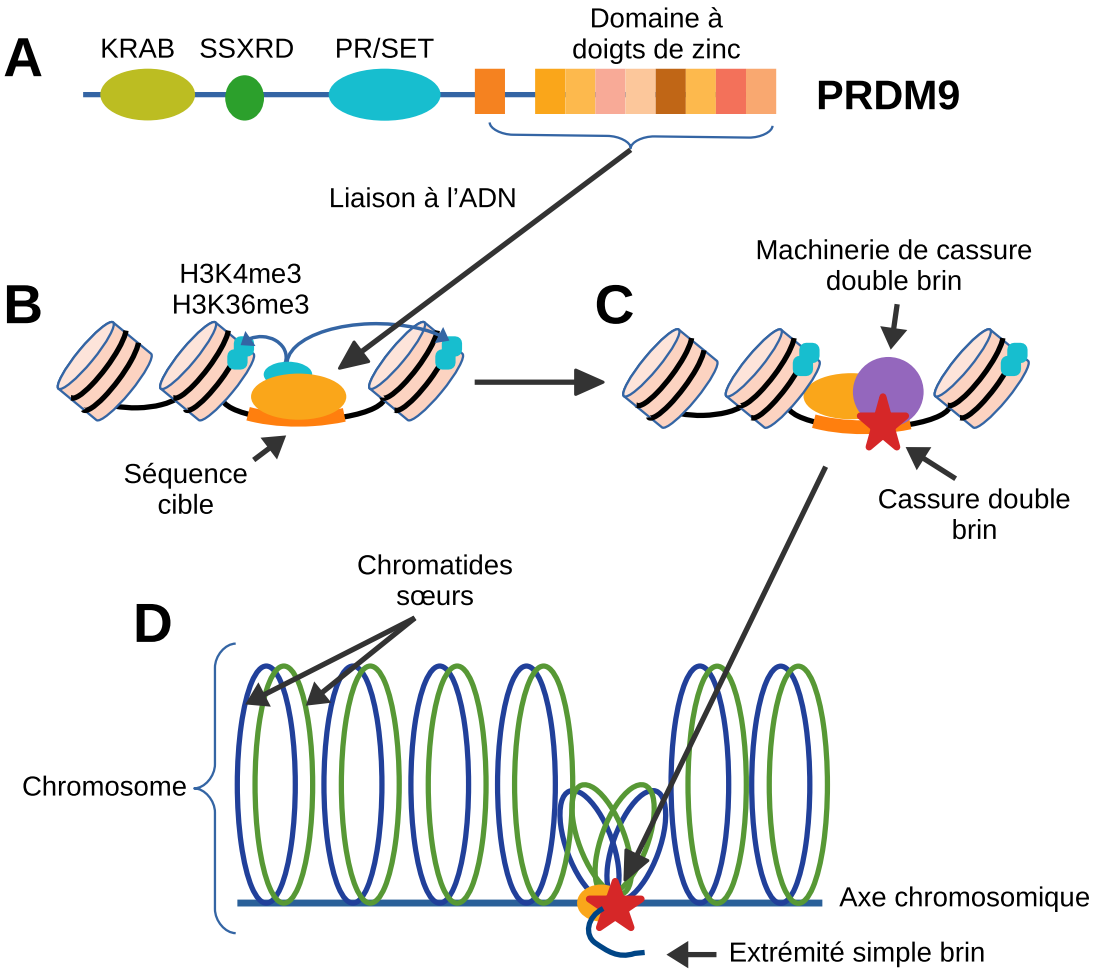

Figure 1.4: Mécanisme d'action de PRDM9. *PRDM9 est une protéine composée de 4 domaines (KRAB, SSXRD, PR/SET et Domaine à doigts de zinc). Elle reconnaît avec son domaine à doigts de zinc un motif ADN spécifique et vient s'y lier. Le domaine PR/SET triméthyle alors les lysines 4 et 36 des histones 3 environnants (petits carrés bleu turquoise). La protéine ramène le site auquel elle est liée sur l'axe chromosomique, et les marqueurs H3K4me3 et H3K36me3 entraînent le recrutement de la machinerie de cassure double-brin (cercle violet) qui réalise la cassure (étoile rouge). Cette cassure génère une extrémité simple brin qui va rechercher la séquence homologue servant à réparer la cassure.*

1.3.3 Paradoxe des points chauds de recombinaison

Toutefois, cette conversion génique pose un problème lorsque l'un des deux sites homologues comporte une mutation qui ne permet plus la liaison avec PRDM9, ou qui diminue l'affinité de liaison. Par la suite, les sites mutés seront appelés sites inactifs. Dans ce cas précis, la protéine va préférentiellement se lier au site actif, menant ainsi au remplacement de cette séquence par le site inactif se trouvant sur le chromosome homologue. On assiste ainsi à de la conversion génique biaisée en faveur du site inactif. À l'échelle de la population, ce biais de conversion mène à une augmentation de fréquence des versions mutantes, ou inactives, des sites de fixation, ainsi qu'à leur fixation dans la population. Ce phénomène mène à l'érosion progressive des sites de liaison de PRDM9 à l'échelle du génome. Cette autodestruction des hotspots est communément appelé le « paradoxe de conversion des points chauds » [83]. Le paradoxe des points chauds implique la disparition de la recombinaison à l'échelle du génome, en un temps assez court (de l'ordre de 10 000 à 100 000 générations), et soulève donc la question de son maintien sur le long terme. À noter qu'à l'origine, le paradoxe des points chauds a été conceptualisé de façon purement théorique, sans le concours de PRDM9 mais de manière assez globale [83]. Ce phénomène est ainsi plus généralement attendu dès lors que le processus de recombinaison implique la reconnaissance d'un motif ADN précis qui se situe à l'endroit même de la cassure, et peut éventuellement disparaître par remplacement par la séquence homologue, si différente, lors de la phase de réparation. Dans quelle mesure ce phénomène serait important dans les espèces sans *PRDM9* est encore une question ouverte.

Le paradoxe des points chauds a suscité beaucoup de réflexions pour sa résolution dans le contexte de *PRDM9*, modèles que je vais présenter par la suite. À cette fin, je présente dans la section suivante la nature et le rôle des modèles mathématiques en biologie évolutive.

1.4 L'élaboration de modèles d'explication

1.4.1 Utilité des modèles mathématiques en biologie évolutive

En 2014, Servedio *et al.* [84] ont réalisé une revue concernant l'utilité des modèles mathématiques dans le domaine de la biologie évolutive. Mon sujet de thèse fait partie de ces modèles théoriques, appelés modèles de preuves de concepts (tout comme les modèles sur lesquels je me base, qui seront présentés dans la suite de l'introduction). Je vais tenter de résumer dans cette partie les points essentiels de cette revue, afin de mettre en perspective les travaux qui seront présentés dans les prochains chapitres, à la lumières de ceux déjà réalisés.

Depuis leur introduction au début du 20ème siècle par Fisher, Wright et Haldane avec la "synthèse évolutionniste moderne", les modèles mathématiques ont pris une importance grandissante dans le domaine de la biologie évolutive. Aujourd'hui, ils sont devenus une forme à part entière de preuve de concept biologique, au même titre que l'observation ou l'expérimentation.

Cette approche théorique est essentielle en biologie évolutive, du fait de l'implication de grandes périodes évolutives, difficilement testables empiriquement, et de la forte complexité des processus à l'échelle de la population, rendant laborieux la formulation d'hypothèses sur les mécanismes impliqués et leurs impacts. Les modèles mathématiques ont par exemple permis de grandes avancées sur différentes questions, comme l'origine du sexe ou le processus de spéciation.

En réalité, il existe deux grands types de modèles mathématiques, et il est important de bien comprendre cette différence pour la suite. Il existe des modèles que l'on pourrait qualifier de prédictifs, et d'autres plus conceptuels. Les premiers sont des modèles que l'on ne peut construire que si l'on dispose d'une bonne connaissance des mécanismes impliqués. Ils sont utilisés lorsqu'il devient important de réaliser des prédictions et comparaisons empiriques. Ces modèles permettent une description quantitative du processus étudié. Ils sont assez souvent faciles à appréhender, et donc bien compris par l'ensemble de la communauté.

Les modèles conceptuels, quant à eux, sont souvent plus abstraits et fournissent une vision qualitative du processus étudié. Ces modèles plus théoriques permettent de vérifier la logique d'un modèle verbal et, en particulier, la cohérence globale de ses hypothèses. En effet, le modèle mathématique nécessite de poser toutes les hypothèses, mêmes celles initialement implicites, ce qui a pour effet de détecter les potentielles erreurs logiques dans le raisonnement verbal. Une fois les hypothèses posées, les résultats qui en découlent fournissent des informations précises sur les conséquences de chaque hypothèse. De manière analogue à la recherche empirique, qui teste des hypothèses en comparant les données recueillies aux résultats prédits, les modèles théoriques comparent les prédictions mathématiques basées sur des hypothèses précises du modèle verbal aux prédictions de ce même modèle verbal. Ainsi, comme pour la recherche empirique, si les résultats du modèle ne correspondent pas aux prédictions, les hypothèses sur lesquelles est basé le modèle (et non pas le modèle en lui même) sont réfutées.

Ces modèles de preuve de concept n'ont donc pas spécifiquement besoin de calibrations empiriques pour démontrer la logique d'un modèle verbal, même si elles sont recommandées. En effet, les comparaison empiriques sont importantes à toutes les phases de la recherche, du fondement des hypothèses (en mettant en lumière les hypothèses cachées) à la discussion (soulèvement de nouveaux questionnements sur les mécanismes, processus et hypothèses étudiées), en passant par la phase de prédiction (prédictions contre-intuitives).

Pour conclure, les modèles de preuve de concept jouent un rôle crucial de renforcement et de levier d'élaboration de la réflexion en recherche en biologie évolutive, sous réserve d'une explication claire et pédagogique de la part des théoriciens et d'une lecture attentive et d'un esprit ouvert de la part des non-théoriciens.

Il est cependant important de noter que la distinction entre ces deux types de modèles n'est pas absolue. En effet, il peuvent être considérés comme les deux extrémités d'un spectre où l'on commencerait par des modèles conceptuels peu à peu renforcés par de nouvelles idées et hypothèses, et rehaussées petit à petit de calibrations empiriques. On aurait donc une sorte de montée graduelle, allant des modèles très conceptuels aux

modèles de prédictions. Dans ce qui va être montré par la suite, le travail réalisé est en grande partie et avant tout à des fins conceptuelles. Un des objectifs important de mon travail de thèse est de savoir où situer le curseur sur ce spectre de modèles, et comment procéder vis-à-vis des prédictions empiriques.

1.4.2 Premières tentatives d'explication du paradoxe des points chauds

Maintenant que le statut de chaque modèle mathématique a été clarifié, je vais détailler les modèles variés développés pour tenter d'expliquer le paradoxe des points chauds présenté précédemment. Le premier modèle construit pour montrer les différentes facettes du paradoxe est bien entendu celui de Boulton *et al.* de 1997 [83]. Par un travail théorique en génétique des populations, Boulton *et al.* démontrent que la sélection qui doit être appliquée sur les hotspots afin de lutter contre la conversion génique biaisée est beaucoup trop élevée pour être réaliste, et que ni le taux de mutation aux cibles, ni les avantages sélectifs des CO, ni la structure du génome ne sont suffisants pour maintenir le paysage de recombinaison. D'autres modèles mathématiques ont alors émergé. Tous ces modèles font intervenir, en plus du paradoxe de conversion des points chauds, un modificateur qui va moduler le mécanisme moléculaire en réponse à l'extinction de la recombinaison. Les modèles diffèrent par le mode d'action supposé du modificateur. Soit le modificateur agit en *cis*, c'est-à-dire qu'il est proche de la cible et cette dernière détermine elle même sa capacité à subir une DSB. Soit le modificateur agit en *trans*, c'est-à-dire indépendemment de la distance et du chromosome. Soit le modificateur agit en *cis* et en *trans* en même temps. Chacun d'eux tente de résoudre le paradoxe mais ne parvient pas à prendre en compte chacune des observations empiriques liées à la recombinaison. L'article de Ubeda & Wilkins de 2011 [85] résume assez bien les différents modèles construits ainsi que les limites de chacun.

Les modèles en *cis*, c'est-à-dire avec des cibles capables d'ajuster leur propre probabilité de DSB couplé à un modificateur déterminant la probabilite de DSB en des cibles proches, pourraient fonctionner si le crossing-over présente un avantage sélectif [86; 87]. Cependant, cela ne permet pas d'expliquer la sur-transmission du site érodé, ni d'expliquer les changements rapides observés dans le paysage de recombinaison. Les modèles en *trans*, prenant en compte un modificateur définissant la probabilité de DSB en une cible situé à n'importe quelle distance sur son chromosome et sur l'homologue, pourraient fonctionner sans besoin de sélection [88], mais ils ne permettent pas d'expliquer le caractère transitoire du paysage de recombinaison, et suggère une surtransmission des allèles chauds, ce qui est contraire aux observations empiriques. Finalement, les modèles agissant à la fois en *cis* et en *trans* sont proposés sans pour autant être convainquants, car ils supposent des contraintes fortes sur l'emplacement du locus modificateur par rapport à ses cibles.

C'est en 2011, dans ce papier d'Ubeda et Wilkins [85], qu'un modèle cohérent est proposé pour la première fois, permettant d'expliquer le maintien du taux de recombi-

naison malgré l'érosion des points chauds, tout en prenant en compte les observations empiriques de l'époque.

1.4.3 Le modele de Reine Rouge

Ubeda & Wilkins : le constat qualitatif

Le modèle mathématique de Ubeda & Wilkins est modèle où un gène "modificateur" (que l'on apparente à *PRDM9*) agit en *cis* (sur des cibles situées sur le même chromosome que le locus du modificateur), et en *trans* (sur des cibles situées sur l'homologue) sur des cibles subissant des DSBs. Il implémente à l'échelle du nucléotide les cibles et le domaine de liaison de la protéine "modificatrice". Ainsi, cela lui permet de faire varier l'affinité de liaison de la protéine à ses cibles par mutations, à la fois dans le domaine de liaison à l'ADN et dans la séquence des cibles. La conversion génique est aussi implémentée par la génération de DSB aux sites liés et la réparation par le site homologue. Enfin, la fertilité dépend mathématiquement du nombre de CO réalisés pendant la méiose.

Le mode de fonctionnement du modèle est présenté en figure 1.5. Supposons qu'un seul allèle ségrège dans la population (dans le cas de plusieurs allèles chacun subit un cycle similaire). Au départ, il y a un allèle dans la population qui reconnaît un certain nombre de sites à travers le génome. Les sites de fixation de cet allèle *PRDM9* sont d'abord érodés par mutation et conversion génique biaisée en faveur des sites inactifs. La disparition des points chauds de recombinaison qui en résulte compromet le bon fonctionnement de la méiose et diminue fortement la fertilité des individus. Dans ce contexte, de nouveaux allèles *PRDM9*, reconnaissant des cibles différentes dans le génome, permettent de restaurer la recombinaison, et sont donc positivement sélectionnés. Ils finissent donc par remplacer l'allèle ou les allèles résidant actuellement dans la population. Ce modèle est analogue à un modèle de Reine Rouge [89] couramment utilisé en biologie évolutive pour les systèmes proies prédateurs [90; 91] ou hôte pathogène [92; 93] par exemple mais cette fois ci, il s'applique à l'échelle intra-génomique. Cette dynamique émerge donc de la compétition entre les deux forces antagonistes que sont l'érosion des cibles par mutation et conversion génique en faveur du site non cassé, et la sélection positive agissant sur le locus *PRDM9*.

Ce modèle a l'avantage de prendre en considération et d'expliquer toutes les observations empiriques de l'époque, tout en modélisant l'affinité de liaison cible/modificateur de manière finie. Je vais résumer dans la partie suivante quelles sont les observations empiriques dont il est question ici.

Élément empirique en faveur du modèle de Reine Rouge

Plusieurs études ont donné des résultats qui allaient dans le sens du modèle de Reine Rouge intra-génomique, tel que proposé par Ubeda & Wilkins [85]. Tout d'abord, des études réalisées sur des espèces proches phylogénétiquement ont montré qu'il y a peu de similarités dans les paysages de recombinaison entre ces espèces, comme entre l'humain et

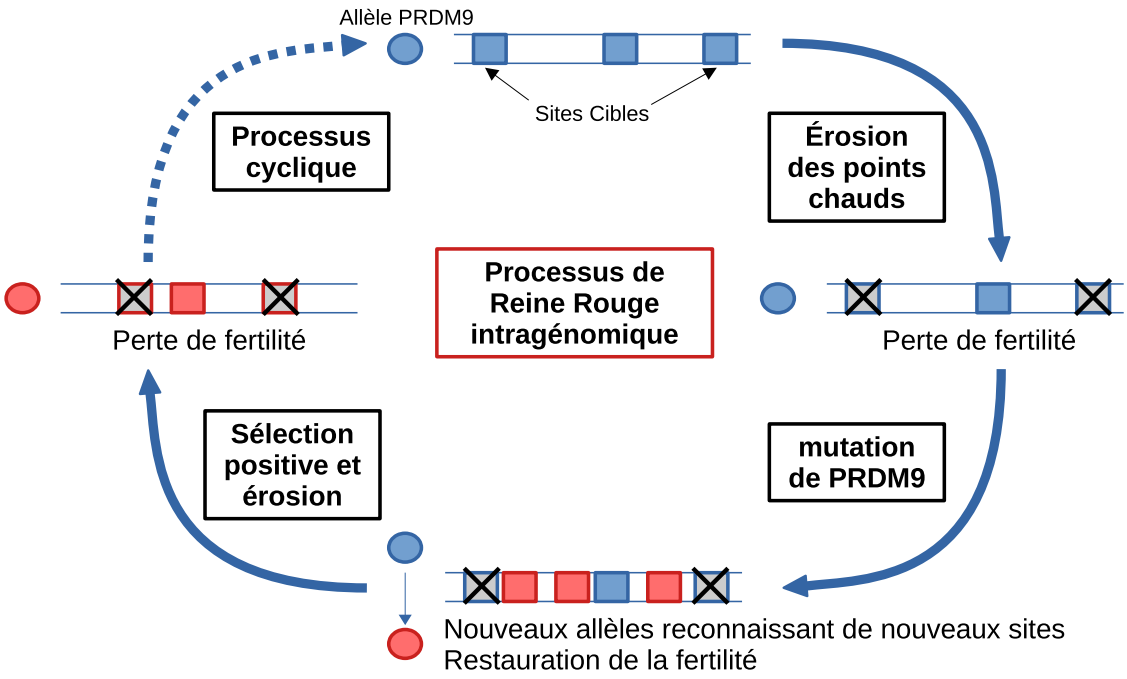

Figure 1.5: Processus de Reine Rouge intra-génomique. *Le processus de Reine Rouge proposé pour résoudre le paradoxe des points chauds fonctionne comme suit : on commence avec un allèle reconnaissant des cibles à travers le génome, il érode progressivement ces cibles par l'action de la conversion génique biaisée en faveur des sites mutés, ce qui génère de la perte de fertilité. Dans cet état, lorsqu'un nouvel allèle apparaît par mutation au locus PRDM9, ce dernier est positivement sélectionné car il reconnaît de nouvelles cibles, ce qui restaure la fertilité. Cet allèle va alors envahir la population, éroder ses cibles, puis être remplacé, et ainsi de suite.*

le chimpanzé [58; 31] ou entre l'humain moderne et Denisova [48]. De plus, il a été montré que le domaine à doigts de zinc, codé par un mini-satellite, mute très rapidement [94] ce qui génère de nouveaux allèles *PRDM9*. Ces nouveaux allèles émergent soit par la création d'un nouveau doigt de zinc par mutations des résidus permettant la liaison à l'ADN, soit par évolution concertée due aux CO inégaux dans la séquence codée par minisatellite, ce qui génère de nouveaux allèles possédant moins de diversité en terme de doigt de zinc. Enfin, des analyses de génomique comparative ont permis d'établir, d'une part, que les fixations de substitutions non-synonymes dans le domaine à doigts de zinc de PRDM9 sont plus nombreuses que les substitutions synonymes, ce qui suggère la présence d'une forte sélection positive agissant sur *PRDM9* [68; 95], et d'autre part, que cette sélection positive semble n'être présente que chez les espèces dont PRDM9 possède à la fois un domaine SET (responsable de la méthylation des histones) et un domaine KRAB intacts [53]. Aussi, *PRDM9* a été montré comme très polymorphe chez différentes espèces comme la souris (plus d'une centaine d'allèles [68; 69]) ou l'humain ([57; 64; 67; 95]), avec deux allèles dominants (A et B), et d'autres allèles en très faible fréquence dans la population.

Le modèle de Reine Rouge proposé en 2010 prend donc en compte l'instabilité des paysages de recombinaison, ainsi que la sélection positive appliquée au locus *PRDM9*, comme observés empiriquement. Cependant, ce modèle manque de précision et d'explications sur certains aspects de la dynamique. Premièrement, le modèle ne permet pas d'expliquer si le fort polymorphisme de *PRDM9* observé empiriquement chez plusieurs taxa est dû au taux de mutation de *PRDM9* ou à une sélection diversifiante. Deuxièmement, la dépendance du régime de Reine Rouge à l'équilibre en terme de niveau d'érosion ou de diversité génétique en fonction des paramètres (taux de mutations aux cibles ou au locus *PRDM9* et taille de population) n'est pas détaillée dans ce modèle. Cette étude pourrait permettre d'étudier la dynamique pour un plus grand nombre de taxa et pas seulement les souris et les humains. En effet, il a été montré que *PRDM9* est ancestral aux métazoaires et que son rôle dans la recombinaison a été retrouvé chez beaucoup de vertébrés [53] (voir box "*PRDM9* à travers l'arbre des métazoaires"). Ce modèle, qui donne une vision qualitative du modèle de Reine Rouge, ne tente pas de calibration empirique.

Latrille *et al.* : Étude de la réponse des mécanismes en fonction des paramètres

Latrille *et al.* [96] ont développé un nouveau modèle théorique en 2017 pour essayer de palier aux manquements du modèle précédent. Ils ont dans un premier temps créé un simulateur de la Reine Rouge intra-génomique exécuté sur un large éventail de conditions afin d'explorer les différents régimes à l'équilibre de la Reine Rouge. Les résultats obtenus ont ensuite été étayés par des approximations analytiques et numériques. L'article comprend aussi une tentative de calibration empirique du modèle avec les données obtenues chez la souris. Ce travail apporte plusieurs points importants à l'étude du modèle de Reine Rouge par rapport au précédent modèle d'Ubeda & Wilkins (2011) [85]. D'une part, une approche originale par champs moyen autoconsistant, issu du domaine de la

physique, permet de trouver quantitativement un équilibre. D'autre part, les auteurs ont réalisé un effort de calibration empirique, ce qui est relativement rare dans les modèles en biologie théorique. Grâce à cela, le modèle de Latrille *et al.* a permis quelques avancées sur la compréhension et le fonctionnement du modèle de Reine Rouge. Premièrement, les simulations ont permis de mieux identifier les différents régimes que peut prendre la dynamique de Reine Rouge, tandis que les calibrations empiriques ont permis de mieux identifier les relations de dépendance entre les statistiques observables telles que la diversité génétique ou le taux de recombinaison moyen en fonction des paramètres intrinsèques du modèle. Enfin, les calibrations empiriques ont montré qu'un fort taux de mutation au locus *PRDM9* et une forte conversion génique biaisée étaient nécessaires pour obtenir une diversité et un niveau d'érosion tels qu'observés empiriquement. Cependant, ce modèle, malgré ses apports importants et originaux vis à vis des anciens modèles, reste néanmoins très peu informé sur les mécanismes moléculaires de la méiose.

1.4.4 Perspectives

Les modèles précédents, bien que permettant une meilleur compréhension du processus de la Reine Rouge, ne sont encore qu'au stade assez conceptuel. En effet, le modèle d'Ubeda et Wilkins est très abstrait et celui de Latrille *et al.*, malgré sa tentative de calibration empirique, n'est pas encore très clair sur les mécanismes moléculaires impliqués. Ainsi, pour aller plus loin du point de vue empirique et des conséquences phénoménologiques du modèle, que faut il prendre en considération ? En particulier, les précédents modèles ne précisent pas comment l'érosion des paysages de recombinaison implique des baisses de fertilité. Ainsi, l'une des faiblesses des modèles proposés précédemment concerne le lien exact, pour le moment posé de manière arbitraire, entre le niveau d'érosion et la perte de fertilité pour lequel il n'y a aucune explication précise. Or, sur ce point, de grandes avancées empiriques à partir de 2016 on permis de donner des pistes sur les mécanismes impliqués. Ces avancées réalisées chez des hybrides ont d'ailleurs mis en lumière le rôle que *PRDM9* pourrait jouer dans la stérilité hybride. Développer ces points très importants demandent donc de faire un petit détour par les mécanismes de la stérilité hybride.

1.5 *PRDM9* et stérilité hybride

Il s'est avéré que *PRDM9* joue un rôle dans la stérilité hybride, et c'est en comprenant ce qu'il se passe chez les hybrides qu'on a compris les mécanismes mêmes de la recombinaison avec *PRDM9* en général. Aussi, dans cette section, je vais brièvement introduire la stérilité hybride avant d'en venir au rôle spécifique que joue *PRDM9* dans cette forme d'isolement reproductif, puis je préciserai ce qu'on en a déduit sur le mécanisme de la recombinaison dépendante de *PRDM9*.

PRDM9 à travers l'arbre des métazoaires

PRDM9 est un gène ancestral des métazoaires, conservé chez beaucoup d'espèces, mais perdu chez d'autres [95; 53] (cf figure 1.6). On le retrouve chez la plupart des mammifères étudiés (par exemple chez la souris, les primates dont l'humain, et chez les bovidés), mais pas chez les canidés, où *PRDM9* est devenu un pseudogène [43; 97]. Plus largement, le gène semble avoir été perdu chez les oiseaux, crocodiles et amphibiens, ou du moins, la protéine codée semble avoir perdu des domaines fonctionnels chez certains poissons et chez les monotrèmes. Chez les primates, un paralogue de *PRDM9* a été identifié (*PRDM7* [98]).

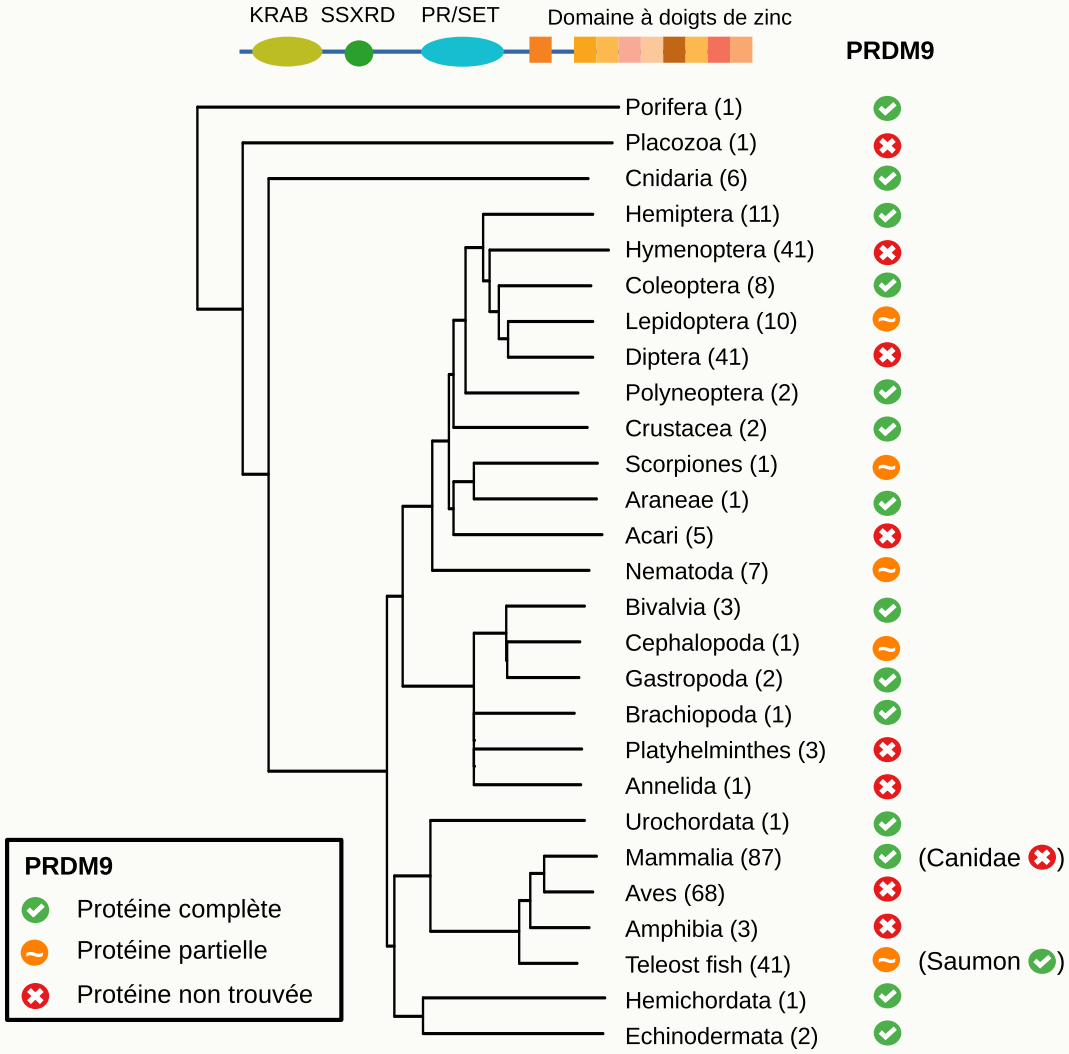

Figure 1.6: Arbre phylogénétique de la présence/absence d'homologues de PRDM9 chez les métazoaires. *Les clades possédant au moins un orthologue de PRDM9 complet (comprenant les domaines KRAB, SSXRD, SET et doigts de zinc) sont indiqués avec un check dans un cercle vert (le nombre de génomes analysés est donné entre parenthèses). Les homologues PRDM9 partiels (KRAB ou SSXRD manquants) sont indiqués par des tildes dans un cercle orange. Enfin, les clades dans lesquels aucun homologue de PRDM9 n'a été trouvé sont indiqués par une croix dans un cercle rouge. Cette figure est adaptée d'une figure non publiée de Laurent Duret.*

1.5.1 Stérilité hybride

Le processus de spéciation permet de passer d'une situation avec une espèce (population où tous les individus sont interfertiles ainsi que leur descendants) à deux ou plusieurs espèces (groupe de populations dont les croisements à une génération donnée ou dans les quelques générations suivantes sont stériles ou non viables). Les barrières reproductives peuvent être antérieures à la fertilisation (prézygotique) ou postérieures (postzygotique). Dans le premier cas, on distingue les isolements se produisant avant l'accouplement, comme des barrières dues à l'habitat, la période d'accouplement ou le comportement, des isolements post-accouplements comprenant les barrières dues à des incompatibilités entre gamètes ou entre mécanismes biologiques. Dans le deuxième cas, l'isolement reproductif se retrouve chez les hybrides issus de croisements entre deux populations. Ces hybrides, ou leurs descendants, sont stériles ou non viables. *PRDM9* semblant être impliqué dans la stérilité hybride, je vais par la suite me focaliser uniquement sur cette forme d'isolement reproductif.

Modèles génétiques

Depuis longtemps, on a identifié la stérilité hybride comme un mécanisme clé de la spéciation. La stérilité hybride a été décrite par Darwin comme un mécanisme réduisant le flux de gènes entre des taxons apparentés [99]. En effet, un processus assez simple pour en arriver à cette stérilité hybride est le suivant. On part d'une population unique qui devient structurée après l'apparition d'une barrière, par exemple géographique. Cette barrière a pour effet de réduire partiellement ou totalement la migration entre les populations, ce qui, au niveau génétique, se traduit par une baisse de flux de gènes et donc d'homogénéisation génétique indépendante des deux populations.

Malgré l'intérêt massif porté à cet isolement reproductif chez plusieurs espèces, il reste encore beaucoup à comprendre. Depuis le milieu du 20ème siècle, le modèle généralement accepté pour expliquer génétiquement la stérilité hybride est le modèle d'incompatibilités de Bateson-Dobzhansky-Muller (BDMI, communément appelé le modèle Dobzhansky-Muller) [100; 101; 102; 103; 104]. Ce modèle explique la stérilité hybride comme une incompatibilité entre les gènes en interaction qui découle de leur évolution indépendante dans deux populations (détails de l'explication de ce modèle dans la Box "Modèle des incompatibilités Bateson-Dobzhansky-Muller"). Ce processus est graduel, et l'accumulation des ces incompatibilités va donc mener à une baisse progressive de la fertilité des hyrbides, voire même à de la stérilité totale. À noter qu'en parallèle, d'autres traits comportementaux peuvent évoluer et être sélectionnés pour éviter ces croisements défaillants. Ce renforcement accroît alors l'évolution indépendante et les incompatibilités, ce qui mène à terme à la formation de deux espèces indépendantes.

Modèle des incompatibilités Bateson-Dobzhansky-Muller

Dans la première moitié du siècle dernier, plusieurs scientifiques ont proposé la première explication génétique de la stérilité hybride, un modèle qui maintenant porte leurs noms, le modèle d'incompatibilité Bateson-Dobzhansky-Muller (BDMI). Ce modèle explique la stérilité hybride comme des incompatibilités entre gènes dans le contexte hybride qui ne se retrouvent pas en contexte unipopulationnel (cf figure 1.7). Plus précisément, nous considérons deux gènes en interaction dont les allèles A et B sont fixés dans la population ancestrale (AABB). Après séparation en deux populations distinctes sans possibilité d'interaction entre elles, les deux gènes subiront des mutations différentes. Certaines de ces mutations peuvent finir par se fixer dans l'une ou l'autre des deux populations en raison de la dérive génétique ou de la sélection naturelle en réponse au nouvel environnement auquel cette population est confrontée. On pourrait donc se retrouver avec un allèle a dans la population 1 pour le premier gène (aaBB), et un allèle b dans la population 2 pour le second gène (AAbb). Si l'hybridation devait se produire entre les deux populations, les hybrides (AabB) se retrouveraient dans une configuration allélique encore inconnue du point de vue de la sélection naturelle (a/b), qui pourrait conduire à des incompatibilités entre les deux gènes.

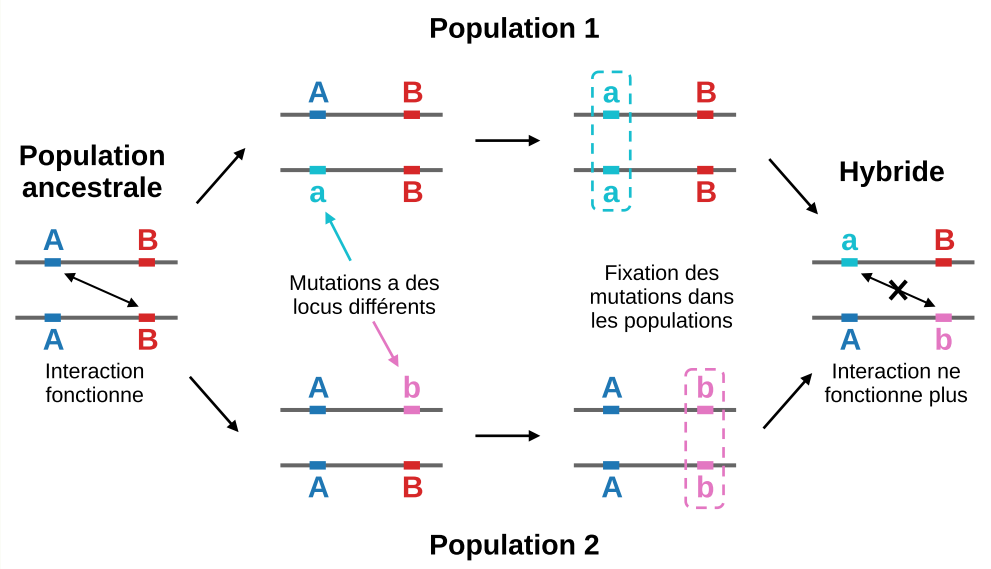

Figure 1.7: **Schéma du modèle d'incompatibilités de Bateson-Dobzhansky-Muller**

Gène de spéciation

Les gènes impliqués dans les incompatibilités décrites précédemment ont été nommés gènes de spéciation. Néanmoins, même si de nombreuses études ont été menées chez de nombreuses espèces (en particulier chez la Drosophile [105; 106] mais aussi chez d'autres

eucaryotes), il est encore très compliqué d'identifier de tels « gènes de spéciation », puisqu'on ne sait pas si ces gènes sont responsables de la stérilité hybride aux premiers stades de la spéciation ou s'ils ont évolué après la spéciation comme mécanisme de renforcement. À ce jour, peu de gènes de spéciation ont été découverts. La majorité a été trouvée chez la Drosophile comme OdsH [107; 108], Ovd (Overdrive) [109] ou Nup96 [110], mais aucun d'entre eux ne semble avoir un gène partenaire avec lequel ils pourraient interagir et générer une incompatibilité comme expliqué dans le modèle BDMI. Un gène candidat a été trouvé chez la souris, le seul chez les vertébrés, qui n'est autre que *PRDM9* [75; 111].

1.5.2 *PRDM9* en contexte hybride

Comme précisé à plusieurs reprises dans les parties précédentes, *PRDM9* semble être un gène susceptible d'induire de la stérilité hybride. Je vais donc présenter dans cette partie comment on a découvert l'implication de ce gène dans la stérilité hybride, et ce que cela implique sur la compréhension du mécanisme de recombinaison et ses conséquences sur les modèles.

Historique

Chez la souris *Mus muscuslus*, on distingue plusieurs sous espèces très structurées. Il y a en particulier *Mus musculus castaneus* se situant en Asie, qui a divergé il y a environ 500 000 ans [112; 113] des sous espèces *Mus musculus musculus* (*M. m. musculus*) et *Mus musculus domesticus* (*M. m. domesticus*) qui ont migré pour leur part à l'ouest. Lors de cette migration vers l'Europe, une population est passée par la côte méditerranéenne, *Mus musculus domesticus* et l'autre, *Mus musculus musculus*, est passée par le Nord. Il y a quelques milliers d'années, pendant l'Holocène (revue de Macholan *et al.* (2012) [114]), ces deux populations se sont retrouvées dans une zone hybride en Europe centrale (cf figure 1.8).

C'est dans cette zone hybride qu'en 1969 Ivanyi *et al.* [115] ont détecté de la stérilité hybride (arrêt précoce de la spermatogénèse) par croisement de souris sauvages de la sous espèces *M. m. musculus* avec des souris de la souche C57BL/10 de *M. m. domesticus*. Quelques années plus tard, en 1974, Forejt et Ivanyi [116] on découvert que le locus *Hst1* (Hybrid sterility 1) situé sur le chromosome 17 de la souris était impliqué dans la stérilité hybride chez certains croisements des sous-espèces *M. m. musculus* et *M. m. domesticus* [116]. Par la suite, plusieurs études sur des hybrides issus de croisements entres ces deux sous espèces ont été réalisées. En particulier, des croisements entre PWD/Ph (PWD) de *M. m. musculus* dérivée de la lignée sauvage [117] et C57BL/6J (B6), une souche de laboratoire de *M. m. domesticus*, ont montré que chez l'hybride F1 des croisements PWDxB6 (femelle PWD et mâle B6), les mâles sont stériles mais pas les femelles (comme le stipule la loi de Haldane, cf box "La règle de Haldane"). Le croisement inverse (B6xPWD) montre cependant des mâles semi-fertiles et des femelles totalement fertiles. Ce n'est qu'à la fin des années 2000 que le locus Hst1 a été attribué

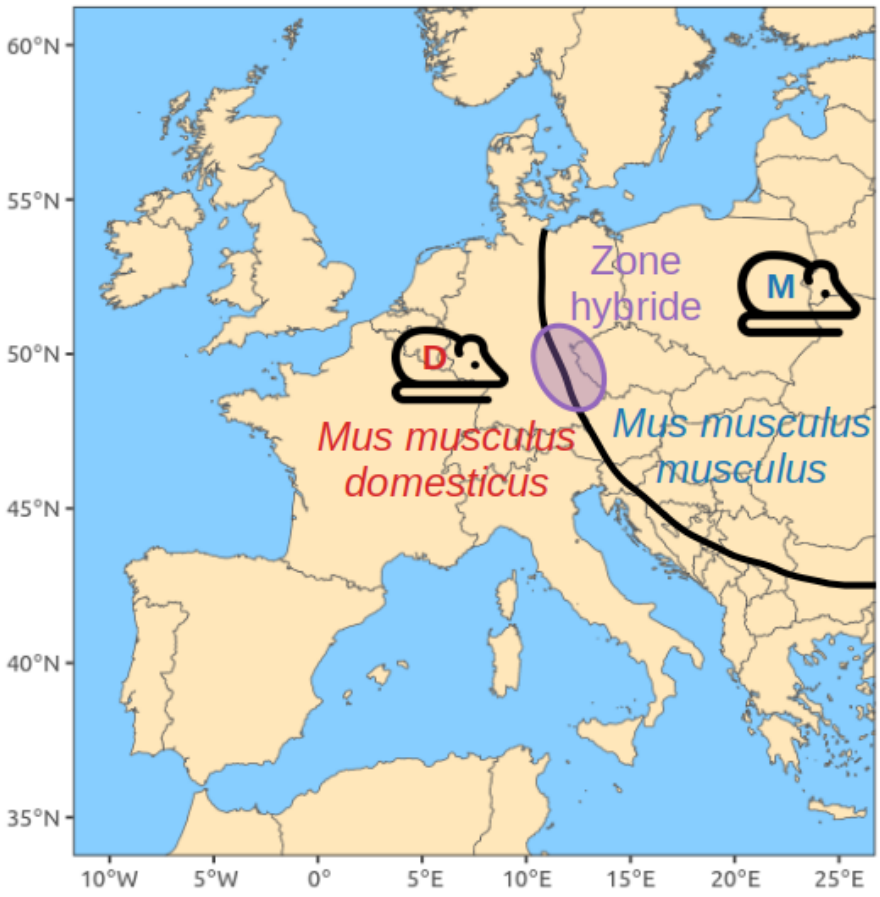

Figure 1.8: Carte de zone hybride (ovale violet) entre Mus musculus musculus (souris M en bleu) et Mus musculus domesticus (souris D en rouge). *Icônes de souris créées par Ayub Irawan - Flaticon.com*

au gène *Prdm9* [75] et a d'ailleurs par la même occasion été détecté comme gène avec une fonction de triméthylation des histones, corroborant des études précédentes [62] sur le même locus alors appelé Meisetz (Meiotic gene with SET/PR domain and Zinc fingers). Le gène *Prdm9* a donc été proposé comme gène de stérilité hybride, voire même de spéciation, chez la souris [75].

La règle de Haldane

Selon la règle de Haldane, formulée en 1922 par le biologiste évolutionniste britannique J. B. S. Haldane, « lorsque dans la progéniture F1 de deux races animales différentes, un sexe est absent, rare ou stérile, ce sexe est le sexe hétérozygote (sexe hétérogamique) » soit chez les mâles (XY (mammifères) ou le sexe de type XO (drosophile)) ou chez les femelles (ZW (oiseaux) ou le sexe de type Z0 (papillon)) [118].
Cette règle a été observée empiriquement chez de nombreuses espèces de différents groupes taxonomiques tels que les mammiferes, les diptères, les orthoptères, les amphibiens, les oiseaux, les lépidoptères ou les reptiles (revue de Schilthuizen *et al.* (2011) [119]).

Plusieurs théories ont été avancées afin d'expliquer cette règle avec deux règles qui semblent être majoritaires :
- La théorie de la dominance [120; 121; 122] : Les hybrides hétérogamiques sont affectés par tous les allèles X-liés (récessifs ou dominants) ce qui provoque des incompatibilités de Bateson-Dobzhansky-Muller (cf encadré "Modèle des incompatibilités Bateson-Dobzhansky-Muller").
- "The faster male theory" [123] : En supposant que les gènes mâles évoluent plus rapidement en raison de la sélection sexuelle, la stérilité est plus souvent observée chez les hybrides mâles hétérogamétiques.

Observations empiriques

Plusieurs études se sont focalisées sur la compréhension du mécanisme impliqué dans cette stérilité hybride. En 2010, on a découvert que la fonction du gène *PRDM9* était de cibler la position des points chauds [57; 58], et que l'activité de ce gène était associée à l'érosion des cibles [58]. C'est d'ailleurs cette force d'érosion qui a été proposée comme cause de la sélection positive pour les nouveaux allèles *PRDM9*, ce qui entraîne la Reine Rouge intra génomique (cf section 1.4.3). Cette érosion a également été suspectée de jouer un rôle dans l'observation de stérilité hybride. En particulier, l'érosion différentielle de sites de fixation de PRDM9 entre sous-espèces (différents allèles *PRDM9* érodant chacun leurs sites au sein de chaque sous-espèce) aboutit à des configurations très asymétriques chez les individus hybrides [124]. En effet, chez un hybride, chacun des deux allèles *PRDM9* a érodé ses cibles dans la population d'origine, mais non dans l'autre population. Chez un individu diploïde, quand PRDM9 se fixe à un locus, il est susceptible de se fixer sur les deux copies de la paire de chromosomes. Lorsque c'est le cas, on parle alors de liaison symétrique de PRDM9 (cf figure 1.9). À l'inverse, dans le cas où la protéine se fixerait sur un seul des chromosomes, on parle de liaison asymétrique (cf figure 1.9). Ainsi,

lorsqu'une des cibles à une position donnée est érodée, caractérisée par la présence de polymorphisme sur le site de liaison, cela empêche la liaison de la protéine à cette cible, ce qui peut être cause de liaison asymétrique. Par conséquent, lors d'érosion différentielle sur les chromosomes homologues, il y a liaison globalement asymétrique de PRDM9 à travers l'ensemble du génome.

Or, l'asymétrie de liaison semble être impliquée dans la stérilité. En effet, des études de Gregorová et Forejt en 2000 [117] et Gregorová *et al.* en 2018 [125] menées sur deux lignées de souris différentes, *M. m. domesticus* et *M. m. musculus*, ont permis d'observer, par mesure expérimentale des taux d'asymétrie de liaison PRDM9 par Chip Seq chez une variété d'hybride pour différentes paires d'allèles *Prdm9*, une corrélation entre taux d'asymétrie et stérilité. De façon analogue, Valiskova *et al.* en 2022 [126] ont récemment observé de la stérilité hybride chez des hybrides mâles issus de croisements de *M. m. musculus* et *Mus musculus castaneus*. Or, des observations cytogénétiques montrent que les chromosomes les plus asymétriquement liés sont moins souvent correctement appariés en métaphase I [127; 111; 128]. Enfin, des études de sauvetage de l'hybride par transgénèse réalisées par Davies *et al.* en 2016 [111] ont montré que *Prdm9* était fortement lié à la stérilité de l'hybride. En effet, l'introduction par transgénèse chez la souris d'un allèle *Prdm9* possédant le domaine à doigts de zinc humain, donc d'un allèle qui n'a pas été présent et n'a pas érodé ses cibles de façon asymétrique dans l'une ou l'autre des deux populations, restaure à la fois une bonne symétrie de liaison et des niveaux élevés de fertilité chez la souris.

Il faut cependant noter que ces phénomènes de stérilité hybride ne sont pas généralisés à tous les croisements. En effet, seulement certains croisements portant des allèles *Prdm9* spécifiques sont responsable de stérilité. On peut citer entre autres les croisements entre allèles $Prdm9^{dome2}$ et $Prdm9^{dom3}$ chez *M. m. musculus* et les allèles $Prdm9^{msc1}$, $Prdm9^{msc2}$ et $Prdm9^{msc5}$ chez *M. m. domesticus*. Il semblerait que l'interaction avec un allèle spécifique au locus Hstx2 (Hybrid sterility X Chromosome 2) situé sur le chromosome X soit aussi impliquée dans la stérilité hybride [129; 130; 131; 132; 126].

Proposition du modèle de liaison symétrique

À la lumière de ces résultats empiriques récents, un modèle a été proposé, selon lequel PRDM9 aurait en fait un double rôle pendant la méiose : la protéine PRDM9 serait non seulement responsable du recrutement de la machinerie de cassure double-brin, mais également, grâce à sa liaison symétrique, elle faciliterait l'appariement entre deux chromosomes homologues. En effet, seulement 200 DSB se réalisent sur les 5000 sites liés par PRDM9 dans un méiocyte [80]. Comme indiqué plus haut, chaque cassure entraîne la formation d'une extrémité simple brin, qui doit ensuite trouver la région à laquelle s'apparier, afin d'être réparée, soit sur le chromosome homologue (cas le plus courant et essentiel à la réussite de la méiose), soit sur la chromatide sœur [25] (en particulier sur les chromosomes X et Y lors de méioses mâles, ou ailleurs dans le génome lorsque la cassure peine à se réparer avec l'homologue). Cette étape de la méiose est critique car c'est elle

qui va permettre aux deux chromosomes homologues de s'associer, de réaliser un CO et de former leur synapse [57]. Aussi, cette étape semble être le facteur limitant puisque la recherche d'homologue se réalise sur de très longues séquences. Dans ce contexte, PRDM9 agirait pour ramener les sites cibles auxquels elle se lie sur l'axe chromosomique. C'est ici que la symétrie joue un rôle clé : lorsque PRDM9 est liée symétriquement, les deux loci homologues sont chacun rapprochés de l'axe chromosomique, facilitant la recherche de la séquence complémentaire de l'extrémité simple brin [52]. A l'inverse, dans le cas où PRDM9 ne se fixe que sur un seul des deux loci homologues, une éventuelle cassure double-brin serait soit réparée préférentiellement avec la chromatide sœur plutôt qu'avec l'homologue, trop éloigné, soit réparée plus tard, préférentiellement en NCO, une fois que la synapse a eu lieu (et ce, grâce à un appariement initié en un autre site, symétriquement lié), soit pas réparée du tout, ce qui déclencherait l'échec de la méiose. Ainsi, selon ce modèle, la fixation symétrique de PRDM9 serait nécessaire pour l'appariement des chromosomes homologues, étape indispensable à la formation de CO, et donc à la réussite de la méiose [125]. L'asymétrie des liaisons PRDM9 serait une nouvelle cause de stérilité hybride s'ajoutant à la différence de caryotypes, d'hétérogamétie ou de mauvaises associations d'allèles. Cette protéine pourrait aussi jouer un rôle important au début du processus de spéciation [75; 95], par l'intermédiaire de cette stérilité hybride due aux liaisons asymétriques [111].

À ce stade, je vais résumer tout ce que l'on sait sur la recombinaison dépendante de *PRDM9*. On sait que *PRDM9* joue un rôle de localisation des points chauds de recombinaison à travers le génome. De plus, ces points chauds sont soumis à un paradoxe (celui du maintien des hotspots dans le génome malgré leur autodestruction) qui a été résolu mathématiquement, d'abord par le modèle de Reine Rouge proposé par Ubeda & Wilkins [85], puis amélioré par Latrille *et al.* en 2017 [96]. Cependant, ces modèles restent silencieux sur la provenance de la perte de la fertilité. Par ailleurs, on a trouvé un rôle de la symétrie de liaison de PRDM9 aux points de recombinaison. Il devient alors nécessaire de réconcilier le processus de Reine-Rouge avec les nouvelles connaissances sur les mécanismes moléculaires obtenues en contexte hybride. C'est pourquoi nous avons décidé de centrer mon travail de thèse sur la création d'un modèle prenant en compte ces deux informations, constituant ainsi mon Axe 1 (article sur BioRxiv [133] soumis a PLOS Genetics). En parallèle de mon travail, Zackary Baker a développé son propre modèle [134] qui sera discuté à la fin de l'Axe 1. Dans un deuxième temps, j'ai adapté mon modèle afin d'étudier si les mécanismes moléculaires de symétrie de PRDM9 provoquaient bien de la stérilité hybride.

1.6 Plan de thèse

Mon travail de thèse a consisté en l'étude théorique de l'impact des mécanismes moléculaires de PRDM9 pendant la méiose sur la dynamique évolutive de la recombinaison. Il se décompose en deux axes, l'un en population unique, le second en contexte hybride.

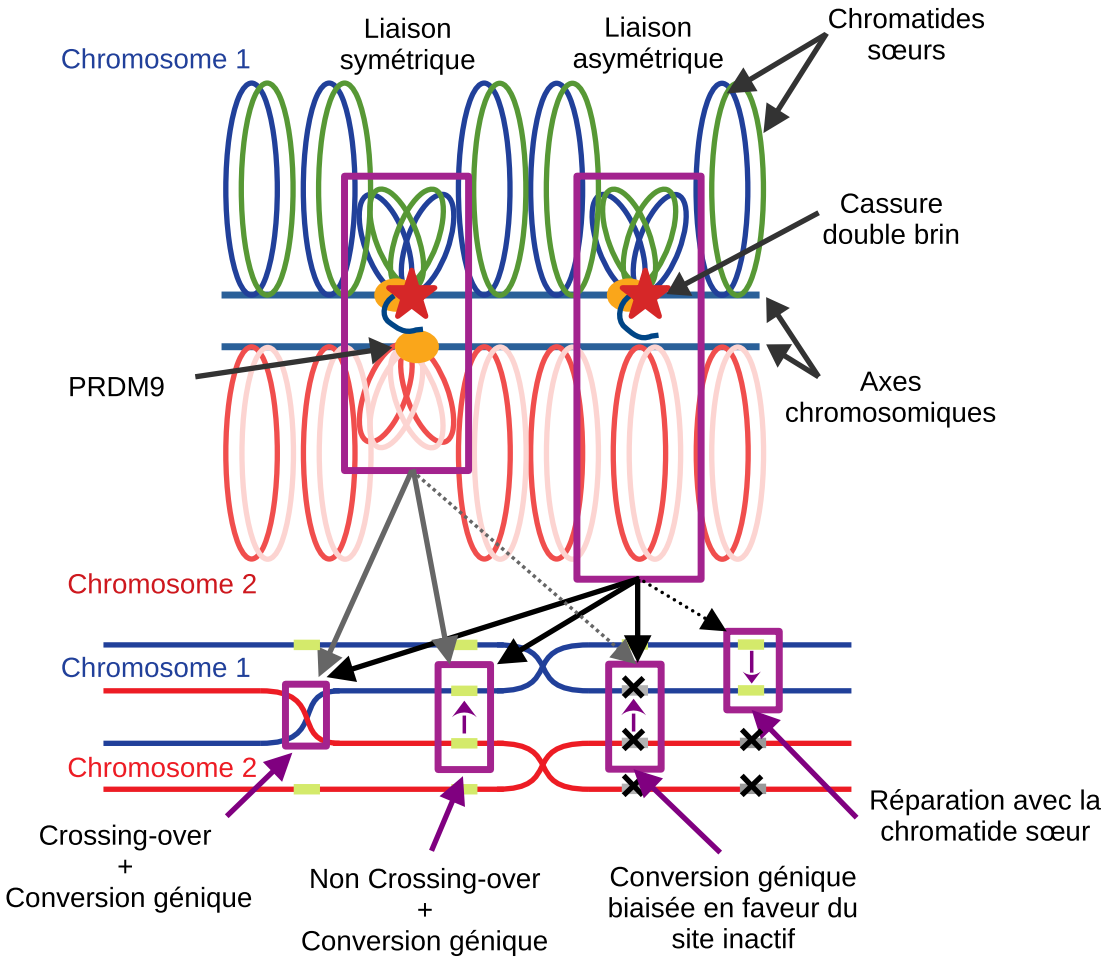

Figure 1.9: Liaison de PRDM9 et modes de réparation. *La liaison de PRDM9 peut être symétrique, liaison sur les deux chromosomes homologues à la même position, essentielle pour la réussite de la méiose, soit asymétrique, liaison sur un seul des deux homologues. Dans le premier cas, les deux cibles sont liées donc actives ce qui donne naissance soit à un évènement de type crossing-over soit non crossing-over. Dans les deux cas on observe de la conversion génique mais lorsque les deux sites sont identiques, elle n'a pas d'impact. Dans le cas de liaison asymétrique, la réparation peut résulter en un crossing-over ou en un évènement non crossing-over associé à de la conversion génique biaisée en faveur du site non cassé qui peut parfois être inactif (possibilité plus rare dans le cas de liaison symétrique). La liaison asymétrique peut aussi être réparée dans de rare cas par la chromatide soeur.*

1.6.1 Premier Axe

Dans un premier Axe, intitulé "Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis : a model based exploration of the *PRDM9* intra-genomic Red Queen", nous avons développé un simulateur informatique modélisant l'évolution du gène *PRDM9* et de ses cibles chez les individus d'une population unique évoluant sur plusieurs dizaines de milliers de générations. Ce travail de simulation, complémenté par des développements analytiques basés sur le principe de champs-moyen auto-consistant, ont permis d'identifier la cause de la sélection positive appliquée sur les nouveaux allèles *PRDM9*, et de quantifier l'impact des paramètres du modèle sur la dynamique de Reine Rouge. De plus, une tentative de calibration empirique a été réalisée. Cette calibration démontre qu'il reste des zones grises sur lesquelles on manque encore de connaissances empiriques qui seront discutées.

Cet axe a fait l'objet d'un article soumis sur BioRxiv [133] puis au journal PLOS Genetics.

1.6.2 Deuxieme Axe

Dans un second temps, nous avons adapté le simulateur développé dans l'Axe 1 afin de le faire tourner en contexte bi-populationnel, tout en générant de manière récurrente des individus hybrides issus de croisements entre ces populations. Cette fois-ci, le simulateur a démontré la présence de stérilité hybride dans certaines configurations de paramètres et a permis d'en identifier la cause. Cependant, cette stérilité hybride prédite par le modèle ne semble pas aussi marquée et récurrente que ce qui est observé empiriquement. La calibration empirique réalisée lors de cette étude, bien que soumise aux mêmes limitations que dans l'Axe 1, permet néanmoins de suggérer que les phénomènes de stérilité hybride induits par PRDM9 jouent probablement un rôle marginal dans la spéciation.

1.6.3 Discussion et perspectives

Enfin, le manuscrit est clôturé par une discussion générale. J'y ai tout d'abord résumé les résultats importants des deux axes présentés auparavant. S'en suit une discussion sur plusieurs pistes d'études essentielles à une meilleure compréhension des tenants et aboutissants des différentes composantes du modèle. Enfin, ce chapitre se termine par une discussions de questions plus générales liées a mon travail de modélisation de la dynamique évolutive de la recombinaison.

Partie II

Études

2

Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis

Alice Genestier[1], Laurent Duret[1], Nicolas Lartillot[1]

[1]Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France

Contents

Avant propos

Cet article a été soumis sur BioRxiv et à PLOS Genetics chez qui il est toujours en révision. Étant rédigé en anglais, je vais tout d'abord en faire un résumé rapide en français avant de présenter le manuscrit dans sa totalité.

Résumé

La dissection moléculaire de la recombinaison méiotique chez les mammifères, combinée à des études comparatives et de génétique des populations, a révélé une dynamique évolutive complexe caractérisée par des points chauds de recombinaison de courte durée, dont l'emplacement exact est déterminé par la protéine de liaison à l'ADN PRDM9. Pour expliquer cette dynamique évolutive rapide, un modèle dit de la "Reine rouge" intra-génomique a été proposé, basé sur l'interaction entre deux forces antagonistes : une conversion génétique biaisée, médiée par les cassures double-brin, entraînant l'extinction des points chauds (le paradoxe de la conversion des points chauds), suivie d'une sélection positive favorisant les allèles mutants de *PRDM9* qui reconnaissent de nouveaux motifs de séquence. Bien que ce modèle prédise de nombreuses observations empiriques, les causes exactes de la sélection positive agissant sur les nouveaux allèles de *PRDM9* ne sont pas encore bien comprises. Dans cette direction, des expériences sur des hybrides de souris ont suggéré que, en plus de cibler les cassures double brin, *PRDM9* a un autre rôle pendant la méiose. Plus précisément, la liaison symétrique de PRDM9 (liaison simultanée au même site sur les deux homologues) faciliterait la recherche d'homologie et, par conséquent, l'appariement des homologues. Bien que découverte chez les hybrides, cette seconde fonction de *PRDM9* pourrait également être impliquée dans la dynamique évolutive observée au sein des populations. Pour répondre à cette question, nous présentons ici un modèle théorique de la dynamique évolutive de la recombinaison méiotique intégrant les connaissances actuelles sur la fonction moléculaire de PRDM9. Notre travail de modélisation donne des indications importantes sur les forces sélectives qui régissent le renouvellement des points chauds de recombinaison. Plus précisément, la réduction de la liaison symétrique de PRDM9 causée par la perte de sites de liaison à haute affinité induit une sélection positive nette qui provoque l'apparition de nouveaux allèles de *PRDM9* reconnaissant de nouvelles cibles. Le modèle offre également de nouvelles perspectives sur l'influence du dosage du gène *PRDM9*, qui peut paradoxalement entraîner une sélection négative sur les nouveaux allèles *PRDM9* entrant dans la population, entraînant leur éviction et réduisant ainsi la variation permanente à ce locus.

Résumé étendu

Introduction

Comme présenté dans l'introduction générale de ce manuscrit, le modèle de Reine-Rouge intragénomique pour expliquer la dynamique évolutive du paysage de recombinaison a été précédemment étudié grâce à des modèles théoriques [85; 96]. Ce phénomène de Reine Rouge a été expliqué comme le résultat de la compétition perpétuelle entre les forces d'érosion des cibles de *PRDM9* faisant baisser la fertilité et la sélection positive agissant sur les nouveaux allèles *PRDM9* qui la restaurent. Cependant, ces modèles sont défaillants sur certains points que nous avons tenté de corriger dans ce chapitre. Premièrement, les modèles réalisés, à l'exception de celui de Baker et al (2022) [134] discuté à la fin de ce chapitre, ne prennent pas en compte les nouvelles connaissances liées au mécanisme moléculaire de PRDM9, en particulier sa liaison symétrique qui semble essentielle au bon appariement des chromosomes homologues. De plus, les modèles précédemment publiés ne proposent aucune explication du mécanisme causant la sélection positive agissant sur *PRDM9*. C'est dans l'optique de répondre à cette question et de comprendre l'impact des mécanismes moléculaires de PRDM9 sur la dynamiques de Reine Rouge, que nous avons développé un simulateur ainsi que des développements analytiques.

Matériel et Méthode

Nous avons développé un simulateur informatique simulant une population d'individus diploïdes subissant à chaque génération des mutations au locus *PRDM9* et aux sites cibles de ce gène, et générant des méioses d'individus pris aléatoirement dans le but de créer la génération suivante. Chaque méiose implémente les mécanismes moléculaires de la recombinaison dépendante de *PRDM9*, comprenant la liaison des protéines à leurs cibles respectives en fonction de l'affinité de ces dernières, la création d'un certain nombre de cassures double brins au niveau des sites liés, l'obligation de former au moins une cassure double brin dans un site également lié par PRDM9 au même site sur le chromosome homologue, et la réalisation des crossing over et des conversions géniques aux sites cassés. Ce simulateur est exécuté sur plusieurs dizaines de milliers de générations, durant lesquelles des statistiques descriptives essentielles à l'étude du mécanisme sont régulièrement calculées (la fréquence des allèles, leurs niveaux d'érosion, ou encore leur fertilité). En complément, afin de mieux comprendre les liens entre paramètres du modèle et statistiques descriptives, et d'identifier les valeurs à l'équilibre de ces statistiques, nous avons détaillé des développements analytiques basés sur le principe de champ moyen auto-consistant.

Résultats

Dans un premier temps, nous avons identifié deux régimes distincts de la Reine Rouge, monomorphe et polymorphe, caractérisé respectivement par la présence d'un ou plusieurs

allèles en même temps dans la population. L'étude approfondie de cette dynamique a permis d'identifier que la sélection positive agissant sur les nouveaux allèles *PRDM9* vient de l'érosion des sites de haute affinité générant des baisses de fertilité chez les vieux allèles. En parallèle, l'étude de la variation des différentes statistiques descriptives en fonction des paramètres du modèle a permis d'identifier que la diversité dépend majoritairement du taux de mutation au locus *PRDM9*, et que le niveau d'érosion et la fertilité dépendent des taux de mutations à la fois aux cibles et au locus du gène.

Ensuite, nous avons étudié les effets du dosage génétique de PRDM9 sur la dynamique de Reine Rouge. Ce dosage, caractérisés par une concentration de PRDM9 chez les homozygotes deux fois plus grande pour un allèle donné que chez un hétérozygote, s'avère responsable d'un phénomène d'éviction des jeunes allèles au profit des vieux allèles homozygotes présents en haute fréquence dans la population. Ce phénomène est problématique car il agit contre la diversité *PRDM9*.

Enfin, nous avons tenté une calibration empirique du modèle. Malgré la prise en compte des paramètres empiriques observés chez la souris, le modèle a eu du mal à prédire des niveaux de diversité, d'érosion des cibles, de sélection positive et d'haplo-insuffisance des allèles qui soient simultanément cohérents avec les connaissances empiriques actuelles. En particulier, Le dosage génère de l'éviction agissant contre la diversité génétique, ce qui nous a amené à étudier l'impact d'autres mécanismes moléculaires non pris en compte précédemment. La diversité retrouve des niveaux raisonnables lorsque qu'on augmente le nombre de méioses réalisables par individus lors d'un évènement de reproduction donné, mais cela baisse la sélection positive, générant un régime de Reine Rouge quasi neutre. L'augmentation de DSB, quant à elle, a permis à la fois l'observation d'une bonne diversité et de sélection positive, mais sans réussir à atteindre des niveaux d'haplo-insuffisance empiriquement.

Discussion

Notre modèle montre que la liaison symétrique de PRDM9 est essentielle à la dynamique de Reine Rouge et à la compréhension de la sélection positive agissant sur les nouveaux allèles *PRDM9*. L'implémentation du dosage génétique a permis l'observation d'un phénomène agissant contre la diversité *PRDM9* par l'éviction des jeunes allèles hétérozygotes. Ce phénomène est d'ailleurs la cause principale de notre incapacité à prédire correctement les valeurs de diversité et de sélection positive observée empiriquement chez la souris.

Cependant, il existe plusieurs pistes qui pourraient agir pour la diversité malgré la présence de dosage. En effet, une autre étude réalisée par Baker et al [134] en parallèle de mon travail, a utilisé une distribution d'affinité différente de la mienne, distribution qui pourrait jouer contre l'éviction due au dosage. Cependant, leur modèle prenant aussi en compte la compétition entre cibles, il fonctionne un peu différemment. Il devient donc important de mieux comprendre et dissocier les différents rôles que peuvent jouer la distribution d'affinité, le dosage et la concentration de PRDM9 dans la cellule pendant la méiose.

Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis : a model based exploration of the *PRDM9* intra-genomic Red Queen

Alice Genestier[1], Laurent Duret[1], Nicolas Lartillot[1,*]

1 Laboratoire de Biometrie et Biologie Evolutive, UMR CNRS 5558, Universite Lyon 1, Villeurbanne, France

* nicolas.lartillot@univ-lyon1.fr

2.1 Abstract

Molecular dissection of meiotic recombination in mammals, combined with population-genetic and comparative studies, have revealed a complex evolutionary dynamics characterized by short-lived recombination hotspots, whose exact location is determined by the DNA-binding protein PRDM9. To explain these fast evolutionary dynamics, a so-called intra-genomic Red Queen model has been proposed, based on the interplay between two antagonistic forces: biased gene conversion, mediated by double-strand breaks, resulting in hotspot extinction (the hotspot conversion paradox), followed by positive selection favoring mutant *PRDM9* alleles recognizing new sequence motifs. Although this model predicts many empirical observations, the exact causes of the positive selection acting on new *PRDM9* alleles is still not well understood. In this direction, experiment on mouse hybrids have suggested that, in addition to targeting double strand breaks, PRDM9 has another role during meiosis. Specifically, PRDM9 symmetric binding (simultaneous binding at the same site on both homologues) would facilitate homology search and, as a result, the pairing of the homologues. Although discovered in hybrids, this second function of *PRDM9* could also be involved in the evolutionary dynamics observed within populations. To address this point, here, we present a theoretical model of the evolutionary dynamics of meiotic recombination integrating current knowledge about the molecular function of PRDM9. Our modeling work gives important insights into the selective forces driving the turnover of recombination hotspots. Specifically, the reduced symmetrical binding of PRDM9 caused by the loss of high affinity binding sites induces a net positive selection eliciting new *PRDM9* alleles recognizing new targets. The model also offers new insights about the influence of the gene dosage of PRDM9, which can paradoxically result in negative selection on new *PRDM9* alleles entering the population, driving their eviction and thus reducing standing variation at this locus.

2.2 Author summary

Meiosis is an important step in the eukaryotic life cycle, leading to the formation of gametes and implementing genetic mixing by recombination of paternal and maternal genomes. A key step of meiosis is the pairing of homologous chromosomes, which is required in order to distribute them evenly into the gametes. Chromosome pairing will also determine the exact position at which paternal and maternal chromosomes will exchange material. Research on the molecular basis of meiosis has revealed the role of a key gene, *PRDM9*. The protein encoded by *PRDM9* binds to specific DNA sequences, by which it determines the location of recombination points. Symmetric binding of the protein (at the same position on the homologous chromosomes) also facilitates chromosome pairing. This molecular mechanism, however, has paradoxical consequences, among which the local destruction of the DNA sequences recognized by PRDM9, leading to their rapid loss at the level of the population over a short evolutionary time. In order to better understand why recombination is maintained over time despite this process, we have developed a simulation program implementing a model taking into account these molecular mechanisms. Our model makes realistic predictions about recombination evolution and confirms the important role played by *PRDM9* during meiosis.

2.3 Introduction

In eukaryotes, meiosis is a fundamental step in the reproduction process, allowing the formation of haploid cells from a diploid cell. This process requires the success of a key step, namely, the pairing of homologous chromosomes. Correct pairing is essential for proper segregation of chromosomes in daughter cells. In addition, it allows for the formation of cross-overs, thus implementing recombination and generating new combinations of alleles. Finally, meiosis and recombination are at the heart of questions of hybrid sterility and speciation [111; 75; 95]. However, the evolutionary dynamics of meiotic recombination still remains poorly understood. A correct understanding of this dynamics requires explicit description of the population genetics processes, on one side, and the molecular mechanisms of meiosis, on the other side. In the present work, we present an attempt in this direction, using theoretical and simulation models.

In mammals and many other eukaryotes, recombination points are not uniformly distributed along the chromosomes [41]. Instead, crossovers frequently occur at the same positions in independent meioses, into regions of the genome called recombination hotspots, where the frequency of crossing-over occurrence is 10 to 100 times higher than in the rest of the genome [33; 37]. Recombination hotspots are typically 1 to 2 kb long, and they are often located outside of genes. More than 30,000 hotspots were found in humans [49; 41], and around 40,000 in mice [42]. These hotspots are characterized in humans by the presence of a sequence motif (13-mer) determining up to 40% of crossing-overs [57]. This motif has helped to identify the *PRDM9* gene as the gene responsible for hotspot location, [57; 58]. *PRDM9* encodes a DNA-binding protein, and the hotspots therefore

correspond to strong binding sites of this protein.

Once bound to its target site, the PRDM9 protein trimethylates surrounding histones (H3K4me3 and H3K36me3 [62; 135]), inducing the recruitment of the double strand break (DSB) machinery near the target site [17] (for a review see [59]). DSB induction, followed by DNA resection, produces a single-stranded end that searches for its homologous sequence on the other chromosome and is then repaired, using the sequence at the same locus on the homologous chromosome as the template. This repair leads to the conversion of the sequence located near the break (often overlapping the site targeted by PRDM9 [25]). Some of these gene conversion events lead to the formation of crossing-overs (CO), while others are repaired without exchange of flanking regions (non crossing-overs, NCO). In some cases, repair is done not with the homologue but with the sister chromatid [25].

Crucially, when allelic variation exists at a given target motif that modulates the binding affinity for PRDM9, DSBs will form more frequently on the allele for which PRDM9 has the highest affinity (the 'hot' allele). Given that DSBs are repaired by using the intact homologue as a template, this process leads to the preferential replacement of 'hot' alleles by alleles for which PRDM9 has a lower affinity. This mechanism of biased gene conversion favors the transmission of mutations that inactivate hotspots, leading to an increase in the frequency of these mutant inactive versions of binding sites, as well as their fixation in the population. This self-destruction phenomenon, commonly called the "hotspot conversion paradox" [83], leads to the progressive inactivation, hereafter called erosion, of PRDM9 binding sites at the genome scale, over short evolutionary times (in the order of 10,000 to 100,000 generations [48]), which therefore raises the question of the maintenance of recombination in the long term.

As a solution to the paradox, a model has been proposed [85], which works as follows: the disappearance of recombination hotspots compromises the proper functioning of meiosis and greatly reduces the fertility of individuals. In this context, new *PRDM9* alleles, recognizing new target sites already present by chance in the genome and thus restoring recombination, would be positively selected. They would eventually replace the alleles currently segregating in the population. By analogy with the so-called Red Queen dynamics [89] typically displayed by prey-predator or host-pathogen systems, this model has been called the intragenomic Red Queen model of recombination hotspot evolution [85]. It predicts a rapid evolutionary dynamics of recombination landscapes, as well as a strong positive selection on the DNA binding domain of *PRDM9*.

Several empirical observations support the Red Queen model. First, fine scale recombination landscapes differ between closely related species, like between humans and chimpanzees [58; 31], or between modern humans and Denisovan [48], suggesting a fast turnover of recombination hotspots. Second, the DNA binding domain of PRDM9 is a zinc finger domain, consisting of an array of 7 to 10 zinc fingers. This domain is encoded by a minisatellite, which mutates rapidly [94] by a combination of point mutations and unequal crossing over between the sequence repeats. This allows for the rapid accumulation of new combinations of zinc fingers, and thus of new alleles recognizing different hotspots, providing the necessary mutational input for the Red Queen

to run. In part because of this high mutation rate, *PRDM9* is typically characterized by a high genetic diversity in natural populations [67; 69; 68; 136; 74] . Finally, non-synonymous substitutions in the zinc finger domain of *PRDM9* are more frequent than synonymous substitutions, suggesting the presence of a strong positive selection acting on the DNA binding domain [68; 95].

Studies combining computer simulations and theoretical analyses have also given many arguments in favor of the Red Queen model [85],[96]. However, current theoretical models remain silent on two essential points. First, what exact mechanism explains the positive selection acting on *PRDM9* to restore recombination? Current models invoke a decline in fertility caused by the erosion of recombination, but without providing a precise explanation of this point. Second, these models do not provide any explanation, at this stage, of the link between the intra-genomic Red Queen and the phenotype of hybrid sterility induced by *PRDM9* in the mouse [111; 137].

In this respect, a series of molecular and comparative analyses conducted on the mouse provide some clues. First, the differential erosion of PRDM9 binding sites between mouse subspecies (due to different *PRDM9* alleles eroding their respective target sites within each subspecies) leads to asymmetrical binding configurations in hybrid individuals [124]. Specifically, in a F1 hybrid, each of the two *PRDM9* alleles has eroded its targets in the genome of its parental strain, but not in the other strain's genome. Each *PRDM9* allele will therefore tend to bind preferentially to the still active target sites present on the chromosomes inherited from the other parent, but not to the homologous but eroded sites on the chromosome from its own parent. These asymmetrical binding patterns of PRDM9 across the whole genome are suspected to be involved in the sterility phenotype. Indeed, Chip-Seq experiments have uncovered a correlation between PRDM9 binding asymmetry rates in a variety of mouse hybrids for different pairs of *PRDM9* alleles and hybrid sterility [117; 125]. Cytogenetic observations have also shown that chromosomes that are more asymmetrically bound are less often correctly paired during metaphase I [125]. Finally, the introduction by transgenesis in mice of a *PRDM9* allele possessing the human zinc finger domain, which has never been present in the mouse species complex and has then not eroded its targets asymmetrically in either of the two populations, restores both a good binding symmetry and high levels of fertility [111].

In the light of these empirical results, a model has been proposed [111], according to which *PRDM9* would in fact have a dual role during meiosis: in addition to being responsible for the recruitment of the DSB machinery, it would also facilitate the pairing between the homologous chromosomes, thanks to its symmetrical binding, that is, its simultaneous binding to the same target site on both homologues. More precisely, PRDM9 is thought to act to bring the sites on which it is bound to the chromosomal axis. The mechanisms involved in this step are not very clear, although it would seem that H3K4me3 markers [35] and the SSXRD and KRAB domains of PRDM9 are implicated[77]. This is where symmetry would play a key role (Fig 2.1): when PRDM9 is bound symmetrically, the two homologous loci are each brought closer to the chromosomal axis, which makes the complementary sequence of the homologue more directly accessible to the single-stranded

end produced by resection of the DSB [52]. In contrast, in the case where PRDM9 binds only on one of the two homologous loci, a possible DSB would be repaired only later on, either by the homologue as a non CO event or by the sister chromatid [25].

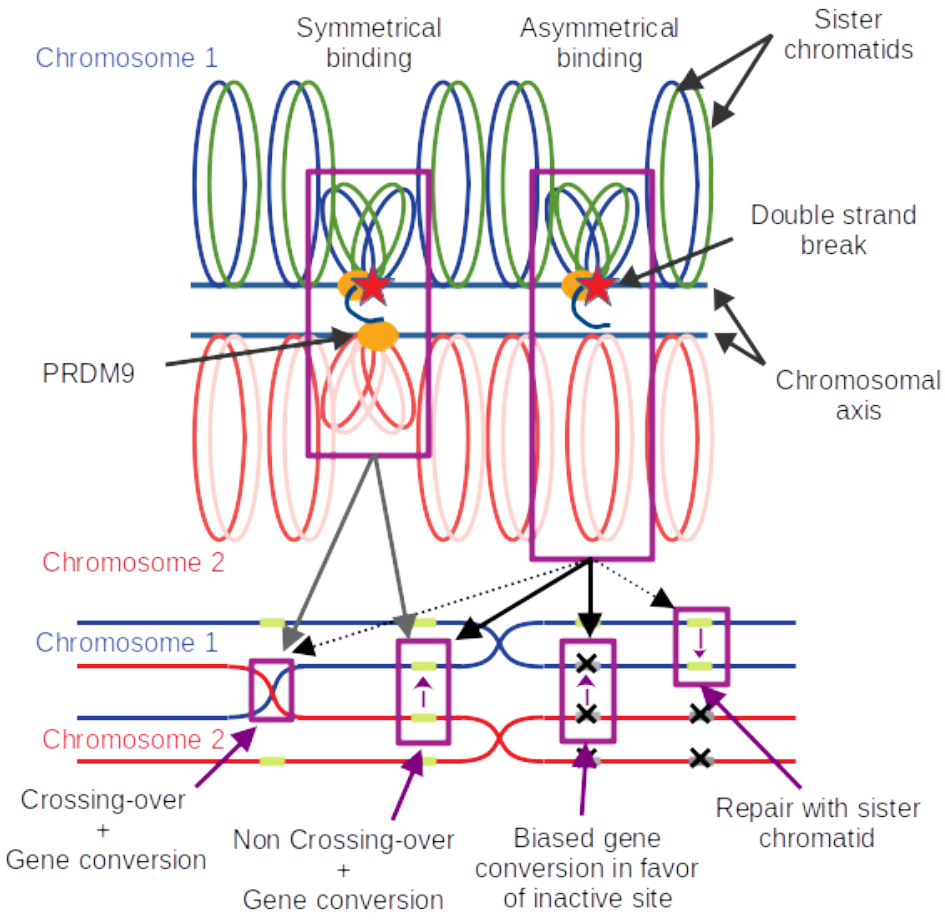

Figure 2.1: **A model of how symmetrical PRDM9 binding facilitates chromosome pairing.** *Upon binding DNA at a specific target motif, PRDM9 (orange oval) brings the DNA segment close to the chromosomal axis. Some of the sites bound by PRDM9 may then undergo a DSB (red star). Resection of the DSB generates a single-stranded end, which will search for a complementary sequence to use as a template for repair. In the case where PRDM9 is symmetrically bound (i.e. on both homologues, case on the left-hand side), the templates provided by the two sister chromatids of the homologue are more directly accessible, thus facilitating homology search and pairing with the homologue. The break can then be repaired either as a CO or a NCO event, in both cases, implementing gene conversion at the broken site. In the case of asymmetrical PRDM9 binding (case shown on the right-hand side), the homologue is less directly accessible, preventing efficient homologue engagement. The broken site is assumed to be repaired later on, once the homologues have synapsed (and this, thanks to other DSBs occurring at symmetrically bound sites somewhere else on the same pair of chromosomes), as NCO events. In the case where the homologue bears an inactive binding site at the position corresponding to the DSB, the NCO will effectively implement biased gene conversion in favor of the inactive version.*

This mechanistic model based on the symmetrical binding of PRDM9 provides a globally coherent explanation of the hybrid sterility phenotype. A last question remains open however: could this dual role of *PRDM9* also explain the Red Queen evolutionary dynamics of recombination observed within a single population? Of note, in the hybrid, failure of meiosis is a consequence of macroscopic asymmetric sequence patterns due to differential erosion in the two parental lineages. Such macroscopic asymmetric sequence patterns are unlikely within a single population. On the other hand, statistical binding asymmetries might nevertheless occur and play a role.

To investigate this question, in this paper, we revisit the theoretical modeling of the intra-genomic Red Queen of *PRDM9*-dependent recombination. In contrast to previous work [96; 85] (but see recent work of Baker *et al.* [134]), we explicitly model the molecular mechanism of meiosis and, more specifically, the function currently hypothesized for *PRDM9*. Our specific aim was to test whether the combined effects of biased gene conversion and symmetry provide sufficient ingredients for running an intra-genomic Red Queen, and this, under empirically reasonable parameters values.

2.4 Results

To investigate the evolutionary dynamics of *PRDM9*-dependent recombination, we developed a simulation program modeling the evolution of a randomly mating population of diploid individuals and accounting for the key molecular processes involved in meiotic recombination (Fig 2.2). The genome consists of a single chromosome, bearing a locus encoding the *PRDM9* gene (Fig 2.2A). The locus mutates at a rate u, producing new alleles (Fig 2.2B). Each allele recognizes a set of binding sites randomly scattered across the genome, of varying binding affinity. Binding sites undergo inactivating mutations at a rate v (Fig 2.2B). At each generation, sexual reproduction entails the production of gametes that are themselves obtained by explicitly implementing the process of meiosis on randomly chosen individuals (Fig 2.2C-F).

Meiosis is modeled step by step. First, the PRDM9 proteins binds to their target sites, according to a simple chemical equilibrium (Fig 2.2C). The probability of occupation of a site will thus depend on the binding affinity of PRDM9 for this site. It may also depend on PRDM9 concentration, which can itself depend on the genotype of the individual (either homozygous for single *PRDM9* allele, or heterozygous for two distinct *PRDM9* alleles), through gene dosage effects. Then, a small number of double strand breaks are induced at some of the sites bound by PRDM9 (Fig 2.2D). At that step, a key assumption of the model is that meiosis will succeed only if at least one of those DSBs occurs at a site that is also bound by PRDM9 on at least one of the two chromatids of the homologous chromosome, in which case a single cross-over is performed at the site of one of the DSBs fulfilling this requirement. Finally, all DSBs are repaired using the homologue (Fig 2.2E), upon which meiosis is assumed to proceed successfully, producing four gametes, one of which is chosen at random for reproduction (Fig 2.2F).

Of note, because of the symmetry requirement, the probability of success of meiosis

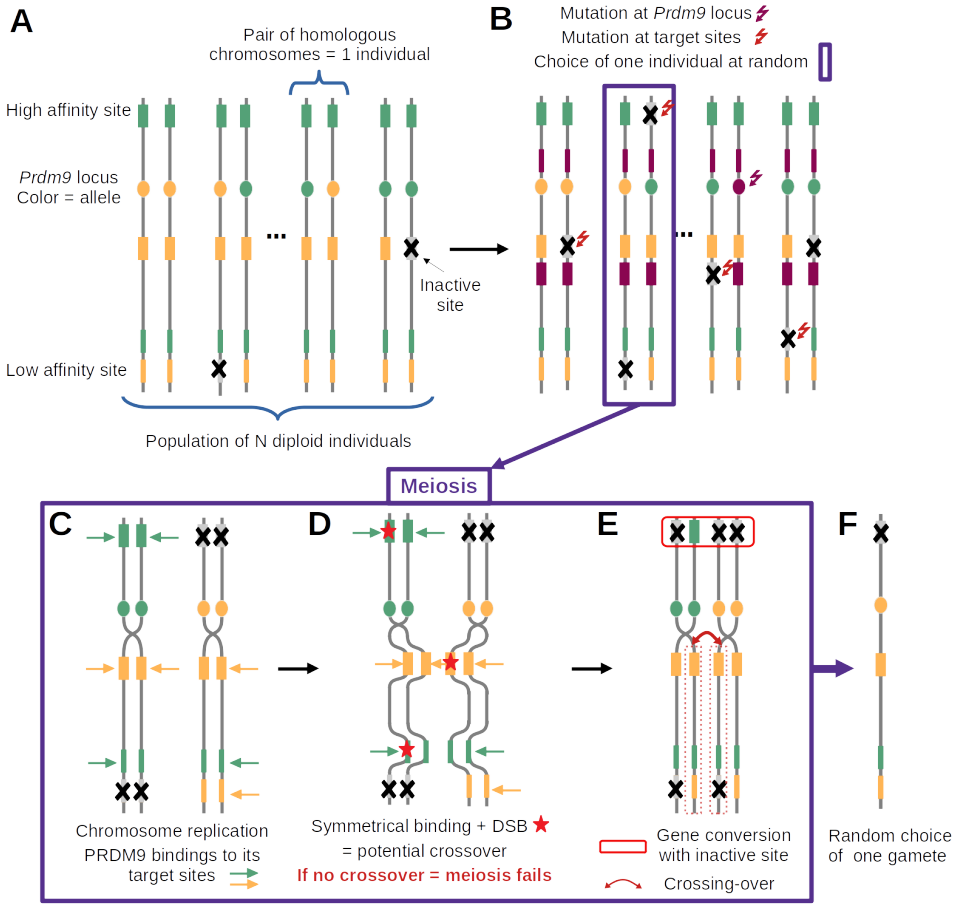

Figure 2.2: **Diagram summarizing the main features of the model and the successive steps of the simulation cycle.** *(A) The model assumes a population of N diploid individuals (2N chromosomes, vertical lines). Each chromosome has a PRDM9 locus (filled oval, with a different color for each allele) and, for each PRDM9 allele, a set of target sites (filled rectangles, with color matching their cognate PRDM9 allele) of variable binding affinity (variable width of the filled rectangles). (B) Mutations at the PRDM9 locus create new alleles (here, purple allele), while mutations at the target sites inactivate PRDM9 binding (grey sites with a cross). (C-E) meiosis; (C) Chromosome replication and binding of PRDM9 to its target sites (horizontal arrows). (D) Induction of DSBs at a small number of randomly chosen sites bound by PRDM9 (red stars) and search for DSBs at symmetrically bound sites (symmetrical DSBs); if no symmetrical DSB is found, meiosis fails; otherwise, one symmetrical DSB is chosen uniformly at random (here, on the yellow binding site). (E) Completion of the crossing-over (red curved arrow) and repair of all DSBs using the homologous chromosome (red box on top); (F) random choice of one of the 4 gametes of the tetrad, which will contribute to the next generation.*

will depend on the genotype of the individuals on which it is attempted, which will thus induce differences in fertility between individuals. In addition, explicit repair of the DSBs by the homologue effectively implements the process of gene conversion at the binding sites. The main question is then to what extent these two aspects of the molecular mechanism are susceptible to influence the evolutionary dynamics.

The overall procedure is run for several tens of thousands of generations, during which

several summary statistics are monitored: the frequency of each *PRDM9* allele and the resulting genetic diversity at the *PRDM9* locus, the mean proportion of binding sites of *PRDM9* alleles that are currently active, their mean affinity across the genome, the mean probability of symmetrical binding and the mean probability of success of meiosis. The model parameters and the monitored statistics are summarized in Table 2.1

Table 2.1: **Descritption of input parameters and output variables**

Parameters	Description	Value
u	Mutation rate at the *PRDM9* locus	2×10^{-6} to 5×10^{-3}
v	Mutation rate at the targets	2×10^{-6} to 5×10^{-3}
N	Population size	$5,000$
h	Number of targets recognised by a new allele	400
d	Mean number of DSB per chromosome pair per meiocyte	6
y_i	Affinity of target i	variable
g	Gene conversion rate	variable
n_{mei}	Number of meioses allowed per individual before reproduction failure	1 to 5

Variables	Description
f_i	Frequency of allele i [a]
θ_i	Proportion of target sites still active for allele i [a]
z_i	Level of erosion of allele i [a]
w_i	Mean fertility of allele i [a,b]
q_i	Mean probability of symmetrical binding of allele i [a,b]
ρ	Net rate of erosion per generation
x_i	Occupancy probability at site i at equilibrium
α	Linear response of the log-fitness as a function of erosion
τ	Mean time between successive invasions of the population by new *PRDM9* alleles
μ	Population-scaled mutation rate of *PRDM9* ($4Nu$)
D	*PRDM9* diversity in the population at equilibrium
s_0	mean selection coefficient acting on new *PRDM9* alleles at equilibrium
σ_0	Relative difference in fertility between homozygotes and heterozygotes for young alleles
$\sigma_{\bar{z}}$	Relative difference in fertility between homozygotes and hemizygotes for alleles of level of erosion $\bar{z}$

[a] These variables also change over time; [b] the mean is over all individuals carrying this allele (with a weight of 1 for heterozygotes and a weight of 2 for homozygotes).

The results section is divided in three main parts. First, a simple version of the model is presented, in which PRDM9 gene dosage is ignored (that is, assuming that the number of proteins produced by a given *PRDM9* allele is the same in a homozygote and in a heterozygote for this allele). Although empirically questionable, this assumption offers a simpler basis for understanding key features of the model and of the resulting evolutionary dynamics. Then, in a second part, we introduce PRDM9 gene dosage and work out its consequences. Finally, we attempt an empirical calibration of the model

based on current knowledge in the mouse, so as to test its predictions.

2.4.1 Intragenomic Red Queen

We first ran the model without gene dosage, and with a low mutation rate at the *PRDM9* locus ($u = 5 \times 10^{-6}$). Note that the parameter values used here are not meant to be empirically relevant. Instead, the aim is to illustrate the different regimes produced by the model. Assuming a low mutation rate for *PRDM9* results in few, rarely more than one, alleles segregating at any given time in the population (Fig 2.3A). We call this a *monomorphic* regime, after Latrille *et al* [96]. The dynamic is as follows. First, an allele appears in the population and invades, until reaching a frequency close to 1. Meanwhile, the proportion θ of active binding sites for this allele decreases, due to the erosion of these sites by inactivating mutations and biased gene conversion. Once the allele has eroded a fraction of around 20% of its sites, it is quickly replaced by a newly arisen PRDM9 allele that recognizes a different hotspot sequence motif. This rapid replacement clearly suggests the presence of strong positive selection. The newly invading allele then erodes its target until being replaced by a new allele which in turn follows a similar trajectory, and so on.

A key aspect of the model is that it does not explicitly invoke a fitness function. Instead, the positive selection that seems to be operating on new alleles (Fig 2.3) is an emerging property of the mechanism of meiosis. This positive selection can be more precisely understood by examining how the rate of PRDM9 symmetrical binding (defined by Eq (2.15) and Eq (2.14)) and the mean fertility of a typical allele (Eq (2.16)) evolve over the lifespan of this allele (Fig 2.3D and 2.3E), and how this relates to the level of erosion (Eq (2.13), Fig 2.3B) and the mean affinity of the remaining non-eroded sites (Fig 2.3C).

First, we observe a clear correlation between allele frequency and proportion of active (non-eroded) target sites. When the frequency of an allele increases in the population, the proportion of still active sites for this allele decreases. This erosion (Fig 2.3B) seems to occur at a rate which is directly proportional to the frequency of the allele in the population (Fig 2.3A). This is expected: the more frequent an allele, the more often it will be implicated in a meiosis and the more opportunities its target sites will have to undergo gene conversion events.

Second, erosion results in a decrease of the mean affinity of the sites that are still active (Fig 2.3C). This reflects the fact that sites of high affinity are more often bound by PRDM9, and thus are more often converted by inactive mutant versions.

Third, sites of lower affinity are also less often symmetrically bound (i.e. bound simultaneously on both homologues). The key quantity that captures this effect is the conditional rate of symmetrical binding (q). Since DSBs are chosen uniformly at random among all bound sites by PRMD9, q corresponds to the probability that a DSB will occur in a symmetrically bound site and will thus contribute to a successful pairing of the homologues. This probability is a monotonous function of the mean affinity of the binding sites for the PRDM9 protein (Fig 2.3D). The mean value of q over all target sites

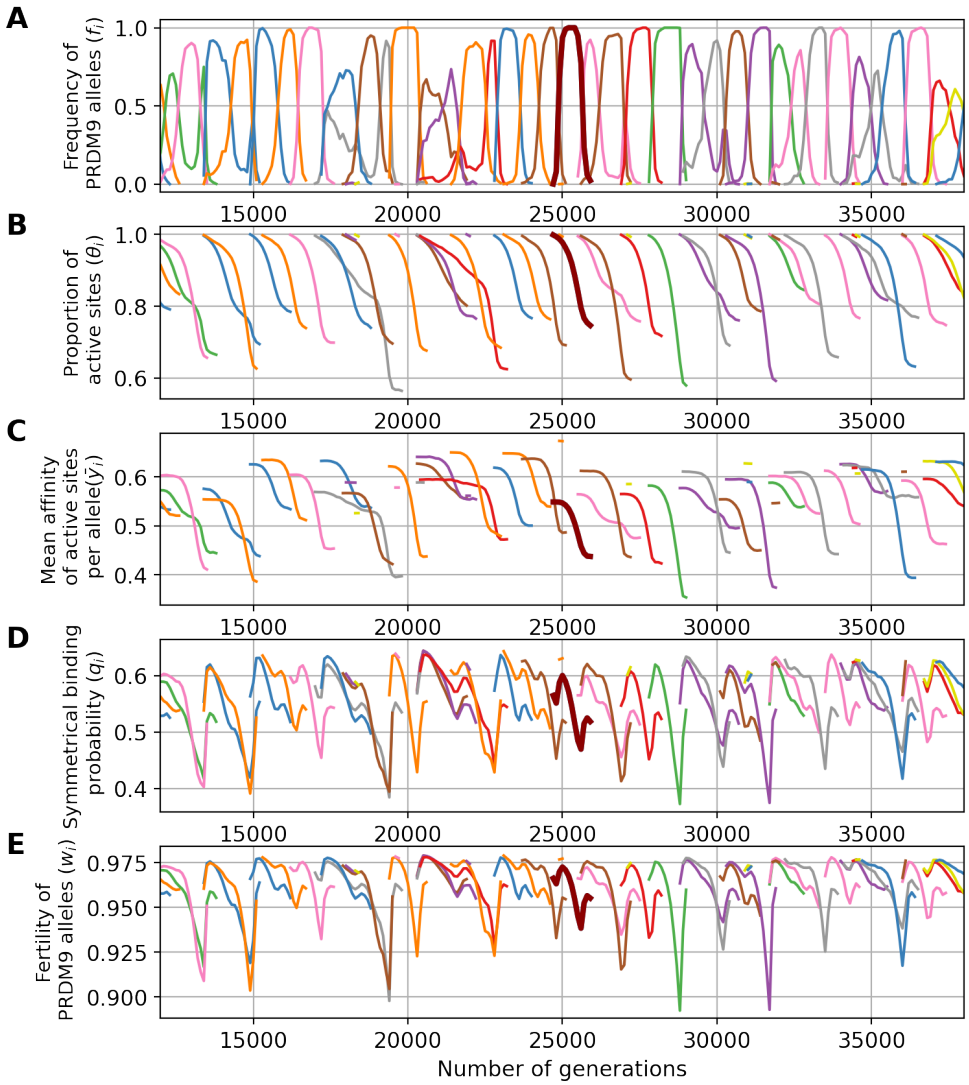

Figure 2.3: **A simulation trajectory showing a typical evolutionary dynamics, under a monomorphic regime ($\mathbf{u = 5 \times 10^{-6}}$ and $\mathbf{v = 5 \times 10^{-5}}$).** *In all panels, each color corresponds to a different allele. Note that a given color can be reassigned to a new allele later in the simulation. Successive panels represent the variation through time of (A) the frequency of PRDM9 alleles and their corresponding (B) proportion of active sites, (C) mean affinity, (D) probability of symmetrical binding and (E) fertility. The thick line singles out the trajectory of a typical allele.*

of a given *PRDM9* allele is thus lower for older alleles (Fig 2.3D).

Finally, since, in our model, meiosis requires at least one DSB in a symmetrically bound site, the mean fertility of an older allele is lower (Fig 2.3E). Hence, new alleles (young alleles) will be positively selected at the expense of old alleles and will ultimately replace them in the population.

To assess to what extent the requirement of symmetrical binding impacts the evolutionary dynamics of *PRDM9*, we performed additional simulations with a simpler model, in which DSBs can be repaired as COs even when PRDM9 binds only one of the two homologues (i.e. this corresponds to a model where PRDM9 is required for the formation

of DSBs, but not for chromosome pairing). With this setting, the model still predicts a turnover of *PRDM9* alleles, but with unreasonably high, in fact nearly complete, levels of erosion (Supplementary Figure 5.2). Fundamentally, *PRDM9* alleles persist in the population until they have no more sites to bind, at which point they cannot anymore recruit DSBs and are thus selectively eliminated for this trivial reason. Thus, a key result obtained here is that the requirement of symmetrical binding for chromosome pairing, combined with preferential erosion of high affinity sites, is sufficient for creating differences in fitness (fertility) between old and young alleles, which in turn will spontaneously induce a Red Queen dynamics at the level of the population, while ensuring moderate levels of hotspot erosion.

Of note some stochastic deviations from this typical life-cycle for a *PRDM9* allele are sometimes observed, such as an allele being first outcompeted by a subsequent allele but then showing a rebound in frequency when the competitor has itself eroded a large fraction of its target sites. Such deviations are relatively rare and do not seem to fundamentally change the overall regime. It should also be noted that we observe an increase in the rate of symmetrical binding and in the mean fertility at the beginning and at the end of the life of the alleles. The reason for this is that these two summary statistics are defined, for each allele, as a mean over all diploid genotypes carrying this allele segregating in the population. As a result, when old alleles have declined to a low frequency, they often find themselves in a heterozygous state with new alleles, which restores the rate of symmetrical binding and thus the fertility of the corresponding diploid individual. Likewise, when a new allele appears in the population, it is in a heterozygous state with an older allele, which gives a lower rate of symmetrical binding and fertility than being in the homozygous state.

The simulation shown above (Fig 2.3) was run under a low mutation rate, hence resulting in a *monomorphic* regime. Running the simulation under higher mutation rates for *PRDM9* (higher u) results in a *polymorphic* regime, where many alleles segregate together at any given time. In this regime, of which a typical simulation is shown in Fig 2.4, the Red Queen process is also operating, except that many alleles are simultaneously segregating (Fig 2.4A), at a lower frequency. As in the previously shown monomorphic regime, each allele undergoes erosion (Fig 2.4B), primarily of its higher affinity sites (Fig 2.4C), again causing a decrease in symmetrical binding rate (Fig 2.4D) and fertility (Fig 2.4E). Owing to the high mutation rate, however there is a much lower erosion, leading to a lower decrease in fertility than in a monomorphic regime. Note that in Fig 2.4, the scale of the axes of the ordinates are not the same as in Fig 2.3 (see Supplementary Figure 5.1 for a figure with same scale on the y-axis).

Control simulations (without the requirement of symmetry) run in the polymorphic case result in less extreme erosion levels than control simulations in the monomorphic regime, although still much stronger than in the simulations in which symmetrically binding is required (Supplementary Figure 5.1). The overall dynamics under this control appears to be neutral, with a turnover caused by mutational input of new alleles and loss of old alleles by drift.

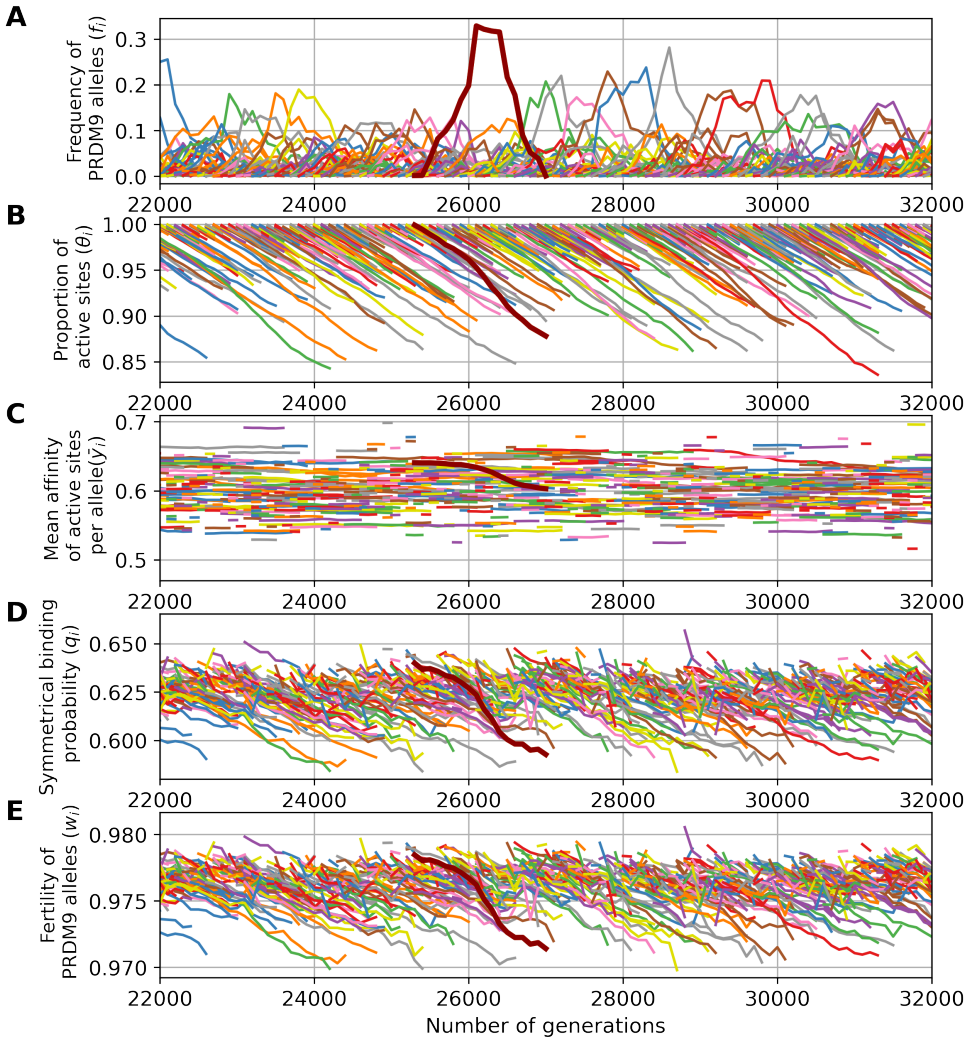

Figure 2.4: **A simulation trajectory showing a typical evolutionary dynamics, under a polymorphic regime ($u = 5 \times 10^{-4}$ and $v = 5 \times 10^{-5}$).** *In all panels, each color corresponds to a different allele. Note that a given color can be reassigned to a new allele later in the simulation. Successive panels represent the variation through time of (A) the frequency of PRDM9 alleles and their corresponding (B) proportion of active sites, (C) mean affinity, (D) probability of symmetrical binding and (E) mean fertility. The thick line singles out the trajectory of a typical allele. Note that the scale of the axes of the ordinates are not the same as in Fig 2.3.*

2.4.2 *PRDM9* diversity and erosion

Scaling experiments

The simulations shown in the previous section have helped to establish that the molecular details of the implication of PRDM9 during meiosis (such as depicted in Fig 2.1) are sufficient to induce a Red Queen dynamics. However, it remains to be understood how the equilibrium regime quantitatively depends on the parameters of the model. To address this issue, scaling experiments were performed (see methods). These experiments are helpful for determining how the different characteristics of the model at equilibrium (the allelic diversity at the *PRDM9* locus, the average level of erosion of the target sites or the

fertility of the individuals) respond to variation in the values of the model parameters. The results of these scaling experiments are shown in Fig 2.5 (in blue), along with the predictions of an analytical approximation (in orange), introduced further below.

A first scaling experiment was carried out on the mutation rate u at the *PRDM9* locus. It immediately appears that u has a direct impact on the standing diversity (Fig 2.5A), which is roughly proportional to Nu, the population-scaled mutation rate. Increasing u also reduces the equilibrium erosion level (Fig 2.5B). This can be explained as follows. The equilibrium set point of the Red Queen is fundamentally the result of a balance between erosion and invasion. The rate of invasion by new alleles can be expressed as $2NuP_{inv}$ where P_{inv} corresponds to the probability of invasion. In this expression, if u increases, this also increases the rate of replacement of old alleles thus shifting the equilibrium towards weaker equilibrium erosion. These lower erosion levels imply a higher rate of symmetrical binding and a higher mean fertility (Fig 2.5C).

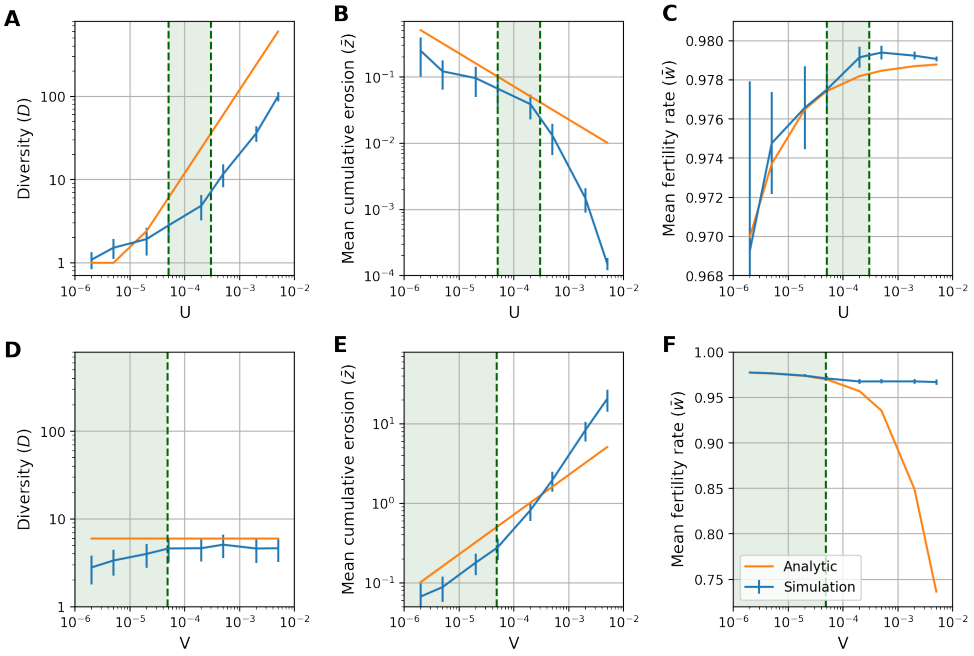

Figure 2.5: **Scaling of key summary statistics at equilibrium, as a function of the mutation rate u at the *PRDM9* locus and the mutation rate v at the target sites.** *The statistics are: the PRDM9 diversity D as a function of u (A) and v (D); the mean erosion $\bar{z}$ (i.e. the mean fraction of target sites that have been inactivated) at equilibrium as a function of u (B) and v (E); the mean fertility of the population $\bar{w}$ as a function of u (C) and v (F). On each graph, the mean (blue line) and standard variation (blue vertical bars) over a simulation are displayed against the prediction of the analytical approximation (orange line). The area colored in green corresponds to the range of parameters for which the analytical model verifies the assumptions of a high diversity ($1 < 4Nu < 100$), a low erosion ($\bar{z} < 0.5$) and strong selection on new PRDM9 alleles ($4Ns_0 > 3$).*

A second scaling experiment was performed, now varying the rate v of target inactivation. From this experiment, it appears that v has a very weak impact on standing

diversity (Fig 2.5D). On the other hand, it has an important impact on equilibrium erosion levels (Fig 2.5E) and on fertility (Fig 2.5F), with an opposite effect, compared to what was observed with u. This comes from the fact that the rate of erosion is equal to $4Nvg$. Increasing v will therefore result in an increase of the rate of erosion at equilibrium.

To summarize, the equilibrium set point of the Red Queen dynamic is directly impacted by the two parameters u and v, whose effects are opposed. In the case of diversity, u (or, more precisely, Nu) is the key determining factor, while v seems to have minor impact. These results are consistent with the results obtained by Latrille et al. [96].

Comparison with analytical developments

In parallel with the simulator, the model was also explored analytically, using the self-consistent mean-field approach originally described in Latrille *et al.* [96]. The complete analytical developments can be found in Supporting informations sections 5.1.1 to 5.1.3. Here, a simplified presentation is given, introducing the main intuitions and formulae.

For these analytical developments to be tractable, three preliminary conditions were introduced. First, a polymorphic regime ($Nu >> 1$) is assumed for *PRDM9*. This assumption makes it possible to neglect the presence and the impact on the dynamics of homozygous individuals, which facilitate the approximations. Second, we assume a strong selection regime ($4Ns_0 >> 1$, where s_0 is the mean selection coefficient acting on new *PRDM9* alleles). This allows us to rely on deterministic approximations for the typical trajectory of *PRDM9* alleles. Finally, we assume a weak erosion of the target sites at equilibrium, which allows us to linearize the equations.

First, to a good approximation, the rate of erosion of the targets of an allele depends on its frequency in the population :

$$\frac{d\theta}{dt} \approx -\rho f \theta, \tag{2.1}$$

where $\rho = \frac{Nvd}{2h} = 4Nvg$ and $g = \frac{d}{8h}$ is the net rate of erosion per generation. As a result, the cumulated erosion of an allele of age t is:

$$z(t) = \rho \int_0^t f(t')\, dt', \tag{2.2}$$

where $f(t')$ is the frequency of the allele at time t'. The quantity z can thus be seen as the intrinsic age of the allele (i.e. an allele ages more rapidly if more frequent in the population). Because of selection, the frequency of an allele changes as a function of its erosion, or intrinsic age z.
Assuming weak erosion ($z << 1$) allows one to linearize the differential equation describing the evolution of the frequency of an allele, which gives :

$$\frac{d \ln f}{dt} \approx -\frac{\alpha}{2}(z - \bar{z}). \tag{2.3}$$

Eq 2.3 says that the variation in frequency of an allele is proportional to the difference between the erosion of the targets for this allele (z) and the mean erosion for all other

alleles in the population ($\bar{z}$). The proportionality coefficient, α, corresponds to the linear response of the log-fitness as a function of erosion. It can be expressed as a function of the mechanistic parameters of the model (Supporting materials section 1 Eq 56). Thus, if an allele is younger, the allele has eroded less targets than the average ($z < \bar{z}$) and will increase in frequency. Conversely when the allele reaches and then exceeds the mean erosion level, its frequency will start to decrease.

The previous equation shows the evolution of the frequency of a typical allele in the population, but it depends on the mean level of erosion $\bar{z}$, which is unknown. To determine its value, we can apply a self-consistent argument, essentially saying that all alleles entering the population have this typical dynamic. Using this self-consistent argument leads to the following explicit expression for $\bar{z}$ [96] :

$$\bar{z} \approx \sqrt{\frac{\rho}{\mu\alpha}} \approx \sqrt{\frac{vg}{u\alpha}}. \tag{2.4}$$

Here, we find three constants of the model. There is first of all ρ, the net erosion constant, which by increasing will contribute to increasing the mean equilibrium erosion level $\bar{z}$. Then, in the denominator, there is $\mu = 4Nu$, the population-scaled mutation rate of *PRDM9*, which by increasing leads to a more frequent emergence of new alleles. This restricts the lifetime of alleles in the population and thus decreases the average cumulative erosion rate. Finally, we find again the constant α. If the latter increases, this means that the log-fitness will decrease more quickly as a function of erosion. As a result, the allele will be more quickly eliminated (i.e. with less erosion), which lowers the equilibrium erosion level $\bar{z}$.

Note that this equation is the same as that presented in Latrille *et al* [96]. However, in [96], α and g were phenomenological parameters, set arbitrarily, while in our mechanistic model, they emerge directly from the molecular mechanisms of meiosis. As such, they can be expressed as functions of the model parameters (d, h, N or v, see Supporting matrials section 1 Eq 56 and Eq 34).

From there, we can express the mean equilibrium quantities over the population and across the whole simulation as a function of $\bar{z}$. Namely $\bar{\theta}$, the mean proportion of still active target sites, $\bar{w}$, the mean fertility of the alleles, and the diversity D of *PRDM9* in the population at equilibrium :

$$\begin{cases} \bar{\theta} & \approx & 1 - \bar{z} \\ \bar{w} & \approx & 1 - e^{-d\bar{q}} \\ D & \approx & 24Nu \end{cases} \tag{2.5}$$

In the expression for $\bar{w}$, $\bar{q}$ corresponds to the mean probability of symmetrical binding at equilibrium. It depends on $\bar{z}$.

The analytical approximations presented here are plotted on Fig 2.5A to F (orange curves), against the results obtained directly using the simulation program (blue curves). The model and the analytical approximation give qualitatively similar results in the range of parameters validating all the conditions (in practice, we consider that the analytical

results should be valid in the following intervals for the model parameters: $1 < 4Nu < 100$, $\bar{z} < 0.5$ and $4Ns_0 > 3$). Concerning *PRDM9* diversity, substantial differences are observed between the simulation results and the analytical approximations, up to a factor of 10, in the scaling of u (Fig 2.5A). However, the nature of the regime, polymorphic (many alleles segregating at the same time in the population) or monomorphic (only one allele present in the population at a time), is correctly predicted. In particular, we can say that the nature of the regime is directly and mostly determined by Nu and the level of erosion has almost no influence (Fig 2.5D). Finally, the analytical approximations are less accurate for low and high u or high v. These correspond to strong erosion regimes (low u and high v) or to regimes with weak selection (high u), for which the assumption of the analytical developments are not met.

2.4.3 Taking into account *PRDM9* gene dosage

The previous results were obtained with a model assuming the same concentration of the PRDM9 protein product of a given allele in individuals that are either homozygous or heterozygous for this allele. Yet, in reality, gene dosage seems to be an important aspect of the regulation of *PRDM9* expression [138]. To account for this fact, gene dosage was introduced in the simulation model. This was done by assuming that a homozygote produces twice as many protein products as a heterozygote for a given allele. The main consequence of introducing gene dosage is that, in homozygotes, the occupancy of a site of a given affinity is increased, compared to a heterozygote, leading to a higher probability of symmetrical binding and fertility. This could have an important impact on the Red Queen dynamics.

To illustrate this point, Fig 2.6 contrasts the results obtained with and without gene dosage in an otherwise identical parameter configuration. We observe a drastic change of regime between the two settings. While the simulation without gene dosage gives a polymorphic regime, the simulation with gene dosage gives a strict monomorphic regime with an extremely dominant allele staying at a frequency close to 1 during several thousands of generations, a time period during which new alleles apparently cannot invade the population.

What happens here is that, due to gene dosage, homozygotes have a fitness advantage over heterozygotes. Since alleles at high frequency are more often present in a homozygous diploid genotype, they have an advantage over low frequency alleles. And so, paradoxically, old (but not too old) alleles at high frequency can have a higher mean fitness than new alleles that just appeared in the population, and this, in spite of their higher levels of erosion. This last point is confirmed by measuring the selection coefficient associated with new alleles (Fig 2.6F): during a phase of domination of one allele, new alleles are strongly counter-selected. When the dominant allele becomes too old, its homozygote advantage is no longer strong enough to compensate for its erosion. At this point, new alleles become positively selected (e.g. at around $15,000$ generations on Fig 2.6F). The old allele then quickly disappears and all other alleles competes for invasion.

This transient polymorphic regime is unstable however: as soon as one of the young alleles reaches a higher frequency than its competitors, it benefits from a homozogous advantage, ultimately evicting all competitors and thus becoming the new dominant allele.

The example on Fig 2.6 corresponds to only one particular parameter settings showing that gene dosage can act against diversity. To more systematically assess the conditions under which gene dosage is expected to have an impact on the qualitative regime of the model, a perturbative development was conducted, starting from the analytical results shown above (for the complete analytical developments, see Supporting materials section 5.1.4). Here, perturbative means that it is valid only when the impact of dosage is weak. However it will also give the conditions on the model parameters for this to be the case. Thus, with dosage, the evolution of the frequency f of an allele through time is now given by:

$$\frac{d\ln f}{dt} = -\frac{\alpha^{het}}{2}(z-\bar{z}) + \sigma_0(f-\bar{f}), \tag{2.6}$$

where α^{het} corresponds to the α coefficient (slope of the log-fitness as a function of z) for a heterozygote and :

$$\sigma_0 = \frac{w^{hom}(0) - w^{het}(0,0)}{w^{het}(0,0)} \tag{2.7}$$

is the relative difference in fertility between homozygotes and heterozygotes for young alleles (or homozygous advantage).

Thus, unlike what was observed without dosage, the fate of an allele is now determined by a competition between two different effects : an age effect and a frequency effect. The age effect (everything else being equal, older alleles tend to have a lower fitness than younger alleles) was already present without dosage (Eq 2.3). The frequency effect (everything being equal, more frequent alleles benefit from a homozygous advantage and will therefore tend to become even more frequent) is the specific contribution of dosage. Importantly, the frequency effect (which is proportional to σ_0) systematically acts against diversity.

Based on Eq 2.6, we can determine when dosage will play an important role. First, σ_0 should be large. Second, there should be enough time between successive invasions for eviction to take place. An analytical approximation for the time between invasions is given by $\tau = \frac{2}{\mu\alpha\bar{z}}$. Altogether we predict a qualitative change induced by gene dosage on *PRDM9* standing diversity when Nu is large (i.e. the regime would have been polymorphic without dosage) and $\sigma_0\tau$ is also large (i.e. eviction due to homozygous advantage is sufficiently strong and has enough time to occur). Conversely, we predict a polymorphic regime when Nu is large and when the dosage effects are negligible, i.e. when $\sigma_0\tau$ is small.

These predictions were verified by conducting two scaling experiments, exploring a broad range of values for the mutation rates at the *PRDM9* locus u and at the targets v (Supplementary Figure 5.3), or varying the mean number of DSB (d) and the mean binding affinity $\bar{y}$ of the target sites (Supplementary Figure 5.4). In all cases, these experiments confirm that the regime is polymorphic in the presence of dosage if $4Nu > 10$ and $\sigma_0\tau < 3$.

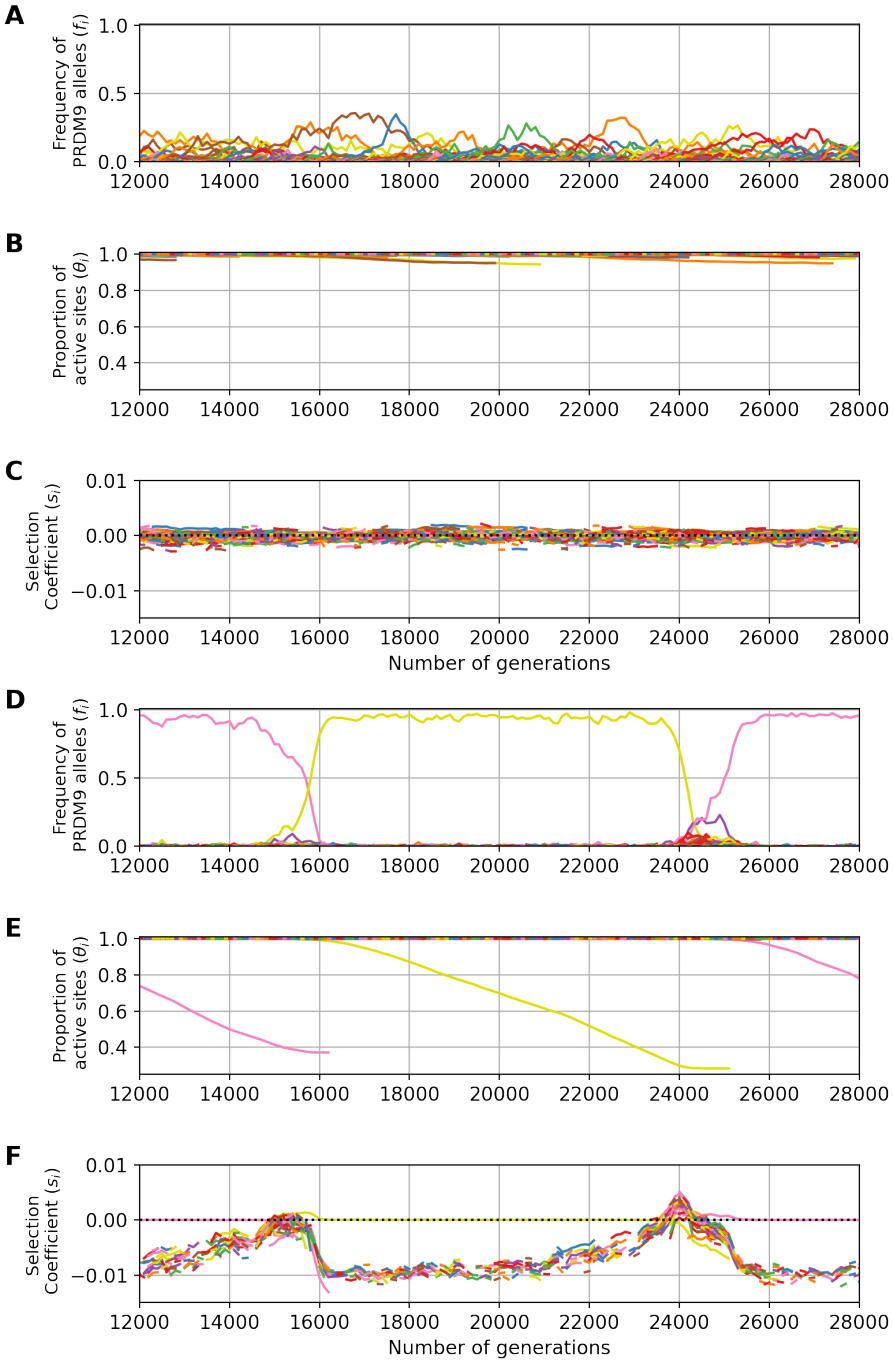

Figure 2.6: **Example of a *PRDM9* dynamic without gene dosage (A) to (C) and with gene dosage (D) to (F) (mutation rates $\mathbf{u = 5 \times 10^{-4}}$ at the *PRDM9* locus, and $\mathbf{v = 2 \times 10^{-6}}$) at the target sites.** *In all panels, each color corresponds to a different allele. Note that a given color can be reassigned to a new allele later in the simulation. Successive panels represent the variation through time of (A),(D) the frequency of PRDM9 alleles and their corresponding (B),(E) proportion of active sites and (C),(F) selection coefficient.*

Altogether, gene dosage entails additional constraints, which tend to substantially restrict the conditions for observing a high diversity of *PRDM9* alleles in the population.

2.4.4 Empirical calibrations of the model

Empirical input parameters

Finally an empirical calibration of the model was attempted, using parameter values roughly estimated based on current knowledge in mammals and, more specifically, in the mouse. The effective population size (N_e) of different *Mus musculus* subspecies such as *Mus m. musculus* or *Mus m. domesticus* is around 10^5 [139]. The number of hostspots recognized per *PRDM9* alleles across the genome (h) is estimated between $15,000$ and $20,000$ [34]. Knowing that the mouse possesses 20 pairs of chromosomes, this corresponds approximately to $h \approx 1000$ sites per chromosome. The mean number of DSBs (d) performed per individual is estimated between 200 and 250 per meiosis [81], which correspond to approximately 10 DSBs per chromosome pair per meiosis. An exponential distribution for the affinities has a good fit to observed distribution obtained from Chip-seq experiments [137] (Supplementary Figure 5.5). The mean, however, is unknown but can be determined by assuming an average of $5,000$ targets bound per meiocyte [80] out of $20,000$ available targets. We obtain a mean of approximately 0.2.

Concerning the target mutation rate, considering that the mutation rate per nucleotide per generation across the mouse genome is $5.4.10^{-9}$ [140] and that the motifs characterising most of the hotspots in the mouse are several tens of nucleotides long [124], the inactivating mutation rate per generation (v) at the targets can be estimated at 10^{-7} [96]. The other mutation rate to determine is the one at the *PRDM9* locus (u). Owing to the mutational instability of the minisatellite encoding the zinc finger domain, this mutation rate is high and has been estimated at (10^{-5}) in humans [94]. However, this is the raw mutation rate (including mutation that lead to either non-functional or functionally equivalent alleles). In contrast, the mutation rate u of the model is the net functional mutation rate (probability of creating a new functional allele recognizing entirely different targets), and the net rate is likely to be smaller than the raw experimental estimate. Accordingly, and as in Latrille et al [96], we used $u = 3.10^{-6}$.

Finally, since a population size of $N = 10^5$ is too large for running the simulator, a standard scaling argument was used, by setting $N = 5.10^3$ and multiplying u and v by a factor 20 (i.e. using between $u = 6.10^{-5}$ to $u = 6.10^{-4}$ and $v = 2.10^{-6}$). The other parameters are set for the smallest chromosome in mouse, so with lower h and d than the mean in the entire genome ($h = 800$, $d = 8$, $\bar{y} = 0.2$). This rescaling leaves approximately invariant the following quantities: D, z, σ and s_0. The settings just presented represent our reference for empirical confrontation. Based on this reference, several variations of the model were also explored, which are described below.

Model predictions

The simulator was calibrated based on these rough empirical estimates for its parameters, then run and monitored for several key summary statistics, specifically, the genetic diversity of *PRDM9* (D), the proportion of eroded sites ($\bar{z}$), the mean haplo-insufficiency

at the birth of the allele (σ_0^{het}), the mean haplo-insufficiency over alleles sampled at the equilibrium regime ($\sigma_{\bar{z}}^{hemi}$) and the selection coefficient experienced by new *PRDM9* alleles (s_0). The predictions are shown in Table 2.2 and examples of Red Queen dynamics are shown in Figs 2.7 and 2.8.

Table 2.2: **Empirical calibration experiments.**

u	$\bar{y}$	d	c_{hom}	n_{mei}	D	$\bar{z}$	σ_0	$\sigma_{\bar{z}}$	s_0
3×10^{-6}	0.2	8	2	1	1	0.25	2.9×10^{-2} [a]	5.2×10^{-2} [a]	-2.2×10^{-2}
3×10^{-6}	0.44	8	1	1	2.8	0.05	1.5×10^{-5} [a]	-4.9×10^{-7} [a]	5.5×10^{-4}
3×10^{-6}	0.3	8	1.5	1	1.1	0.22	1×10^{-2} [a]	1.5×10^{-2} [a]	-6×10^{-3}
3×10^{-5}	0.2	8	2	1	1.2	0.23	2.9×10^{-2} [a]	4.6×10^{-2} [a]	-2.2×10^{-2}
3×10^{-6}	2	8	2	1	2.1	0.23	5.1×10^{-4} [a]	7.8×10^{-4} [a]	-8.5×10^{-5}
3×10^{-6}	0.2	8	2	5	2.2	0.22	2.9×10^{-2} [a]	6.9×10^{-2} [a]	6.5×10^{-6}
3×10^{-6}	0.2	24	2	1	2.5	0.26	5.7×10^{-5} [a]	3.3×10^{-3} [a]	1.5×10^{-4}

Equilibrium values of output summary statistics at equilibrium (*PRDM9* diversity D,mean level of erosion $\bar{z}$, mean haplo-insufficiency at the birth of the allele (σ_0), mean haplo-insufficiency at the equilibrium ($\sigma_{\bar{z}}$) and the mean selection coefficient experienced by new *PRDM9* alleles (s_0) as a function of the input parameters (mutation rate at *PRDM9* locus (u), mean of the affinity distribution ($\bar{y}$) , number of DSB par meiosis (d), dosage coefficient (c_{hom}) and maximum number of meiosis allowed for each individual), and with fixed values for the other parameters ($d = 8$, $h = 800$ and $v = 10^{-7}$)
[a] Haplo-insufficiency is here formally defined as the relative difference in success rate of meiosis between homozygotes and hemizygotes. This is equivalent to the relative difference in fertility, except in the case where $n_{mei} = 5$, where fitness and rate of successful meiosis are not proportional.

The first line of Table 2.2 reports the results obtained from simulations run under the parameter values corresponding to our reference for empirical comparison (see also Fig 2.7 for a typical simulation trajectory). The level of erosion (z) predicted by the model is consistent with what is known in the mouse (between 20% and 50% of erosion [137; 124; 111]). However, the predicted *PRDM9* diversity appears to be too low: these simulations result in a monomorphic regime (*i.e.* $D \sim 1$), whereas the *PRDM9* diversity observed in natural populations of mice is of the order of $D = 6$ to $D = 18$ [137] (See Supplementary materials section 5.2). Of note, the diversity predicted by the model is the diversity of functionally different alleles (owing to the assumption made by our model of non-overlapping sets of targets for different alleles). In reality, closely related alleles can share a substantial fraction of their targets [137]. However, even accounting for this redundancy, the empirical functional diversity is still typically greater than 1, of the order of 2 to 6 in mouse subspecies [137] (See Supplementary materials section 5.2). Thus, the monomorphic regime obtained here seems to be in contradiction with empirical observations.

This low diversity can be explained by the fact that we are in a range of parame-

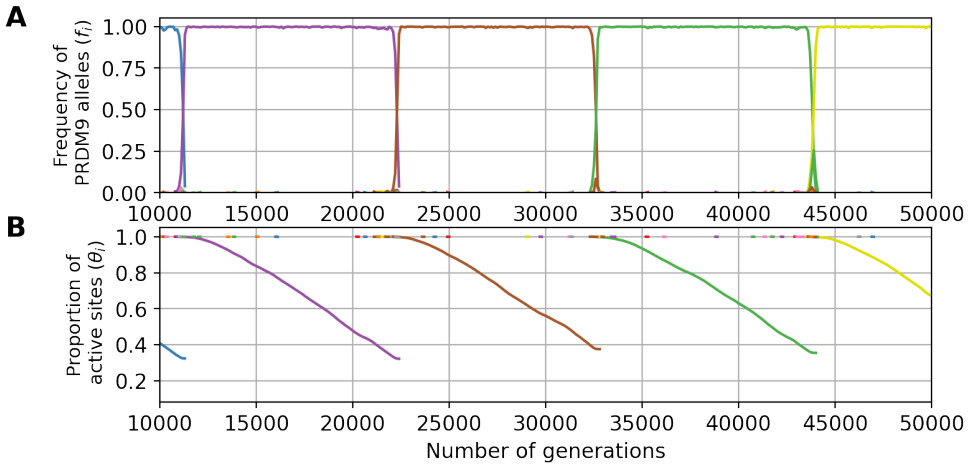

Figure 2.7: **An example evolutionary trajectory under the fitness scheme allowing for only one meiosis per individual ($n_{mei} = 1$, and with parameters $u = 6 \times 10^{-5}$ and $\bar{y} = 0.2$).** *Variation through time of the frequency of each PRDM9 allele (A) and their proportion of active sites (B).*

ters involving an eviction regime due to gene dosage effects (compare Fig 2.7A and Fig 2.6D,E,F), such that one allele dominates in the population, while all the other alleles are counter-selected except during the short phases of allelic replacement (on average, the mean selection coefficient experienced by new alleles, s_0, is negative, Table 2.1).

The fact that eviction is caused by gene dosage can be further verified by running the model without gene dosage (i.e. $c = 1$, Table 2.2, line 2), in which case higher levels of diversity are produced. Intermediate levels of gene dosage ($c = 1.5$, Table 2.2, line 3), on the other hand, give results that are essentially identical to those obtained with a dosage directly proportional to the number of gene copies ($c = 2$). Gene dosage also results in a non-negligible haplo-insufficiency, especially in old alleles (Table 2.2 column $\sigma_{\bar{z}}$), with a reduction of a few percents in the success of meiosis in hemizygotes (or, equivalently, heterozygotes with two alleles of same age), compared to homozygotes. Interestingly, such levels of haplo-insufficiency are comparable to those observed in old alleles in the mouse (B6, C3H [138]). The predicted haplo-insufficiency of young alleles σ_0 is weaker (Table2.2). Such an age-dependency for the impact of gene dosage was previously suggested [111]. Nevertheless, at least in its current form and under those parameter values, the model does not predict an empirically reasonable regime.

Increasing the *PRDM9* functional mutation rate (u) by as much as a factor 10 is not sufficient to get out of the eviction regime, resulting instead in a predicted *PRDM9* diversity still very close to one (fourth row of Table 2.2). Alternatively, increasing the mean affinity (or, equivalently, the concentration of PRDM9 in the cell) allows for higher levels of *PRDM9* equilibrium diversity while maintaining empirically reasonable levels of erosion (fifth row of Table 2.2). However, the mean number of sites bound by PRDM9 in a meiocyte would then predicted to be too large ($\sim 20,000$), compared to current empirical knowleldge ($\sim 5,000$ [80]). In addition, in this parameter regime, the model predicts very low levels of haplo-insufficiency ($\sigma_{\bar{z}} < 1 \times 10^{-3}$ for old alleles), substantially

lower than empirically observed levels in the mouse ($\sigma_{\bar{z}} > 1 \times 10^{-2}$ [138]).

The model is naive in several other aspects. First, gametes may often not be limiting and, as a result, the fitness of an individual may not be proportional to the probability of success of meiosis, such as assumed by the model thus far. A less-than-linear relation between fitness and success of meiosis can be modeled by increasing the number of meiosis that an individual can attempt before being declared sterile. Allowing for 5 attempts, thus mimicking a situation where an up to 80% reduction in the number of gametes would not substantially affect the reproductive success of individuals (sixth row of Table 2.2), we observe a higher functional diversity with a still reasonable level of erosion. The predicted reduction in meiosis success in hemizygotes is also consistent with the ones observed empirically in mice ($\sim$ 2 to 5% [138]). However, the fitness now reacts more weakly to variation in the success of meiosis, and as a result, the model is running in an nearly-neutral regime, with a mean scaled selection coefficient acting on new alleles entering the population (s_0) smaller than $1/N_e = 10^{-5}$, such that the turnover of recombination landscapes is mostly driven by the neutral mutational turnover at the *PRDM9* locus. Although this could be seen as a possible working regime for the evolutionary dynamics of recombination given the high mutation rate at this locus, it is incompatible with the empirical support previously found for a positive selection acting on *PRDM9* [95; 68].

Alternatively, the version of the model considered thus far assumes that the total number of DSBs induced along the chromosome does not depend on the subsequent steps of the process. In reality, DSBs are tightly regulated, in a way that may entail a negative feedback inhibiting the creation of further DSBs once the chromosome has managed to synapse. For that reason, the mean number of DSBs per successful meiosis, which is what is empirically measured [17], may be substantially smaller than the maximum number of DSBs allowed before a meiocyte undergoes apoptosis. Yet, the success rate of meiosis depends on the maximum, not on the mean number. As a way to indirectly account for this, we performed a last simulation allowing for $d = 24$ DSBs (instead of 8, last row of Table 2.2, Fig 2.8). Running the model under this configuration results in an evolutionary dynamics which is not in the eviction regime, with moderate levels of diversity ($D = 2.5$), reasonable levels of erosion ($\bar{z} = 0.26$), and strong positive selection on new *PRDM9* alleles ($s_0 > 10^{-4}$). On the other hand, the haplo-insuffiency predicted for old alleles is now weaker than that observed empirically [138].

Altogether, these results show that the requirement of symmetric binding can result in a Red Queen maintaining erosion at moderate levels under empirically reasonable parameter values. On the other hand, there is only a narrow window for which eviction due to dosage is avoided and a sufficiently high *PRDM9* diversity is maintained, while still having haplo-insufficiency for old alleles and strong selection acting on *PRDM9*.

2.5 Discussion

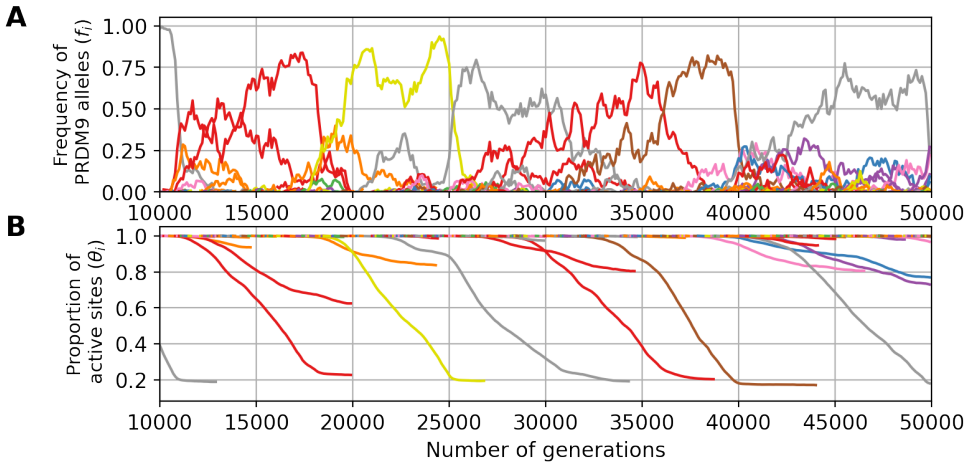

Figure 2.8: **An example evolutionary trajectory with 24 DSBs ($d = 24$, $u = 3 \times 10^{-6}$ and $\bar{y} = 0.2$).** *Variation through time of the frequency of each PRDM9 allele (A) and their proportion of active sites (B).*

2.5.1 Fundamental role of the symmetrical binding of PRDM9

The first studies on the function of *PRDM9* uncovered its role in targeting DSBs at specific sites, by its capacity to set histone marks to recruit the DSB machinery [62; 141; 76; 142]. More recently, Davies *et al.* (2016) discovered that this gene plays another important role during meiosis, namely, by facilitating chromosome pairing [111]. In this process, the symmetrical binding of PRDM9 to its target sites plays a key role. By bringing to the chromosomal axis the sites on the homologue, PRDM9 presents them to the single-stranded DNA produced by the resection around the DSB, thereby facilitating the search for the complementary sequence (Fig 2.1). This second function of *PRDM9*, combined with the differential erosion of the target sites between sub-species (which is itself a direct consequence of the first function of *PRDM9*, the recruitment of the DSB machinery), was found to be sufficient for explaining most aspects of the hybrid sterility pattern [111].

Here, we show that, similarly, these two aspects of the function of *PRDM9* during meiosis provide the necessary ingredients for explaining the intragenomic Red Queen occurring in the context of a single population – thus giving a globally coherent picture bridging the gap between the molecular mechanisms and the evolutionary dynamics of recombination. Using scaling experiments, we have also characterized what determines the steady state regime. Overall, we recover the main results of Úbeda & Wilkins 2011 [85] and Latrille *et al.* 2017 [96], although now with a mechanistic foundation. In particular, *PRDM9* diversity is mostly determined by the mutational input at the *PRDM9* locus u, while the mean erosion and fertility at equilibrium are the result of a balance between erosion (v, g) and selection of new alleles (u, α).

2.5.2 Impact of gene dosage of *PRDM9* on the Red Queen dynamics

Gene dosage of *PRDM9* implies that a homozygote has a PRDM9 concentration for its allele which is twice that of each of the two alleles of a heterozygote. By the law of mass action, everything else being equal, increased dosage results in an increased occupancy of the targets and therefore an increased probability of symmetrical binding and a higher fertility. This clearly impacts the Red Queen dynamics for certain combinations of parameters, by acting against diversity and leading to a monomorphic regime characterized by the eviction of minor alleles (mostly found in a heterozygous state) by the currently dominant allele (mostly present in a homozygous state). The higher the selection coefficient associated to the homozygous advantage (captured here by σ_0), the stronger the effect against diversity. This effect is mitigated when PRDM9 is in high concentration in the cell. Indeed, a large concentration of PRDM9 makes the binding probability at a target of a given affinity less responsive to gene dosage (Supplementary Figure 5.6), thus reducing the fertility gap between a homo- and a heterozygote (Supplementary Figure 7). Based on these observations, we therefore hypothesize that the *PRDM9* expression level has been selected at a sufficiently high level so as to limit the impact of gene dosage. This is an interesting hypothesis that could be studied in the future.

2.5.3 Comparison with Baker *et al.*'s model

In a recent article, Baker *et al.* [134] also propose a mechanistic model of the *PRDM9* Red Queen. Several of their conclusions agree with ours. Most notably, Baker *et al.* also recognize that the symmetric binding of PRDM9 provides the key ingredient for running the Red Queen. However, the two studies differ on some important points.

First, in our model, the amount of PRDM9 protein is not limiting (that is, most PRDM9 molecules are not bound to the target sites), whereas Baker *et al.* assume strong competition between targets for PRDM9 proteins (most molecules are bound). Of note, gene dosage and competition are two distinct aspects of the steady-state equilibrium of PRDM9 binding. Gene dosage is a direct consequence of the law of mass action (site occupancy is an increasing function of the free concentration of PRDM9, see Supplementary Figure 5.1), a phenomenon that can happen even when PRDM9 is not limiting, as shown here. Competition between targets, on the other hand, relates to the fact that the inactivation of some of the binding sites results in extra PRDM9 molecules available for other sites to bind, an effect that is negligible if PRDM9 is in excess (as in our model), but important otherwise (as in Baker et al's model [134]). Whether PRDM9 is limiting ultimately depends on the total concentration of PRDM9 relative to the mean affinity and total number of sites, a question that still remains open. In any case, our model shows that competition between targets is not a necessary feature to drive the evolutionary dynamics of *PRDM9*.

The second difference concerns the affinity distribution. Baker *et al.* use a two-heat distribution. Here we use an exponential distribution (Supplementary Figure 5.5), which

is motivated by results from Chip-seq experiments [137], although with some uncertainty regarding the shape of the distribution in the low affinity range, which is not captured by Chip-seq. The use of these different affinity distributions has clear consequences on the behavior of models with respect to dosage. Indeed, increasing the gene dosage can have two opposing effects. On the one hand, a higher dosage increases the symmetrical binding at sites of intermediate affinity. On the other hand, it also results in more sites of low affinity being bound, and this, most often asymmetrically. Depending on the balance between these two effects, different outcomes are obtained. In our case, where the affinity distribution has a moderate variance, the increased symmetrical binding of sites of intermediate affinity always wins. As a result, there is always an advantage to increasing dosage. That is, homozygotes always have an advantage over heterozygotes, which creates an eviction regime that tends to play against diversity. A contrario, in Baker's model, the two-heat affinity distribution results in a non-monotonous dependency, with a turning point, such that the dosage is in favor of homozygotes for young alleles, but in favor of heterozygotes for old alleles. This turning point is one potential solution to the problem of eviction. Here, we present an alternative solution, which is captured in our model by the statistics $\sigma_0\tau$. Intuitively, eviction does not take place if the homozygote advantage is sufficiently weak and erosion sufficiently rapid, such that ageing alleles are replaced before eviction has enough time to take place.

As it turns out, it is not so easy to find empirically reasonable parameter configurations such that the eviction regime is avoided, although this could be the consequence of other aspects of the biology of meiosis and reproduction being missed by the model. On the other hand, our model predicts that old alleles should be haplo-insufficient, thus in agreement with what is observed in the mouse (for allele B6 and C3H [138]). This haplo-insufficiency of old alleles is not observed systematically in Baker's *et al.* model, but only occasionally and in small populations.

Altogether it would be useful to empirically investigate the haplo-insufficiency for young and old alleles over a broader range of alleles and species, in order to determine whether a lifelong homozygous advantage is systematic or occasional. More globally, it will be important to unravel the exact roles of affinity distribution, PRDM9 concentration and competition between targets in the evolutionary dynamics. Finally, all this raises the question of a possible evolution of the *PRDM9* expression level, the affinity (whether in its distribution or in its mean) and the number of targets recognized per allele.

2.5.4 Current limitations and perspectives

In addition to those discussed above, the model introduced here has other limitations. First, in its current form, the model implements only limited variance among alleles in their strength at birth. Yet, in reality, some alleles are dominant over others [36], such that, in a heterozygote for a strong and a weak allele, DSBs are more often produced at the target sites of the stronger allele. Of note, although some dominance is expected to emerge purely as a consequence of erosion (with younger alleles being on average

dominant over older ones), a substantial part of it appears to be instead related to intrinsic differences between alleles regardless of their age [143; 25]. If stronger alleles have an advantage over weaker alleles, either because they promote a higher binding symmetry or just because of their higher penetrance, then weaker alleles should be less likely to invade. As a result, the population would tend to be dominated by stronger alleles. In good approximation, this would amount to running the model under a lower effective mutation rate (now to be taken as the rate at which new functional and sufficiently strong *PRDM9* alleles are being produced by mutation) but otherwise would not fundamentally change the evolutionary dynamics.

Second, the model currently assumes that all DSBs are repaired with the homologue. In reality, some are repaired with the sister chromatid [25]. This simplifying assumption, however, should have a moderate impact on our conclusions, as it essentially amounts to a change in the rate of erosion, which can be accounted for by considering a lower mutation rate v at the target sites. Our experiments suggest that small changes in v do not fundamentally impact the dynamics.

Although deserving more careful examination, the limitations just mentioned are therefore probably minor. Perhaps a more fundamental issue is to fully understand how the eviction regime induced by gene dosage effects can be avoided, so as to predict reasonable levels of *PRDM9* diversity, while having a Red Queen regime driven by strong positive selection on *PRDM9*. As it stands, the model appears to be stuck in a dilemma, such that either dosage has a substantial impact on fertility, but then eviction takes place and the model fails at explaining empirically observed levels of *PRDM9* diversity, or the effects of dosage are made weaker by assuming globally less stringent conditions for achieving a good fertility (such as allowing for an excess in PRDM9 protein, or in gametes or in DSBs), but then the fitness differences between old and new alleles also become small and the model approaches a nearly-neutral regime, in which the turnover of *PRDM9* alleles is primarily driven by the mutational turnover at the locus. Although the latter regime may not be so unreasonable from an evolutionary point of view, it fails at explaining the positive selection observed on *PRDM9*, or its haplo-insufficiency, depending on the detailed model configuration (last two rows of Table 2.2).

This dilemma is still in need of a robust and convincing explanation. In this respect, as mentioned above, the questions of the affinity distribution and the concentration of PRDM9 relative to the affinity of the target sites should be explored in depth. Alternatively, the empirical regime of the Red Queen may possibly alternate between long nearly-neutral epochs, with occasional bouts of positive selection whenever the current alleles become too eroded. Such occasional epsiodes of positive selection could be sufficient to induce the empirically observed patterns of accelerated evolution of the zinc finger domain at the non-synonymous level. The impact of Hill-Robertson interference and its dissipation on the evolutionary dynamics of *PRDM9* could also contribute to maintaining a high diversity at the *PRDM9* locus, a point that may also need to be investigated. Finally, population structure could play a role, by maintaining different pools of alleles in different sub-populations connected by recurrent migration, thus resulting in a large

diversity at the metapopulation level, and this, in spite of non-negligible gene dosage effects. Population structure is also pointing toward the other big question still in need of a model-based exploration: the potential role of *PRDM9* in hybrid sterility and speciation.

2.6 Materials and methods

2.6.1 The model

The simulation model is graphically summarized in Fig 2.2, and its key parameters are listed in Table 2.1. This model assumes a population of N diploid individuals, whose genetic map is composed of a single chromosome. A *PRDM9* locus is located on this chromosome. Each *PRDM9* allele has a set of h binding sites, each of which has an intrinsic binding affinity for its cognate PRDM9 protein. The positions of binding sites along the chromosome are drawn uniformly at random and their affinities are drawn according to an exponential law of parameter $\bar{y}$ (Supplementary Figure 5.5). Each target site is associated to a unique allele and different sites cannot overlap. In practice, this is implemented by encoding the chromosome as an array of L slots, such that, upon creation of a new *PRDM9* allele by mutation, each binding site of this allele chooses one of the available slots uniformly at random. Given the composition of the population of the current generation, the next generation, of the same population size, is generated as follows.

Mutations

The *PRDM9* locus mutates with a probability u per gene copy. Given a mutation, a new functional *PRDM9* allele is created to which are associated h new sites along the genome, according to the procedure just described. Next, each target site recognized by each allele currently present in the population mutates with a probability v. This type of mutation results in a complete inactivation of the target site. Of note, all target sites are assumed to be monomorphic for the active variant at the time of the birth of the corresponding *PRDM9* allele. As a result, the loss of target sites is entirely contributed by gene conversion acting on inactivating mutations that have occurred after the birth of the allele. Also, we neglect new target sites that might arise by mutation, which is reasonable since hotspot alleles newly arisen by mutation would be at a very high risk of being lost by conversion.

Meiosis and reproduction

A meiosis is attempted on a randomly selected individual, according to the following steps. Of note, the meiosis can fail (in which case we assume that the meiocyte undergoes apoptosis) for multiple reasons, all of which will be described below. In the simulations presented here, unless stated otherwise, whenever meiosis fails, then a new individual is chosen at random from the population and a new meiosis is attempted.

First, the two homologous chromosomes are replicated, thus creating a set of 4 chromatids. Then, PRDM9 binds to its target site according to an occupation probability determined by the chemical equilibrium. A binding site i has an affinity

$$K_i = \frac{[PS_i]}{[P]_{free}[S_i]} = \frac{x_i}{[P]_{free}(1-x_i)}, \tag{2.8}$$

where $[PS_i] = x_i$ and $[S_i] = 1 - x_i$ are the proportions of target site i (across meiocytes) which are occupied by PRDM9 or free, respectively. Thus:

$$x_i = \frac{[P]_{free}K_i}{1 + [P]_{free}K_i}. \tag{2.9}$$

We assume that PRDM9 is not limiting, meaning that most PRDM9 molecules are free $[P]_{free} \approx [P]_{tot}$; there is therefore a total absence of competition between binding sites. Thus, if we define the rescaled affinity as $y_i = [P]_{tot}K_i$, then the occupancy probability can be re-written :

$$x_i = \frac{y_i}{1 + y_i}. \tag{2.10}$$

Based on this equation, for each target site, binding is randomly determined, by drawing a Bernoulli random variable of parameter x_i for site i. Of note, at a given target locus, there are four instances of the binding site (one on each of the two sister chromatids for each of the two homologues), and binding is determined independently for each of those instances.

Once the occupation status of all binding sites has been determined, the total number k of sites bound by PRDM9 over the four chromatids, is calculated. If $k = 0$, meiosis fails. Otherwise, each site bound by PRDM9 undergoes a DSB with a probability equal to $p = \min\left(1, \frac{d}{k}\right)$, with d being a parameter of the model. In most experiments, we use $d = 6$. Thus, in the most frequent case where $k > d$, an average number of d DSBs are being produced. This procedure aims to model the regulation of the total number of DSBs through the genome, which in mammals seems to be independent from PRDM9 binding [17; 111].

Next, the four chromatids are scanned for symmetrically bound sites, which the model assumes are essential for chromosome pairing (see introduction). If no such site is detected, meiosis fails. Otherwise, one of the symmetrically bound sites is uniformly picked at random and becomes the initiation site for a CO. Thus, only one CO is performed per meiosis (and this, in order to model CO interference), and all other DSBs are repaired as NCO events. Note that in our model we do not allow for the possibility of DSB repair by the sister chromatid.

A successful meiosis therefore produces 4 chromatids, with exactly one CO between two of them, and some events of gene conversion at all sites that have undergone a DSB. Of note, in the presence of inactive versions of the binding sites (created by mutations, see above), these gene conversion events are effectively implementing the hotspot conversion paradox [83]. Finally one chromatid is randomly picked among the tetrad to become a gamete. The whole procedure is repeated, starting from the choice of a new individual until another gamete is obtained. Then these two gametes merge and create a new diploid

individual of the next generation. All these steps, from the choice of an individual until the choice of a chromatid among the tetrad, are performed as many time as needed until the complete fill of the next generation.

gene dosage

In the case where the gene dosage of PRDM9 is taken into account, it is assumed that, for a given *PRDM9* allele, a homozygote will produce twice the concentration of the corresponding protein compared to a heterozygote. By assuming that $[P]_{tot}^{homo} = 2[P]_{tot}^{het}$, the probability that a PRDM9 protein binds to a site i of rescaled affinity y_i is

$$x_i = \frac{cy_i}{1 + cy_i} \tag{2.11}$$

where $c = 1$ for a heterozygote and $c = 2$ for a homozygote.

2.6.2 Summary statistics

Several summary statistics were monitored during the run of the simulation program. They were used, first, to evaluate the time until convergence to the equilibrium regime (burn-in). The burn-in was taken in excess, so as to be adapted to all simulation settings. In practice, the burn-in is set at 10000 generations over a total of $50,000$ generations. Second, the summary statistics were averaged over the entire run (burn-in excluded), thus giving a quantitative characterization of the equilibrium regime, as a function of the model parameters. The main statistics are the *PRDM9* diversity in the population (D), the mean proportion of binding sites that are still active per allele θ, the mean symmetrical binding probability (q) and the mean fertility (w).

Diversity is defined as the effective number of *PRDM9* alleles, or in other words as the inverse of the homozygosity [96]:

$$D_t = \frac{1}{\sum\limits_i f_{i,t}^2} \tag{2.12}$$

where $f_{i,t}$ is the frequency of allele i at time t. With this definition, when K alleles segregate each at frequency $\frac{1}{K}$ in the population, the diversity D is equal to K.

The mean proportion of binding sites that are still active per allele is defined as:

$$\theta_t = \sum_i f_{i,t}\theta_{i,t} \tag{2.13}$$

where $\theta_{i,t}$ is the proportion of sites that are still active for allele i at time t. Equivalently, $z_{i,t} = 1 - \theta_{i,t}$ gives a measure of the level of erosion of the target sites of allele i and, by extension $z_t = 1 - \theta_t$ gives the mean erosion level.

The probability of symmetrical binding q corresponds to the mean probability of having a PRDM9 protein bound on at least one of the two chromatids of the homologous chromosome at a certain position, given that a DSB has occured at this very position.

In the case of a homozygous individual, this quantity can be obtained analytically for a given complete diploid genotype and is given by:

$$q^{hom} = \frac{2 < x^2 > - < x^3 >}{< x >}, \tag{2.14}$$

where $x = \frac{cy}{1+cy}$ is the occupancy of a binding site of affinity y and, for any j, $< x^j >$ is the mean of x^j over all sites, taking into account their affinity distribution. In the case of a heterozygous individual, with two *PRDM9* alleles 1 and 2, the symmetrical binding rate is:

$$q^{het} = \frac{2 < x^2 >_1 - < x^3 >_1 + 2 < x^2 >_2 - < x^3 >_2}{< x >_1 + < x >_2}, \tag{2.15}$$

where the subscripts 1 and 2 correspond to averages over sites of allele 1 and 2, respectively. This statistic was then averaged over all diploid complete genotypes present in the population at a given generation.

For a given genotype, the fertility can be computed analytically. Here we assumed that fertility in proportional to the rate of success of meiosis. Thus, it is equivalent to 1-(the mean probability of failure of meiosis) characterized by the absence of a DSB in a symmetric site. In turn, the number of DSBs in a symmetric site follows a Poisson law with parameter dq where d is the average number of DSBs per meiocyte and q is the mean probability of symmetrical binding for this allele. Thus, the mean fertility of an allele can be expressed as follows :

$$w = 1 - e^{-dq} \tag{2.16}$$

For a given allele, and at a given time, this statistic was then averaged over all diploid complete genotypes carrying this allele present in the population at a given generation (and then averaged over all generations of the run).

2.6.3 Scaling experiments

In order to visualize how the equilibrium regime varies according to the model parameters, first, a central parameter configuration was chosen, which will represent a fixed reference across all scaling experiments. Then only one parameter at a time (or two for bi-dimensional scaling) is systematically varied over a range of regularly spaced values on both sides of the reference configuration. This parameter is fixed along the entire simulation and is variable only between simulations. For each parameter value over this range, a complete simulation is run. Once the equilibrium is reached for a given simulation setting, the summary statistics described above are averaged over the entire trajectory, excluding the burn-in. These mean equilibrium values, which characterize and quantify the stationary regime of the model, were finally plotted as a function of the varying parameter(s).

The central parameters were chosen as follows. First, the parameters of population size N was set to $5,000$, corresponding to the maximum population size that can be afforded computationally. Then the number of sites recognized as target sites by each

PDRM9 allele's protein was set to $h = 400$ in order to get closer to the average number of target sites found on the smallest chromosome of the mouse (around 800 sites) while limiting the memory requirements. The mean for the affinity distribution of the target was set to $\bar{y} = 6$. The parameter d representing the number average of DSBs in a meiocyte is set at 6 which corresponds to the approximate number of DSBs found on the smallest chromosome of the mouse. For uni-dimensional scaling, the parameter v, the mutation rate at target sites, was set to 2×10^{-6} and the parameter u was set to 5×10^{-5}. For the bi-dimensional scaling of d and $\bar{y}$, $u = 2 \times 10^{-4}$ and $v = 5 \times 10^{-5}$.

2.7 Data accessibility

The model, and the codes to generate the figures can be found at https://github.com/alicegenestier/Red_Queen_PRDM9_Panmictic.git. The model was implemented in C++ and all figures were created using Python.

2.8 Acknowledgments

All simulations of this work were performed using the computing facilities of the CC LBBE/PRABI.

2.9 Competing interest

The authors have declared that no competing interests exist.

2.10 Funding

Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec.

Bibliographie

[1] Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature. 2016;530(7589):171–176. doi:10.1038/nature16931. *Cited at pages 26, 29, 30, 39, 41, 57, 58, 60, 65, 76, 80, 89, 91, 96, 97, 98, 108, 109, 122*

[2] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science. 2009;323(5912):373–375. doi:10.1126/science.1163601. *Cited at pages 14, 26, 28, 30, 39, 80, 81, 96*

[3] Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLOS Genetics. 2009;5(12):e1000753. doi:10.1371/journal.pgen.1000753. *Cited at pages 21, 23, 30, 39, 41, 59, 81, 108*

[4] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science. 2005;310(5746):321–324. doi:10.1126/science.1117196. *Cited at pages 12, 39*

[5] Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics. 2001;29(2):217–222. doi:10.1038/ng1001-217. *Cited at pages 11, 12, 39*

[6] Kauppi L, Jeffreys AJ, Keeney S. Where the crossovers are: recombination distributions in mammals. Nature Reviews Genetics. 2004;5(6):413–424. doi:10.1038/nrg1346. *Cited at pages 11, 39*

[7] McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science. 2004;304(5670):581–584. doi:10.1126/science.1092500. *Cited at pages 12, 39*

[8] Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome. Genetics. 2012;191(3):757–764. doi:10.1534/genetics.112.141036. *Cited at pages 12, 39*

[9] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science. 2010;327(5967):836–840. doi:10.1126/science.1183439. *Cited at pages 13, 21, 28, 30, 39, 117*

[10] Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Mei-

otic Recombination. Science. 2010;327(5967):876–879. doi:10.1126/science.1182363. *Cited at pages 13, 21, 28, 39, 40*

[11] Forejt J, Iványi P. Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.). Genetics Research. 1974;24(2):189–206. doi:10.1017/S0016672300015214. *Cited at pages 26, 79*

[12] Hayashi K, Yoshida K, Matsui Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature. 2015;438(7066):374–378. doi:10.1038/nature04112. *Cited at pages 13, 14, 28, 40, 60*

[13] Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. PLOS Genetics. 2016;12(6):e1006146. doi:10.1371/journal.pgen.1006146. *Cited at page 40*

[14] Lange J, Yamada S, Tischfield SE, Pan J, Kim S, Zhu X, et al. The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. Cell. 2016;167(3):695–708.e16. doi:10.1016/j.cell.2016.09.035. *Cited at pages 6, 11, 12, 14, 40, 59, 65*

[15] Grey C, Baudat F, Massy Bd. PRDM9, a driver of the genetic map. PLOS Genetics. 2018;14(8):e1007479. doi:10.1371/journal.pgen.1007479. *Cited at pages 13, 40*

[16] Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nature Communications. 2019;10:3900. doi:10.1038/s41467-019-11675-y. *Cited at pages 8, 29, 40, 42, 63, 115*

[17] Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997;94(15):8058–8063. doi:10.1073/pnas.94.15.8058. *Cited at pages 16, 18, 40, 65, 80, 106, 123*

[18] Lesecque Y, Glémin S, Lartillot N, Mouchiroud D, Duret L. The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. PLOS Genetics. 2014;10(11):e1004790. doi:10.1371/journal.pgen.1004790. *Cited at pages 12, 21, 40*

[19] Úbeda F, Wilkins JF. The Red Queen theory of recombination hotspots. Journal of Evolutionary Biology. 2011;24(3):541–553. doi:10.1111/j.1420-9101.2010.02187.x. *Cited at pages 18, 19, 21, 30, 36, 40, 41, 43, 60, 80, 107*

[20] Van Valen L. Molecular evolution as predicted by natural selection. Journal of Molecular Evolution. 1974;3(2):89–101. doi:10.1007/BF01796554. *Cited at pages 19, 40*

[21] Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science. 2012;336(6078):193–198. doi:10.1126/science.1216872. *Cited at pages 10, 12, 21, 40*

[22] Jeffreys AJ, Cotton VE, Neumann R, Lam KWG. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. Proceedings of the National Academy of Sciences. 2013;110(2):600–605. doi:10.1073/pnas.1220813110. *Cited at pages 21, 40, 56*

[23] Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ. 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. 2011;108(30):12378–12383. doi:10.1073/pnas.1109531108. *Cited at pages 13, 21, 41*

[24] Kono H, Tamura M, Osada N, Suzuki H, Abe K, Moriwaki K, et al. Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. DNA Research. 2014;21(3):315–326. doi:10.1093/dnares/dst059. *Cited at pages 13, 21, 41, 108, 136*

[25] Buard J, Rivals E, Segonzac DDd, Garres C, Caminade P, Massy Bd, et al. Diversity of Prdm9 Zinc Finger Array in Wild Mice Unravels New Facets of the Evolutionary Turnover of this Coding Minisatellite. PLOS ONE. 2014;9(1):e85021. doi:10.1371/journal.pone.0085021. *Cited at pages 13, 21, 41, 59, 108*

[26] Vara C, Capilla L, Ferretti L, Ledda A, Sánchez-Guillén RA, Gabriel SI, et al. PRDM9 Diversity at Fine Geographical Scale Reveals Contrasting Evolutionary Patterns and Functional Constraints in Natural Populations of House Mice. Molecular Biology and Evolution. 2019;36(8):1686–1700. doi:10.1093/molbev/msz091. *Cited at page 41*

[27] Alleva B, Brick K, Pratto F, Huang M, Camerini-Otero RD. Cataloging Human PRDM9 Allelic Variation Using Long-Read Sequencing Reveals PRDM9 Population Specificity and Two Distinct Groupings of Related Alleles. Frontiers in Cell and Developmental Biology. 2021;9. *Cited at pages 13, 41, 108*

[28] Latrille T, Duret L, Lartillot N. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017;372(1736):20160463. doi:10.1098/rstb.2016.0463. *Cited at pages 21, 30, 36, 41, 43, 46, 51, 52, 56, 60, 66, 83, 85, 107, 125, 128, 129*

[29] Smagulova F, Brick K, Pu Y, Camerini-Otero RD, Petukhova GV. The evolutionary turnover of recombination hot spots contributes to speciation in mice. Genes & Development. 2016;30(3):266–280. doi:10.1101/gad.270009.115. *Cited at pages 41, 56, 57, 62, 80, 91, 96, 97, 108, 109, 112, 115, 136*

[30] Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, et al. PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. PLOS Genetics. 2015;11(1):e1004916. doi:10.1371/journal.pgen.1004916. *Cited at pages 28, 41, 56, 57, 81, 89, 108, 115*

[31] Gregorová S, Forejt J. PWD/Ph and PWK/Ph Inbred Mouse Strains of Mus cyrchar
cyrt. musculus Subspecies-a Valuable Resource of Phenotypic Variations and Genomic Polymorphisms. Folia biologica. 2000;46:31–41. *Cited at pages 26, 29, 41, 80, 81*

[32] Gregorova S, Gergelits V, Chvatalova I, Bhattacharyya T, Valiskova B, Fotopulosova V, et al. Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. eLife. 2018;7:e34282. doi:10.7554/eLife.34282. *Cited at pages 29, 30, 41, 80, 81, 89, 96, 107*

[33] Grey C, Clément JAJ, Buard J, Leblanc B, Gut I, Gut M, et al. In vivo binding of PRDM9 reveals interactions with noncanonical genomic sites. Genome Research. 2017;27(4):580–590. doi:10.1101/gr.217240.116. *Cited at pages 11, 14, 41*

[34] Parvanov ED, Tian H, Billings T, Saxl RL, Spruce C, Aithal R, et al. PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis. Molecular Biology of the Cell. 2017;28(3):488–499. doi:10.1091/mbc.E16-09-0686. *Cited at pages 14, 41*

[35] Borde V, de Massy B. Programmed induction of DNA double strand breaks during meiosis: setting up communication between DNA and the chromosome structure. Current Opinion in Genetics & Development. 2013;23(2):147–155. doi:10.1016/j.gde.2012.12.002. *Cited at pages 12, 14, 30, 42, 80*

[36] Baker Z, Przeworski M, Sella G. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence; 2022. Available from: https://www.biorxiv.org/content/10.1101/2022.09.27.509707v1. *Cited at pages 30, 36, 37, 43, 61, 80, 96, 107, 110, 112, 115*

[37] Baker CL, Petkova P, Walker M, Flachs P, Mihola O, Trachtulec Z, et al. Multimer Formation Explains Allelic Suppression of PRDM9 Recombination Hotspots. PLoS Genetics. 2015;11(9):e1005512. doi:10.1371/journal.pgen.1005512. *Cited at pages 53, 58, 59, 62, 99, 108, 115*

[38] Salcedo T, Geraldes A, Nachman MW. Nucleotide Variation in Wild and Inbred Mice. Genetics. 2007;177(4):2277–2291. doi:10.1534/genetics.107.079988. *Cited at page 56*

[39] Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012;485(7400):642–645. doi:10.1038/nature11089. *Cited at pages 11, 56, 122*

[40] Kauppi L, Barchi M, Lange J, Baudat F, Jasin M, Keeney S. Numerical constraints and feedback control of double-strand breaks in mouse meiosis. Genes & Development. 2013;27(8):873–886. doi:10.1101/gad.213652.113. *Cited at pages 14, 56*

[41] Baker CL, Walker M, Kajita S, Petkov PM, Paigen K. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Research. 2014;24(5):724–732. doi:10.1101/gr.170167.113. *Cited at pages 14, 29, 56, 58*

[42] Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Research. 2015;25(8):1125–1134. doi:10.1101/gr.186148.114. *Cited at page 56*

[43] Wu H, Mathioudakis N, Diagouraga B, Dong A, Dombrovski L, Baudat F, et al. Molecular Basis for the Regulation of the H3K4 Methyltransferase Activity of PRDM9. Cell Reports. 2013;5(1):13–20. doi:10.1016/j.celrep.2013.08.035. *Cited at page 60*

[44] Eram MS, Bustos SP, Lima-Fernandes E, Siarheyeva A, Senisterra G, Hajian T, et al. Trimethylation of Histone H3 Lysine 36 by Human Methyltransferase PRDM9 Protein. The Journal of Biological Chemistry. 2014;289(17):12177–12188. doi:10.1074/jbc.M113.523183. *Cited at pages 14, 60*

[45] Grey C, Barthès P, Friec GCL, Langa F, Baudat F, Massy Bd. Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. PLOS Biology. 2011;9(10):e1001176. doi:10.1371/journal.pbio.1001176. *Cited at page 60*

[46] Diagouraga B, Clément JAJ, Duret L, Kadlec J, Massy Bd, Baudat F. PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites. Molecular Cell. 2018;69(5):853–865.e6. doi:10.1016/j.molcel.2018.01.033. *Cited at pages 11, 12, 14, 62, 80, 91, 98, 115*

[47] Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, et al. Factors Influencing Meiotic Recombination Revealed by Whole Genome Sequencing of Single Sperm. Science (New York, NY). 2019;363(6433):eaau8861. doi:10.1126/science.aau8861. *Cited at pages 63, 115*

3

A theoretical investigation of the role of *PRDM9* in hybrid sterility

Alice Genestier[1], Laurent Duret[1], Nicolas Lartillot[1]

[1]Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France

Contents

Avant propos

Ce chapitre présente un projet de manuscrit en anglais, en vue d'une publication future, portant sur le rôle possible de *PRDM9* dans la stérilité hybride. Je vais donc résumer rapidement l'article, puis présenter le manuscrit en entier.

Résumé

La stérilité hybride est une forme d'isolement reproducteur post-zygotique survenant lorsque des parents fertiles de deux populations différentes donnent naissance à un hybride stérile. Cet isolement reproductif peut être un des facteurs conduisant à terme à la spéciation. Le gène *PRDM9* a été proposé comme un gène potentiel de spéciation chez les mammifères. Ce gène possède une double fonction molèculaire. Tout d'abord, la protéine codée par ce gène localise les points chauds de recombinaison grâce à sa fonction de triméthyltransférase des histones, conduisant au recrutement de la machinerie de cassure double brin. Deuxièmement, *PRDM9* a un rôle de facilitation de l'appariement des chromosomes homologues par sa liaison symétrique, c-à-d la liaison en la même position sur les deux homologues. Cette dernière fonction a été découverte lors d'expériences sur des hybrides issus de croisements entre deux sous-espèces de souris *Mus musculus musculus musculus* et *Mus musculus domesticus*, qui étaient stériles et chez qui un taux élevé de liaison asymétrique de PRDM9 a été observé. Enfin, ce double rôle de *PRDM9* lors de la méiose a été testé analytiquement et avec des simulations dans une population panmictique, ce qui a permis de découvrir, entre autres, l'origine de la sélection positive agissant sur *PRDM9*. Cependant, le rôle possible de *PRDM9* dans la stérilité hybride n'a pas encore été testé par des modèles théoriques. Dans cet article, nous avons adapté le modèle de population unique (introduit dans notre article précédent), en un modèle bi-populationnel avec génération d'hybrides, afin d'étudier la fréquence et l'intensité de la stérilité hybride en fonction de plusieurs paramètres moléculaires. Notre modèle prédit la stérilité hybride dans certains régimes. Cette stérilité est causée par la liaison asymétrique de PRDM9, la symétrie étant elle-même due à l'érosion différentielle des sites de liaison dans les deux sous-populations. Cette stérilité hybride semble toutefois plutôt marginale et transitoire, ce qui ne plaide pas pour un rôle majeur de *PRDM9* dans la spéciation.

Résumé étendu

Introduction

Dans le chapitre 2, nous avons vu que les mécanismes moléculaires de PRDM9, et en particulier sa liaison symétrique facilitant l'appariement des chromosomes homologues, étaient nécessaires à la réalisation de la Reine Rouge. Notre étude a également permis de comprendre l'origine de la sélection positive agissant sur les nouveaux allèles *PRDM9*. Cependant, l'hypothèse que cette liaison symétrique joue un rôle crucial dans la bonne

réussite de la méiose a été initialement découverte en 2016 par Davies et al [111], lors d'observations de forts taux d'asymétrie de PRDM9 couplés à une stérilité observés chez les hybrides issus de croisements entre deux sous espèces de souris. Cette hypothèse a depuis reçu beaucoup de soutien empirique de la part de plusieurs études, principalement réalisées chez la souris. Indépendamment de ces travaux, la stérilité hybride causée par *PRDM9* a mené à émettre l'hypothèse de *PRDM9* comme gène de spéciation. Or, ce rôle que semble jouer *PRDM9* dans la stérilité hybride n'a pas encore été testé par l'intermédiaire de modèles théoriques. Pour adresser ce point, dans ce chapitre, nous avons adapté le modèle que nous avions précédemment utilisé en population unique afin d'étudier le rôle de PRDM9 dans la stérilité hybride et tenter de quantifier son impact.

Matériel et Méthode

Comme précisé dans l'introduction, le modèle utilisé est le même que celui développé dans le chapitre 2. Ce modèle a été adapté de manière à représenter l'histoire évolutive d'espèces dont deux populations se verraient séparées par une barrière (par exemple géographique), pendant un certain temps, et réaliseraient des hybrides lors de nouveaux contacts. Notre modèle a tourné pendant une dizaine de milliers de générations en population unique faisant référence à la population ancestrale qui se sépare ensuite en deux sous populations de même taille évoluant indépendamment l'une de l'autre. Le processus intrinsèque du modèle, avec la création de nouvelles générations, est inchangé par rapport à la première version en population unique. Cependant, nous avons ajouté la possibilité de réaliser des hybrides issus de croisements entre des individus des deux sous-populations. L'étude des mêmes statistiques descriptives que dans le chapitre précédent est réalisée à la fois chez les hybrides et chez deux populations qui se comportent chacune comme une population unique. Nous avons ensuite comparé les taux de fertilité chez les hybrides et chez les deux populations afin d'identifier leurs différences en fonction des paramètres du modèle.

Résultats

Comme attendu, on observe une Reine Rouge indépendante dans chaque sous-population, et nous observons par ailleurs de la stérilité hybride. Cette dernière vient de l'érosion différentielle des cibles dans les deux populations ne possédant pas les même allèles, ce qui génère un taux d'asymétrie plus important qu'en intra-population. Cette asymétrie, qui est ici une asymétrie de séquence, diminue la probabilité pour une cassure double brin de se réaliser dans un site symétriquement lié, mécanisme nécessaire à la réussite de la méiose.

Cependant, la stérilité hybride induite par ce mécanisme semble plutôt rare et souvent de très faible intensité. En effet, la condition nécessaire à l'observation de gros pics de stérilité est la présence chez l'hybride de deux allèles qui sont tous deux agés dans leur population respectives. Cette condition ne semble arriver que dans des régimes de faible taux de mutation au locus *PRDM9*, et de fort taux de mutation au niveau des

cibles. Cette stérilité marginale et transitoire ne semble pas indiquer un rôle majeur de PRDM9 dans la stérilité hybride.

Comme réalisé dans la partie de calibration empirique du chapitre précédent, nous avons pris en compte le dosage génétique de PRDM9. Celui-ci, de part son effet d'éviction, permet aux hybrides d'avoir des niveaux de stérilité plus importants que sans dosage. Ces baisses de fertilité sont dues à la fois à la plus forte érosion des allèles homozygotes en forte fréquence dans les populations, et au désavantage hétérozygote des hybrides comparés aux parents qui sont souvent homozygotes.

Cependant, le niveau de diversité inra-populationnel prédit par le modèle n'étant toujours pas cohérent avec les observations empiriques, nous avons tout d'abord augmenté le nombre de méioses par individu avant qu'on ne le déclare stérile. Cela a pour effet de donner des niveaux de succès de la méiose chez l'hybride bien plus faibles. Ils sont donc plus cohérents empiriquement, mais les niveaux de fertilité des individus correspondant sont au contraire très élevés. De plus, la Reine Rouge est toujours quasi neutre dans cette situation. Enfin, nous avons augmenté le nombre de DSB par méiose, ce qui ne permet pas d'obtenir des niveaux de stérilité raisonnables, sauf dans des cas de régimes monomorphes dont le taux de mutation des cibles est très élevé.

Discussion

Au vu des résultats présentés ci-dessus, il devient clair que la stérilité hybride observée est plutôt marginale et transitoire, ce qui ne plaide pas pour un rôle majeur de *PRDM9* dans la spéciation. Cependant, plusieurs paramètres peuvent jouer en faveur de la stérilité hybride. En effet, la distribution d'affinité des cibles, la concentration de PRDM9 dans la cellule (qui peut être source de compétition entre cibles), ou la dominance de certains allèles par rapport à d'autres, pourraient impacter directement ou indirectement le taux de liaison symétrique chez l'hybride, ce qui impacterait directement la fertilité.

Il existe par ailleurs un autre phénomène non pris en compte dans le modèle présenté ici : la migration entre les deux populations. La migration pourrait avoir deux conséquences. La première touche les populations elles-mêmes, qui auraient un apport régulier de nouveau matériel génétique provenant de l'autre population. Ce flux de gènes pourrait, une fois la barrière de stérilité hybride dépassée, augmenter la diversité *PRDM9* intra-populationnelle. On pourrait par exemple observer des introgressions adaptatives d'allèles (*PRDM9* seraient sélectionnés positivement du fait d'une potentielle dominance). La deuxième conséquence, qui touche les hybrides, serait que les deux Reines Rouges ne seraient plus totalement indépendantes, et qu'à un certain point, les hybrides n'auraient plus d'érosion différentielle assez forte pour causer de l'asymétrie et de la stérilité hybride.

Axe 2

Alice GENESTIER[1], Laurent DURET[1], Nicolas LARTILLOT[1*]

[1] Universite Lyon 1, CNRS, UMR 5558, Laboratoire de Biometrie et Biologie Evolutive, Villeurbanne, France

Corresponding author :

[*] nicolas.lartillot@univ-lyon1.fr

3.1 Abstract

Hybrid sterility is a form of post-zygotic reproductive isolation occurring when fertile parents of two different populations give birth to a sterile hybrid. This reproductive isolation can ultimately contribute to speciation. The genetic basis of hybrid sterility is still not well understood. In mammals, only one candidate speciation gene has been suggested, *PRDM9*. *PRDM9* is known to have a double molecular function during meiosis. First, the protein encoded by this gene determines the location of hotspots, thanks to its histone tri-methyltransferase activity, which recruit the double strand break machinery. Second, PRDM9 has a role of facilitation of homologous chromosome pairing by its symmetrical binding, i.e. binding at the same position on both homolgues. This last function was discovered during experiments in hybrids from crosses between two subspecies of mice *Mus musculus musculus* and *Mus musculus domesticus* which were sterile and showed a high rate of asymetrical binding of PRDM9. Finally, this double role of PRDM9 during meiosis has been tested analytically and with simulations in a panmictic population However, this dual role in hybrid sterility has not yet been tested by theoretical models. In this work, we have adapted the single population model used in our previous article, in a bi-populational model with generation of hybrids, in order to study the frequency and intensity of hybrid sterility as a function of several molecular parameters. Our model predicts hybrid strility in certain regimes, sterility caused by asymetrical PRDM9 binding due to independent erosion of binding sites in the two sub-populations. This hybrid sterility seems rather marginal and transient which, however, in spite of the limitations of our analysis, does not plead for a major role of PRDM9 in speciation.

3.2 Introduction

Speciation is a concept whose ins and outs are still unclear despite loads of studies. Some aspects of it are nevertheless well known and described in the literature. Indeed, this process of speciation may be due to the isolation of populations in different geographical areas, like allopatric, or within the same geographical area, known as sympatric speciation, with intermediate forms between these two extremes. In all cases, populations evolve in different ways, eventually leading to reproductive isolation (reviewed in Seehausen *et al.* 2014 [144]). The latter may occur either before fertilization, due to differences in behavior, anatomy or geographical area - known as prezygotic isolation - or after fertilization, postzygotic isolation, resulting in non-viable or non-fertile hybrids. In this latter case, the isolation can either be extrinsic, caused, as for prezygotic isolation, by divergent selection and linked to ecological or behavioral factors leading to a decrease in mating success, or intrinsic, caused by independent-environmental genetic incompatibilities.

Hybrid sterility is often seen as a key step in the early phase of speciation, as it can lead to a reinforcement of pre-zygotic isolation mechanisms, so as to avoid the fitness cost then associated with the choice of genetically distant mates. Although extensively studied in various eukaryotes, the genetic and molecular mechanisms of hybrid sterility are still poorly understood. However, some similar characteristics are found across many species. This is particularly true of an observation made by Haldane in 1922 [118], which later became known as Haldane's rule, which states that "When in the F1 offspring of two different animal races one sex is absent, rare, or sterile, that sex is the heterozygous sex (heterogametic sex)" either in males (XY (mammals) or XO-type sex (Drosophila fruit flies)) or in females (ZW (birds) or Z0-type sex (moth)). To date, there are several more or less controversial explanations for this law, depending on the species concerned (dominance theory [120] or Faster male theory [123] (reviewed in [119])).

Regardless of Haldane's rule, a genetic explanation for the hybrid sterility mechanism was given by several scientists independently, and is currently known as the Dobzhansky-Muller model [101; 102; 103; 145; 104]. This model explains hybrid sterility as incompatibilities between genes that have diverged after independent evolution in two distinct subspecies or populations. More precisely, by considering 2 (or more) interacting loci, different alleles have become fixed at each of these loci in each population, giving rise in the hybrid to combinations of alleles never tested by natural selection before in the parental populations, sometimes resulting in hybrid sterility.

Some studies have then gone beyond theoretical models or ideas, by obtaining empirical evidence on the genetic basis of hybrid sterility in model species. In particular, the genes responsible for hybrid sterility, more generally known as speciation genes, are not so easy to detect. Only a small number have been detected in metazoans. In Drosophila, the organism in which most studies have been carried out, a small number of these genes have been detected, including *OdsH* [107; 108], *Ovd* (Overdrive) [109] or *Nup96* [110]. In vertebrates, only one candidate gene has been detected in mice, called *Prdm9*.

This gene, discovered at the locus Hybrid sterility 1 (*Hst1*) [116] and nowadays also

called Meisetz, causes meiosis arrest in F1 hybrid males from crosses between PWD females and C57BL/6J (or simply B6) males [117]. These hybrids show an arrest of meiosis during prophase 1 due to a lack of DSB repair leading to failure of synapsis between homologous chromosomes [75][125]. This hybrid sterility caused by this gene was subsequently found in other mouse crosses [137; 131; 146] and in other species such as cattle breeds [147].

PRDM9, PR domain zinc finger protein 9, encodes a histone methyltransferase enzyme with multiple functions acting during early meiotic prophase. It determines recombination points across the genome by recognition of specific DNA sequences by its zinc finger, then recruits the double-strand break (DSB) machinery by trimethylation of surrounding histones (H3K4me3 [51; 52] and H3K36me3 [36]). These breaks are then repaired with the homologous chromosome, giving rise to crossover (CO) or non-crossover (NCO) events. However, via DSB induced biased gene conversion (dBGC), *PRDM9* is also responsible for the progressive erosion of its binding sites and thus of the recombination landscape, a phenomenon known as the "hotspot conversion paradox" [83]. Moreover, this gene is under strong positive selection and is one of the fastest-evolving genes in mammals. An intra-genomic Red Queen model was then proposed to explain the long-term maintenance of the recombination rate despite the hotspot paradox [85]. This model would be the result of a never-ending race between two antagonistic forces, namely erosion and positive selection, acting on *PRDM9*.

More recently, new theoretical models have been developed to study the Red Queen mechanism mediated by *PRDM9*, through the implementation of *PRDM9*'s second molecular function. This function corresponds to the symmetrical binding, i.e. at the same site on both homologous chromosomes, of the PRDM9 protein, which appears to facilitate the pairing of homologous chromosomes during meiosis. Interestingly, this second molecular function of PRDM9, which was originally discovered during experiments to uncover the cause of the hybrid sterility observed in certain mouse crosses by Davies *et al.* in 2016 [111], has made it possible, through new theoretical models acting in single populations, to better identify the forces driving the Red Queen and their causes. In particular, two parallel models, by Baker *et al.* 2022 [134] and Genestier *et al.* 2023 [133], have uncovered what underpins the positive selection acting on the new *PRDM9* alleles, a force counterbalancing the erosion whose origin had not yet been understood. This selection force comes from the loss of symmetrical binding in old alleles that have eroded most of their high affinity target leading to a greater asymmetry in the genome which decreases their fertility. These models have also uncovered different Red Queen regimes, depending on factors such as the binding affinity of the protein to its targets, and gene dosage. From these models, several questions emerge, including the role of this symmetry in a bi-population context.

In fact, as mentioned above, asymmetrical binding of PRDM9 to its targets seems to have an impact on the hybrid fertility of certain mouse crosses. Generally speaking, in the context of two populations that separated several hundred thousand years ago, as was the case for the *Mus musculus musculus* and *Mus musculus domesticus* subspecies

[112; 113] , the genomes evolve independently. In the context of recombination, *PRDM9* evolved independently in both populations, generating different recombination and target erosion landscapes. This has been observed in different mouse subspecies [124]. When these populations come together, they form a secondary contact zone where they can attempt to reproduce and form hybrids. However, because of the asymmetrical erosion patterns between the two populations, PRDM9 proteins tend to bind to the homologous chromosome where their targets have not been eroded, generating asymmetry. These appear to be the cause of hybrid sterility in several crosses of mice [117],[125], giving some support for seeing *PRDM9* as a speciation gene [75][95]. However, this situation has not yet been modeled theorically. Models could nevertheless help to determine whether this asymmetry is really a flagrant cause of hybrid sterility or whether other factors are necessary, and to determine whether this strength is sufficient for *PRDM9* to be considered a speciation gene.

In order to try to answer this question concerning the impact of the mechanical role of *PRDM9* in hybrid sterility, we adapted our simulator, initially developed in single population [133], in the context of two populations simulating a scenario similar to that which led to the hybrid zone of secondary contact between the two subspecies *Mus m. musculus* and *Mus m. domesticus*.

3.3 Material and methods

3.3.1 The model

In this section, we present the Wright-Fisher model used for the study of the evolutionary dynamics of *PRDM9* in the context of 2 populations as shown on figure 3.1. This model is an extension of the model presented in Genestier *et al.* (2023) [133]. It was adapted so as to account for population subdivision.

The model simulates a scenario similar to what happened for the two sub-species *Mus musculus musculus* and *Mus musculus domesticus*. It is articulated as follows : at the beginning, a panmictic population of size N evolves until reaching its stationnary regime such as presented in the paper of Genestier *et al.* of 2023 [133] . In brief, each individual in the population is diploid and possesses one pair of homologous chromosomes. Each chromosome contains a *PRDM9* locus (colored circle on the schema) where the color correspond to a specific allele, and binding sites (colored rectangles) recognized by each allele (h site per allele). These sites have different binding affinity (y) with their allele (rectangle thickness). At each generation the *PRDM9* locus mutates at rate u and the sites are inactivated at rate v. Then to fill the next generation, meioses are performed from individuals chosen randomly. Each meiosis is characterized by PRDM9 binding to its targets (colored arrows), formation of DSBs (red stars, d DSBs per meiocyte), search for symmetrical binding and CO formation (blue rectangles).

Then, at some time T, to implement the introduction of a geographical barrier between two populations of the same species, the ancestral population is split in two of the

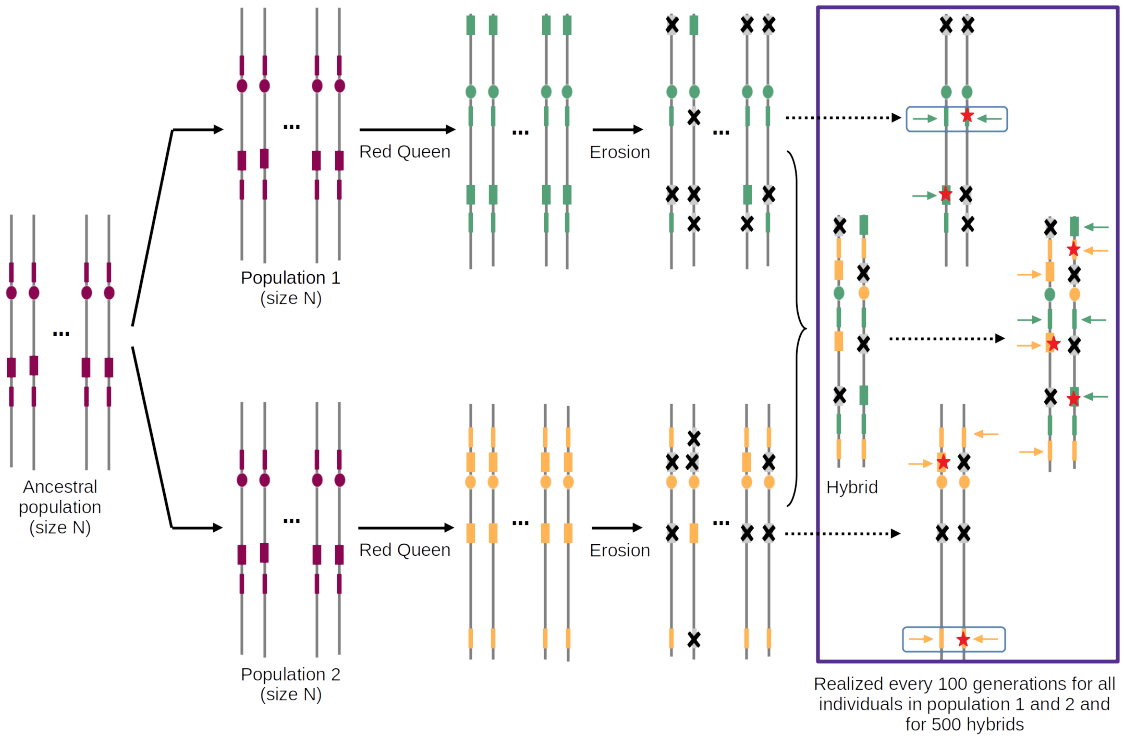

Figure 3.1: Schema of the bi-population model and PRDM9 asymmetrical bindings in hybrids between two populations. *This diagram shows the mechanism of the model. The model assumes a population of N diploid individuals (2N chromosomes, vertical lines). Each chromosome has a PRDM9 locus (filled oval, with a different color for each allele) and, for each PRDM9 allele, a set of target sites (filled rectangles, with color matching their cognate PRDM9 allele) of variable binding affinity (variable width of the filled rectangles). Mutations at the PRDM9 locus create new alleles, while mutations at the target sites inactivate PRDM9 binding (grey sites with a cross). To pass from one generation to another, several individuals are randomly picked and their meiosis is performed as described in the paper of Genestier et al. 2023 [133]. Here, the simulator model a pancmictic ancestral population which evolves to equilibrium and is then separated into two populations of the same size. Each population evolve independently from each other. They both make their own Red Queen by eroding the targets associated with their respective alleles and selectionning new PRDM9 alleles. Several hybrids are regularly generated from the two independent populations. During meiosis, PRDM9 proteins (horizontal arrows which color correspond to the allele producing this protein) bind to their targets according to their affinity. Then a certain number of DSBs is induced at a small number of randomly chosen sites bound by PRDM9 (red stars) and search for DSBs at symmetrically bound sites (symmetrical DSBs). Sequence asymmetries are observed, with an active target (colored rectangle) facing an inactive target (gray rectangle). PRDM9 binding prioritizes high-affinity sites on the chromosome homolog in the hybrid (binding in trans). This leads to a greater number of bonds, particularly asymmetrical ones, in the hybrid than in the parental populations. Double-strand breaks are therefore less likely to occur at a symmetrically bound sites in the hybrid, resulting in lower hybrid fertility than in parental populations.*

same size N. During this step we generate twice as many gametes as in normal time in order to create 2 populations of N individuals each. Then, we let these two populations evolve independently from each other during another 40000 generations.

Some variants of the model are also implemented, differing in the details of the mechanism such as varying gene dosage, increasing the number of meioses per individual or the number of DSB per meiosis.

Creation of hybrids

In order to study the extent and evolution of hybrid sterility between the two populations throughout the simulation, we generate a certain number of hybrids (in this paper 500) every 100 generations. These hybrids are generated as follows : one individual is taken at random in population 1, then its meiosis is attempted (as described in Genestier *et al.* 2023 [133]). If it fails, another individual is randomly picked in the population and its meiosis is attempted. This is repeated until the meiosis succeeds. Then, one of the four generated gametes is randomly picked and set aside. Then, a second chromosome is generated in the same way but with one or several individuals sampled from population 2. A hybrid genotype is created with these two chromosomes which in turn attempts to realize a meiosis. All these steps are done for a total of the 500 hybrids made and the success rate of meiosis and thus the mean fertility is computed over all the hybrids. These hybrids are not reintroduced in neither of the two populations. This then allows the comparison of descriptive statistics of this model for hybrids and sub-populations, just after splitting into two populations, as well as over the long term.

3.3.2 Summary statistics

Several summary statistics were monitored every 100 generations during the run of the simulation program. They were averaged over the entire run, thus giving a quantitative characterization of the equilibrium regime, as a function of the model parameters.

The main statistics are the following : first, the *PRDM9* diversity in the population (D), the mean haplo-insufficiency of young alleles (σ_0) and of average alleles in the population ($\sigma_{\bar{z}}$) and the scaled selection coefficient ($4Ns_0$). Second, for each allele, the mean erosion of its target sites ($\bar{z}$), the mean fertility (w).

Diversity (D)

The diversity is defined as the effective number of *PRDM9* alleles, or in other words as the inverse of the homozygosity [96]:

$$D = \frac{1}{\sum_i f_i^2} \tag{3.1}$$

where $f_{i,t}$ is the frequency of each allele i.

With this definition, when K alleles segregate each at frequency $\frac{1}{K}$ in the population, the diversity D is equal to K.

Mean cumulative erosion ($\bar{z}$)

The cumulated erosion of an allele of age t is defined as:

$$z(t) \approx \rho \int_0^t f(X)\, dX \tag{3.2}$$

where $f(X)$ is the frequency of the allele at time X and $\rho = \frac{Nvd}{2h} \approx \frac{Nvg}{2h}$, with g as the gene conversion rate, is the net rate of erosion per generation.

Mean fertility (w)

For a given genotype, the fertility can be computed analytically. Here we assumed that fertility in proportional to the rate of success of meiosis. Thus, it is equivalent to 1-(the mean probability of failure of meiosis) characterized by the absence of a DSB in a symmetric site. In turn, the number of DSBs in a symmetric site follows a Poisson law with parameter dq where d is the average number of DSBs per meiocyte and q is the mean probability of symmetrical binding for this allele. Thus, the mean fertility of an allele can be expressed as follows :

$$w = 1 - e^{-dq} \tag{3.3}$$

For a given allele, and at a given time, this statistic was then averaged over all diploid complete genotypes carrying this allele present in the population at a given generation (and then averaged over all generations of the run). This statistics is also computed for each hybrid and averaged over all the hybrids too (periodically and then over the whole run). Note that if the meiosis success rate is decoupled from fertility (several meioses allowed per individual), the mean fertility of an allele is

$$w_{meiosis\_number} = 1 - (1 - w)^{meiosis\_number} \tag{3.4}$$

Mean Sigma (σ_0 and $\sigma_{\bar{z}}$)

In presence of gene dosage, the homozygotes have a higher concentration of *PRDM9* protein associated to the allele than in heterozygotes. This lead to an increase in the fertility in homozygotes compared to heterozygotes. This difference of fertility, also defined as the haplo-insufficiency, can be represented by the σ statistic : σ_0 for young alleles and $\sigma_{\bar{z}}$ for an average allele in the population. In both cases, σ is defined as the relative difference in fitness between homozygote and heterozygote (relative to heterozygote).

It can be expressed as follows:

$$\sigma_0 = \frac{w^{hom}(0) - w^{het}(0,0)}{w^{het}(0,0)} \tag{3.5}$$

and

$$\sigma_{\bar{z}} = \frac{w^{hom}(\bar{z}) - w^{hemi}(0,\bar{z})}{w^{het}(0,\bar{z})} \tag{3.6}$$

Scaled selection coefficient ($4Ns_0$)

A key quantity which is particularly informative about the exact selective regime induced on *PRDM9* by the overall evolutionnary dynamic is the mean selection acting on new alleles entering the population. This statistics is computed as $4Ns_0$ with

$$s_0 = ln(w^*) - ln(\bar{w}) \tag{3.7}$$

where w^* is the mean fertility of the allele at time t and

$$\bar{w} = \sum_i f_i w_i^* \tag{3.8}$$

is the mean fertility of all *PRDM9* alleles in the population.

In particular, if $4Ns_0 >> 1$ the system is under positive selection, else if $4Ns_0 << -1$ the system is under negative selection and if $|4Ns_0| << 1$ the system is neutral.

3.3.3 Scaling experiments

In order to study how the overall Red Queen dynamics and hybrid sterility vary according to the model parameters, a central configuration of parameters is determined and will be fixed for all simulations of the same scaling experiments. First, the population size N is set to $1,000$. Then, the number of binding sites per allele was set to $h = 800$ which roughy correspond to the number of sites per allele on the smallest chromosome in mice. The parameter d representing the average number of DSBs in a meiocyte is set at $d = 8$ to approximate the average number of DSBs on the smallest chromosome in mice. In one scaling experiment, where we look at the impact of an increased number of DSB on hybrid sterility, the number of DSB is set to $d = 24$. Then two parameters at a time systematically vary over a range of regularly spaced values on both sides of the reference configuration. In the case of one dimensional scaling, only one parameter varies at a time. This parameter is fixed along the entire simulation and is variable only between simulations. For each parameter value over this range, a complete simulation is run. For the bidimentional scaling, the central parameter values are set to empirically motivated values, already used in previous studies [96][133] , i.e. the *PRDM9* locus mutation rate is set to $u = 3 \times 10^{-6}$ and the target mutation rate is set to $v = 10^{-7}$ (which are then rescaled in order to match $N = 1000$ in our simulations, to $u = 3 \times 10^{-4}$ and $v = 10^{-5}$). Once the equilibrium is reached for a given simulation setting, the summary statistics described above are averaged over the entire trajectory, excluding the burnin. These mean equilibrium values, which characterize and quantify the stationary regime of the model, were finally plotted as a function of the varying parameter(s).

Parameters	Description	Value
u	Mutation rate at the *PRDM9* locus	2×10^{-6} to 5×10^{-3}
v	Mutation rate at the targets	2×10^{-6} to 5×10^{-3}
N	Population size	$5,000$
h	Number of targets recognised by a new allele	400
d	Mean number of DSB per chromosome pair per meiocyte	6 to 24
y_i	Affinity of target i	variable
g	Gene conversion rate	variable
n_{mei}	Number of meiosis allowed per individual before reproduction failure	1 to 5

Variables	Description
f_i	Frequency of allele i [1]
θ_i	Proportion of target sites still active for allele i [1]
z_i	Level of erosion of allele i [1]
w_i	Fertility of allele i [1]
q_i	Probability of symmetrical binding of allele i [1]
ρ	Net rate of erosion per generation
D	Diversity D of *PRDM9* in the population at equilibrium
σ_0	Relative difference in fertility between homozygotes and heterozygotes for young alleles
$\sigma_{\bar{z}}$	Relative difference in fertility between homozygotes and hemizygotes for alleles of level of erosion $\bar{z}$
$4Ns_0$	Scaled selection coefficient

Table 3.1: **Descritption of input parameters and output variables.** *These variables also change over time*

3.4 Results

In our previous work [133], we developed a simulation program for investigating the evolutionary dynamics of *PRDM9*-dependent recombination in a single randomly mating population. Here we use this model to study the impact of the molecular mechanisms of *PRDM9*-driven recombination in a hybrid context. The model in a single population showed that the PRDM9 symmetrical binding is sufficient to run the Red-Queen process and that the loss of fertility is due to the loss of symmetrical binding along time due to hostpots erosion by target mutation and biased gene conversion. Also, two types of regime where observed, one characterized by few alleles segregating at a time in the population called monomorphic regime, and one with several alleles segregating together, called polymorphic regime. Moreover, by implementing the gene dosage, we discovered that it was a force acting against diversity. Finally, the empirical calibrations carried out showed that the regime of Red Queen strongly depends on the different parameters of the model and that this model still lacked theoretical exploration and knowledge on the molecular mechanisms before being able to provide empirically realistic results for all summary statistics.

This model is here adapted so as to consider two populations with the generation of hybrids (see methods). The results section is thus divided in two main parts. First, the

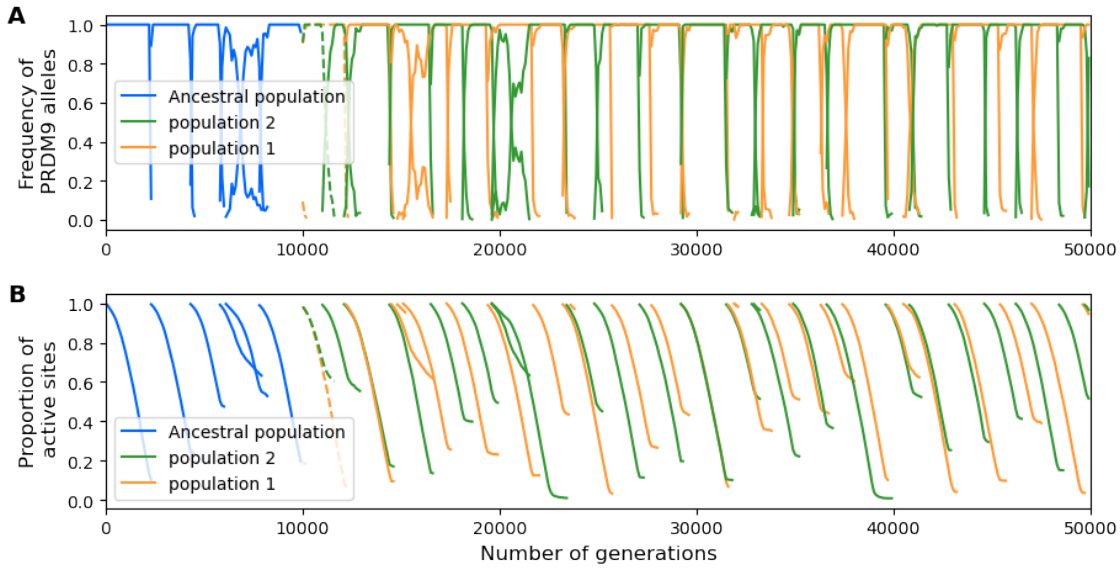

Figure 3.2: Evolution of frequency (top) and proportion of active sites (bottom) for each allele of the bi-population model as a function of time (number of generations) for a monomorphic regime *($u = 3 \times 10^{-8}$* ***and*** *$v = 10^{-6}$**);** *(A) Allele frequency: several Red Queen cycles are observed for the initial population, and after 10,000 generations (steady state of the panmictic model), we separate into two subpopulations. Independent evolution (desynchronized Red Queen) is observed, due to the lack of genetic exchange between these two populations; (B) Allele activity: decreasing activity until allele disappearance, independent Red Queen.*

results for two simple simulations, one in a monomorphic regime and one in a polymorphic regime, are presented and we analyse the dynamics of fertility both in parental population and in hybrids. Second, we explore the equilibrium regime under different conditions by varying the mutation rates u and v and by allowing gene dosage, for a higher DSB rate or by decoupling fertility and meiosis success thanks to the increase in the number of meiosis allowed for each individual, as already tested in our previous paper.

3.4.1 Hybrid sterility

First, the model was run with parameters of mutations rate at the *PRDM9* locus $u = 3 \times 10^{-8}$ and at the target site $v = 10^{-6}$ (cf figure 3.2).

On this graph, we can see a monomorphic regime (only one allele at a time in the population) the Red Queen dynamics of the ancestral population (in blue from 0 to 10,000 generations), then the separation into two distinct populations evolving independently of each other between 10,000 and 50,000 generations (population 1 in orange and population 2 in green). Independent envolution in the two populations is shown by the rapid desynchronisation of the dynamic of allelic trunover at the *PRDM9* locus.

Then, in order to study hybrid sterility, we examined what would happen if hybrids were made between the two subspecies, at different times since the separation (every 100 generations). At each time, the fertility of hybrids and of individuals in each subpopulation is computed by just attempting a meiosis for a sample of 500 hybrids creating

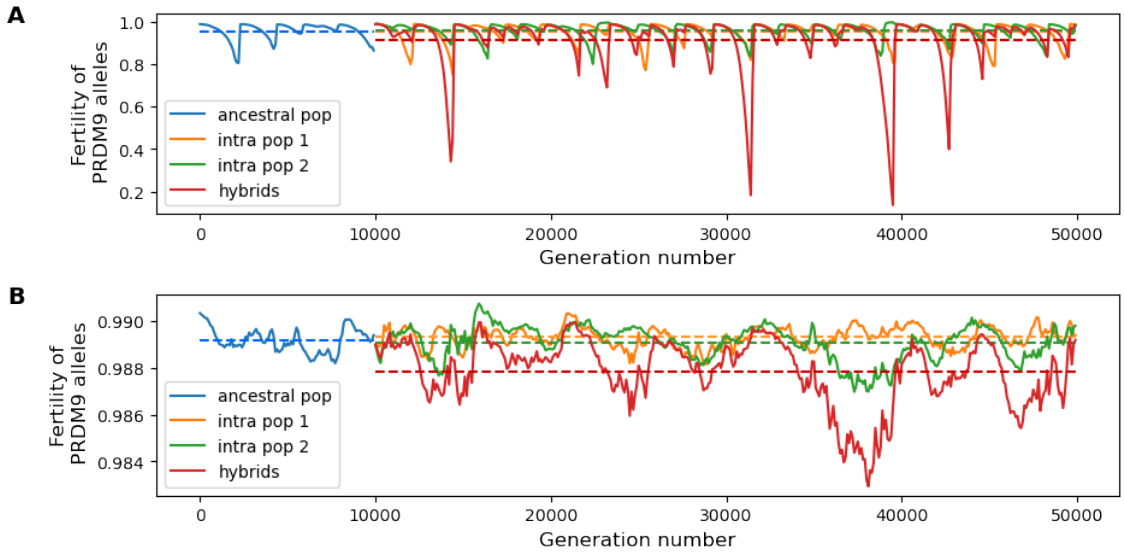

Figure 3.3: Evolution of mean fertility in populations and hybrids as a function of time in (A) a monomorphic regime *($u = 3 \times 10^{-8}$,* ***$v = 10^{-6}$)*** ***and in (B) a polymorphic regime*** *($u = 3 \times 10^{-6}$,* ***$v = 10^{-7}$)****. We observe several cycles of Queen Red for the initial population (blue), then at equilibrium, we see the separation into 2 sub-populations (green and orange). Red corresponds to the hybrid case. The curves for hybrids are relatively close to the curves for the parent populations, even though the probabilities for hybrids are lower than for intra-populations. In graph (A), we can see that fertility drops significantly in hybrids, which are therefore mostly sterile.*

by randomly sampling parents from the two populations and then measuring the rate of success of meiosis over them. Figure 3.3A and B show the evolution through time of the mean fertility in the ancestral population (blue), sub-population 1 (orange), sub-popualtion 2 (green) and in hybrids (red).

First, we focus on the figure 3.3A, representing the evolution of fertility in the simulation shown in figure 3.2. We notice the presence of oscillations of the mean fertility for both populations and for the hybrids. Those observed in the populations (blue curve for ancestral population and orange and green curves for sub-populations) are due to the intra-genomic Red Queen operating independently in each of them. In brief, current dominant alleles erode their targets. This generates a loss of symmetrical binding and thus a decrease in fertility. This is the reason why the old alleles are then being replaced by new alleles.

Next, we focus on the fertility dynamic of hybrids. In general the mean fertility of hybrids is very close to the mean fertility calculated in the two parental populations and is thus relatively high. In this case, hybrids are fertile. However, hybrids sometimes show much lower fertility than intra-population individuals. These sharp drops in relative fertility illustrate hybrid sterility events such as those observed empirically in *Mus musculus*. Note that this may correspond to two cases. A first case is where the hybrid receive the same allele from both of its parents, but then the corresponding sites on each chromosome are eroded differently. A second case is where the hybrid receives two different alleles that have each eroded their own target sites on their own chromosome.

The first case is only possible shortly after the separation of the two populations, because the dominant allele at that precise moment will differentially erode its targets in the two independent populations, whereas, later on, the new *PRDM9* alleles that appear in each of the two populations will necessarily be different. In this last case (different alleles in the two sub-populations), there are only low affinity targets sites left for each allele on its associated chromosome (in *cis*). However, on the homologous chromosome (in *trans*), the high affinity target sites attributed to this allele have not experienced erosion in their population of origin and are therefore still active. A large fraction of PRDM9 molecules will bind to those asymmetric sites and will undergo DSBs. As a result, a large proportion of DSBs will occur on asymmetrical sites. This leads to a decrease in their fertility. These results thus recapitulate the various aspects of *PRDM9*-induced hybrid sterility such as observed in the experiments of Baker *et al.* in 2015 [124], Gregorova *et al.* in 2018 [125] and Davies *et al.* in 2016 [111].

We also notice some cases where, in spite of a low fertility in one of the parental populations, the hybrid itself shows a high fertility. This is especially the case when the hybrid benefits from an allele that has recently appeared in the other parental population. We can therefore deduce that the age of the alleles received by the hybrid has an influence on its fertility.

Altogether, the necessary condition for hybrid sterility to be observed is that the two alleles should have both eroded a large fraction of their (high affinity) target site in their respective population. As a consequence, under this specific succession regime, hybrid sterility is transient but sometimes very strong, leading to an overall mean fertility lower in hybrids than in intra-populations.

If the mutation parameter at *PRDM9* locus u increases, the regime progressively changes from monomorphic (a single allele in the population at a given time) to polymorphic (several alleles segregating together in the population). Note that, in this new regime, the mean erosion level and the mean intra-population fertility are not the same as in the previously shown monomorphic regime. In this regime shown in figure 3.3B where $u = 3 \times 10^{-6}$ and $v = 10^{-7}$, alleles generally erode much less of their targets on average than in the previous regime, and the associated fertility drops are therefore smaller. As a result, hybrids still have, on average, a lower fertility than intra-populations, but the difference is much smaller than in previous regime. Furthermore, here, the average drop in hybrid fertility is linked to certain combinations of alleles that have both eroded their targets in the parental populations, generating in this particular case a greater drop in hybrid fertility than in the other, more numerous cases that do not specifically result in hybrid sterility. These cases may be more frequent than in the previous regime, but of lower intensity, resulting in a more diffuse drop in hybrid fertility throughout the simulation, unlike the sterility peaks observed in the monomorphic regime (Figure 3).

Finally, in order to measure the impact of hybrid sterility, a histogram of the distribution of average fertility was drawn in figure 3.4 for hybrids and intra-populations, observed throughout the simulation after divergence. The tail of the distribution is wider on the left for the hybrid, showing that hybrids are more likely to have a lower fertility

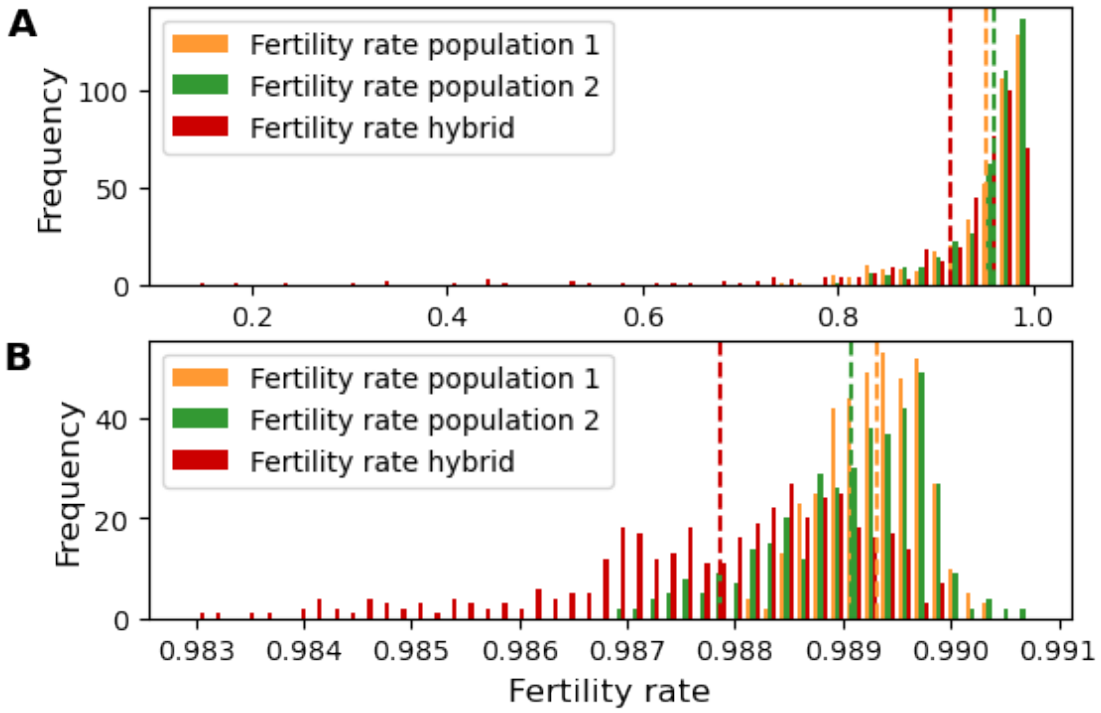

Figure 3.4: Histogram of the frequency distribution of hybrids and intrapopulation fertility rates (A) under a monomorphic regime *($u = 3 \times 10^{-8}$,* ***$v = 10^{-6}$)*** ***and (B) under a polymorphic regime*** *($u = 3 \times 10^{-6}$,* ***$v = 10^{-7}$)****. Note that the scale of the x-axis is not the same between the two histograms, to see what happens in both cases. The distribution of intra-population fertility rates is shifted to the right compared with the probabilities in the hybrid (which has a long tail on the left). Hybrids are therefore more likely to have low fertility events than intrapopulations, and in a more pronounced way. The vertical dotted bars correspond to the averages: the intra-population averages are very close to each other, as the system is in equilibrium, while the hybrid average is shifted to the left, showing that, in general, hybrids have lower fertility than intra-population individuals.*

than intra-population individuals. But, as we saw earlier, there are rare cases where individuals within a population have much lower than average fertility. This is why we can also observe a left-hand tail for both populations, but less extensive than for the hybrid.

To clearly identify the difference between the mean intrapopulation fertility and the mean hybrid fertility, these means were plotted on the histogram. The averages for the two sub-populations are very close to each other, if not superimposed, which shows that the system is at equilibrium. On the other hand, the hybrid average is shifted to the left. This difference between the hybrid and intra-population averages is 0.039 in this monomorphic regime (a 4.1% drop for the hybrid compared with the intra-population, figure 3.4A).

Note that under polymorphic conditions, the difference is 0.001 (or 0.1% drop) (cf figure 3.4B, be careful, the x axis scale are not the same between figure A and B). These differences are small, probably not significant enough, especially for fertility to induce a speciation barrier.

In summary, *PRDM9*-induced hybrid sterility is indeed observed with our model, in

both regimes but at different intensities and over longer or shorter periods. Indeed, in a monmorphic regime with high erosion rate, we observe high hybrid sterility pics along the simulation but these pics are really rare and do not last more than several hundreds of generations. On the contrary, in a case of a polymorphic regime with a lower erosion rate, the hybrid sterility is less marked but is present for several thousands of generations. We also discovered two ways to obtain this sterility depending on whether the observed sterility peak is associated with an ancestral allele or two different alleles having eroded their targets in their respective genome. This second cause of sterility corresponds to what has been observed empirically by several studies [111][36][137].

3.4.2 Two dimensional scaling

In the previous section, only two examples were shown in order to explain the phenomena behind hybrid sterility. This already shows that the regime and therefore the consequences for hybrid sterility vary according to the u and v parameters. To get a more comprehensive picture of the evolution of hybrid sterility as a function of these two mutational parameters, we carried out bi-dimensional scaling experiments (shown in Figure 3.5). Supplementary figures 6.1, 6.2 and 6.3 show additional statistics monitored during those simulations.

Without gene dosage

We first consider the simplest model for which we observed levels of diversity and activity in both populations similar to those already found in our first article (Supplementary Figure 6.1). Thus, the diversity increases when u increases and the activity increases when v increases or u decreases (Supplementary Figure 6.2). Average fertility, the equivalent of meiosis success rate here, in both populations and in hybrids, is shown in Figure 3.5 A,B and C. It can be seen that when u is high or v is low, intra-population fertility is very close to 1 (0.99) and there is no significant drop in fertility in hybrids on average. This is in particular the case for the central configuration corresponding to empirical estimated parameters in the mouse ($u = 3 \times 10^{-6}$ and $v = 10^{-7}$ which, after rescaling for $N = 1000$ gives $u = 3 \times 10^{-4}$ and $v = 10^{-5}$). There are occasional very small drops in hybrid fertility, as shown in figure 3.3 B.

It's only when u becomes low enough ($u < 3 \times 10^{-6}$) and v high enough ($v > 10^{-7}$) that we see significant differences in fertility between intra-population and hybrids. These differences are more pronounced in these cases, as the low mutation rate at the *PRDM9* locus (u) enables alleles to remain in the population for longer periods, and therefore to erode more of their targets, coupled with a high target mutation rate (v), resulting in stronger and faster erosion. If erosion is strong and more regular within a population, the drop in average hybrid fertility will be more pronounced. Thus, in this case, the hybrid fertility drops are mostly determined by the mean erosion level in the parental populations. An overall observation about hybrid sterility here is that it

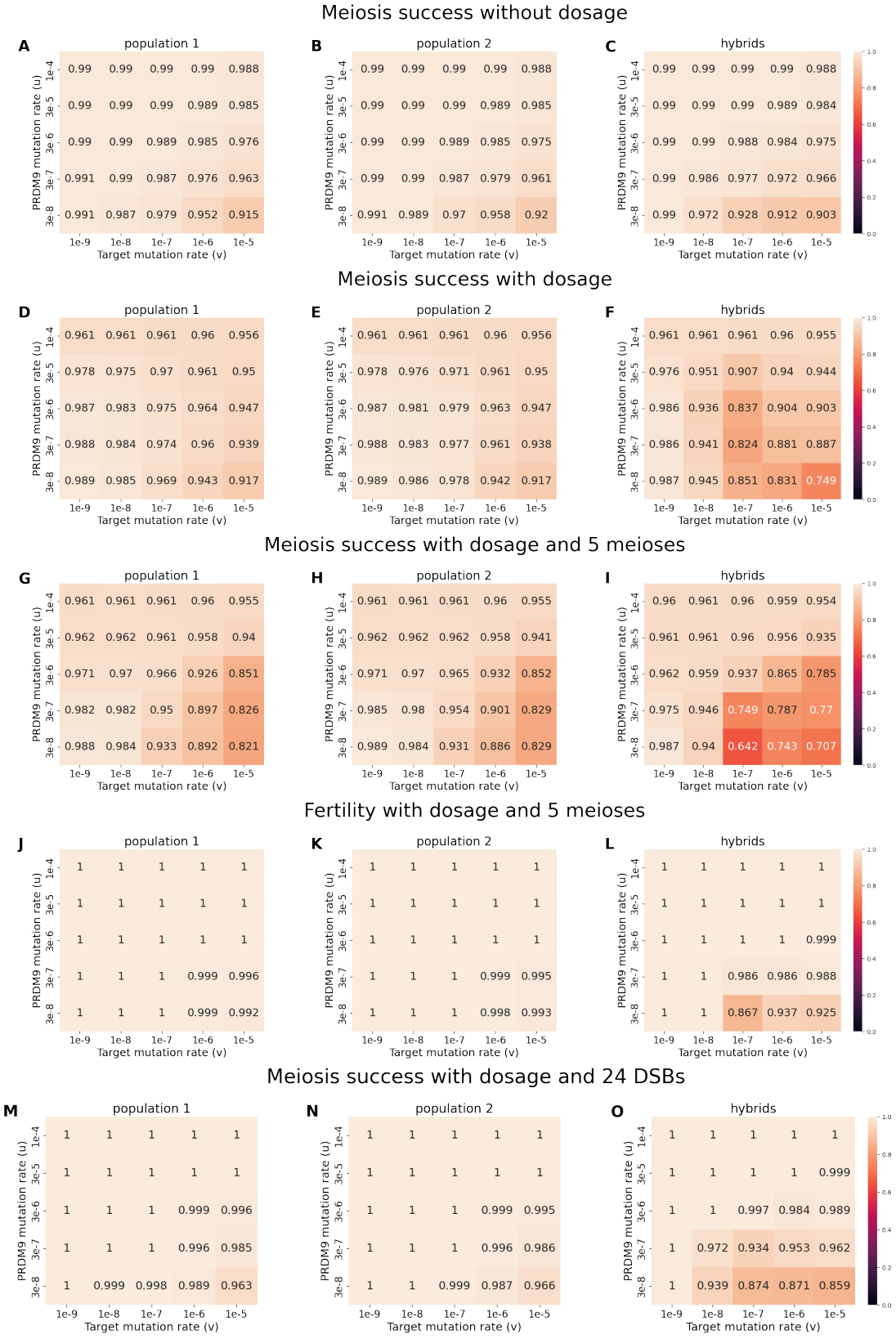

Figure 3.5: Bi-dimentional scaling of meiosis success rate and fertility rate in function of the mutation rate at PRDM9 locus *(u)* ***and mutations rate at target sites*** *(v)* ***in both populations (left and middle heatmpas) and hybrids (right heatmaps)*** *for each case : (A, B, C) without dosage, (D, E, F) with dosage, (G, H, I) with dosage and 5 meiosis, (J, K, L) fertility rate with dosage and 5 meiosis and (M, N, O) meiosis success rate with dosage and 24 DSBs.*

is mostly marginal over a broad range of regimes.

Effect of gene dosage

The case we've just presented is a fairly basic one, enabling us to identify the mechanisms involved in hybrid sterility and their consequences as a function of the parameters of the model. However, the model does not take into account other empirical features, such as gene dosage, which we already implemented in our previous uni-populational model [133]. To compensate for this, a second scaling experiment was carried out in the same way as above, but this time taking into account *PRDM9* gene dosage. Taking into account PRDM9 gene dosage means that in homozygotes, the PRDM9 concentrtion associated to this allele is twice the concentration of the same allele in a heterozygote. As a consequence, this increases the binding probability and thus the fertility of homozygotes compared to heterozygotes.

In our previous study in a panmictic context, as well as in the two parental populations in the present simulation (cf Supplementary Figure 6.1C and D), this add-on created in certain cases an eviction regime where one allele dominates the population due to its homozygous advantage compared to new alleles in heterozyous states that were negatively selected. We recover this observation here.

Most simulations are under an eviction regime. This can be illustrated first, by the diversity, close to 1 (cf Supplementary Figure 6.1C and D), and second, by the scaled selection coefficient $4Ns_0$, which indicates regimes under strong selection, either positive if $4Ns_0 >> 1$, or negative if $4Ns_0 << -1$ (cf Supplementary Figure 6.3C and D). We can see that, except from simulations with high erosion rate, all other simulations have a mean $4Ns_0 << -1$ showing a strong negative selection and thus an eviction regime. Additionnally, we show the percentage of time along the whole simulation when the regime is under negative selection (cf Supplementary Figure 6.4). We can then witness the generalised character of this eviction regime, except form the upper right simulations.

Note that, the bi-dmentional scaling for mean $4Ns_0$ at equilibrum shows some exceptional regimes when $v \geq 10^{-5}$. First, the simulations with low u and high v shows high mean selection but an overall negative selection around 60% of the simulation. This is due to rare very strong peaks of positive selection and the rest of the time the regime is under negative selection. Second, the simulations with high u and v display only positive selection all along the simulation due to a high enough polymorphism and erosion mimicking a regime without gene dosage.

Figure 3.5 D, E and F show that with gene dosage, intra-population and hybrid fertility decreases are more widespread than before. If we first look at the central configuration, we see a drop of fertility about 17% in hybrids compared to 2% in parental populations as opposed to 0.1% without gene dosage under the same parameter configuration. Indeed, the eviction regime allows an allele to segregate longer in the population due to its homozygous advantage, which helps counter the force of erosion. In this way, the allele can continue to segregate at high frequency, with fertility close to 1, despite

strong erosion. However, in the hybrid, this homozygous advantage generates two causes of sterility. First, if the hybrid receive young alleles already at homozygous state in the parental populations, even if they don't have eroded much of their targets, the allele fertility will be higher in homozygotes found in parental populations than in heterozygous hybrids. This is caused by the homozygous advantage that gives a higher fertility to the parents compared to the heterozygous hybrids. Second, if the hybrid receive old alleles from their parents, strong erosion in each of the parental populations generates a strong asymmetry implying a sharp drop in fertility in hybrids. Also, the effects of hybrid sterility are stronger in lower right (u low and v high) because of these mutual effects of asymmetry and dosage.

In summary, the eviction regime seen in most parameter configurations increases intra-population and hybrid sterility, except when v is too low, where there is too few erosion for a drop in fertility to be visible either within the population or in the hybrid. However, this drop of fertility observed in hybrid is in fact mostly due to the disadvantage induced by the heterozygosity at the *PRDM9* locus. This phenomenon, combined with the accentuated high parental erosion of old alleles (Supplementary Figure 6.2C and D) involving greater asymmetry in hybrids, generates higher sterility levels in hybrids than in parental populations.

Decoupling fertility and meiosis success

The previous model with gene dosage is more realistic that the previous version without dosage. However, because of eviction, it products a low diversity at odds with empirical observations. To addres this point, in Genestier *et al.* 2023 [133], we explored two elaborations in this model. First, we decided to take into account the fact that, in reality, gamete are often non limiting, by allowing an individual to perform up to 5 meioses before being declared sterile. In this way, fertility and rate of success of meiosis are not anymore proportional. Thus, we can observe what happens both in terms of meiosis success and in terms of fertility rate.

This decoupling restores reasonable levels of diversity (cf Supplementary Figure 6.1E, and F) and of erosion of the targets (Supplementary Figure 6.2E and F). Also, as shown in figure 3.5 G, H and I, the intra-population and hybrid meiosis failure rates seem to be accentuated for $v \geq 10^{-7}$ and for $u \leq 3.10^{-6}$, where the meiosis failure rate is between 20 and 40%. However, by checking the corresponding fertility rates (3.5 J, K and L) there is barely any decrease in fertility. Note that the central calibration shows only a slight decrease in hybrids meiosis success.

The meiosis success rate does not represent the average fertility in the population, as individuals have the possibility of achieving up to 5 meiosis before being declared sterile. In other words, a drop in the meiosis success rate will have little impact on the individual's fertility. Thus, the observed meiosis success rate in intra-population illustrates the ability of individuals in the parent populations to survive even with a rate of meiosis of the order of 60%. This has the effect of counteracting eviction by restoring the diversity of

PRDM9 but also the individuals can from now on survive with higher rates of erosion which generates a stronger asymmetry in the hybrid and therefore a lower meiosis success.

As shown in our previous study [133], this version of the model implies a neutrality of certain regimes ($4Ns_0$ close to 0 Supplementary Figure 6.3E and F during the major part of the simulation Supplementary Figure 6.5). This explains why there is no decrease in fertility in parental populations. However, in the hybrids of the simulations with low u and high v, fertility is lower than 1. In these cases, there is once again a strong erosion in both parent populations, where alleles segregate without selection or fertility constraints until they have eroded too many targets which leads to strong asymmetry in the hybrids, resulting in a very low meiosis success rate. This strong erosion has also an impact in the parental population where the meiosis success is too low and generates loss of fitness (Supplementary Figure 6.6) allowing for new alleles to be under strong positive selection.

In summary, decoupling the meiosis success rate and the fertility restores reasonable levels of diversity and erosion. However, due to the weak response of fertility to a lower success rate of meiosis, there is less observed hybrid sterility than in the case examined above (Figure 3.5D, E and F) in which fertility is proportional to meiosis success rate. This is not what could correctly explain the empirically observed prevalence of hybrid sterility.

Increasing the DSB rate

Another way in which the model might be naive is that it assumes that a fixed number of DSBs are produced first and the resulting single-stranded DNA (ssDNA) engage in the search for homology. In reality, the DSBs are induced progressively along the genome. If a ssDNA find its homologue, then the generation of new DSBs is stopped and meiosis can continue. On the contrary, if no ssDNA finds its homologue other DSBs are generated until one finds its homologue or the meiosis turns to apoptosis after a certain time. In this context, more DSBs can be generated on average than the number observed in successful meioses. Our current model does not allow the progressive generation of a different number of DSBs. So, by simplification, instead of producing only 8 DSBs per meiosis, we increase this number to 24 DSB. This number is fixed and it could be seen as the upper limit of the number of DSBs that can be realized in a meiosis before apoptosis, but in reality, on average, far fewer DSBs are observed, in the order of 10.

On the one hand, this increase in DSB number seems to restaure diversity (Supplementary Figure 6.1G and H) and to increase the mean erosion level for simulations under low u (Supplementary Figure 6.2). On the other hand, we can see on figure 3.5M, N and O that individuals, whether hybrids or not, have a high meiosis success rate, except when u is low, where hybrid sterility is present but does not exceed 15%. Also, in these cases, the regime is under positive selection (as shown on Supplementary Figure 6.3G and H) from 20 to 100% of the time depending on the simulation. The rest of the time, these regimes are under neutral selection. Note that the central simulation does not show hybrid sterility. Figure 3.6 gives an example of the evolution of intra-population

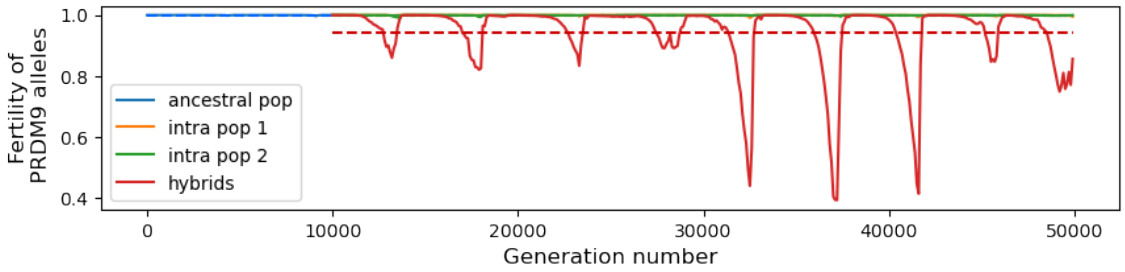

Figure 3.6: Evolution of mean fertility in populations and hybrids as a function of time in a monomorphic regime with dosage and 24 DSBs *($u = 3\times10^{-7}$, $v = 10^{-7}$). No significant drop in fertility is observed throughout the simulation for ancestral (blue) and parent (orange and green) populations. However, very sharp declines in fertility are observed in hybrids (red).*

and hybrid fertility throughout the simulation when $u = 3 \times 10^{-7}$ and $v = 10^{-7}$, in the case of a slight positive selection regime.

In this simulation, while hybrids have sharp drops in fertility, their parent are completely fertile. This phenomenon is what is observed empirically. However, this phenomenon is accompanied by a severe erosion without this having a major impact on intra-populational fertility. This strong erosion is the consequence, on the one hand, of the homozygous advantage strongly present in this homozygous regime (weak u and moderate v) and, on the other hand, of the increase in the number of DSB, reducing the loss of fertility, which also explains the quasi neutrality regime oberved under this model. In the hybrid, on the other hand, we find the mutual disadvantages of dosage and asymetrical erosion. Moreover, the erosion being quite severe, the number of asymetric sites is substantial and the increase from 8 to 24 DSB does not allow to counteract this asymmetry, implying hybrid sterility.

3.5 Discussion

The *PRDM9* gene was initially discovered as a gene inducing hybrid sterility and several studies carried out in several mouse crosses [111; 137; 125; 131; 146] gave convincing results in this direction. For this reason, the gene has been proposed as a speciation-inducing gene, the only candidate to date in vertebrates [75]. Interestingly, it is by discovering its role in hybrid sterility that the mechanism of action of *PRDM9* during meiosis was understood. Indeed, the hypothesis of the symmetrical binding of PRDM9 being essential for the success of meiosis was proposed by Davies *et al.* in 2016 [111] in this context. In a second step, models taking into account this new mechanism have been developed in the case of a single panmictic population [134; 133]. However, this model with the molecular mechanism of PRDM9 had not yet been tested in a multi-population context with hybrid formation, which was finally realized the present article.

3.5.1 Observation of marginal and transient hybrid sterility

In accordance with the status of *PRDM9* as a gene inducing hybrid sterility, the structured population model effectively allowed the identification of a decline in fertility in the hybrid compared to that of the two populations. However, these cases of hybrid sterility appear to be rather marginal and of relatively low intensity, occurring only in fairly precise configurations characterized by old alleles being present in each of the two parental populations. Thus, altogether, the model, in its current state, does not show an important role of the *PRDM9* role in speciation. If this had been the case, we would have more often found fairly significant fertility decreases in hybrids, which would then possibly induce a reproductive barrier that can lead to speciation.

However, despite the occasional presence of hybrid sterility in empirical studies [137], some studies show that this sterility does not appear as rare and as low as what is predicted by our model [146][131]. However, one might think that the cases of hybrid sterility discovered by other authors could correspond to the rare cases highlighted in my model (specific association of alleles causing meiosis to fail) [111][130]. These observations suggest that hybrid sterility is highly dependent on hybrid *PRDM9* alleles and erosion background in the genome of different populations.

Moreover, according to the model presented here, the phenomenon of hybrid sterility is only transient. Indeed, once a new allele appears in either population, as it has never eroded in any of the populations, the hybrid possessing this allele will benefit from an increase in symmetrical binding for this *PRDM9* allele, which will compensate for the loss of symmetrical binding induced by the other old allele. Therefore, if the *PRDM9* gene is related to speciation, it can only be at very particular times and associated with other speciation forces. The hypothesis of transient hybrid sterility was previously suggested by other authors [111]. My work here gives further support to this hypothesis.

Finally, these phenomena of hybrid sterility have only been observed, for the moment in our simulations, in a succession regime. This is due to the fact that it is simpler to have a strong differential erosion of targets leading to hybrid sterility in succession. However, this does not mean that there is no hybrid sterility in the polymorphic regime but rather that it would possibly be rarer and weaker than in the succession regime. Recent studies [146][131] show that in wild mice populations with high *PRDM9* diversity, hybrid sterility is not so prevalent, which is consistent with our results.

In accordance with what has been tested in panmictic population in the paper of Genestier *et al.* [133], we first decoupled the rate of success of meiosis and the fertility of individuals. This allowed the observation of a sharp decrease in the rate of meiosis success, up to 40%, however not impacting the individual's fertility. These results match with the observations of AbuAlia *et al.* 2023 [146] which find fertile hybrids showing a quite high percentage of asynapsis. In a second experiment, we increased the number of DSB per meiocyte. Simulations continue to give decrease in hybrid fertility only in monomorphic regimes with strong erosion. However, unlike other simulations showing a decrease in hybrid fertility associated with a loss of fertility in the parent populations (low

but present sterility), in the case of an increase in the number of meiosis per individual or the number of DSB per meiocyte, there is sometimes no decrease in fertility in the parents when the hybrid shows a sterile phenotype. This characteristic is consistent with empirical observations showing total fertility in parents and sterility, sometimes quite high, in hybrids [111; 36].

Finally, in general terms, the central simulation of scaling experiments, corresponding to the empirical calibration simulation, gives in all cases fairly high average levels of fertility in both parent and hybrid populations, except for the regime with eviction in the presence of dosage which does not seem empirically coherent. However, it is important to note that the model used here, as shown in the panmictic population study [133], is not quite satisfactory in terms of its empirical predictions at the moment, due to a lack of consideration of other aspects of meiosis and reproductive biology, which are now discussed.

3.5.2 Perspectives for empirical relevance

Our current model, although taking into account a number of biological constraints linked to recombination, has difficulty reconciling three phenomena observed empirically: strong *PRDM9* diversity, strong selection on young alleles at this locus and the observation of hybrid sterility. The first two points (strong *PRDM9* diversity, strong selection on young alleles at this locus) which were addressed in our first article [133], highlighted the shortcomings of the current model with regard to the biological aspects of meiosis and recombination, and the need for empirical research to determine the effects of the various mechanisms involved in the process.

Also, in a bi-populationnal context, there is a new way to increase the PRDM9 diversity independent of the mutation rate at the *PRDM9* locus. Indeed, migration between two populations is characterized genetically by the intake of new genetic material in a population from another. This migration would have several effects depending on its intensity. First, in cases of high migration between the two populations, gene flow would be so imortant that the population would behave as a single panmictic population of the size of the two combined populations. In this case, the boundary between the two populations would become blurred, the hybrid would somehow lose hybrid status but would resemble any other individual from one of the two populations and the recombination landscapes would be almost homogeneous between the two populations. It would therefore be interesting to determine the migration threshold from which hybrids are no longer distinguishable from individuals of the two populations.

However, in a reverse case, when migration is low, this could, in cases where hybrids are able to exceed the potential sterility barrier, allow the contribution of new *PRDM9* alleles. In some cases, we could observe the invasion of some new "good" alleles brought by migration. This would then characterize adaptive introgressions. It would therefore be a new lever allowing the increase of *PRDM9* diversity in the population. We could thus study the strength of selection of a new allele in one or the other of the two populations according to its level of erosion to identify the probability of invasion of this

allele appeared either by mutation or by migration. This specific case could happen in non-neutral Red Queen regimes.

Moreover, in the case where an allele was introduced in the other population, and whose genetic background would have become homogeneous over the generations in the population would allow to act against the hybrid sterility in the hybrid individuals possessing this allele. Indeed, we could first observe homozygous hyrbids which would therefore have the advantage of dosage which could significantly increase their fertility. In the heterozygous case, an increase in hybrid fertility could also be observed. Indeed, hybrid sterility comes from the asymetric binding to high affinity sites which decreases the probability of making a DSB on a site of lower affinity but bound symmetrically. However, if for one of the two alleles of the hybrid, all the targets of high affinity are eroded (erosion in the popualtion of origin and in the new one), it increases the probability of DSB in a symmetrically bound site of low affinity. Note that, due to the other allele which would always have high affinity targets on a single chromosome, the fertility would not be fully restored. Migration could therefore increase intra-population diversity but this would tend to act against hybrid sterility.

There are, however, other parameters that can work in favor of hybrid sterility. Indeed, For example, the exact form of the affinity distribution of the targets which could counteract dosage effects, and the concentration of free PRDM9 in the cell which may imply competition between targets, could both also have an impact on the observation of hybrid sterility. Indeed, in the case of an affinity distribution with a marked split between high and low affinity targets, or in the case of strong competition between targets, hybrids could show stronger sterility patterns and often with lower levels of intra-population erosion.

Another biological mechanism that has not been taken into account here, and which could impact sterility in hybrids due to inter-allele dominance, is the possibility of polymerization of the PRDM9 protein when it binds to its targets [138; 148]. In particular, in cases of dominance of an old allele (therefore having eroded a lot in its original population) whose proteins would systematically dimerize with the proteins of the other younger allele (with high affinity targets), we could observe a surplus of binding to the asymmetrically high-affinity targets of the old allele, and a lack of binding to even the symmetrically active, high-affinity targets of the younger allele. This form of dominance could generate sterility even in hybrids with a young allele which, in the current model, is fertile because it reduces asymmetrical erosion.

3.6 Data accessibility

The model can be found at https://github.com/alicegenestier/Red_Queen_PRDM9_Panmictic.git. The model was implemented in C++ and all figures were created using Python.

3.7 Competing interest

The authors have no competing interests.

3.8 Funding

Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec.

3.9 Acknowledgment

All simulations of this work were performed using the computing facilities of the CC LBBE/PRABI.

Bibliographie

[1] Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nature Reviews Genetics;15(3):176–192. doi:10.1038/nrg3644. *Cited at page 79*

[2] Haldane JB. Sex ratio and unisexual sterility in hybrid animals. Journal of genetics. 1922;12:101–109. *Cited at pages 28, 79*

[3] Muller HJ. Bearing of the Drosophila work on systematics. The new systematics,. 1940; p. 185–268. *Cited at pages 28, 79*

[4] Wu CI, Johnson NA, Palopoli MF. Haldane's rule and its legacy: Why are there so many sterile males? Trends in Ecology & Evolution. 1996;11(7):281–284. doi:10.1016/0169-5347(96)10033-1. *Cited at pages 28, 79*

[5] Schilthuizen M, Giesbers MCWG, Beukeboom LW. Haldane's rule in the 21st century. Heredity;107(2):95–102. doi:10.1038/hdy.2010.170. *Cited at pages 28, 79*

[6] Dobzhansky T, Beadle GW. Studies on Hybrid Sterility IV. Transplanted Testes in Drosophila Pseudoobscura. Genetics. 1936;21(6):832–840. *Cited at pages 24, 79*

[7] Dobzhansky T. Genetics and the Origin of Species. 11. Columbia university press; 1982. *Cited at pages 24, 79*

[8] Muller HJ. Isolation mechanisms, evolution and temperature. Biology Symposium. 1942;6:71–125. *Cited at pages 24, 79*

[9] Muller HJ, Pontecorvo G. Recessive genes causing interspecific sterility and other disharmonies between Drosophila melanogaster and simulans. Genetics. 1942;27:157. *Cited at page 79*

[10] Orr HA. Dobzhansky, Bateson, and the Genetics of Speciation. Genetics. 1996;144(4):1331–1335. doi:10.1093/genetics/144.4.1331. *Cited at pages 24, 79*

[11] Ting CT, Tsaur SC, Wu ML, Wu CI. A Rapidly Evolving Homeobox at the Site of a Hybrid Sterility Gene. Science. 1998;282(5393):1501–1504. doi:10.1126/science.282.5393.1501. *Cited at pages 26, 79*

[12] Wu CI, Ting CT. Genes and speciation. Nature Reviews Genetics. 2004;5(2):114–122. doi:10.1038/nrg1269. *Cited at pages 26, 79*

[13] Phadnis N, Orr HA. A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids. Science. 2009;323(5912):376–379. doi:10.1126/science.1163934. *Cited at pages 26, 79*

[14] Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature;423(6941):715–719. doi:10.1038/nature01679. *Cited at pages 26, 79*

[15] Forejt J, Iványi P. Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.). Genetics Research. 1974;24(2):189–206. doi:10.1017/S0016672300015214. *Cited at pages 26, 79*

[16] Gregorová S, Forejt J. PWD/Ph and PWK/Ph Inbred Mouse Strains of Mus cyrchar
cyrt. musculus Subspecies-a Valuable Resource of Phenotypic Variations and Genomic Polymorphisms. Folia biologica. 2000;46:31–41. *Cited at pages 26, 29, 41, 80, 81*

[17] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science. 2009;323(5912):373–375. doi:10.1126/science.1163601. *Cited at pages 14, 26, 28, 30, 39, 80, 81, 96*

[18] Gregorova S, Gergelits V, Chvatalova I, Bhattacharyya T, Valiskova B, Fotopulosova V, et al. Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. eLife. 2018;7:e34282. doi:10.7554/eLife.34282. *Cited at pages 29, 30, 41, 80, 81, 89, 96, 107*

[19] Smagulova F, Brick K, Pu Y, Camerini-Otero RD, Petukhova GV. The evolutionary turnover of recombination hot spots contributes to speciation in mice. Genes & Development. 2016;30(3):266–280. doi:10.1101/gad.270009.115. *Cited at pages 41, 56, 57, 62, 80, 91, 96, 97, 108, 109, 112, 115, 136*

[20] Mukaj A, Piálek J, Fotopulosova V, Morgan AP, Odenthal-Hesse L, Parvanov ED, et al. Prdm9 Intersubspecific Interactions in Hybrid Male Sterility of House Mouse. Molecular Biology and Evolution. 2020;37(12):3423–3438. doi:10.1093/molbev/msaa167. *Cited at pages 29, 80, 96, 97, 117*

[21] AbuAlia KF, Damm E, Ullrich KK, Mukaj A, Parvanov E, Forejt J, et al.. Natural variation in Prdm9 affecting hybrid sterility phenotypes; 2023. Available from: https://www.biorxiv.org/content/10.1101/2023.01.17.524418v1. *Cited at pages 80, 96, 97, 117, 118*

[22] Seroussi E, Shirak A, Gershoni M, Ezra E, de Abreu Santos DJ, Ma L, et al. Bos taurus–indicus hybridization correlates with intralocus sexual-conflict effects of PRDM9 on male and female fertility in Holstein cattle. BMC Genetics. 2019;20(1):71. doi:10.1186/s12863-019-0773-5. *Cited at page 80*

[23] Borde V, Robine N, Lin W, Bonfils S, Géli V, Nicolas A. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. European Molecu-

lar Biology Organization Journal. 2009;28:99 – 111. doi:10.1038/emboj.2008.257. *Cited at pages 12, 80*

[24] Borde V, de Massy B. Programmed induction of DNA double strand breaks during meiosis: setting up communication between DNA and the chromosome structure. Current Opinion in Genetics & Development. 2013;23(2):147–155. doi:10.1016/j.gde.2012.12.002. *Cited at pages 12, 14, 30, 42, 80*

[25] Diagouraga B, Clément JAJ, Duret L, Kadlec J, Massy Bd, Baudat F. PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites. Molecular Cell. 2018;69(5):853–865.e6. doi:10.1016/j.molcel.2018.01.033. *Cited at pages 11, 12, 14, 62, 80, 91, 98, 115*

[26] Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997;94(15):8058–8063. doi:10.1073/pnas.94.15.8058. *Cited at pages 16, 18, 40, 65, 80, 106, 123*

[27] Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature. 2016;530(7589):171–176. doi:10.1038/nature16931. *Cited at pages 26, 29, 30, 39, 41, 57, 58, 60, 65, 76, 80, 89, 91, 96, 97, 98, 108, 109, 122*

[28] Baker Z, Przeworski M, Sella G. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence; 2022. Available from: https://www.biorxiv.org/content/10.1101/2022.09.27.509707v1. *Cited at pages 30, 36, 37, 43, 61, 80, 96, 107, 110, 112, 115*

[29] Genestier A, Duret L, Lartillot N. Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis : a model based exploration of the PRDM9 intra-genomic Red Queen; 2023. Available from: https://www.biorxiv.org/content/10.1101/2023.03.08.531712v2. *Cited at pages 30, 32, 80, 81, 82, 83, 85, 86, 93, 94, 95, 96, 97, 98, 107*

[30] Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F. The Evolution of House Mice. Annual Review of Ecology and Systematics. 1993;24(1):119–152. doi:10.1146/annurev.es.24.110193.001003. *Cited at pages 26, 81*

[31] Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, et al. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Molecular Ecology. 2008;17(24):5349–5363. doi:10.1111/j.1365-294X.2008.04005.x. *Cited at pages 26, 81*

[32] Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, et al. PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-

Specific Initiation of Meiotic Recombination. PLOS Genetics. 2015;11(1):e1004916. doi:10.1371/journal.pgen.1004916. *Cited at pages 28, 41, 56, 57, 81, 89, 108, 115*

[33] Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLOS Genetics. 2009;5(12):e1000753. doi:10.1371/journal.pgen.1000753. *Cited at pages 21, 23, 30, 39, 41, 59, 81, 108*

[34] Latrille T, Duret L, Lartillot N. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017;372(1736):20160463. doi:10.1098/rstb.2016.0463. *Cited at pages 21, 30, 36, 41, 43, 46, 51, 52, 56, 60, 66, 83, 85, 107, 125, 128, 129*

[35] Balcova M, Faltusova B, Gergelits V, Bhattacharyya T, Mihola O, Trachtulec Z, et al. Hybrid Sterility Locus on Chromosome X Controls Meiotic Recombination Rate in Mouse. PLOS Genetics. 2016;12(4):e1005906. doi:10.1371/journal.pgen.1005906. *Cited at pages 29, 97, 118*

[36] Baker CL, Petkova P, Walker M, Flachs P, Mihola O, Trachtulec Z, et al. Multimer Formation Explains Allelic Suppression of PRDM9 Recombination Hotspots. PLoS Genetics. 2015;11(9):e1005512. doi:10.1371/journal.pgen.1005512. *Cited at pages 53, 58, 59, 62, 99, 108, 115*

[37] Schwarz T, Striedner Y, Horner A, Haase K, Kemptner J, Zeppezauer N, et al. PRDM9 forms a trimer by interactions within the zinc finger array. Life Science Alliance. 2019;2(4). doi:10.26508/lsa.201800291. *Cited at pages 99, 115*

Partie III

Conclusion

4

Discussion & perspectives

Contents

4.1 Introduction et résumé des résultats

L'évolution de la recombinaison méiotique a fait l'objet de nombreuse études théoriques ([149; 150; 151; 152; 83; 153] et revues de [154; 155; 156]). Toutefois ces travaux se sont avant tout focalisés sur la question de la dissipation de la liaison génétique entre loci sous sélection, ignorant de ce fait les enjeux tout aussi importants du rôle de la recombinaison dans la méiose (et plus particulièrement dans l'appariement des chromosomes). Ce n'est que plus récemment que des travaux théoriques ont considéré les enjeux méiotiques, en

se concentrant plus particulièrement sur le gène *PRDM9* [85; 96; 134; 133]. En effet, il s'agit d'un des mécanismes impliqués dans la recombinaison méiotique les mieux connus actuellement. Cette dernière catégorie d'études est d'ailleurs celle dans laquelle s'inscrit mon travail de thèse mais avec deux particularités originales. Premièrement, les modèles théoriques présentés dans ce mémoire prennent en compte les mécanismes moléculaires de la méiose et de la recombinaison avec un niveau de détail qui n'avait jusqu'à présent pas été considéré (à l'exception du travail récent de Baker *et al.* [134]). Deuxièmement, ce modèle fait l'effort d'une calibration empirique, aspect assez rarement pris en compte dans le cadre des modèles de preuve de concepts théoriques.

Dans cette section, je vais tout d'abord résumer les résultats importants obtenus lors de ma thèse, puis je discuterai différentes pistes de réflexions à traiter avant de pouvoir obtenir des modèles moins conceptuels plus directement applicables aux données empiriques. Enfin, je terminerai sur des questionnements plus larges liés à mon travail de thèse, tel que présenté dans ce manuscrit.

4.1.1 Le modèle et ses objectifs

Mon travail de thèse consistait en la réalisation d'un modèle combinant à la fois les principes de génétique des populations et les mécanismes moléculaires de la recombinaison réalisés pendant la méiose chez des mammifères possédant *PRDM9*. L'une des grosses différences apportées par mon modèle, comparé aux anciens, réside dans l'implémentation de la liaison symétrique de PRDM9, considérée comme essentielle pour la bonne réussite de la méiose [125]. Ce modèle a été implémenté tout d'abord dans un contexte uni-populationnel afin de vérifier (i) si la liaison symétrique suffisait à faire tourner le processus de Reine Rouge intra-génomique proposé comme solution au paradoxe des points chauds, (ii) s'il était devenu possible d'identifier la cause de la sélection positive appliquée sur les nouveaux allèles *PRDM9* entrant dans la population, et (iii) si ce modèle, calibré avec des paramètres empiriques, donnait des valeurs raisonnables pour toutes les statistiques descriptives pour lesquelles des informations empiriques étaient disponibles. Dans un deuxième temps, ce modèle a été implémenté dans un contexte bi-populationnel, avec la création d'hybrides, afin de vérifier (i) si l'asymétrie de PRDM9 causait bien de la stérilité hybride, et dans quelles conditions, et (ii) si cette stérilité hybride était plutôt récurrente et puissante (ce qui conforterait l'idée de *PRDM9* comme gène de spéciation), ou si la stérilité serait, au contraire, plutôt marginale et transitoire.

4.1.2 Sélection positive et asymétrie

Le modèle réalisé en population panmictique a permis l'observation du processus de Reine Rouge, qui fonctionne par la compétition entre, d'une part, l'érosion des cibles des allèles *PRDM9* présents dans la population, et d'autre part, la sélection positive agissant sur les nouveaux allèles. En particulier, on a pu déterminer que cette sélection positive était causée par l'érosion préférentielle des sites de haute affinité chez les vieux allèles, ce qui entraîne une asymétrie de liaison de la protéine aux sites de plus faible affinité.

Cette baisse de taux de liaison symétrique impacte donc négativement la fertilité des individus portant ces vieux allèles.

4.1.3 Le problème de l'effet dosage

De plus, l'implémentation du dosage génétique a généré une prédiction contre-intuitive. En effet, le dosage, qui correspond au fait d'avoir une concentration de PRDM9 associée à un allèle spécifique doublée chez l'homozygote par rapport a l'hétérozygote, crée un avantage homozygote (ou désavantage hétérozygote). Ainsi, les vieux allèles à l'état homozygote vivent plus longtemps avec une érosion plus forte, mais les jeunes allèles, qui arrivent forcément à l'état hétérozygote, ont des difficultés à envahir la population du fait de leur dosage génétique défavorable. Ces nouveaux allèles sont alors contre-sélectionnés, ce qui induit un régime d'éviction des nouveaux allèles en faveur d'un autre allèle présent en forte fréquence homozygote dans la population. Le dosage a donc cet inconvénient de jouer contre la diversité et d'étendre la plage des régimes monomorphes.

4.1.4 Calibrations empiriques

Enfin, nous avons tenté une calibration empirique en essayant de prendre en compte un certain nombre d'observations, comprenant la forte diversité de *PRDM9* empiriquement observée chez de nombreuses espèces [137; 69; 68; 74; 71; 72; 73; 70], l'érosion faible à modérée des points chauds [124; 111], la sélection positive agissant sur *PRDM9* [95; 68]; et l'haplo-insuffisance de certains allèles [124; 138], tout en maintenant les différents paramètres du modèle dans des fourchettes de valeurs empiriquement raisonnables. Au bout du compte, notre modèle est coincé entre toutes ces contraintes, ce qui rend la fenêtre d'action très petite. Dans l'état actuel des choses, malgré la prise en compte de différents phénomènes tels que la non-limitation des gamètes ou la potentielle régulation du nombre de DSB, le modèle a permis de se rapprocher des observations empiriques sans pour autant réussir à les prédire correctement, simultanément pour tous les paramètres étudiés, que ce soit en contexte panmictique ou en contexte hybride. En particulier, la statistique la plus problématique semble être la diversité *PRDM9*. En effet le modèle a du mal à prédire une diversité comparable à celle observée empiriquement. De nouveau, cette composante est fondamentalement liée aux problèmes d'effet de dosage. De façon plus globale, le modèle semble être pris en étau entre plusieurs injonctions contradictoires. Soit on vise un régime dans lequel l'impact du dosage sur la fertilité est suffisamment fort (et ce, afin d'expliquer l'haplo-insuffisance des allèles vieux), mais on génère alors de l'éviction et on ne retrouve pas de forte diversité. Soit on diminue l'aspect du dosage (et ce afin de sortir du régime d'éviction), mais les baisses de fertilités et donc la sélection positive sont alors moins importantes, et on se retrouve dans un régime de Reine Rouge quasi-neutre.

4.1.5 Stérilité hybride

L'une des questions centrales concernant *PRDM9* concerne la stérilité des hybrides, stérilité dans laquelle ce gène semble être impliqué. Notre modèle adapté en contexte bi-populationnel a tout d'abord montré qu'il y a bien de la stérilité hybride causée par de l'asymétrie, mais contrairement à la population panmictique où la baisse de fertilité vient d'une asymétrie statistique de liaison, c'est une asymétrie de séquence due à l'érosion qui génère de l'asymétrie de liaison. Cette forme d'asymétrie est celle décrite dans les études qui ont permis de découvrir le rôle de la symétrie de PRDM9 pendant la méiose [111].

Par ailleurs, le modèle montre que la différence de fertilité observée entre les populations parentales et les hybrides sont accentuées par l'effet du dosage, donnant aux hybrides, toujours à l'état hétérozygotes, un désavantage vis-à-vis des parents se trouvant à l'état homozygote dans leur population respective. Enfin, nous avons tenté, là aussi, une calibration empirique. Bien que l'interprétation des résultats ne soit pas définitive, du fait des imperfections du modèle en l'état actuel, on peut tout de même fortement suggérer que la stérilité engendrée par *PRDM9* est en réalité assez marginale, ce qui ne plaide donc pas pour un rôle majeur de *PRDM9* dans la spéciation.

Au vu des résultats présentés ci-dessus, il devient nécessaire d'approfondir les différents points de blocage et d'incertitude, ce que je vais tenter de faire dans la section suivante.

4.2 Les différentes facettes des modalités d'action moléculaire de *PRDM9*

Il reste encore plusieurs paramètres qui ne sont pas totalement déterminés empiriquement, en particulier la distribution d'affinité des points chauds et la concentration de PRDM9 dans la cellule. Ces paramètres jouent un rôle crucial dans la modulation de la dynamique de Reine-Rouge, d'où l'importance de mieux les comprendre.

4.2.1 La distribution d'affinité

Comme rappelé ci-dessus, une des raisons fondamentales pour laquelle on n'arrive pas à prédire une forte diversité *PRDM9* est l'effet dosage. Ce dernier est en fait intrinsèquement lié à la distribution d'affinité.

Dans notre modèle, la distribution d'affinité des cibles utilisée suit une loi de type exponentiel (cf Firgue 4.1A). Cette distribution est justifiée par la distribution des DSBs déterminée par expérience de Chip-seq [137]. Par cette distribution continue et de variance modérée, les homozygotes sont toujours avantagés par rapport aux hétérozygotes. Ce fait est à l'origine du phénomène d'éviction. Cependant, si cette distribution parait assez bien adaptée concernant les sites de haute affinité, la validité de cette distribution pour les sites de faible affinité, difficile à détecter pas Chip-Seq, est moins évidente. Il peut donc être utile d'explorer d'autres types de distribution, qui pourraient éventuellement pallier le problème de l'éviction.

Distribution binaire (Baker *et al.* 2015)

À la différence de ce que j'ai présenté ici, Baker *et al.* (2022) utilisent une distribution binaire [134] (cf Firgue 4.1B). Celle-ci découpe la distribution en deux : les sites de très haute affinité, en faible quantité, et les sites de faibles intensités, en très grande quantité. À première vue, cette distribution ne semble pas réaliste car non continue. Cependant, elle a l'avantage de faire une claire distinction entre les sites de haute et de faible affinité, ce qui donne des résultats intéressants, que je ne retrouve pas dans mon modèle. Du fait que les allèles vieux saturent les sites de haute affinité en dosage hétérozygote, doubler le dosage va avoir pour conséquence d'augmenter la liaison asymétrique sur les sites de faible affinité, qui sont les seuls encore disponibles. Les homozygotes sont donc désavantagés lorsqu'ils sont vieux. Pour les allèles jeunes, au contraire, le dosage est favorable aux homozygotes. Cette inversion de l'effet du dosage fait que le régime d'éviction est peu présent dans le modèle de Baker *et al.* (2022) [134].

Distribution de type inverse

La distribution binaire de Baker *et al.* 2022 possède un caractère phénoménologique non dépeint par notre distribution actuelle, de par la scission plus marquée entre les rares cibles de hautes affinité comparée à la masse de sites de faible affinité. D'autre part, une distribution d'affinité continue, telle que supposée dans notre étude, est empiriquement plus défendable. De ce fait, on pourrait imaginer une distribution conciliant ces deux caractéristiques.

En particulier on pourrait imaginer une distribution plutôt sous la forme $\frac{1}{\epsilon+x}$ avec ϵ proche de 0^+ (cf Firgue 4.1C). Cette distribution a l'avantage d'être plus marquée dans la différence de proportion de cibles de haute ou de faible affinité, tout en restant continue. Cette distribution pourrait donc s'appliquer pour les cibles de faible affinité, associée à la distribution exponentielle observée pour les cibles de haute affinité. Elle pourrait de plus permettre de résoudre le problème de l'éviction, tout en étant empiriquement bien justifiée.

Distribution déterminée par un motif ADN

Un des inconvénients de la distribution qui vient d'être suggérée est qu'elle est purement phénoménologique. Elle n'est donc pas motivée clairement par des considérations mécanistes. Dans cette direction, on pourrait utiliser une distribution d'affinité directement obtenue à partir de considérations combinatoires sur le nombre de match et mismatch entre séquence ADN et motif reconnu par le domaine à doigts de zinc, tout en modélisant la décroissance de l'affinité avec le nombre de mismatch (cf Firgue 4.1D). Cette distribution aurait l'avantage d'implémenter directement une baisse d'affinité progressive en fonction des mutations réalisées au niveau des cibles, plutôt qu'une simple inactivation totale, comme c'est le cas dans le modèle actuel. Ce mécanisme ralentirait l'érosion car contrairement à notre modèle actuel où les sites possèdent un état binaire

(actif ou inactif après avoir subit une mutation), la perte d'affinité est continue, et donc l'inactivation des cibles est plus progressive. On s'attendrait donc à observer une Reine Rouge plus lente que ce qui est prédit par le modèle actuel. Cependant, cette distribution d'affinité nécessiterait plusieurs changements dans le modèle, dont un codage du génome à l'échelle du nucléotide, que ce soit au niveau des cibles ou au niveau du locus *PRDM9*. Cela permettrait par ailleurs de prendre en compte d'autres mécanismes, tels que la possibilité d'avoir des allèles fonctionnellement équivalents et/ou reconnaissant des cibles qui se chevauchent.

C'est par la comparaison avec le modèle de Baker *et al.* que nous avons pu déterminer que la distribution d'affinité pourrait jouer contre l'éviction due aux effets dosages. Cependant, le modèle de Baker *et al.* assume un autre processus que nous n'avons pas implémenté dans notre modèle et qui fait que le modèle ne fonctionne pas exactement de la même manière. Cet autre composante est la compétition entre cibles.

4.2.2 La compétition entre cibles

Le phénomène de compétition entre cibles consiste en une concentration de PRDM9 limitante dans la cellule, ce qui implique que les sites de liaison sont en compétition pour la liaison de la protéine. Ce phénomène dépend donc de la concentration de la protéine dans la cellule, et de l'affinité des cibles.

Afin de bien comprendre les enjeux, il faut revenir au principe biochimique, qui stipule que la constante d'association entre deux composés chimiques, ici des protéines libres de PRDM9 (de concentration $[P]_{libre}$) qui se lient à des séquences ADN cibles d'affinité i (de proportion $[S_i]$), s'écrit :

$$K_i = \frac{[PS_i]}{[P]_{libre}[S_i]} = \frac{x_i}{[P]_{libre}(1 - x_i)}, \tag{4.1}$$

où $[PS_i] = x_i$ est la fraction du nombre total de molécules PRDM9 qui est lié aux cibles. En inversant cette équation, cette fraction x_i s'écrit :

$$x_i = \frac{[P]_{libre}K_i}{1 + [P]_{libre}K_i}. \tag{4.2}$$

De plus, si on introduit la concentration totale de PRDM9 dans la cellule, notée $[P]_{tot}$, et la concentration de protéine PRDM9 occupant des cibles $[P]_{occup}$, on obtient la relation suivante :

$$[P]_{occup} = [P]_{tot} - [P]_{libre} = \sum_i x_i. \tag{4.3}$$

L'équation (4.3) introduit un couplage entre cibles. En effet, d'après cette équation, si on augmente le nombre de cibles liées, on diminue mécaniquement le nombre de protéines libres susceptibles de se lier, ce qui, d'après l'équation (4.2), diminue la probabilité de lier plus de cibles. C'est ce couplage qui caractérise la compétition entre cibles.

Cependant, si on se place dans un cas où l'affinité moyenne est faible, ou dans un cas où la concentration $[P]_{tot}$ est forte, la proportion de protéines liées par rapport

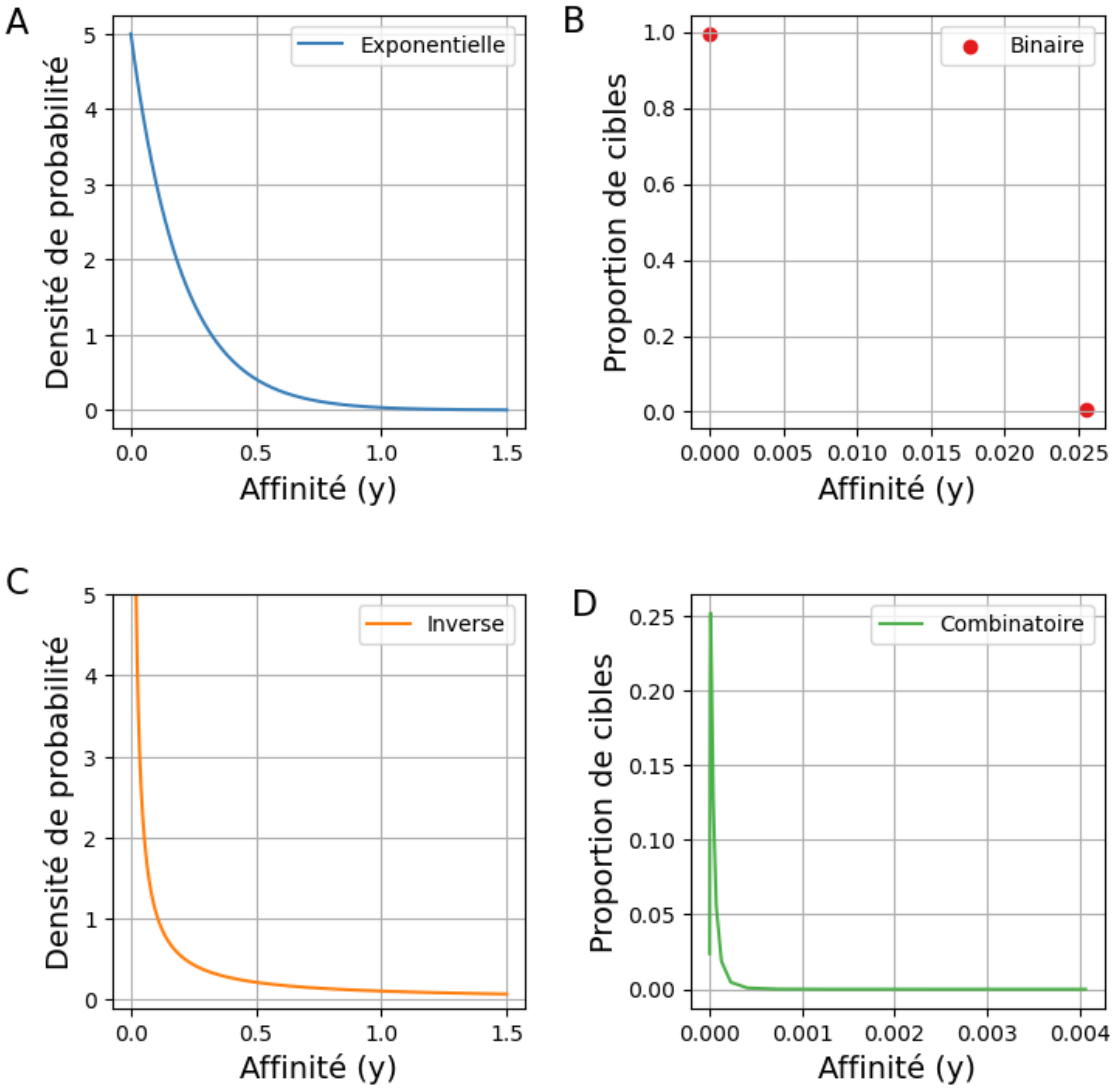

Figure 4.1: Différents types de distribution d'affinité des cibles de PRDM9. *(A) Distribution Exponentielle : c'est la distribution utilisée dans notre modèle et qui colle aux données des expériences de Chip-seq [137], tout du moins en ce qui concerne les cibles de haute affinité. (B) Distribution binaire : c'est le type de distribution utilisée dans l'article de Baker et al. [134]. (C) Distribution Inverse : distribution continue montrant un marquage plus fort entre les sites de haute et basse affinité. (D) Distribution déterminée par un motif ADN : distribution basée sur des considérations mécanistes à l'échelle du nucléotide. À noter que les axes des ordonnées de graphiques (A) et (C) correspondent à une densité de probabilité alors que pour les graphiques (B) et (D) on est sur une proportion de cibles. Enfin, les axes des abscisses représentant l'affinité des cibles ne sont pas tous à la même échelle car elles n'ont pas été construites de la même manière, ce qu'il est important de retenir est plutôt l'allure générale des distributions.*

à la concentration totale est relativement faible ($[P]_{occup} << [P]_{tot}$). La concentration de protéines libres est donc proche de la concentration totale ($[P]_{libre} \approx [P]_{tot}$), et se trouve alors très peu influencée par l'occupation des cibles. Dans ce cas, la compétition entre cibles devient négligeable. Ainsi, la probabilité de liaison d'une protéine à une cible devient :

$$x_i \approx \frac{[P]_{tot} K_i}{1 + [P]_{tot} K_i} = \frac{y_i}{1 + y_i} \tag{4.4}$$

Dans ce cas, on peut définir l'affinité relative d'une cible i comme $y_i = [P]_{tot} K_i$, et paramétrer le modèle directement en fonction de y_i. Enfin, en présence de dosage, on obtient :

$$x_i = \frac{c y_i}{1 + c y_i} \tag{4.5}$$

où c est le coefficient de dosage prenant les valeurs 1 ou 2 suivant si on se trouve chez un hétérozygote ou un homozygote. On retrouve ici la formulation utilisée dans le premier manuscrit.

Par rapport à la dérivation qui vient d'être présentée, et qui est donc valide quand PRDM9 n'est pas limitant, la compétition entre cibles a deux conséquences. Premièrement, si on augmente la concentration totale ($[P]_{tot}$) en la multipliant par 2, comme c'est supposé être le cas chez les homozygotes par rapport aux hétérozygotes, les concentrations de protéines libres ($[P]_{libre}$) et de protéines liées ($[P]_{occup}$) vont elles aussi augmenter mais pas d'un même facteur ($c < 2$, cf Box "Variations de la concentration libre et de la concentration liée de PRDM9 en fonction des variations de la concentration totale dans la cellule" page 114). La compétition entre cibles a donc tendance à modérer les effets du dosage, mais sans en changer la direction.

Deuxièmement, lorsque des cibles sont érodées, cela libère des protéines, et donc la concentration libre de PRDM9 augmente, ce qui permet de lier plus de cibles de faibles affinités. Or, dans notre modèle actuel, la probabilité de liaison à une cible ne dépend pas du nombre de cibles déjà érodées. Ce mécanisme pourrait permettre, dans le cas d'une distribution continue comme la notre, de ralentir la perte d'asymétrie pour des cibles d'affinité moyenne, et donc de ralentir la perte de fertilité en fonction de l'érosion, par rapport à ce qui est prédit en l'absence de compétition.

Box Variations de la concentration libre et de la concentration liée de PRDM9 en fonction des variations de la concentration totale dans la cellule

D'après l'équation (4.3), on exprime la concentration de protéines libres comme

$$[P]_{libre} = [P]_{tot} - \sum_i x_i. \tag{4.6}$$

Ainsi, en dérivant selon $[P]_{tot}$ on obtient

$$\frac{\partial [P]_{libre}}{\partial [P]_{tot}} = 1 - \sum_i \frac{\partial x_i}{\partial [P]_{tot}}. \tag{4.7}$$

Or en décomposant $\frac{\partial x_i}{\partial [P]_{tot}}$ en

$$\frac{\partial x_i}{\partial [P]_{tot}} = \frac{\partial x_i}{\partial [P]_{libre}} \frac{\partial [P]_{libre}}{\partial [P]_{tot}} \tag{4.8}$$

l'équation (4.7) devient

$$\frac{\partial [P]_{libre}}{\partial [P]_{tot}} = 1 - \frac{\partial [P]_{libre}}{\partial [P]_{tot}} \sum_i \frac{\partial x_i}{\partial [P]_{libre}} = \frac{1}{1 + \sum_i \frac{\partial x_i}{\partial [P]_{libre}}}. \tag{4.9}$$

De plus, comme $\frac{\partial x_i}{\partial [P]_{libre}} > 0$, on en déduit que

$$0 < \frac{\partial [P]_{libre}}{\partial [P]_{tot}} < 1. \tag{4.10}$$

Ainsi, la concentration libre de PRDM9 augmente quand la concentration totale augmente mais moins rapidement.
Concernant la concentration de protéines liées, en dérivant selon $[P]_{tot}$ l'équation (4.3), on obtient :

$$\frac{\partial [P]_{occup}}{\partial [P]_{tot}} = 1 - \frac{\partial [P]_{libre}}{\partial [P]_{tot}} \tag{4.11}$$

qui, combinée a l'encadrement (4.12) donne

$$0 < \frac{\partial [P]_{occup}}{\partial [P]_{tot}} < 1. \tag{4.12}$$

Ainsi, la concentration de PRDM9 liées augmente quand la concentration totale augmente mais moins rapidement.

4.2.3 La dominance entre allèles *PRDM9*

Une troisième composante qui pourrait jouer en faveur de la sélection positive de certains allèles, et jouer contre des effets dosages d'autres allèles serait la dominance entre allèles *PRDM9*. De manière générale, le nombre total de DSB est régulé à travers le génome, et les DSB sont induites aléatoirement sur les cibles liées par les protéines PRDM9 d'un allèle ou de l'autre. Ainsi, la dominance d'un allèle, i.e. la capacité de diriger la localisation de la majorité des cassures double brins, viendrait du fait que cet allèle se

lie à plus de cibles que l'autre. Cette dominance n'a pour le moment pas encore été modélisée. Cependant, il existe quelques pistes sur les causes de cette dominance, et plusieurs paramètres pouvant l'influencer.

Tout d'abord, une partie des effets de dominance est possiblement due à des questions d'érosion [36], de sorte que les allèles ayant moins érodé se lient plus, et donc induisent plus de DSB que les allèles plus vieux (exemple de l'allèle *Cst* (*Mus musculus castaneus*) qui est dominant face à *Dom2* (souche B6 *Mus musculus domesticus*), plus vieux, mais récessif face à 13R qui est plus jeune, [138; 137]). Cependant, il semble y avoir aussi une dominance intrinsèque, qui serait liée au fait que les allèles n'ont pas, au départ, le même nombre de cibles, ou la même affinité moyenne pour celles-ci (exemple de l'allèle *Cst* qui est dominant par rapport à l'allèle humanisé dans le génome de la souris, alors que ce dernier n'a jamais érodé chez la souris [143; 25]). Les effets de dominance pourraient également être dus à des différences de niveau d'expression entre allèles, mais cette éventualité semble a priori moins probable empiriquement (pas de différence d'expression entre l'allèle *Cst* et humanisé chez la souris [25]), dans la mesure où la régulation du niveau d'expression n'est pas directement liée à l'évolution du domaine à doigt de zinc. Il semble donc logique de supposer que les différents allèles *PRDM9* ont le même niveau d'expression.

Dans un contexte où PRDM9 agit sous forme de monomères (un point que l'on réexaminera ci-dessous), les prédictions diffèrent selon que PRDM9 est limitant (régime de compétition forte entre cibles) ou non. En substance, les effets de dominance sont naturellement prédits dans un contexte ou PRDM9 est en excès, comme supposé dans le cadre de mon modèle. En revanche, en compétition forte (tel que dans le modèle de Baker *et al.* (2022) [134]), on s'attend en principe à une absence de dominance. En effet, dans ce dernier cas, comme PRDM9 est limitant, toutes les protéines libres se lient à une cible, et si les deux allèles ont le même niveau d'expression, ils vont se lier au même nombre de cibles, ce qui n'est pas cohérent avec la dominance qui demanderait qu'un allèle se lie à plus de cibles que l'autre pour avoir plus de chance de subir des DSB. Cette observation est importante, car elle constitue un argument pour penser que, empiriquement, PRDM9 n'est pas limitant.

Toutefois, un autre aspect, non discuté jusqu'à présent, rend les choses potentiellement plus complexes : la possibilité que PRDM9 pourrait fonctionner en dimères (voire même en trimères [148]). Cette possibilité est mentionnée dans la littérature, mais sur la base de données empiriques [124; 138] qui demanderaient encore confirmation). Dans ce cas de polymérisation des protéines PRDM9, même en situation de compétition entre cibles, on pourrait observer de la dominance entre allèles. En effet, en se liant entre elles, deux protéines de deux allèles différents ne peuvent se lier qu'à une seule cible. Il suffit donc qu'un allèle ait des cibles d'assez haute affinité pour sa protéine pour observer un surplus de liaisons pour un allèle au détriment du deuxième. Ainsi, même si toutes les protéines sont liées, il n'y a plus forcement le même nombre de cibles liées par chaque allèle.

Notons qu'être dominant ne signifie pas forcement que l'allèle induira une meilleure

fertilité que l'allèle récessif. En effet, un allèle qui serait dominant, donc dirigerait la localisation de la majorité des DSB, mais sans avoir beaucoup de cibles liées symétriquement n'aurait pas forcément une meilleure probabilité d'invasion qu'un allèle récessif ayant un bon taux de liaison symétrique. La sélection d'un allèle dépendrait donc à la fois de sa dominance intrinsèque sur d'autres allèles et de son fort taux de symétrie par rapport aux autre allèles présents dans la population.

4.2.4 Vers un modèle global du mode d'action moléculaire de PRDM9

En conclusion, tous les points évoqués ci-dessus mériteraient d'être soigneusement explorés, aussi bien de façon empirique que théorique. L'enjeu ultime serait de pouvoir aboutir à un modèle global du mode d'action biochimique et moléculaire de *PRDM9*. Cela demanderait, d'une part, une meilleure connaissance empirique des détails biochimiques et de la valeur de certains paramètres, comme la distribution d'affinité des cibles ou la concentration de PRDM9 libre dans la cellule, qui restent encore inconnus ou insuffisamment caractérisés. D'autre part, un gros travail de modélisation est nécessaire, en s'inspirant des différentes pistes évoquées ci-dessus, pour réussir à démêler les différentes conséquences logiques de ces divers aspects du fonctionnement moléculaire.

4.3 Reine Rouge neutre ou sélective

Comme présenté dans la partie 1.4, il est encore difficile à ce stade de concilier plusieurs observations empiriques. En particulier, il est difficile de prédire simultanément une forte diversité *PRDM9* et une forte sélection positive sur le domaine à doigt de zinc. Ainsi, les régimes impliquant un niveau de diversité raisonnable empiriquement sont souvent des modèles qui présentent un niveau de sélection positive envers *PRDM9* relativement faible, voire même nul ou transitoire. Dans ces cas précis, caractérisés entre autre par la présence de gamètes non limitants, le processus de Reine Rouge est neutre ou quasi-neutre, au sens ou le turnover des allèles *PRDM9* au cours du temps n'est plus médié par la sélection, mais par la seule pression mutationnelle combinée à la perte d'anciens allèles par dérive. Cette dynamique neutre découle de la prise en compte de processus diminuant l'effet dosage pour lutter contre le régime d'éviction, mais cela a aussi l'effet de limiter la perte de fertilité des individus à des cas extrêmes d'érosion.

Face à ce dilemme, notre premier réflexe a été d'incriminer le modèle. Toutefois, à bien y réfléchir, il se pourrait que cette neutralité de la Reine Rouge ne soit finalement pas entièrement fausse. On pourrait en effet se retrouver empiriquement dans une alternance entre régime neutre et régime sous sélection positive. De fait, le domaine à doigts de zinc mute très rapidement, mais ces changements peuvent être découpés en deux cas distincts. Premièrement, le domaine à doigt de zinc change lorsqu'un nouveau doigt de zinc apparaît. Or, le taux d'apparition d'un nouveau doigt de zinc est assez faible car il faut pour cela que la mutation se passe très précisément à l'une des

trois positions impliquées dans la liaison à l'ADN, c'est-à-dire les positions -1, 3 et 6 de l'hélice α du doigt de zinc [57; 65; 66]. En réalité, le domaine à doigts de zinc mute très rapidement du fait de la structure en minisatellite qui code pour ce domaine. En effet, les minisatellites, sont connus pour être des régions qui mutent plus souvent que le reste du génome. De plus, il sont constitués de répétitions en tandem de séquences dont le motif original peut varier et atteindre plusieurs dizaines de nucléotides de long. Ces répétitions sont des régions génomiques ayant subit des CO inégaux et continuant d'en subir, ce qui implique que les doigts de zinc vont avoir tendance à s'homogénéiser dans le domaine par évolution concertée.

Ainsi, on pourrait imaginer une dynamique de Reine Rouge *PRDM9* qui agirait comme suit : en premier, un allèle possédant un tout nouveau doigt de zinc reconnaîtrait de nouveaux motifs et serait donc sélectionné positivement. L'allèle serait alors soumis à l'homogénéisation de ses doigts de zinc formant alors plusieurs allèles fonctionnellement différents mais possédant des doigts de zinc similaire entre eux. Une fois les allèles fortement homogénéisés, l'évolution serait alors stoppée et un nouveaux doigt de zinc qui pourrait apparaître par mutation serait alors positivement sélectionné. On aurait donc des phases de sélections positives de nouveaux doigts de zinc entrecoupées de phase d'évolution concertée des doigts de zinc constituant une dynamique plus neutre.

4.4 La stérilité hybride et la spéciation

PRDM9 a été identifié comme étant le premier gène potentiel de spéciation chez les vertébrés. Durant ce travail de thèse, nous avons réalisé un modèle prenant en compte les différents mécanismes d'action de *PRDM9* pendant la méiose et en particulier le mécanisme de liaison symétrique dont le rôle a été initialement découvert lors d'expérimentation chez des hybrides stériles de souris. Nos résultats ne vont cependant pas dans le sens d'un rôle majeur de *PRDM9* comme gène de spéciation. En effet, on prédit plutôt de la stérilité hybride de manière marginale et transitoire dans des régimes assez restreints, caractérisés par des forts taux de mutations des cibles et un faible taux de mutation au locus *PRDM9*. Au vu de ce qui a été discuté précédemment dans un contexte panmictique, plusieurs facteurs pourraient impacter le modèle et changer cette prédiction. Cependant, il semblerait qu'augmenter la diversité (ce que l'on cherche à faire dans le modèle uni-populationnel) a tendance à diminuer l'érosion moyenne des cibles, ce qui diminue la probabilité d'asymétrie chez les hybrides, et donc leur stérilité. De plus, il a été découvert que dans les populations sauvages de souris, un niveau de diversité élevé ne montre pas de cas récurent et intense de stérilité hybride [131; 146].

Il existe cependant certains mécanismes non pris en compte dans ce modèle, qui pourraient expliquer les niveaux de stérilité hybride observés empiriquement.

4.4.1 Les effets Dobzhansky-Muller

L'un des modèles proposés pour expliquer génétiquement la stérilité hybride menant à de la spéciation est le modèle de Dobzhansky-Muller présenté dans la Box "Modèle des incompatibilités Bateson-Dobzhansky-Muller" de l'introduction. Pour que ce phénomène soit effectif, il faut des interactions entre différents loci géniques. Dans notre cas, nous avons uniquement un seul locus qui interagit avec des cibles à travers le génome. Le phénomène observé avec *PRDM9* et ses cibles ressemble à une incompatibilité entre allèles, mais cette incompatibilité s'applique au même locus sur les chromosomes homologues, et son intensité dépend des niveaux d'érosion respectifs des deux allèles dans les populations parentales. Cette incompatibilité n'est pas définitive mais plutôt transitoire. Ce n'est pas celle décrite par le modèle de Dobzhansky-Muller.

Le gène *PRDM9* semble interagir avec le locus de stérilité hybride *Hstx2* se trouvant sur la région proximale du chromosome X chez *Mus musculus*. Les hybrides issus de croisements PWDxB6 sont stériles [129; 130]. De plus, le locus *Hstx2* se comporte comme un point froid de recombinaison, car il présente des niveaux faibles de H3K4me3 médiés par PRDM9, et des marqueurs DMC1 sur les DSBs. Or, des études récentes n'ont pas montré de baisse de fertilité chez les hybrides due à l'incompatibilité entre *PRDM9* et *Hstx2* [146].

Ce mécanisme d'incompatibilité n'est pas pris en compte dans notre modèle, mais pourrait potentiellement expliquer certaines baisses de fertilité observées par ailleurs chez les hybrides. Cette incompatibilité avec le locus *Hstx2* est encore en discussion, mais il se pourrait que d'autres incompatibilités n'aient pas encore été trouvées.

Remarquons que les phénomènes de baisse de fertilité, même transitoires, peuvent être accompagnés d'évolutions comportementales au sein des deux populations parentales, en vue d'éviter l'accouplement hybride. Cela génère alors une nouvelle barrière, cette fois-ci comportementale et pré-zygotique, dans les zones de second contact, ce qui permet l'évolution indépendante et la potentielle formation de nouvelles incompatibilités génétiques.

4.4.2 Migration et introgression d'allèles

Dans un contexte bi-populationnel, un autre phénomène pourrait moduler la diversité *PRDM9* : la migration. En effet, dans le modèle bi-populationnel étudié dans cette thèse, les deux populations ne se mélangent jamais, et donc ne partagent jamais de matériel génétique (pas de flux de gènes). Les hybrides sont seulement créés pour étudier la différence de fertilité entre eux et les populations parentes, sans qu'ils soient réintroduits par la suite dans l'une ou l'autre des populations. Cependant, s'il était possible d'échanger des individus entre populations, on pourrait faire apparaître de nouveaux allèles *PRDM9* dans chacune des populations sans avoir à augmenter le taux de mutation au locus *PRDM9*.

L'augmentation de diversité médiée par la migration dépend de plusieurs facteurs. Premièrement, l'intensité de la migration entre les populations permettrait de mod-

uler l'apport de nouveau matériel génétique d'une population à l'autre. Pour cela il faudrait en premier lieu trouver le point à partir duquel les deux populations se comportent comme une seule et même population (de taille efficace deux fois plus grande que les sous-populations d'origine). En effet, dans des cas de migration forte, l'apport de nouveaux migrants est tellement important que la limite génétique entre les deux populations est fortement estompée, voire même indétectable, du fait du fort brassage génétique inter-populationnel.

Dans le cas de faible migration, le nouvel allèle apporté par le migrant devra surmonter plusieurs barrières avant d'être réellement introduit et jouer un rôle dans le calcul de la diversité de *PRDM9*. En effet, dans un premier temps, l'allèle devra surmonter la barrière de la stérilité, si elle existe. Ici, deux cas se présentent. Soit l'allèle est jeune (il est récemment apparu par mutation dans la population d'origine), et se comporte comme tel dans la population de destination, ce qui limite la stérilité hybride. Soit l'allèle est vieux dans la population d'origine (il possède donc déjà une forte érosion dans son génome), ce qui génère alors potentiellement de la stérilité hybride. Si cette stérilité n'est pas trop forte, l'allèle peut potenitellement passer au bout de quelques générations la barrière de la stérilité. Une seconde barrière est celle de la dérive. L'allèle entrant doit, une fois débarrassé de la stérilité hybride, augmenter significativement en fréquence pour peser dans le calcul de la diversité. Pour cela, il doit soit avoir de la chance de passer à travers la dérive, soit être sélectionné positivement. Dans des régimes non neutres, on pourrait donc observer des introgressions adaptatives, c'est-a-dire des allèles d'une population sélectionnés positivement dans une autre population ou leur niveau d'érosion est nul. On pourrait par exemple s'attendre à ce qu'un allèle, qui aurait été sélectionné dans une population, le soit aussi dans une seconde suite à un évènement de migration.

4.5 Questionnements plus larges

4.5.1 Les effets Hill-Robertson

Dans notre modèle actuel, la recombinaison est considérée uniquement comme favorisant l'appariement des chromosomes. Or la recombinaison est connue pour générer de nouvelles combinaisons d'allèles. Elle permet en particulier de casser la liaison génétique entre différents loci, c'est-à-dire casser la transmission mutuelle de deux loci proches physiquement sur un même chromosome, et donc permet de limiter les effets Hill-Robertson (voir la Box "Déséquilibre de Liaison et effets Hill-Robertson") qui y sont associés. Ces effets caractérisent le fait que la liaison génétique entre des sites sous pression de sélection réduit la sélection globale dans le reste du génome. De plus, ces effets ont été proposés comme la raison de l'évolution du paysage de recombinaison.

Plusieurs simulations et modèles théoriques ont montré que ces effets généraient de la sélection indirecte sur des loci dit "modificateurs de recombinaison" [157; 151; 158], c'est-à-dire agissant directement sur la recombinaison, comme par exemple *PRDM9* qui détermine les points de recombinaisons. En particulier, dans le cas de *PRDM9*, des

allèles reconnaissant des cibles situées entre des loci liés génétiquement pourraient être sélectionnés pour cette raison, et donc envahiraient la population. Ce processus générerait donc de la sélection diversifiante, ce qui manque actuellement à notre modèle pour lutter contre le régime d'éviction sans avoir à augmenter le taux de mutation au locus *PRDM9*.

Ce mécanisme est testable avec notre modèle actuel en implémentant des loci tout le long du génome, qui seraient chacun sujet à sélection, positive ou négative, et d'intensités différentes. On pourrait alors examiner dans quelle mesure les effets d'interférence entre ces loci sous sélection généreraient de la sélection positive sur le locus *PRDM9*, spécifiquement en faveur des allèles reconnaissant des cibles à travers le génome susceptibles de casser au mieux les liaisons génétiques freinant l'adaptation en ces loci.

Pour conclure, introduire des problématiques de Hill-Robertson dans le cadre de mon modèle permettrait de revisiter les questions théoriques qui ont été explorées depuis longtemps, mais cette fois-ci dans un contexte où les mécanismes moléculaires sont bien spécifiés, et où les considérations du rôle de la recombinaison dans le bon accomplissement de la méiose (pour des raisons d'appariement et de ségrégation des chromosomes) sont et déjà bien modélisées. De ce fait, un tel travail de modélisation offrirait une opportunité unique d'articuler de façon soigneuse les enjeux méiotiques et génétiques de l'évolution de la recombinaison.

Déséquilibre de Liaison et effets Hill-Robertson

Le **déséquilibre de liaison** correspond à une association statistique entre les allèles de 2 ou plusieurs loci. En l'absence de déséquilibre de liaison, la fréquence d'observation d'haplotypes à la génération t+1 est égale au produit des fréquences des allèles constituant cet haplotype à la génération t. Au contraire, lors de présence de déséquilibre de liaison, certains haplotypes vont être plus fréquents que d'autres à la génération t+1. Dans le cas d'un surplus d'haplotype par rapport à ce qui est attendu au hasard, on a un déséquilibre positif; dans le cas d'un défaut d'haplotype, le déséquilibre est négatif. On observe donc que certaines combinaisons d'allèles sont plus souvent transmises que d'autres. Plus les loci sont proches physiquement l'un de l'autre sur le chromosome, plus le déséquilibre est élevé. Pour casser cette liaison génétique, il faut qu'il y ait recombinaison et en particulier des CO entre ces loci.

Lorsque la recombinaison est faible, le déséquilibre de liaison entre loci proches crée des effets qu'on appelle **effets Hill-Robertson**.
- L'interférence Hill-Robertson [159] se traduit en la fixation par dérive d'un certain haplotype due à la présence de mutations faiblement avantageuses, subissant une sélection de même intensité, à différents loci dans différents haplotypes. À terme, seulement certaines mutations auront été fixées, et les autres perdues.
- Le balayage sélectif [160] correspond à la fixation d'une mutation faiblement délétère à un certain locus, en même temps qu'une mutation fortement avantageuse située sur un autre locus qui lui est lié génétiquement. Cela mène à une perte locale de polymorphisme.
- À l'inverse, la sélection d'arrière plan [161] correspond à la perte d'une mutation faiblement avantageuse ou neutre à un certain locus en même temps qu'une mutation modérément délétère située sur un autre locus qui lui est lié génétiquement.

4.5.2 Évolution des paramètres

Le modèle utilisé dans ces travaux de thèse fait une hypothèse d'invariabilité des différents paramètres moléculaires au cours du temps. Autant que possible, les valeurs imposées à ces paramètres étaient justifiées empiriquement. Autoriser l'évolution de certains de ces paramètres pourraient cependant permettre, d'une part, de mieux comprendre pourquoi ces paramètres s'avèrent avoir ces valeurs empiriques, et d'autre part, d'étudier les différentes phases d'évolution du système de recombinaison méiotique dépendant de *PRDM9*.

Dans une version du modèle prenant en compte le niveau d'expression de *PRDM9* dans la cellule, qui permettrait entre autre d'étudier des régimes avec PRDM9 limitant impliquant de la compétition entre cibles, il serait intéressant de donner la possibilité de faire évoluer ce paramètre. En particulier, au vu des conséquences des effets dosage et de l'affaiblissement de ceux-ci quand le taux d'expression augmente (cf Figure supplémentaire 7 de l'article 1), on pourrait s'attendre à ce que le niveau d'expression ait évolué à un niveau assez élevé pour limiter les effets dosage.

Empiriquement, et comme présenté précédemment, la protéine PRDM9 possède un domaine constitué de plusieurs doigts de zinc reconnaissant chacun un motif ADN particulier. Afin de simplifier, nous avons implémenté un modèle discret où un allèle *PRDM9* reconnaît un nombre h de cibles à travers le génome, mais sans détermination par des séquences ADN précises. Dans une version améliorée où le modèle serait codé à l'échelle du nucléotide, les cibles reconnues par chaque doigt de zinc, ainsi que les mutations ou réarrangements dans le minisattelite codant pour le domaine à doigt de zinc de PRDM9, seraient directement issues du fonctionnement du modèle, et non plus posé arbitrairement comme dans la version actuelle. Cette façon de faire permettrait d'étudier à la fois l'évolution du nombre et de la diversité des doigts de zinc, ainsi que le nombre et l'affinité des cibles reconnues. En particulier, on pourrait voir apparaître de la dominance entre allèles reconnaissant des nombres de cibles différents ou d'affinité différentes. Cela permettrait aussi de tester l'hypothèse d'alternance entre la Reine Rouge neutre et la Reine Rouge sélective présentée plus haut.

Un dernier paramètre sur lequel on pourrait jouer serait de faire évoluer le nombre de DSB réalisées pendant la méiose. En effet, plus on fait de DSB, plus on a de chance a priori d'apparier les chromosomes homologues. Cependant, un trop fort taux de DSB est désavantageux et dangereux, car les DSB peuvent donner lieu de des réarrangements ou des pertes de morceaux de chromosomes, ce qui est à terme contre productif. En lien avec cette évolution du nombre de DSB, et comme testé d'une manière assez simplifiée dans notre modèle, on pourrait aussi implémenter une réalisation progressive et régulée des DSB à travers le génome, ce qui mènerait soit à un succès de la méiose dans un cas de DSB dans une liaison symétrique, soit à l'échec de la méiose après la réalisation d'un certain nombre maximal de cassures. Ce modèle aurait donc l'avantage de déterminer le nombre de DSB optimal réalisé pendant une méiose en fonction du niveau d'érosion de l'allèle, tout en gardant à l'esprit que le nombre de DSB maximal réalisable par méiose

est fixe et indépendant de *PRDM9*.

4.5.3 Pourquoi *PRDM9* pour la recombinaison ?

En 2009 a été découvert le rôle de *PRDM9* dans la triméthylation des histones proches d'un motif ADN précis. Ces changements locaux aboutissent au recrutement de la protéine SPO11 pour la formation des cassures double brins, et ce majoritairement en dehors des promoteurs des gènes et îlots CpG. De plus, certaines lignées de souris dont le phénotype a été modifié pour invalider le gène *PRDM9* (souris knock-out ou $PRDM9^{-/-}$) montrent des forts taux d'asynapsie et un arrêt méiotique complet [162]. Par ailleurs, chez ces souris mutantes, la recombinaison est redirigée vers les sites enrichis en H3K4me3 non médié par PRDM9, incluant les promoteurs des gènes et d'autres sites fonctionnels [34; 162; 163]. Ces régions correspondent aux points chauds de recombinaison chez des espèces telles que la levure, les oiseaux ou le chien, qui n'ont pas *PRDM9*, sans que cela ne suffise forcement pour restaurer le bon fonctionnement de la méiose. À partir de ces observations, l'idée a été proposée que le rôle de *PRDM9* était avant tout de rediriger la recombinaison en dehors des gènes [34].

Or, cette hypothèse, qui présuppose qu'il est plus viable de recombiner en dehors des promoteurs de gènes, possède plusieurs contre-arguments. En effet, de nombreuses espèces eukaryotes ont un système de recombinaison "par défaut", le même retrouvé chez les souris knock-out, qui cible les DSBs (et donc crée des hot spots de recombinaison) dans les sites d'initiation de la transcription. Ces espèces sont celles présentées dans l'introduction comme espèces à paysage de recombinaison stable. On peut aussi noter le cas particulier d'une femme qui, bien qu'homozygote pour un allèle nul de *PRDM9*, est tout à fait fertile et possède une descendance viable et fertile [163]. Ces espèces ou individus ne semblent pas avoir de problème particulier avec la non présence de *PRDM9*. De plus, on a pu observer chez plusieurs espèces une perte de fonctionnalité de *PRDM9*, voire même une perte totale du gène, sans que cela ne semble poser de problème particulier [43; 45].

Le travail de modélisation réalisé pendant cette thèse permet d'aboutir à une autre manière de voir *PRDM9*. De part sa capacité à lier l'ADN, le ramener sur l'axe chromosomique et recruter SPO11, *PRDM9* fonctionne comme un facilitateur très efficace de l'appariement des chromosomes. Cette hypothèse, qui avait d'abord été proposée par Davies *et al.* en 2016 [111], représente un changement de paradigme très intéressant.En effet, cette nouvelle vision des choses permet d'expliquer à la fois la stérilité hybride et le moteur sélectif de la Reine Rouge en intra-population. En substance, ma contribution principale au domaine de recherche de l'évolution de la recombinaison méiotique dépendante de*PRDM9*, est la modélisation prenant à la fois en compte les mécanismes moléculaires de la recombinaison méiotique et les principes de la génétique des populations afin de faire le lien entre les phénomènes observés en populations unique et en contexte hybride.

Dans ce contexte, le mécanisme de facilitation de la recombinaison par la liaison

symétrique suffit à donner une explication de l'existence de *PRDM9*. En effet, le facteur limitant de la méiose semble être très clairement la question de l'appariement des chromosomes homologues, mécanisme extrêmement délicat, surtout quand les génomes sont gros. Afin de répondre à cet enjeu de taille, les eucaryotes semblent avoir inventé toute une série de mécanismes moléculaires et cellulaires qui visent à augmenter les chances d'un appariement correct menant à une bonne synapse, et donc à une bonne ségrégation de chromosomes homologues dans les cellules filles. Ainsi, le rôle de *PRDM9* dans la recombinaison semble être l'un de ces systèmes. Ce système s'avère être efficace, mais il est tout de même soumis au paradoxe de conversion des points chauds [83], du fait même de son action de recrutement de la machinerie de cassures double brins au niveau de la séquence ADN reconnue par PRDM9. Cela a pour conséquence la dynamique de Reine Rouge, hypothèse actuellement prédominante pour la résolution du paradoxe des points chauds dépendant de PRDM9, ce qui implique une déficience cyclique de *PRDM9*.

Pour conclure, au vu de ce qui a été dit précédemment, il devient possible de spéculer sur les raisons des pertes récurrentes de *PRDM9*, et de les modéliser en améliorant le modèle qui a été développé pendant ma thèse. En effet, il est tout a fait faisable de développer une nouvelle version de mon modèle dans laquelle l'ensemble des autres mécanismes de recombinaison visant à augmenter l'appariement des chromosomes homologues seraient représentés par une probabilité non nulle de réaliser l'appariement des chromosomes en l'absence de liaison symétrique de PRDM9. Du fait que ces mécanismes alternatifs sont eux-mêmes susceptibles d'évoluer, cette probabilité pourrait elle-aussi évoluer. Notamment, lorsque cette probabilité devient grande, les déficiences récurrentes de *PRDM9* liées à la Reine Rouge pourraient aboutir à sa perte occasionnelle [53], comme observé, par exemple, chez les canidés (chez qui *PRDM9* est devenu un pseudo-gène [97] suite à une perte de fonctionnalité assez récente), ou à plus grande échelle évolutive, chez les oiseaux [45].

Partie IV
Appendices

5

Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis - Supplementary Materials

Contents

Supplementary materials

5.1 Derivation of the analytical approximations

5.1.1 Introduction

In this document, we give the detailed derivation of the analytical approximation for the stationary regime of the Red Queen process. The derivation proceeds along similar lines as in Latrille *et al.* [96]. It relies on a self-consistent mean field argument. In brief, the derivation assumes the following hypotheses :

- Wright-Fisher model (*i.e.* non-overlapping generations) with mutation and selection
- Panmictic population (random mating)
- Constant population size N
- Highly polymorphic model
- Weak erosion
- Strong selection implying that genetic drift can be ignored

The first four hypotheses are entailed by the simulator. The last three are additional assumptions that are made in order to make the analytical derivation feasible. The analytical results will therefore be valid only in regimes in which these three assumptions are met.

Based on these hypotheses, we first express the frequency $f(t)$ and the proportion of active site , $\theta(t)$ through time of a typical *PRDM9* allele. We will obtain that the erosion level of an allele is $z = 1-\theta$ and thus $\bar{z} = 1-\bar{\theta}$, the mean erosion level of the population. The frequency trajectory of a typical allele depends on the mean erosion level of the population $\bar{z} = 1 - \bar{\theta}$. Relying on the argument that the population itself is composed of such typical *PRDM9* alleles successively invading, with a mean time interval between successive invasions noted τ, itself depending on the mean fitness of the population, we can obtain a self-consistent expression for $\bar{z}$ (or equivalently, for τ), from where we can then express all summary statistics of interest.

5.1.2 Analytical developments

Studied model

First, assuming negligible random drift, the evolution of the frequency through time $f(t)$ of a typical allele is deterministic and is given by:

$$\frac{df}{dt} = \frac{w^* - \bar{w}}{\bar{w}} f, \tag{5.1}$$

where w^* is the mean fertility of the allele at time t and

$$\bar{w} = \sum_i f_i w_i^* \tag{5.2}$$

is the mean fertility of all *PRDM9* alleles in the population. This equation means that selection will depend on differential erosion.
In turn, erosion depends on the frequency trajectory of the allele. To decouple these effects, we start by observing that the fraction of active target sites for the allele varies approximately as :

$$\frac{d\theta}{dt} \approx -\rho f \theta \tag{5.3}$$

In this equation, ρ is the erosion rate per generation. It can be expressed as :

$$\rho = (2Nv)(2g) \tag{5.4}$$

where $g = \frac{d}{8h}$ is the mean gene conversion rate (*i.e.* the probability that a site undergoes a DSB and a repair). Thus, ρ is equal to :

$$\rho = \frac{Nvd}{2h} \tag{5.5}$$

Equation (5.3) expresses the fact that the rate of extinction of target sites is equal to the mutation rate at the level of the population ($2Nv$) multiplicated by the probability of fixation of the inactive mutant. Assuming strong gene conversion ($4Ng >> 1$), this probability is well approximated by twice the conversion rate, *i.e.* $2fg$. Note that the rate of erosion given by (5.4) is only approximate. A more accurate expression accounting for the affinity distribution of the sites will be given below.

The intrinsic age of an allele (z)

Therefore,

$$z(t) = \rho \int_0^t f(u)\, du \tag{5.6}$$

can be seen as a measure of the cumulative erosion level of an allele. As such, it can be used as a measure of the intrinsic age of the allele. We can then express the equation describing the evolution of f and θ directly as a function of z, instead of t. This change of variable will entail the following factor :

$$\frac{dz}{dt} = \rho f. \tag{5.7}$$

As mentioned above, our derivation assumes weak erosion. Mathematically, this translates into the assumption that $z << 1$.

Frequency of an allele (f) depending on its age (z) : f(z)

Then, we want to express the evolution of the frequency of an allele in the population as a function on its intrinsic age z:

$$\frac{df}{dz} = \frac{df}{dt}\frac{dt}{dz} = \left[\frac{w^*(z) - \bar{w}}{\bar{w}}\right] . \left[\frac{1}{\rho}\right] = \frac{1}{\rho}\left(\frac{w^*(z)}{\bar{w}} - 1\right) \tag{5.8}$$

However, $\frac{w^*}{\bar{w}}$ is extremely close to 1, so $\frac{w^*}{\bar{w}} - 1$ is close to 0. In addition, we know that, in the vicinity of 0, $x \approx ln(1 + x)$. As a result, we can write

$$\frac{1}{\rho}\left(\frac{w^*(z)}{\bar{w}} - 1\right) \approx \frac{1}{\rho} ln\left(1 + \left(\frac{w^*(z)}{\bar{w}} - 1\right)\right) = \frac{1}{\rho} ln\left(\frac{w^*(z)}{\bar{w}}\right) \tag{5.9}$$

Working in the weak erosion limit ($z << 1$) allows us to linearize $ln(w)$ in the vicinity of 0 :

$$ln\left(\frac{w^*(z)}{\bar{w}}\right) = ln(w^*(z)) - ln(\bar{w}) \approx ln(w(0,0)) - \frac{\alpha}{2}(z+\bar{z}) - ln(w(0,0)) + \frac{\alpha}{2}(\bar{z}+\bar{z}) = -\frac{\alpha}{2}(z-\bar{z}) \tag{5.10}$$

where:

$$\alpha = \left| \frac{\partial ln(w)}{\partial z} \right|_{(z=0)} \tag{5.11}$$

is the slope at the origin of $ln(w)$. It depends on the mechanistic parameters, in a way that will be determined further below.
Then, by replacement in equation (5.8) :

$$\boxed{\frac{df}{dz} \approx -\frac{\alpha}{2\rho}(z - \bar{z})} \tag{5.12}$$

Integrating equation (5.12), with the constraint that $f(0) = 0$, gives :

$$f(z) = -\frac{\alpha}{4\rho} z[z - 2\bar{z}]. \tag{5.13}$$

The function f has the shape of a concave parabola and $f(z) = 0$ when $z = 0$ and $z = 2\bar{z}$.

Mean age of an allele in the population ($\bar{z}$) : a self-consistent derivation

Equations (5.12) and (5.13) depend on $\bar{z}$, and thus we now need to express $\bar{z}$ as a function of the model parameters.
The mean allele age of the population $\bar{z}$ is equal to :

$$\bar{z} = \sum_i (f_i z_i). \tag{5.14}$$

Relying on a tiling argument (Latrille *et al.*, 2017 [96]) this can also be expressed as :

$$\bar{z} = \frac{1}{\tau} \int_0^{+\infty} f(t) z(t)\, dt. \tag{5.15}$$

Equation (5.15) expresses the idea that, at stationarity, the allele frequency distribution at a given time point (eq. (5.14)) is equivalent to the distribution of frequencies at which a typical allele has segregated over its entire life normalized by the mean waiting time τ between successive invasions (see Latrille *et al.*, 2017 figure 6 [96]).
We then do a change of variable from t to z in the integrand:

$$\bar{z} = \frac{1}{\tau} \int_0^{z(\infty)} z f \frac{dt}{dz}\, dz = \frac{1}{\tau} \int_0^{z(\infty)} z f \left(\frac{dz}{dt} \right)^{-1} dz, \tag{5.16}$$

then, we replace $\frac{dz}{dt}$ by its expression in equation (5.3) :

$$\bar{z} = \frac{1}{\tau} \int_0^{z(\infty)} z f \frac{1}{f\rho}\, dz = \frac{1}{\tau\rho} \int_0^{z(\infty)} z\, dz = \frac{1}{\tau\rho} \frac{z(\infty)^2}{2}. \tag{5.17}$$

In order to determine $z(\infty)$, we express the constrain that $\sum_i (f_i) = 1$. This expression can also be written as :

$$1 = \sum_i (f_i) = \frac{1}{\tau} \int_0^{\infty} f(t)\, dt = \frac{1}{\tau} \int_0^{z(\infty)} f \frac{dt}{dz}\, dz = \frac{1}{\rho\tau} \int_0^{z(\infty)} f \frac{1}{f}\, dz = \frac{1}{\rho\tau} \int_0^{z(\infty)} dz = \frac{z(\infty)}{\rho\tau} \tag{5.18}$$

So

$$z(\infty) = \rho\tau, \tag{5.19}$$

such that, by replacement :

$$\bar{z} = \frac{\rho\tau^2}{2\rho\tau} = \frac{\rho\tau}{2}. \tag{5.20}$$

We now need to express τ, which is the inverse of the invasion rate of a new allele in the population. The rate of invasion is equal to the rate of mutation at the population level ($2Nu$) multiplicated by the invasion probability. Assuming strong selection, this probability is well approximated by $2s_0$.

$$\tau^{-1} = (2Nu).(2s_0) \tag{5.21}$$

where u is the mutation rate at the Prdm9 locus and s_0 is the selection coefficient of a new allele in the population. Based on equation (5.10), s_0 can be expressed as $s_0 = \frac{\alpha}{2}\bar{z}$. Thus :

$$\tau^{-1} = 4Nu\frac{\alpha}{2}\bar{z} = \mu\frac{\alpha}{2}\bar{z}, \tag{5.22}$$

Where $\mu = 4Nu$. If we replace this in equation (5.20) we finally obtain :

$$\bar{z} = \frac{\rho}{2\tau^{-1}} = \frac{\rho}{2}\frac{2}{\mu\alpha\bar{z}} = \frac{\rho}{\mu\alpha\bar{z}} \iff \bar{z}^2 = \frac{\rho}{\mu\alpha} \iff \boxed{\bar{z} = \sqrt{\frac{\rho}{\mu\alpha}}} \tag{5.23}$$

We thus recover the main result of the derivation given in Latrille *et al.* [96], although now, the compound parameters ρ and α depend on the mecanistic details of our model. We have already seen that $\rho = \frac{Nvd}{2h}$. On the other hand, we still need to express α.

Slope at the origin of the fertility rate : α

As mentioned above, α is the slope of $ln(w)$ at the origin (*i.e.* for two new alleles of age $z_1 = z_2 = 0$):

$$\alpha = \left|\frac{\partial \ln(w(z_1, z_2))}{\partial z_1}\right|_{(z_1=0,z_2=0)} = \left|\frac{1}{w}\frac{\partial w(z_1, z_2)}{\partial z_1}\right|_{(z_1=0,z_2=0)} = \left|\frac{\partial w(z_1, z_2)}{\partial z_1}\right|_{(z_1=0,z_2=0)}, \tag{5.24}$$

since $w \approx 1$.

Thus, in order to find an explicit expression for α, we need to express w according to the parameters.

Assuming that gametes are not limiting, the fitness of an individual is equal to the rate of success of meiosis, which is itself equal to the probability of having at least one DSB in a symmetrical bound site. So, we can write w as 1 - the probability of having no DSB in symmetrical bound site (1 - probability of failure of the meiosis). The number of DSBs in symmetrical bound sites is approximately Poisson of mean $dq(z_1, z_2)$, where d is the mean number of DSBs and $q(z_1, z_2)$ is the probability that a DSB occurs in a symmetrically bound site in an individual heterozygous for two *PRDM9* alleles of age

z_1 and z_2. Thus, the probability of 0 DSB in a symmetrically bound site is $e^{-dq(z_1,z_2)}$. So, $w(z_1, z_2)$ can be expressed as :

$$w(z_1, z_2) = 1 - e^{-dq(z_1,z_2)} \tag{5.25}$$

Substituting in equation (5.24), we obtain :

$$\alpha = \left| \frac{\partial w(z_1, z_2)}{\partial z_1} \Big|_{(z_1=0,z_2=0)} \right| = \left| d \, \frac{\partial q(z_1, z_2)}{\partial z_1} \Big|_{(z_1=0,z_2=0)} e^{-dq(z_1=0,z_2=0)} \right| \tag{5.26}$$

We see here that, in order to obtain an explicit formula for α, it is necessary to express $q(z_1, z_2)$ as a function of the model parameters and then compute its derivative as a function of z.

Probability of symmetrical binding : q

For a given site of affinity y, the probability that PRDM9 is bound is given by $x = \frac{cy}{1+cy}$ where $c = 1$ or $c = 2$. The conditional probability of symmetrical binding at that site (conditional on at least one of the four chromatids being bound) is then equal to :

$$q = \frac{2x^2 - x^3}{x} \tag{5.27}$$

This expression is valid for a single site. To compute the mean probability of symmetrical binding over the genome, we need to average the numerator and the denominator separately over the affinity distribution across sites. Of note this distribution itself depends on the age z of the allele, and the mean over the distribution of a given function $B(y)$ is noted $< B >_z$. Also, the mean q over the genome depends on the *Prdm9* genotype of this individual. In the case of a homozygote, thus possessing twice the same *Prdm9* allele of age z, q is equal to:

$$q^{hom}(z) = \frac{2 < x^2 >_z - < x^3 >_z}{< x >_z} \tag{5.28}$$

If the individual is heterozygous for *Prdm9* with two alleles of age z_1 and z_2, q is equal to:

$$q^{het}(z_1, z_2) = \frac{2 < x^2 >_{z_1} - < x^3 >_{z_1} + 2 < x^2 >_{z_2} - < x^3 >_{z_2}}{< x >_{z_1} + < x >_{z_2}} \tag{5.29}$$

The mean over the affinity distribution can be more precisely expressed as :

$$< B >_z = \int B(y)\theta_y(z)\varphi(y)\, dy. \tag{5.30}$$

In this equation $\varphi(y)$ is the affinity distribution of an allele at birth and $\theta_y(z)$ is the fraction of active sites with a given affinity y recognised by an allele of age z. Thus, $\theta_y(z)\varphi(y)dy$ is the total number of target sites still active with an affinity $y \pm dy$.
The integrals of the form (5.30) can be obtained numerically. However, they depend on $\theta_y(z)$, which we therefore need to determine.

Fraction of active sites with a given affinity y recognised by an allele of an age z : $\theta_y(z)$

Here our aim is to compute more precisely the proportion of target sites of a given affinity y that are still active, for an allele of age z. We note this quantity $\theta_y(z)$

By an argument similar to that used for equation (5.4):

$$\frac{d\theta_y}{dt} = -(2Nv).(2fg_y^{het})\theta_y \tag{5.31}$$

where $g_y^{het} = g_y^{het}(z_1, z_2)$ is the gene conversion rate at sites of affinity y in a genotype (z_1, z_2). Of note, we consider only heterozygotes for *PRDM9* since we work under the assumption of a highly polymorphic regime. In turn :

$$g_y^{het} = d.\frac{\frac{cy}{1+cy}}{4h\left[\left\langle\frac{cy}{1+cy}\right\rangle_{z_1} + \left\langle\frac{cy}{1+cy}\right\rangle_{z_2}\right]} \tag{5.32}$$

The fraction on the right-hand side of equation (5.32) is the proportion of sites of affinity y among all sites that are bound by either one of the two *PRDM9* alleles.

So, if we replace it in equation (5.31), we obtain :

$$\frac{d\theta_y}{dt} = -2Nv.d\frac{\frac{cy}{1+cy}}{4h\left[\left\langle\frac{cy}{1+cy}\right\rangle_{z_1} + \left\langle\frac{cy}{1+cy}\right\rangle_{z_2}\right]}.2f(t)\theta_y(z) = -Nv\frac{d}{h}.\frac{\frac{cy}{1+cy}}{\left[\left\langle\frac{cy}{1+cy}\right\rangle_{z_1} + \left\langle\frac{cy}{1+cy}\right\rangle_{z_2}\right]}.f(t)\theta_y(z). \tag{5.33}$$

If we suppose that $<\frac{cy}{1+cy}>_{z_1}$ and $<\frac{cy}{1+cy}>_{z_2}$ don't vary too much as a function of z ($\approx<\frac{cy}{1+cy}>_0$), then we have:

$$\frac{d\theta_y}{dt} = -Nv\frac{d}{h}\frac{\frac{cy}{1+cy}}{2\left\langle\frac{cy}{1+cy}\right\rangle_0}.f(t)\theta_y(z) \tag{5.34}$$

Then, by setting $\rho = \frac{Nvd}{2h}$ we can re-express the equation (5.37) as

$$\frac{d\theta_y}{dt} = -\rho\frac{\frac{cy}{1+cy}}{\left\langle\frac{cy}{1+cy}\right\rangle_0}.f(t)\theta_y(z) \tag{5.35}$$

And at this stage we can pose $\gamma(y) = \frac{\frac{cy}{1+cy}}{\left\langle\frac{cy}{1+cy}\right\rangle_0}$, which gives the following equation:

$$\frac{d\theta_y}{dt} = -\rho f(t)\gamma(y)\theta_y. \tag{5.36}$$

Now, we can replace $\rho f(t)$ by $\frac{dz}{dt}$:

$$\frac{d\theta_y}{dt} = -\frac{dz}{dt}\gamma(y)\theta_y \Leftrightarrow \frac{d\theta_y(z)}{dz} = -\gamma(y)\theta_y \tag{5.37}$$

Finally if we integrate this equation, we obtain the fraction of active sites with an affinity y recognized by an allele of age z :

$$\theta_y(z) = \theta_y(0)e^{-\gamma(y)z} \tag{5.38}$$

By definition, $\theta_y(0)$ is the fraction of active sites with an affinity y recognized by an allele of age 0. Thus $\theta_y(0) = 1$, and as a result, equation (5.34) becomes

$$\boxed{\theta_y(z) = e^{-\gamma(y)z}} \tag{5.39}$$

Under the condition of weak erosion, we can simplify equation (5.39) which gives

$$\theta_y(z) \approx 1 - \gamma(y)z \tag{5.40}$$

and by averaging over the affinity, we get

$$< \theta_y(z) > \approx 1 - z \tag{5.41}$$

Expression of α

Thanks to this equation, we are now able to express the successive moments $< x^m >_z$ for $m = 1$, 2 and 3.

$$< x^m >_z = \int \left(\frac{cy}{1+cy}\right)^m \theta_y(z)\varphi(y)\,dy = \int \left(\frac{cy}{1+cy}\right)^m e^{-\gamma(y)z}\varphi(y)\,dy \tag{5.42}$$

To express α, we need $q(0,0)$ and $\frac{\partial q(z_1,z_2)}{\partial z_1}\Big|_{z_1=0,z_2=0}$.
For $q(0,0)$ we have the following expression

$$q(0,0) = \frac{2 < x^2 >_0 - < x^3 >_0 + 2 < x^2 >_0 - < x^3 >_0}{< x >_0 + < x >_0} = \frac{2 < x^2 >_0 - < x^3 >_0}{< x >_0} \tag{5.43}$$

with $< B >_0 = \int B(y)\varphi(y)$.

And for $\frac{\partial q(z_1,z_2)}{\partial z_1}\Big|_{z_1=0,z_2=0}$, we first need to determine the general expression of $\frac{\partial < B >_z}{\partial z}$ and then express all the moments $< x^n >$ that we need.

$$\frac{\partial < B >_z}{\partial z} = \int B(y)\varphi(y)(-\gamma(y))e^{\gamma(y)z}\,dy = - < \gamma(y)B >_z \tag{5.44}$$

And with $\gamma(y) = \frac{\frac{cy}{1+cy}}{\left\langle \frac{cy}{1+cy} \right\rangle_0} = \frac{x}{<x>_0}$ we obtain :

$$\frac{\partial < B >_z}{\partial z} = -\frac{< xB >_z}{< x >_0} \tag{5.45}$$

So

$$\frac{\partial < B >_z}{\partial z}\Big|_{z=0} = -\frac{< xB >_0}{< x >_0} \tag{5.46}$$

In particular, with $B = x^n$ we have

$$\frac{\partial < x^n >_z}{\partial z}\Big|_{z=0} = -\frac{< x^{n+1} >_0}{< x >_0} \tag{5.47}$$

If we come back to the expression of $\frac{\partial q(z_1,z_2)}{\partial z_1}\Big|_{z_1=0,z_2=0}$, we obtain

$$\beta = \frac{\partial q(z_1,z_2)}{\partial z_1}\Big|_{z_1=0,z_2=0} = \frac{\left[-2\frac{<x^3>_0}{<x>_0} + \frac{<x^4>_0}{<x>_0}\right] 2 < x >_0 + \frac{<x^2>_0}{<x>_0}(4 < x^2 >_0 - 2 < x^3 >_0)}{4 < x >_0^2} \tag{5.48}$$

Replacing in the equation (5.26) and after some simplifications, we obtain the final expression of α

$$\alpha = \left| d \left[\frac{-2 < x^3 >_0 + < x^4 >_0 + \frac{2 < x^2 >_0^2}{< x >_0} - \frac{< x^3 >_0 < x^2 >_0}{< x >_0}}{2 < x >_0^2} \right] e^{-d \frac{2 < x^2 >_0 - < x^3 >_0}{< x >_0}} \right| \tag{5.49}$$

which can also be written as follows

$$\boxed{\alpha = \left| d\beta e^{-dq} \right|} \tag{5.50}$$

5.1.3 Summary statistics

Based on these analytical developments, we can now obtain analytical expressions for the summary statistics.

Diversity : D

The *Prdm9* diversity, written D, is defined as

$$D = \left(\sum_i f_i^2 \right)^{-1} \tag{5.51}$$

We now need to calculate $\sum_i f_i^2$. For that, we can use again the tiling principle :

$$\sum_i f_i^2 \approx \frac{1}{\tau} \int_0^\infty f^2(t)\, dt = \frac{1}{\tau} \int_0^{z(\infty)} f^2(z) \frac{dt}{dz}\, dz = \frac{1}{\tau} \int_0^{z(\infty)} f^2(z) \frac{1}{\rho f(z)}\, dz = \frac{1}{\rho\tau} \int_0^{z(\infty)} f(z)\, dz \tag{5.52}$$

Replacing f(z) by its expression from equation (5.13) and integrating over z gives :

$$\frac{1}{\rho\tau} \int_0^{z(\infty)} -\frac{\alpha}{4\rho} z(z - 2\bar{z})\, dz = -\frac{\alpha}{4\rho^2\tau} \left[\int_0^{z(\infty)} z^2\, dz - 2\bar{z} \int_0^{z(\infty)} z\, dz \right] - \frac{\alpha}{4\rho^2\tau} \frac{z(\infty)^3}{3} + \frac{\alpha\bar{z}}{2\rho^2\tau} \frac{z(\infty)^2}{2} \tag{5.53}$$

But we know that $\bar{z} = \frac{\rho\tau}{2}$ and $z(\infty) = \rho\tau$. So if we replace it in the equation (5.53)

$$-\alpha \left[\frac{\rho^3\tau^3}{12\rho^2\tau} - \frac{\rho^3\tau^3}{8\rho^2\tau} \right] = -\alpha \left[\frac{\rho\tau^2}{12} - \frac{\rho\tau^2}{8} \right] = -\alpha \left[\frac{2\rho\tau^2}{24} - \frac{3\rho\tau^2}{24} \right] \tag{5.54}$$

So we obtain

$$\sum_i f_i^2 = \frac{\alpha\rho\tau^2}{24} \tag{5.55}$$

But, we also know that $\tau = \frac{2}{\mu\alpha\bar{z}}$ and $\bar{z} = \sqrt{\frac{\rho}{\mu\alpha}}$. So then $\tau = \frac{2}{\mu\alpha} \sqrt{\frac{\mu\alpha}{\rho}}$ and $\tau^2 = \frac{4}{\mu\alpha\rho}$. Which leads to the equation

$$\sum_i f_i^2 = \frac{4\alpha\rho}{24\mu\alpha\rho} = \frac{1}{6\mu} \tag{5.56}$$

Where $\mu = 4Nu$, so finally we have

$$\sum_i f_i^2 = \frac{1}{24Nu} \tag{5.57}$$

And thus

$$\boxed{D = 24Nu} \tag{5.58}$$

Mean age : $\bar{z}$

Using equation (5.50), $\bar{z}$ can be more directly expressed as :

$$\boxed{\bar{z}} = \sqrt{\frac{\rho}{\mu\alpha}} = \sqrt{\frac{Nvd}{2h}\frac{1}{4Nu}\frac{e^{dq}}{d\beta}} = \boxed{\sqrt{\frac{ve^{dq}}{8hu\beta}}} \tag{5.59}$$

with q given by the equation (5.43)

Mean activity : $< \bar{\theta} >$

If we start from the equation (5.39) replacing z by $\bar{z}$ we have :

$$\overline{\theta_y(z)} = \theta_y(\bar{z}) = e^{-\gamma(y)\bar{z}} \approx 1 + (-\gamma(y)\bar{z}) \approx 1 - \gamma(y)\bar{z} \tag{5.60}$$

Averaging over the affinity distribution :

$$\boxed{< \theta_y(\bar{z}) > \approx 1 - \bar{z}} \tag{5.61}$$

Mean probability of symmetrical binding : $\bar{q}$

If we start from the equation (5.29) replacing z by $\bar{z}$ we have :

$$\boxed{\bar{q}} = q^{het}(\bar{z},\bar{z}) = \frac{2 < x^2 >_{\bar{z}} - < x^3 >_{\bar{z}} + 2 < x^2 >_{\bar{z}} - < x^3 >_{\bar{z}}}{< x >_{\bar{z}} + < x >_{\bar{z}}} = \boxed{\frac{2 < x^2 >_{\bar{z}} - < x^3 >_{\bar{z}}}{< x >_{\bar{z}}}} \tag{5.62}$$

The moments $< x^m >_{\bar{z}}$ are given by equation (5.42) an can be evaluated by numerical integration.

Mean fertility rate : $\bar{w}$

If we start from the equation (5.25) replacing z by $\bar{z}$ we have :

$$\boxed{\bar{w} = w(\bar{z},\bar{z}) = 1 - e^{-dq(\bar{z},\bar{z})}} \tag{5.63}$$

With $\bar{q} = \frac{2<x^2>_{\bar{z}} - <x^3>_{\bar{z}}}{<x>_{\bar{z}}}$.

5.1.4 Perturbative development accounting for genetic dosage

The evolution of the frequency of an allele in the population as a function of its age has the same general expression as without dosage :

$$\frac{df}{dz} = \frac{1}{\rho}\left(\frac{w^*(z) - \bar{w}}{\bar{w}}\right) \tag{5.64}$$

However, for the fitness, we now account for the contribution of homozygotes:

$$w^*(z) = f(t)w^{hom}(z) + (1 - f(t))w^{het}(z,\bar{z}) \tag{5.65}$$

Linearizing in the vicinity of 0 for z:

$$w^*(z) = fw^{hom}(z)+(1-f)w^{het}(z,\bar{z}) \approx fw^{hom}(0)(1-\alpha^{hom}z)+(1-f)w^{het}(0,0)(1-\frac{\alpha^{het}}{2}(z+\bar{z})) \tag{5.66}$$

Our development is perturbative in the sense that it assumes that gene dosage has a weak impact, and this, because homozygotes are assumed to be rare, i.e. because f is small ($f << 1$). Combined with the weak erosion assumption ($\bar{z} << 1$), this means that we can ignore terms of the order of zf. Thus, equation (5.66) simplifies to :

$$w^*(z) \approx fw^{hom}(0) + (1-f)w^{het}(0,0) - w^{het}(0,0)\frac{\alpha^{het}}{2}(z+\bar{z}) \tag{5.67}$$

Averaging over the population gives the mean fitness:

$$\bar{w} \approx \bar{f}w^{hom}(0) + (1-\bar{f})w^{het}(0,0) - w^{het}(0,0)\frac{\alpha^{het}}{2}(\bar{z}+\bar{z}) \tag{5.68}$$

Finally :

$$\frac{df}{dz} = \frac{1}{\rho}\left(\frac{w^*(z)-\bar{w}}{\bar{w}}\right) \approx \frac{1}{\rho}\left[\frac{w^{hom}(0)-w^{het}(0,0)}{w^{het}(0,0)}(f-\bar{f}) - \frac{\alpha^{het}}{2}(z-\bar{z})\right] \tag{5.69}$$

If we express $\sigma(0) = \frac{w^{hom}(0)-w^{het}(0,0)}{w^{het}(0,0)}$, we obtain :

$$\boxed{\frac{df}{dz} \approx \frac{1}{\rho}\left[\sigma(0)(f-\bar{f}) - \frac{\alpha^{het}}{2}(z-\bar{z})\right]} \tag{5.70}$$

5.2 Measure of Prdm9 diversity in Mus musculus sub-species

5.2.1 Introduction

Kono *et al.* (2014) genotyped the *Prdm9* ZF-encoding exon in 105 individuals, from 4 subspecies of mice (see Table S2 in [69]). In total, they identified 56 alleles encoding different ZF arrays. To analyze *Prdm9* diversity in natural populations, we focused on data from wild-captured mice (total=79 individuals). Kono *et al.* classified *Prdm9* alleles in 12 clusters (Ca, Cb, Cc, Cd, Ce, Da, Db, Dc, Dd, Ma, Mb, t). It should be noted that some of these alleles that belong to a same cluster share a large fraction of their target sites. For instance, *Prdm9* allele C3H (Da1 in [69]) is closely related to B6 (Da2), and these two alleles share ˜30% of their targets ([137]). Similarly, PWD (Ma7) and MOL, which both belong to the Ma cluster share about 13% of their targets ([137]). Thus, the total *Prdm9* allelic diversity probably over-estimates the functional diversity (in terms of *Prdm9* target sites).
We therefore computed two estimates of *Prdm9* diversity:

- Dmax: diversity measured by considering all *Prdm9* alleles individually
- Dmin: diversity measured after grouping of *Prdm9* alleles per cluster

5.2.2 Summary

Species	NbIndividuals	NbAlleles	Heterozygous	NbClusters	Dmax	Dmin
M_m_molossinus	10	3	20.0%	2	1.80	1.60
M_m_domesticus	20	12	10.0%	4	6.30	2.24
M_m_castaneus	24	23	50.0%	7	12.52	5.82
M_m_musculus	25	27	48.0%	7	18.12	2.53

The number of individuals genotyped in *M. m. molossinus* is too small to estimate *Prdm9* diversity. In the three other subspecies, the total *Prdm9* diversity ranges from Dmax=6.3 to Dmax=18.12. If we consider the *Prdm9* diversity in terms of clusters of alleles, then diversity ranges from Dmin=2.24 to Dmin=5.82.

The true functional *Prdm9* diversity (in terms of target sites) is most probably between Dmin and Dmax.

5.3 Supplementary figures

S1 Fig. A simulation trajectory under a polymorphic regime ($u = 5 \times 10^{-4}$ and $v = 5 \times 10^{-5}$) with same scale as Fig3 in the main text. In all panels, each color corresponds to a different allele. Note that a given color can be reassigned to a new allele later in the simulation. Successive panels represent the variation through time of (A) the frequency of each *PRDM9* allele and its corresponding (B) the proportion of active sites, (C) the mean affinity of active sites, (D) the probability of symmetrical binding and (E) the fertility. The thick line singles out the trajectory of a typical allele.

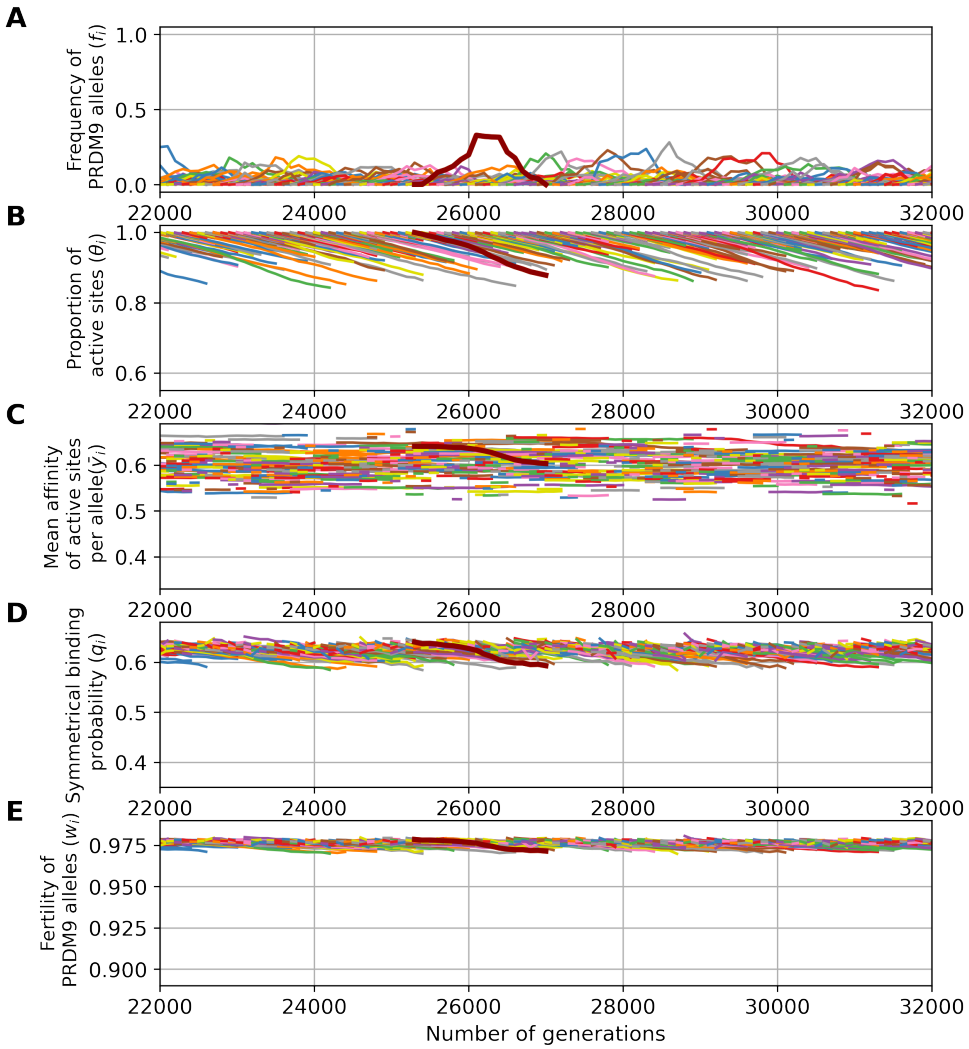

Figure 5.1

S2 Fig. Two simulation trajectories under the control model allowing for chromosome pairing and success of meiosis without requiring symmetrical binding of PRDM9. (A) and (B) correspond to the monomorphic regime (as in Fig 3, $\mathbf{u = 5 \times 10^{-6}}$ and $\mathbf{v = 5 \times 10^{-5}}$), while (C) and (D) correspond to the polymorphic regime (as in Fig 4, $\mathbf{u = 5 \times 10^{-4}}$ and $\mathbf{v = 5 \times 10^{-5}}$). In all panels, each color corresponds to a different allele. Note that a given color can be reassigned to a new allele later in the simulation. Successive panels represent the variation through time of (A) and (C) the frequency of each *PRDM9* allele and (B) and (D) its corresponding proportion of active sites.

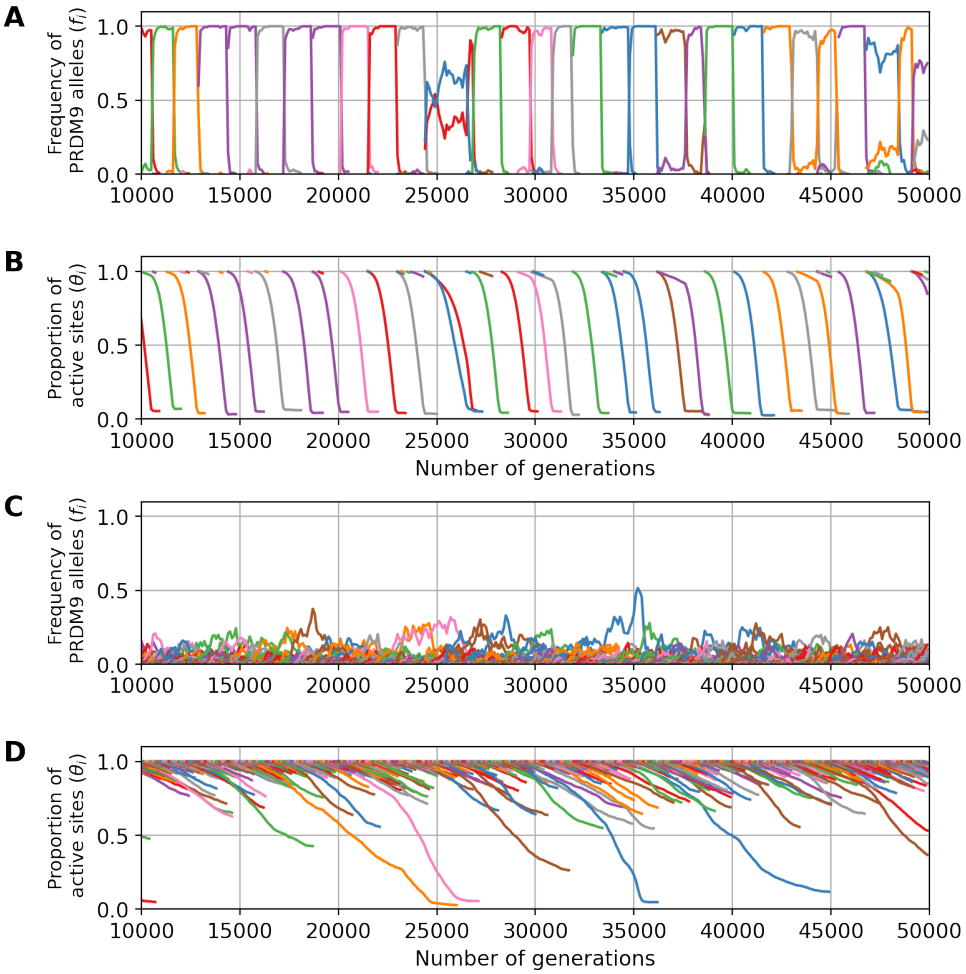

Figure 5.2

S3 Fig. Scaling experiment showing the conditions for a change of regime (from polymorphic to monomorphic) upon introducing gene dosage. A and B: equilibrium levels of *PRDM9* diversity without (A) and with (B) gene dosage. C and D: summary statistics that are predictive of the type of regime observed in the presence of dosage, namely: (C) $\sigma_0\tau$ (selection coefficient associated to dosage multiplied by the time between successive invasions by new *PRDM9* alleles), and (D) $4Nu$. The regime is predicted to be polymorphic in the presence of dosage if $4Nu > 10$ and $\sigma_0\tau < 3$.

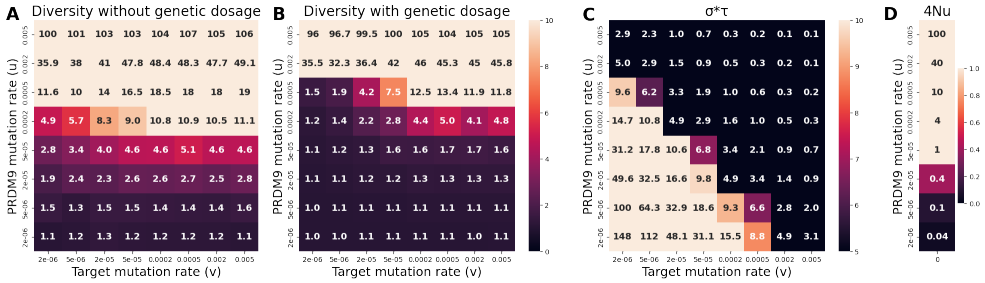

Figure 5.3

S4 Fig. Scaling experiment showing the conditions for a change of regime (from polymorphic to monomorphic) upon introducing gene dosage. A and B: equilibrium levels of *PRDM9* diversity without (A) and with (B) gene dosage. C: summary statistic predictive of the type of regime observed in the presence of dosage, defined as $\sigma_0\tau$ (selection coefficient associated to dosage multiplied by the time between successive invasions by new *PRDM9* alleles). The regime is predicted to be polymorphic in the presence of dosage if $\sigma_0\tau < 3$.

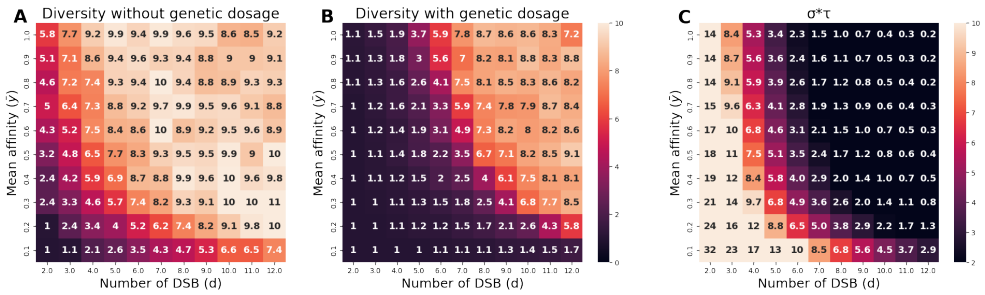

Figure 5.4

S5 Fig. Exponential law for affinity distribution (with mean $y = 0.2$). The continuous blue line corresponds to the affinity distribution for young alleles and the dotted red line corresponds to the affinity distribution for old alleles.

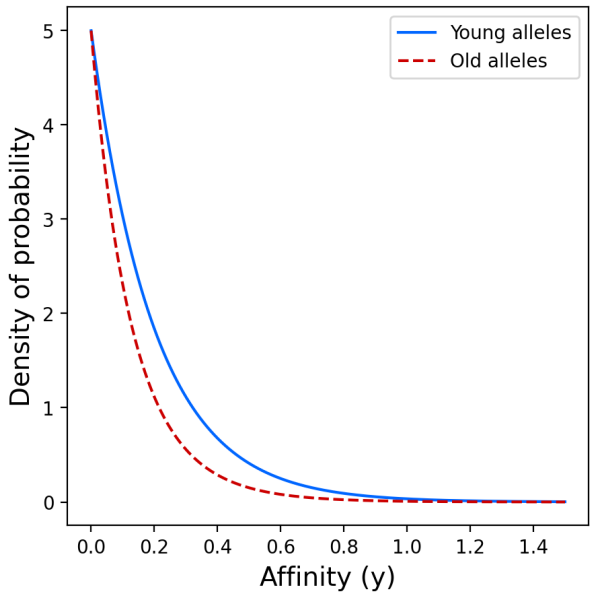

Figure 5.5

S6 Fig. Binding probability as a function of the concentration of free PRDM9 protein molecules for homozygotes and heterozygotes (mean affinity y = 0.6). The binding probability for homozygotes is always higher than that for heterozygotes, but the difference between them decreases when the free PRDM9 concentration increases.

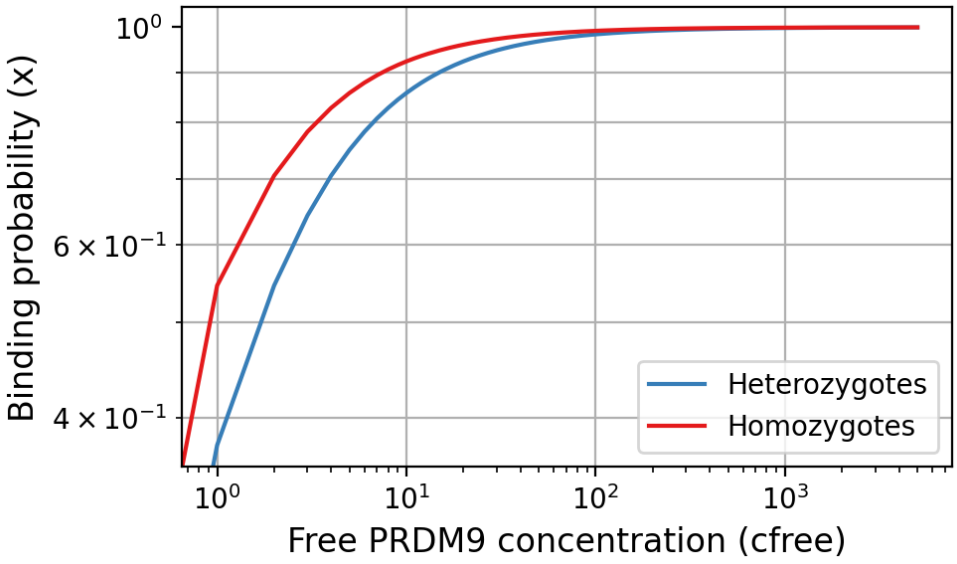

Figure 5.6

S7 Fig. Selection coefficient associated to gene dosage (σ_0) as a function of the concentration of PRDM9 in the cell and the mean affinity of the target sites ($\bar{y}$).

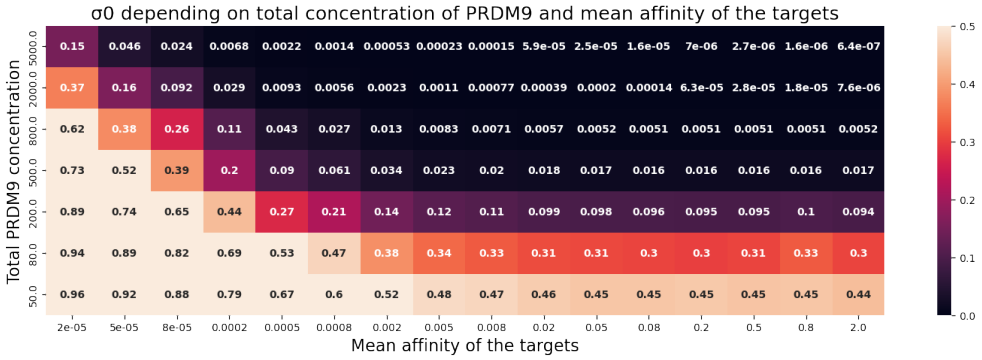

Figure 5.7

Bibliographie

[1] Latrille T, Duret L, Lartillot N. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017;372(1736):20160463. doi:10.1098/rstb.2016.0463. *Cited at pages 21, 30, 36, 41, 43, 46, 51, 52, 56, 60, 66, 83, 85, 107, 125, 128, 129*

[2] Kono H, Tamura M, Osada N, Suzuki H, Abe K, Moriwaki K, et al. Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. DNA Research. 2014;21(3):315–326. doi:10.1093/dnares/dst059. *Cited at pages 13, 21, 41, 108, 136*

[3] Smagulova F, Brick K, Pu Y, Camerini-Otero RD, Petukhova GV. The evolutionary turnover of recombination hot spots contributes to speciation in mice. Genes & Development. 2016;30(3):266–280. doi:10.1101/gad.270009.115. *Cited at pages 41, 56, 57, 62, 80, 91, 96, 97, 108, 109, 112, 115, 136*

6

A theoretical investigation of the role of *PRDM9* in hybrid sterility - Supplementary Materials

Supplementary materials

Supplementary figures

S1 Fig. Bi-dimensional scaling of Prdm9 diversity (D) in function of the mutation rate at *Prdm9* locus (u) and mutations rate at target sites (v) in population 1 (left panels) and population 2 (right panels). When it is close to one, the regime is monomorphic, when it is higher, the regime is polymorphic. When u is high, we are in a polymorphic regime and v does not change the diversity so much (panels A and B). The dosage tends to act against diversity meaning that the monomorphic regime are found for higher u than without dosage (panels C and D). The non limitation of gametes (panels E and F) and the increase in DSB number (panels G and H) restore the diversity levels as without dosage.

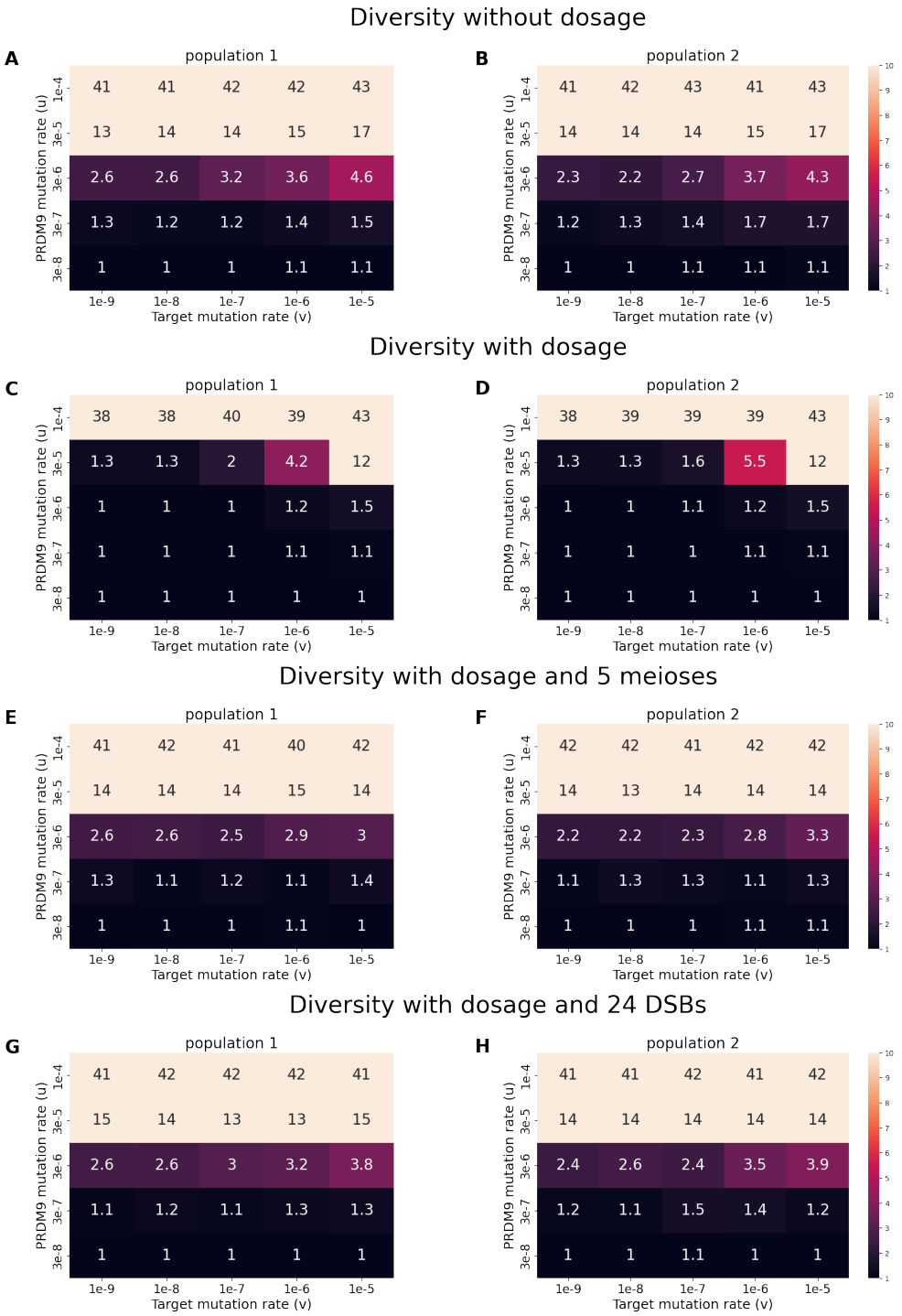

Figure 6.1

S2 Fig. Bi-dimensional scaling of mean activity of Prdm9 target sites (θ) in the population along the whole simulation in function of the mutation rate at *Prdm9* locus (u) and mutations rate at target sites (v) in population 1 (left panels) and population 2 (right panels). When it is close to one, their is no or few erosion. Low level of activity are found for high U and low v (panels A and

B). The dosage effect allows for a higher level of erosion than without dosage (panels C and D). The non limitation of gametes (panels E and F) and the increase in DSB number (panels G and H) restore the activity levels for high u and low v but accentuate the erosion for low u and high v.

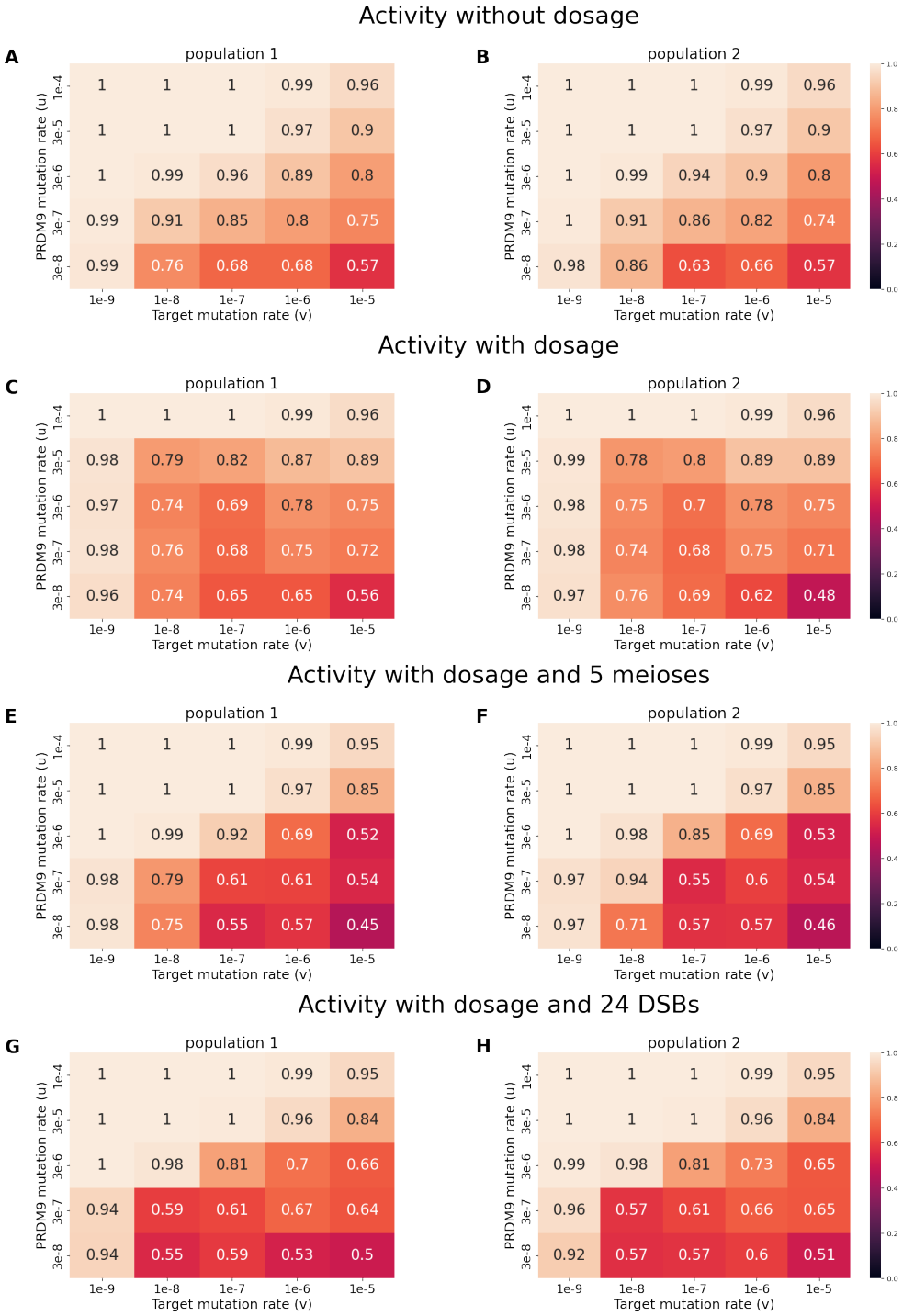

Figure 6.2

S3 Fig. Bi-dimensional scaling of scaled selection coefficient ($4Ns_0$) in function of the mutation rate at *Prdm9* locus (u) and mutations rate at target sites (v) in population 1 (left panels) and population 2 (right panels). Numbers close to 0 indicate no selection regimes, positive numbers indicate positive selection and negative one indicate negative selection. These are means among all the simulation but some oscillations can occur between positive and negative selection along time.

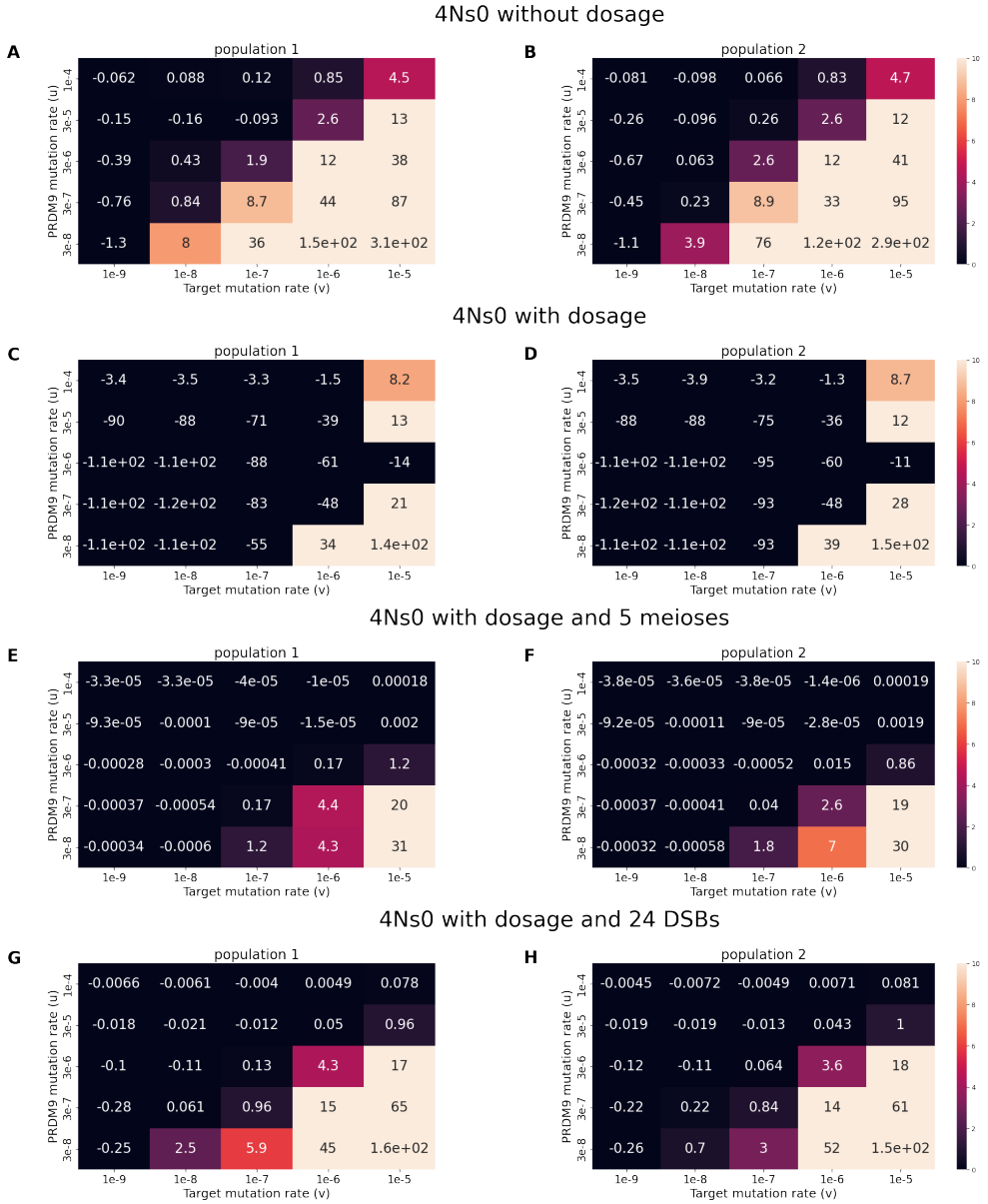

Figure 6.3

S4 Fig. Bi-dimensional scaling of the percentage of time in the whole simulation when the regime is under negative selection when it is run with genetic dosage ($4Ns_0 < -1$) in function of the mutation rate at *Prdm9* locus (u) and mutations rate at target sites (v) in population 1 (A) and population 2 (B).

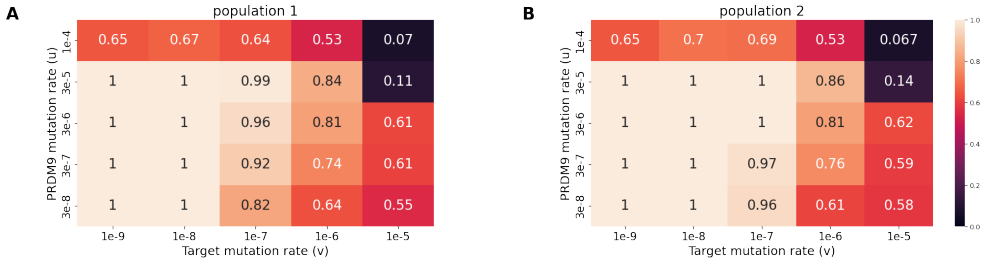

Figure 6.4

S5 Fig. Bi-dimensional scaling of the percentage of time in the whole simulation when the regime is under neutral selection when it is run with non-limiting gametes ($-1 < 4Ns_0 < 1$) in function of the mutation rate at *Prdm9* locus (u) and mutations rate at target sites (v) in population 1 (A) and population 2 (B).

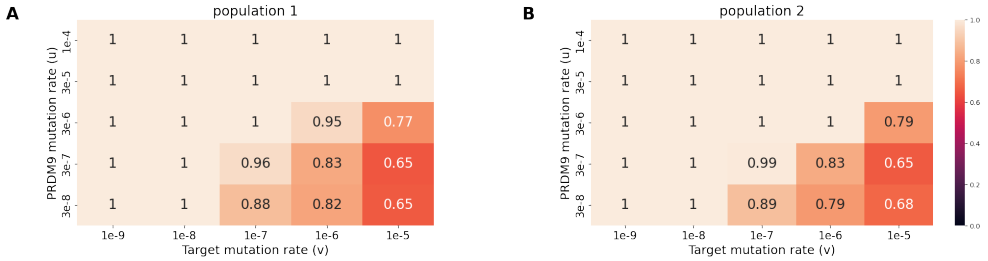

Figure 6.5

S6 Fig. Individual fertility in function of meiosis success when fertility correspond to meiosis success (blue) and when individuals can perform up to 5 meiosis before being declared sterile. The horizontal dotted line correspond to the 0.99 threshold for fertility, and the vertical dotted line correspond to the corresponding meiosis success rate associated to the fertility level for 1 or 5 meiosis per individuals. Ths figure show that when individuals can realise up to 5 meiosis, the loss of fertility is still lower than 1% for a meiosis success of 0.6%.

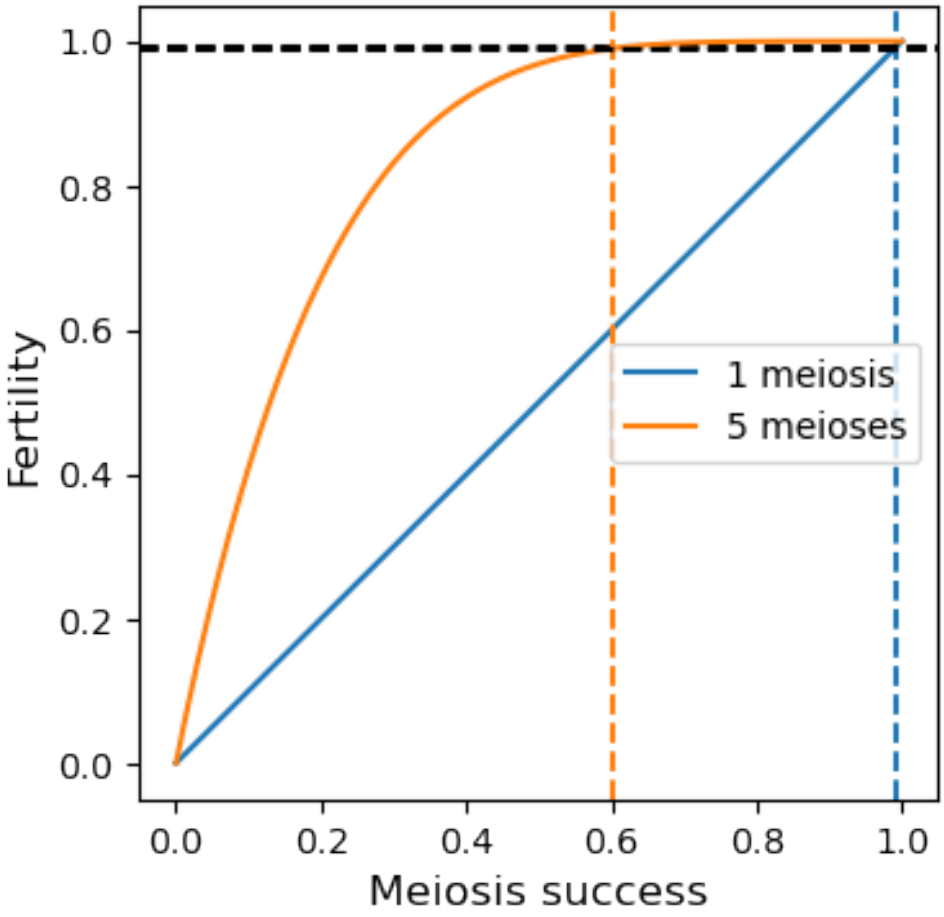
1.0
0.8
0.6
0.4
0.2
0.0
Fertility
0.0
0.2
0.4
0.6
0.8
1.0
Meiosis success
1 meiosis
5 meioses

Figure 6.6

Bibliographie

[1] Skottsberg CJF. Nils Eberhard Svedelius, 1873-1960. Biographical Memoirs of Fellows of the Royal Society. 1961;7:294–312. doi:10.1098/rsbm.1961.0023. *Cited at page 4*

[2] Barton NH, Charlesworth B. Why Sex and Recombination? Science. 1998;281(5385):1986–1990. doi:10.1126/science.281.5385.1986. *Cited at page 4*

[3] Otto SP, Lenormand T. Resolving the paradox of sex and recombination. Nature Reviews Genetics. 2002;3(4):252–261. doi:10.1038/nrg761. *Cited at page 4*

[4] Ohkura H. Meiosis: An Overview of Key Differences from Mitosis. Cold Spring Harbor Perspectives in Biology. 2015;7(5):a015859. doi:10.1101/cshperspect.a015859. *Cited at page 5*

[5] Hassold T, Hall H, Hunt P. The origin of human aneuploidy: where we have been, where we are going. Human Molecular Genetics. 2007;16(R2):R203–R208. doi:10.1093/hmg/ddm243. *Cited at page 6*

[6] Zickler D, Kleckner N. THE LEPTOTENE-ZYGOTENE TRANSITION OF MEIOSIS. Annual Review of Genetics. 1998;32(1):619–697. doi:10.1146/annurev.genet.32.1.619. *Cited at pages 6, 8, 14*

[7] Baudat F, Imai Y, de Massy B. Meiotic recombination in mammals: localization and regulation. Nature Reviews Genetics. 2013;14(11):794–806. doi:10.1038/nrg3573. *Cited at pages 6, 10, 11, 13, 14*

[8] Dernburg AF, McDonald K, Moulder G, Barstead R, Dresser M, Villeneuve AM. Meiotic Recombination in C. elegans Initiates by a Conserved Mechanism and Is Dispensable for Homologous Chromosome Synapsis. Cell. 1998;94(3):387–398. doi:10.1016/S0092-8674(00)81481-6. *Cited at page 6*

[9] Rubin T, Macaisne N, Vallés AM, Guilleman C, Gaugué I, Dal Toe L, et al. Premeiotic pairing of homologous chromosomes during Drosophila male meiosis. Proceedings of the National Academy of Sciences. 2022;119(47):e2207660119. doi:10.1073/pnas.2207660119. *Cited at page 6*

[10] Morgan TH. NO CROSSING OVER IN THE MALE OF DROSOPHILA OF GENES IN THE SECOND AND THIRD PAIRS OF CHROMOSOMES. The Biological Bulletin. 1914;26(4):195–204. doi:10.2307/1536193. *Cited at page 6*

[11] McKee BD. Homologous pairing and chromosome dynamics in meiosis and mitosis. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. 2004;1677(1):165–180. doi:10.1016/j.bbaexp.2003.11.017. *Cited at page 6*

[12] McKee BD, Yan R, Tsai JH. Meiosis in male Drosophila. Spermatogenesis. 2012;2(3):167–184. doi:10.4161/spmg.21800. *Cited at page 6*

[13] Scherthan H, Weich S, Schwegler H, Heyting C, Härle M, Cremer T. Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. Journal of Cell Biology. 1996;134(5):1109–1125. doi:10.1083/jcb.134.5.1109. *Cited at page 6*

[14] Paigen K, Petkov P. Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics. 2010;11(3):221–233. doi:10.1038/nrg2712. *Cited at page 6*

[15] Keeney S, Lange J, Mohibullah N. Self-Organization of Meiotic Recombination Initiation: General Principles and Molecular Pathways. Annual Review of Genetics. 2014;48(1):187–214. doi:10.1146/annurev-genet-120213-092304. *Cited at page 6*

[16] De Massy B. Initiation of Meiotic Recombination: How and Where? Conservation and Specificities Among Eukaryotes. Annual Review of Genetics. 2013;47(1):563–599. doi:10.1146/annurev-genet-110711-155423. *Cited at pages 6, 12, 14*

[17] Lange J, Yamada S, Tischfield SE, Pan J, Kim S, Zhu X, et al. The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. Cell. 2016;167(3):695–708.e16. doi:10.1016/j.cell.2016.09.035. *Cited at pages 6, 11, 12, 14, 40, 59, 65*

[18] Cole F, Keeney S, Jasin M. Evolutionary conservation of meiotic DSB proteins: more than just Spo11. Genes & Development. 2010;24(12):1201–1207. doi:10.1101/gad.1944710. *Cited at pages 6, 14*

[19] Hunter N. Meiotic Recombination: The Essence of Heredity. Cold Spring Harbor Perspectives in Biology. 2015;7(12):a016618. doi:10.1101/cshperspect.a016618. *Cited at page 6*

[20] Smagulova F, Gregoretti IV, Brick K, Khil P, Camerini-Otero RD, Petukhova GV. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature. 2011;472(7343):375–378. doi:10.1038/nature09869. *Cited at page 8*

[21] San Filippo J, Sung P, Klein H. Mechanism of Eukaryotic Homologous Recombination. Annual Review of Biochemistry. 2008;77(1):229–257. doi:10.1146/annurev.biochem.77.061306.125255. *Cited at page 8*

[22] Hinch AG, Becker PW, Li T, Moralli D, Zhang G, Bycroft C, et al. The Configuration of RPA, RAD51, and DMC1 Binding in Meiosis Reveals the Nature of Critical Recombination Intermediates. Molecular Cell. 2020;79(4):689–701.e10. doi:10.1016/j.molcel.2020.06.015. *Cited at page 8*

[23] Holloway JK, Booth J, Edelmann W, McGowan CH, Cohen PE. MUS81 Generates a Subset of MLH1-MLH3–Independent Crossovers in Mammalian Meiosis. PLOS Genetics. 2008;4(9):e1000186. doi:10.1371/journal.pgen.1000186. *Cited at page 8*

[24] Cole F, Kauppi L, Lange J, Roig I, Wang R, Keeney S, et al. Homeostatic control of recombination is implemented progressively in mouse meiosis. Nature cell biology. 2012;14(4):424–430. doi:10.1038/ncb2451. *Cited at page 8*

[25] Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nature Communications. 2019;10:3900. doi:10.1038/s41467-019-11675-y. *Cited at pages 8, 29, 40, 42, 63, 115*

[26] Takeo S, Hawley RS. Rumors of Its Disassembly Have Been Greatly Exaggerated: The Secret Life of the Synaptonemal Complex at the Centromeres. PLOS Genetics. 2012;8(6):e1002807. doi:10.1371/journal.pgen.1002807. *Cited at page 8*

[27] Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nature Genetics. 2002;31(3):241–247. doi:10.1038/ng917. *Cited at page 10*

[28] Payseur BA, Nachman MW. Microsatellite Variation and Recombination Rate in the Human Genome. Genetics. 2000;156(3):1285–1298. doi:10.1093/genetics/156.3.1285. *Cited at page 10*

[29] Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, et al. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research. 2004;14(4):528–538. doi:10.1101/gr.1970304. *Cited at page 10*

[30] Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467(7319):1099–1103. doi:10.1038/nature09525. *Cited at page 10*

[31] Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science. 2012;336(6078):193–198. doi:10.1126/science.1216872. *Cited at pages 10, 12, 21, 40*

[32] Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature. 2008;454(7203):479–485. doi:10.1038/nature07135. *Cited at page 11*

[33] Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics. 2001;29(2):217–222. doi:10.1038/ng1001-217. *Cited at pages 11, 12, 39*

[34] Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012;485(7400):642–645. doi:10.1038/nature11089. *Cited at pages 11, 56, 122*

[35] Grey C, Clément JAJ, Buard J, Leblanc B, Gut I, Gut M, et al. In vivo binding of PRDM9 reveals interactions with noncanonical genomic sites. Genome Research. 2017;27(4):580–590. doi:10.1101/gr.217240.116. *Cited at pages 11, 14, 41*

[36] Diagouraga B, Clément JAJ, Duret L, Kadlec J, Massy Bd, Baudat F. PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites. Molecular Cell. 2018;69(5):853–865.e6. doi:10.1016/j.molcel.2018.01.033. *Cited at pages 11, 12, 14, 62, 80, 91, 98, 115*

[37] Kauppi L, Jeffreys AJ, Keeney S. Where the crossovers are: recombination distributions in mammals. Nature Reviews Genetics. 2004;5(6):413–424. doi:10.1038/nrg1346. *Cited at pages 11, 39*

[38] Lichten M, Goldman ASH. MEIOTIC RECOMBINATION HOTSPOTS. Annual Reviews Genetics. 1995;29:423–444. *Cited at page 11*

[39] Jeffreys AJ, Ritchie A, Neumann R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Human Molecular Genetics. 2000;9(5):725–733. doi:10.1093/hmg/9.5.725. *Cited at page 11*

[40] Lien S, Szyda J, Schechinger B, Rappold G, Arnheim N. Evidence for Heterogeneity in Recombination in the Human Pseudoautosomal Region: High Resolution Analysis by Sperm Typing and Radiation-Hybrid Mapping. The American Journal of Human Genetics. 2000;66(2):557–566. doi:10.1086/302754. *Cited at page 12*

[41] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science. 2005;310(5746):321–324. doi:10.1126/science.1117196. *Cited at pages 12, 39*

[42] Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome. Genetics. 2012;191(3):757–764. doi:10.1534/genetics.112.141036. *Cited at pages 12, 39*

[43] Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Research. 2012;22(1):51–63. doi:10.1101/gr.124123.111. *Cited at pages 12, 23, 122*

[44] Auton A, Li YR, Kidd J, Oliveira K, Nadel J, Holloway JK, et al. Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLOS Genetics. 2013;9(12):e1003984. doi:10.1371/journal.pgen.1003984. *Cited at page 12*

[45] Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, et al. Stable recombination hotspots in birds. Science. 2015;350(6263):928–932. doi:10.1126/science.aad0843. *Cited at pages 12, 122, 123*

[46] Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, et al. Recombination and linkage disequilibrium in Arabidopsis thaliana. Nature Genetics. 2007;39(9):1151–1155. doi:10.1038/ng2115. *Cited at page 12*

[47] Saintenac C, Faure S, Remay A, Choulet F, Ravel C, Paux E, et al. Variation in crossover rates across a 3-Mb contig of bread wheat (Triticum aestivum) reveals the presence of a meiotic recombination hotspot. Chromosoma. 2011;120(2):185–198. doi:10.1007/s00412-010-0302-9. *Cited at page 12*

[48] Lesecque Y, Glémin S, Lartillot N, Mouchiroud D, Duret L. The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. PLOS Genetics. 2014;10(11):e1004790. doi:10.1371/journal.pgen.1004790. *Cited at pages 12, 21, 40*

[49] McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science. 2004;304(5670):581–584. doi:10.1126/science.1092500. *Cited at pages 12, 39*

[50] Lam I, Keeney S. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. Science. 2015;350(6263):932–937. doi:10.1126/science.aad0814. *Cited at page 12*

[51] Borde V, Robine N, Lin W, Bonfils S, Géli V, Nicolas A. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. European Molecular Biology Organization Journal. 2009;28:99 – 111. doi:10.1038/emboj.2008.257. *Cited at pages 12, 80*

[52] Borde V, de Massy B. Programmed induction of DNA double strand breaks during meiosis: setting up communication between DNA and the chromosome structure. Current Opinion in Genetics & Development. 2013;23(2):147–155. doi:10.1016/j.gde.2012.12.002. *Cited at pages 12, 14, 30, 42, 80*

[53] Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, et al. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. eLife. 2017;6:e24133. doi:10.7554/eLife.24133. *Cited at pages 12, 14, 21, 23, 123*

[54] Buard J, Barthès P, Grey C, de Massy B. Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. The EMBO Journal. 2009;28(17):2616–2624. doi:10.1038/emboj.2009.207. *Cited at page 12*

[55] Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. Nature Genetics. 2008;40(9):1124–1129. doi:10.1038/ng.213. *Cited at page 13*

[56] Webb AJ, Berg IL, Jeffreys A. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proceedings of the National Academy of Sciences. 2008;105(30):10471–10476. doi:10.1073/pnas.0804933105. *Cited at page 13*

[57] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science. 2010;327(5967):836–840. doi:10.1126/science.1183439. *Cited at pages 13, 21, 28, 30, 39, 117*

[58] Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science. 2010;327(5967):876–879. doi:10.1126/science.1182363. *Cited at pages 13, 21, 28, 39, 40*

[59] Grey C, Baudat F, Massy Bd. PRDM9, a driver of the genetic map. PLOS Genetics. 2018;14(8):e1007479. doi:10.1371/journal.pgen.1007479. *Cited at pages 13, 40*

[60] Birtle Z, Ponting CP. Meisetz and the birth of the KRAB motif. Bioinformatics. 2006;22(23):2841–2845. doi:10.1093/bioinformatics/btl498. *Cited at page 13*

[61] Lim FL, Soulez M, Koczan D, Thiesen HJ, Knight JC. A KRAB-related domain and a novel transcription repression domain in proteins encoded by SSX genes that are disrupted in human sarcomas. Oncogene. 1998;17(15):2013–2018. doi:10.1038/sj.onc.1202122. *Cited at page 13*

[62] Hayashi K, Yoshida K, Matsui Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature. 2015;438(7066):374–378. doi:10.1038/nature04112. *Cited at pages 13, 14, 28, 40, 60*

[63] Dillon SC, Zhang X, Trievel RC, Cheng X. The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biology. 2005;6(8):227. doi:10.1186/gb-2005-6-8-227. *Cited at page 13*

[64] Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature genetics. 2010;42(10):859–863. doi:10.1038/ng.658. *Cited at pages 13, 21*

[65] Pabo CO, Peisach E, Grant RA. Design and Selection of Novel Cys2His2 Zinc Finger Proteins. Annual Review of Biochemistry. 2001;70(1):313–340. doi:10.1146/annurev.biochem.70.1.313. *Cited at pages 13, 117*

[66] Wolfe SA, Grant RA, Elrod-Erickson M, Pabo CO. Beyond the "Recognition Code": Structures of Two Cys2His2 Zinc Finger/TATA Box Complexes. Structure. 2001;9:717–723. *Cited at pages 13, 117*

[67] Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ. 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. 2011;108(30):12378–12383. doi:10.1073/pnas.1109531108. *Cited at pages 13, 21, 41*

[68] Buard J, Rivals E, Segonzac DDd, Garres C, Caminade P, Massy Bd, et al. Diversity of Prdm9 Zinc Finger Array in Wild Mice Unravels New Facets of the Evolutionary Turnover of this Coding Minisatellite. PLOS ONE. 2014;9(1):e85021. doi:10.1371/journal.pone.0085021. *Cited at pages 13, 21, 41, 59, 108*

[69] Kono H, Tamura M, Osada N, Suzuki H, Abe K, Moriwaki K, et al. Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. DNA Research. 2014;21(3):315–326. doi:10.1093/dnares/dst059. *Cited at pages 13, 21, 41, 108, 136*

[70] Steiner CC, Ryder OA. Characterization of Prdm9 in Equids and Sterility in Mules. PLOS ONE. 2013;8(4):e61746. doi:10.1371/journal.pone.0061746. *Cited at pages 13, 108*

[71] Ahlawat S, De S, Sharma P, Sharma R, Arora R, Kataria RS, et al. Evolutionary dynamics of meiotic recombination hotspots regulator PRDM9 in bovids. Molecular Genetics and Genomics. 2017;292(1):117–131. doi:10.1007/s00438-016-1260-6. *Cited at pages 13, 108*

[72] Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. PLOS Genetics. 2012;8(7):e1002854. doi:10.1371/journal.pgen.1002854. *Cited at pages 13, 108*

[73] Damm E, Ullrich KK, Amos WB, Odenthal-Hesse L. Evolution of the recombination regulator PRDM9 in minke whales. BMC Genomics. 2022;23(1):212. doi:10.1186/s12864-022-08305-1. *Cited at pages 13, 108*

[74] Alleva B, Brick K, Pratto F, Huang M, Camerini-Otero RD. Cataloging Human PRDM9 Allelic Variation Using Long-Read Sequencing Reveals PRDM9 Population Specificity and Two Distinct Groupings of Related Alleles. Frontiers in Cell and Developmental Biology. 2021;9. *Cited at pages 13, 41, 108*

[75] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science. 2009;323(5912):373–375. doi:10.1126/science.1163601. *Cited at pages 14, 26, 28, 30, 39, 80, 81, 96*

[76] Eram MS, Bustos SP, Lima-Fernandes E, Siarheyeva A, Senisterra G, Hajian T, et al. Trimethylation of Histone H3 Lysine 36 by Human Methyltransferase PRDM9 Protein. The Journal of Biological Chemistry. 2014;289(17):12177–12188. doi:10.1074/jbc.M113.523183. *Cited at pages 14, 60*

[77] Parvanov ED, Tian H, Billings T, Saxl RL, Spruce C, Aithal R, et al. PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis. Molecular Biology of the Cell. 2017;28(3):488–499. doi:10.1091/mbc.E16-09-0686. *Cited at pages 14, 41*

[78] Imai Y, Baudat F, Taillepierre M, Stanzione M, Toth A, de Massy B. The PRDM9 KRAB domain is required for meiosis and involved in protein interactions. Chromosoma. 2017;126(6):681–695. doi:10.1007/s00412-017-0631-z. *Cited at page 14*

[79] Sommermeyer V, Béneut C, Chaplais E, Serrentino ME, Borde V. Spp1, a Member of the Set1 Complex, Promotes Meiotic DSB Formation in Promoters by Tethering Histone H3K4 Methylation Sites to Chromosome Axes. Molecular Cell. 2013;49(1):43–54. doi:10.1016/j.molcel.2012.11.008. *Cited at page 14*

[80] Baker CL, Walker M, Kajita S, Petkov PM, Paigen K. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Research. 2014;24(5):724–732. doi:10.1101/gr.170167.113. *Cited at pages 14, 29, 56, 58*

[81] Kauppi L, Barchi M, Lange J, Baudat F, Jasin M, Keeney S. Numerical constraints and feedback control of double-strand breaks in mouse meiosis. Genes & Development. 2013;27(8):873–886. doi:10.1101/gad.213652.113. *Cited at pages 14, 56*

[82] Duret L, Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics. 2009;10(1):285–311. doi:10.1146/annurev-genom-082908-150001. *Cited at page 14*

[83] Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997;94(15):8058–8063. doi:10.1073/pnas.94.15.8058. *Cited at pages 16, 18, 40, 65, 80, 106, 123*

[84] Servedio MR, Brandvain Y, Dhole S, Fitzpatrick CL, Goldberg EE, Stern CA, et al. Not Just a Theory—The Utility of Mathematical Models in Evolutionary Biology. PLOS Biology. 2014;12(12):e1002017. doi:10.1371/journal.pbio.1002017. *Cited at page 16*

[85] Úbeda F, Wilkins JF. The Red Queen theory of recombination hotspots. Journal of Evolutionary Biology. 2011;24(3):541–553. doi:10.1111/j.1420-9101.2010.02187.x. *Cited at pages 18, 19, 21, 30, 36, 40, 41, 43, 60, 80, 107*

[86] Coop G, Myers SR. Live Hot, Die Young: Transmission Distortion in Recombination Hotspots. PLOS Genetics. 2007;3(3):e35. doi:10.1371/journal.pgen.0030035. *Cited at page 18*

[87] Peters AD. A Combination of cis and trans Control Can Solve the Hotspot Conversion Paradox. Genetics. 2008;178(3):1579–1593. doi:10.1534/genetics.107.084061. *Cited at page 18*

[88] Archetti M. A selfish origin for recombination. Journal of Theoretical Biology. 2003;223(3):335–346. doi:10.1016/S0022-5193(03)00102-4. *Cited at page 18*

[89] Van Valen L. Molecular evolution as predicted by natural selection. Journal of Molecular Evolution. 1974;3(2):89–101. doi:10.1007/BF01796554. *Cited at pages 19, 40*

[90] Dieckmann U, Marrow P, Law R. Evolutionary cycling in predator-prey interactions: population dynamics and the red queen. Journal of Theoretical Biology. 1995;176(1):91–102. doi:10.1006/jtbi.1995.0179. *Cited at page 19*

[91] Dercole F, Ferriere R, Rinaldi S. Chaotic Red Queen coevolution in three-species food chains. Proceedings of the Royal Society B: Biological Sciences. 2010;doi:10.1098/rspb.2010.0209. *Cited at page 19*

[92] Decaestecker E, Gaba S, Raeymaekers JAM, Stoks R, Van Kerckhoven L, Ebert D, et al. Host–parasite 'Red Queen' dynamics archived in pond sediment. Nature. 2007;450(7171):870–873. doi:10.1038/nature06291. *Cited at page 19*

[93] Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, et al. Antagonistic coevolution accelerates molecular evolution. Nature. 2010;464(7286):275–278. doi:10.1038/nature08798. *Cited at page 19*

[94] Jeffreys AJ, Cotton VE, Neumann R, Lam KWG. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. Proceedings of the National Academy of Sciences. 2013;110(2):600–605. doi:10.1073/pnas.1220813110. *Cited at pages 21, 40, 56*

[95] Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLOS Genetics. 2009;5(12):e1000753. doi:10.1371/journal.pgen.1000753. *Cited at pages 21, 23, 30, 39, 41, 59, 81, 108*

[96] Latrille T, Duret L, Lartillot N. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017;372(1736):20160463. doi:10.1098/rstb.2016.0463. *Cited at pages 21, 30, 36, 41, 43, 46, 51, 52, 56, 60, 66, 83, 85, 107, 125, 128, 129*

[97] Muñoz-Fuentes V, Di Rienzo A, Vilà C. Prdm9, a Major Determinant of Meiotic Recombination Hotspots, Is Not Functional in Dogs and Their Wild Relatives, Wolves and Coyotes. PLoS ONE. 2011;6(11):e25498. doi:10.1371/journal.pone.0025498. *Cited at pages 23, 123*

[98] Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M, Ciccarelli FD. Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. BMC Evolutionary Biology. 2007;7:187. doi:10.1186/1471-2148-7-187. *Cited at page 23*

[99] Darwin C. On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life; 1859. *Cited at page 24*

[100] Bateson W. Heredity and Variation in Modern Lights. In: Seward AC, editor. Darwin and Modern Science: Essays in Commemoration of the Centenary of the Birth of Charles Darwin and of the Fiftieth Anniversary of the Publication of The Origin of Species. Cambridge Library Collection - Darwin, Evolution and Genetics. Cambridge: Cambridge University Press; 1909. p. 85–101. Available from: https://www.cambridge.org/core/books/darwin-and-modern-science/heredity-and-variation-in-modern-lights/6105CC0E0388ECDCEEA76EE779E278BE. *Cited at page 24*

[101] Dobzhansky T, Beadle GW. Studies on Hybrid Sterility IV. Transplanted Testes in Drosophila Pseudoobscura. Genetics. 1936;21(6):832–840. *Cited at pages 24, 79*

[102] Dobzhansky T. Genetics and the Origin of Species. 11. Columbia university press; 1982. *Cited at pages 24, 79*

[103] Muller HJ. Isolation mechanisms, evolution and temperature. Biology Symposium. 1942;6:71–125. *Cited at pages 24, 79*

[104] Orr HA. Dobzhansky, Bateson, and the Genetics of Speciation. Genetics. 1996;144(4):1331–1335. doi:10.1093/genetics/144.4.1331. *Cited at pages 24, 79*

[105] Dobzhansky T. Experiments on Sexual Isolation in Drosophila. Proceedings of the National Academy of Sciences. 1951;37(12):792–796. doi:10.1073/pnas.37.12.792. *Cited at page 25*

[106] Mayr E. Animal Species and Evolution. In: Animal Species and Evolution. Harvard University Press; 1963.Available from: https://www.degruyter.com/document/doi/10.4159/harvard.9780674865327/html. *Cited at page 25*

[107] Ting CT, Tsaur SC, Wu ML, Wu CI. A Rapidly Evolving Homeobox at the Site of a Hybrid Sterility Gene. Science. 1998;282(5393):1501–1504. doi:10.1126/science.282.5393.1501. *Cited at pages 26, 79*

[108] Wu CI, Ting CT. Genes and speciation. Nature Reviews Genetics. 2004;5(2):114–122. doi:10.1038/nrg1269. *Cited at pages 26, 79*

[109] Phadnis N, Orr HA. A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids. Science. 2009;323(5912):376–379. doi:10.1126/science.1163934. *Cited at pages 26, 79*

[110] Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature. 2003;423(6941):715–719. doi:10.1038/nature01679. *Cited at pages 26, 79*

[111] Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature. 2016;530(7589):171–176. doi:10.1038/nature16931. *Cited at pages 26, 29, 30, 39, 41, 57, 58, 60, 65, 76, 80, 89, 91, 96, 97, 98, 108, 109, 122*

[112] Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F. The Evolution of House Mice. Annual Review of Ecology and Systematics. 1993;24(1):119–152. doi:10.1146/annurev.es.24.110193.001003. *Cited at pages 26, 81*

[113] Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, et al. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Molecular Ecology. 2008;17(24):5349–5363. doi:10.1111/j.1365-294X.2008.04005.x. *Cited at pages 26, 81*

[114] Macholán M, Baird SJE, Munclinger P, Piálek J. Evolution of the House Mouse. Cambridge University Press; 2012. *Cited at page 26*

[115] Iványi P, Vojtísková M, Démant P, Micková M. Genetic factors in the ninth linkage group influencing reproductive performance in male mice. Folia biologica. 1969;15(6):401–421. *Cited at page 26*

[116] Forejt J, Iványi P. Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.). Genetics Research. 1974;24(2):189–206. doi:10.1017/S0016672300015214. *Cited at pages 26, 79*

[117] Gregorová S, Forejt J. PWD/Ph and PWK/Ph Inbred Mouse Strains of Mus cyrchar
cyrt. musculus Subspecies-a Valuable Resource of Phenotypic Variations and Genomic Polymorphisms. Folia biologica. 2000;46:31–41. *Cited at pages 26, 29, 41, 80, 81*

[118] Haldane JB. Sex ratio and unisexual sterility in hybrid animals. Journal of genetics. 1922;12:101–109. *Cited at pages 28, 79*

[119] Schilthuizen M, Giesbers MCWG, Beukeboom LW. Haldane's rule in the 21st century. Heredity. 2011;107(2):95–102. doi:10.1038/hdy.2010.170. *Cited at pages 28, 79*

[120] Muller HJ. Bearing of the Drosophila work on systematics. The new systematics,. 1940; p. 185–268. *Cited at pages 28, 79*

[121] Orr HA, Turelli M. Dominance and Haldane's Rule. Genetics. 1996;143(1):613–616. *Cited at page 28*

[122] Turelli M, Orr HA. The dominance theory of Haldane's rule. Genetics. 1995;140(1):389–402. doi:10.1093/genetics/140.1.389. *Cited at page 28*

[123] Wu CI, Johnson NA, Palopoli MF. Haldane's rule and its legacy: Why are there so many sterile males? Trends in Ecology & Evolution. 1996;11(7):281–284. doi:10.1016/0169-5347(96)10033-1. *Cited at pages 28, 79*

[124] Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, et al. PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. PLOS Genetics. 2015;11(1):e1004916. doi:10.1371/journal.pgen.1004916. *Cited at pages 28, 41, 56, 57, 81, 89, 108, 115*

[125] Gregorova S, Gergelits V, Chvatalova I, Bhattacharyya T, Valiskova B, Fotopulosova V, et al. Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. eLife. 2018;7:e34282. doi:10.7554/eLife.34282. *Cited at pages 29, 30, 41, 80, 81, 89, 96, 107*

[126] Valiskova B, Gregorova S, Lustyk D, Šimeček P, Jansa P, Forejt J. Genic and Chromosomal Components of Prdm9-Driven Hybrid Male Sterility in Mice (Mus musculus). Genetics. 2022; p. iyac116. doi:10.1093/genetics/iyac116. *Cited at page 29*

[127] Flachs P, Mihola O, Šimeček P, Gregorová S, Schimenti JC, Matsui Y, et al. Interallelic and Intergenic Incompatibilities of the Prdm9 (Hst1) Gene in Mouse Hybrid Sterility. PLOS Genetics. 2012;8(11):e1003044. doi:10.1371/journal.pgen.1003044. *Cited at page 29*

[128] Forejt J. Asymmetric breaks in DNA cause sterility. Nature. 2016;530(7589):167–168. doi:10.1038/nature16870. *Cited at page 29*

[129] Bhattacharyya T, Reifova R, Gregorova S, Simecek P, Gergelits V, Mistrik M, et al. X Chromosome Control of Meiotic Chromosome Synapsis in Mouse Inter-Subspecific Hybrids. PLOS Genetics. 2014;10(2):e1004088. doi:10.1371/journal.pgen.1004088. *Cited at pages 29, 118*

[130] Balcova M, Faltusova B, Gergelits V, Bhattacharyya T, Mihola O, Trachtulec Z, et al. Hybrid Sterility Locus on Chromosome X Controls Meiotic Recombination Rate in Mouse. PLOS Genetics. 2016;12(4):e1005906. doi:10.1371/journal.pgen.1005906. *Cited at pages 29, 97, 118*

[131] Mukaj A, Piálek J, Fotopulosova V, Morgan AP, Odenthal-Hesse L, Parvanov ED, et al. Prdm9 Intersubspecific Interactions in Hybrid Male Sterility of House Mouse. Molecular Biology and Evolution. 2020;37(12):3423–3438. doi:10.1093/molbev/msaa167. *Cited at pages 29, 80, 96, 97, 117*

[132] Forejt J, Jansa P, Parvanov E. Hybrid sterility genes in mice (Mus musculus): a peculiar case of PRDM9 incompatibility. Trends in Genetics. 2021;37(12):1095–1108. doi:10.1016/j.tig.2021.06.008. *Cited at page 29*

[133] Genestier A, Duret L, Lartillot N. Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis : a model based exploration of the PRDM9 intra-genomic Red Queen; 2023. Available from: https://www.biorxiv.org/content/10.1101/2023.03.08.531712v2. *Cited at pages 30, 32, 80, 81, 82, 83, 85, 86, 93, 94, 95, 96, 97, 98, 107*

[134] Baker Z, Przeworski M, Sella G. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence; 2022. Available from: https://www.biorxiv.org/content/10.1101/2022.09.27.509707v1. *Cited at pages 30, 36, 37, 43, 61, 80, 96, 107, 110, 112, 115*

[135] Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. PLOS Genetics. 2016;12(6):e1006146. doi:10.1371/journal.pgen.1006146. *Cited at page 40*

[136] Vara C, Capilla L, Ferretti L, Ledda A, Sánchez-Guillén RA, Gabriel SI, et al. PRDM9 Diversity at Fine Geographical Scale Reveals Contrasting Evolutionary Patterns and Functional Constraints in Natural Populations of House Mice. Molecular Biology and Evolution. 2019;36(8):1686–1700. doi:10.1093/molbev/msz091. *Cited at page 41*

[137] Smagulova F, Brick K, Pu Y, Camerini-Otero RD, Petukhova GV. The evolutionary turnover of recombination hot spots contributes to speciation in mice. Genes & Development. 2016;30(3):266–280. doi:10.1101/gad.270009.115. *Cited at pages 41, 56, 57, 62, 80, 91, 96, 97, 108, 109, 112, 115, 136*

[138] Baker CL, Petkova P, Walker M, Flachs P, Mihola O, Trachtulec Z, et al. Multimer Formation Explains Allelic Suppression of PRDM9 Recombination Hotspots. PLoS Genetics. 2015;11(9):e1005512. doi:10.1371/journal.pgen.1005512. *Cited at pages 53, 58, 59, 62, 99, 108, 115*

[139] Salcedo T, Geraldes A, Nachman MW. Nucleotide Variation in Wild and Inbred Mice. Genetics. 2007;177(4):2277–2291. doi:10.1534/genetics.107.079988. *Cited at page 56*

[140] Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Research. 2015;25(8):1125–1134. doi:10.1101/gr.186148.114. *Cited at page 56*

[141] Wu H, Mathioudakis N, Diagouraga B, Dong A, Dombrovski L, Baudat F, et al. Molecular Basis for the Regulation of the H3K4 Methyltransferase Activity of PRDM9. Cell Reports. 2013;5(1):13–20. doi:10.1016/j.celrep.2013.08.035. *Cited at page 60*

[142] Grey C, Barthès P, Friec GCL, Langa F, Baudat F, Massy Bd. Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. PLOS Biology. 2011;9(10):e1001176. doi:10.1371/journal.pbio.1001176. *Cited at page 60*

[143] Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, et al. Factors Influencing Meiotic Recombination Revealed by Whole Genome Sequencing of Single Sperm. Science (New York, NY). 2019;363(6433):eaau8861. doi:10.1126/science.aau8861. *Cited at pages 63, 115*

[144] Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nature Reviews Genetics. 2014;15(3):176–192. doi:10.1038/nrg3644. *Cited at page 79*

[145] Muller HJ, Pontecorvo G. Recessive genes causing interspecific sterility and other disharmonies between Drosophila melanogaster and simulans. Genetics. 1942;27:157. *Cited at page 79*

[146] AbuAlia KF, Damm E, Ullrich KK, Mukaj A, Parvanov E, Forejt J, et al.. Natural variation in Prdm9 affecting hybrid sterility phenotypes; 2023. Available from: https://www.biorxiv.org/content/10.1101/2023.01.17.524418v1. *Cited at pages 80, 96, 97, 117, 118*

[147] Seroussi E, Shirak A, Gershoni M, Ezra E, de Abreu Santos DJ, Ma L, et al. Bos taurus–indicus hybridization correlates with intralocus sexual-conflict effects of PRDM9 on male and female fertility in Holstein cattle. BMC Genetics. 2019;20(1):71. doi:10.1186/s12863-019-0773-5. *Cited at page 80*

[148] Schwarz T, Striedner Y, Horner A, Haase K, Kemptner J, Zeppezauer N, et al. PRDM9 forms a trimer by interactions within the zinc finger array. Life Science Alliance. 2019;2(4). doi:10.26508/lsa.201800291. *Cited at pages 99, 115*

[149] Charlesworth B. RECOMBINATION MODIFICATION IN A FLUCTUATING ENVIRONMENT. Genetics. 1976;83(1):181–195. doi:10.1093/genetics/83.1.181. *Cited at page 106*

[150] Feldman MW, Christiansen FB, Brooks LD. Evolution of recombination in a constant environment. Proceedings of the National Academy of Sciences. 1980;77(8):4838–4841. doi:10.1073/pnas.77.8.4838. *Cited at page 106*

[151] Otto SP, Barton NH. The Evolution of Recombination: Removing the Limits to Natural Selection. Genetics. 1997;147(2):879–906. doi:10.1093/genetics/147.2.879. *Cited at pages 106, 119*

[152] Otto SP, Feldman MW. Deleterious Mutations, Variable Epistatic Interactions, and the Evolution of Recombination. Theoretical Population Biology. 1997;51(2):134–147. doi:10.1006/tpbi.1997.1301. *Cited at page 106*

[153] Lenormand T, Otto SP. The Evolution of Recombination in a Heterogeneous Environment. Genetics. 2000;156(1):423–438. doi:10.1093/genetics/156.1.423. *Cited at page 106*

[154] Felsenstein J. THE EVOLUTIONARY ADVANTAGE OF RECOMBINATION. Genetics. 1974;78(2):737–756. doi:10.1093/genetics/78.2.737. *Cited at page 106*

[155] Smith JM. The evolution of recombination. Journal of Genetics. 1985;64(2):159–171. doi:10.1007/BF02931144. *Cited at page 106*

[156] Feldman MW, Otto SP, Christiansen FB. Population Genetic Perspectives on the Evolution of Recombination. Annual Review of Genetics. 1996;30(1):261–295. doi:10.1146/annurev.genet.30.1.261. *Cited at page 106*

[157] Felsenstein J, Yokoyama S. THE EVOLUTIONARY ADVANTAGE OF RECOMBINATION. II. INDIVIDUAL SELECTION FOR RECOMBINATION. Genetics. 1976;83(4):845–859. doi:10.1093/genetics/83.4.845. *Cited at page 119*

[158] Iles MM, Walters K, Cannings C. Recombination Can Evolve in Large Finite Populations Given Selection on Sufficient Loci. Genetics. 2003;165(4):2249–2258. doi:10.1093/genetics/165.4.2249. *Cited at page 119*

[159] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966;8(3):269–294. doi:10.1017/S0016672300010156. *Cited at page 120*

[160] Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974;23(1):23–35. doi:10.1017/S0016672300014634. *Cited at page 120*

[161] Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–1303. doi:10.1093/genetics/134.4.1289. *Cited at page 120*

[162] Mihola O, Pratto F, Brick K, Linhartova E, Kobets T, Flachs P, et al. Histone methyltransferase PRDM9 is not essential for meiosis in male mice. Genome Research. 2019;29(7):1078–1086. doi:10.1101/gr.244426.118. *Cited at page 122*

[163] Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science. 2016;352(6284):474–477. doi:10.1126/science.aac8624. *Cited at page 122*