The evolution of recombination in self-fertilising species: theoretical approach and genomic study of linkage disequilibrium between deleterious mutations in several Angiosperm species

Roman Stetsenko

Thèse de doctorat de Sorbonne Université

Ecole doctorale 227 :
Sciences de la Nature et de l'Homme : Evolution et Ecologie
MNHN - Sorbonne Université

En vue de l'obtention du grade de
Docteur de Sorbonne Université

L'évolution de la recombinaison chez les espèces autogames : approche théorique et étude génomique du déséquilibre de liaison entre mutations délétères chez plusieurs espèces d'Angiospermes

The evolution of recombination in self-fertilising species: theoretical approach and genomic study of linkage disequilibrium between deleterious mutations in several Angiosperm species

Présentée et soutenue publiquement par

Roman Stetsenko

Le 18 Septembre 2023

Membres du jury

Sylvain GANDON	Directeur de recherche, CNRS Montpellier	Rapporteur
Susan JOHNSTON	Research fellow, University of Edinburgh	Rapportrice
Guillaume ACHAZ	Professeur, Université Paris Cité	Examinateur
Claire MÉROT	Chargée de recherche, CNRS Rennes	Examinatrice
Denis ROZE	Directeur de recherche, CNRS Roscoff	Directeur de thèse
Henrique TEOTÓNIO	Maître de conférence, ENS Paris	Membre invité

Remerciements/Acknowledgements

Je tiens à remercier Sylvain Gandon et Susan Johnston d'avoir accepté d'être rapporteurs de ma thèse ainsi que Guillaume Achaz et Claire Mérot d'avoir accepté de faire partie de mon jury de thèse.

Un grand merci à Denis pour ces "presque" quatre années d'encadrement. J'ai apprécié la liberté que tu m'as laissée d'explorer différents aspects de ma thèse, quitte à s'aventurer vers des approches que tu maîtrisais moins, ce qui m'a incité à collaborer avec d'autres personnes dans divers labos. J'ai également apprécié ton infinie patience pour expliquer, même plusieurs fois s'il le fallait, les concepts, les développements mathématiques et les programmes de simulation. Cela a toujours été un plaisir de discuter de science (mais pas que) avec toi. Enfin, j'ai beaucoup appris de ta rigueur scientifique, qu'elle soit conceptuelle ou technique et de ton souci du détail.

Thanks, Henrique, for your co-supervision, for having welcomed me in your team and for having allowed this collaboration with Tom during which I've learned a lot about experimental evolution in *C. elegans*. Merci, Tom, pour les discussions autour de nos travaux respectifs et les comités de suivi de thèse communs. C'était très intéressant et j'ai appris énormément de choses en suivant votre expérience très ingénieuse avec Henrique pour tester des attendus théoriques sur l'évolution de la recombinaison.

Je tiens à remercier les membres de mon comité de thèse pour avoir donné de leur temps et pour leurs conseils sur l'organisation du travail de thèse.

Merci, Sylvain, de m'avoir fait découvrir le monde de la génomique des populations, de m'avoir accueilli dans ton équipe et d'avoir pris de ton temps pour me présenter le jeu de données, ainsi que pour les nombreuses discussions qui ont été essentielles concernant la partie génomique de la thèse. Merci, Claire, pour ton aide précieuse sur l'analyse des variants structuraux et d'avoir pris de ton temps pour discuter des résultats.

Merci aux membres de l'IRL 3614 de m'avoir accueilli dans l'équipe et d'avoir contribué à ce que ma thèse se passe dans les meilleures conditions, à Myriam et Christophe pour votre enthousiasme pour la génétique des populations mais aussi pour la lutte politique ainsi qu'à Barbara pour ta bonne humeur et ton aide indispensable. Merci, Jérôme et Lucie, pour votre agréable compagnie dans le bureau 201.2.

Thanks, Tianlin, for having provided me with your genome of *Capsella rubella* and for your valuable advice in genomic analyses.

Merci, Lauric, pour tes précieux conseils en génomique et en bioinfo qui m'ont été bien utiles.

Merci, Lise, de m'avoir aidé pendant ton stage à analyser les données d'*Arabisposis thaliana* et de t'être confrontée à la génétique des populations avec ta formation en bioinformatique.

Merci aux membres du groupe informel "Théorie et modélisation" pour les *journal clubs* et les discussions autour des recherches de chacun. Merci au personnel de la station qui a contribué de près ou de loin à ce que je puisse travailler dans les meilleures conditions, notamment le service ABiMS qui m'a permis de réaliser mes analyses.

Merci à toutes celles et ceux avec qui j'ai pu passer de bons moments à la station ou en dehors, notamment Sylvie, Seb, Charlotte, Nico, Pélagie, Aline, Clara, P-G, Louise, Emma, Amine, Sam, Louison, Aurélien, Jean, Lisa, Jeremy, Yasmine, Sonia, Iris, Nathan, Rémi, Victoire, Antonin, Tanweer, Théo, Camille, Vittoria et beaucoup d'autres.

Merci à Sonia J. de m'avoir aidé pour l'installation à St-Pol et pour les nombreux coups de mains.

Et enfin, merci à ma famille et à mes amis d'un peu partout, pour leur soutien, y compris dans les moments plus difficiles.

Résumé étendu

La recombinaison génétique est un élément constitutif du cycle sexué des Eucaryotes, et est souvent considérée comme l'un des principaux avantages de la reproduction sexuée. Elle est la conséquence de crossing-over se produisant pendant la méiose, permettant la formation de gamètes recombinants. Or, le nombre et la position de ces crossing-over le long du génome est très variable entre espèces, ainsi qu'entre individus au sein d'une même espèce. Malgré le fait que des contraintes mécaniques liées à la ségrégation des chromosomes pendant la méiose peuvent imposer un nombre minimal et maximal de crossing-over, plusieurs résultats empiriques montrent que les taux de recombinaison peuvent évoluer rapidement. De plus, on observe en général des taux de recombinaison élevés chez les Eucaryotes. Une des premières explications pour le maintien de la recombinaison met en avant le fait que celle-ci serait avantagée car elle permettrait d'augmenter la variance du succès reproducteur entre individus, favorisant ainsi l'adaptation. Cependant, la recombinaison n'augmente pas nécessairement la variance du succès reproducteur et, depuis les années 1960, de nombreux travaux théoriques se sont attachés à étudier les conditions sous lesquelles la recombinaison peut être maintenue par son effet sur les combinaisons génétiques. Malgré les avancées majeures dans la compréhension des mécanismes pouvant favoriser la recombinaison qu'ont permises ces travaux, ils ne permettent toujours pas d'expliquer les forts taux de recombinaison maintenus chez les Eucaryotes. En outre, les modèles actuels sur l'évolution de la recombinaison font généralement des hypothèses simplificatrices sur la variation des taux de recombinaison le long du génome et son architecture génétique. Incorporer des hypothèses plus réalistes dans ces modèles pourrait permettre de mieux comprendre comment des taux de recombinaisons élevés peuvent être maintenus par la sélection naturelle. De plus, la plupart de ces modèles se sont concentrés sur des organismes à fécondation croisée et l'évolution de la recombinaison chez les espèces se reproduisant par autofécondation a été beaucoup moins étudiée. De façon intéressante, on observe souvent des taux de recombinaison plus élevés chez les espèces hermaphrodites autogames par rapport aux espèces allogames, suggérant que la recombinaison est plus fortement favorisée chez les autogames.

Une première partie de la thèse a consisté à développer des modèles théoriques permettant de mieux comprendre cet effet de l'autofécondation sur l'évolution de la recombinaison. Un premier modèle a considéré l'évolution d'un gène "modificateur" affectant le nombre moyen de crossing-over par chromosome, en présence de mutations délétères

se produisant le long des chromosomes. Les approximations analytiques ainsi que les résultats de simulation montrent que, sous des valeurs de paramètres réalistes, la sélection pour la recombinaison est généralement plus forte chez les espèces autogames ; par ailleurs, cette sélection est principalement due au déséquilibre de liaison négatif entre mutations délétères généré par l'effet Hill-Robertson (un effet stochastique lié à la taille finie des populations). Ces résultats s'expliquent par le fait que la recombinaison est moins efficace pour créer de nouvelles combinaisons génétiques quand l'homozygotie générée par l'autofécondation est plus importante. Des taux de recombinaison plus élevés sont donc sélectionnés car ils permettent d'augmenter la variance du succès reproducteur entre descendants, rendant la sélection plus efficace pour éliminer les mutations délétères. Ce modèle pourrait donc expliquer la corrélation positive entre taux de recombinaison et taux d'autofécondation. Un deuxième modèle de simulation a exploré des scénarios plus réalistes concernant l'architecture génétique de la variation des taux de recombinaison, ainsi que la distribution des crossing-over le long des chromosomes. Les résultats montrent notamment que des taux de recombinaison plus élevés peuvent être maintenus lorsque cette distribution n'est pas uniforme, ou lorsque l'effet des modificateurs de recombinaison est restreint à une portion de chromosome. Ces résultats suggèrent que la structure du génome et son caractère non-uniforme ainsi que l'architecture génétique de la variation des taux de recombinaison sont des éléments importants pour comprendre l'évolution de la recombinaison.

La deuxième partie de la thèse a porté sur une estimation empirique du déséquilibre de liaison entre mutations délétères (une composante importante de la sélection pour la recombinaison), en utilisant des données génomiques issues de populations naturelles de la plante allogame *Capsella grandiflora*, ainsi que des plantes fortement autogames *Arabidopsis thaliana* et *Capsella orientalis*. Cette étude met en avant plusieurs biais méthodologiques pouvant générer du déséquilibre de liaison positif entre mutations délétères. Une première source de biais se produit lorsque l'analyse est restreinte aux mutations présentes en faible fréquence. En effet, des mutations en fréquences similaires ont tendance à être en déséquilibre de liaison positif. Une deuxième source de biais peut être générée par la présence de duplications présentes à l'état polymorphe dans l'échantillon, et absentes du génome de référence. Ainsi, le déséquilibre de liaison positif entre mutations délétères observé lors de précédents travaux (notamment chez *C. grandiflora*) résulte probablement en grande partie de ces biais. Néanmoins, même avoir pris en compte ces différents biais, du déséquilibre de liaison positif persiste entre mutations délétères et pourrait être causé par des duplications non-détectées, d'autres biais méthodologiques ou par des effets épistatiques. Chez l'espèces autogame *A. thaliana*

(chez qui les duplications présentes à l'état polymorphe sont plus faciles à détecter) on observe néanmoins du déséquilibre de liaison positif entre mutations délétères, pouvant être causé par des effets épistatiques ou par la forte structure spatiale observée chez cette espèce.

Les implications de ces différents résultats théoriques et empiriques sont enfin discutées afin d'en dégager des perspectives pour des travaux futurs.

Contents

Part I

Introduction

1 Recombination: at the origin of sexual reproduction

1.1 Variation in fitness as the fuel for natural selection

Genetic variation – defined as differences in the DNA sequence between individuals in a population – is one of the fundamental ingredients of evolutionary change whether it is caused by neutral of selective processes (Nei et al., 2010). The ultimate source of genetic variation is the imperfectness of the process of DNA replication that generates mutations during cell division. These mutations can consist in the change of one nucleotide into another (point mutation) or in the change of larger portions of DNA that can alter the structure of the genome (structural variation; Mérot et al. 2020). New genotypes that differ from the others by at least one mutation are called variants. The majority of mutations have almost no consequence on the reproductive success of individuals (and are thus effectively neutral) and their frequency is submitted to random fluctuations caused by genetic drift. Other mutations are submitted to natural selection to various degrees, most of them decreasing the reproductive success (fitness) of individuals (deleterious mutations) and a small proportion increasing it (beneficial mutations) (Eyre-Walker and Keightley, 2007). Because genetic drift is increased in smaller populations while selection may change with the environment, the same mutation can have different fates depending on the context where it appears. Furthermore, it is important to note that selection does not act directly on the genotype but on the phenotype of individuals. The phenotype is the result of the interaction between a large number of genes and environmental effects. The field of quantitative genetics aims at studying the evolution of traits resulting from the effect of a large number of genes (quantitative traits) that vary continuously and thus explores how selection acts on the genotype through the phenotype. In the present thesis we adopt the approach of population genetics which often considers genes with independent effects on fitness and makes the simplifying assumption of a direct relationship between genotype and fitness (ignoring the phenotypic level). However, the limit between population genetics and quantitative genetics can be blurry as the interaction between genes on fitness (considered as a phenotype), epistasis, is often incorporated

into populations genetics models. Whether a phenotypic level is explicitly considered or not, the genetic variation that matters for natural selection is variation in fitness. As first described by Darwin and Wallace (1858), heritable variation in fitness between individuals in a population is the fuel of natural selection. A more modern formulation by Fisher's fundamental theorem of natural selection states that selection increases the mean fitness of a population proportionally to the additive genetic variance in fitness present in this population (Fisher, 1930; Frank and Slatkin, 1992). The additive genetic variance in fitness represents the variance in fitness among individuals in a population due to the additive component of the alleles they carry. This measure represents the evolvability of a population, namely, its potential to respond to selection (Houle, 1992; Hansen et al., 2011). The concept of evolvability is related to the concept of adaptation, referring to the increase of the mean fitness of a population in response to selective pressures (Burt, 1995). Populations generally face changes of their environment – abiotic or biotic – that decrease mean fitness, so that alleles or combinations of alleles that were previously deleterious or neutral can become advantageous. Even in a stable environment, populations need to counter the decrease in mean fitness caused by the constant occurrence of deleterious mutations (mutation load; Lynch and Gabriel 1990; Burt 1995).

1.2 Sex and recombination

The exchange of genetic material between individuals via various forms of sex and recombination has been hypothesised as a mechanism that has evolved at least in part to regenerate variance in fitness (Otto, 2009). However, the question of the relative importance of different selective forces maintaining sex and recombination remains highly debated (West et al. 1999; Gouyon 1999; Otto 2009; see also section *Modelling the evolution of recombination rates* in the present introduction). Indeed, sexual reproduction can have many important costs such as the cost of males in anisogamous species, or costs associated with the mating process (Lehtonen et al., 2012) and it is not yet totally clear whether these costs can be compensated by indirect benefits associated with the production of new genotypes.

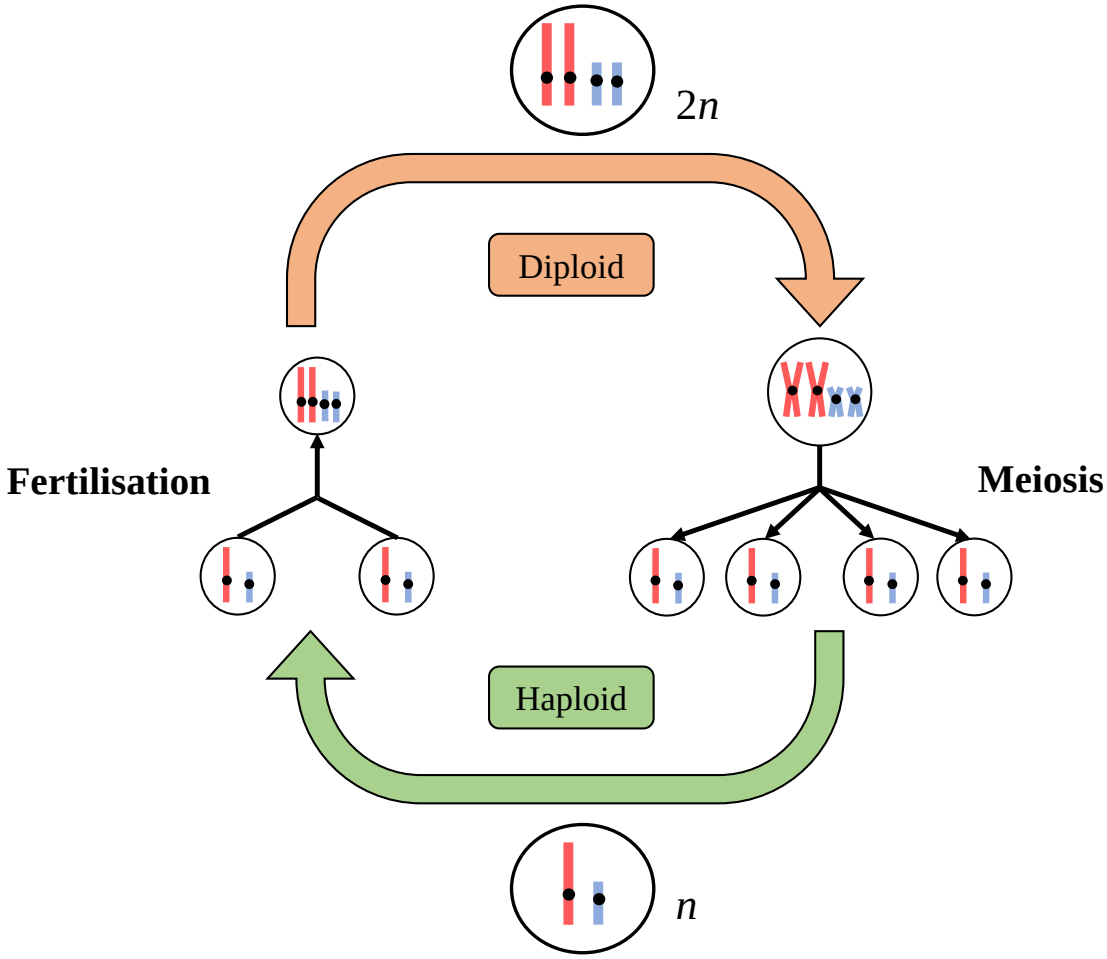


Figure 1: The sexual life cycle of eukaryotes characterised by the alternation between a haploid (n) and a diploid phase ($2n$). Genetic material is exchanged between individuals in the form chromosomes that combine in pairs after fertilisation and separate into haploid cells during meiosis. Note that the chromosomes of the mother cell are replicated before meiosis.

Sex can be broadly defined as the the exchange of genetic material between two individuals. It is present in all major groups of organisms such as prokaryotes that can unidirectionally exchange DNA via different mechanisms such as conjugation, transformation or transduction (Redfield, 2001; Narra and Ochman, 2006). Viruses can also display a form of sex when they co-infect a host cell (Pérez-Losada et al., 2015). However, the extent and the evolutionary consequences of sex and recombination in those groups remain debated (Redfield, 2001; Vos, 2009; Pérez-Losada et al., 2015). Eukaryotes have evolved a typical life cycle where sex involves an alternation between a phase in which the genome is in a single copy (n, haploid) and a phase in which it is in two copies ($2n$, diploid; Figure 1). Two haploid cells fuse during fertilisation to produce a diploid cell, and a diploid cell gives rise to haploid cells during meiosis. The relative duration and degree of development of these phases vary importantly across eukaryotes, with the diploid phase dominating in many plants and animals, while in other eukaryotic organisms either the diploid or the haploid phase may dominate, or development may be equally important in both phases (Valero et al., 1992; Mable and Otto, 1998). In the sexual life cycle of eukaryotes, at each generation an individual combines half of its genome (in the form of chromosomes) with another individual. Genes present on the same chromosome are thus physically linked and a first level of genetic shuffling (called segregation) occurs during meiosis when the two copies of each chromosome (maternally or paternally inherited) are independently distributed between the daughter cells during anaphase I (Figure 2).

A second level of genetic shuffling, called meiotic recombination, occurs through the pairing of homologous chromosomes during prophase I, visible as chiasmata. This pairing can result into the reciprocal exchange of homologous portions of chromosome, called crossover (CO; see Frame 1 for further details). Crossovers generate recombinant chromosomes presenting a mix of paternally and maternally inherited alleles. In eukaryotes, homologous recombination can also occur during mitosis as a mechanism for DNA repair, although it usually does not result in crossover (Lafave and Sekelsky, 2009). It has been hypothesised that homologous recombination first evolved as a DNA-repair mech-

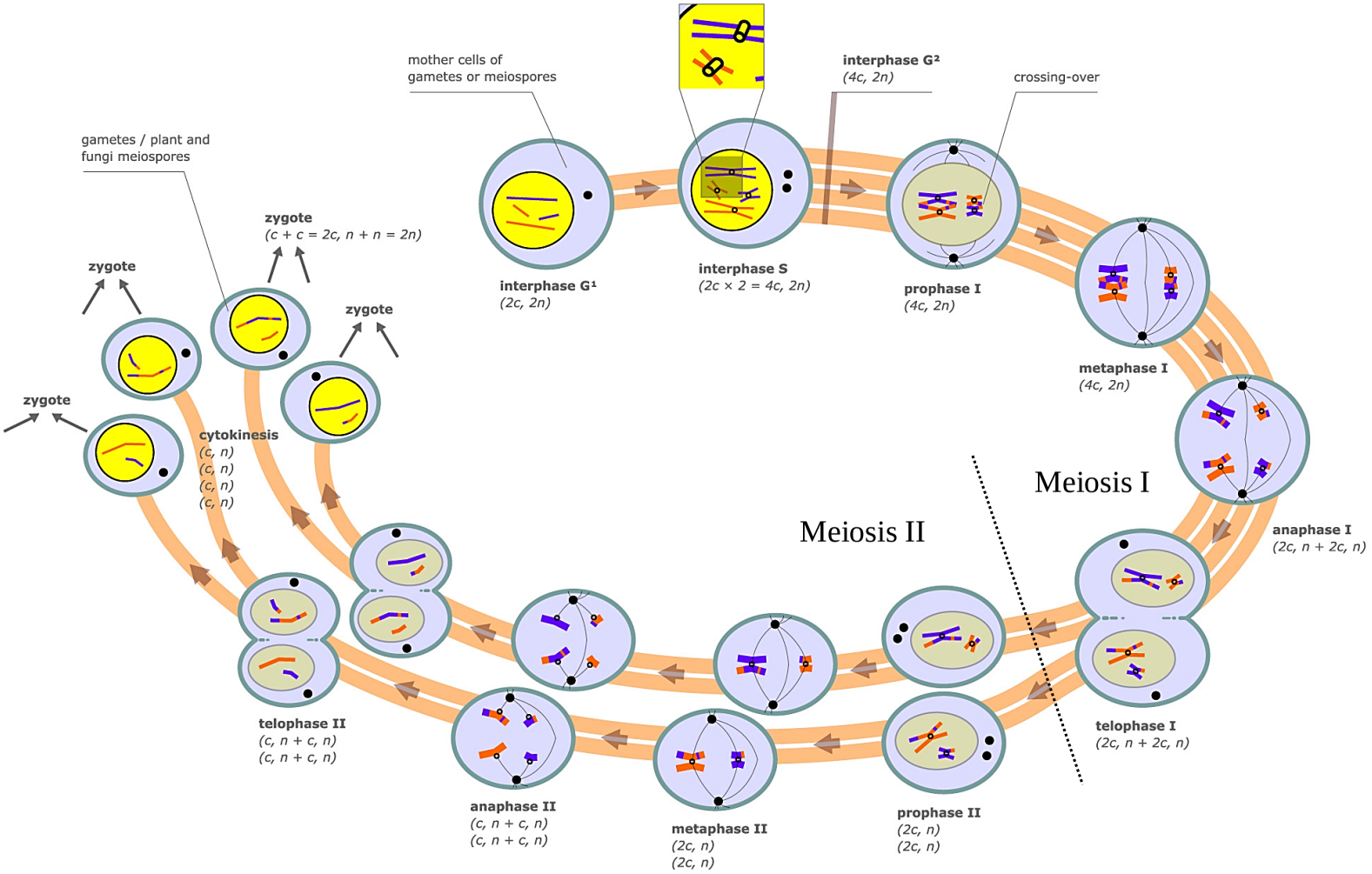


Figure 2: Schematic representation of meiosis in a cell with two pairs of chromosomes (a short and a long one) with different colours according to the maternal or paternal origin of chromosomes. At each step, the quantity of DNA (c) and the ploidy (n) are indicated. The duplication of the genome occurs during the interphase (the two sister chromatids remaining tied by the centromere) followed by two successive divisions: meiosis I and meiosis II. During meiosis I, homologous chromosomes first pair and usually form chiasmata that can result in crossovers (CO; prophase I). The two homologous chromosomes are then split into different daughter cells (metaphase, anaphase and telophase I). During meiosis II the two sister chromatids of each chromosome are separated into two cells yielding four haploid cells in total. Modified from Wikimedia Commons | Marek Kultys (www.marekkultys.com) | CC BY-SA 3.0.

anism and was then co-opted to ensure genetic shuffling during meiosis (Cavalier-Smith, 2002). Hereafter we will refer to recombination as the meiotic recombination process of eukaryotes.

Frame 1: Molecular mechanisms of crossover formation

Crossover formation starts at the beginning of prophase I with the formation of multiple double strand breaks (DSBs) along the chromosome by the protein SPO11 binding to DNA (Figure 3). Although DSBs can occur without it, SPO11 catalyses the formation of DSBs and is a highly conserved protein among eukaryotes (de Massy, 2013). Proteins RAS51 and DMC1 allow single-end invasion of the excised chromatid into the homologous region of the other chromatid (Zickler and Kleckner, 2015). A series of less conserved proteins can then resolve this strand invasion by a reciprocal exchange of homologous DNA called crossover (CO). It can also be resolved by a non-CO, resulting in gene conversion in which a portion of the donor chromosome is copied into the homologous region of the other chromosome. A small fraction of DSBs gives COs and the majority of them are referred to as class I COs as they are formed by a particular set of proteins (ZMM). Class I COs are subjected to CO interference, the phenomenon by which the formation of a CO prevents the formation of other COs nearby.

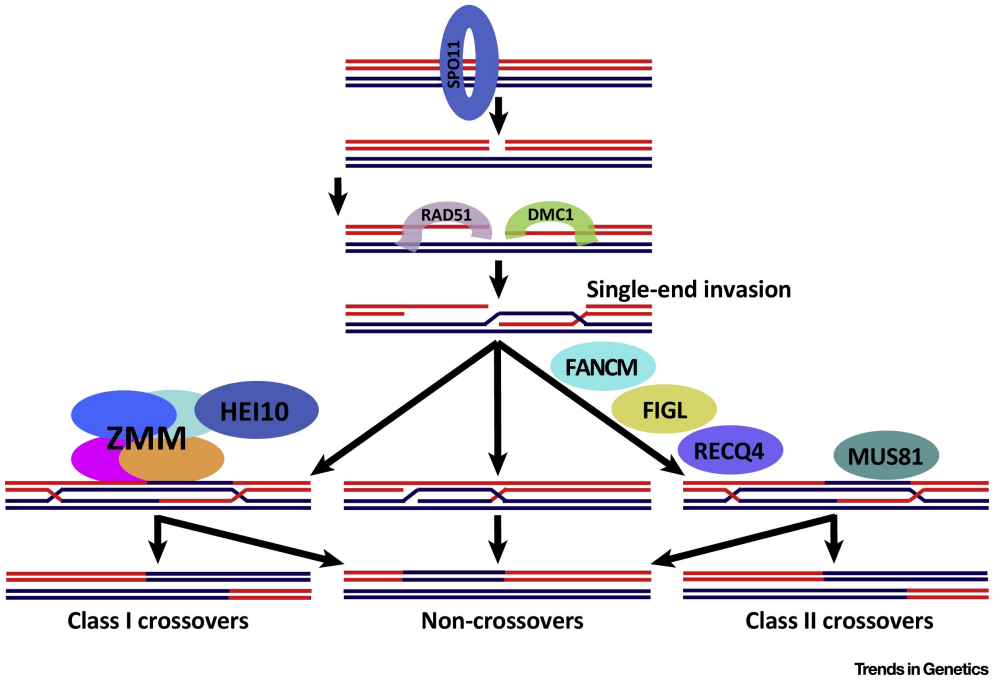


Figure 3: Simplified diagram of crossover formation where only one chromatid for each chromosome is represented. From Zelkowski et al. (2019).

2 Recombination rate variation and its genetic bases

2.1 Measuring recombination rates

The effect of COs on the genetic shuffling between pairs of loci on the same chromosome is measured as a recombination rate which is the proportion of recombinant genotypes following meiosis. Imagine two loci on the same chromosome with the following genotype: AB/ab with alleles A and B on one chromosome and a and b on the other chromosome. Among a large number of meiotic products, the proportion of segregating recombinant genotypes Ab and aB corresponds to the recombination rate between the two loci. Two loci segregate independently when they are located on different chromosomes, in which case the recombination rate equals 0.5. However, the recombination rate between two loci is not completely informative about the average number of CO occurring between them, defined as the genetic distance – measured in Morgans. Indeed, for recombination to occur between two loci on the same chromosome their has to be an odd number of CO between them. Moreover, the probability distribution of COs can be affected by several phenomena such as CO interference. Because of this, the genetic distance can be computed from the recombination rate using mapping functions (Zhao and Speed, 1996). The number and approximate position of COs can also be directly inferred from the observation of chiasmata during prophase I. Recombination rates can be estimated from genomic data at different scales by genotyping a large number of gametes from a single individual, direct relatives of a pedigree or different individuals of the same population (Peñalba and Wolf, 2020). In order to compare recombination between different genomic regions, individuals or species, recombination rate can be expressed as an average number of CO (or map length) per unit of physical distance (often in cM/Mb).

2.2 Recombination rate variation

Measures of recombination rates in experimental and natural populations have shown that recombination rates vary at different scales: within and between chromosomes,

within and between individuals, between sexes and between populations and taxa (Stapley et al. 2017; Haenel et al. 2018; Peñalba and Wolf 2020; Sardell and Kirkpatrick 2020; Figure 4). Crossovers do not occur randomly along chromosomes and recombination rates can vary at fine genomic scales (a few kb) from several orders of magnitude, between regions of very low (coldspots) or very high (hotspots) recombination. At the chromosome scale, patterns in the recombination landscape (variation of the recombination rate along the chromosome) are observed, with often higher recombination rates near telomeres, while recombination is typically suppressed around centromeres (Haenel et al., 2018; Brazier and Glémin, 2022). Average recombination rates between chromosomes can also greatly differ in the same genome (e.g. in Angiosperms; Brazier and Glémin 2022). Differences in recombination rates on autosomes are often observed between the males and females of the same species, called heterochiasmy (Lenormand, 2003). In the extreme case, recombination can be absent in one sex – a phenomenon called achiasmy – as described in many arthropod species (Satomura et al., 2019). A shared pattern of most vertebrates, plants and mollusks is a higher recombination rate towards telomeres in males wile recombination is more uniformly distributed along chromosomes in females (Sardell and Kirkpatrick, 2020). The differences in recombination landscapes between males and females could either be a byproduct of differences between male and female meiosis or be adaptive although no model is yet fully satisfying (Sardell and Kirkpatrick, 2020). Finally, recombination rates can vary by several orders of magnitude between large phylogenetic groups, with SAR (Stramenopiles-Alveolates-Rhizaria) having the highest genome-wide recombination rates and conifers the lowest (Stapley et al., 2017).

2.3 Direct and indirect selective forces acting on recombination

The literature on the evolution of recombination classically partitions selective forces acting on recombination rates among direct and indirect forces. Direct selection corresponds to direct fitness effects caused by changes in recombination rates, which may stem from physical constrains imposing limits to the number of CO per meiosis, while

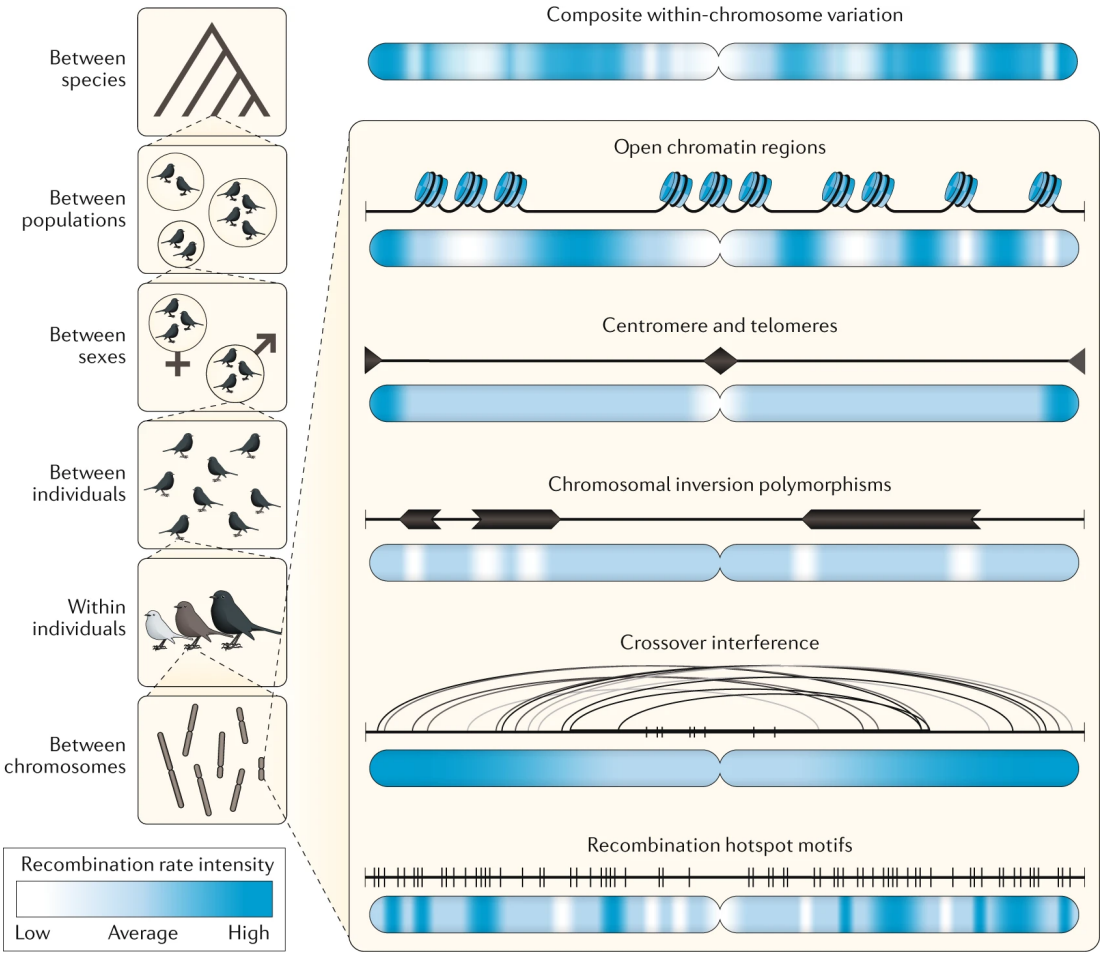


Figure 4: The different scales at which recombination rates can vary. From Peñalba and Wolf (2020). See main text for details.

indirect selection stems from the effect of recombination in breaking or creating genetic combinations.

Direct forces

Recombination landscapes are caused by the distribution of COs during meiosis which can stem from different molecular constrains. Crossovers tend to occur in open chromatin regions, so that factors modifying chromatin are associated with recombination rate variation; in particular, open chromatin patterns determine the position of recombination hotspots (Brachet et al., 2012). In many vertebrates, the position of hotspots is determined by the well described transcription factor PRDM9 that binds specific DNA sequences and creates an open chromatin environment favouring DSBs (Gray et al., 2018; Kenneth Paigen and Petkov, 2018). PRDM9 is active in most vertebrates and the position of the sequence binding site determines the position of recombination hotspots. The position of these hotspots is poorly conserved between species as it evolves rapidly due to biased gene conversion – generated by COs – at the sequence binding site (Latrille et al., 2017; Genestier et al., 2023). In vertebrates that have lost PRDM9 (birds, dogs, reptiles) hotspots tends to occur in gene promoter regions like in plants and yeast (Auton et al., 2013; Choi and Henderson, 2015; Singhal et al., 2015). This recombination centered on gene promoters could explain in part the positive correlation between gene density and recombination rate in plants (Haenel et al., 2018; Brazier and Glémin, 2022). In other eukaryotes such as *Drosophila* and *Caenorhabditis elegans*, hotspots are absent (Kaur and Rockman, 2014; Smukowski Heil et al., 2015). Overall, explanations for the evolution of these different systems determining the fine-scale position of COs are still lacking.

At the chromosome scale, it has been proposed that the periphery-bias in recombination rate (higher density of COs near the telomeres) could come from the fact that homology search and DSBs are initiated from the telomeres (Zickler and Kleckner, 2015; Haenel et al., 2018). In addition, lower recombination rates in the centromeric region are caused by the kinetochores suppressing DSBs and COs (Fernandes et al., 2019).

Crossover interference, by preventing COs to occur too close to each other, spreads the location of COs and reduces their number, which may have evolved to limit the number of CO per chromosome (Libuda et al., 2013; Otto and Payseur, 2019). Indeed, too many COs during meiosis can lead to the non-disjonction of chromosomes and aneuploidy (Koehler et al., 1996). However, this upper limit does not seem to be absolute as it was shown that a greatly increased number of CO does not seem to impair fitness significantly in mutants of *Arabidopsis thaliana* (Fernandes et al., 2018). Structural variants can also influence recombination such as inversions that play a central role in recombination suppression between sex chromosomes (Kirkpatrick, 2010). Finally, an obligate crossover per bivalent per meiosis is generally thought to be required to ensure the proper segregation of homologs during meiosis I. This constraint is not absolute, however, since in achiasmate species one sex performs meiosis without any CO. As the number of CO per chromosome does not vary much (typically from 1 to 3, 4; Fernandes et al. 2018) the difference in average recombination rate between chromosomes within a species is mainly explained by their size (Stapley et al., 2017; Haenel et al., 2018; Brazier and Glémin, 2022).

Indirect forces

Although mechanical contraints may act on the number of COs, substantial variation in recombination rates can be observed between individuals of the same population. Indeed, variation in the genome-wide and/or local recombination rate between individuals has been described either in model species (e.g fruit flies; Singh et al. 2015), domesticated species (e.g. house mice, pigs, cattle, sheep, maize, honey bees; Dumont et al. 2009; Bauer et al. 2013; Ma et al. 2015; Petit et al. 2017; Kawakami et al. 2019; Brekke et al. 2022) or in natural populations (e.g. humans, red deer, Soay sheep, wild mice; Wang et al. 2017; Johnston et al. 2016, 2018; Halldorsson et al. 2019). This variation in recombination has a heritable genetic basis, with heritability ranging from 8% (Johnsson et al., 2021) to 48% (Dumont et al., 2009). Some of the genes found in the regions that contribute to variation in the recombination rate are shared among several mammals

such as PRDM9, RNF212 or HEI10. Therefore, heritable variation in recombination rates exists within populations, and natural selection could potentially act upon this variation.

Moreover, a series of experiments have shown that recombination rates can evolve rapidly in response to strong selection applied on recombination rate itself or on other traits (reviewed in Otto and Barton 2001). For example, Korol and Iliadi (1994) reported in *Drosophila melanogaster* more than 50% increase in recombination rates over 50 generations of artificial selection for geotaxis, regardless of the direction of selection (positive of negative geotaxis, i.e. tendency to fly towards or away from the ground). The increase in recombination in response to strong selection could be interpreted as indirect selection on recombination, that is, as a way of increasing variation in fitness (Otto and Barton, 2001). It has long been hypothesised that sex and recombination are advantageous because they increase genetic variation and thus facilitate adaptation (Weismann, 1889). However, a large amount of theoretical work developed since the 60s has shown that this explanation is in fact not trivial: in particular, recombination does not always increase variation, and increasing variation is not always advantageous (reviewed by Otto 2009).

3 Modelling the evolution of recombination rates

A common approach to understand the selective forces acting on recombination rates is to use recombination modifier models, that represent a gene affecting recombination rates between other loci affecting fitness. Recombination modifiers are not mere theoretical constructs, since as we saw in the previous section, genetic variation for recombination rates exists within natural populations (reviewed in Gray and Cohen 2016; Stapley et al. 2017; Zelkowski et al. 2019). Although changing recombination rates may have direct fitness effects (see previous section), the recombination modifier gene is generally considered as neutral in order to isolate the indirect component of selection for recombination. The simplest recombination modifier models consider a recombination modifier

locus and two selected loci, in a haploid population. The modifier locus has a wild type allele coding for a recombination rate r between the two selected loci and a mutant allele coding for a different recombination rate $r + \delta$ – with $\delta > 0$ or $\delta < 0$ if the mutant allele increases or decreases recombination – while alleles affecting fitness segregate at the selected loci. This type of model can be used to explore under which conditions a mutant allele increasing recombination can be favoured (*i.e.*, increase in frequency). Because the modifier locus has no direct fitness effect, changes in allele frequencies at this locus are only due to genetic associations with the selected loci (hitchhiking effects). In particular, a modifier allele that tends to produce advantageous genotypes will then increase in frequency with these genotypes. This type of model has been used since the 60s under various assumptions, to explore the conditions under which recombination should be favoured, and what should be the evolutionary stable recombination rates.

A first series of works explored the fate of recombination modifiers in infinitely large, panmictic populations (no genetic drift, no population structure or migration, no inbreeding) at equilibrium around a fitness optimum (stabilizing selection; Kimura 1956; Nei 1967; Feldman 1972). In this case, modifiers decreasing the recombination rate are always favoured, as recombination decreases the mean fitness of offspring (Figure 5A). In other words, recombination is disfavoured at equilibrium because is breaks the advantageous genetic combinations built up by selection, a phenomenon known as the "reduction principle" (Feldman et al., 1997). However, it is not clear whether natural populations can truly reach a fitness optimum, even in a constant environment (Lenski et al., 2015).

A series of later models thus extended these first models to the case of directional selection, and found that selection on recombination depends on how alleles are associated within genomes and how their fitness effects interact (Feldman et al., 1980; Kondrashov, 1984; Charlesworth, 1990; Barton, 1995; Otto and Feldman, 1997). Indeed, the only effect of recombination is to break genetic associations among loci (Felsenstein, 1974). Genetic associations are classically measured by the linkage disequilibrium (LD). If one considers two loci with alleles a/A and b/B segregating at the first and the second locus, respectively, LD between alleles a and b is defined as $D_{ab} = p_{ab} - p_a p_b$, where p_{ab} is the

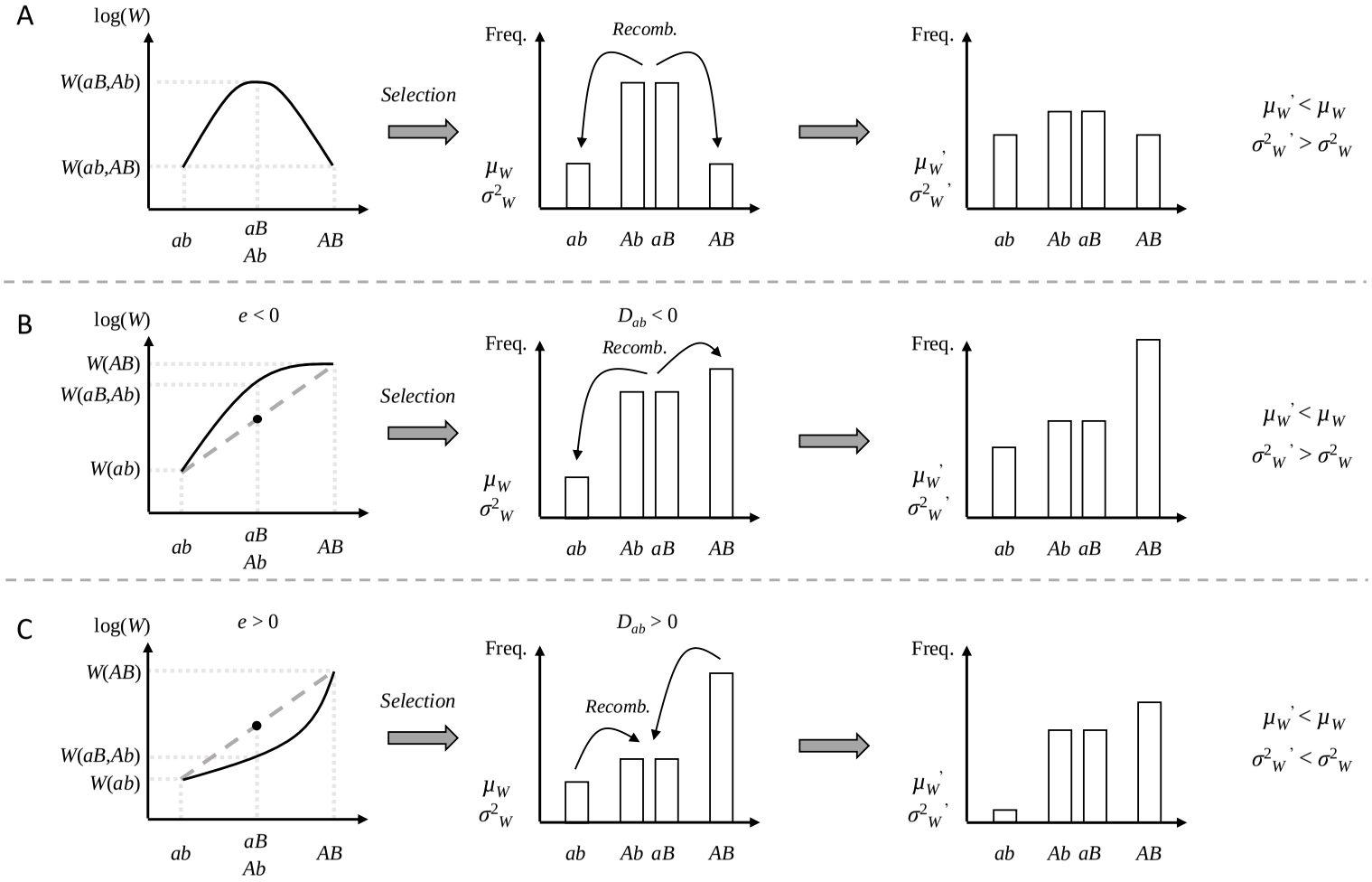


Figure 5: Effect of recombination in an infinitely large population where haplotype frequencies were built up by different types of selection (left panels): stabilising (A), directional with negative epistasis (B) or directional with positive epistasis (C). Fitness landscapes are represented on a log scale so that the grey dashed line represents a pure multiplicative landscape. Middle panels represent the effect of selection on haplotype frequencies – with a certain mean (μ_W) and variance in fitness (σ^2_W) – and right panels represent the effect of recombination on haplotype frequencies – with a new mean (μ'_W) and variance ($\sigma^{2\prime}_W$) in fitness. A: Under stabilising selection, recombination is disfavoured because high-fitness haplotypes (Ab, aB) are in relative excess and recombination tends to generate lower-fitness haplotypes (ab, AB) decreasing the average fitness of offspring. B: Under directional selection with weak negative epistasis recombination is favoured because the increase in the variance in fitness that it generates compensates the decrease in mean fitness of offspring. B: Under directional selection with positive epistasis recombination is disfavoured because it decreases both the mean and the variance in fitness among offspring.

frequency of haplotypes ab in the population, and p_a, p_b the frequencies of alleles a and b at each locus (Lewontin and Kojima, 1960; Weir, 1996). D_{ab} thus equals zero when allele a is not more (or less) associated with allele b than expected based on the frequency of allele b in the population, while D_{ab} is positive (negative) when the frequency of the ab haplotype is higher (lower) than expected based on allele frequencies. If alleles a and b are deleterious, positive LD corresponds to a relative excess of genotypes with extreme fitnesses (ab, AB) whereas negative LD corresponds to a relative excess of genotypes with intermediate fitnesses (aB, Ab; Figure 5C, B). Selection is more efficient at purging deleterious alleles and fixing advantageous alleles when LD is positive, as the variance in fitness within the population is higher. By contrast, negative LD tends to impede selection by reducing the variance in fitness. Recombination thus tends to be favored when LD is negative (as it increases the variance in fitness), and disfavored when LD is positive. This effect of recombination on the variance in fitness is sometimes referred to as the "long-term effect" of recombination. But under which conditions should LD between selected loci be positive or negative? In infinite, randomly mating populations, the only possible source of LD is epistasis. Positive epistasis means that the deleterious alleles a and b tend to compensate each other when combined, leading to a higher frequency of the ab haplotype than expected based on the frequencies of a and b ($D_{ab} > 0$), while negative epistasis means that the deleterious effects of a and b tend to reinforce each other when combined, leading to a lower frequency of the ab haplotype ($D_{ab} < 0$). Therefore, positive epistasis tends to generate positive LD, while negative epistasis generates negative LD (strictly, this is true when epistasis is measured on a multiplicative scale, that is, as a deviation from multiplicative effects of alleles at different loci; Felsenstein, 1965). From the reasoning above, one would thus expect that recombination should be favored under negative epistasis (as it increases the variance in fitness among offspring), but disfavored under positive epistasis (as it decreases the variance in fitness). However, breaking LD also has an effect on the mean fitness of offspring in the presence of epistasis: in particular, one can show that breaking genetic associations that have been built by selection is always disadvantageous in the short term. This effect on the mean fitness of offspring is

sometimes referred to as the "short-term effect" of recombination, and tends to disfavor recombination under both positive and negative epistasis. As a consequence, recombination is always disfavoured when epistasis is positive, as it increases the frequency of intermediate haplotypes (on average less fit than extreme haplotypes) and decreases the variance in fitness, thus suffering from a long and a short-term disadvantage (Figure 5C). However, recombination may be favoured when epistasis is weakly negative, so that the long-term advantage of increasing the variance in fitness outweighs the short therm disadvantage of decreasing mean fitness (Figure 5B). The threshold value of negative epistasis depends on the fitness effects of alleles a and b and on the recombination rates between the modifier and the selected loci (Barton, 1995). Similar results were found in the case of diploid populations, as long as mating is random (Charlesworth et al., 1990; Barton, 1995). However, empirical evidence for negative epistasis between selected loci remains inconclusive: epistasis is highly variable among pairs of loci, but does not seem to be negative on average (Rice, 2002; de Visser and Elena, 2007). Moreover, variable epistasis between pairs of loci is predicted to select against recombination in infinitely large populations even when epistasis is weakly negative on average (Otto and Feldman, 1997). Therefore, negative epistasis seems unlikely to explain the maintenance of high recombination rates among the majority of eukaryotic species.

Another mechanism that can generate LD among selected loci is known as selective interference or the Hill-Robertson effect (Hill and Robertson, 1966; Felsenstein, 1974), and involves finite population size. Indeed, drift in finite populations causes random fluctuations in haplotype frequencies and thus in LD. While drift may generate either positive or negative LD, situations where LD is positive tend to be rapidly dissipated by selection (because the variance in fitness is then higher), while situations where LD is negative tend to persist longer (because selection is less efficient at changing genotype frequencies), yielding negative LD on average at selection-mutation-drift equilibrium. Recombination is thus favoured because it increases the variance in fitness by bringing deleterious mutations into the same haplotypes. The effect of selective interference in favouring recombination has been studied either by simulation (Felsenstein and Yoko-

hama, 1976; Otto and Barton, 2001; Keightley and Otto, 2006) or analytically (Barton and Otto, 2005; Roze and Barton, 2006; Roze, 2021) showing that this advantage holds under various population sizes – even very large – making it a likely general explanation for the maintenance of recombination. While most models on the Hill-Robertson effect considered haploid populations, Roze (2021) showed that recombination is also generally favored in finite diploid populations, despite the fact that the LD between selected loci becomes positive when deleterious alleles are sufficiently recessive ($h < 0.25$).

Besides selective interference, other general mechanisms that can favour recombination involve temporal or spatial changes in selection. In particular, recombination can be favoured when epistasis fluctuates in sign over time, which occurs in some models of coevolution between species, in particular host-parasite interactions (Charlesworth, 1976; Barton, 1995; Peters and Lively, 1999; Gandon and Otto, 2007; Peters and Lively, 2007; Salathé et al., 2008). Indeed, coevolutionary dynamics may generate fluctuations in LD and epistasis in the host and in the parasite (Peters and Lively, 1999; Gandon and Otto, 2007). If these fluctuations are sufficiently rapid, LD and epistasis have more often opposite signs than the same sign, and recombination is favoured because it increases the mean fitness of offspring (Peters and Lively, 1999; Gandon and Otto, 2007; Salathé et al., 2009). In line with this "Red Queen hypothesis", some experimental studies on plants and on the red flour beetle tend to show that hosts exposed to parasite have higher recombination rates (Fischer and Schmid-Hempel, 2005; Kovalchuk et al., 2003; Andronic, 2012; Kerstes et al., 2012), although other studies on mice and flour beetle found no effect of parasites (Greeff and Schmid-Hempel, 2010; Dumont et al., 2015).

4 The evolution of recombination in selfing species

The majority of models on the evolution of recombination consider randomly mating populations, and the effect or the mating system has been less investigated. The capacity to self-fertilise in hermaphroditic organisms – where the same individual produces male and female gametes – is found in numerous groups of eukaryotes such as plants,

fungi, nematodes or gastropods (Jarne and Charlesworth, 1993; Jarne and Auld, 2006; Billiard et al., 2012). Selfing and its genetic consequences has been extensively studied in plants where it is present in numerous families at various degrees (Barrett, 2002). Interestingly, cytological data from several Angiosperm genera show that selfing species tend to have a higher number of chiasmata than their outcrossing relatives, suggesting that higher recombination rates have evolved in more highly selfing species (Roze and Lenormand, 2005; Ross-Ibarra, 2007). Similar results were found from genetic maps, pointing at higher recombination rates in the highly selfing *Arabidopsis thaliana* compared to its outcrossing relative *Arabidopsis lyrata* (Kuittinen et al., 2004; Hansson et al., 2006; Kawabe et al., 2006). As an extreme form of inbreeding, one of the main effects of selfing is to increase the rate of coalescence between lineages present in the same individual (1/2 for each selfing event), decreasing the effective population size and increasing homozygosity (Pollak, 1987; Nordborg, 2000b). It has been proposed that the higher recombination rates observed in selfing species could be caused by the fact that heterozygosity prevents crossovers. Indeed, experiments on yeast showed that induced heterozygosity tends to prevent recombination (Borts and Haber, 1987). However, more recent data in yeast suggest that strong divergence is needed to prevent recombination (Chen and Jinks-Robertson, 1999), while a recent experiment of induced polymorphism in inbred lines of *A. thaliana* shows that heterozygosity does not affect the recombination landscape at large genomic scales (several megabase; Lian et al. 2022). Moreover, data from tertraploid rye (Benavente and Sybenga, 2004) and from crosses between *A. thaliana* strains with different levels of divergence show that crossovers may occur preferentially in heterozygous regions at smaller genomic scale (Ziolkowski et al., 2015; Blackwell et al., 2020). Because recombination between homozygous sites does not have any genetic effect, another hypothesis is that recombination could be higher in selfing species to compensate for its low efficacy at breaking LD (Nordborg, 2000b; Wright et al., 2008). However, indirect selection for recombination should vanish at very high selfing rates (due to the very high homozygosity of individuals) so that the effect of selfing on selection for recombination may be non-monotonous. The effect of partial selfing on

selection for recombination is thus not trivial; furthermore, previous simulation results have suggested that recombination may be favoured in different parameter ranges in partially selfing and in outcrossing populations (Charlesworth et al., 1977; Holsinger and Feldman, 1983).

The mechanisms by which recombination can be favoured in partially selfing populations have been described by Roze and Lenormand (2005) using a 3-locus deterministic model (infinitely large populations). The results showed that correlations in homozygosity among loci generated by partial selfing generate a selective force on recombination that is absent under random mating. These correlations in homozygosity, also known as identity disequilibria (ID) are due to the fact that, if an individual is homozygous at a given locus, it has a higher probability to have been produced by selfing and therefore to be also homozygous at other loci (conversely, if an individual is heterozygous at a given locus, it has a higher probability to be also heterozygous at other loci). Because selfing increases the frequency of double homozygotes, ID vanishes under full outcrossing and full selfing and is maximized at intermediate selfing rates. Recombination, tends to break these correlations in homozygosity, by increasing the frequency of genotypes that are homozygous at one locus and heterozygous at the other among selfed offspring (Figure 6). In the absence of epistasis but with dominance ($h \neq 1/2$), those intermediate genotypes have a reduced fitness compared to the average of double heterozygotes and double homozygotes, the difference being $(1-2h)^2 s^2/8$ (Roze, 2009). Therefore, recombination is disfavoured because the intermediate genotypes that it tends to generate are less fit on average (Figure 6A). However, when a particular form of epistasis known as dominance-by-dominance epistasis $e_{d\times d}$ is negative (that is, when double homozygotes have a lower fitness than expected on the basis of single homozygotes), recombination is favoured because the intermediate genotypes that it tends to generate are now fitter on average (Figure 6B). Therefore, recombination can be favoured under partial selfing because it can increase the mean fitness of offspring, provided that $e_{d\times d}$ is sufficiently negative. Furthermore, this selective force is stronger than indirect forces acting on selection for recombination in randomly mating populations. Although evidence for negative

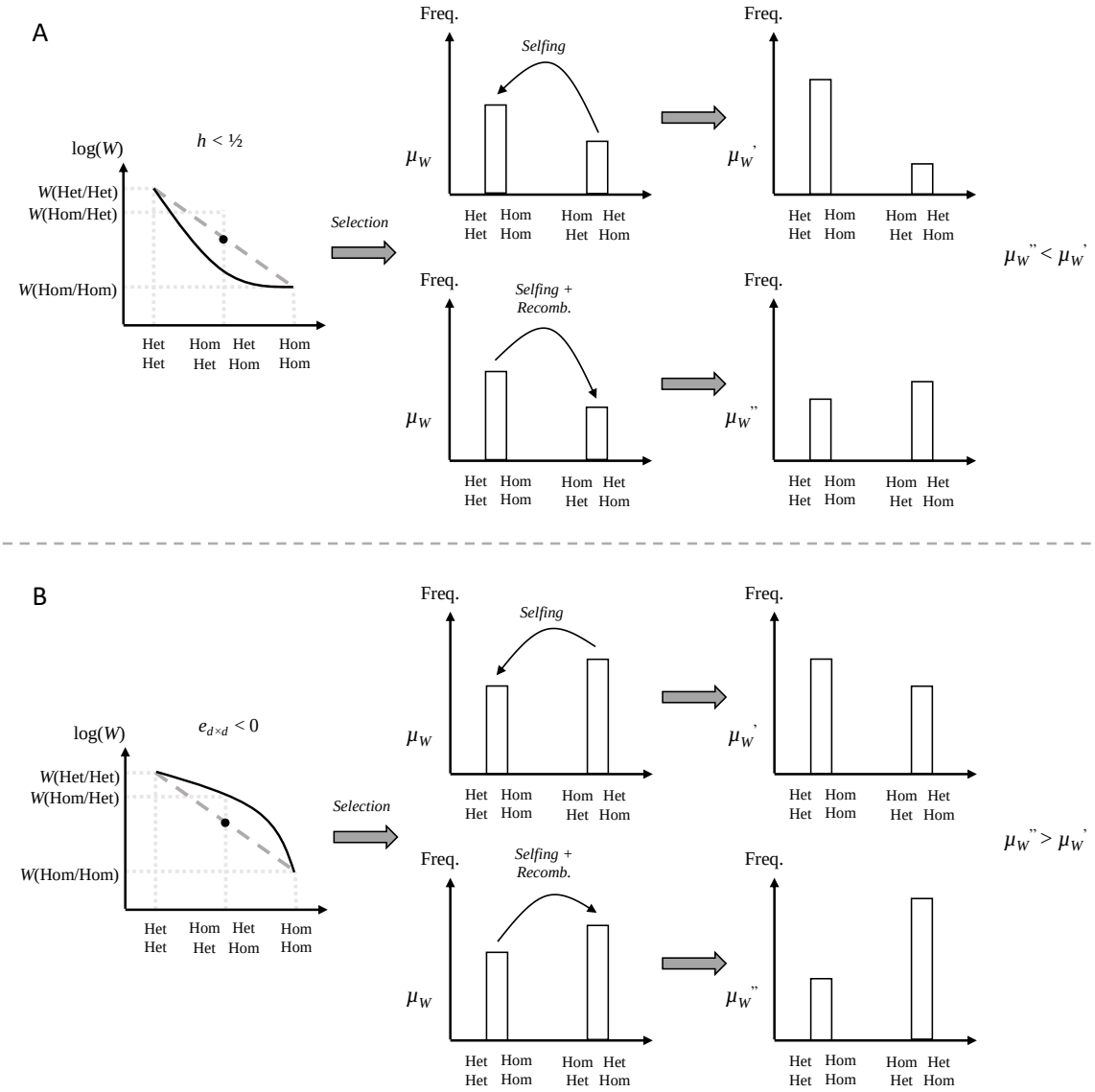


Figure 6: Effect of recombination in an infinitely large population of partially selfing diploids where genotype frequencies were build up by different fitness landscapes with directional selection (left panels). A: In the absence of epistasis, dominance causes the fitness landscape to be positively curved and recombination, by breaking correlations in homozygosity created by partial selfing, decreases the mean fitness of offspring because intermediate genotypes (homozygous at one locus and heterozygous at the other locus) are less fit on average than extreme genotypes (double homozygous and double heterozygous). Therefore, recombination is disfavoured under these assumptions. B: With sufficient negative dominance-by-dominance epistasis, the fitness landscape is negatively curved and recombination, by breaking correlations in homozygosity created by partial selfing, increases the mean fitness of offspring because intermediate genotypes are fitter on average than extreme genotypes. Therefore, recombination is favoured under these assumptions.

dominance-by-dominance epistasis between mutations produced by mutagenesis has been found in *Drosophila* (Sharp and Agrawal, 2016), data on this form of epistasis remain scarce, however. Moreover, indirect selection on recombination caused by its effect on identity disequilibria should vanish under high selfing rates, and it is thus unclear if this mechanism can explain the higher recombination rates observed in highly selfing species. The results of the deterministic model of Roze and Lenormand (2005) are based on quasi-linkage equilibrium approximations (QLE) that only hold when effective recombination rates are large, and thus break down when the selfing rate is not small, or when loci are tightly linked. Yet tightly linked loci should be the ones contributing most to indirect selection for recombination – unless selfing is strong – and it is thus not obvious to predict how recombination should evolve in species with moderate or high selfing rates based on this model. Furthermore, this model and the previous simulation models (Charlesworth et al., 1977; Holsinger and Feldman, 1983) are deterministic, considering infinitely large populations. However, as explained in the previous section, selective interference in finite populations is an important force generating selection for recombination, which may be particularly important in selfing species due to their reduced effective population size.

5 Open questions on the evolution of recombination

Important advances have been made in our understanding of the maintenance of high recombination rates in eukaryotes. On the one hand, theoretical models have characterized and quantified a number of possible evolutionary forces that may favour recombination. On the other hand, advances have also been made by empiricists in describing genetic variation for recombination at different scales, as well as the mechanisms that may constrain the location and number of COs. Despite this, the vast majority of the models on the evolution of recombination have considered very simple assumptions, often assuming a few selected loci (typically 2), a single recombination modifier locus, no cost of recombination and no obligate CO. Recent models incorporated the genetic architecture of the position of COs mediated by PRDM9 (Latrille et al., 2017; Genestier et al.,

2023). However these models were designed to understand the evolutionary dynamics of PRDM9-mediated hotspots and do not quantify the strength of selection for recombination. A recent model computed the overall strength of selection for recombination generated by selective interference between deleterious mutations segregating on a whole chromosome and found that, under realistic parameter values, indirect selection for recombination becomes extremely weak when one crossover (CO) per chromosome occurs on average (Roze, 2021). Therefore, it seems difficult to explain the range of the number of COs per chromosome observed among eukaryotes (typically between 1 and 3 or 4) from current models on the evolution of recombination due to its effect on genetic variation (Stapley et al., 2017; Fernandes et al., 2018; Brazier and Glémin, 2022). This discrepancy between models and data could be resolved by introducing more realistic features into the models, for example, CO interference, an obligate CO, a non-uniform recombination rate along the chromosome, a non-uniform density of genes along the chromosome or multiple recombination modifier loci.

In addition, the possible strength of indirect selection for recombination has been little explored from an empirical perspective, while the current wealth of genomic data offers interesting possibilities. For example, several models have shown that recombination may be favoured when deleterious (or advantageous) mutations are in negative LD (either due to selective interference or to negative epistasis). A first step to test if recombination is maintained due to indirect selection would be to check if such negative associations are indeed present within genomes. A series of recent works measured LD between putatively deleterious mutations from genomic data in human, fruit fly, the outcrossing plant *Capsella grandiflora* and the fungus *Schizophyllum commune* (Sohail et al., 2017; Garcia and Lohmueller, 2021; Sandler et al., 2021; Ragsdale, 2022; Stolyarova et al., 2022). However, they found contrasted results: LD between putatively deleterious variants was either positive or negative, while LD between putatively neutral variants was positive although LD among neutral variants should be zero on average. Moreover, these studies often use different estimates of LD and generally focus on rare mutations, which may generate a statistical bias towards finding positive LD (Good, 2022). These results

are thus difficult to interpret, and the possible sources of bias and technical artifacts that may affect such estimates should be better explored. Moreover, to our knowledge, LD between deleterious mutations has never been measured in selfing species where one expects to find possibly stronger LD due to the decrease in effective recombination rates and effective population size caused by selfing (Pollak, 1987; Nordborg, 2000b).

Finally, the effect of the mating system on the evolution of recombination should be better explored, as the only analytical model on partially selfing populations obtained results that are valid under weak selfing rates only, and ignores selective interference (Roze and Lenormand, 2005). Extending this model to low effective recombination rates and finite population size may provide new insights on the effect of the mating system on the evolution of recombination. In addition, the relationship between recombination and selfing rate would benefit from being updated with new data from genetic maps in plants (but also other organisms) as the correlation has mainly been obtained from cytological data on the number of chiasmata per bivalent, that are only indirect measures of recombination rates.

6 Organization of the thesis

The thesis is divided into three parts:

1. In a first part, recombination modifier models are used to quantify the indirect selection on the average number of CO per chromosome generated by deleterious mutations occurring along the chromosome:

   (a) In Chapter 1 analytical approximations and simulations are used to investigate the effect of arbitrary selfing rates on indirect selection for recombination, in finite populations.

   (b) Chapter 2 presents the results of simulations exploring more realistic scenarios concerning the genetic architecture of recombination rate variation, as well as the distribution of crossovers and mutations along chromosomes.

2. In a second part, genomic data from different Angiosperm species are used to estimate the linkage disequilibrium between deleterious mutations, an important component of models on the evolution of recombination in outcrossing and selfing species:

   (a) In Chapter 3 the linkage disequilibrium between deleterious mutations is estimated using genomic data from a natural populations of the outcrossing plant *Capsella grandiflora*.

   (b) Chapter 4 extends this analysis to genomic data from two highly selfing plant species, *Capsella orientalis* and *Arabidopsis thaliana*.

3. Finally, the third part draws general conclusions from all chapters as well as perspectives for future works.

Part II

Theoretical aspects: modelling the effect of the mating system and heterogeneities along chromosomes on the evolution of recombination

In this part, recombination modifier models are used to quantify the indirect selection on the average number of CO per chromosome generated by deleterious mutations occurring along the chromosome. In Chapter 1, the effect of selfing on indirect selection for recombination in finite populations is explored using analytical approximations and simulations. Chapter 2 presents simulation results on the evolution of recombination under more realistic assumptions concerning the genetic architecture of recombination rate variation, as well as the distribution of crossovers and mutations along chromosomes.

Chapter 1

The evolution of recombination in self-fertilizing organisms

Attached paper : Stetsenko R., Roze D. (2022). The evolution of recombination in self-fertilizing organisms. Genetics

OXFORD GENETICS

GENETICS, 2022, 222(1), iyac114
https://doi.org/10.1093/genetics/iyac114
Advance Access Publication Date: 5 August 2022
Investigation
Highlighted Article

The evolution of recombination in self-fertilizing organisms

Roman Stetsenko [1,2] Denis Roze [1,2,*]

[1]CNRS, IRL 3614 Evolutionary Biology and Ecology of Algae, 29688 Roscoff, France,
[2]Station Biologique de Roscoff, Sorbonne Université, 29688 Roscoff, France

*Corresponding author: Station Biologique de Roscoff, Place Georges Teissier, CS90074 29688 Roscoff, France. Email: roze@sb-roscoff.fr

Abstract

Cytological data from flowering plants suggest that the evolution of recombination rates is affected by the mating system of organisms, as higher chiasma frequencies are often observed in self-fertilizing species compared with their outcrossing relatives. Understanding the evolutionary cause of this effect is of particular interest, as it may shed light on the selective forces favoring recombination in natural populations. While previous models showed that inbreeding may have important effects on selection for recombination, existing analytical treatments are restricted to the case of loosely linked loci and weak selfing rates, and ignore the stochastic effect of genetic interference (Hill–Robertson effect), known to be an important component of selection for recombination in randomly mating populations. In this article, we derive general expressions quantifying the stochastic and deterministic components of selection acting on a mutation affecting the genetic map length of a whole chromosome along which deleterious mutations occur, valid for arbitrary selfing rates. The results show that selfing generally increases selection for recombination caused by interference among mutations as long as selection against deleterious alleles is sufficiently weak. While interference is often the main driver of selection for recombination under tight linkage or high selfing rates, deterministic effects can play a stronger role under intermediate selfing rates and high recombination, selecting against recombination in the absence of epistasis, but favoring recombination when epistasis is negative. Individual-based simulation results indicate that our analytical model often provides accurate predictions for the strength of selection on recombination under partial selfing.

Keywords: epistasis; genetic interference; meiosis; multilocus model; mating systems

Introduction

Genetic recombination lies at the heart of the sexual life cycle, and is often considered as one of the main evolutionary benefits of sexual reproduction (Otto 2021). However, considerable variation for the rate and position of meiotic crossovers along chromosomes exists within and between species (Kong *et al.* 2010; Johnston *et al.* 2016; Ritz *et al.* 2017; Stapley *et al.* 2017; Brand *et al.* 2018; Samuk *et al.* 2020), showing that recombination can evolve over short timescales. Rapid changes in recombination rates have also been observed during artificial selection experiments, in response to selection on recombination itself or on other traits (reviewed in Otto and Barton 2001), or following major genomic rearrangements such as whole genome duplications (e.g. Wright *et al.* 2015). Although substantial progress has been achieved over recent years in our understanding of the molecular mechanisms governing crossover formation (Gray and Cohen 2016; Zelkowski *et al.* 2019), the evolutionary forces acting on recombination in natural populations remain elusive. In most species, at least 1 crossover per bivalent seems required to ensure the proper segregation of chromosomes during meiosis I, while data on human trisomies suggest that homologs may fail to separate when they are entangled by too many crossovers (Koehler *et al.* 1996). These mechanistic constraints probably set lower and upper bounds to the genetic map length of chromosomes, but may leave some space for evolutionary change to occur. The evolution of recombination may also be affected by indirect selective forces, stemming from the effect of recombination on genetic variation. In particular, higher recombination rates may be favored when negative linkage disequilibria (LD) between selected loci exist within populations (i.e. when beneficial alleles at some loci tend to be associated with deleterious alleles at other loci), as recombination then increases the variance in fitness among offspring and the efficiency of natural selection (Otto and Lenormand 2002; Agrawal 2006). Different possible sources of negative LD have been identified, including epistatic interactions among loci (Charlesworth 1990; Barton 1995) and the Hill–Robertson effect, that tends to generate negative LD between selected loci from the random fluctuations of genotype frequencies occurring in finite populations (Hill and Robertson 1966; Felsenstein 1974; Otto and Barton 1997; Barton and Otto 2005; Roze and Barton 2006). Analytical and simulation models have shown that the stochastic component of selection for recombination (due to the Hill–Robertson effect) may be stronger than deterministic components generated by epistasis even when population size is rather large, especially when linkage is tight (Otto and Barton 2001; Keightley and Otto 2006; Roze 2021).

An interesting pattern observed in several genera of flowering plants is that self-fertilizing species tend to have higher chiasma frequencies than their outcrossing relatives (Roze and Lenormand 2005; Ross-Ibarra 2007). Detailed comparisons between the genetic maps of the selfing *Arabidopsis thaliana* and its

outcrossing relative *Arabidopsis lyrata* also point to higher recombination rates in *A. thaliana* (Kuittinen *et al.* 2004; Hansson *et al.* 2006; Kawabe *et al.* 2006). A possible explanation for higher recombination rates in selfers could be that polymorphism between homologs hinders recombination (Borts and Haber 1987). However, this hypothesis does not stand up to closer scrutiny, as available data suggest that substantial levels of divergence are needed to prevent meiotic recombination (Chen and Jinks-Robertson 1999), while data from tetraploid rye (Benavente and Sybenga 2004) and from crosses between *A. thaliana* strains with different levels of divergence show that crossovers may occur preferentially in heterozygous regions (Ziolkowski *et al.* 2015; Blackwell *et al.* 2020), possibly due to a positive effect of mismatches among homologs on crossover initiation (Blackwell *et al.* 2020). Because recombination between homozygous loci has no genetic effect, selfing reduces the efficiency of recombination in breaking LD (Nordborg 2000; Wright *et al.* 2008), and one may thus expect that increased rates of recombination could evolve to compensate for this effect. However, indirect selection for recombination should vanish under complete selfing (as heterozygosity should then be extremely rare), and the effect of selfing on selection for recombination may thus be nonmonotonic. Furthermore, it is not immediately obvious that models for the evolution of recombination under random mating can be directly transposed to the case of partial selfers. Indeed, simulation models have shown that recombination may be favored under different conditions in partially selfing than in outcrossing populations, though the exact mechanisms remained unclear (Charlesworth *et al.* 1977, 1979; Holsinger and Feldman 1983).

A 3-locus model of the effect of partial selfing on the evolution of recombination was analyzed by Roze and Lenormand (2005). The results showed that correlations in homozygosity across loci caused by partial selfing generate a selective force on recombination that is absent under random mating. By breaking correlations in homozygosity, recombination is favored when dominance-by-dominance epistasis is negative (meaning that double homozygotes have a lower fitness than expected based on the fitness of single homozygotes), as recombination increases the mean fitness of offspring produced by selfing. In the absence of dominance-by-dominance epistasis (or when it is positive), recombination is generally disfavored. The analysis also showed that even a small selfing rate has important consequences, with the effect of breaking correlations in homozygosity quickly becoming the dominant source of indirect selection on recombination. However, the method used by Roze and Lenormand (2005) only holds when effective recombination rates are large, and thus breaks down when the selfing rate is not small, or when loci are tightly linked—yet tightly linked loci should be the ones contributing most to indirect selection for recombination, unless selfing is strong. Furthermore, this model and the previous simulation models mentioned above are deterministic, considering infinite populations. From previous results on selection for recombination in finite, randomly mating populations (Otto and Barton 2001; Keightley and Otto 2006; Roze 2021), and from the fact that the effective size of highly selfing populations may be strongly reduced by interference effects among loci (Glémin and Ronfort 2013; Roze 2016), it seems likely that the Hill–Robertson effect should be an important component of selection for recombination in selfing organisms, but this has not been quantified.

In this article, we provide a general analysis of selection for recombination caused by interactions among deleterious alleles, in populations with arbitrary selfing rates. In a first step, we revisit Roze and Lenormand's (2005) deterministic 3-locus model, and show that linkage disequilibrium is the main source of selection for recombination in the case of tightly linked loci or under strong selfing, while the effect of correlations in homozygosity stays negligible. In a second step, we explore how the stochastic component of selection for recombination (Hill–Robertson effect) is affected by the mating system, by extending Roze's (2021) finite population model to partial selfing. Last, we extrapolate the results from the stochastic and deterministic 3-locus models in order to quantify the overall strength of selection acting on a modifier allele increasing the map length of a whole chromosome, and compare the predictions obtained with results from individual-based simulations. The results confirm that the Hill–Robertson effect is often the main component of selection for recombination when the selfing rate is high, usually generating stronger benefits of recombination as selfing increases, but not always. Deterministic effects may become important under intermediate selfing rates and when the chromosomal map length is sufficiently high, and may either increase or decrease selection for recombination depending on the sign and magnitude of the different components of epistasis.

Methods

Our baseline model is the 3-locus deterministic recombination modifier model with partial selfing considered by Roze and Lenormand (2005), that we reanalyze in order to obtain more accurate results for arbitrary selfing and recombination rates. In a second step, this model is extended to include the effect of random drift in finite populations. Finally, extrapolations are used to predict the overall strength of selection acting on a modifier affecting the genetic map length of a whole chromosome.

The 3-locus model

Genetic architecture

The model considers 2 selected loci (each with 2 alleles, A, a and B, b respectively) and a recombination modifier locus (with 2 alleles M and m). Throughout the following, a, b, and m will also be used to refer to the 3 loci, in order to keep the notation simple. All notations used are summarized in Table 1. Alleles a and b are deleterious and affect fitness by a factor $1-s$ in homozygotes and $1-sh$ in heterozygotes (s and h thus correspond to the selection and dominance coefficients of alleles a and b). Deleterious alleles are generated by mutation at a rate u per generation; back mutation is ignored. Throughout the article, we assume that h is significantly greater than zero and that $u \ll s$, so that the equilibrium frequency of deleterious alleles remains small. As in Roze and Lenormand (2005), epistasis between alleles a and b is decomposed into 3 components (see Table 2): additive-by-additive epistasis $e_{a\times a}$ represents the effect of the interaction between 2 deleterious alleles, 1 at each selected locus (either on the same or on different chromosomes), while additive-by-dominance epistasis $e_{a\times d}$ represents the effect of the interaction between 3 deleterious alleles, and dominance-by-dominance epistasis $e_{d\times d}$ the effect of the interaction between 4 deleterious alleles. The recombination modifier locus affects the baseline recombination rate r_{ab} between the 2 selected loci, so that individuals with genotypes MM, Mm, and mm at the modifier locus have recombination rates r_{ab}, $r_{ab}+\delta r_{ab}h_m$, and $r_{ab}+\delta r_{ab}$, respectively, with δr_{ab} denoting the effect of the modifier and h_m the dominance coefficient of allele m. Note that because recombination only affects the genotype of meiotic products when it occurs between heterozygous loci, any effect of the modifier on recombination rates between itself and the selected loci (r_{ma}, r_{mb}) will not

Table 1. Parameters and variables of the model.

σ	Selfing rate
s, h, $\tilde{h}$	Selection, dominance, and effective dominance coefficients of deleterious alleles
$e_{a\times a}$, $e_{a\times d}$, $e_{d\times d}$, $\tilde{e}$	Additive-by-additive, additive-by-dominance, dominance-by-dominance epistasis, and effective epistasis between deleterious alleles
β	Charlesworth *et al.*'s (1991) synergistic epistasis coefficient
ϵ	Strength of selection (scaling parameter)
u, U	Deleterious mutation rate per locus, and per chromosome
r_{ij}, $\tilde{r}_{ij}$	Recombination rate and effective recombination rate between loci i and j
δr_{ab}, $\tilde{\delta} r_{ab}$	Effect and effective effect of the modifier on r_{ab}
h_m	Dominance coefficient of allele m at the modifier locus
N, N_e	Census and effective population size
R, R_{ES}	Chromosome map length and evolutionarily stable map length
δR	Effect of the modifier on R
$D_{\mathbb{U},\mathbb{V}}$	Genetic association between the sets $\mathbb{U}$ and $\mathbb{V}$ of loci present on the maternally and paternally inherited haplotypes of an individual
$\alpha_{\mathbb{U},\mathbb{V}}$	Effect of selection on the sets $\mathbb{U}$ and $\mathbb{V}$ of loci present on the maternally and paternally inherited haplotypes of an individual
W, $\overline{W}$	Fitness of an individual and average fitness
$F = \sigma/(2-\sigma)$	Inbreeding coefficient (probability of identity by descent at 1 locus)
ϕ_{ab}	Joint probability of identity by descent at loci a and b
$G_{ab} = \phi_{ab} - F^2$	Identity disequilibrium between loci a and b
p_j, q_j	Frequencies of the lower- and uppercase allele at locus j
n	Mean number of deleterious alleles per chromosome

Table 2. Fitness matrix in the 3-locus model.

	AA	Aa	aa
BB	1	$1-hs$	$1-s$
Bb	$1-hs$	$(1-hs)^2 + e_{a\times a}$	$(1-hs)(1-s) + 2e_{a\times a} + e_{a\times d}$
bb	$1-s$	$(1-hs)(1-s) + 2e_{a\times a} + e_{a\times d}$	$(1-s)^2 + 4e_{a\times a} + 4e_{a\times d} + e_{d\times d}$

generate any indirect selection at the modifier locus (e.g. Barton 1995; Otto and Barton 1997). The results given throughout the article are valid for any ordering of the 3 loci along the chromosome (i.e. either m—a—b or a—m—b).

Genetic associations

The change in frequency at the modifier locus involves different forms of genetic associations between alleles among loci, which are defined as follows. As in Barton and Turelli (1991) and Kirkpatrick *et al.* (2002), we define indicator variables $X_{j,\varnothing}$ and $X_{\varnothing,j}$ that equal 1 if the lowercase allele j (where j may be m, a, or b) is present (or 0 if absent) on the maternally ($X_{j,\varnothing}$) or paternally ($X_{\varnothing,j}$) inherited gene of an individual (note that the terms "maternally" and "paternally inherited" are simply used to differentiate the 2 haplotypes of an individual). The average of these indicator variables over all individuals in the population gives the frequency of allele j for maternally and paternally inherited genes, $p_{j,\varnothing}$ and $p_{\varnothing,j}$, respectively. The frequency of allele j in the population, p_j, is thus given by $(p_{j,\varnothing} + p_{\varnothing,j})/2$. Centered variables $\zeta_{j,\varnothing}$ and $\zeta_{\varnothing,j}$ are defined as:

$$\zeta_{j,\varnothing} = X_{j,\varnothing} - p_{j,\varnothing}, \quad \zeta_{\varnothing,j} = X_{\varnothing,j} - p_{\varnothing,j}. \tag{1}$$

The genetic association between the sets of alleles $\mathbb{U}$ and $\mathbb{V}$ (that may be $\varnothing$, a, b, m, ab, ma, mb, mab) present on the maternally and paternally inherited haplotypes of the same individual is given by:

$$D_{\mathbb{U},\mathbb{V}} = \mathbb{E}[\zeta_{\mathbb{U},\mathbb{V}}], \tag{2}$$

with $\mathbb{E}$ the average over all individuals in the population, while

$$\zeta_{\mathbb{U},\mathbb{V}} = \left(\prod_{j\in\mathbb{U}} \zeta_{j,\varnothing}\right)\left(\prod_{k\in\mathbb{V}} \zeta_{\varnothing,k}\right). \tag{3}$$

Because our model does not include any sex-of-origin effect, we always have $D_{\mathbb{U},\mathbb{V}} = D_{\mathbb{V},\mathbb{U}}$. Associations between alleles on the same haplotype $D_{\mathbb{U},\varnothing} = D_{\varnothing,\mathbb{U}}$ will be denoted $D_{\mathbb{U}}$ for simplicity. For example, $D_{a,a}$ measures the excess of homozygotes for allele a (departure from Hardy–Weinberg equilibrium), while D_{ab} is the linkage disequilibrium between alleles a and b.

Life cycle

Each generation starts by selection between newly formed diploid individuals, followed by meiosis and syngamy. The effect of selection on allele frequencies and genetic associations can be computed using the multilocus genetics framework of Kirkpatrick *et al.* (2002). For this, the fitness or an individual relative to the mean fitness of the population is written as:

$$\frac{W}{\overline{W}} = 1 + \sum_{\mathbb{U},\mathbb{V}} \alpha_{\mathbb{U},\mathbb{V}}(\zeta_{\mathbb{U},\mathbb{V}} - D_{\mathbb{U},\mathbb{V}}), \tag{4}$$

where the "selection coefficients" $\alpha_{\mathbb{U},\mathbb{V}}$ represent the effect of selection acting on the sets of loci $\mathbb{U}$ and $\mathbb{V}$ present on the maternally and paternally inherited haplotypes of an individual. Since we do not assume any sex-of-origin effect we have $\alpha_{\mathbb{U},\mathbb{V}} = \alpha_{\mathbb{V},\mathbb{U}}$, while $\alpha_{\mathbb{U},\varnothing}$ will be denoted $\alpha_{\mathbb{U}}$ for simplicity. The combined effect of selection at loci a and b can thus be represented by 9 coefficients: α_a, α_b represent the effective strength of selection against the deleterious alleles a and b, $\alpha_{a,a}$, $\alpha_{b,b}$ the effect of dominance at the 2 loci, while α_{ab}, $\alpha_{a,b}$, $\alpha_{ab,a}$, $\alpha_{ab,b}$, and $\alpha_{ab,ab}$ represent epistatic interactions, measured as deviations from additivity. Throughout the article, we assume that selection is weak (s is of order ϵ, where ϵ is a small term) while epistasis is weaker ($e_{a\times a}$, $e_{a\times d}$, $e_{d\times d}$ of order ϵ^2). Assuming that deleterious alleles stay at low frequency,

and under the fitness matrix given by Table 2, $\alpha_{\mathbb{U},\mathbb{V}}$ coefficients are, to leading order:

$$\begin{aligned}&\alpha_a = \alpha_b \approx -sh, \quad \alpha_{a,a} = \alpha_{b,b} \approx -s(1-2h),\\ &\alpha_{ab} = \alpha_{a,b} \approx e_{\mathrm{a\times a}} + (sh)^2, \quad \alpha_{ab,a} = \alpha_{ab,b} \approx e_{\mathrm{a\times d}} + s^2h(1-2h),\\ &\alpha_{ab,ab} \approx e_{\mathrm{d\times d}} + s^2(1-2h)^2.\end{aligned} \tag{5}$$

The effect of selection on genetic associations (in terms of $\alpha_{\mathbb{U},\mathbb{V}}$ coefficients) is given by equations 9 and 15 in Kirkpatrick *et al.* (2002), while the change in frequency of the modifier is given by:

$$\Delta p_m = \sum_{\mathbb{U},\mathbb{V}} \alpha_{\mathbb{U},\mathbb{V}} D_{m\mathbb{U},\mathbb{V}}. \tag{6}$$

After selection, individuals produce gametes to form the zygotes of the next generation. During syngamy, a proportion σ of fertilizations involves 2 gametes produced by the same parent (selfing), while a proportion $1-\sigma$ involves gametes sampled at random from the whole population (outcrossing). Recombination and syngamy do not change allele frequencies, but do change genetic associations, their effect being given by equations 13 and 14 in Roze and Lenormand (2005). These equations, together with the equations describing the effect of selection on allele frequencies and genetic associations, have been implemented in a *Mathematica* notebook (available as Supplementary material) that can be used to automatically generate recursions on these variables.

Approximations

Throughout the article, we assume that the effect of the recombination modifier (δr_{ab}) is weak and compute all results to the first order in δr_{ab}. Our 3-locus diploid model can be described by 36 genotype frequencies (leading to 35 independent variables), or alternatively by 3 allele frequencies (p_m, p_a, p_b) and 32 genetic associations. A separation of timescales argument can be used to reduce this large number of variables: in particular, when selection is weak relative to recombination, allele frequencies change slowly while genetic associations are rapidly eroded by recombination. In this case, one can show that genetic associations quickly reach a quasi-equilibrium value, which can be computed by assuming that allele frequencies remain constant (Barton and Turelli 1991; Nagylaki 1993; Kirkpatrick *et al.* 2002). In this quasi-linkage equilibrium (QLE) state, associations can be expressed in terms of allele frequencies and of the parameters of the model, and these expressions can then be plugged into Equation (6) to obtain the change in frequency of the modifier. This is the approach used in Roze and Lenormand (2005) to quantify the strength of selection for recombination under partial selfing, assuming that recombination rates are sufficiently large (relative to the strength of selection) and that the selfing rate is not too large (as selfing reduces the effect of recombination). However, more accurate expressions can in principle be obtained in situations where deleterious alleles are maintained at mutation—selection balance: indeed, in this case changes in allele frequencies depend solely on the effect of the recombination modifier, and the QLE approximation thus requires only that δr_{ab} is sufficiently small relative to effective recombination rates (Roze 2014; Gervais and Roze 2017; Roze 2021). This is the approach used in the present article to obtain approximations that remain valid under low effective recombination (i.e. when recombination rates r_{ij} or when the outcrossing rate $1-\sigma$ are of order ϵ).

The Hill–Robertson effect

An expression for the strength of selection for recombination due to the Hill–Robertson effect between 2 deleterious alleles in a randomly mating, diploid population was derived in Roze (2021). Generalizing this analysis to partial selfing would be extremely tedious, as a very large number of stochastic moments of genetic associations and allele frequencies would need to be computed. However, in the case of tightly linked loci (r_{ij} of order ϵ), which should be the ones contributing most to selection for recombination unless the selfing rate is very high, separation-of-timescales arguments can be used to show that the effects of selection against deleterious alleles, recombination and drift under partial selfing can be predicted by replacing the dominance coefficient h, recombination rates r_{ij} and the population size N by the effective coefficients $h(1-F)+F$, $r_{ij}(1-F)$ and $N/(1+F)$ in the expressions obtained under random mating, where $F=\sigma/(2-\sigma)$ is the inbreeding coefficient (e.g. Nordborg 1997; Glémin and Ronfort 2013; Roze 2016). We thus introduced these effective coefficients into the expression derived in Roze (2021) in order to explore how self-fertilization affects selection for recombination generated by the Hill–Robertson effect. As we will see, comparisons with multilocus simulations indicate that this approach often yields correct results.

Multilocus extrapolation

Following Roze (2021), the 3-locus analysis can be extrapolated to predict the overall strength of selection on a modifier affecting the genetic map length R of a whole chromosome (by an amount δR). For simplicity, we assume that the modifier is located at the mid-point of a linear chromosome and that the position of each crossover is sampled in a uniform distribution along the chromosome (no interference). Deleterious mutations occur at a rate U per haploid chromosome per generation, and we assume that all mutations have the same selection and dominance coefficients s and h, the position of each new mutation being sampled in a uniform distribution along the chromosome (infinite site model). We make the simplifying assumption that epistatic effects are identical between all pairs of loci across the genome. Neglecting the effect of interactions between more than 2 mutations (for simplicity), the strength of indirect selection for recombination can be obtained by integrating the result from the 3-locus analysis over the genetic map (see Supplementary material). When the mean number of deleterious alleles per chromosome (n) is large and with epistasis, more accurate expressions for selection coefficients $\alpha_{\mathbb{U},\mathbb{V}}$ at loci j and k segregating for deleterious alleles are given by:

$$\begin{aligned}&\alpha_j \approx -sh + 2ne_{\mathrm{a\times a}}, \quad \alpha_{j,j} \approx -s(1-2h) + 2ne_{\mathrm{a\times d}},\\ &\alpha_{jk} = \alpha_{j,k} \approx e_{\mathrm{a\times a}} + \alpha_j\alpha_k, \quad \alpha_{jk,k} \approx e_{\mathrm{a\times d}} + \alpha_j\alpha_{k,k}, \quad \alpha_{jk,jk} \approx e_{\mathrm{a\times d}} + \alpha_{j,j}\alpha_{k,k}\end{aligned} \tag{7}$$

(see Supplementary material). An approximate expression for the mean number of deleterious alleles per chromosome n (taking into account the effects of epistasis and selfing) can be obtained from equation 27 in Abu Awad and Roze (2020). We also computed the overall strength of selection for recombination under the diploid model of synergistic epistasis considered by Charlesworth *et al.* (1991), in which the fitness of individuals is given by $W=\exp[-(\alpha\tilde{n}+\beta\tilde{n}^2/2)]$, with $\tilde{n}=hn_{\mathrm{he}}+n_{\mathrm{ho}}$ and where n_{he} and n_{ho} are the numbers of heterozygous and homozygous mutations present in the genome of the individual. As shown in Abu Awad and Roze (2020), this is equivalent to setting

$e_{a\times a} = -\beta h^2$, $e_{a\times d} = -\beta h(1-2h)$, and $e_{d\times d} = -\beta(1-2h)^2$ in the present model.

Individual-based simulations

Our simulation program (written in C++ and available from Zenodo) is equivalent to the program used in Roze (2021), with the addition of partial selfing and the different forms of epistasis. It represents a population of N diploids, each carrying a pair of linear chromosomes. At each generation, the number of new deleterious mutations per chromosome is drawn from a Poisson distribution with parameter U, and their position along the chromosome is drawn from a uniform distribution between 0 and 1. The fitness of each individual is computed as:

$$W = W_c(1-sh)^{n_{he}}(1-s)^{n_{ho}}(1+e_{a\times a})^{n_2}(1+e_{a\times d})^{n_3}(1+e_{d\times d})^{n_4} \quad (8)$$

where n_{he} and n_{ho} are the number of heterozygous and homozygous deleterious mutations in the genome of the individual, and where n_2, n_3, and n_4 are the number of interactions between 2, 3, and 4 deleterious alleles at 2 loci, given by:

$$\begin{aligned} n_2 &= \frac{1}{2}n_{he}(n_{he}-1) + 2n_{he}n_{ho} + 2n_{ho}(n_{ho}-1),\\ n_3 &= n_{he}n_{ho} + 2n_{ho}(n_{ho}-1),\\ n_4 &= \frac{1}{2}n_{ho}(n_{ho}-1) \end{aligned} \quad (9)$$

(e.g. Abu Awad and Roze 2020). The term W_c corresponds to a direct fitness effect of the chromosome map length R; as in Roze (2021), it is set to $W_c = \exp(-cR)$, so that c measures a direct fitness cost per crossover. This parameter reflects the fertility cost of having too many crossovers (see Introduction); it also ensures that R does not evolve toward very large values, and allows simple comparisons with analytical predictions.

To produce individuals of the next generation, parents are sampled according to their fitness, each new individual being produced by selfing with probability σ. The recombination modifier locus is located at the midpoint of the chromosome, with an infinite number of possible alleles coding for different values of R (the map length of the individual being given by the average of its 2 modifier alleles). One replicate was performed for each parameter set. During the first 20,000 generations R is fixed to 1 in order to reach mutation–selection balance for deleterious alleles. Then, for an extra 5×10^6 generations (increased up to 5×10^7 generations for high values of σ), mutations are introduced at a rate 10^{-4} at the modifier locus, each mutation multiplying the value of R by a random number drawn from a Gaussian distribution with mean 1 and variance 0.04. To allow for large effect mutations, a proportion 0.05 of mutations have an additive effect on R drawn from a uniform distribution between −1 and 1 (the new value being set to zero if it is negative). The average map length $\overline{R}$, average fitness $\overline{W}$, average number of deleterious mutations per chromosome n, and number of fixed mutations are recorded every 500 generations (fixed mutations are removed from the population in order to reduce execution time). The equilibrium value of map length is computed as the time average of $\overline{R}$, after removing the first 5×10^5 generations to allow $\overline{R}$ to equilibrate. In order to study the effect of variable selection coefficients of deleterious mutations, we modified the above baseline simulation program so that each new mutation is associated with a value of s drawn from a log-normal distribution (ensuring $s>0$): the value of $\ln(s)$ is drawn from a Gaussian distribution with variance sd^2 and mean $\ln(\overline{s}) - sd^2/2$ (so that the average selection coefficient $\overline{s}$ stays constant for different values of sd).

Results

The deterministic 3-locus model

We first reiterate the result obtained by Roze and Lenormand (2005) under high effective recombination in a more general version, and then provide a new analysis for the case of weak effective recombination.

High effective recombination

As found by Roze and Lenormand (2005), the joint effects of selfing and the recombination modifier generate an association $D_{mab,ab}$ even in the absence of selection. This stems from the fact that recombination tends to break correlations in homozygosity between loci generated by partial selfing (see also Roze 2009). The association $D_{mab,ab}$ has the sign of $-\delta r_{ab}$, reflecting the fact that the modifier allele increasing recombination tends to be found on genetic backgrounds in which the correlation in homozygosity between loci a and b is relatively weaker. At QLE and to the first order in δr_{ab}, it is given by:

$$D^0_{mab,ab} \approx -\frac{2\delta r_{ab}(1-2r_{ab})F(1-F)[H_m(1+2F\gamma_{ab})+F(\gamma_{ma}+\gamma_{mb}-\gamma_{ab})]}{[1+F(\gamma_{ma}+\gamma_{mb}+\gamma_{ab})](1+2F\gamma_{ab})^2} pq_{mab}, \quad (10)$$

where the superscript "0" indicates that the effect of selection at loci a and b is neglected in this expression, and where $H_m = h_m + p_m(1-2h_m)$, $\gamma_{ij} = r_{ij}(1-r_{ij})$ and $pq_{mab} = p_m q_m p_a q_a p_b q_b$ (with $q_j = 1-p_j$). Equation (10) generalizes equation 32 in Roze and Lenormand (2005) to arbitrary h_m and any ordering of the 3 loci along the chromosome. It shows that $D^0_{mab,ab}$ vanishes under random mating ($F=0$) and full selfing ($F=1$), because in these cases the mating system does not generate any correlation in homozygosity among loci.

The other associations in Equation (6) are generated by selection against the deleterious alleles (of order ϵ) and by the modifier effect. At QLE, associations $D_{mi,i}$, $D_{mij,i}$, and $D_{mi,ij}$ (where i, j are either a or b) are of order ϵ, while associations D_{mi}, $D_{m,i}$, D_{mij}, $D_{m,ij}$, and $D_{mi,j}$ are of order ϵ^2 (expressions for these associations are given in Appendix A). As a result, one obtains from Equations (5) and (6) that the change in frequency of the modifier is, to leading order:

$$\Delta p_m \approx \alpha_{a,a}D_{ma,a} + \alpha_{b,b}D_{mb,b} + \alpha_{ab,ab}D^0_{mab,ab}, \quad (11)$$

the association $D_{ma,a}$ being given by:

$$D_{ma,a} \approx \alpha_{b,b}F D^0_{mab,ab}, \quad (12)$$

(and symmetrically for $D_{mb,b}$). $D_{ma,a}$ is positive when $\delta r_{ab} > 0$ and $h < 1/2$ (since $\alpha_{b,b}$ and $D^0_{mab,ab}$ are both negative in this case), reflecting the fact that the modifier allele m increasing recombination tends to be found on more homozygous backgrounds at locus a. Indeed when $h < 1/2$, homozygous genotypes at loci a and b have, on average, a lower fitness than heterozygous genotypes, and the correlation in homozygosity generated by partial selfing increases the efficiency of selection, lowering the frequency of homozygous genotypes. As the modifier allele increasing recombination tends to break this correlation, it is associated to a relative excess in homozygosity at each selected locus. This effect disfavors recombination (since homozygotes have a lower fitness than heterozygotes on average), which is reflected by the first 2 terms of Equation (11). Equations (11) and (12) give:

$$\Delta p_m \approx (2\alpha_{a,a}\alpha_{b,b}F + \alpha_{ab,ab})D^0_{mab,ab}, \quad (13)$$

which, using Equation (5), becomes:

$$\Delta p_m \approx [s^2(1-2h)^2(1+2F) + e_{d\times d}]D^0_{mab,ab}, \quad (14)$$

yielding Equation 36 in Roze and Lenormand (2005). Equations (10) and (14) show that increased recombination is favored when dominance-by-dominance epistasis ($e_{d\times d}$) is sufficiently negative. Indeed, in this case, breaking correlations in homozygosity (thus increasing the frequency of genotypes that are homozygous at one selected locus and heterozygous at the other) tends to increase the mean fitness of offspring.

Low effective recombination

In the previous analysis, terms in $r_{mi}(1-F)$ appear in the denominators of the expressions for D_{mi} and $D_{m,i}$ at QLE (see Appendix A), causing these expressions to diverge (i.e. tend to infinity) as effective recombination rates $r_{mi}(1-F)$ tends to zero. Similarly, terms in $r_{mab}(1-F)$ appear in the denominators of D_{mab}, $D_{m,ab}$, $D_{ma,b}$, and $D_{mb,a}$, where $r_{mab} = (r_{ma} + r_{mb} + r_{ab})/2$ is the probability that at least 1 recombination event occurs between the 3 loci. This indicates that these associations should play a more important role in the case of tightly linked loci. In order to explore this regime, we reanalyzed the model in the case where all recombination rates are of order ϵ (see Supplementary material). This analysis shows that the effect of the linkage disequilibrium D_{ab} between deleterious alleles (which was negligible under high effective recombination) becomes predominant when loci are tightly linked. A general expression for D_{ab} at equilibrium in terms of $\alpha_{U,V}$ coefficients is given by:

$$D_{ab} \approx \frac{\tilde{\alpha}_{ab} - \tilde{\alpha}_a\tilde{\alpha}_b + 2FG_{ab}(\alpha_a + \alpha_{a,a})(\alpha_b + \alpha_{b,b})}{\tilde{r}_{ab} - \tilde{\alpha}_a - \tilde{\alpha}_b} pq_{ab}, \quad (15)$$

where the tilde denotes "effective coefficients": $\tilde{r}_{ab} = r_{ab}(1-F)$, $\tilde{\alpha}_a = \alpha_a(1+F) + \alpha_{a,a}F$ (and similarly for $\tilde{\alpha}_b$), while $\tilde{\alpha}_{ab} = \alpha_{ab}(1+\phi_{ab}) + 2\alpha_{a,b}F + (\alpha_{ab,a} + \alpha_{ab,b})(F + \phi_{ab}) + \alpha_{ab,ab}\phi_{ab}$, where ϕ_{ab} is the probability of joint identity-by-descent at the 2 loci (which is approximately F when r_{ab} is small, see Appendix A). Finally, G_{ab} in Equation (15) refers to the identity disequilibrium between the 2 loci, defined as $G_{ab} = \phi_{ab} - F^2$ (Weir and Cockerham 1973), and thus approximately equal to $F(1-F)$ under tight linkage. Equation (15) simplifies to $(\alpha_{ab} - \alpha_a\alpha_b)pq_{ab}/(r_{ab} - \alpha_a - \alpha_b)$ in the absence of selfing; this is equivalent to the result obtained by Barton (1995) under strong recombination (equation 9b in Barton 1995), except that α_a and α_b now appear in the denominator, due to our assumption that r_{ab} is small (of order ϵ). With selfing, 2 important differences appear: (1) the recombination rate and selection coefficients are replaced by effective coefficients $\tilde{r}_{ab}$, $\tilde{\alpha}_a$, $\tilde{\alpha}_b$, $\tilde{\alpha}_{ab}$ (since increased homozygosity affects both the effects of recombination and of selection, as explained in Appendix B), and (2) an extra term, involving the identity disequilibrium G_{ab}, appears in the numerator and tends to generate positive linkage disequilibrium between deleterious alleles.

Using the expressions for $\alpha_{U,V}$ coefficients given by Equation (5), Equation (15) simplifies to:

$$D_{ab} \approx \frac{s^2(1-h)^2(1+2F)G_{ab} + \tilde{e}}{\tilde{r}_{ab} + 2s\tilde{h}} pq_{ab}, \quad (16)$$

with $\tilde{e} = e_{a\times a}(1 + 2F + \phi_{ab}) + 2e_{a\times d}(F + \phi_{ab}) + e_{d\times d}\phi_{ab}$ and $\tilde{h} = h(1-F) + F$. The fact that partial selfing generates positive linkage disequilibrium between deleterious alleles in the absence of epistasis has been noticed in previous analytical and simulation studies (Roze and Lenormand 2005; Kamran-Disfani and Agrawal 2014), and may be understood as follows. Lineages with different histories of inbreeding coexist in a partially selfing population: lineages that have been inbred for many generations tend to be very homozygous (at all loci), while lineages that have been inbred for fewer generations tend to be less homozygous. This is the basis of correlations in homozygosity among loci, represented by the identity disequilibrium G_{ab}. Because homozygosity increases the efficiency of selection, the frequency of deleterious alleles tends to be lower within lineages that have been inbred for longer (purging), and higher in less inbred lineages, resulting in positive linkage disequilibrium between deleterious alleles. As shown by Equation (16), D_{ab} may become negative when epistasis is negative and sufficiently strong, the relative importance of $e_{a\times d}$ and $e_{d\times d}$ increasing as the selfing rate increases.

The fact that the allele associated with higher recombination tends to erode D_{ab} more rapidly in turn generates genetic associations between the modifier and the selected loci. In particular, we have:

$$D_{mab} \approx -\frac{\delta\tilde{r}_{ab}H_mD_{ab}}{\tilde{r}_{mab} - \tilde{\alpha}_a - \tilde{\alpha}_b} pq_m, \quad D_{ma} \approx \frac{\tilde{\alpha}_b D_{mab}}{\tilde{r}_{ma} - \tilde{\alpha}_a} \quad (17)$$

with $\delta\tilde{r}_{ab} = \delta r_{ab}(1-F)$, $\tilde{r}_{mab} = r_{mab}(1-F)$, and again $\tilde{\alpha}_a = \tilde{\alpha}_b \approx -s\tilde{h}$, $H_m = h_m + p_m(1-2h_m)$ (D_{mb} is given by a symmetric expression). These equations take the same form as under random mating (e.g. equations 9c and 11 in Barton 1995) and are interpreted in the same way: when $D_{ab} > 0$, AB and ab haplotypes are in excess in the population, and the allele increasing recombination (allele m if $\delta r_{ab} > 0$) tends to reduce this excess, and thus becomes more associated with Ab, aB haplotypes (which is reflected by a negative value of D_{mab}). By contrast, when $D_{ab} < 0$, the allele increasing recombination becomes more associated with ab, AB haplotypes (which is reflected by a positive value of D_{mab}). When $D_{mab} < 0$, selection is less efficient in the m background (because the frequency of extreme haplotypes AB and ab is lower in this background), leading to positive values of D_{ma}, D_{mb} (deleterious alleles are more frequent in the m background). When $D_{mab} > 0$, selection is more efficient in the m background (because the frequency of extreme haplotypes AB and ab is higher in this background), and m thus becomes better purged from deleterious alleles (which is reflected by negative values of D_{ma}, D_{mb}).

Under weak recombination, separation of timescale arguments can be used to express the other associations that appear in Equation (6) in terms of D_{mab}, $D_{ma,}$ and D_{mb} (e.g. Roze 2016); one obtains:

$$\begin{aligned} D_{mab,ab} \approx D_{mab,a} \approx D_{ma,ab} \approx D_{m,ab} \approx D_{ma,b} \approx F D_{mab}, \\ D_{m,a} \approx D_{ma,a} \approx F D_{ma}, \quad D_{m,b} \approx D_{mb,b} \approx F D_{mb}. \end{aligned} \quad (18)$$

Plugging these into Equation (6) yields:

$$\Delta p_m \approx \tilde{\alpha}_a D_{ma} + \tilde{\alpha}_b D_{mb} + \tilde{\alpha}_{ab} D_{mab}, \quad (19)$$

which again takes the same form as under random mating (equation 8 in Barton 1995). From Equations (17) and (19), one obtains:

$$\Delta p_m \approx -\frac{\delta\tilde{r}_{ab}H_mD_{ab}pq_m}{\tilde{r}_{mab} - \tilde{\alpha}_a - \tilde{\alpha}_b}\left[\tilde{\alpha}_{ab} + \tilde{\alpha}_a\tilde{\alpha}_b\left(\frac{1}{\tilde{r}_{ma} - \tilde{\alpha}_a} + \frac{1}{\tilde{r}_{mb} - \tilde{\alpha}_b}\right)\right], \quad (20)$$

equivalent to equation 12 in Barton (1995). Given that the equilibrium frequency of deleterious alleles is approximately $u/(s\tilde{h})$, thus of order u/ϵ, Equations (16) and (20) show that under weak

effective recombination (i.e. when $\tilde{r}$ coefficients are of order ϵ), selection for recombination generated by the linkage disequilibrium D_{ab} is of order $\delta r_{ab}u^2/\epsilon$, thus stronger than selection for recombination generated by the term in $D^0_{mab,ab}$ seen in the previous subsection (of order $\delta r_{ab}u^2$, see Equation (14)). When $\tilde{r}$ coefficients are not small (of order 1), however, indirect selection generated by D_{ab} becomes of order $\delta r_{ab}u^2\epsilon^2$ (and thus negligible). A general expression for the change in frequency of the modifier, valid under both weak and strong effective recombination, can thus be obtained by summing Equations (14) and (20). We noticed that in some cases, more accurate expressions can be obtained by taking into account indirect effects of $D^0_{mab,ab}$ on other genetic associations between the modifier and selected loci (these terms should be negligible under both weak and strong effective recombination, but improve the precision of our approximations for intermediate values of recombination rates); these expressions are given in Appendix C. Two further improvements to our approximations are made in Appendix C, leading to more accurate expressions when epistasis is of the same order of magnitude as the strength of selection ($e_{a\times a}$, $e_{a\times d}$, $e_{d\times d}$ of order ϵ) and when the selfing rate is high (see Supplementary material for derivations).

Figure 1 and Supplementary Fig. 1 show that our approximations provide correct predictions for the change in frequency at the modifier locus, for all values of the selfing rate between 0 and 1. The dots in these figures correspond to the results of deterministic simulations, obtained by iterating exact recurrence equations for the 36 genotype frequencies: allele M is fixed during the first 3,000 generations to reach mutation–selection balance at the selected loci, then allele m is introduced in frequency 0.01 and the population is allowed to evolve for an extra 1,000 generations, $\Delta p_m/pq_{mab}$ being averaged over the last 500 generations (see Supplementary material). As can be seen in Fig. 1, indirect selection is mostly driven by D_{ab} when recombination rates are small (left figures) or when the selfing rate approaches 1. In the absence of epistasis, the identity disequilibrium generates positive D_{ab} which disfavors recombination (as breaking positive D_{ab} reduces the variance in fitness, see Fig. 1, a and b). Negative epistasis generates negative D_{ab}, which favors increased recombination when effective recombination rates are sufficiently small (Equation (20) and Fig. 1, c–f). The relative effect of the term in $D^0_{mab,ab}$ becomes more important in the case of loosely linked loci and for intermediate selfing rates, either when epistasis is absent (in which case it disfavors recombination, see Equation (14) and Fig. 1b), or when the dominance-by-dominance component of epistasis is important (in which case it favors increased recombination when $e_{d\times d} < 0$, Fig. 1f). Finally, selection on recombination vanishes under complete selfing ($\sigma = 1$), as recombination is ineffective in that case.

The Hill–Robertson effect

An expression for the strength of selection for recombination caused by the Hill–Robertson effect between 2 deleterious alleles in a diploid, randomly mating population was derived by Roze (2021). This expression consists of a sum of terms corresponding to different mechanisms generating selection for recombination, that all take the form $\delta r_{ab}(sh)^2u^2/[N_e\prod_{i=1}^{5}(r_{\mathbb{U}_i}+x_i sh)]$ where N_e is the effective population size, $\mathbb{U}_i$ is either mab, ma, mb, or ab, and x_i equals 1, 2, 3, or 4. In the case of loosely linked loci ($r_{\mathbb{U}_i} \gg sh$), the terms in sh in the denominator may be neglected and the strength of selection for recombination increases as sh increases—being roughly proportional to $(sh)^2$. In the case of tightly linked loci ($r_{\mathbb{U}_i} \ll sh$), however, recombination rates $r_{\mathbb{U}_i}$ may be neglected in the denominator and selection for recombination now decreases as sh increases—being roughly proportional to $1/(sh)^3$.

In order to extend these results to partially selfing populations, we replaced δr_{ab}, $r_{\mathbb{U}_i}$ and h by the effective parameters $\delta\tilde{r}_{ab} = \delta r_{ab}(1-F)$, $\tilde{r}_{\mathbb{U}_i} = r_{\mathbb{U}_i}(1-F)$, and $\tilde{h} = h(1-F)+F$ in the expression derived in Roze (2021). The effect of increasing selfing (thus increasing F) on the strength of selection for recombination due to the Hill–Robertson effect can be understood using the same reasoning as above. When effective recombination rates are large relative to the strength of selection against deleterious alleles (selfing rate not too large, $r_{\mathbb{U}_i}$ not too small so that $\tilde{r}_{\mathbb{U}_i} \gg s\tilde{h}$), selection for recombination becomes approximately proportional to $\delta\tilde{r}_{ab}(s\tilde{h})^2/[N_e\prod_{i=1}^{5}\tilde{r}_{\mathbb{U}_i}] = \delta r_{ab}(s\tilde{h})^2/[N_e(1-F)^4\prod_{i=1}^{5}r_{\mathbb{U}_i}]$, which increases as F increases (mostly due to the decreased effective recombination rates $\tilde{r}_{\mathbb{U}_i}$ at the denominator, but also to the increase in $\tilde{h}$). When effective recombination rates are small relative to the strength of selection (high selfing and/or tightly linked loci, so that $\tilde{r}_{\mathbb{U}_i} \ll s\tilde{h}$), selection for recombination is approximately proportional to $\delta\tilde{r}_{ab}/[N_e(s\tilde{h})^3] = \delta r_{ab}(1-F)/[N_e(s\tilde{h})^3]$, which decreases as F increases (mostly due to the decreased effect of the modifier $\delta\tilde{r}_{ab}$, but also to the increase in $\tilde{h}$). One thus predicts that increasing selfing should generally have a nonmonotonic effect on selection for recombination due to the Hill–Robertson effect (selection for recombination first increasing, and then decreasing toward zero as F increases from 0 to 1), the position of the maximum depending on sh and on the values of recombination rates (however, selfing may always decrease selection for recombination when sh is sufficiently strong and/or linkage sufficiently tight). These verbal predictions are illustrated by Fig. 2. Note that selfing also has the additional effect of reducing N_e by a factor $1/(1+F)$ (Pollak 1987; Nordborg 2000), thus increasing the strength of the Hill–Robertson effect, but this effect stays minor relative to the effect of selfing on effective recombination rates. Selfing may cause stronger reductions in N_e when deleterious alleles are segregating at many linked loci, however, through background selection effects (Glémin and Ronfort 2013; Roze 2016).

These results also provide us with some understanding of the relative importance of stochastic and deterministic sources of selection for recombination: when effective recombination rates are large, indirect selection due to the Hill–Robertson effect is of order $\delta r_{ab}u^2\epsilon^2/N_e$, and should thus be negligible relative to the deterministic component (of order $\delta r_{ab}u^2$). When effective recombination rates are small (of order ϵ), however, selection due to the Hill–Robertson effect is now of order $\delta r_{ab}u^2/(N_e\epsilon^3)$, and may thus become of the same order of magnitude or stronger than deterministic terms (which are then of order $\delta r_{ab}u^2/\epsilon$, as shown in the previous subsection).

Multilocus extrapolation

The 3-locus analysis can be extended to compute the overall strength of indirect selection acting on a modifier affecting the genetic map length of a whole chromosome. Neglecting the effect of interactions between deleterious alleles at more than 2 loci, this can be done by integrating the result from the 3-locus model over all possible positions of alleles a and b. As in Roze (2021), we consider a linear chromosome with map length R (in Morgans), along which deleterious mutations occur at a rate U per chromosome per generation. The modifier is located at the mid-point of the chromosome, allele m increasing map length by an amount $\delta R/2$ when heterozygous and δR when homozygous (see *Methods*). The strength of selection for allele m, defined as $s_m = \Delta p_m/\left(\frac{1}{2}p_m q_m\right)$, can be decomposed into 3 terms:

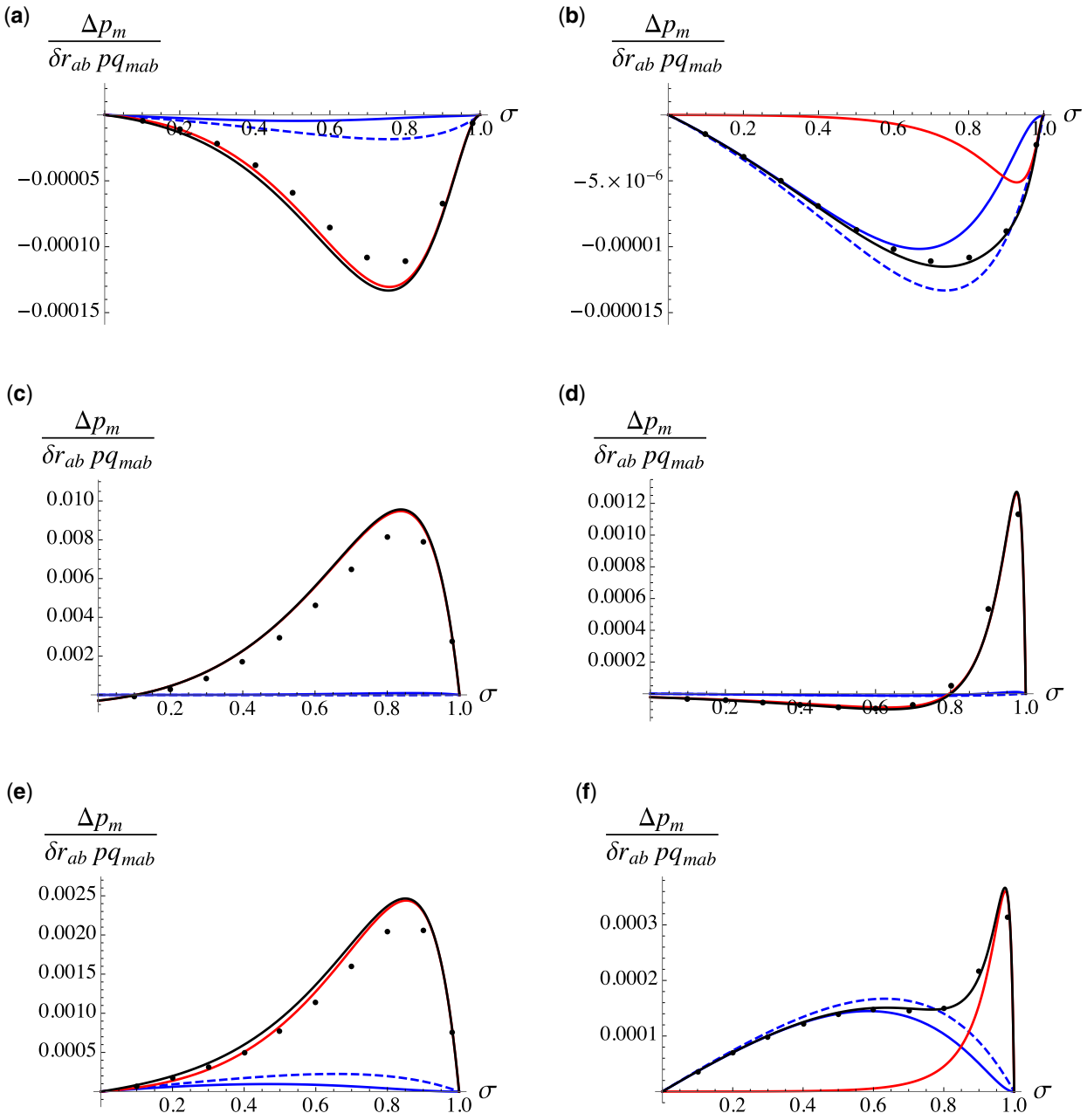


Fig. 1. Effect of the selfing rate σ on the change in frequency of allele m in the deterministic 3-locus model (scaled by $\delta r_{ab} pq_{mab}$) with no epistasis (a and b), negative additive-by-additive epistasis ($e_{a\times a} = -0.001$; c and d) and negative dominance-by-dominance epistasis ($e_{d\times d} = -0.001$; e and f), and for different recombination rates (a, c, e: $r_{ma} = r_{ab} = 0.01$; b, d, f: $r_{ma} = r_{ab} = 0.1$). Dots correspond to deterministic simulation results (see text for details) and black curves to the result obtained from Equations (6) and (C1)–(C11), which is the sum of a term generated by $D_{ab} - D_{a,b}$ (red curves) and a term generated by $D^0_{mab,ab}$ (blue curves); dashed blue curves correspond to the strong recombination approximation (Equation (11)). Note that the red and black curves are nearly indistinguishable in (c) and (d). The effect of additive-by-dominance epistasis ($e_{a\times d}$) is similar to the effect of additive-by-additive epistasis ($e_{a\times a}$) for these parameter values, and is shown in Supplementary Fig. 1. Loci are in the order m—a—b, parameters values are: $s = 0.01$, $h = 0.2$, $h_m = 0.5$, while $\delta r_{ab} = 0.01$ and $u = 10^{-5}$ in the simulations.

$$s_m = s_{\text{direct}} + s_{\text{det}} + s_{\text{HR}} \tag{21}$$

where s_{direct} corresponds to the effect of direct selective pressures acting on R (due to any direct effect of R on the fitness of individuals), s_{det} to indirect selection caused by deterministic interactions between deleterious alleles, and s_{HR} to indirect selection caused by the Hill–Robertson effect. Assuming a direct fitness cost of crossovers so that fitness decreases as e^{-cR} as R increases, the direct selection term is given by $s_{\text{direct}} = -c\delta R(1 + F)$ to the first order in δR (e.g. Gervais and Roze 2017). From the analysis above, the term s_{det} can be further decomposed into $s_{D_{ab}} + s_{D^0_{mab,ab}}$, where $s_{D_{ab}}$ corresponds to the overall effect of deterministically generated LD (D_{ab}) between deleterious alleles (either by epistasis or by identity disequilibria), given by Equations (15) and (20), while $s_{D^0_{mab,ab}}$ corresponds to the overall effect of associations of the form $D_{mab,ab}$ generated by correlations in homozygosity and by the modifier effect, given by Equations (10) and (13). Because $s_{D_{ab}}$ should be mostly driven by tightly linked loci, recombination rates are approximated by genetic distances in Equations (15) and (20) (before integrating over the genetic map), while δr_{ab} is approximated by $\delta R r_{ab}/R$ (Roze 2021). By contrast, loosely linked loci can make a stronger contribution to $s_{D^0_{mab,ab}}$, and the

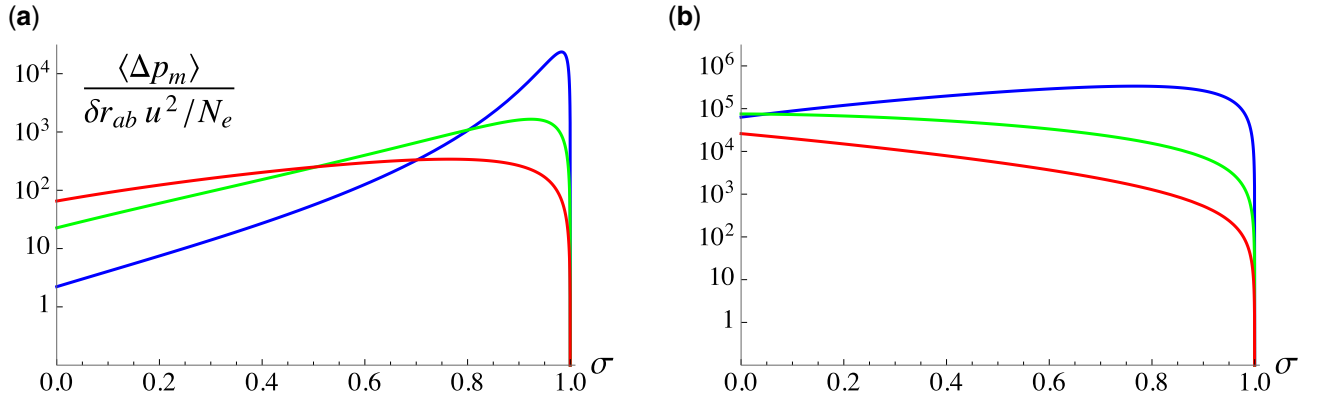


Fig. 2. Effect of the selfing rate σ on the expected change in frequency of allele m (scaled by $\delta r_{ab} u^2/N_e$ and on a log scale) generated by the Hill–Robertson effect between alleles a and b, for different values of the strength of selection against deleterious alleles (blue: $s = 0.005$; green: $s = 0.02$; red: $s = 0.05$) and recombination rates (a: $r_{ma} = r_{ab} = 0.05$; b: $r_{ma} = r_{ab} = 0.005$). The value of σ maximizing selection for recombination decreases as s increases and as recombination decreases. When selection is sufficiently strong and linkage sufficiently tight, selection for recombination decreases monotonously as σ increases (as can in seen in (b) for $s = 0.02$, $s = 0.05$). Loci are in the order m—a—b, the dominance coefficient of deleterious alleles is set to $h = 0.2$.

recombination rate between loci i and j in Equation (10) is thus expressed in terms of the genetic distance d_{ij} between these loci using Haldane's mapping function $r_{ij} = [1 - e^{-2d_{ij}}]/2$ (Haldane 1919), while δr_{ab} is approximated by $\delta R e^{-2\tilde{d}_{ab}} \tilde{d}_{ab}/R$ (see Supplementary material). Note that when U is not small (so that the mean number of deleterious alleles per chromosome n may be large), and with epistasis, the expressions for $\alpha_{\mathbb{U},\mathbb{V}}$ coefficients given by Equation (7) must be used instead of Equation (5).

Concerning the integration of the term generated by the Hill–Robertson effect, Roze (2021) showed that using scaled recombination rates $\rho_{\mathbb{U}} = r_{\mathbb{U}}/(sh)$ eliminates sh from the integrand, sh only appearing in the integration limits, given by $R/(2sh)$. The same method can be used when $r_{\mathbb{U}}$, δr_{ab} and h are changed to effective parameters in order to incorporate the effect of selfing. In particular, defining scaled recombination rates $\tilde{\rho}_{\mathbb{U}} = \tilde{r}_{\mathbb{U}}/(s\tilde{h})$, one finds that the overall strength of indirect selection generated by the Hill–Robertson effect is given by the same expression as under random mating (equation 3 in Roze 2021), except that the integration limits become $R(1-F)/(2s\tilde{h})$ and that the whole expression is multiplied by a factor $1/(1-F)^2$. When $R(1-F) \gg s\tilde{h}$ (which implies that the selfing rate is not too large and s is sufficiently small), the integral can be approximated by the same integral taken between 0 and infinity, which is ≈1.8 (Roze 2021), yielding:

$$s_{HR} \approx \frac{1.8\delta R U^2}{N_e R^3 (1-F)^2}. \quad (22)$$

Equation (22) is equivalent to equation 1 in Roze (2021), replacing R by $\tilde{R} = R(1-F)$ and δR by $\delta\tilde{R} = \delta R(1-F)$. Note that this approximation is not valid when F approaches 1, however, as $R(1-F) \gg s\tilde{h}$ cannot hold in this case. Finally, the effective population size N_e may be significantly lowered by background selection effects when U is not small. Assuming that the reduction in N_e is mostly due to tightly linked loci, one obtains that $N_e \approx [N/(1+F)] \exp[-2U/(\tilde{R} + 2s\tilde{h})]$ (e.g. Hudson and Kaplan 1995; Roze 2016), where N is the census population size.

The evolutionarily stable map length (R_{ES}) corresponds to the value of R for which direct and indirect selection balance each other, that is, $s_m = 0$. Figure 3 shows R_{ES} as a function of the selfing rate σ, for different values of the direct cost of recombination c. The curves correspond to the analytical predictions, obtained

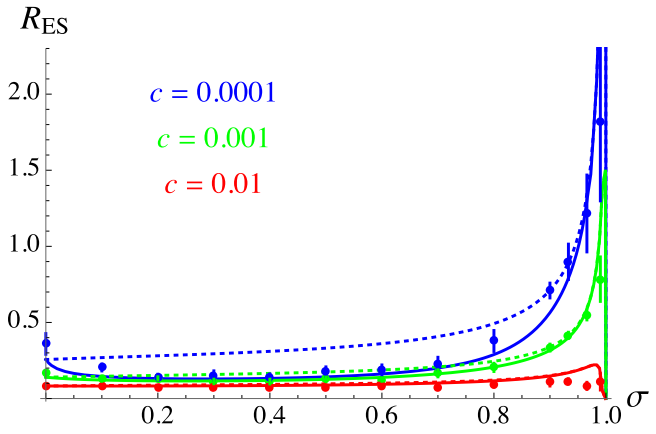


Fig. 3. Effect of the selfing rate σ on the evolutionarily stable map length R_{ES}, for different values of the cost of recombination c. Solid curves correspond to the analytical predictions obtained by solving $s_{direct} + s_{det} + s_{HR} = 0$ for R, where s_{det} and s_{HR} are obtained by integrating the expressions for the strength of indirect selection generated by deterministic effects (for s_{det}) and by the Hill–Robertson effect (for s_{HR}) over the genetic map (see Supplementary material). Dotted curves correspond to the predictions obtained when ignoring indirect selection caused by deterministic effects (i.e. solving $s_{direct} + s_{HR} = 0$ for R). Dots correspond to individual-based simulation results; in this and the following figures, error bars are obtained by dividing the simulation output (after removing the first 5×10^5 generations) into 10 batches and calculating the variance of the average map length per batch, error bars measuring ±1.96 SE. Parameter values are $N = 20,000$, $U = 0.1$, $s = 0.02$, $h = 0.2$, $e_{a\times a} = e_{a\times d} = e_{d\times d} = 0$.

using *Mathematica* by numerically integrating the results of the 3-locus model over the genetic map to obtain s_{det} and s_{HR} for a range of values of R, and finding the value of R for which $s_m = 0$ by interpolation (see Supplementary material). Note that the more precise approximations given in Appendix C have been used to compute s_{det}, but using Equations (10), (13), (15), and (20) often yields similar results. Figure 3 shows that the ES map length generally increases as the selfing rate increases, due to stronger indirect selection caused by the Hill–Robertson effect. While indirect selection vanishes under full selfing ($\sigma = 1$), the analytical model predicts that the maximum map length is reached for values of σ very close to 1, in particular when $c = 10^{-3}$ and $c = 10^{-4}$. In the absence of epistasis, the deterministic component of indirect selection selects against recombination ($s_{det} < 0$).

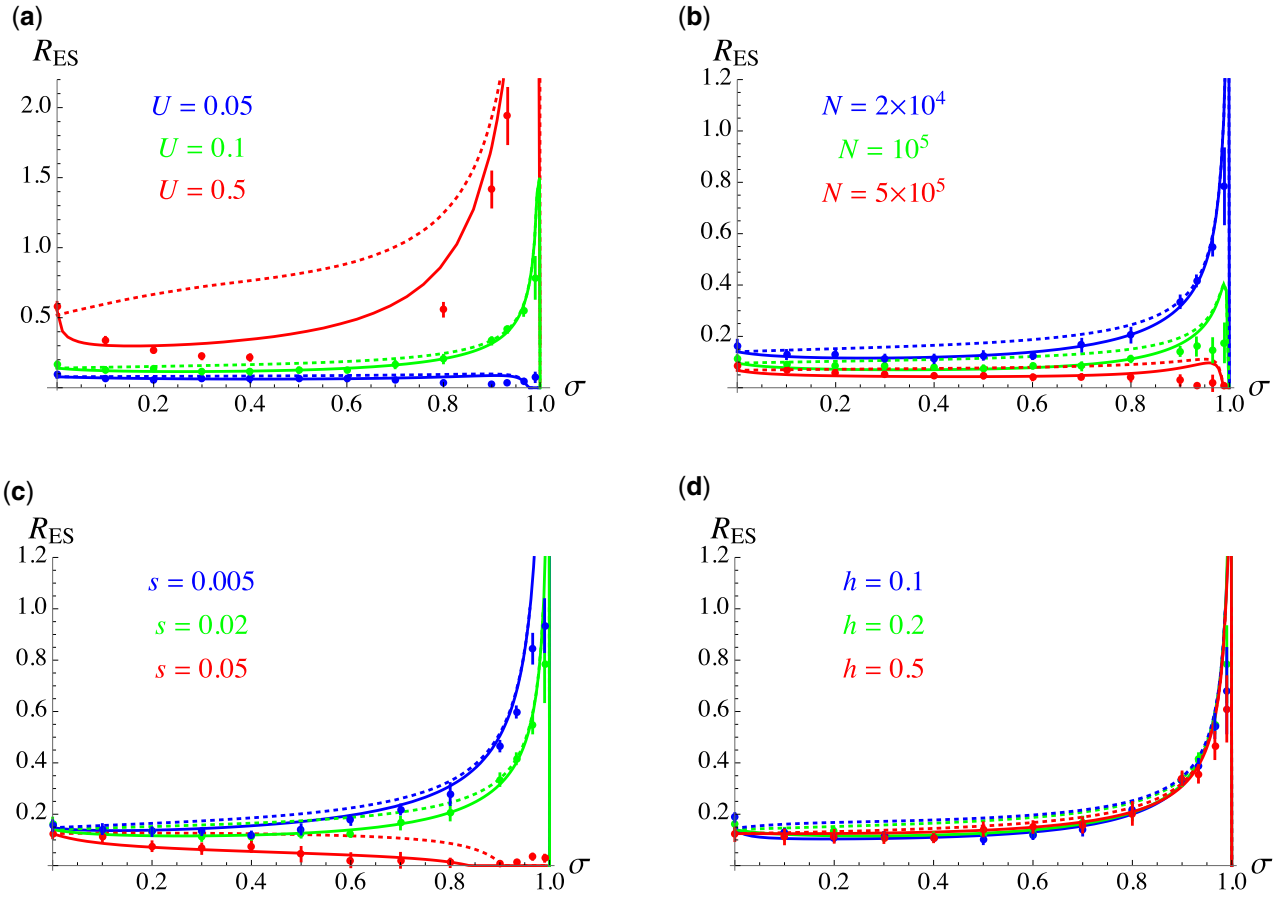


Fig. 4. Effect of the selfing rate σ on the evolutionarily stable map length R_{ES}, for different values of the mutation rate U, population size N, strength of selection, and dominance coefficient of deleterious alleles (s, h). Dots correspond to simulation results, curves have the same meaning as in Fig. 3, and default parameter values are as in Fig. 3 with $c = 0.001$ (no epistasis).

Figure 3 shows that this deterministic component stays negligible relative to the Hill–Robertson effect for parameter values leading to low R_{ES} ($c = 0.01$), while its relative effect becomes more important for parameter values leading to higher R_{ES} ($c = 10^{-4}$). This agrees with the prediction that the Hill–Robertson effect becomes stronger than deterministic effects when effective recombination rates are sufficiently small. In the absence of epistasis, we generally found that s_{det} is mainly driven by $s_{D_{ab}}$, the effect of $s_{D^0_{mab,ab}}$ staying negligible (not shown). Figure 3 also shows that extrapolations from our 3-locus model often provide accurate predictions of the evolutionarily stable map length observed in the simulations, with discrepancies appearing for high values of the selfing rate. These discrepancies may be due to the fact that using effective recombination coefficients to transpose the result obtained under random mating to the case of a partially selfing population (as we did to compute s_{HR}) should strictly only hold when loci are sufficiently tightly linked (e.g. Padhukasahasram *et al.* 2008; Roze 2016), while loosely linked loci may significantly contribute to selection for recombination when the selfing rate is high. They may also be caused by higher-order genetic associations (involving more than 2 selected loci) that are not taken into account in the analysis.

Figure 4 shows the effects of the deleterious mutation rate U, population size N, selection and dominance coefficients of deleterious alleles (s, h) on the evolutionarily stable map length (Supplementary Fig. 2 shows the same results with R on a log scale). As predicted, increasing U and/or decreasing N leads to stronger effects of Hill–Robertson interference between deleterious alleles, favoring higher values of R (see also Roze 2021). As can be seen in Fig. 4a, our analytical approximations overestimate the ES map length when U is high: this is probably due to the effect of higher-order genetic associations. Furthermore, in some simulations with $U = 0.5$, R fell to a very small value at some point during the simulation, which led to a quick accumulation of heterozygous mutations, the only surviving individuals being heterozygous for haplotypes carrying different deleterious alleles in repulsion. This accumulation of slightly deleterious mutations at the heterozygous state in a low recombining region has previously been described as pseudo-overdominance (e.g. Charlesworth and Charlesworth 1997; Pálsson and Pamilo 1999; Waller 2021). This occurred for $0.5 \leq \sigma \leq 0.7$, in which case the simulation was terminated, as the number of segregating mutations quickly became very large. Figure 4c shows that selection for recombination is stronger when deleterious alleles are more weakly selected, as already found by Roze (2021); furthermore, above a given value of s, the equilibrium map length decreases as the selfing rate increases (which can be understood from the reasoning given in the previous subsection and Fig. 2). By contrast, the dominance coefficient h of deleterious alleles has only little effect on the ES map length (Fig. 4d). Supplementary Figs. 3 and 4 show that the results are not significantly affected by introducing variability in the selection coefficients of deleterious alleles into the simulation program, nor by limiting to 100 or 1,000 the number of loci at which deleterious mutations may occur. Supplementary Fig. 5 shows that the strength of selection for recombination due to the Hill–

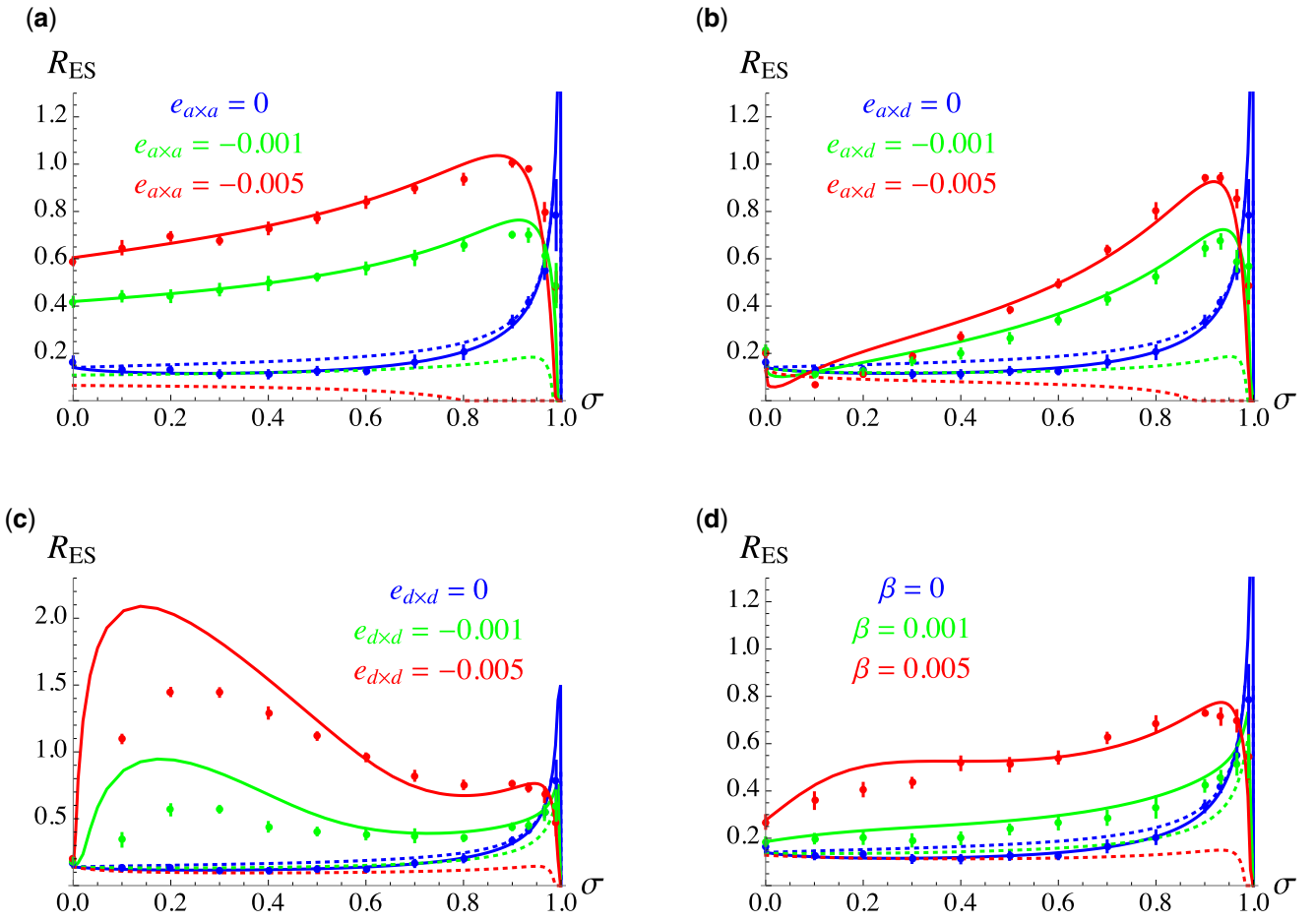


Fig. 5. Effect of the selfing rate σ on the evolutionarily stable map length R_{ES}, for different values of additive-by-additive (a), additive-by-dominance (b), and dominance-by-dominance (c) epistasis, and different values of Charlesworth *et al.*'s (1991) synergistic epistasis coefficient β (d). Note the different scale of the *y*-axis in (c). Dots correspond to simulation results, curves have the same meaning as in Fig. 3, and default parameter values are as in Fig. 3 with $c = 0.001$.

Robertson effect is accurately predicted by Equation (22) when selection against deleterious alleles is sufficiently weak, as long as the selfing rate is not too high.

Figure 5 shows the effect of each form of diploid epistasis ($e_{a\times a}$, $e_{a\times d}$, $e_{d\times d}$) and of Charlesworth *et al.*'s (1991) synergistic epistasis coefficient β (that combines the 3 forms of epistasis, see *Methods*) on the ES map length. Note that only negative values of epistasis were considered, as combinations of deleterious mutations quickly become beneficial when epistasis is positive and U is not very small. Negative epistasis tends to lower the strength of the Hill–Robertson effect (dotted curves in Fig. 5), by increasing the effective strength of selection against deleterious alleles. Nevertheless, the deterministic effects generated by negative epistasis increase the overall strength of selection for recombination. Supplementary Fig. 6 shows that this increase is driven by the term in D_{ab} in the case of negative additive-by-additive ($e_{a\times a}$) and additive-by-dominance ($e_{a\times d}$) epistasis, the term in $D^0_{mab,ab}$ being negligible when $e_{a\times a} < 0$, while it disfavors recombination when $e_{a\times d} < 0$ (one can show that this last effect is generated by the first term of Equation (13), since effective dominance coefficients $a_{j,j}$ are increased by $e_{a\times d}$, as shown by Equation (7)). In the case of dominance-by-dominance epistasis ($e_{d\times d}$), the large increase in recombination observed for moderate values of the selfing rate is generated by the term in $D^0_{mab,ab}$, and thus corresponds to the benefit of recombination previously described by Roze and Lenormand (2005). One can note that the model overestimates the strength of selection for recombination in this case. This is possibly due to the fact that, while the effect of $e_{d\times d}$ on effective dominance coefficients $a_{j,j}$ is negligible as long as epistasis is sufficiently weak, one can show that $a_{j,j}$ increases with $e_{d\times d}$ when dominance-by-dominance epistasis becomes the main source of selection against deleterious alleles (see Appendix G in Roze 2009), reducing selection for recombination through the first term of Equation (13).

Discussion

Digging into the causes of general empirical patterns such as the positive correlation between selfing rate and chiasma frequency observed within several families of flowering plants may help us to gain a better understanding of the selective forces affecting the evolution of recombination rates. While there is no obvious reason why the mechanistic constraints associated with chromosomal segregation during meiosis should differ between outcrossing and selfing species, the mating system of organisms does affect the benefit of recombination through its effect on genetic variation. Although indirect selective forces acting on recombination are expected to vanish under complete selfing (as heterozygosity should then be very rare), the results presented in this article show that intermediate selfing rates may either increase or decrease selection for recombination caused by interference (Hill–Robertson effect) among deleterious mutations,

depending on parameter values. Roughly, selfing leads to stronger selection for increased chromosomal map length as long as $sh \ll R(1-F)$ (mostly due to the fact that selfing reduces effective recombination rates, thus increasing the strength of genetic associations), while selfing decreases selection for recombination when $sh \gg R(1-F)$, due to the fact that changes in recombination rates have little effect on genetic associations in this regime. Given that most deleterious mutations seem to have weak fitness effects (e.g. Charlesworth 2015), selection for recombination should thus be increased by selfing, and be maximized for selfing rates slightly below 1 (as illustrated by Figs. 3 and 4).

In agreement with previous results (Otto and Barton 2001; Keightley and Otto 2006; Roze 2021), we found that interference is often the main driver of selection for recombination when linkage is tight (or when the selfing rate is strong), while deterministic effects tend to become more important when recombination is frequent. In the absence of epistasis, the variance in the degree of inbreeding among individuals caused by partial selfing (associated with a more efficient purging of deleterious alleles in more inbred lineages) generates positive associations among deleterious alleles, selecting against recombination in infinite populations. As a result, the equilibrium map length may be lower under moderate selfing rates than under random mating in large, highly recombining populations, as can be seen in Fig. 3 with $c = 10^{-4}$. As in standard models of infinite, randomly mating populations (Charlesworth 1990; Barton 1995), negative epistasis between mutations tends to favor recombination by generating negative LD between deleterious alleles. This effect is often maximized at high selfing rates, again due to the fact that selfing reduces effective recombination rates, thus increasing the magnitude of LD. Furthermore, components of epistasis involving dominance ($e_{a\times d}$, $e_{d\times d}$) also contribute to generating linkage disequilibrium (through an effective epistasis parameter) when the selfing rate is greater than zero.

Under high effective recombination, Roze and Lenormand (2005) showed that correlations in homozygosity among loci generated by partial selfing (identity disequilibria) are the main drivers of indirect selection on recombination rates, favoring higher recombination when dominance-by-dominance epistasis is negative (as recombination then benefits from a short-term advantage, by increasing the mean fitness of offspring). The results of the present article temper the relative importance of this effect, by showing that it becomes negligible relative to the effect of deterministically generated LD and Hill–Robertson interference under high selfing and in the case of tightly linked loci. Nevertheless, moderate selfing rates may favor a significant increase in chromosomal map length through this short-term benefit of breaking identity disequilibria, provided that dominance-by-dominance interactions are negative and represent a major component of epistasis (Fig. 5c). In principle, the average sign and overall importance of dominance-by-dominance effects can be inferred from the shape of the relation between the degree of inbreeding of individuals and their fitness (Crow and Kimura 1970, p. 80), negative $e_{d\times d}$ causing a faster than linear decline in fitness with inbreeding. This method was used on several plant species and did not yield any clear evidence for negative $e_{d\times d}$ (Willis 1993; Falconer and Mackay 1996); however, the methodology used (involving experimental crosses to increase the degree of inbreeding of individuals) generates a bias against finding negative $e_{d\times d}$, as deleterious alleles may have been purged from the more highly inbred lines. Using an experimental protocol that avoids this bias, Sharp and Agrawal (2016) found evidence for negative $e_{d\times d}$ (on viability) between EMS-induced mutations in *Drosophila melanogaster*. While more work is needed to assess the generality of this result, previous experimental studies showed that epistasis is typically quite variable among pairs of loci (de Visser and Elena 2007; Kouyos *et al.* 2007; Martin *et al.* 2007). As shown by Otto and Feldman (1997), recombination tends to be less favored when epistasis is variable, and it would thus be of interest to extend our model to more realistic fitness landscapes including distributions of epistasis.

While the indirect benefits of increased crossover rates may be strong when recombination is rare (e.g. Keightley and Otto 2006), they typically become rather weak under frequent recombination (Roze 2021), and one may thus wonder to what extent our model can explain a positive effect of selfing on chiasma frequency when at least 1 crossover per chromosome occurs during meiosis. When our model is modified so that 1 crossover per bivalent necessarily occurs (leading to a minimum map length of 50 cM, that is, $R = 0.5$) and letting R evolve above 0.5 using a similar model as before, one indeed observes very limited effects of indirect selection on the evolutionarily stable map length (except for high selfing rates) in the absence of epistasis, for a deleterious mutation rate per chromosome of $U = 0.1$ (Fig. 6). Substantial increases in recombination under partial selfing can be favored in the presence of negative $e_{d\times d}$, however, chromosomal map length being maximized for moderate selfing rates in this case (Fig. 6a). Furthermore, higher chromosomal mutation rates lead to important increases in R at high selfing rates due to stronger Hill–Robertson effects, as can be seen on Fig. 6b for $U = 0.5$. More generally, equilibrium values of R with 1 obligate crossover per bivalent should be approximately the same as in Figs. 3–5 when setting to 0.5 the points falling below $R = 0.5$ (with some slight differences caused by the fact that the distribution of the number of crossovers per chromosome is not Poisson anymore in this modified model).

Are these results consistent with empirical data on the effect of selfing on recombination? Supplementary Fig. 7 shows chiasma count data from several families of Angiosperms (comparing closely related species with contrasted mating systems) that were used to generate Figure 1 in Roze and Lenormand (2005). As can be seen on Supplementary Fig. 7, the increase in chromosome map length in selfing species compared with their outcrossing relatives is generally of the order 20–30 cM, which seems consistent with the results obtained here for moderate chromosomal mutation rates, either due to the Hill–Robertson effect or to negative values of $e_{a\times a}$, $e_{a\times d}$ or Charlesworth *et al.*'s (1991) synergistic epistasis coefficient β. The results obtained under negative $e_{d\times d}$ seem less consistent with the empirical patterns, as they show that R should be maximized for moderate selfing rates (Figs. 5 and 6). As the data do not include any precise selfing rate estimate, it is difficult to assess whether the observed patterns are more consistent with a gradual increase in R as selfing increases (as predicted for example under negative $e_{a\times a}$, $e_{a\times d}$ or β) or with a sharper increase at higher selfing rates (as predicted under the action of the Hill–Robertson effect alone). Further insights could be gained by comparing the genetic maps of closely related species for which estimates of selfing rates are available.

As already noted in Roze (2021), the fact that chromosomal map length may evolve toward values greater than 0.5 even in the absence of selfing (as can be seen on Supplementary Fig. 7) may seem at odds with the predictions of our model. Several explanations can be proposed to explain this discrepancy. First, the direct fitness cost of increasing R above 0.5 may be very low, as suggested by recent experiments on *A. thaliana* mutants in which map length is increased up to 7 to 8-fold without any clear effect on fertility

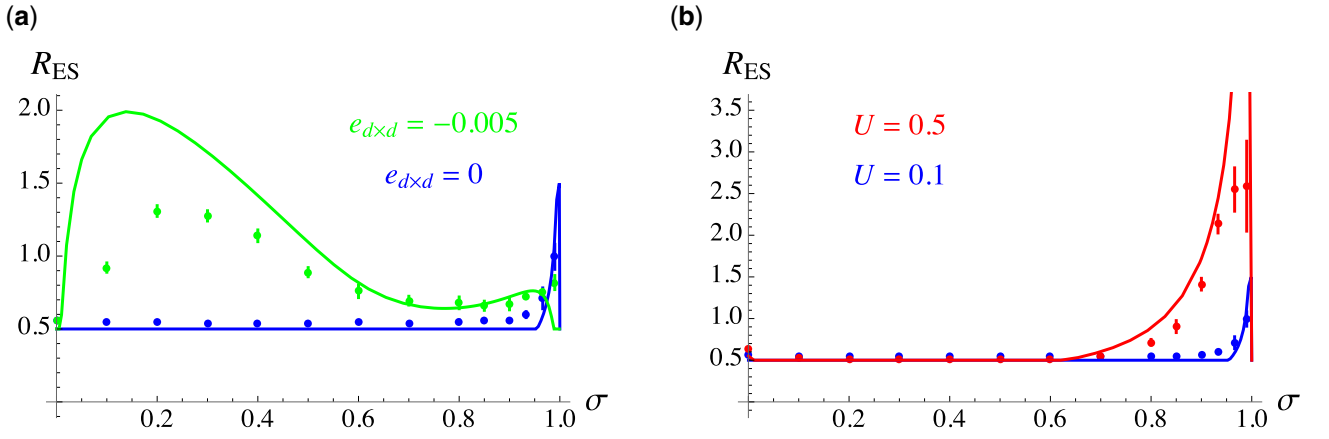


Fig. 6. Effect of the selfing rate σ on the evolutionarily stable map length R_{ES}, when at least 1 crossover per chromosome occurs during meiosis (leading to a minimal map length of $R = 0.5$). Dots correspond to simulation results, curves correspond to analytical predictions; default parameter values are as in Fig. 3 with $c = 0.001$. a) Blue: no epistasis, green: $e_{d\times d} = -0.005$. b) Blue: $U = 0.1$, red: $U = 0.5$.

(Fernandes *et al.* 2018). Second, relaxing the (unrealistic) assumption that crossovers occur anywhere along the chromosome with the same probability may generate stronger effects of indirect selection on the evolutionary stable map length. In particular, data from *Caenorhabditis elegans* and humans suggest that structural constraints may impose restrictions on the possible localization of obligate crossovers along chromosomes (Koehler *et al.* 1996; Ottolini *et al.* 2015; Altendorfer *et al.* 2020); the fact that crossovers are often more frequent in subtelomeric regions (in plants and animals) may possibly reflect such constraints (e.g. Haenel *et al.* 2018). Restricting the position of the obligate crossover in a given chromosomal region should increase the strength of indirect selection acting on a modifier allele increasing recombination in other regions, particularly if the modifier is located in regions with lower recombination. Effects of nonuniform positions of crossovers along chromosomes and crossover interference will be explored in a future work. Third, sweeps of beneficial mutations may also increase the strength of indirect selection for recombination (Hartfield *et al.* 2010; Roze 2021). Note that this may also enhance the effect of selfing on recombination, since transitions toward predominant selfing are often followed by a number of phenotypic changes such as reduced flower size ("selfing syndrome," e.g. Cutter 2019) that may be seen as adaptations to the new mating system, thus involving the spread of newly beneficial alleles that may (at least transiently) favor higher recombination rates. Recombination rates may also be further increased after transitions to selfing due to stronger Hill–Robertson effects caused by population bottlenecks and extinction/recolonization dynamics that often characterize selfing populations (e.g. Guo *et al.* 2009; Willi *et al.* 2018; Orsucci *et al.* 2020).

Interestingly, predominantly selfing species often maintain low rates of outcrossing in nature (e.g. Bonnin *et al.* 2001; Bomblies *et al.* 2010), which may reflect the effect of selective forces favoring the maintenance of recombination. Indeed, deterministic and stochastic multilocus models showed that individuals outcrossing at a low rate are often selectively favored over complete selfers in conditions where recombination is advantageous (negative epistasis or finite population size, Charlesworth *et al.* 1991; Kamran-Disfani and Agrawal 2014). However, higher effective recombination rates could in principle be achieved by increasing either the outcrossing rate of individuals or the number of crossovers at meiosis, and one could imagine that one solution or the other may be favored depending on the different types of direct and indirect selective forces that may act on outcrossing versus recombination modifiers. Exploring the joint evolution of outcrossing and recombination rates in predominantly selfing species would thus be of interest, both from a theoretical and empirical perspective.

Data availability

All derivations are provided in the *Mathematica* notebook available as Supplementary material at doi.org/10.5281/zenodo.6783728, along with the C++ codes used to run the simulations.

Supplemental material is available at *GENETICS* online.

Acknowledgments

We thank 4 anonymous reviewers for helpful comments, and the Bioinformatics and Computing Service of Roscoff's Biological Station (Abims platform) for computing time.

Funding

This work was funded by the Agence Nationale pour la Recherche (SelfRecomb project: ANR-18-CE02-0017-02, and GenAsex project: ANR-17-CE02-0016-01).

Conflicts of interest

None declared.

Appendix A: High effective recombination

Under high effective recombination, the associations between the modifier and the selected loci involved in Equation (6) are given by (see *Mathematica* notebook for derivation):

$$D_{mab,a} \approx D_{ma,ab} \approx F(\alpha_b + \alpha_{b,b})D^0_{mab,ab} \quad \text{(A1)}$$

$$D_{ma} \approx \frac{(1+2Fr_{ma})[2F(\alpha_a+\alpha_{a,a})\alpha_{b,b}+\alpha_{ab,b}+\alpha_{ab,ab}]}{r_{ma}(1-F)} D^0_{mab,ab} \quad \text{(A2)}$$

$$D_{m,a} \approx \frac{F(1+2r_{ma})}{1+2Fr_{ma}} D_{ma} \quad \text{(A3)}$$

$D_{mab,b}$, $D_{mb,ab}$, $D_{mb,}$ and $D_{m,b}$ being given by symmetric expressions. D_{mab}, $D_{m,ab}$, and $D_{ma,b}$ are given by:

$$\begin{aligned} D_{mab} \approx & \frac{(1+2Fr_{ma})(1+2Fr_{mb})}{(1-F)[r_{mab}+2F(r_{ma}r_{mb}+r_{ma}r_{ab}+r_{mb}r_{ab})+6F^2r_{ma}r_{mb}r_{ab}]} \\ & \times[(1+2Fr_{ab})T_{ab}D^0_{mab,ab} \\ & -\delta r_{ab}H_m(D_{ab}-D_{a,b})pq_m-\delta r_{ab}(1-H_m)(D_{mab,m}-D_{ma,mb})] \end{aligned} \quad \text{(A4)}$$

$$D_{m,ab} \approx \frac{F[(1+2Fr_{ma})(1+2Fr_{mb})+(1-F)(r_{ma}+r_{mb}+4Fr_{ma}r_{mb})]}{(1+2Fr_{ma})(1+2Fr_{mb})} D_{mab} \quad \text{(A5)}$$

$$\begin{aligned} D_{ma,b} \approx & \frac{F}{(1-F)[r_{mab}+2F(r_{ma}r_{mb}+r_{ma}r_{ab}+r_{mb}r_{ab})+6F^2r_{ma}r_{mb}r_{ab}]} \\ & \times[(1+2Fr_{ma})[(1+2Fr_{mb})(1+2Fr_{ab}) \\ & +(1-F)(r_{mb}+r_{ab}+4Fr_{mb}r_{ab})]T_{ab}D^0_{mab,ab} \\ & -(1+2Fr_{mb})[1-(1-3F)r_{ma}][\delta r_{ab}H_m(D_{ab}-D_{a,b})pq_m \\ & +\delta r_{ab}(1-H_m)(D_{mab,m}-D_{ma,mb})]] \end{aligned} \quad \text{(A6)}$$

($D_{mb,a}$ being given by a symmetric expression), with $H_m = h_m + (1-2h_m)p_m$, and:

$$T_{ab} = \alpha_{ab}+\alpha_{ab,a}+\alpha_{ab,b}+\alpha_{ab,ab}+2F(\alpha_a+\alpha_{a,a})(\alpha_b+\alpha_{b,b}) \quad \text{(A7)}$$

$$D_{ab}-D_{a,b} \approx \frac{(1-F)[\tilde{\alpha}_{ab}-\tilde{\alpha}_a\tilde{\alpha}_b+2FG_{ab}(\alpha_a+\alpha_{a,a})(\alpha_b+\alpha_{b,b})]}{r_{ab}(1-F)} pq_{ab} \quad \text{(A8)}$$

$$D_{mab,m}-D_{ma,mb} \approx \frac{F(r_{ma}+r_{mb}-r_{ab}-2r_{ma}r_{mb})}{1+F(r_{ma}+r_{mb}+r_{ab}-2r_{ma}r_{mb})}(D_{ab}-D_{a,b})pq_m. \quad \text{(A9)}$$

The identity disequilibrium G_{ab} is given by $\phi_{ab}-F^2$, with:

$$\phi_{ab} = F\left[1-\frac{2(1-F)r_{ab}(1-r_{ab})}{1+2Fr_{ab}(1-r_{ab})}\right]. \quad \text{(A10)}$$

Appendix B: Interpreting effective selection coefficients

The form of the coefficients $\tilde{\alpha}_a$ and $\tilde{\alpha}_{ab}$ representing the effective strength of selection against deleterious alleles and the effective epistasis between pairs of deleterious alleles in a partially selfing population can be understood as follows. Assuming that deleterious alleles stay rare in the population ($u \ll s$), we can neglect homozygotes for deleterious alleles produced by outcrossing, and consider that homozygosity for deleterious alleles necessarily implies identity-by-descent among these alleles. Then, if an individual carries allele a on one of its haplotypes, with probability F it also carries allele a on its second haplotype. In that case, the fitness effect of the deleterious allele (α_a) is increased by α_a due to the presence of a on the second haplotype, and by $\alpha_{a,a}$ due to the interaction among those 2 deleterious alleles: therefore, the effective strength of selection experienced by deleterious alleles is $\tilde{\alpha}_a = (1+F)\alpha_a + F\alpha_{a,a}$ (which, using Equation (5), yields $\tilde{\alpha}_a \approx -s\tilde{h}$). A similar reasoning can be used to compute the effective epistasis coefficient. With probability ϕ_{ab}, an individual is identical-by-descent at both loci a and b; with probability $F-\phi_{ab}$ it is identical-by-descent at the first locus and not at the second (or at the second and not at the first), while with the complementary probability $1-2F+\phi_{ab}$ it is identical-by-descent at neither locus. Therefore, an individual carrying alleles a and b on one of its haplotypes is $aabb$ with probability ϕ_{ab}, $aaBb$ (or $Aabb$) with probability $F-\phi_{ab}$, and $AaBb$ with probability $1+2F-\phi_{ab}$. In the first case ($aabb$), the fitness effect of the interaction between a and b (α_{ab}) is increased by the interaction between a and b present on the other haplotype (α_{ab}), twice the interaction between deleterious alleles in trans ($\alpha_{a,b}$), twice the additive-by-dominance interactions ($\alpha_{ab,a}$, $\alpha_{ab,b}$), and by the dominance-by-dominance interaction ($\alpha_{ab,ab}$). In the second case ($aaAb$ or $Aabb$), it is increased by the interaction between deleterious alleles in trans ($\alpha_{a,b}$) and by the additive-by-dominance interaction ($\alpha_{ab,a}$ or $\alpha_{ab,b}$). This yields:

$$\begin{aligned} \tilde{\alpha}_{ab} &= \phi_{ab}(2\alpha_{ab}+2\alpha_{a,b}+2\alpha_{ab,a}+2\alpha_{ab,b}+\alpha_{ab,ab}) \\ &\quad +(F-\phi_{ab})(2\alpha_{ab}+2\alpha_{a,b}+\alpha_{ab,a}+\alpha_{ab,b})+(1-2F+\phi_{ab})\alpha_{ab} \\ &= (1+\phi_{ab})\alpha_a+2F\alpha_{a,b}+(F+\phi_{ab})(\alpha_{ab,a}+\alpha_{ab,b})+\phi_{ab}\alpha_{ab,ab}. \end{aligned} \quad \text{(B1)}$$

Appendix C: General QLE approximations

The following expressions combine approximations obtained under different regimes (high effective recombination, weak recombination, high selfing, epistasis of order ϵ^2 or ϵ, see Supplementary material). The different associations affecting the change in frequency of the modifier (Equation (6)) are given by (with $D^0_{mab,ab}$ given by Equation (10)):

$$D_{mab,ab} \approx D^0_{mab,ab} + FD_{mab} \quad \text{(C1)}$$

$$D_{mab,a} \approx D_{ma,ab} \approx F[(\alpha_b+\alpha_{b,b})D^0_{mab,ab}+D_{mab}] \quad \text{(C2)}$$

$$D_{m,ab} \approx D_{ma,b} \approx D_{mb,a} \approx FD_{mab} \quad \text{(C3)}$$

$$D_{ma,a} \approx F[\alpha_{b,b}D^0_{mab,ab}+D_{ma}] \quad \text{(C4)}$$

$$D_{m,a} \approx FD_{ma} \quad \text{(C5)}$$

$$\begin{aligned} D_{mab} \approx & \frac{(1+2Fr_{ma})(1+2Fr_{mb})}{Y_{mab}}[(1+2Fr_{ab})T_{ab}D^0_{mab,ab} \\ & -\delta r_{ab}H_m(D_{ab}-D_{a,b})pq_m-\delta r_{ab}(1-H_m)(D_{mab,m}-D_{ma,mb})] \end{aligned} \quad \text{(C6)}$$

$$D_{ma} \approx \frac{(1+2Fr_{ma})[[2F(\alpha_a+\alpha_{a,a})\alpha_{b,b}+\alpha_{ab,b}+\alpha_{ab,ab}]D^0_{mab,ab}+(\tilde{\alpha}_b+\tilde{\alpha}_{ab})D_{mab}]}{r_{ma}(1-F)-(1+2Fr_{ma})\tilde{\alpha}_a} \quad \text{(C7)}$$

($D_{mab,b}$, $D_{mb,ab}$, $D_{mb,b}$, $D_{m,b}$, and D_{mb} being given by symmetric expressions), with:

$$T_{ab} = \alpha_{ab}+\alpha_{ab,a}+\alpha_{ab,b}+\alpha_{ab,ab}+2F(\alpha_a+\alpha_{a,a})(\alpha_b+\alpha_{b,b}) \quad \text{(C8)}$$

$$\begin{aligned} Y_{mab} = & (1-F)[r_{mab}+2(r_{ma}r_{mb}+r_{ma}r_{ab}+r_{mb}r_{ab})+6r_{ma}r_{mb}r_{ab}] \\ & -(1+2Fr_{ma})(1+2Fr_{mb})(1+2Fr_{ab})(\tilde{\alpha}_a+\tilde{\alpha}_b+\tilde{\alpha}_{ab}) \end{aligned} \quad \text{(C9)}$$

$$D_{ab}-D_{a,b} \approx \frac{(1-F)[\tilde{\alpha}_{ab}-\tilde{\alpha}_a\tilde{\alpha}_b+2FG_{ab}(\alpha_a+\alpha_{a,a})(\alpha_b+\alpha_{b,b})]}{r_{ab}(1-F)-(1+2Fr_{ab})(\tilde{\alpha}_a+\tilde{\alpha}_b+\tilde{\alpha}_{ab})} pq_{ab} \quad \text{(C10)}$$

$$D_{mab,m}-D_{ma,mb} \approx \frac{F(r_{ma}+r_{mb}-r_{ab}-2r_{ma}r_{mb})}{1+F(r_{ma}+r_{mb}+r_{ab}-2r_{ma}r_{mb})}(D_{ab}-D_{a,b})pq_m. \quad \text{(C11)}$$

The change in frequency of the modifier is given by $\Delta p_m = \sum_{\mathbb{U},\mathbb{V}} \alpha_{\mathbb{U},\mathbb{V}} D_{m\mathbb{U},\mathbb{V}}$ where the double sum is over all elements of the set $\{\varnothing, a, b, ab\}$, and thus decomposes into a term generated by $D^0_{mab,ab}$ (that becomes predominant under high effective recombination) and a term generated by $D_{ab}-D_{a,b}$ (that becomes predominant under weak recombination or strong selfing).

Literature cited

Abu Awad D, Roze D. Epistasis, inbreeding depression and the evolution of self-fertilization. Evolution. 2020;74(7):1301–1320.

Agrawal AF. Evolution of sex: why do organisms shuffle their genotypes? Curr Biol. 2006;16(17):R696–R704.

Altendorfer E, Láscarez-Lagunas LI, Nadarajan S, Mathieson I, Colaiácovo MP. Crossover position drives chromosome remodeling for accurate meiotic chromosome segregation. Curr Biol. 2020;30(7):1329–1338.

Barton NH. A general model for the evolution of recombination. Genet Res. 1995;65(2):123–144.

Barton NH, Otto SP. Evolution of recombination due to random drift. Genetics. 2005;169(4):2353–2370.

Barton NH, Turelli M. Natural and sexual selection on many loci. Genetics. 1991;127(1):229–255.

Benavente E, Sybenga J. The relation between pairing preference and chiasma frequency in tetrasomics of rye. Genome. 2004;47(1):122–133.

Blackwell AR, Dluzewska J, Szymanska-Lejman M, Desjardins S, Tock AJ, Kbiri N, Lambing C, Lawrence EJ, Bieluszewski T, Rowan B, *et al.* MSH2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. EMBO J. 2020;39(21):e104858.

Bomblies K, Yant L, Laitinen RA, Kim S-T, Hollister JD, Warthmann N, Fitz J, Weigel D. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. PLoS Genet. 2010;6(3):e1000890.

Bonnin I, Ronfort J, Wozniak F, Olivieri I. Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). Mol Ecol. 2001;10(6):1371–1383.

Borts RH, Haber JE. Meiotic recombination in yeast: alteration by multiple heterozygosities. Science. 1987;237(4821):1459–1465.

Brand CL, Cattani MV, Kingan SB, Landeen EL, Presgraves DC. Molecular evolution at a meiosis gene mediates species differences in the rate and patterning of recombination. Curr Biol. 2018;28(8):1289–1295.

Charlesworth B. Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet Res. 1990;55(3):199–221.

Charlesworth B. Causes of natural variation in fitness: evidence from studies of *Drosophila* populations. Proc Natl Acad Sci USA. 2015;112(6):1662–1669.

Charlesworth B, Charlesworth D. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet Res. 1997;70(1):63–73.

Charlesworth B, Morgan MT, Charlesworth D. Multilocus models of inbreeding depression with synergistic selection and partial self-fertilization. Genet Res. 1991;57(2):177–194.

Charlesworth D, Charlesworth B, Strobeck C. Effects of selfing on selection for recombination. Genetics. 1977;86(1):213–226.

Charlesworth D, Charlesworth B, Strobeck C. Selection for recombination in partially self-fertilizing populations. Genetics. 1979;93(1):237–244.

Chen W, Jinks-Robertson S. The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. Genetics. 1999;151(4):1299–1313.

Crow JF, Kimura M. An Introduction to Population Genetics Theory. New York: Harper and Row; 1970.

Cutter AD. Reproductive transitions in plants and animals: selfing syndrome, sexual selection and speciation. New Phytol. 2019;224(3):1080–1094.

de Visser JAGM, Elena SF. The evolution of sex: empirical insights into the roles of epistasis and drift. Nat Rev Genet. 2007;8(2):139–149.

Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. Harlow: Addison Wesley Longman; 1996.

Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737–756.

Fernandes JB, Séguéla-Arnaud M, Larchevêque C, Lloyd AH, Mercier R. Unleashing meiotic crossovers in hybrid plants. Proc Natl Acad Sci USA. 2018;115(10):2431–2436.

Gervais C, Roze D. Mutation rate evolution in partially selfing and partially asexual organisms. Genetics. 2017;207(4):1561–1575.

Glémin S, Ronfort J. Adaptation and maladaptation in selfing in outcrossing species: new mutations versus standing variation. Evolution. 2013;67(1):225–240.

Gray D, Cohen PE. Control of meiotic crossovers: from double-strand break formation to designation. Annu Rev Genet. 2016;50(1):175–210.

Guo YL, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc Natl Acad Sci USA. 2009;106(13):5246–5251.

Haenel Q, Laurentino TG, Roesti M, Berner D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol Ecol. 2018;27(11):2477–2497.

Haldane JBS. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet. 1919;8:299–309.

Hansson B, Kawabe A, Preuss S, Kuittinen H, Charlesworth D. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 1 and 2 and the corresponding *A. thaliana* chromosome 1: recombination rates, rearrangements and centromere location. Genet Res. 2006;87(2):75–85.

Hartfield M, Otto SP, Keightley PD. The role of advantageous mutations in enhancing the evolution of a recombination modifier. Genetics. 2010;184(4):1153–1164.

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966;8(3):269–294.

Holsinger KE, Feldman MW. Linkage modification with mixed random mating and selfing: a numerical study. Genetics. 1983;103(2):323–333.

Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605–1617.

Johnston SE, Bérénos C, Slate J, Pemberton JM. Conserved genetic architecture underlying individual recombination rate variation in a wild population of Soay sheep (*Ovis aries*). Genetics. 2016;203(1):583–598.

Kamran-Disfani A, Agrawal AF. Selfing, adaptation and background selection in finite populations. J Evol Biol. 2014;27(7):1360–1371.

Kawabe A, Hansson B, Forrest A, Hagenblad J, Charlesworth D. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 6 and 7 and *A. thaliana* chromosome IV: evolutionary history, rearrangements and local recombination rates. Genet Res. 2006;88(1):45–56.

Keightley PD, Otto SP. Interference among deleterious mutations favours sex and recombination in finite populations. Nature. 2006;443(7107):89–92.

Kirkpatrick M, Johnson T, Barton NH. General models of multilocus evolution. Genetics. 2002;161(4):1727–1750.

Koehler KE, Hawley RS, Sherman S, Hassold T. Recombination and nondisjunction in human and flies. Hum. Mol. Genet. 1996;5(Suppl. 1):1495–1504.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Bragi Walters G, Jonasdottir A, Gylfason A, Kristinsson KT, *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467(7319):1099–1103.

Kouyos RD, Silander OK, Bonhoeffer S. Epistasis between deleterious mutations and the evolution of recombination. Trends Ecol Evol. 2007;22(6):308–315.

Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppälä J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. Genetics. 2004;168(3):1575–1584.

Martin G, Elena SF, Lenormand T. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat Genet. 2007;39(4):555–560.

Nagylaki T. The evolution of multilocus systems under weak selection. Genetics. 1993;134(2):627–647.

Nordborg M. Structured coalescent processes on different time scales. Genetics. 1997;146(4):1501–1514.

Nordborg M. Linkage disequilibrium, gene trees and selfing: and ancestral recombination graph with partial self-fertilization. Genetics. 2000;154(2):923–929.

Orsucci M, Milesi P, Hansen J, Girodolle J, Glémin S, Lascoux M. Shift in ecological strategy helps marginal populations of shepherd's purse (*Capsella bursa-pastoris*) to overcome a high genetic load: competition avoidance and colonization. Proc Roy Soc Lond B. 2020;287:20200463.

Otto SP. Selective interference and the evolution of sex. J Hered. 2021;112(1):9–18.

Otto SP, Barton NH. The evolution of recombination: removing the limits to natural selection. Genetics. 1997;147(2):879–906.

Otto SP, Barton NH. Selection for recombination in small populations. Evolution. 2001;55(10):1921–1931.

Otto SP, Feldman MW. Deleterious mutations, variable epistatic interactions, and the evolution of recombination. Theor Popul Biol. 1997;51(2):134–147.

Otto SP, Lenormand T. Resolving the paradox of sex and recombination. Nat Rev Genet. 2002;3(4):252–261.

Ottolini CS, Newnham LJ, Capalbo A, Natesan SA, Joshi HA, Cimadomo D, Griffin DK, Sage K, Summers MC, Thornhill AR, et al Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. Nat Genet. 2015;47(7):727–737.

Padhukasahasram B, Marjoram P, Wall JD, Bustamante CD, Nordborg M. Exploring population genetics models with recombination using efficient forward-time simulations. Genetics. 2008;178(4):2417–2427.

Pálsson S, Pamilo P. The effects of deleterious mutations on linked, neutral variation in small populations. Genetics. 1999;153(1):475–483.

Pollak E. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics. 1987;117(2):353–360.

Ritz KR, Noor MAF, Singh ND. Variation in recombination rate: adaptive or not? Trends Genet. 2017;33(5):364–374.

Ross-Ibarra J. Genome size and recombination in angiosperms: a second look. J Evol Biol. 2007;20(2):800–806.

Roze D. Diploidy, population structure and the evolution of recombination. Am Nat. 2009;174(S1):S79–S94.

Roze D. Selection for sex in finite populations. J Evol Biol. 2014;27(7):1304–1322.

Roze D. Background selection in partially selfing populations. Genetics. 2016;203(2):937–957.

Roze D. A simple expression for the strength of selection on recombination generated by interference among mutations. Proc Natl Acad Sci USA. 2021;118:e2022805118.

Roze D, Barton NH. The Hill–Robertson effect and the evolution of recombination. Genetics. 2006;173(3):1793–1811.

Roze D, Lenormand T. Self-fertilization and the evolution of recombination. Genetics. 2005;170(2):841–857.

Samuk K, Manzano-Winkler B, Ritz KR, Noor MAF. Natural selection shapes variation in genome-wide recombination rate in *Drosophila pseudoobscura*. Curr Biol. 2020;30(8):1517–1528.

Sharp NP, Agrawal AF. The decline in fitness with inbreeding: evidence for negative dominance-by-dominance epistasis in *Drosophila melanogaster*. J Evol Biol. 2016;29(4):857–864.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Phil Trans R Soc B. 2017;372(1736):20160455.

Waller DM. Addressing Darwin's dilemma: can pseudo-overdominance explain persistent inbreeding depression and load? Evolution. 2021;75(4):779–793.

Weir BS, Cockerham CC. Mixed self and random mating at two loci. Genet Res. 1973;21(3):247–262.

Willi Y, Fracassetti M, Zoller S, Van Buskirk J. Accumulation of mutational load at the edges of a species range. Mol Biol Evol. 2018;35(4):781–791.

Willis JH. Effects of different levels of inbreeding on fitness components in *Mimulus guttatus*. Evolution. 1993;47(3):864–876.

Wright KM, Arnold B, Xue K, Šurinová M, O'Connell J, Bomblies K. Selection on meiosis genes in diploid and tetraploid *Arabidopsis arenosa*. Mol Biol Evol. 2015;32(4):944–955.

Wright SI, Ness RW, Foxe JP, Barrett SCH. Genomic consequences of outcrossing and selfing in plants. Int J Plant Sci. 2008;169(1):105–118.

Zelkowski M, Olson MA, Wang M, Pawlowski W. Diversity and determinants of meiotic recombination landscapes. Trends Genet. 2019;35(5):359–370.

Ziolkowski PA, Berchowitz LE, Lambing C, Yelina NE, Zhao X, Kelly KA, Choi K, Ziolkowska L, June V, Sanchez-Moran E, *et al.* Juxtaposition of heterozygous and homozygous regions causes reciprocal crossover remodelling via interference during Arabidopsis meiosis. eLife 2015;4:e03708.

Communicating editor: A. Agrawal

Chapter 2

The evolution of recombination with variable recombination rate along the genome

1 Introduction

The question of the evolution of sex and recombination has been an active field of research in evolutionary biology. In particular, theoretical population genetics has produced a rich literature proposing a variety of mechanisms that could explain why most eukaryotes maintain high recombination rates (Otto and Lenormand, 2002; Agrawal, 2006b). One of the most general mechanisms proposed to explain the maintenance of recombination is know as selective interference or the Hill-Robertson effect (Hill and Robertson, 1966; Felsenstein, 1974; Barton and Otto, 2005; Keightley and Otto, 2006; Roze, 2021; Otto, 2021). Selective interference corresponds to the fact that the joint effects of genetic drift and selection generate an excess of genetic combinations with a mix of deleterious and advantageous variants (negative linkage disequilibrium), decreasing the variance in fitness and thus the efficacy of selection. Recombination is then favored because it tends to break negative LD, increasing the variance in fitness and thus the efficacy of

selection. The strength of selection to increase the genetic map length of a chromosome due to interference among deleterious mutations has been quantified in previous works (Roze, 2021; Stetsenko and Roze, 2022). While selection for recombination can be strong when recombination is rare, it decreases rapidly as recombination increases, and typically becomes quite small when one crossover (CO) per chromosome occurs on average. Given that the range of the number of COs per chromosome in eukaryotes is typically comprised between 1 and 3 or 4 (Stapley et al., 2017; Fernandes et al., 2018; Brazier and Glémin, 2022), one may thus conclude that the effect of indirect selection on the evolution of recombination may be negligible. However, the models cited above make a number of simplifying assumptions: in particular, a uniform distribution of mutations and COs along chromosomes, and no interference among COs (the number of COs per chromosome being Poisson distributed). Yet, recombination rates are known to vary along genomes at multiple scales (Ritz et al., 2017; Stapley et al., 2017; Peñalba and Wolf, 2020) and genome-wide recombination rates are the result of different selective pressures (direct and indirect) that vary along the genome because of the heterogeneity of different genomic features (e.g. gene density, methylation patterns, inversions, recombination hotspots). Furthermore, the number of COs per chromosome is typically not Poisson distributed: in particular, a minimum of one CO per bivalent occurs in most species, and is thought to be required to ensure proper chromosomal segregation. However, one could imagine that a similar mechanical constraint may act on the position of this obligate CO along the chromosome: for example, a CO in the subtelomeric region may prove optimal for chromosomal segregation, or alternatively may result from that fact that the pairing of homologs starts from the telomeres (Haenel et al., 2018; Brazier and Glémin, 2022). If the obligate CO tends to occur in a specific chromosomal region (e.g., near the extremities), this should increase the strength of indirect selection for extra COs in other chromosomal regions, due to interference among selected loci in these regions.

Another simplifying assumption made in Roze (2021) and Stetsenko and Roze (2022) is that chromosomal map length is controlled by a single locus (with an infinite number of possible alleles). However, recent works on the genetic basis of recombination rates

variation tend to show that recombination is either a polygenic trait (encoded by multiple loci across the genome) or an oligogenic trait (mainly encoded by a few loci) according to the species considered (e.g. Kong et al. 2014; Johnston et al. 2016 but see Zelkowski et al. 2019 for a review). Roze (2021) showed that considering multiple modifier loci that collectively affect the map length of the chromosome does not affect the chromosome map length at equilibrium. However, different results may be obtained in the case of multiple recombination modifiers, each having a local effect on a given portion of chromosome.

In this chapter, I present preliminary results on the effect of a non-uniform recombination rate along the chromosome and of multiple recombination modifier loci on the indirect strength of selection for recombination. The simulation results show that more COs can be maintained in randomly mating populations when the recombination modifier locus is located away from the obligate CO and when the modifier locus is located in a gene-dense region. The number of COs also increases with the number of modifier loci, each having a local effect on a portion of chromosome.

2 Model

The baseline program is written in C++ and is a modified version of the program from Stetsenko and Roze (2022) except that either a non-uniform recombination rate or multiple recombination modifier loci were introduced. The program models a population of N diploids each carrying a pair of linear chromosomes where deleterious mutations are introduced uniformly along the chromosome at a rate U per generation per chromosome. Each mutation affects fitness by a factor $1-sh$ in the heterozygous state and $1-s$ in the homozygous state. Mutations at different sites have multiplicative effects (no epistasis). A direct fitness effect of the chromosome map length R, W_c is introduced and set to $W_c = \exp(-cR)$, so that c measures a direct fitness cost per CO (reflecting a possible fertility cost of having too many COs). For each parameter set an evolutionarily stable map length is computed as the average map length among all individuals after reaching equilibrium. For further details about the simulation program see Chapter 1, Methods

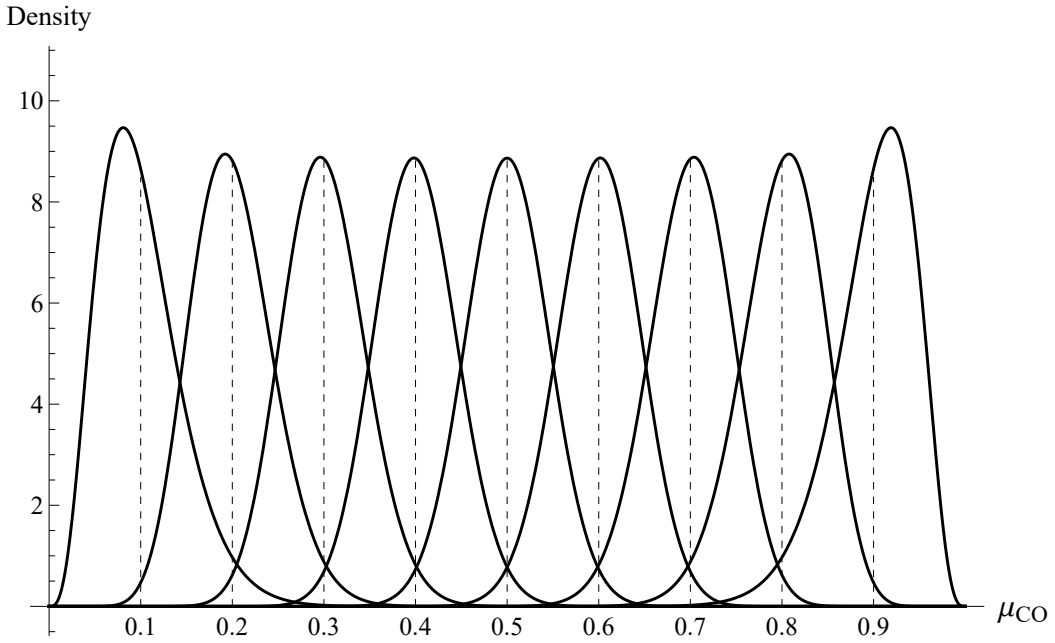


Figure 2.1: Probability density function for the position of the obligate CO along the chromosome, modelled by a beta distribution with variance $\sigma^2_{CO} = 0.002$ and mean μ_{CO} taking the following values: 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9.

section. The effect of the mating system has started to be explored by running the simulations under different selfing rates, although only the case of randomly mating populations is presented here.

2.1 Non-uniform recombination rate

The baseline program was first modified to add an obligate CO per bivalent (so that chromosomal map length is at least 0.5 Morgan), whose position can be restricted to a given chromosomal region. In particular, the position of this obligate CO was drawn from a beta distribution with mean μ_{CO} and variance σ^2_{CO} (Figure 2.1). The number of additional COs is controlled by a single modifier locus (whose position can vary along the chromosome), with an infinite number of possible alleles coding for different values of R. Mutations at the modifier locus are introduced at a constant rate. As in Roze (2021) and Stetsenko and Roze (2022), the number of additional COs is drawn from a Poisson distribution (with parameter equal to the average of the two modifier alleles),

and the position of each of these additional COs is drawn from a uniform distribution along the chromosome. A different version of this program was used to study the effect of heterogeneity in the distribution of coding sequences along chromosomes: in this case, deleterious mutations were assumed to occur only within a particular region of the chromosome.

2.2 Multiple recombination modifier loci with local effects

The baseline program was also modified to consider a number n_{mod} of recombination modifier loci evenly spaced along the chromosome. Each modifier is located at the mid-point of a segment of chromosome (of physical length $1/n_{\text{mod}}$ times the total length of the chromosome), and only affects the map length (average number of COs) of this segment. The total chromosome map length is thus the sum of map lengths encoded by all modifier loci along the chromosome. Each modifier locus initially codes for a local map length of $R_{\text{init}}/n_{\text{mod}}$ where R_{init} (fixed to 1) is the initial map length of the chromosome, and mutates at a rate 10^{-4} per generation. As in previous programs (Roze, 2021; Stetsenko and Roze, 2022) large effect mutations are introduced with probability 0.05 (see Chapter 1 for further details).

3 Results

3.1 Non-uniform recombination rate

We found – under the parameter values explored – that an evolutionarily stable (ES) map length R_{ES} higher than 0.5 can evolve when the the obligate CO is sufficiently far from the recombination modifier locus (Figure 2.2). Indeed, indirect selection for recombination primarily stems from selective interference between selected loci close to the recombination modifier locus (Roze, 2021). Forcing the obligate CO to fall close to the modifier locus drastically decreases selective interference in its vicinity, strongly reducing indirect selection for extra COs. Because recombination rates stay very low in other chromosomal regions, deleterious mutations may accumulate in these regions.

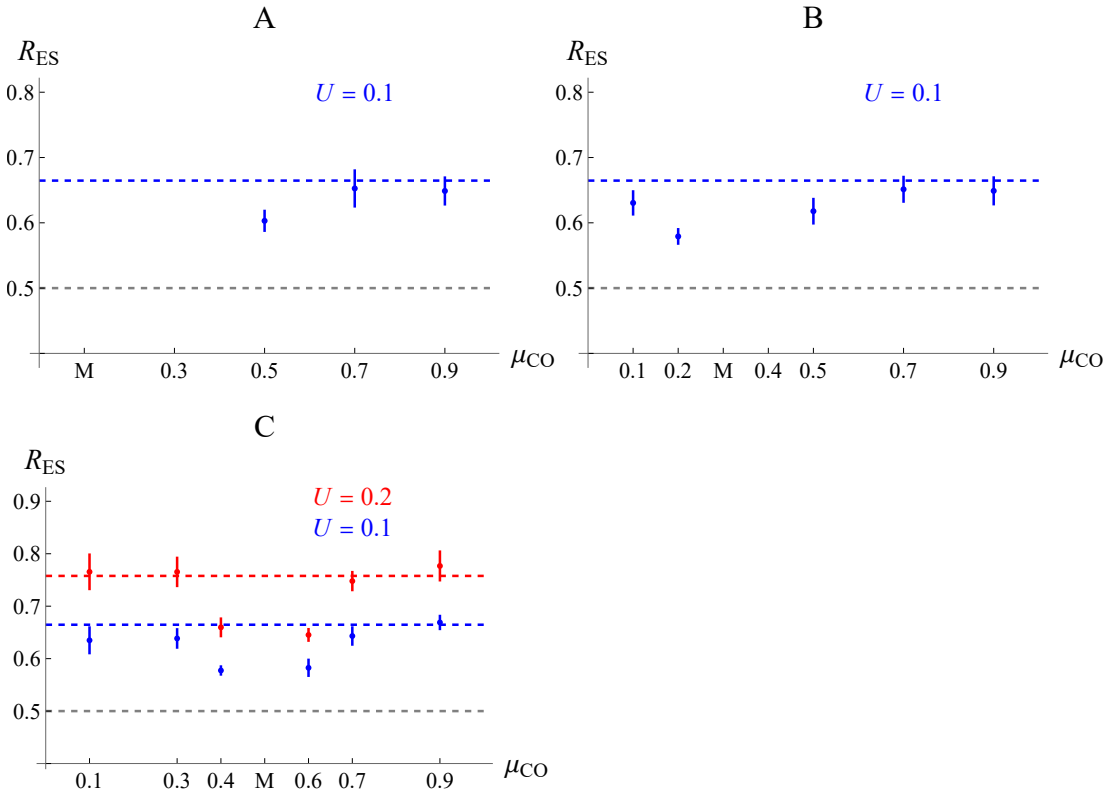


Figure 2.2: Effect of the mean position of the obligate CO μ_{CO} on the evolutionarily stable map length R_{ES}, for different positions of the recombination modifier locus M (A: 0.1; B: 0.3; C: 0.5) and for different values of the deleterious mutation rate U (blue: $U = 0.1$, red: $U = 0.2$). The dashed grey line represents the minimum map length of 0.5 (only the obligate CO), while the coloured dashed line correspond to 0.5 plus the evolutionarily stable map length obtained when COs are uniformly distributed along the chromosome and the modifier locus is at the midpoint of the chromosome (obtained from Roze 2021; with colors corresponding to the mutation rate). Missing points correspond to simulations that had to be stopped due to an accumulation of deleterious mutations. Other parameters are: $\sigma^2_{CO} = 0.002$; $s = 0.02$; $h = 0.2$, $c = 0.001$; $N = 20,000$.

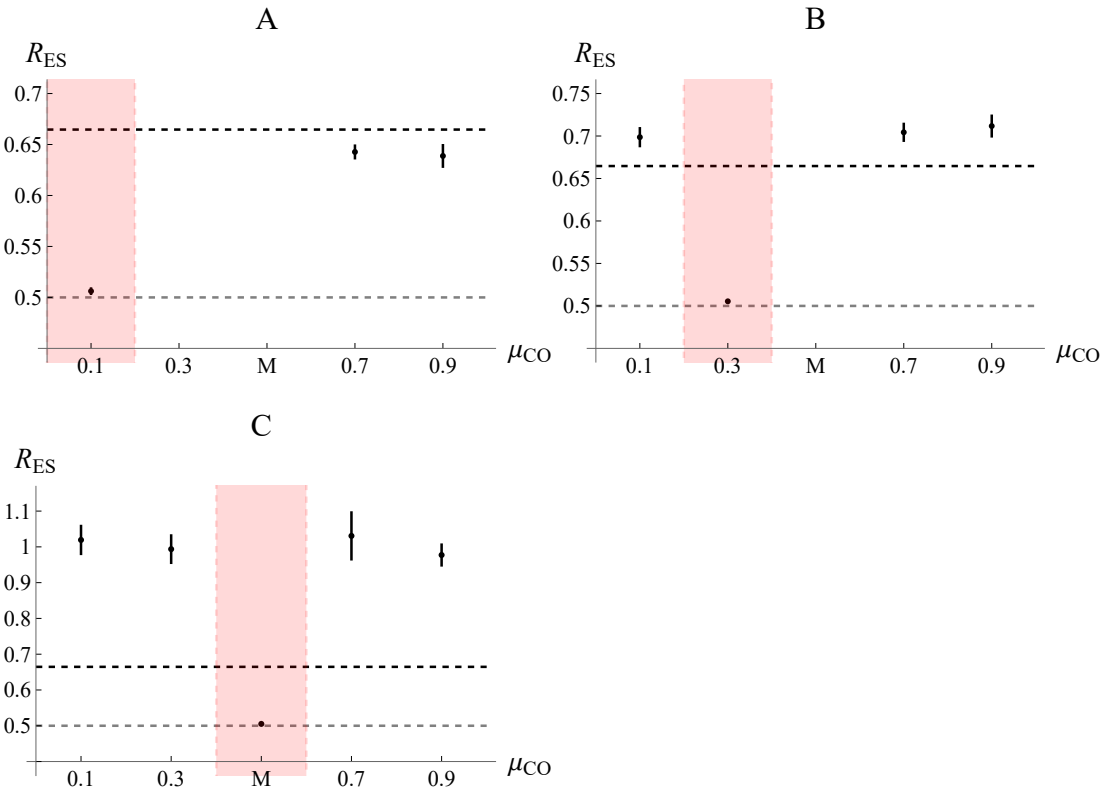


Figure 2.3: Effect of the mean position of the obligate CO μ_{CO} on the evolutionarily stable map length R_{ES}, for different regions of the chromosome where deleterious mutations can occur (red area, A: 0-0.2; B: 0.2-0.4; C: 0.4-0.6). The dashed grey line represents the minimum map length of 0.5 (only the obligate CO), while the coloured dashed line correspond to 0.5 plus the evolutionarily stable map length obtained when COs and deleterious mutations are uniformly distributed along the chromosome, and the modifier locus is at the midpoint of the chromosome (obtained from Roze 2021). Missing points correspond to simulations that had to be stopped due to an accumulation of deleterious mutations. Other parameters are: $\sigma^2_{CO} = 0.002$; $s = 0.02$; $h = 0.2$, $c = 0.001$; $N = 20,000$, $U = 0.1$.

When the obligate CO falls sufficiently far from the modifier locus its effect on selective interference in the vicinity of the modifier locus is negligible, and extra COs are selected for because they decrease selective interference in this region. In this case, the ES number of extra COs is roughly the same as R_{ES} in Roze (2021) (no obligate CO, COs uniformly distributed along the chromosome and modifier locus at the midpoint of the chromosome) with the same parameter values (blue and red dashed lines in Figure 2.2).

Restricting deleterious mutations to occur on a specific portion of the chromosome only can indicate the influence of this portion on indirect selection acting on the modifier. As shown by Figure 2.3A, the population evolves towards a regime where deleterious mutations accumulate (due to low recombination rates) when the obligate CO is located between the region where deleterious mutations occur and the modifier locus. This result stems from the fact that COs occurring between deleterious mutations and the modifier locus break linkage disequilibrium between selected loci and the recombination modifier locus, decreasing the effect of indirect selection. Contrary to the case of a uniform density of mutations along the chromosome (see above; Figure 2.2), when the obligate CO falls sufficiently far from the modifier locus, the ES number of extra COs can be lower or higher than R_{ES} in Roze (2021) for the same parameter values (Figure 2.3). The extra R_{ES} depends on the relative position of the modifier locus and the mutation dense region: increasing the density of selected loci near the modifier increases R_{ES} (Figure 2.3C), while R_{ES} decreases when selected loci are located in a different chromosomal region (Figure 2.3A). This is due to the fact that indirect selection is mostly generated by selected loci that are located near the modifier locus (so that LD between the modifier and these loci can be maintained over a larger number of generations).

3.2 Multiple recombination modifier loci

As shown by Figure 2.4A, introducing multiple recombination modifier loci – each controlling the local map length in its vicinity – increases the average chromosome map length at equilibrium. However, selective interference has little effect on this increase as increasing the deleterious mutation rate U only slightly shifts the intercept, keeping the

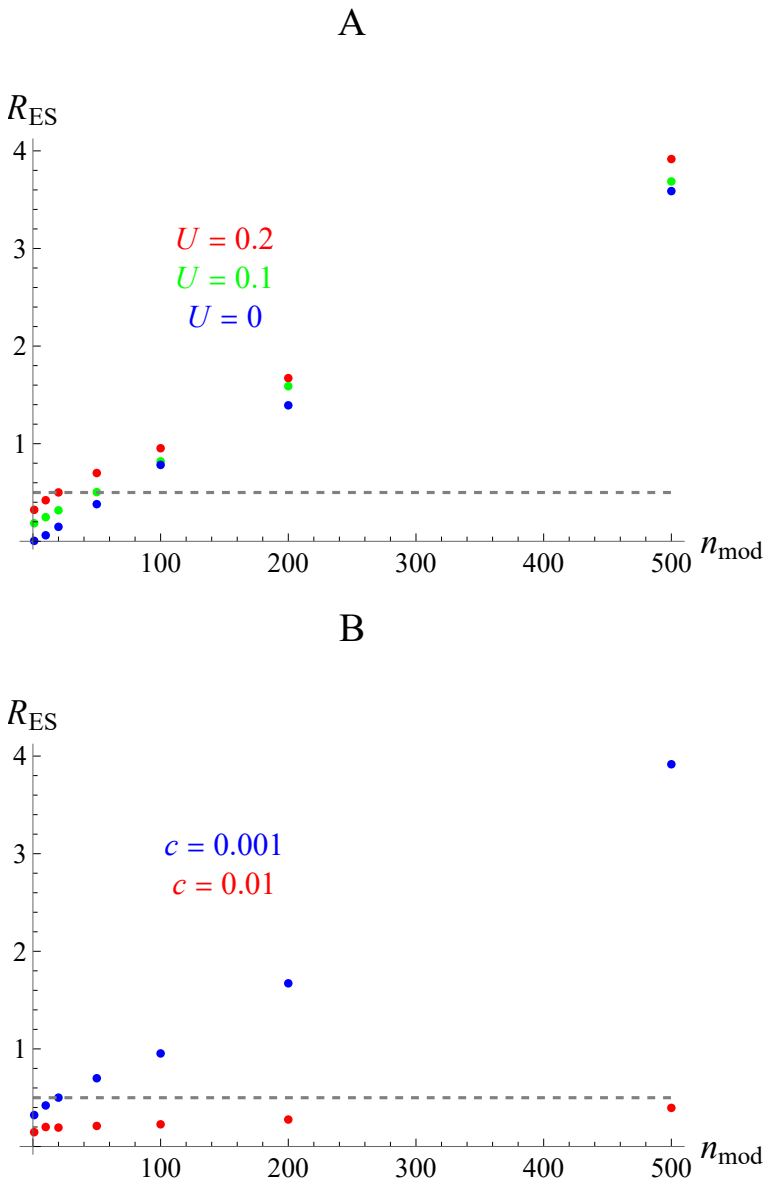


Figure 2.4: Effect of the number of recombination modifier loci n_{mod} on the average chromosome map length at equilibrium R_{ES} for different deleterious mutation rates U (A) and different direct fitness costs of COs c (B). The dashed grey line represents the total map length with 1 obligate CO. Other parameters are: $s = 0.05$; $h = 0.2$, $c = 0.001$; $N = 10,000$

slope of the relationship unchanged. In particular, the increase in map length R_{ES} with the number of modifier loci n_{mod} is seen even in the absence of deleterious mutations ($U = 0$). Moreover, the magnitude of this effect mostly depends on the direct fitness cost per CO c, as the slope of the relationship between R_{ES} and n_{mod} decreases when c increases (Figure 2.4B). This stems from the fact that, as the direct fitness cost of recombination c is proportional to the map length R, it is thus shared between all modifier loci and direct selection against recombination at each modifier locus can become of the same order of magnitude (or lower) as the effect of drift when the number of modifier loci is sufficiently large. This increases the variance of local map length due to random fluctuations of the local R coded by each modifier locus and therefore increases the total map length of the chromosome as R cannot be negative. Despite the dominant effects of mutation and drift at each modifier locus, indirect selection still seems to have an effect as the total map length increases with the deleterious mutation rate U (Figure 2.4A). However, more work is needed to assess whether this increase indeed results from indirect selection, or from the fact that deleterious mutations tend to decrease N_e at linked loci (background selection) and thus amplify the effect of drift. Under the lower value of c explored ($c = 0.001$), 50 modifier loci are sufficient to maintain approximately 1 CO per meiosis per bivalent and several hundred modifier loci maintain high total map lengths ($\approx$ 8 COs per meiosis per bivalent for $n_{\mathrm{mod}} = 500$; Figure 2.4B). However, these results should depend on the values of the mutation rate per modifier locus and population size.

4 Discussion

Those preliminary results show that assuming a heterogeneous recombination rate and/or a heterogeneous mutation rate along chromosomes can substantially increase the number of COs maintained at equilibrium. Those results were found assuming that the obligate CO is restricted to a given region of the chromosome, while the position of each supplementary CO is uniformly distributed. This assumption could be relaxed by letting the position of COs evolve, which could help to better understand the selective

pressures shaping recombination landscapes that are observed in eukaryotes (Haenel et al., 2018; Brazier and Glémin, 2022). Indeed, the distribution of COs along the genome is currently mostly explained by mechanical constraints during meiosis, whereas indirect selection could also play a role in shaping this distribution. In particular, in most species the presence of a CO tends to reduce the probability that another CO occurs nearby – a phenomenon known as CO interference. While interference may have evolved for mechanical reasons (e.g., to reduce the entanglement between homologs), previous models have shown that indirect selection may also have played a role, as interference may reduce the frequency of double COs between selected loci, thus increasing the recombination rate between those loci (Goldstein et al., 1992; Otto and Payseur, 2019). However, these models only assume an interference modifier locus acting on recombination rates between three selected loci, and similarly as for the evolution of recombination our model could help to shed light on the evolution of the position of COs at the scale of a whole chromosome. As indirect selection for recombination mostly stems from the effect of deleterious mutations located in the vicinity of the recombination modifier loci, the evolution of the recombination landscape should strongly depend on the relative positions of modifier loci (affecting the number or position of COs, or interference) and deleterious mutations.

These results also outline the need for a better empirical understanding of the possible effects of recombination modifier loci. Indeed, different evolutionary dynamics may occur depending on whether modifiers tend to have local or global effects on recombination rates. As shown in Roze 2021, assuming multiple modifier locus that collectively affect the total map length of the chromosome may not change the evolutionary stable chromosome map length, compared to the case of a single modifier locus. This is due to the fact that, assuming a uniform distribution of mutations and COs along chromosomes, the strength of indirect selection acting on a mutation that affects the total map length of the chromosome does not depend much on its position along the chromosome. Different results may be obtained when mutations and COs are not uniformly distributed along chromosomes, however. Furthermore, Figure 2.4 shows that when recombination

is controlled by multiple modifier loci with local effects, the number of COs maintained at equilibrium may evolve mostly by mutation and drift (the strength of selection acting on each modifier decreasing as the number of modifiers increases), and may thus reach large values. Those results echo the drift-barrier hypothesis, that had been initially proposed to explain the negative correlation between effective population size and mutation rate (Sung et al., 2012; Lynch et al., 2016). According to this hypothesis, selection may be inefficient at maintaining traits at their optimal value in species with low effective population size N_e, in particular when traits are coded by many loci with small effects. When a trait has a lower bound value, a negative correlation between the trait value and N_e is expected (Lynch, 2020). The drift-barrier hypothesis could possibly partly explain the negative correlation observed between N_e and genome map length over a wide taxonomic scale (Buffalo, 2021), although it could also be caused by lower interference effects among selected loci at larger N_e (and therefore less indirect selection for recombination; Roze 2021). However, the fact that many other traits also correlate with N_e (the mutation rate in particular; Sung et al. 2012) should be taken into account in such broad scale comparisons.

Part III

Measuring the linkage disequilibrium between deleterious mutations from genomic data: possible biases and effect of the mating system

In this part, genomic data from different Angiosperm species are used to estimate the linkage disequilibrium between deleterious mutations, an important component of models on the evolution of recombination in outcrossing and selfing species. In Chapter 3, the linkage disequilibrium between deleterious mutations is estimated using genomic data from a natural population of the outcrossing plant *Capsella grandiflora*. In Chapter 4, this analysis is extended to genomic data from two highly selfing plant species, *Capsella orientalis* and *Arabidopsis thaliana*.

Chapter 3

Linkage disequilibrium between deleterious mutations in outcrossing species

Attached paper : Stetsenko R., Duan T., Mérot C., Glémin S., Roze D. (2023). Possible causes of positive linkage disequilibrium between putatively deleterious variants. Manuscript in preparation.

1 Introduction

Deleterious mutations may be abundant within natural populations (Lynch et al., 1999; Charlesworth, 2015), and are thought to be involved in various evolutionary processes such as the evolution of aging (Flatt and Partridge, 2018), mating systems (Lande and Schemske, 1985; Charlesworth, 2006), sex and recombination (Otto and Lenormand, 2002; Agrawal, 2006a; Otto, 2009). The linkage disequilibrium between deleterious alleles at different loci (measuring their degree of association within genomes) can play an important role in several of these processes, the evolution of sex and recombination in particular (Felsenstein, 1974; Otto and Lenormand, 2002; Otto, 2021). Given two alleles a and b at two polymorphic loci, the linkage disequilibrium (LD) is classically defined as: $D_{ab} = p_{ab} - p_a p_b$, where p_{ab} is the frequency of haplotypes ab in the population, and p_a, p_b the frequencies of alleles a and b at each locus (Lewontin and Kojima, 1960; Weir, 1996). D_{ab} thus equals zero when allele a is not more (or less) associated with allele b than expected based on the frequency of allele b in the population, while D_{ab} is positive (negative) when the frequency of the ab haplotype is higher (lower) than expected based on allele frequencies. If a and b are deleterious alleles, selection is more efficient when LD is positive since the variance in fitness in the population is higher than when LD is negative. In an infinitely large and panmictic population and in the absence of epistasis, LD is expected to be null at equilibrium as it is decreased by a factor $1 - r_{ab}$ per generation, where r_{ab} is the recombination rate between the two loci (Lewontin and Kojima, 1960). Epistasis between selected loci generates LD of the same sign as the sign of epistasis, when measured on a multiplicative scale (Felsenstein, 1965): positive epistasis means that the deleterious alleles a and b tend to compensate each other when combined, leading to a higher frequency of the ab haplotype than expected based on the frequencies of a and b ($D_{ab} > 0$), while negative epistasis means that the deleterious effects of a and b tend to reinforce each other when combined, leading to a lower frequency of the ab haplotype ($D_{ab} < 0$). In partially inbred populations, positive LD is expected even in the absence of epistasis, due to the coexistence within the same population of

lineages with different levels of inbreeding: relatively more inbred lineages from which deleterious alleles have been purged more efficiently, and relatively less inbred lineages that are more loaded with deleterious alleles (Roze and Lenormand, 2005; Kamran-Disfani and Agrawal, 2014; Stetsenko and Roze, 2022). Furthermore, random drift tends to generate negative LD between deleterious alleles in haploid populations, in the absence of epistasis (selective interference or the Hill-Robertson effect: Hill and Robertson, 1966; Felsenstein, 1974; Otto, 2021). The same effect occurs in diploids when deleterious alleles have additive effects (no dominance); however, dominance introduces additional effects that generate positive LD when deleterious alleles are sufficiently recessive ($h < 0.25$, Roze, 2021).

The squared linkage disequilibrium ${D_{ab}}^2$ — or the scaled quantity $r^2 = {D_{ab}}^2 / (p_a q_a p_b q_b)$, where $q_a = 1 - p_a$, $q_b = 1 - p_b$ — is often measured in genomic studies in order to assess the range of genetic distances over which genetic associations (regardless of their sign) are maintained, informative on a broad range of evolutionary processes (e.g., the effect of recombination, drift, selection, population structure, the mating system; reviewed by Slatkin, 2008). In a panmictic population, the expected squared LD (denoted $\left\langle {D_{ab}}^2 \right\rangle$ hereafter) between neutral sites is approximately $(10 + \rho) \langle p_a q_a p_b q_b \rangle / [(2 + \rho)(11 + \rho)]$ where $\rho = 4 N_\mathrm{e} r_{ab}$ and N_e is the effective population size (Ohta and Kimura, 1971; Hill and Weir, 1988; McVean, 2002), which tends to $\langle p_a q_a p_b q_b \rangle / \rho$ when ρ is large. For example, genetic associations are typically maintained up to approximately 100 kb in humans (Reich et al., 2002) and 1 kb in *Drosophila melanogaster* (MacKay et al., 2012). Genetic associations may be maintained over longer distances in inbred populations, such as the highly selfing nematode *Caenorhabditis elegans* (Cutter, 2006; Wright et al., 2008). This is due to the fact that homozygosity caused by inbreeding reduces the effect of recombination by a factor $1 - F$, and reduces effective population size by a factor $1/(1 + F)$, where F is the inbreeding coefficient (Golding and Strobeck, 1980; Nordborg and Donnelly, 1997; Nordborg, 2000a). Selective sweeps can also increase the variance in LD at linked neutral loci (McVean, 2007).

More recently, a number of studies have tried to assess the overall sign and magni-

tude of LD (D_{ab}) between putatively deleterious alleles segregating within populations. These studies typically contrast different types of mutations: synonymous, missense, and loss-of-function (LoF) mutations (Sohail et al., 2017; Garcia and Lohmueller, 2021; Sandler et al., 2021; Ragsdale, 2022; Stolyarova et al., 2022). Sohail et al. (2017) estimated the genome-wide LD (corresponding to the sum of LD among all pairs of mutations) for each class of mutation from the under- or overdispersion of the distribution of the number of mutations per genome (given that this distribution should be approximately Poisson in the absence of LD), in *Drosophila melanogaster* and humans. They found that genome-wide LD is positive between synonymous (rare) mutations, less positive between missense mutations and negative between LoF (the more deleterious) mutations, which they interpreted as a possible effect of negative epistasis between deleterious mutations. The *D. melanogaster* dataset was later reanalyzed by Sandler et al. (2021), who also analyzed an additional dataset from a single population of the outcrossing plant *Capsella grandiflora*. They found genome-wide positive LD for all classes of (rare) mutations except LoF mutations in *D. melanogaster* which displayed negative LD; however, the latter was found to be non-significant. Additionally, they computed LD for different classes of distances (in bp) between pairs of loci, showing that at short distances, LD between LoF mutations is more positive than between synonymous mutations. Sandler et al. (2021) hypothesized that this pattern may be explained by positive intragenic epistasis, additional LoF mutations having little effect once a gene is already disrupted by a first LoF mutation. Garcia and Lohmueller (2021) found that non-synonymous rare variants exhibit more negative LD than synonymous rare variants in human populations, and interpreted this result as the consequence of selective interference (Hill-Robertson effect) among deleterious alleles. However, the measure of LD used (D', which is scaled differently by products of allele frequencies for positive and negative LD) makes it difficult to compare their results with those obtained in previous studies. Ragsdale (2022) used genomic data from 15 human populations to measure LD within protein-coding genes, and found positive LD between derived synonymous and missense mutations, but slightly negative LD between LoF mutations. LD among missense variants was found to

be more positive than LD between synonymous variants in the case of variants present in the same conserved gene domain, while the opposite pattern was observed in the case of variants that are not present within the same conserved domain (controlling for the distance in bp among variants). This was interpreted as a possible effect of positive epistasis (compensatory effects) among missense mutations segregating within the same domain. Finally, Stolyarova et al. (2022) found higher positive LD among nonsynonymous than among synonymous variants within populations of the hypervariable fungus *Schizophyllum commune*, which was again interpreted as evidence for positive epistasis among segregating deleterious mutations.

Importantly, in many of these studies LD was measured among rare variants, since deleterious alleles are expected to stay at low frequency within populations. However, a recent theoretical study by Good (2022) showed that such a conditioning on allele frequencies generates a bias towards positive LD, which is stronger in the case of neutral alleles than in the case of deleterious alleles. Additionally, Sandler et al. (2021) used a simulation model to show that past admixture may also generate positive LD among rare neutral variants, the effect of admixture being much reduced in the case of deleterious variants. These results may thus explain the positive LD observed among synonymous variants, and show that observing less positive LD among missense than among synonymous rare variants does not necessarily imply the existence of a selective force generating negative LD among deleterious alleles within populations (Sandler et al., 2021). However, they do not explain situations in which LD among closely linked missense variants is more positive than among synonymous variants (Sandler et al., 2021; Ragsdale, 2022; Stolyarova et al., 2022). Another possible source of spurious LD in genomic data analyses may stem from structural variants segregating within populations. In particular, a recent study by Jaegle et al. (2023) showed that polymorphic duplications may be the source of a high level of pseudo-heterozygosity in the raw data of the 1001 Arabidopsis Genome Project. While such pseudo-heterozygosity may be relatively easy to detect in the case of highly selfing species such as *A. thaliana*, it may be more difficult to detect in outcrossing species, and may generate false signals of LD among variants present in

the duplicated sequence (when the duplication is not present in the reference genome). Because structural variation may be more abundant within populations than previously realized (e.g. in *A. thaliana* Jiao, 2020 but for a review see Mérot et al. 2020), it seems important to take this possibility into account.

In this article, we further explore possible sources of bias towards positive LD in genomic data. Using an analytical argument and coalescent simulations, we show that positive LD is expected (on average) among mutations present at similar frequencies within a population or a sample (not only rare mutations). By re-analyzing the sequence data from *Capsella grandiflora* used by Sandler et al. (2021), we show that LD among derived neutral mutations disappears in the absence of any conditioning on allele frequency. While positive LD is still observed among putatively deleterious variants, we show that at least part of this positive LD may be explained by polymorphic structural variants segregating within the population. We then discuss other possible sources of positive LD among selected mutations.

2 Methods

2.1 Coalescent simulations

The effect of computing LD and other two-locus moments on subsets of neutral variants present in a given frequency class in a sample of n sequences was checked using coalescent simulations. We simulated the ancestral recombination graph (ARG) using Hudson's algorithm (Hein et al., 2004, p. 139-144) with parameters $\rho = 4Nr$ and $\theta = 4Nu$, where N is population size, r the recombination rate and u the mutation rate ($\rho = 40$ and $\theta = 100$ were used for all results presented here, while sample size was set to $n = 180$). Pairs of segregating sites were classified into 100 classes of genetic distance, and the linkage disequilibrium D_{ab} among derived alleles (and other two-locus moments) was averaged over 20 batches of 5,000 coalescent trees (10^5 trees in total). 95% confidence intervals of the ratio $\langle D_{ab} \rangle \,/\, \langle p_a q_a p_b q_b \rangle$ were obtained using the bootstrap method for each distance class, by resampling 1000 times over batch averages. In order to assess the

effect of ancestral state misspecification, we also included a parameter μ corresponding to the probability that a variant is misspecified.

2.2 Population genomic data

We retrieved the whole-genome sequencing data from 182 individuals of the obligate outcrossing plant *Capsella grandiflora* from Northern Greece (Josephs et al., 2015) that was used by Sandler et al. (2021). This dataset is suitable for testing predictions on LD patterns from classical population genetics model as it is composed of a large number of individuals sampled from a single population with low genetic structure of an obligately outcrossing species, limiting the effects of population structure and inbreeding. In order to compare LD patterns between species with different mating systems (next chapter), we also retrieved whole-genome sequencing data from 33 individuals of the highly selfing *Capsella orientalis* from Central Asia, provided by Sylvain Glémin (Ågren et al., 2014; Huang et al., 2018; Kryvokhyzha et al., 2019). The list of individuals is available in Table S2.1 (*C. grandiflora*) and Table S3.2 (*C. orientalis*). In order to polarize mutations we used a whole genome assembly of the related Brassicaceae *Neslia paniculata* (Slotte et al., 2013). Raw reads from both species were trimmed using Trimmomatic v0.39 (Bolger et al., 2014) and mapped on a new (not yet published) reference genome of *Capsella rubella*, Cr145 (Tianlin Duan, personal communication) using bwa v0.7.17 (Li and Durbin, 2010). The Cr145 genome has a total length of 158,571,049 bp divided into 129 contigs with an N50 of 17,600,677 bp. The mean coverage per individual was 39.99X for *C. grandiflora* and 28.82X for *C. orientalis*. Duplicate reads where marked using the *MarkDuplicates* option of Picard Tools v2.23.5 (broadinstitute.github.io/picard/) and removed from subsequent analysis. We obtained a VCF file containing 20,745,523 called sites.

2.3 SNP calling

Single nucleotide polymorphisms (SNPs) were called using GATK and following the GATK Best Practices (Van der Auwera et al., 2013). However, adjustments were

made to the filtering criteria in the *HaplotyeCaller* option based on the distribution of quality scores with filters set to: *MQ<50, SOR>3, QD<2, FS>60, MQRankSum<-5, ReadPosRankSum<-8, ReadPosRank-Sum>8* (Figure S2.5). After site filtration 13,602,125 called sites remained in the VCF file. Then, the VCF file was split between *C. grandiflora* and *C. orientalis* individuals. Data from *C. orientalis* individuals will be analysed in the next chapter. In addition, we filtered out sites with extreme mean coverage (>71.65X; Figure S2.6) and genotypes with low coverage (<5X). This was done to filter out the lowest confidence genotypes and regions with a high number of reads due to repetitive sequences. Monomorphic sites, sites with more than 2 alleles and sites with a high number of missing data (>50%) were also filtered out to obtain 8,964,707 biallelic SNPs. Based on a PCA with the filtered independent SNPs (using the function *indep-pairwise* of Plink with parameters: window size = 50 kb, step size = 100 variants and r^2 threshold = 0.1; Purcell et al. 2007), we excluded from LD computation 11 genetically divergent individuals (Figure S2.7). Most of these divergent individuals also present outlier proportions of missing data (Figure S2.8). In order to be able polarize alleles based on their ancestral or derived state, we only kept sites for which information on the sequence of the outgroup (*Neslia paniculata*) was available, yielding 6,935,951 biallelic SNPs.

2.4 SIFT annotation

The SIFT4G algorithm (Vaser et al., 2016) was used to catogorize mutations based on their potential fitness effect. This algorithm uses multiple protein alignments across a database of proteins from a large diversity of taxa. If the site is conserved across the alignment and the query nucleotide is present in a low number of sequences it is inferred as deleterious whereas if the site is not conserved across the alignment or the query nucleotide is present in a higher proportion of sequences it is inferred as neutral. A score from 0 for the most deleterious variants to 1 for the most neutral variants is attributed by SIFT to each possible nucleotide at each site. We built a SIFT library for the Cr145 *C. rubella* reference genome using the UniRef100 protein database (uniprot.org/uniref/). By default, SIFT4G categorizes variants with a SIFT score ≤ 0.05 as deleterious and

> 0.05 as tolerated. In order to distinguish potentially mildly deleterious mutations from neutral mutations, we split the latter category into "mildly deleterious mutations" ($0.05 <$ SIFT score < 1) and "neutral mutations" (SIFT score $= 1$), while mutations with a SIFT score ≤ 0.05 were categorized as "deleterious". In order to be able to polarize LD among sites at which alleles with SIFT score < 1 are segregating based on their inferred deleterious effect, sites at which more than one allele with SIFT score < 1 are segregating were not taken into account.

2.5 Detecting potential structural variants

An R script (available at github.com/RomanStet/LD_deleterious_mutations) was used to detect potential polymorphic duplications, that may generate pseudo-heterozygosity patterns and give rise to spurious LD. Indeed, if a duplication is present in a subset of individuals only and absent from the reference genome, reads from the duplicated sequence will be mapped onto the original sequence, and mutations present in the duplicated sequence will thus give rise to spurious heterozygosity at the corresponding sites in the original sequence (Jaegle et al., 2023; Figure S2.12). Such duplications may thus be detected from neighboring sites presenting similar patterns of heterozygosity (*i.e.*, the same individuals are heterozygous). Mutations that are fixed in the duplicated sequence but absent from the original sequence will thus always appear in the heterozygous state (Figure S2.12, case 1), while mutations that are fixed in the original sequence but absent from the duplicated sequence will appear in the heterozygous state in the same individuals as case 1 mutations, while they will appear in the homozygous state in all other individuals (Figure S2.12, case 2). Our R script thus uses data from the VCF file to detect blocks of genome carrying sites that present similar patterns of heterozygosity, and at which only two genotypes are present in the sample: heterozygous and homozygous for the ancestral allele (case 1), or heterozygous and homozygous for the the derived allele (case 2). In order to take into account mutations that may not be fixed within the duplicated or original sequence (and genotyping errors), two sites falling either into case 1 or case 2 were considered as potentially indicative of a polymorphic duplication when

they shared at least 70% of their heterozygous individuals. Sites are examined sequentially along the chromosome until no case 1 or case 2 pair of sites sharing a common pattern of heterozygosity is found over a distance of 2 kb. Among the large number of genome blocks detected, only a minority may correspond to polymorphic duplications (in particular, the majority of detected blocks only contain 2 or 3 sites sharing a common pattern of heterozygosity). Furthermore, our algorithm should detect haplotype blocks present at low frequency (since these blocks will tend to carry mutations that are mostly present in the heterozygous state) that do not necessarily correspond to potential duplications. We thus only retained blocks over which a minimal number of mutations share a common pattern of heterozygosity, and over which the coverage (over the detected block) in heterozygous individuals that potentially carry the duplication (scaled by the mean coverage of the individual) is significantly higher than the scaled coverage in the other individuals.

Two standard structural variant (SV) detection programs: Smoove and Delly (Rausch et al., 2012) were also used to detect potential duplications in the *C. grandiflora* dataset. Smoove is based on the LUMPY algorithm (Layer et al., 2014), and like Delly, it uses information on paired-ends, split-reads and read-depth to detect SV. We ran both programs on mapped reads (see section on SNP calling above) of each individual of *C. grandiflora* and retained only SV calls labeled as duplication and not tagged as 'low quality' by Delly. We considered that a block detected by the R script was validated by Smoove or Delly if, in at least half of the heterozygous individuals in the block (potentially carrying the duplication), a duplication inferred by Smoove or Delly overlapped with the block in those same individuals.

2.6 Computing LD

Since SNP data are unphased, the two double heterozygous genotypes Ab/aB and ab/AB cannot be distinguished. Therefore, we measured LD between alleles a and b by the composite LD measure Δ_{ab} which is the sum of the association between a and b on the same chromosome (D_{ab}) and the association between a and b on different chromosomes

($D_{a,b}$) in the notation of Kirkpatrick et al., 2002). This composite LD can be computed as:

$$\Delta_{ab} = 2f_{ab/ab} + f_{aB/ab} + f_{Ab/ab} + \frac{1}{2}\left(f_{Ab/aB} + f_{ab/AB}\right) - 2p_a p_b, \tag{3.1}$$

where f_X is the frequency of genotype X in the sample, and p_a, p_b the frequencies of a and b (e.g., p. 126 in Weir, 1996). An expression for Δ_{ab} in terms of indicative variables and allele frequencies using the formalism from Barton and Turelli (1991) and Kirkpatrick et al. (2002) can be found in Appendix S2. In the case of tightly linked loci, $D_{a,b} \approx F\, D_{ab}$, where F is the inbreeding coefficient (Nordborg, 1997; Roze, 2016), so that $\Delta_{ab} \approx D_{ab}\,(1 + F)$. Therefore, Δ_{ab} should be equal to D_{ab} under random mating, and is increased by inbreeding.

When two alleles categorized as neutral by SIFT were segregating at a given locus, they were polarized based on derived state based on the sequence of the outgroup *Neslia paniculata* (*i.e.*, a, b correspond to the derived alleles). At loci were a deleterious (or mildly deleterious) allele and a neutral allele are segregating, we checked that polarizing by SIFT category alone (a, b correspond to the putatively deleterious alleles) or by a combination of SIFT category and derived state (a, b correspond to the putatively deleterious and derived alleles) did not affect the results (see Results section): indeed, most mutations with a SIFT score < 1 are derived. The mean value of Δ_{ab} was computed for every pair of sites at different physical distances up to 1.5 kb (each distance class being 25 bp wide), discarding pairs of sites for which the information on genotype was not available (at both sites) in less than half of the total number of individuals. As for the LD estimator σ_D^2 of Ohta and Kimura (1969), the average LD for a given distance class $\langle\Delta_{ab}\rangle$ was scaled by the average product of allele diversities at both loci (for the same distance class) $\langle p_a q_a p_b q_b\rangle$, where $\langle X\rangle$ denotes the average of quantity X over pairs of sites. 95% confidence intervals were computed using the bootstrap method, by resampling each set of pairs of sites 1,000 times.

3 Results

3.1 Effect of conditioning on frequency on LD between neutral variants: theoretical results

Good (2022) showed that in a finite, panmictic population, the linkage disequilibrium between rare neutral mutations is positive on average. This result can be extended to the case of neutral variants segregating at similar frequencies in a population (not only rare variants). This is a consequence of the variance in LD generated by drift, and may be understood as follows. Due to drift, the linkage disequilibrium between two alleles a and b is sometimes positive, sometimes negative. When it is positive, a and b tend to be found on the same haplotypes (more often than by chance), while when it is negative they tend to be found on different haplotypes (more often than by chance). In the first situation, the frequencies of a and b tend to be more similar than in the second: for example in the extreme case of positive D_{ab} where only ab and AB haplotypes are present in the population, a and b have exactly the same frequency, while in the extreme case of negative D_{ab} where only Ab and aB haplotypes are present, a will be rare if b is frequent (and vice versa). As a consequence, averaging over the stochastic process, two neutral alleles that are present at similar frequencies in a population (both rare or both frequent) have more chances to be in positive than in negative LD, while rare alleles have more chances to be in negative than in positive LD with frequent alleles. One can note that this effect is captured by the moment $\langle (1-2p_a)(1-2p_b) D_{ab} \rangle$ that was computed by Hill and Weir (1988), shown to be generated by the variance in LD $\left\langle {D_{ab}}^2 \right\rangle$ (equation 1 in Hill and Weir, 1988) and approximately equal to $8 \langle p_a q_a p_b q_b \rangle / [(2+\rho)(11+\rho)]$ at equilibrium (where $\rho = 4N_e r_{ab}$). Indeed, this moment may be seen as a covariance between the quantities $(1-2p_a)(1-2p_b)$ and D_{ab}, and the fact that it is positive indicates that D_{ab} tends to be positive when either p_a, $p_b < 0.5$ or p_a, $p_b > 0.5$ (that is, $(1-2p_a)(1-2p_b) > 0$), while D_{ab} tends to be negative when either $p_a < 0.5$, $p_b > 0.5$ or $p_a > 0.5$, $p_b < 0.5$ (that is, $(1-2p_a)(1-2p_b) < 0$).

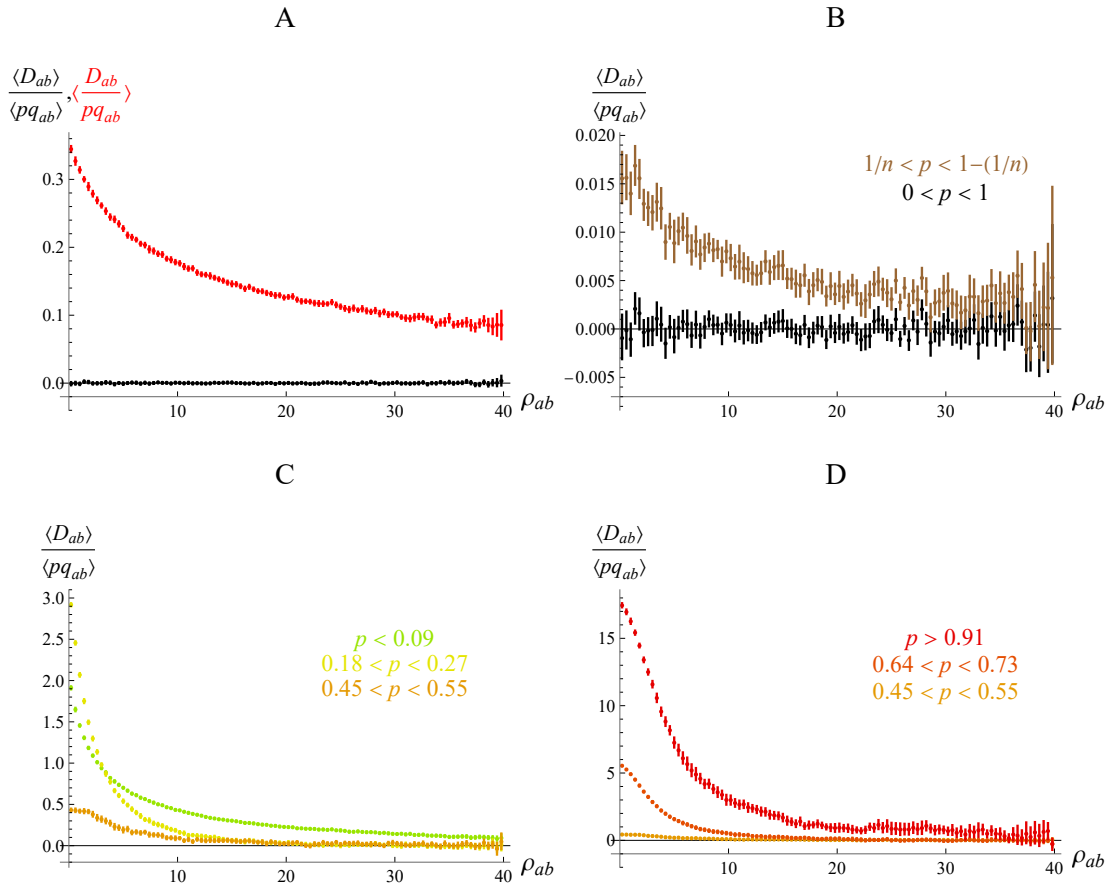


Figure 3.1: Average LD $\langle D_{ab}\rangle$ between derived neutral mutations in the coalescent simulations scaled by the average product of genetic diversities $\langle pq_{ab}\rangle = \langle p_a q_a p_b q_b\rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$. A: Average of D_{ab} (scaled by the average of pq_{ab}, black) and average of D_{ab}/pq_{ab} (red) over all segregating sites for each distance class (no conditioning on frequency). B: Average of D_{ab} (scaled by the average of pq_{ab}) over all segregating sites (black) or excluding singletons (brown). C and D: Average of D_{ab} (scaled by the average of pq_{ab}) over pairs of loci at which the derived allele segregates in a given frequency range. Errors bars correspond to 95% confidence intervals (see Materials & Methods). Parameters values: $\theta = 4N_e u = 100$ (mutation rate); $n = 180$ (sample size).

These predictions are confirmed by our coalescent simulation results (with $\rho = 40$, $\theta = 100$ and $n = 180$). As shown by Figure 3.1, the average LD among derived alleles is zero when computed over all pairs of segregating alleles in the sample (no conditioning on frequency: Figure 3.1A, black dots), while it is positive when computed over pairs of loci at which derived alleles are present at similar frequencies (both rare or both frequent, Figures 3.1C – 3.1D). One can note that the average LD between frequent derived alleles is higher than between rare derived alleles (compare Figures 3.1C and 3.1D). While this is probably caused by the asymmetry of the site frequency spectrum (derived alleles stay rare on average), analytical work would be needed in order to better understand this effect. As can be seen from Figure 3.1B, removing singletons from the analysis already generates a small bias towards positive LD; although this bias is rather small when $n = 180$, it may become more important for smaller sample sizes. Finally, Figure 3.1A shows that contrarily to $\langle D_{ab} \rangle$, the expected value of $D_{ab} / (p_a q_a p_b q_b)$ is strongly positive (when computed over all segregating alleles). By contrast, the expected value of $r = D_{ab} / \sqrt{p_a q_a p_b q_b}$ remains close to zero when computed over all segregating alleles (but again becomes positive when computed over alleles present at similar frequencies, see Figure S2.1).

As shown by Figure 3.2, introducing a given rate μ of misspecification of ancestral/derived state does not affect much the previous results, LD staying zero in the absence of conditioning on allele frequencies, and positive when measured only among 'derived' alleles present at similar frequencies. Misspecification decreases LD among frequent 'derived' alleles (Figure 3.2D), since this class now contains LD measures between rare (and truly) derived alleles. Figure 3.3 shows that polarizing LD based on minor allele frequency (MAF) — that is, computing LD among alleles whose frequency is below a given threshold — generates the same bias towards positive values, which is more important when the threshold frequency is lower. Finally, Figures S2.2 – S2.4 show that while the expected squared LD $\left\langle {D_{ab}}^2 \right\rangle$ is well predicted by $(10 + \rho) \langle p_a q_a p_b q_b \rangle / [(2 + \rho) (11 + \rho)]$ in the absence of any conditioning on allele frequency (Ohta and Kimura, 1971; Hill and Weir, 1988; McVean, 2002), it is again affected

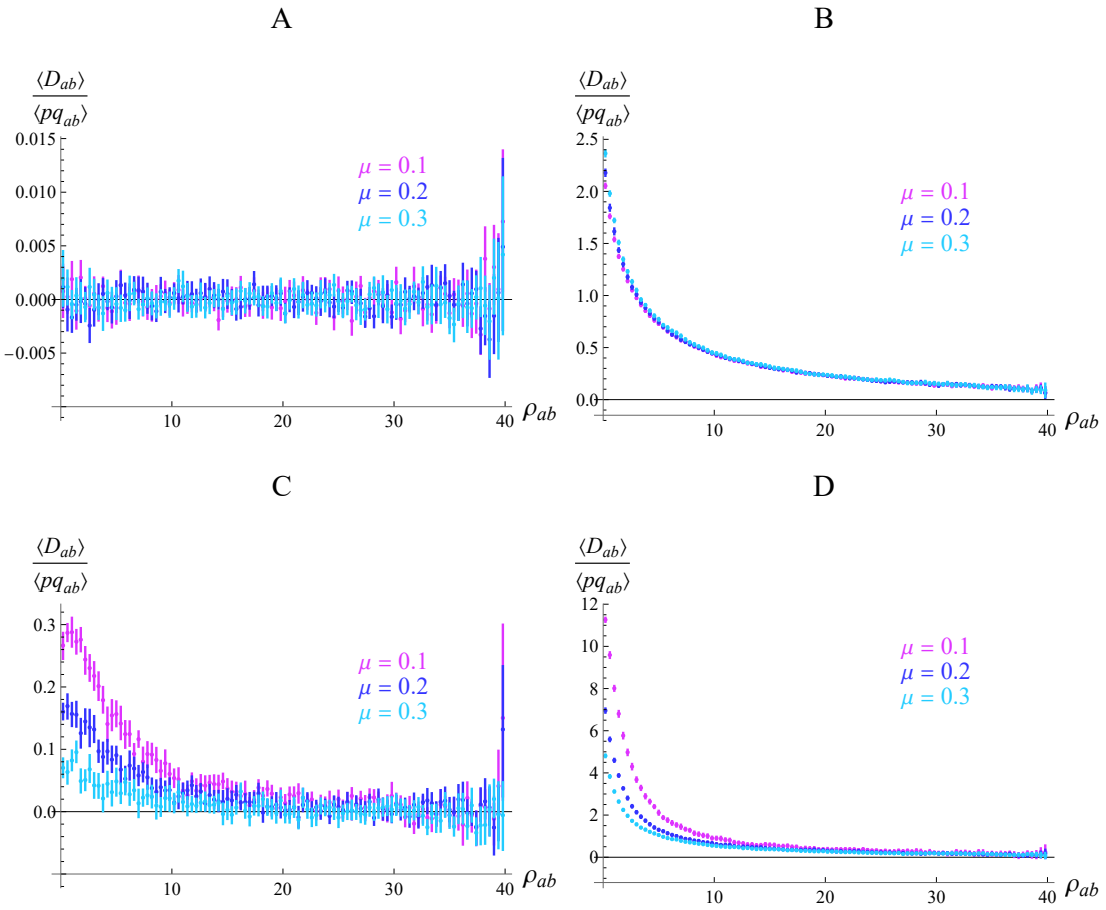


Figure 3.2: Average LD $\langle D_{ab} \rangle$ between derived neutral mutations in the coalescent simulations scaled by the average product of genetic diversities $\langle pq_{ab} \rangle = \langle p_a q_a p_b q_b \rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$, and for different values of the rate of misspecification μ of the ancestral/derived state of alleles. A: Averages over all segregating sites for each distance class (no conditioning on frequency); B: only the lowest frequency class ($p < 0.09$, green points in Figure 3.1C); C: only the middle frequency class ($0.45 < p < 0.55$, orange points in Figure 3.1C and 3.1D); D: only the highest frequency class ($p > 0.91$, red points in Figure 3.1D). Parameter values are as in Figure 3.1.

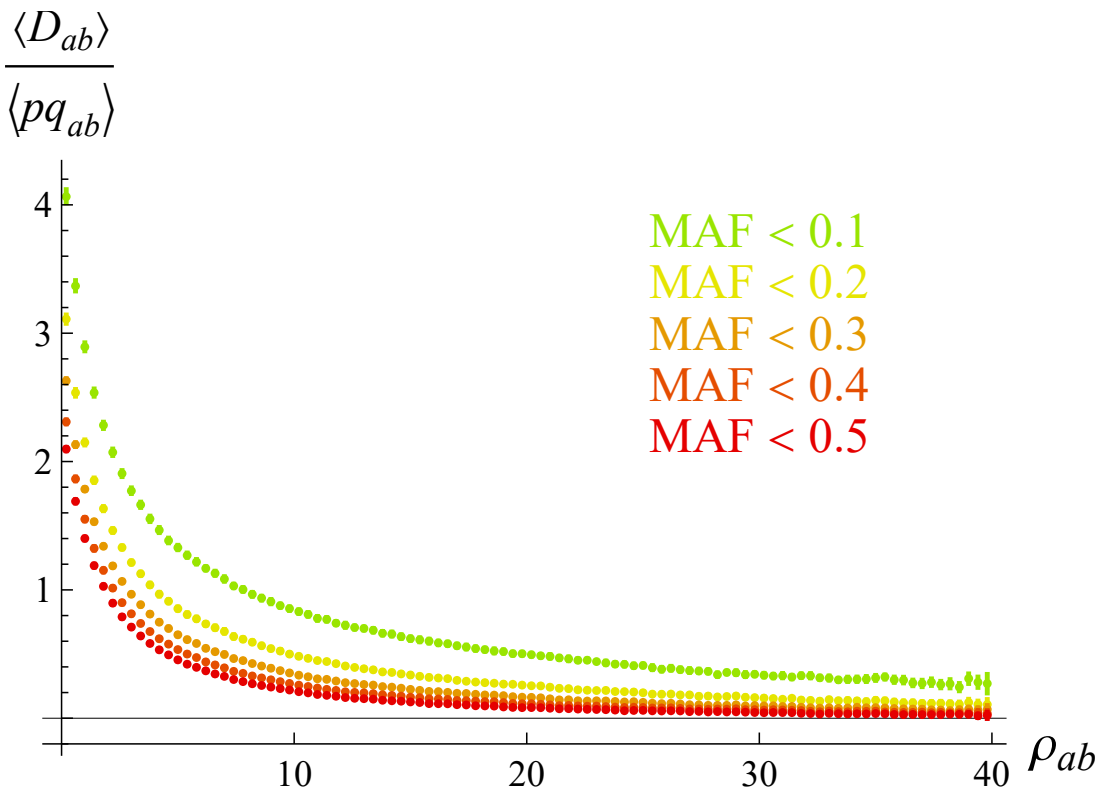


Figure 3.3: Average LD $\langle D_{ab} \rangle$ between derived neutral mutations in the coalescent simulations scaled by the average product of genetic diversities $\langle pq_{ab} \rangle = \langle p_a q_a p_b q_b \rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$. LD is polarized based on the minor allele frequency (MAF) and averaged over pairs of sites with different MAF thresholds. Parameter values are as in Figure 3.1.

by such a conditioning, and is reduced when LD is polarized based on minor allele frequencies (Figure S2.4).

3.2 No LD between neutral mutations, but positive LD between deleterious mutations in the *C. grandiflora* dataset

Alleles at polymorphic loci in the *C. grandiflora* dataset were categorized as 'neutral', 'mildly deleterious' or 'deleterious' based on their SIFT score (Vaser et al., 2016), as explained in the Materials and Methods. The site frequency spectra (SFS) of the different types of derived mutations (deleterious, mildly deleterious or neutral) reflected their SIFT score category as deleterious mutations had an SFS shifted towards rarer frequen-

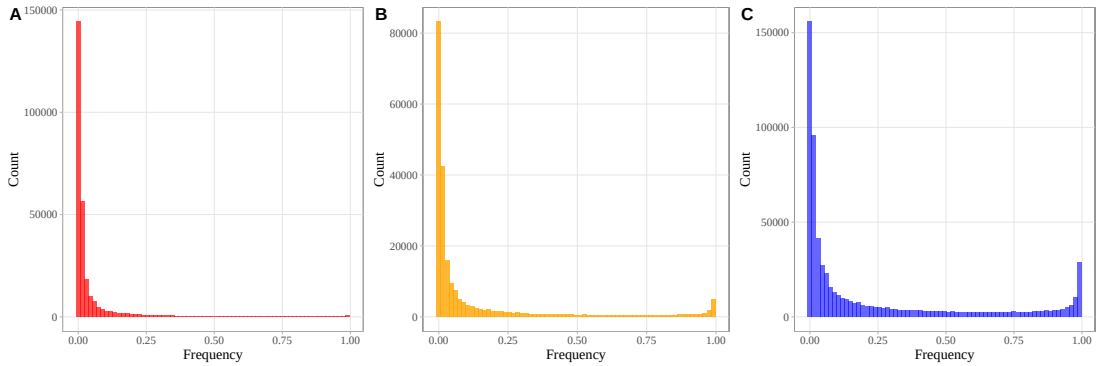


Figure 3.4: Site frequency spectrum for deleterious (A), mildly deleterious (B) and neutral (C) mutations, inferred as derived from the comparison with the outgroup *Neslia paniculata*. Sites at which more than one deleterious or mildly deleterious allele segregates are excluded from the analysis.

cies compared to mildly deleterious mutations and furthermore compared to neutral mutations (Figure 3.4). Including polymorphic loci at which the ancestral allele (inferred from the sequence of the outgroup *Neslia paniculata*) is categorized as deleterious or mildly deleterious (while the derived allele is categorized as neutral) increased the number of high-frequency mildly deleterious mutations (Figure S2.9). Indeed, a proportion of high-frequency mildly deleterious mutations are inferred as ancestral, either because their are actually ancestral to both *N. paniculata* and *C. grandiflora* or because they appeared in the lineage of *N. paniculata* after its split with the lineage of *C. grandiflora*.

As explained in the Materials and Methods, LD between pairs of loci was computed using the composite LD measure Δ_{ab}, that should be equivalent to D_{ab} under random mating. The average LD between derived neutral mutations stays close to zero for all genetic distances, in agreement with our simulation results (Figure 3.5C). By contrast, the average LD between derived deleterious alleles and between derived mildly deleterious alleles is positive and increases with the degree of linkage among loci (Figure 3.5A and 3.5B). Using only the SIFT score to polarize LD among deleterious and mildly deleterious mutations (that is, without taking into account the ancestral/derived state of alleles) has nearly no effect on the results (Figure S2.11). The variance in LD was found to be lower between deleterious mutations compared to between neutral mutations, at least at

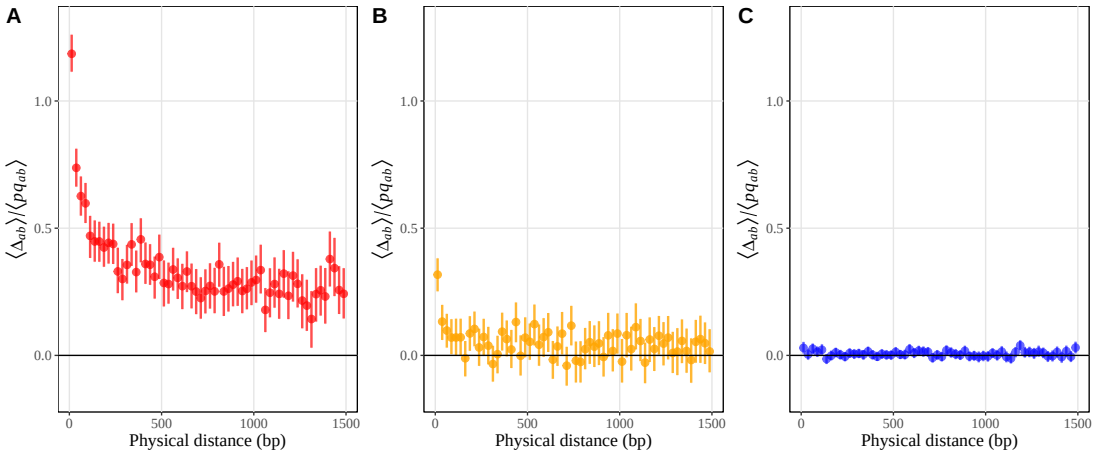


Figure 3.5: Average composite linkage disequilibrium Δ_{ab} (scaled by the average product of allele frequencies) between deleterious (A), mildly deleterious (B) and neutral (C) derived mutations according to physical distance.

very short distances ($<$ 200 kb) whereas mildly deleterious mutations show the highest variance in LD at short distance (Figure S2.10). The lowest variance in LD between deleterious mutations is consistent with predictions from models of purifying selection in finite populations (e.g., Roze, 2021). The highest variance in LD between mildly deleterious mutations may be due to the fact that these mutations have a wider range of SIFT scores, some pairs of mutations having a high, positive LD as observed for deleterious mutations while others have a low LD as observed for neutral mutations.

As shown in Roze (2021), the quantity $\langle p_b \Delta_{ab} \rangle$ should be negative when allele b is neutral while allele a is deleterious: indeed, when allele b tends to be associated with the deleterious allele a (that is, when $D_{ab} > 0$), the frequency of b should decrease. By contrast, $\langle p_a \Delta_{ab} \rangle$ should be zero on average, since being associated with a particular neutral allele should not affect the frequency of a deleterious allele. Figure 3.6 shows that these predictions are confirmed, a negative correlation between the frequency of neutral alleles and their LD with deleterious alleles being observed at short genetic distances.

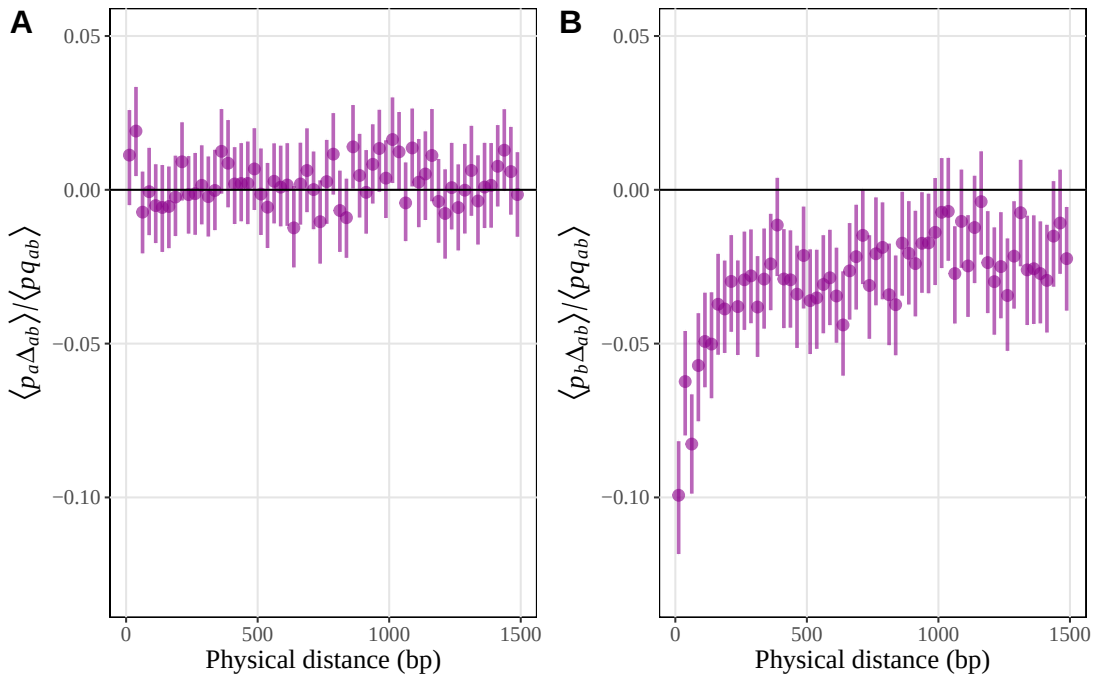


Figure 3.6: Averages of $p_a\Delta_{ab}$ (A) and $p_b\Delta_{ab}$ (B) (scaled by the average product of allele frequencies) over all pairs of derived deleterious alleles a and derived neutral alleles b at a given physical distance.

3.3 Structural variants may contribute to positive LD between deleterious mutations

Structural variants segregating within populations is a possible cause of positive LD among mutations. In particular, Jaegle et al. (2023) showed that polymorphic duplications present in the sequenced individuals but not in the reference genome tend to be frequent in the 1001 Arabidopsis Genomes dataset, generating pseudo-heterozygosity due to the fact that the duplicated sequence is incorrectly mapped onto the original sequence (see Figure S2.12). Mutations present in the duplicated sequence (but not in the original one) would thus appear to be in positive LD due to this incorrect mapping. This effect may be stronger for deleterious than for neutral mutations, because deleterious mutations are less abundant within genomes, and a significant proportion of mutations categorized as deleterious may correspond to mutations that are present in such duplications (while their effect may in fact not be deleterious, since a functional gene copy is maintained elsewhere in the genome). We developed an *R* program to identify potential haplotype blocks that may correspond to such polymorphic duplications (by detecting shared patterns of heterozygosity among closely linked SNPs, see Materials and Methods), that we ran on the SNPs called in *C. grandiflora* individuals. Although the program identified a large number of haplotype blocks (108,896 in total), the majority of them comprised only 2 or 3 SNPs with similar heterozygosity patterns, and were present in a few individuals only. These are probably not caused by structural variants: indeed, the presence of haplotype blocks is expected from the structure of the ancestral recombination graph (Shipilina et al., 2022), and haplotype blocks may be present only in the heterozygous form when they are rare in the population. In the following, we restrict our analyses to blocks comprising more than 3 SNPs with shared heterozygosity patterns, and presenting a significantly increased degree of coverage in heterozygous individuals (as expected when heterozygosity corresponds to pseudo-heterozygosity caused by mapping errors).

We detected 1,023 blocks in which more than 3 SNPs present similar heterozygos-

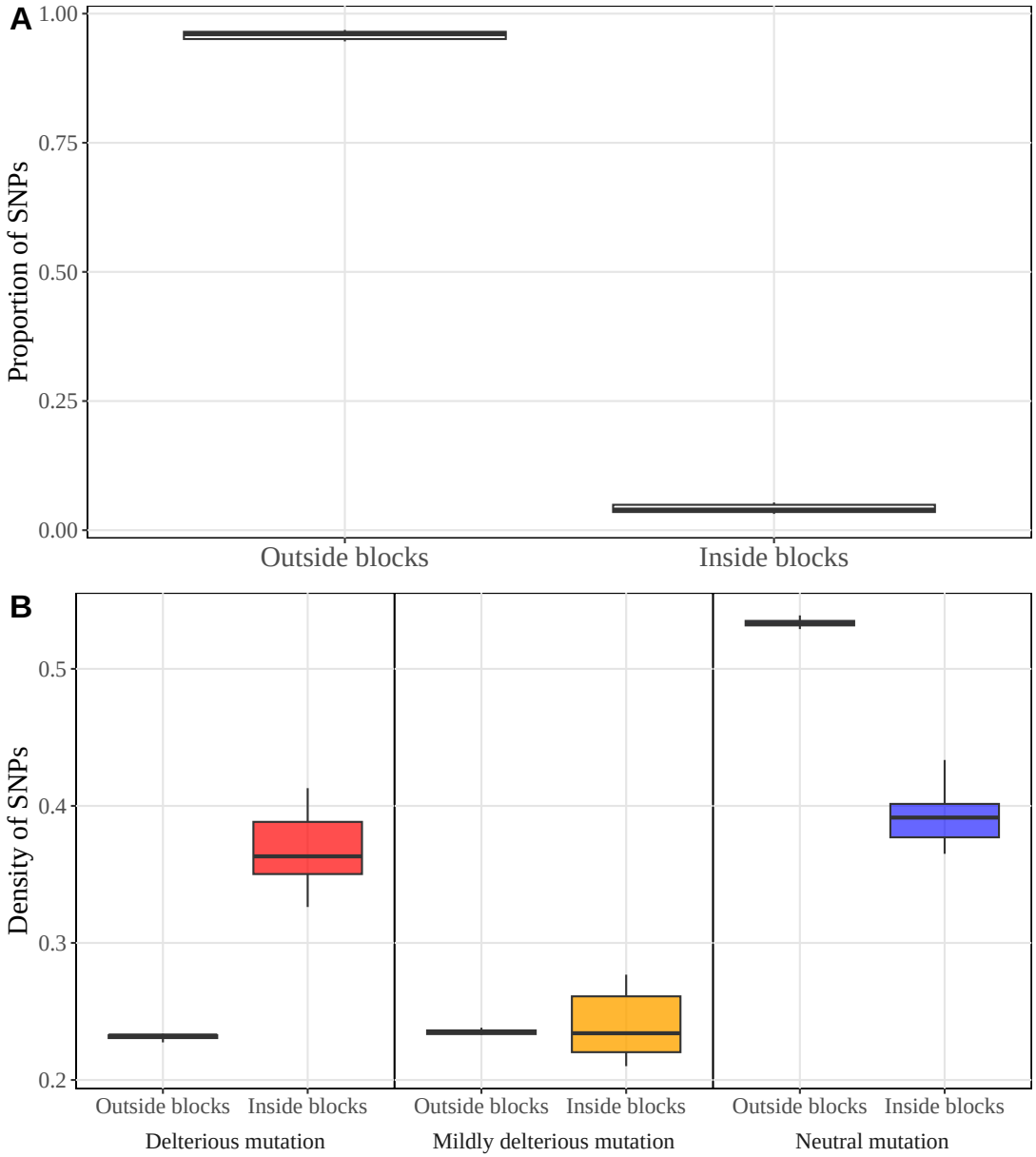


Figure 3.7: A: Proportion of biallelic SNPs annotated by SIFT and having at least one neutral mutation in regions of the genome covered or not covered by blocks. The proportion was computed for each chromosome. B: Density of SNPs carrying different types of mutations in regions of the genome covered or not covered by blocks. This proportion was computed, for each chromosome, by dividing the number of SNPs carrying a deleterious, mildly deleterious or neutral mutation by the total number of those three types of SNPs.

ity patterns (Figure S2.13). Among those, 727 present a significantly higher degree of coverage in heterozygous individuals, and were retained for the following analyses. These blocks are on average 1942 bp long, carry 13.8 SNPs with similar heterozygosity pattern and 6.2 deleterious SNPs (Figure S2.14A, S2.14B, S2.14C). The distribution of the frequency of heterozygous individuals in these blocks is skewed towards low values with 80% of blocks having a frequency of heterozygotes lower than 0.2 (Figure S2.14D). Blocks cover 4.2% of all SNPs annotated by SIFT (Figure 3.7A) and a higher density of deleterious mutations is found in the genomic regions they cover compared to regions not covered by blocks, whereas the density in mildly deleterious and neutral mutations is similar inside and outside blocks or higher outside blocks, respectively (Figure 3.7B). To have an idea of the potential effect of structural variants on LD between deleterious mutations, we compared blocks carrying different numbers of SNPs with deleterious mutations. We considered four nested categories of blocks: blocks with at least 4 SNPs with deleterious mutations, blocks with at least 6 SNPs with deleterious mutations and blocks with at least 12 SNPs with deleterious mutations. Blocks carrying more SNPS with deleterious mutations are also larger on average (Figure 3.8A). Interestingly, blocks with higher number of SNPs with deleterious mutations tend to be in higher frequency on average, possibly due to the fact that blocks in higher frequency have had more time to accumulate deleterious mutations than blocks in lower frequency (Figure 3.8B). This contradicts the expectation that blocks carrying higher numbers of deleterious mutations should be counterselected and thus found at lower frequency. When blocks carrying at least 4, 6 or 12 SNPs with deleterious mutations are excluded from LD computation, the LD between deleterious mutations decreases, while the LD between mildly deleterious or neutral mutations is not affected (Figure 3.9). The highest effect on LD between deleterious mutations is obtained by removing blocks with at least 4 SNPs with deleterious mutations (since a higher number of deleterious mutations are excluded in this case), although LD remains positive around 1.5 kb (Figure 3.9A).

Our block detection algorithm aiming at detecting potential duplications only uses information on shared heterozygosity patterns between SNPs and coverage, and our

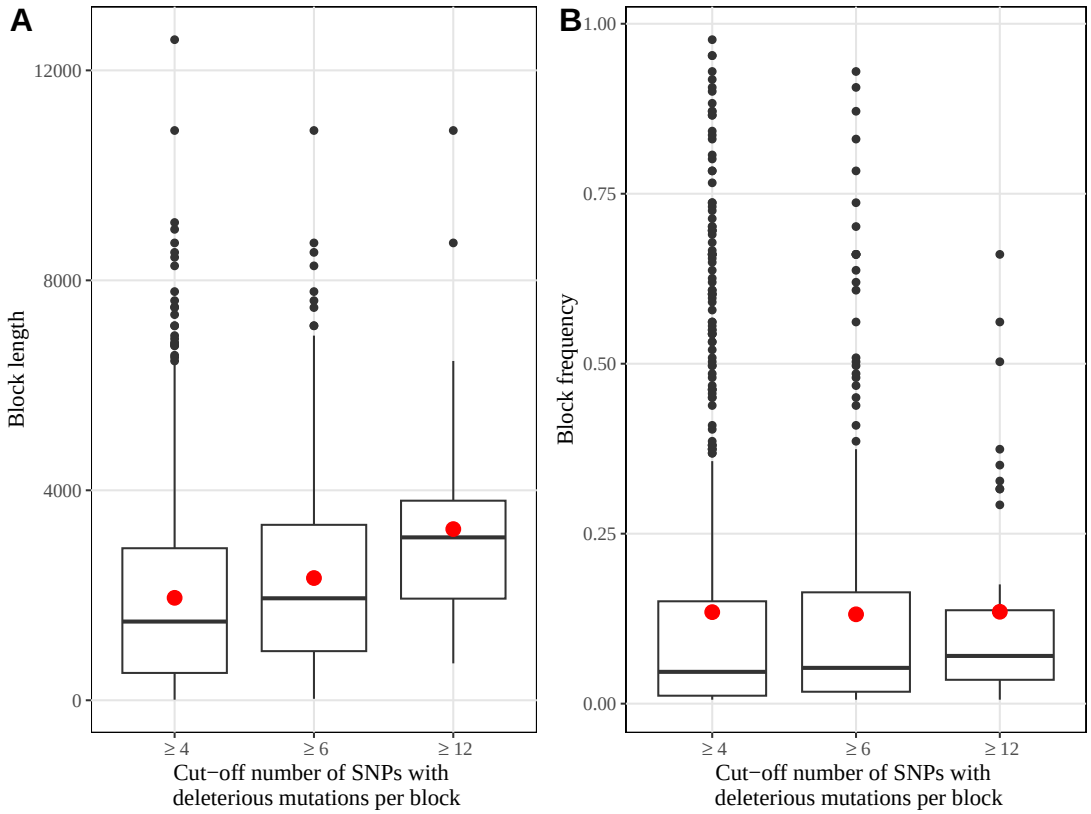


Figure 3.8: Block length (A) and block frequency (B) according to different nested block categories having increasing number of SNPs with deleterious mutations per block. Black lines correspond to the median and red points to the average of the distribution.

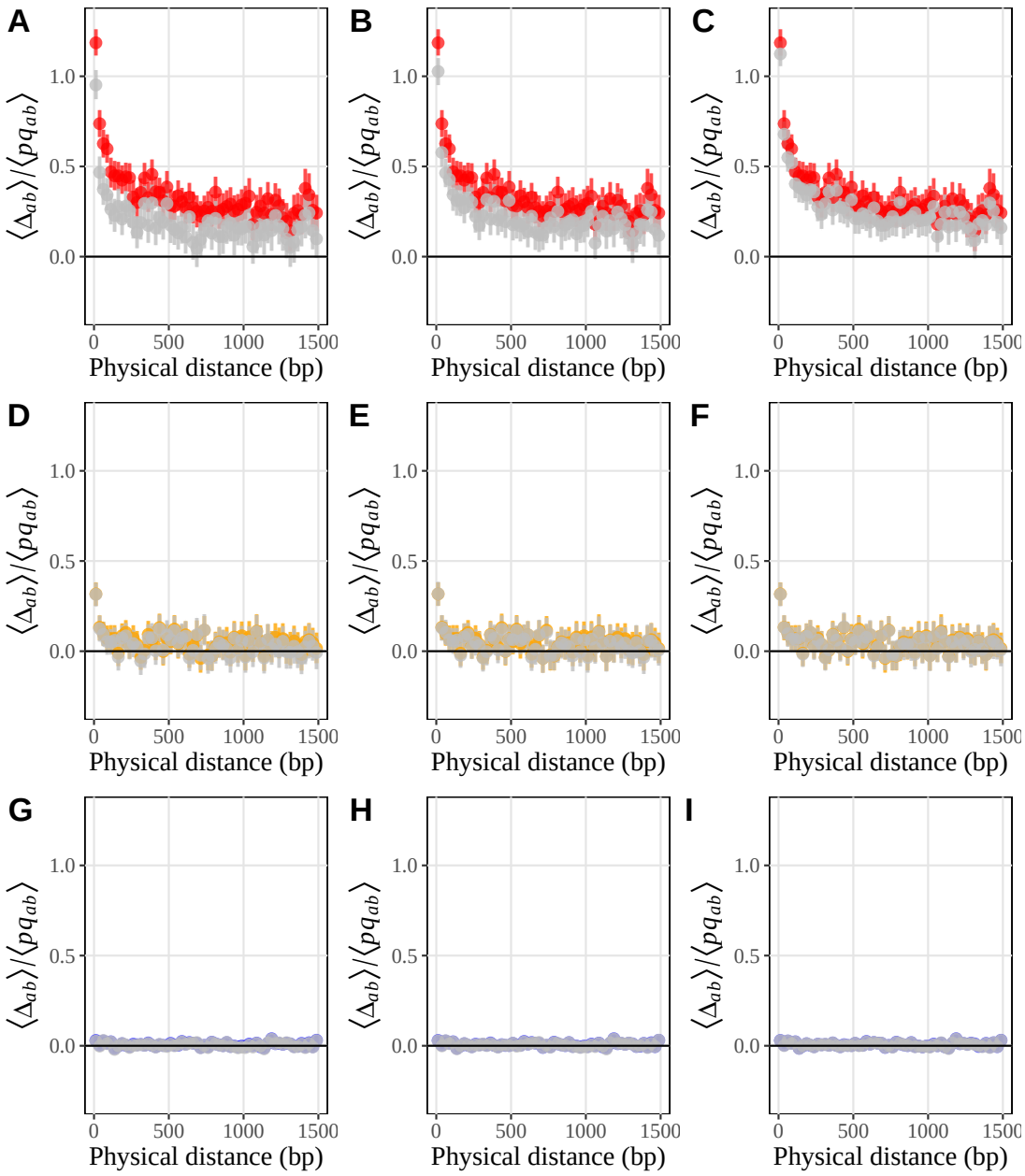


Figure 3.9: Composite linkage disequilibrium Δ_{ab} scaled by the product of allele frequencies between deleterious (A, B and C), mildly deleterious (D, E and F) and neutral (G, H, and I) mutations according to physical distance after removing SNPs covered by different block categories (grey) or considering all SNPs (coloured, as in Figure 5). Block categories are: blocks with at least 4 SNPs with deleterious mutations (A, D and G), blocks with at least 6 SNPs with deleterious mutations (B, E and H) and blocks with at least 12 SNPs with deleterious mutations (C, F and I).

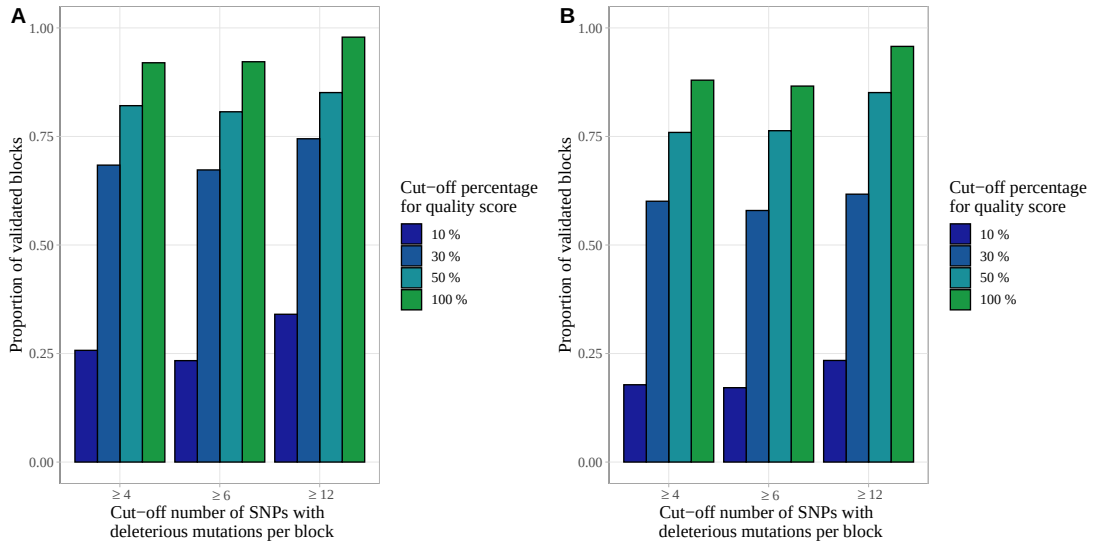


Figure 3.10: Proportion of blocks found using our block detection algorithm that were validated by duplication calls from Delly (A) or Smoove (B). A block is considered as validated if it overlaps with a called duplication in at least half of the individuals carrying the block. Each colour represents a cut-off percentage for duplication calls with the highest quality inferred by Delly or Smoove. This proportion is computed for different block categories with different numbers of SNPs with deleterious mutations per block.

detected blocks may thus result from mapping errors instead of structural variants. In parallel, we thus used standard structural variant (SV) calling programs for short-read sequencing. We used Smoove (github.com/brentp/smoove) and Delly (github.com/dellytools/delly; Rausch et al. 2012) that both detect potential duplications from information on paired-ends, split-reads and read-depth provided in short reads alignments. After running Smoove and Delly on the aligned short-read sequences of *C. grandiflora* individuals (see Methods) we found that, overall, a higher proportion of blocks with at least 12 SNPs with deleterious mutations were validated by Smoove and Delly compared to other block categories (Figure 3.10). After filtering for duplication calls with the highest 10% quality, Smoove validated between 23% and 34% and Delly between 17% and 23% of blocks as duplications.

Overall, our results thus suggest that the positive LD observed between deleterious mutations may at least partly be explained by duplications present in the sequenced

individuals and absent from the reference genome.

4 Discussion

Linkage disequilibrium among deleterious mutations plays an important role in theoretical models on the evolution of sex and recombination, and has recently started to be inferred from genomic data (Sohail et al., 2017; Garcia and Lohmueller, 2021; Sandler et al., 2021; Stolyarova et al., 2022). These recent studies typically apply some form of minor allele frequency filtering (measuring LD among rare variants), and tend to find positive LD (on average) among neutral mutations, often interpreted as an effect of demography or admixture. However, Good (2022) obtained analytical approximations showing that positive LD is expected among rare mutations, this effect being stronger for neutral alleles than for deleterious alleles. Our simulation results confirm and extend these results, showing that positive LD is expected (on average) among alleles present at similar frequencies within a population. Interestingly, LD between neutral derived alleles stays close to zero in the *Capsella grandiflora* dataset used in this paper, in the absence of any filtering on frequency. This indicates that the positive LD among synonymous sites found by Sandler et al. (2021) using the same dataset was probably a consequence of restricting the analysis to low-frequency variants. Our simulations also show that ancestral/derived state misspecification does not have an effect on LD between mutations as long as all frequencies are considered. This can explain why we still find null LD between derived neutral alleles despite the fact that a non-negligible part of SNPs probably have misspecified allelic state, as can be seen by the excess of high frequency neutral mutations. Therefore, when no frequency filter is applied, neutral mutations should display similar LD whether alleles are randomly polarized or polarized by an outgroup sequence.

Our results also show that, contrarily to the case of neutral mutations, LD between putatively deleterious mutations remains positive even in the absence of any conditioning on allele frequency. We explored a possible source of such positive LD, corresponding

to mapping errors yielding pseudo-heterozygosity at SNPs located in a given genomic region. In particular, pseudo-heterozygosity may be generated by duplicated sequences present in some individuals, but absent from the reference genome (Jaegle et al., 2023). Because the program we used to infer the fitness effect of mutations (SIFT4G) uses multiple protein alignments, the mutations detected as deleterious are only potentially deleterious as their fitness effect will depend on the transcription profile of the gene carrying the mutation. This nuance can become very important when inferred alleles actually come from duplicated genes after pseudogenisation where pseudo-deleterious mutations can accumulate. Pseudogenes are common in genomes (up to 20,000 are estimated in the human genome; Harrison et al. 2002) and this would explain the relative high number of deleterious mutations found in the genome of *C. grandiflora* and their higher density in blocks of similar heterozygosity pattern. This could thus explain the positive LD observed among deleterious alleles, if mutations categorized as deleterious by SIFT accumulate in duplicated regions (and therefore tend to be present in the same individuals, corresponding to those carrying the same duplication).

Using the combined information of heterozygosity patterns across SNPs and coverage, we detected 727 blocks, and between 17% and 34% of those blocks were detected as duplications by structural variant detection programs that use information from paired-ends, split-reads and read-depth. However, short-read sequences greatly limit the power to detect duplications and it is difficult to find many regions for which all types of information clearly point to a duplication. The blocks detected by our algorithm could also correspond to mapping errors, that may not necessarily correspond to duplications (also it is not fully obvious why the same mapping error would occur in some individuals but not all). Furthermore, rare haplotype blocks that are only present in the heterozygous state are expected due to the structure of the ancestral recombination graph (Shipilina et al., 2022). This should primarily concern small blocks in low frequency, and concentrating on larger blocks present at higher frequencies should increase the proportion of true duplications. Instead of using information on shared heterozygosity pattern between SNPs, one could use information on F_{IS} combined with the deviation of average

read-depth at each SNP, proposed by McKinney et al. (2017). This method (*HDplot*) has been applied to three species of salmons and to the American lobster to detect duplications (McKinney et al., 2017; Dorant et al., 2020). Although *HDplot* was used on data from genotyping-by-sequencing (GBS) that are not mapped to a reference genome, the signature of pseudo-heterozygosity combined with deviations from average read-depth should be similar in both types of data. Another possibility would be to perform GWAS on each block detected as a potential duplication, using heterozygosity as a trait, as was done by Jaegle et al. (2023); this could also yield additional information on the number and position of the different copies of a duplication in the genome. Without long-read sequencing data, duplications can only be detected by finding the best trade-off between sensitivity and specificity of currently available tools. However, long-read sequencing data are starting to be available (De Coster et al., 2021) which will greatly help detect duplications and removing them from LD analysis.

We found that the positive LD among deleterious mutations is reduced after removing genome blocks identified as potential duplications and carrying at least 4 SNPs with deleterious mutations, indicating that duplications (or, more generally, mapping errors) may contribute to this positive LD. Therefore, we argue that technical artifacts (whether they are caused by duplications or not) known to create spurious positive LD should first be discarded before drawing interpretations based on selective and/or demographic processes generating positive LD among deleterious alleles. A way to bypass the issue of pseudo-heterozygous SNPs would be to measure LD in genomic data obtained from haploid individuals, such as the sequences from haploid embryos from a Zambian population of *Drosophila melanogaster* used by Sohail et al. 2017 and Sandler et al. 2021. Another possibility would be to focus on highly selfing species where heterozygosity is expected to be very rare, and where all SNPs displaying heterozygosity may be removed without affecting the overall LD pattern. We take advantage of published data from two highly selfing species to explore LD between deleterious mutations in the following chapter.

In the absence of a strict criterion to identify mapping errors and other types of artifacts, it is difficult to assess whether all the positive LD between putative deleterious

mutations can be explained by such artifacts or not. If this is not the case, the absence of LD among neutral variants indicates that a form of selective mechanism is likely to be involved. A first possibility, already considered by other authors (Sohail et al., 2017; Sandler et al., 2021; Ragsdale, 2022; Stolyarova et al., 2022) is positive epistasis among deleterious alleles. This may correspond to possible compensatory effects between mutations within the same gene (Davis et al., 2009), or to the fact that once a gene has become non-functional due to a first mutation, additional mutations within the same gene become neutral and can accumulate. Some evidence for possible compensatory effects is given in Ragsdale (2022) and Stolyarova et al. (2022). In particular, without conditioning on allele frequencies, Ragsdale (2022) found positive LD at very short distances (<300 bp) inside genes for both missense and synonymous mutations in several human populations. No difference in LD was observed between synonymous and missense mutations when averaging within genes, but LD was higher between missense mutations inside conserved gene domains compared to synonymous mutations (while the opposite pattern was found outside conserved gene domains), which was interpreted as evidence for positive epistasis among deleterious mutations affecting the same conserved domain. Furthermore, (Stolyarova et al., 2022) showed that, controlling for the physical distance in base-pair among mutations, higher positive LD is observed among mutations affecting aminoacids that are physically close to each other in the protein 3D configuration (than among mutations affecting aminoacids that are farther apart), in two populations of the fungus *Schizophyllum commune*. Another possible source of positive LD among deleterious alleles could be population structure. In particular, different strengths of selection against deleterious alleles at different locations (which could be due to different magnitudes of drift or degrees of inbreeding at those different locations) could generate a higher mutation load in some locations, so that individuals sampled from those locations would carry a higher number of deleterious alleles (Roze, 2009; Roze and Otto, 2012). Finally, positive LD is expected (on average) between deleterious mutations in finite, panmictic populations as long as these mutations are sufficiently recessive, and $Ns \gg 1$ (when N is population size and s the strength of selection against deleterious

alleles; Roze, 2021; Ragsdale, 2022). Overall, more work is thus needed to assess the relative contribution of these different mechanisms and of statistical biases and technical artifacts on positive LD among deleterious mutations within natural populations.

4.1 Data availability

The scripts used to perform genomic analyses are available on GitHub at: github.com/RomanStet/LD_deleterious_mutations.

Chapter 4

Linkage disequilibrium in highly selfing species

An important part of the data analysis on *A. thaliana* presented in this chapter was performed by Lise Porcheron during her M1 internship in the lab.

1 Introduction

We have seen in the previous chapter that linkage disequilibrium (LD) between deleterious mutations seems positive in an outcrossing population with low population structure of *Capsella grandiflora*, even after accounting for the possible effect of structural variants. However, designing a robust method that would detect 100% of pseudo-heterozygosity within short-reads data (either due to structural variants or to other causes of mapping errors) appears difficult. One possible way around this problem is to consider highly selfing populations. Indeed, heterozygosity is expected to be very low in highly selfing species (Wright et al., 2008). Nevertheless, an important proportion of SNPs showed high rates of heterozygosity in the 1001 Genomes project on the highly selfing *Arabidopsis thaliana*, which was shown to result from mapping errors, due (at least in part) to polymorphic duplications that were absent from the reference genome (Jaegle et al., 2023). Because the sequenced accessions of *A. thaliana* were obtained by several generations of

selfing (from an already inbred ancestor), heterozygosity should be extremely low, and SNPs showing heterozygosity can thus safely be discarded from the dataset. Similarly, LD analyses in other highly selfing species can in principle be done by excluding heterozygous SNPs. Measuring LD between deleterious mutations has been the subject of growing interest, but to date has only been performed in outcrossing species (Sohail et al., 2017; Garcia and Lohmueller, 2021; Sandler et al., 2021; Ragsdale, 2022; Stolyarova et al., 2022).

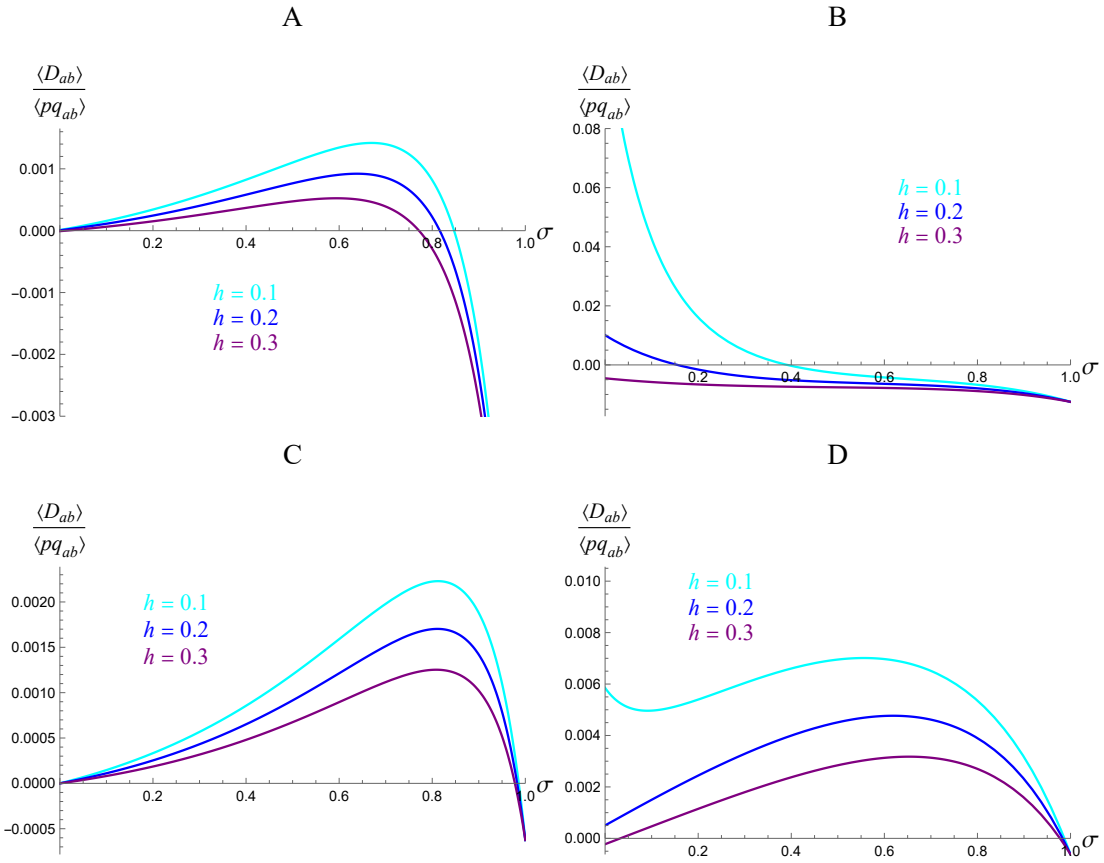


Figure 4.1: Expected linkage disequilibrium D_{ab} between deleterious mutations at mutation-selection-drift balance, scaled by the expected product of allele diversities, as a function of the selfing rate σ and for different values of the dominance coefficient h, recombination rate r_{ab} and population size N. The different parameters are: $r_{ab} = 0.1$ (A, C); $r_{ab} = 10^{-4}$ (B, D); $N = 1,000$ (A, B); $N = 20,000$ (C, D) and $s = 0.02$ for all plots. Curves were drawn from Equation 4.1 and 4.2.

Yet, selfing in itself is expected to have important qualitative and quantitative ef-

fects on LD between deleterious mutations. In randomly mating populations and in the absence of epistasis, LD is only generated by the Hill-Robertson effect (selective interference among loci), but may be positive or negative according to the dominance coefficient of deleterious alleles h: positive when $h < 0.25$, and negative when $h > 0.25$ (Roze, 2021). Furthermore, it increases in absolute value as the recombination rate r_{ab} between the two loci decreases, and as the population size N decreases. The fact that selfing increases homozygosity tends to increase the effective dominance coefficient of deleterious alleles to $\tilde{h} = h\,(1-F) + F$ (where $F = \sigma/\,(2-\sigma)$ is the inbreeding coefficient and σ the selfing rate), as these alleles appear more often in the homozygous state. It also decreases the effective recombination rate to $\tilde{r}_{ab} = r_{ab}\,(1-F)$, and the effective population size to $N_e = N/\,(1+F)$ (Nordborg, 1997; Glémin and Ronfort, 2013; Roze, 2016). Plugging these effective coefficients into the approximation for the LD among deleterious alleles derived in Roze (2021) yields:

$$\langle D_{ab,\mathrm{HR}} \rangle \approx \frac{s^2 \tilde{h}(1-4\tilde{h})}{2N_e(\tilde{r}_{ab}+2s\tilde{h})^2(\tilde{r}_{ab}+3s\tilde{h})}\ \langle pq_{ab} \rangle\,, \tag{4.1}$$

where $\langle D_{ab,\mathrm{HR}} \rangle$ is the average LD between deleterious alleles a and b generated by the Hill-Robertson effect, and $\langle pq_{ab} \rangle$ the average of $p_a\,(1-p_a)\,p_b\,(1-p_b)$. Equation 4.1 indicates that the LD generated by the Hill-Robertson effect should be negative for all values of h in highly selfing populations (due to the fact that $\tilde{h}$ approaches 1), and increases in absolute value due to the decrease in $\tilde{r}_{ab}$ and N_e. An additional effect of partial selfing is that it generates correlations in homozygosity among loci, which can be measured by the identity disequilibrium G_{ab} (Weir and Cockerham, 1973), representing a correlation in identity-by-descent across loci. This can be understood by the fact that different types of lineages coexist within a partially selfing population: more strongly inbred lineages (in which homozygosity is relatively high at all loci), and less inbred lineages which are relatively more heterozygous at all loci. This identity disequilibrium deterministically generates positive LD among deleterious alleles: indeed, deleterious alleles are more efficiently purged from more highly inbred lineages (and are thus less

frequent at all loci in those lineages), while they are less efficiently purged from less inbred lineages and are thus more frequent at all loci within those lineages. From Stetsenko and Roze (2022), an approximation for this effect is given by:

$$\langle D_{ab,\mathrm{det}} \rangle \approx \frac{s^2 \left(1-h\right)^2 \left(1+2F\right) G_{ab}}{\tilde{r}_{ab} + 2s\tilde{h}} \, \langle pq_{ab} \rangle \,, \tag{4.2}$$

which is always positive, and maximized for intermediate selfing rates (as the identity disequilibrium G_{ab} is maximized for intermediate selfing rates). Note that equation 4.2 does not depend on population size, as this effect is deterministic. The overall LD between deleterious alleles (given by the sum of equations 4.1 and 4.2) should thus be negative at very high selfing rates (as the deterministic effect of G_{ab} then vanishes). However, the deterministic term may dominate even at rather high selfing rates when N is sufficiently large (Figure 4.1C, D), yielding positive LD. The stochastic term $D_{ab,\mathrm{HR}}$ becomes relatively stronger at low population size, and as the recombination rate decreases (Figure 4.1A, B), yielding negative LD as soon as the selfing rate is sufficiently large.

We took advantage of the large dataset available from the highly selfing model species *Arabidopsis thaliana* and focused on the most diverse genetic group located around the Iberian Peninsula, called the relict group. This relict group is thought to have survived the last glacial era in a refugium in North Africa (Brennan et al., 2014) and shows less evidence of rapid expansion and bottlenecks compared to non-relict populations (The 1001 Genomes Consortium, 2016) thus decreasing potential demographic effects on LD patterns. In addition to the very high selfing rate typically estimated in *A. thaliana* (around 0.95; e.g. Abbott and Gomes 1989; Stenøien et al. 2005) sequenced individuals are the result of several generations of selfing in controlled conditions, ensuring high homozygosity. We also used the filtered SNPs from 33 *Capsella orientalis* sequences that were mapped on a new reference genome of *Capsella rubella*, and only kept homozygous SNPs. Although an estimation of the selfing rate of *Capsella orientalis* is not available, it is considered high based on the low levels of genetic diversity in this species (e.g.

Hurka et al. 2012; Žerdoner Čalasan et al. 2021). As in the previous chapter, we used SIFT annotations in order to infer the fitness effect of mutations, and measured LD between putatively deleterious and neutral mutations in both species. While the results show positive LD among deleterious mutations in *A. thaliana*, LD does not seem to be significantly different from zero (for all mutation classes) in *C. orientalis*, except for very tightly linked deleterious alleles. We then discuss possible mechanisms that can generate positive LD between deleterious mutations in highly selfing species.

2 Materials & Methods

2.1 Genomic data

We retrieved data from 25 accessions belonging to the "relict" admixture group from the 1001 Genomes database of *Arabidopsis thaliana* (1001genomes.org; The 1001 Genomes Consortium (2016)). This admixture group, whose accessions are located in the Iberian Peninsula, Morocco, Cape Verde Islands and the Canary Islands (Figure 4.2), displays the highest genetic diversity among all European accessions of *A. thaliana* The 1001 Genomes Consortium (2016). Data were retrieved in the form of filtered SNPs with fully homozygous genotypes coming from mapping of the short-read sequences against the TAIR10 reference genome. For *Capsella orientalis*, we mapped the short-read sequences from 33 accessions across its geographic range in Central Asia (Figure 4.3; retrieved from Ågren et al. 2014; Huang et al. 2018; Kryvokhyzha et al. 2019) against a new (and not yet published) reference genome of *Capsella rubella*, Cr145 (Tianlin Duan, personal communication). The variant calling and variant filtration were performed as explained in Chapter 3. Since SNPs from *C. orientalis* had a different distribution of mean read depth per site, we applied a different filter for maximum read depth (>45X; Figure S3.1). Based on a PCA on the filtered independent SNPs (using the function *indep-pairwise* of Plink with parameters: window size = 50 kb, step size = 100 variants and r^2 threshold = 0.1; Purcell et al. 2007), we identified 6 *A. thaliana* (Figure 4.4A) and 8 *C. orientalis* (Figure 4.4B, 4.4C) diverged individuals that were discarded from LD

computation. The list of individuals can be found in Table S3.1 (*A. thaliana*) and Table S3.2 (*C. orientalis*). This yielded 1,818,928 biallelic SNPs in *A. thaliana* and 549,719 in *C. orientalis*. Since the sequences from *C. orientalis* showed an important fraction of SNPs with heterozygosity (52%), we filtered out those SNPs to obtain fully homozygous genomes yielding 285,456 SNPs.

2.2 SIFT annotation

In order to classify mutations by their potential fitness effect we used the SIFT4G algorithm (Vaser et al., 2016) which uses multiple protein alignment across a database of proteins in a large diversity of taxa (See Chapter 3 for details on SIFT4G). We annotated SNPs from *A. thaliana* SIFT database of the TAIR10 reference genome whereas for *C. orientalis* we generated a SIFT database from the new reference genome of *C. rubella* that we used for SNP calling (github.com/pauline-ng/SIFT4G_Create_Genomic_DB). By default, SIFT classifies variants with a SIFT score ≤ 0.05 as deleterious and > 0.05 as tolerated. In order to contrast between the potentially most deleterious mutations and potentially neutral mutations while considering mutations with intermediate fitness effects we stated the following categories: deleterious mutations (SIFT score ≤ 0.05), mildly deleterious mutations ($0.05 <$ SIFT score < 1) and neutral mutations (SIFT score $= 1$).

2.3 Computing LD

Since our filtered dataset only comprises homozygous genomes, we could compute a simple LD estimate D_{ab} measuring the association between alleles a and b in the same individual and given by:

$$D_{ab} = p_{ab} - p_a p_b, \tag{4.3}$$

where p_{ab} is the frequency of double mutant haplotypes and p_a, p_b the frequencies of alleles a and b (Weir, 1996). To compute LD among neutral alleles, we only considered sites that were segregating for neutral alleles (SIFT score = 1). In *A. thaliana*, alleles

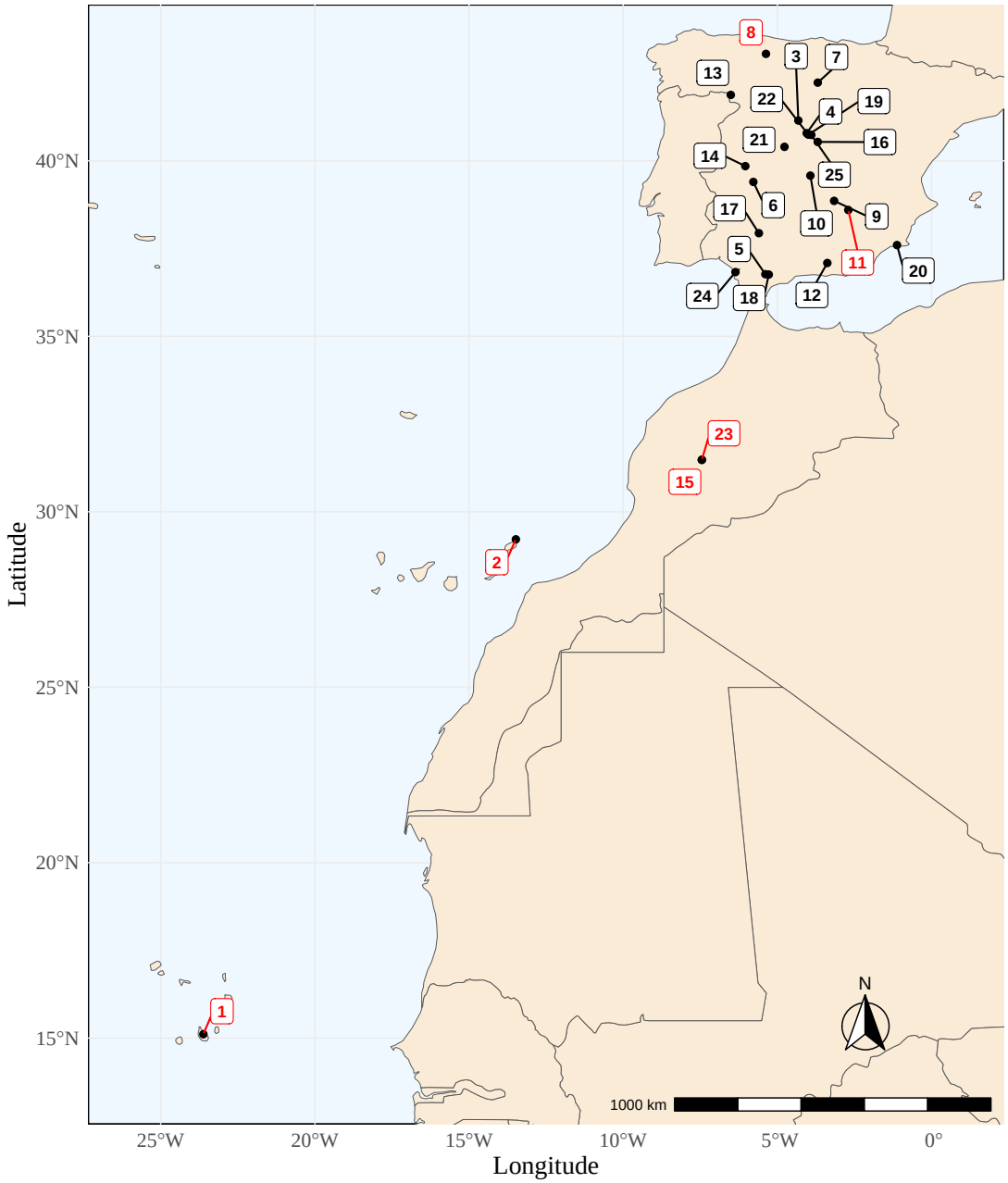


Figure 4.2: Geographic distribution of *A. thaliana* accessions labeled as the relict group from The 1001 Genomes Consortium (2016). Red numbers correspond to genetically divergent individuals that were removed from LD computation. The correspondence between the accession number and the original accession ID is available in Table S3.1.

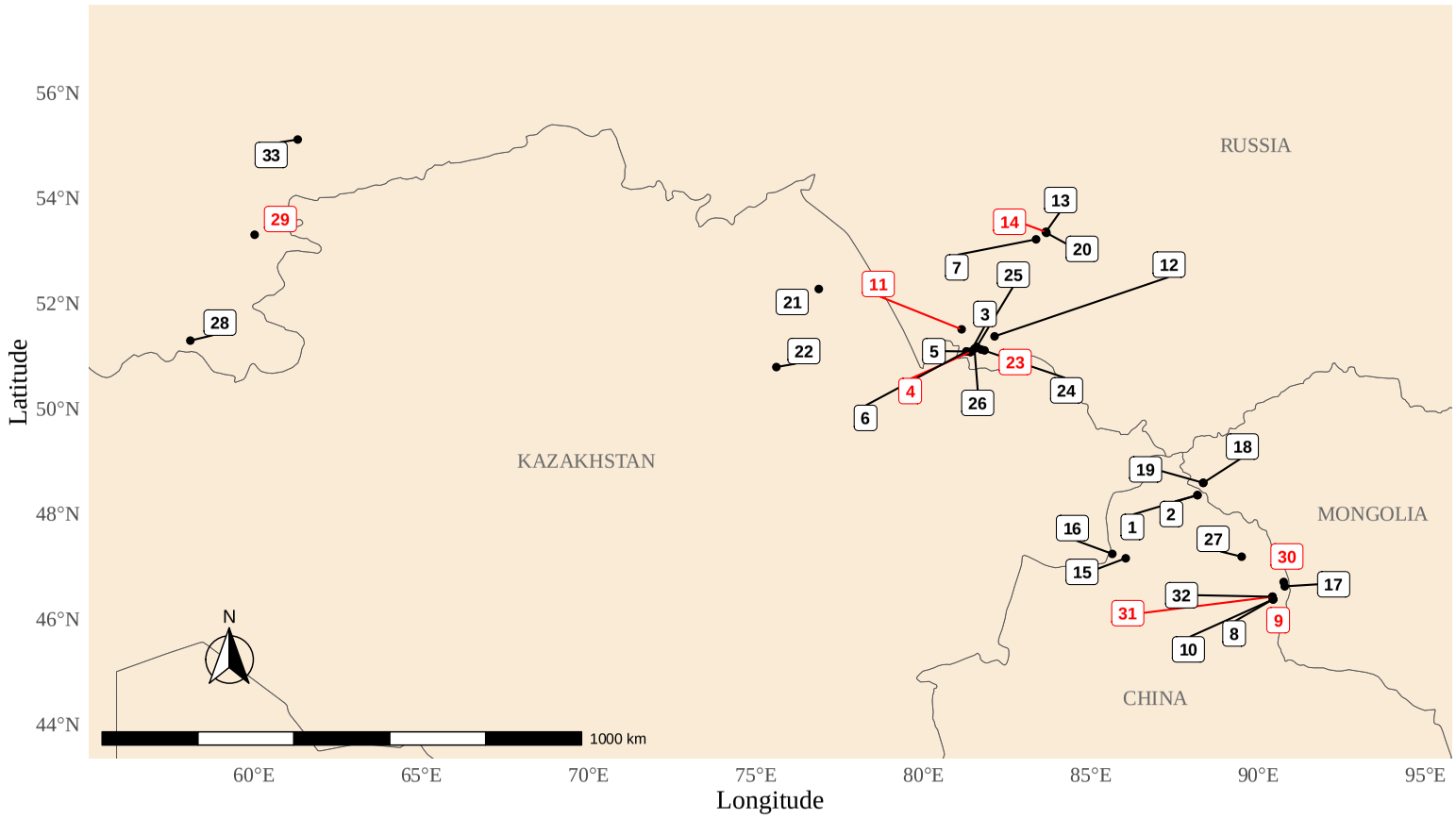


Figure 4.3: Geographic distribution of *C. orientalis* individuals from Ågren et al. (2014); Huang et al. (2018); Kryvokhyzha et al. (2019). Red numbers correspond to genetically divergent individuals that were removed from LD computation. The correspondence between the accession number and the original accession ID is available in Table S3.2.

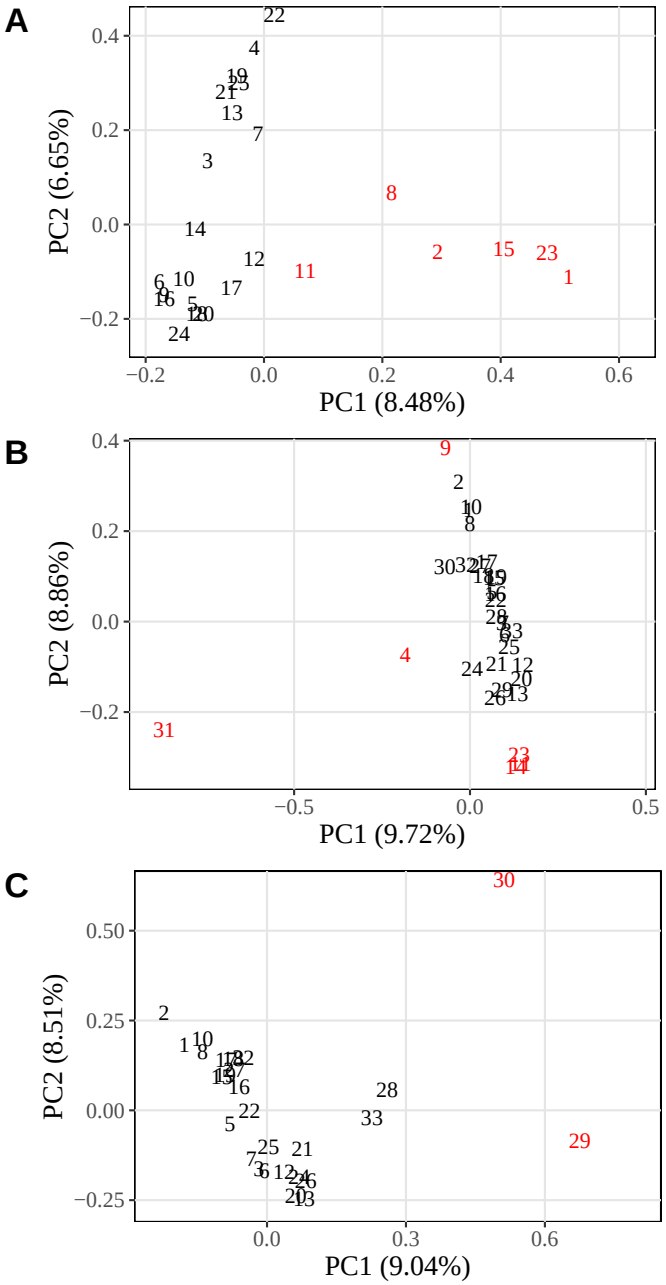


Figure 4.4: The two first components of the PCA on filtered SNPs of the relict individuals of *A. thaliana* (A) and *C. orientalis* individuals (B, C). Red individuals correspond to genetically divergent individuals that were removed for LD computation. Red individuals in plot B where removed to draw plot C. PCA was performed using Plink (Purcell et al., 2007) on independent SNPs extracted with *indep-pairwise* function (window size = 50 kb, step size = 100 variants and r^2 threshold = 0.1).

were randomly polarised for neutral LD computation, while in *C. orientalis* they were polarized based on their inferred derived state, using the outgroup sequence of *Neslia paniculata* (*i.e.*, a and b correspond to derived alleles in equation 4.3). To compute LD between deleterious/mildly deleterious alleles in *A. thaliana*, we only retained sites at which a deleterious/mildly deleterious allele is segregating with a neutral allele, and polarized alleles based on their inferred fitness effect. In *C. orientalis* the same filtering on deleterious/mildly deleterious allele was applied but the analysis was restricted to derived deleterious alleles (*i.e.*, a and b correspond to deleterious and derived alleles in equation 4.3) — extending the analysis to include ancestral deleterious alleles is not expected to change the results (see Chapter 3). We computed the mean value of D_{ab} for every pair of sites at different physical distances up to 3 kb (each distance class being 75 bp wide) and discarded pairs of sites for which the information on the genotype was not available at both sites in less than half of the total number of individuals. As for the LD estimator σ_D^2 of Ohta and Kimura (1969), the average LD for a given distance class $\langle \Delta_{ab} \rangle$ was scaled by the average product of allele diversities at both loci (for the same distance class) $\langle p_a q_a p_b q_b \rangle$, where $\langle X \rangle$ denotes the average of quantity X over pairs of sites in a given distance class. A 95% confidence interval was computed using the bootstrap method, by resampling each set of pairs of sites 1,000 times.

3 Results

In *A. thaliana* the diverged accessions correspond to all accessions outside of the Iberian Peninsula with two additional accessions inside it (Figure 4.2 and Figure 4.4A) and most of them also have a lower proportion of deleterious mutations compared to the other accessions (Figure S3.3A). Unlike in *A. thaliana*, the diverged accessions of *C. orientalis* are spread across the geographic range of the species and are probably accessions whose sequences poorly mapped on the reference genome of *C. rubella*, resulting in a high proportion of missing data after SNP filtering (Figure S3.2B). This does not seem to be the case for diverged accessions in *A. thaliana* as all accessions have comparable

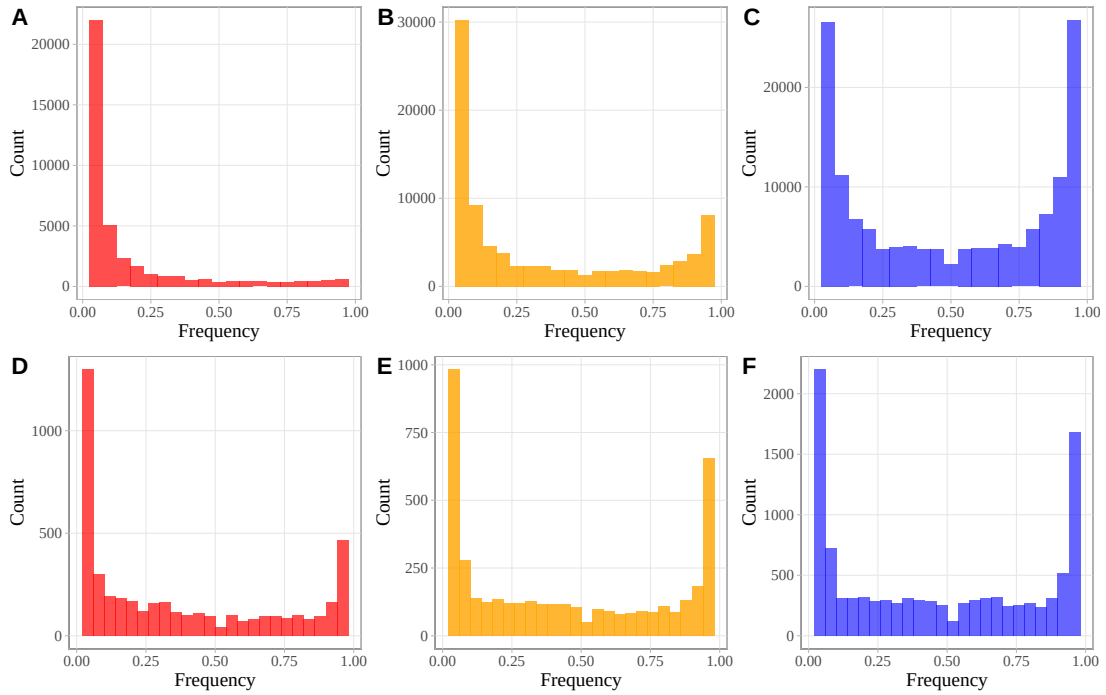


Figure 4.5: Site frequency spectrum for deleterious (A, D), mildly deleterious (B, E) and neutral mutations (C, F) in *A. thaliana* (A, B, C) and *C. orientalis* (D, E, F) accessions. Frequencies were computed for fully homozygous biallelic sites for which at least one neutral mutation segregates. For *A.thaliana* deleterious and mildly deleterious mutations were polarized based on their SIFT score while neutral mutations were randomly polarized. For *C. orientalis* deleterious and mildly deleterious mutations were polarized by their SIFT score and derived state (based on the outgroup sequence of *N. paniculata*) while neutral mutations were polarized based on their derived state.

proportions of missing data (Figure S3.2A). The overall higher proportion of deleterious alleles between *C. orientalis* compared to *A. thaliana* (Figure S3.3) probably reflects the relatively high levels of purifying selection for a selfing species already found in *A. thaliana* (Payne and Alvarez-Ponce, 2018).

Site frequency spectra (SFS) for the different types of mutations (deleterious, mildly deleterious and neutral) reflects their inferred fitness effect as more deleterious mutations display SFS shifted towards lower frequencies (Figure 4.5). The SFS of neutral mutations in *A. thaliana* is symmetrical because alleles were randomly polarised. The deficit of mutations with frequencies around 0.5 is due to the removal of SNPs displaying

heterozygosity, primarily affecting completely heterozygous SNPs (Figure S3.4). Using the information on the outgroup sequence of *N. paniculata* in addition to the SIFT score to polarise mutations in *C. orientalis* reduced the part of high frequency deleterious and mildly deleterious mutations (Figure S3.5). The part of high frequency deleterious and mildly deleterious mutations is still much higher in *C. orientalis* compared to *A. thaliana*.

We retrieved an average null LD between neutral mutations in both species regardless of the physical distance separating them (Figure 4.6C, 4.6F); note that this is expected in the case of *A. thaliana*, since neutral alleles are polarized randomly. In *A. thaliana* we found positive LD between deleterious and mildly deleterious mutations at short distances, deleterious mutations displaying higher positive LD, which is maintained over longer distances (Figure 4.6A, 4.6B). In *C. orientalis*, the results are more noisy due to the lower number of SNPs, but LD does not seem to be significantly different from zero for deleterious and mildly deleterious mutations, except for the smallest distance class in the case of deleterious mutations. Including the diverged accessions of the relict group of *A. thaliana* does not have an effect on LD for any type of mutation (Figure 4.7).

The average squared LD between the different types of mutations is consistent with the effect of selection further reducing the variance in LD between more deleterious variants in both species (Figure S3.6). Higher squared LD is maintained between mutations in *C. orientalis* compared to mutations in *A. thaliana*. In particular, squared LD between neutral mutations reaches a plateau at around 80 kb in *A. thaliana* (Figure S3.7A) while it reaches a plateau at around 200 kb in *C. orientalis* (Figure S3.7B).

4 Discussion

Our results show that positive LD is still retrieved between deleterious mutations at short distances, particularly in relict accessions of *A. thaliana*. This suggests that an important part of this positive LD is caused by other factors than pseudo-heterozygosity generated by polymorphic duplications or other sources of mapping errors, which could

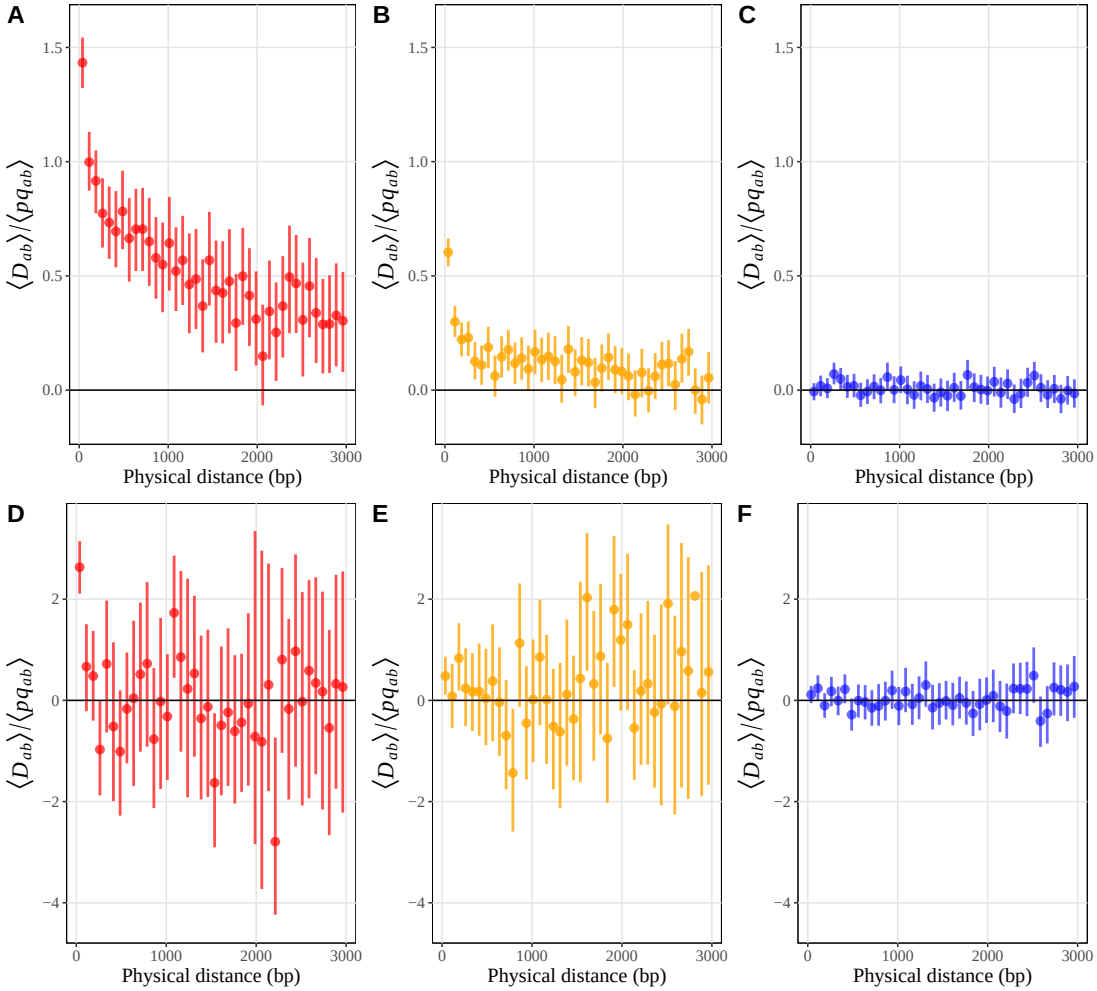


Figure 4.6: Average linkage disequilibrium D_{ab} scaled by the average product of allele diversities between deleterious (A, D), mildly deleterious (B, E) and neutral mutations (C, F), for different classes of physical distance between sites (in base pairs) in *A. thaliana* (A, B, C) and *C. orientalis* (D, E, F). Linkage disequilibrium was computed only between fully homozygous SNPs. For *A. thaliana*, deleterious and mildly deleterious mutations were polarized based on their SIFT score while neutral mutations were randomly polarized. For *C. orientalis* deleterious and mildly deleterious mutations were polarized based on their SIFT score and derived state (based on the outgroup sequence of *N. paniculata*) while neutral mutations were polarized based on their derived state.

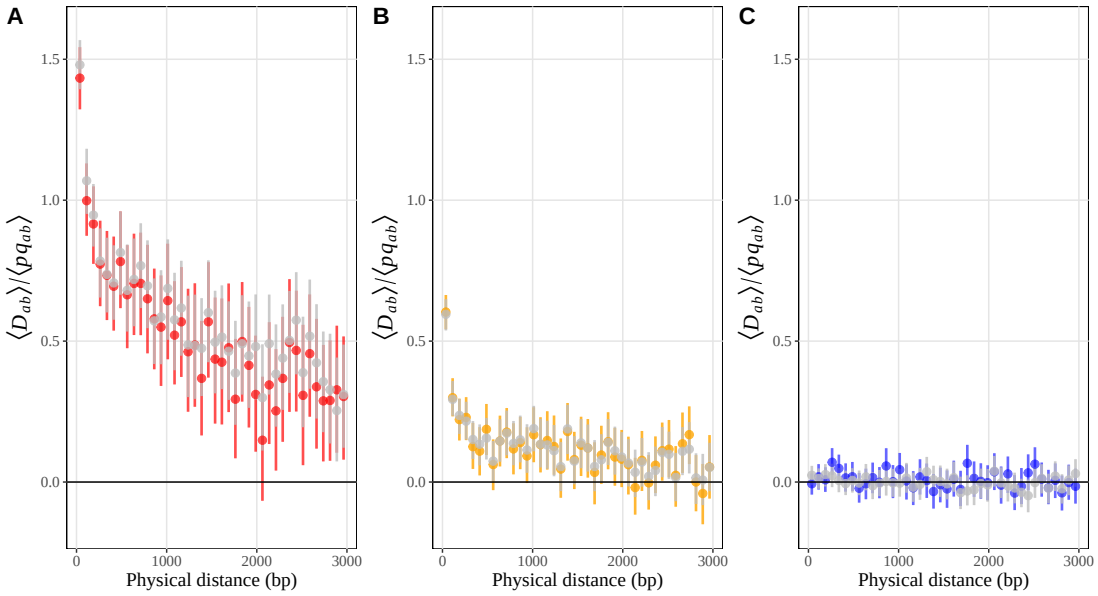


Figure 4.7: Average linkage disequilibrium D_{ab} scaled by the average product of allele diversities between deleterious (A), mildly deleterious (B) and neutral mutations (C), for different classes of physical distance between sites (in base pairs) in *A. thaliana* either without diverged accessions (colour, same as Figure 4.6A, B, C) or including diverged accessions (grey).

have a significant impact on LD patterns in outcrossing species (see Chapter 3). Positive LD between deleterious mutations could be generated by selection in spatially structured populations. Indeed, the variance in inbreeding between demes generated by drift causes deleterious mutations to be more efficiently purged in demes with more inbreeding and less efficiently purged in demes with less inbreeding creating positive LD when measured at the scale of the whole population (Roze, 2009). This mechanism is similar to the one generating positive LD between deleterious mutations in infinitely large populations of partially selfing individuals (Roze and Rousset, 2005; Stetsenko and Roze, 2022). Other factors may generate differences in the efficiency of selection in different spatial locations, such as differences in demographic regimes. Both effects can be particularly strong in highly selfing species, known to be often characterised by extinction-recolonisation dynamics (e.g. Guo et al. 2009; Willi et al. 2018; Orsucci et al. 2020). Moreover, individuals from both species were sampled at large geographical scales, probably covering different genetically divergent populations and increasing the effect of spatial structure. However, we found no difference in LD between deleterious mutations when including diverged accessions of the relict genetic group of *A. thaliana* (among which accessions from outside the Iberian Peninsula). This result could be explained by the fact that most of the genetic structure is found inside the Iberian Peninsula (Brennan et al. 2014; also visible in our PCA). It could thus be interesting to restrict LD measurements inside genetic clusters in the Iberian Peninsula, even though finding a sufficient number of deleterious SNPs among few accessions may be challenging. Sampling and sequencing larger numbers of individuals at a small spatial scale (similarly to what was done by Josephs et al., 2015 in *Capsella grandiflora*) could provide a dataset with minimum genetic structure to remove most of the effect of population structure on LD patterns. However, again, relict populations are found in a very specific environments (The 1001 Genomes Consortium, 2016), with patchy populations probably displaying very low genetic diversity within a patch so that the effect of selfing on LD patterns could be difficult to disentangle from other life-history traits of highly selfing species. The different genetic groups among relict populations in the Iberian Peninsula could also be caused by introgression from other

non-relict populations (The 1001 Genomes Consortium, 2016) and could cause positive LD between deleterious mutations as some accessions may carry specific introgressed combinations (Sandler et al., 2021). As discussed in Chapter 3, positive LD between deleterious mutations may also be caused by positive epistasis, either due to compensatory effects between deleterious mutations, or because strongly deleterious mutations such as loss-of-function mutations allow subsequent deleterious mutations to accumulate within the same gene. Selfing, by decreasing effective recombination is expected to increase the effect of positive epistasis, allowing beneficial sets of mutations to remain associated over longer physical distances.

Finally, we found very different patterns of mean squared LD $\langle D^2_{ab} \rangle$ between the two species, associations being maintained at much longer distances in *C. orientalis* (up to 200 kb; Figure S3.7B) compared to *A. thaliana* (up to 80 kb; Figure S3.7A). As $\langle D^2_{ab} \rangle$ is roughly proportional to $1/4N_e\tilde{r}_{ab}$ the differences in mean squared LD between both species can be due to a lower genome-wide recombination rate in *C. orientalis* and/or lower effective population size. This is in line with the fact that the transition to selfing in the lineage of *C. orientalis* is estimated to be very recent (Late Holocene, around 2.1 kya; Žerdoner Čalasan et al. 2021) compared to the lineage of *A. thaliana* (between 1 Mya and 500 kya with a colonisation of Europe around 80 kya; Durvasula et al. 2017). Because the transition from outcrossing to selfing by the loss of self-incompatibility is often associated with a decrease in population size (John R. Pannell, 2015; Encinas-Viso et al., 2020), *C. orientalis* has probably a much lower historical effective population size than *A. thaliana*.

Part IV

General discussion

The present thesis has presented theoretical approaches exploring the conditions under which recombination may be favoured (in particular, the effect of arbitrary rates of selfing, and the effect of genome structure), and genomic approaches estimating linkage disequilibrium between deleterious mutations, a potentially important component of indirect selection for recombination. In this last section, I discuss the results from all chapters to draw general conclusions and perspectives for future works.

Epistasis and the evolution of recombination

As shown by current models, epistasis on fitness is an important component of indirect selection for recombination: in randomly mating populations, it generates a deterministic force selecting for recombination when it is negative and favours complete linkage when it is positive (Kondrashov, 1984; Charlesworth, 1990; Barton, 1995; Otto and Feldman, 1997). In partially selfing populations, epistatic interactions involving dominance (additive-by-dominance and dominance-by-dominance epistasis) become important, and may generate a short-term benefit for recombination (Roze and Lenormand, 2005). However, in finite populations, the effect of epistasis may be overwhelmed by the stochastic forces selecting for recombination (even in the absence of epistasis) generated by selective interference (Otto and Barton, 2001; Keightley and Otto, 2006; Roze, 2021). However, our model and most previous multilocus models on the evolution of recombination assume that epistasis is the same among all pairs of mutations, whereas empirical results tend to show that epistasis is highly variable (in sign and magnitude) among pairs of loci, and does not tend to be negative on average (de Visser and Elena, 2007; Kouyos et al., 2007; Martin et al., 2007). Otto and Feldman (1997) have shown using a 3-locus model that even when epistasis is negative on average, variable epistasis decreases the parameter space where recombination is favoured in infinite populations. It is possible that the effect of epistasis on the evolution of recombination (relative to the effect of selective interference) becomes more important when distributions of epistasis are considered. Indeed, models assuming fixed epistasis generally consider that epistasis coefficients are small in absolute value (otherwise, selection against deleterious alleles

quickly becomes unrealistically strong due to epistatic interactions with all other deleterious alleles present in the genome). Introducing distributions of epistasis would allow at least some combinations of alleles to display strongly positive or negative epistasis, without affecting the average strength of selection against deleterious alleles. Based on the results of Otto and Feldman (1997), this variance in epistasis should disfavour recombination (at least under random mating), as strong epistasis generates an important short-term cost of recombination. This is in line with the results of Vanhoenacker et al. (2018), showing that sex is disfavored in finite haploid populations under stabilizing selection around a phenotypic optimum (stabilizing selection generates a variance in epistasis, which disfavors recombination). In partially selfing populations, however, it remains unclear how distributions of the different components of epistasis (additive-by-additive, additive-by-dominance and dominance-by-dominance) would affect selection for recombination. It would thus be interesting to quantify the effect of variable epistasis on the overall strength of selection for recombination in partially selfing populations.

However, it remains difficult to determine what a realistic model of multilocus epistasis should be. Indeed, the distribution of epistasis depends on the particular shape of the fitness landscape. Empirical works that assessed the shape of fitness landscapes by measuring the fitness of all possible combinations of a small number of mutations found that they can be highly rugged, meaning that they are made of several local fitness peaks and valleys (e.g. Weinreich et al. 2006; de Visser et al. 2009; see de Visser and Krug 2014 for a review). These complex fitness landscapes are characterized by high levels of sign epistasis, where a mutation can be deleterious or beneficial according to the genetic background in which it appears, meaning that compensatory mutations are needed to reach the global fitness peak (Poelwijk et al., 2007). It appears difficult to obtain general analytical expressions for the strength of selection for recombination in such fitness landscapes. Some simulation works studied the effect of recombination on adaptation on rugged fitness landscapes, but generally consider a small number of selected sites and do not include recombination modifier loci (e.g. de Visser et al. 2009; Moradigaravand and Engelstädter 2012; Cooper and Kerr 2016). Considering simpler models that rep-

resent a single fitness peak, such as the Fisher's geometric model (FGM; Fisher 1930; Tenaillon 2014) could represent a more promising approach. Indeed, the distribution of epistatic effects on fitness generated by mutations in a generalized version of FGM has been shown to match empirical distributions of epistatic effects in *Escherichia coli* and the RNA vesicular stomatitis virus (Martin et al., 2007). This type of fitness landscape involving stabilising selection around a fixed optimum (modelled by a Gaussian or a quadratic function) typically disfavors recombination once the population is adapted. (Charlesworth, 1993; Barton, 1995). However, recombination can be favored when the mean phenotype of the population is displaced from the optimum, either due to a changing environment or to a mutational bias that tends to displace phenotypes away from the optimum (Charlesworth, 1993; Vanhoenacker et al., 2018). Indeed, when the mean phenotype of the population lags behind the optimum, the negative curvature of the fitness landscape generates negative epistasis among beneficial and among deleterious mutations (Martin et al., 2007) and recombination may be favoured as it increases the efficiency of directional selection.

However, it is not clear how the relative effect of deterministic forces and selective interference play in generating selection for recombination in this type of model, particularly in populations adapting to changing environments. It could thus be interesting to quantify these effects, especially in the case of partially selfing species, in which the effect of epistasis on selection for recombination may become important (Stetsenko and Roze, 2022). Different types of environmental changes and fitness functions could be considered, which could change the sign of epistasis in haploids (Gros et al., 2009; Blanquart et al., 2014) or the type of epistasis in diploids ($e_{a\times a}$, $e_{a\times d}$, $e_{d\times d}$; Abu Awad and Roze 2020).

Subsequently, this model could be extended to explore the evolution of recombination in more complex fitness landscapes. Indeed, models studying adaptation in rugged fitness landscapes found that recombination may either promote or hinder adaptation depending on recombination rates, the shape of the fitness landscape, population size or population structure (e.g. Weinreich and Chao 2005; Jain 2010; Altland et al. 2011;

Moradigaravand and Engelstädter 2012; Cooper and Kerr 2016). However, these models are only interested in the time taken by a population to reach a global fitness peak from a local fitness peak through a fitness valley and do not consider indirect selection at a recombination modifier locus. It could thus be of interest to quantify the strength of selection for recombination in these rugged fitness landscapes, under different forms of genetic architecture.

Genome structure and the evolution of recombination

Results from Chapter 2, 3 and 4 highlight the relevance of accounting for genome structure when studying the evolution of recombination. First, because introducing non-uniform recombination or gene density along chromosomes changes the strength of indirect selection for recombination. Indeed, we saw in Chapter 2 that indirect selection can maintain a higher number of crossovers (COs) when the position of the obligate CO is sufficiently far away from the recombination modifier locus and/or when the recombination modifier locus is located in a gene-dense region. We also saw that a higher number of COs can be maintained by the effect of drift when recombination is affected by a large number of modifier loci each having a local effect. Second, because genome structure can bias empirical measures of LD that can be performed to test predictions from recombination modifier models. Indeed, we saw in Chapter 3 that duplications that are not present in the reference genome of *Capsella rubella* probably contribute to the positive LD pattern found between deleterious mutations in the *Capsella grandiflora* population studied by Sandler et al. (2021) and ourselves. A first step before measuring LD between deleterious mutations would thus be to remove pseudo-heterozygous SNPs that may cause spurious positive LD between deleterious mutations.

However, pseudo-heterozygous SNPs can be difficult to distinguish from truly heterozygous SNPs, and the program we used only allowed us to detect the most likely pseudo-heterozygous SNPs. A possibility to avoid this problem is to use haploid data. In Chapter 4 we used sequence data from the two highly selfing species *Arabidopsis thaliana* and *Capsella orientalis* where heterozygosity is expected to be very low, and where het-

erozygous SNPs can be removed without discarding too many true heterozygous SNPs. Although no clear LD pattern was found between deleterious mutations in *C. orientalis*, positive LD was found between deleterious mutations in *A. thaliana*, similarly to what we found in *C. grandiflora*. Positive LD between deleterious mutations in *A. thaliana* is thus caused by other factors than pseudo-heterozygosity generated by mapping errors. In particular, positive LD between selected loci is expected to build up when these loci display positive epistasis on a multiplicative scale (Felsenstein, 1965). As suggested by (Ragsdale, 2022) and (Stolyarova et al., 2022) this positive epistasis could be caused by compensatory mutations within the same gene (Davis et al., 2009). As mentioned in the previous section, compensatory mutations are expected to be pervasive during adaptive walks on rugged fitness landscapes (Poelwijk et al., 2007). Positive epistasis may also result from the accumulation of deleterious mutations in a gene, either following a loss-of-function mutation or because they appear in a gene that does not impact fitness. In particular, if deleterious mutations accumulate in a pesudogene, the resulting signal of positive epistasis is spurious as the mutations that accumulate are effectively neutral regardless of their combination. Indeed, the fitness effect of mutations was inferred using SIFT that only uses the homology of protein sequence in many organisms (Vaser et al., 2016). The number of pseudogenes can be important in genomes (Harrison et al., 2002; Xiao et al., 2016) especially in the Brassicaceae that we studied, whose lineage underwent a recent (less than 20 Mya) whole-genome duplication event (Zhang et al., 2020), which may contribute to generate a significant amount of positive LD between pseudo-deleterious mutations. Although pseudogenes are annotated by bioinformatic programs or based on known pseudogenes from functional studies and removed from genome annotation, some of them may be undetected as annotation programs are not one hundred percent accurate (Xiao et al., 2016; Cheetham et al., 2020). Positive LD between deleterious mutations can also be generated by selection in spatially structured populations, due to differences in the degree of drift or inbreeding (Roze, 2009), or in the efficacy of selection among demes. It is thus difficult to disentangle the respective contributions of these potential sources of positive LD between deleterious mutations

in our dataset. This could nevertheless be achieved by including new data. For example, the data from haploid embryos from a Zambian, panmictic population of *Drosophila melanogaster* (used by Sohail et al. 2017 and Sandler et al. 2021) could be used to remove possibles effects of spatial structure and pseudo-heterozygosity. The potential effect of pseudogenes could be removed by focusing only on known functional genes, or by using available transcriptomic data in *A. thaliana* (Klepikova et al., 2016) and *D. melanogaster* (Huang et al., 2015) in order to approach the real fitness effect of mutations, although some pseudogenes may also be transcribed (Xiao et al., 2016; Cheetham et al., 2020).

Better taking the structure of genomes into account may also help us to better understand LD patterns between deleterious mutations. Indeed, if the positive LD that we observe is caused by positive epistasis, it should be dominated by interactions between mutations within genes and/or protein-coding domains, since we measured LD over short distances. Protein sequences are typically made up of several functional domains, corresponding to conserved and functionally independent elements (Bagowski et al., 2010). Using data from several human populations, Ragsdale (2022) found no difference in LD between synonymous and missense mutations when averaging over all pairs of mutations within genes. However, he found higher positive LD between missense mutations within the same conserved domain compared to synonymous mutations (while the opposite pattern was found outside conserved domains), which was interpreted as evidence for positive epistasis between missence mutations affecting the same domain. At a larger scale, gene products (RNA or proteins) do not interact randomly and large gene interaction networks (e.g. in yeast Costanzo et al. 2019) show that genes preferentially interact within clusters (or modules) at different scales because they are involved in the same metabolic pathway, bioprocess or cellular compartment. Nevertheless, it remains unclear whether an overall relationship between epistasis and genetic proximity exists within genomes. Still, cases are known of changes in recombination that may have occurred in response to specific epistatic interactions. For example, sets of coadapted alleles can be maintained on the same genetic background by the evolution of recombination arrest generating supergenes, as recombination is strongly disfavoured between

them (Thompson and Jiggins, 2014). This recombination suppression can be mediated by different mechanisms, including structural changes such as chromosomal inversions (Schwander et al., 2014). On the contrary, higher recombination rates may be favoured under forms of selection favouring diversity at the level of a set of genes, such as between immunity related genes (Choi et al., 2016; Fulton et al., 2016).

Overall, selection for recombination is thus expected to vary within and between genes because of the joint effects of genome structure and epistatic interactions. Therefore, the discrepancy between patterns of LD that should favour recombination according to modifier models and what is observed in genomic data could be resolved by taking into account the joint effect of genome structure and epistasis. In particular, an extension of the present work could consist in unraveling the pattern of positive LD between deleterious mutations at different genomic scales (within domains, between domains, within genes, between genes) as they correspond to different levels of interaction between mutations and are expected to generate different indirect selective pressure for recombination. In order to make finer predictions on the expected recombination rates at different genomic scales, models that explicitly account for genome structure and modularity of epistatic interactions will thus be needed. Additionally, recombination landscapes at different genomic scales – in particular at the scale of genes – could be used to test those predictions.

However, as mentioned above, a central element that should not be overlooked when studying indirect selection for recombination is the genetic architecture of recombination rate variation. Indeed, as seen in Chapter 2, the number and position of recombination modifier loci relative to the position of COs they encode and gene-dense regions greatly affect the overall strength of selection for recombination. Although a growing number of recombination modifier genes are described, affecting various pathways of CO formation (e.g. in plants; Wang and Copenhaver 2018) it is still difficult to have a clear picture of their relative contribution to recombination rate variation within natural populations. Therefore, more empirically work is needed in that direction, in order to better understand the sign and magnitude of direct and indirect selective pressures that may act on

these genes.

Selfing and the evolution of recombination

In Chapter 1, we found that, under realistic parameter values, indirect selection for recombination generated by deleterious mutations generally increases as the selfing rate increases, and then vanishes as the selfing rate approaches 1. This result stems from the fact that selfing reduces the efficiency of recombination, which increases the strength of indirect selection for recombination as long as the selfing rate is not too high, but suppresses indirect selection when the selfing rate approaches 1, since heterozygosity is so rare that recombination has no effect in this case. This could help to explain two empirical observations: (i) the positive correlation between selfing rates and recombination rates in closely related species of plants (Roze and Lenormand, 2005; Ross-Ibarra, 2007; Nordborg et al., 2005) and (ii) the fact that even highly selfing species retain small rates of outcrossing (e.g. $\approx 5\%$ in *A. thaliana* Stenøien et al. 2005; Bomblies et al. 2010) so that they can recombine.

Concerning the first observation, it would be of interest to obtain more detailed results on recombination rates in species with intermediate selfing rates (in particular, our results often predict similar recombination rates under outcrossing and moderate selfing, unless some forms of negative epistasis are present). In the studies compiled in the metaanalysis of Roze and Lenormand (2005), selfing rate estimates are generally not available, species being classified into broad classes (e.g., outcrossing, mixed-mating, predominantly selfing). Moreover, effective selfing rates can be highly variable among populations within a same species (Whitehead et al., 2018) probably due to a mix of genetic, demographic and ecological factors (Goodwillie et al., 2005; Cheptou et al., 2008), so that it is difficult to define a selfing rate at the scale of the whole species. Moreover, the relationship between recombination and selfing rate could be updated with new data from genetic maps in plants, as the correlation has mainly been obtained from cytological data on the number of chiasmata per bivalent, that are only indirect measures of recombination rates. This correlation seems to be retrieved using genetic map data at the

scale of Angiosperms, when comparing the whole-genome recombination rate of species belonging to different selfing categories (Brazier & Glémin, personal communication). It could be interesting, using this dataset, to investigate whether the correlation between recombination and selfing is also retrieved at smaller phylogenetic scale (i.e. between species in a same genus) as it was found with chiasma count data (Roze and Lenormand, 2005).

Concerning the second observation – that even highly selfing species retain small rates of outcrossing – one can note that in highly selfing species, higher levels of genetic shuffling could be achieved either by increasing the number of COs during meiosis or by increasing the outcrossing rate. However, different selective forces act on selfing rate modifiers (e.g. reproductive assurance, automatic advantage, inbreeding depression, Lande and Schemske 1985; Busch and Delph 2012; Abu Awad and Roze 2020) and on recombination rate modifiers, and it would thus be of interest to model the joint evolution of outcrossing and recombination rates in order to quantify their relative contribution on the evolution of recombination in highly selfing species.

Although our model draws clear expectations concerning recombination rates in selfing populations that seem to match with current data, it is less straightforward to interpret patterns of LD between deleterious mutations in highly selfing species from our model. Indeed, under high selfing rates, LD between deleterious mutations may be negative due to selective interference, or positive due to correlations in homozygosity caused by partial selfing. Because selective interference increases as effective population size decreases, it should dominate in relatively small populations and generate negative LD among deleterious mutations under high selfing, whereas in large populations it may be positive under high selfing as the deterministic effect of correlations in homozygosity may dominate. Although no clear LD pattern was found in the highly selfing *Capsella orientalis*, we found positive LD between deleterious mutations in *A. thaliana*. For *A. thaliana*, this result could thus be in line with the expectation of LD in large highly selfing populations in the absence of epistasis whereas for *C. orientalis* the high level of noise in the results makes them more difficult to interpret.

Individuals from the relict group of *A. thaliana* that we studied are found in patchy habitats (The 1001 Genomes Consortium, 2016) and the population is likely highly spatially structured. Yet, spatial structure has opposing effects on LD: it decreases effective population size thus increasing selective interference (Martin et al., 2006) generating negative LD under high selfing, but the variance in inbreeding among demes creates correlations in homozygosity generating positive LD (Roze, 2009). However, it is unclear whether the stochastic effect generating negative LD or the deterministic effect generating positive LD should dominate in spatially structured populations under high selfing.

Positive LD between deleterious mutations could also be caused by positive epistasis as discussed previously. However, with positive epistasis, positive LD between deleterious mutations is expected to increase roughly proportionally with selfing (Stetsenko and Roze, 2022) so that higher positive LD is expected between deleterious mutations under high selfing compared to under outcrossing. This further complicates the effect of selfing on LD patterns between deleterious mutations in the presence of epistasis and spatial structure in selfing species. New models would thus be required to better predict the sign and magnitude of LD between deleterious mutations due to the interplay between selfing, population size, spatial structure and epistasis.

Are recombination rates always adaptive?

Throughout this thesis we explored indirect selective forces acting on recombination, considering that despite potentially important constraints on the number and position of COs, some room can be left for indirect selection on recombination rates. However, we have seen in Chapter 2 that when considering a large number of recombination modifier loci evenly spaced along the genome and having only a local effect on recombination rates, the total map length of the chromosome is primarily determined by the fitness cost of COs and the number of recombination modifier loci. Indeed, in this case the direct fitness cost of recombination is shared between all modifier loci and direct selection against recombination at each modifier locus becomes of the same order of magnitude (or lower)

as the effect of drift when the number of modifier loci is sufficiently large. Indirect selection acting on each modifier is also reduced, due to the fact that the modifier only affects a small proportion of selected loci. The local map length coded by each modifier locus is thus submitted to random fluctuations, increasing the variance in map length along the chromosome. This also increases the total map length of the chromosome, as map length cannot be negative. The effect of drift on the value of polygenic traits (making selection less efficient at maintaining traits around their optimum value) has been proposed as the "drift-barrier" hypothesis to explain the negative correlation between effective population size and mutation rate observed at a broad phylogenetic scale (Lynch et al., 2016; Lynch, 2020). Under this hypothesis, the lower mutation rates observed in unicellular or small eukaryotes compared with large multicellular organisms would be explained by the lower N_e of these larger organisms, decreasing the efficiency of selection on the DNA repair machinery to maintain replication fidelity (Sung et al., 2012). A similar negative correlation between recombination rate and N_e has been identified among eukaryotes (Buffalo, 2021) and the drift-barrier hypothesis could be an alternative explanation to the effect of selective interference being stronger in populations with smaller N_e. However, many other traits that correlate with N_e (the mutation rate in particular; Sung et al. 2012) also have to be taken into account in such broad scale correlations. Nevertheless, the drift-barrier hypothesis could also represent an alternative explanation to the positive correlation between recombination rate and selfing rate in plants (Roze and Lenormand, 2005) as N_e is decreased by selfing and by the transition from outcrossing to selfing often associated with demographic bottlenecks (Barrett et al., 2014). It would thus be interesting to investigate this alternative mechanism in more detail, by exploring the effect of selfing on the evolution of recombination when chromosome map length is controlled by many loci of small effects.

Conclusion

Considering variable epistatic interactions between mutations could impact the conditions under which recombination is favoured under random mating or selfing. However,

the distribution of epistasis depends on the underlying fitness landscape, and few models have been interested in the evolution of recombination under those conditions. New models would thus be needed to explore the relative importance of deterministic and stochastic forces generating selection for recombination under various fitness landscapes and environmental changes. A first step would be to consider a simple fitness landscape with a single fitness peak that could then be extended to rugged fitness landscapes with multiple peaks.

Genome structure is an important factor that needs to be taken into account in the evolution of recombination as it can affect the strength of selection for recombination, and bias the measures of LD from genomic data, generating a discrepancy between patterns of LD that should favour recombination according to modifier models and what is observed in genomic data. When measuring LD between deleterious mutations, adopting common analysis pipelines would help avoid various biases and would make patterns of LD more comparable across datasets and help to interpret them in terms of selective forces. More work is also needed in the quantitative genetics of recombination rate variation in a wide variety of organisms to get a clearer picture of the genetic architecture of recombination modifiers, and quantify their potential to respond to selection. Models that explicitly account for genome structure and modularity of epistatic interactions would be needed in order to make predictions on the expected recombination rates at different genomic scales.

Indirect selective forces acting on recombination may explain the correlation between recombination rates and selfing rates in plants, mainly found from chiasma count data. This correlation could be updated and refined using more recent data on genetic maps in plants and other organisms. Because changing the level of genetic shuffling in selfing species could also proceed by the evolution of outcrossing, the joint evolution of selfing and recombination could be explored to quantify their relative contribution. Moreover, it is difficult to interpret patterns of LD between deleterious mutations in highly selfing species that display spatial structure and epistasis between mutations. Therefore, modeling the interplay between selfing, population size, spatial structure and epistasis

in generating LD between deleterious mutations will be needed.

Finally, an alternative explanation for the positive correlation between recombination rates and selfing rates in plants is the drift-barrier hypothesis, whereby populations with lower effective sizes could display higher recombination rates when recombination is encoded my many loci of small effect. Since selfing decreases effective population size, recombination rates could be higher in selfing species due to this mechanism. It would thus be interesting to explore the effect of selfing on the evolution of recombination when chromosome map length is controlled by many loci of small effects, in order to quantify the relative contribution of drift and indirect selection in generating a correlation between recombination and selfing.

Appendix

Appendix S1

Supplementary figures from Chapter 1

SUPPLEMENTARY FIGURES

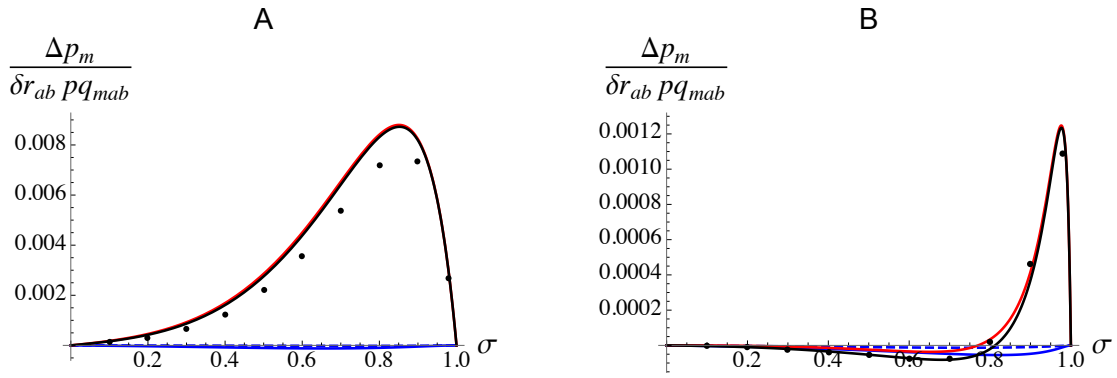


Figure S1. Same as Figure 1 with $e_{\mathrm{a\times d}} = -0.001$, and $r_{ma} = r_{ab} = 0.01$ (A), $r_{ma} = r_{ab} = 0.1$ (B). Other parameter values are as in Figure 1.

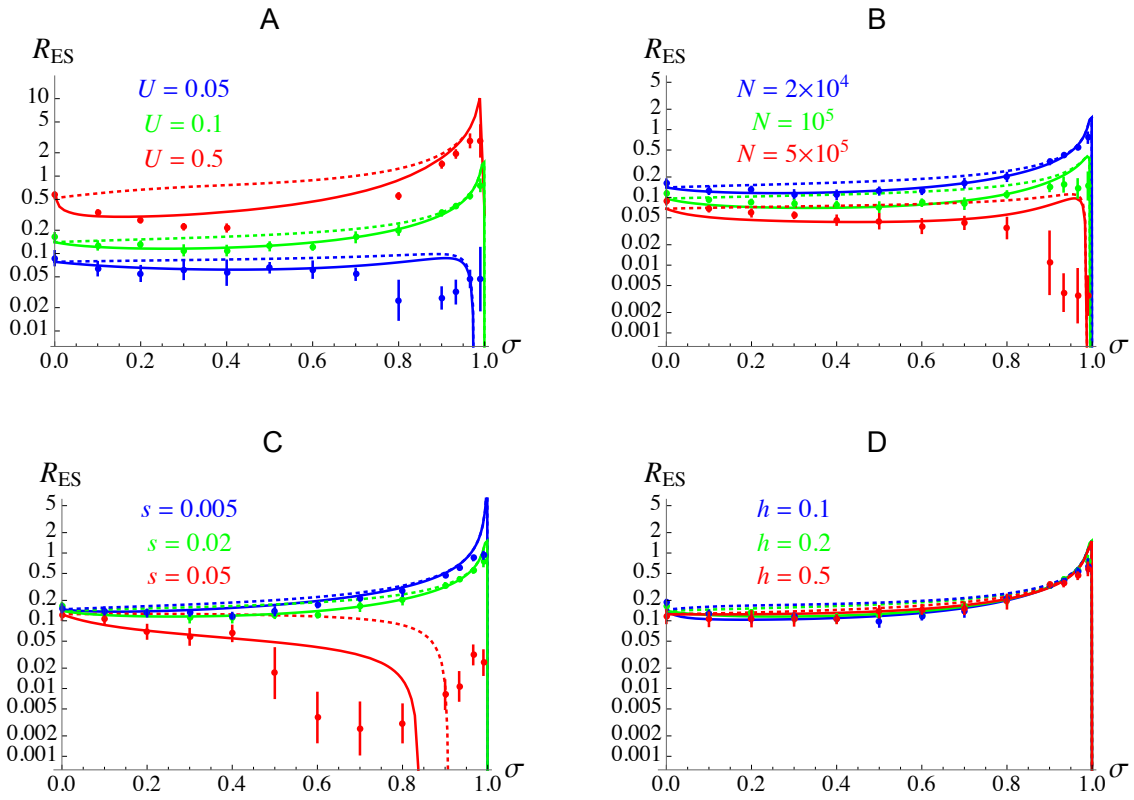


Figure S2. Same as Figure 4, with R on a log scale.

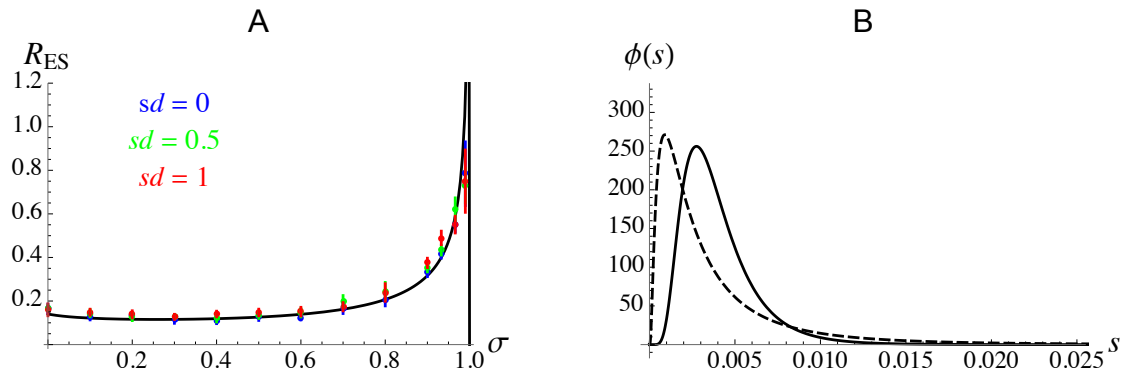


Figure S3. A: effect of the selfing rate σ on the evolutionarily stable map length R_{ES} for different values of the standard deviation sd of the log-normal distribution of selection coefficients s of deleterious alleles (in the simulations). B: distribution of s for $sd = 0.5$ (solid) and $sd = 1$ (dashed). The mean selection coefficient $\overline{s}$ is set at 0.02, and default parameter values are as in Figure 3 with $c = 0.001$. The analytical prediction (curve) in A is the same as in Figure 3 for $c = 0.001$ (fixed s).

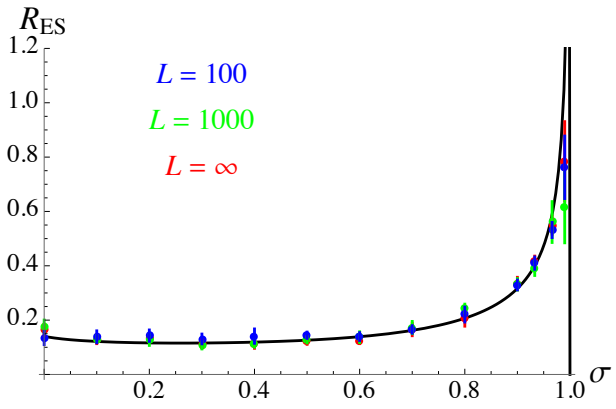


Figure S4. Effect of the selfing rate σ on the evolutionarily stable map length R_{ES} for different numbers of loci L at which deleterious alleles may occur. In simulations with finite L, loci are uniformly spaced along the chromosome, each locus mutating at a rate U/L. Parameter values are as in Figure 3 with $c = 0.001$. The analytical prediction (curve) is the same as in Figure 3 for $c = 0.001$ (infinite number of loci).

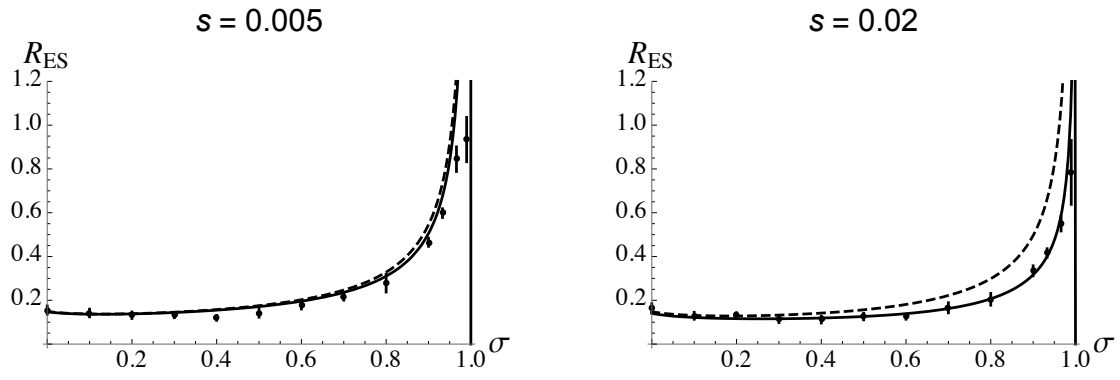


Figure S5. Effect of the selfing rate σ on the evolutionarily stable map length R_{ES} for $s = 0.005$ and $s = 0.02$, and other parameter values as in Figure 4. Solid curves are the same as in Figure 4C, corresponding to the predictions obtained by solving $s_{\mathrm{direct}} + s_{\mathrm{det}} + s_{\mathrm{HR}} = 0$ for R, where s_{det} and s_{HR} are obtained by integrating the expressions for the strength of indirect selection generated by deterministic effects (for s_{det}) and by the Hill-Robertson effect (for s_{HR}) over the genetic map (see Supplementary Material). Dashed curves correspond to the predictions obtained using the same method, but replacing s_{HR} by the approximation given by equation 22: $s_{\mathrm{HR}} \approx 1.8\delta R\, U^2 / \left[N_{\mathrm{e}} R^3 \left(1 - F\right)^2\right]$.

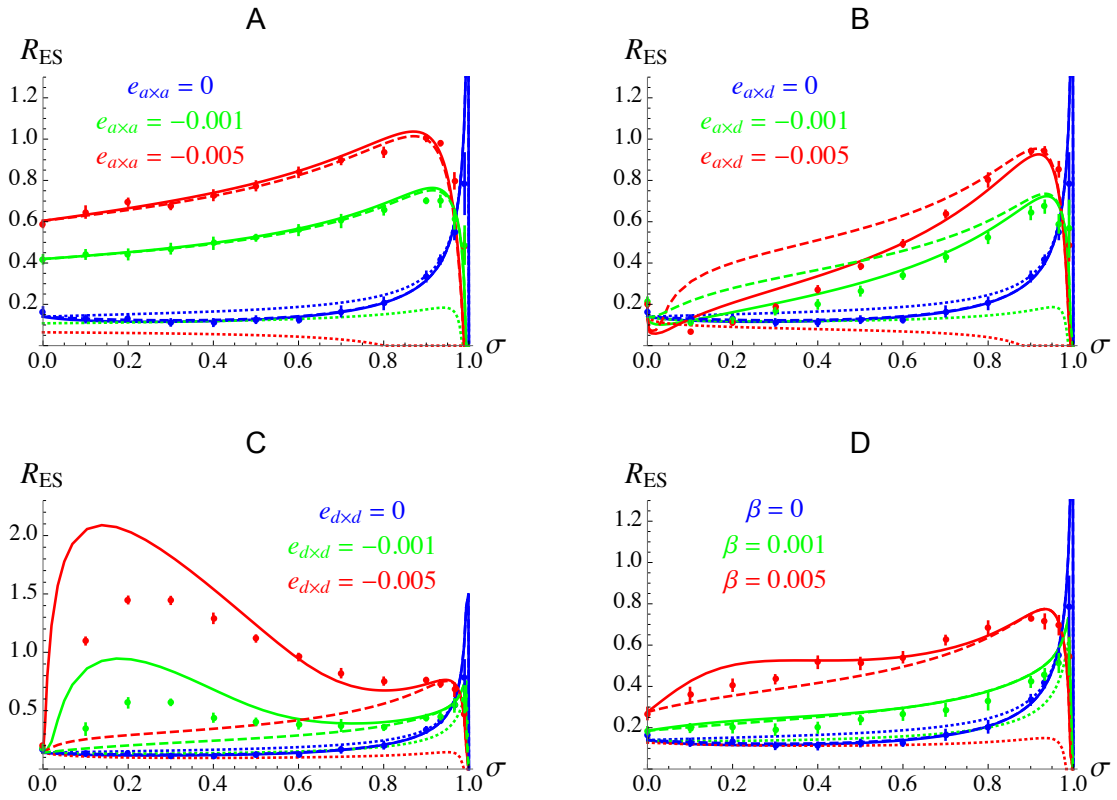
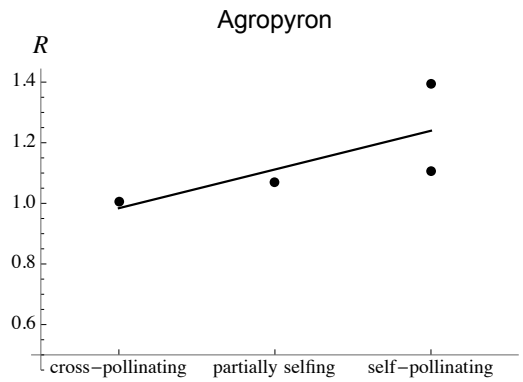
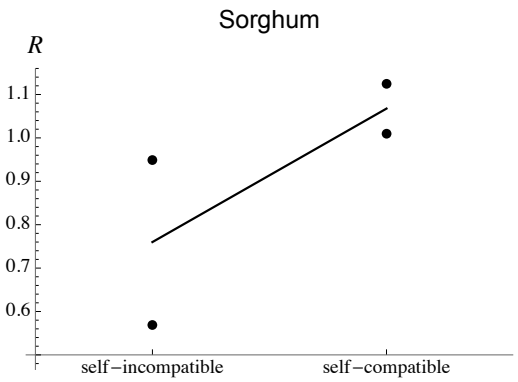
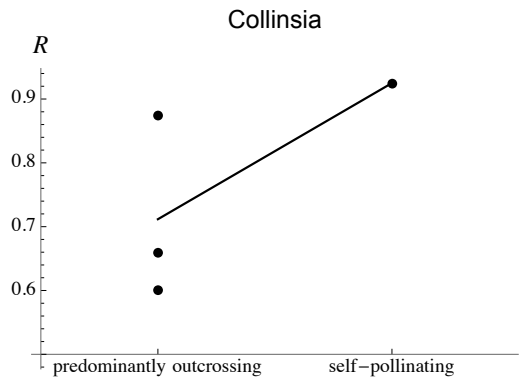


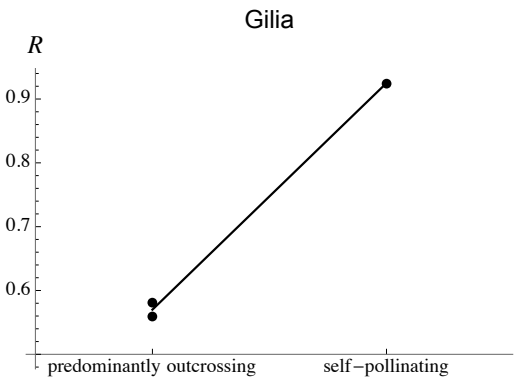
Figure S6. Same as Figure 5, the dashed curves showing the predictions obtained for R_{ES} when ignoring the term generated by $D^0_{mab,ab}$ in the expression for the deterministic source of indirect selection acting on the modifier locus. The difference between the dotted and dashed curves therefore corresponds to the effect of deterministic terms generated by D_{ab}, while the difference between the dashed and solid curves corresponds to the effect of deterministic terms generated by $D^0_{mab,ab}$.

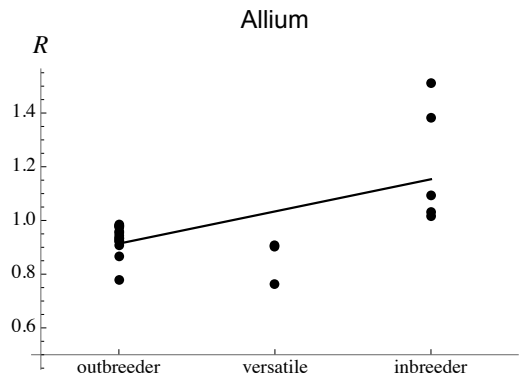
Figure S7 (next page). Cytological data from different families of angiosperms, used to generate Figure 1 in Roze & Lenormand (2005). Each point shows the average chromosomal map length (on y-axes), that is, half the number of chiasmata per bivalent (observed during male meioses) in a given species. The different mating system categories on x-axes correspond to those given in the original articles (see legend of Figure 1 in Roze & Lenormand 2005 for references).

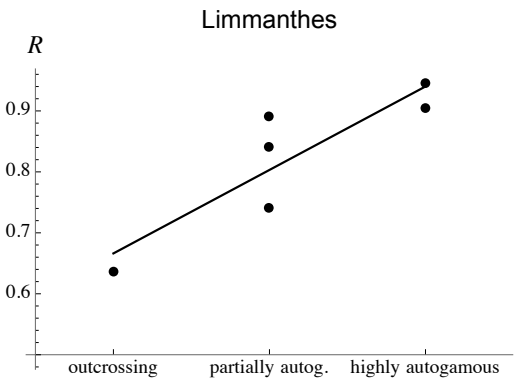

Agropyron
R
1.4
1.2
1.0
0.8
0.6
cross−pollinating
partially selfing
self−pollinating

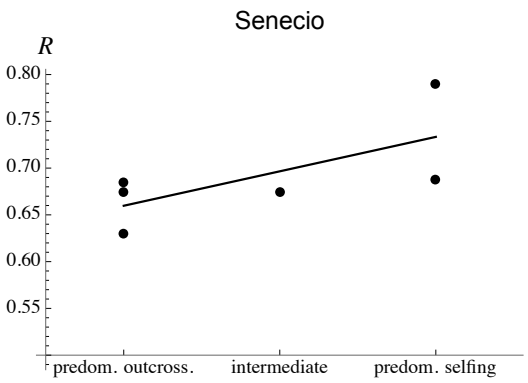

Sorghum
R
1.1
1.0
0.9
0.8
0.7
0.6
self−incompatible
self−compatible

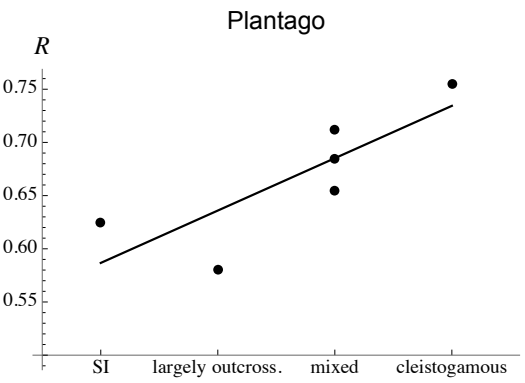

Collinsia
R
0.9
0.8
0.7
0.6
predominantly outcrossing
self−pollinating


Gilia
R
0.9
0.8
0.7
0.6
predominantly outcrossing
self−pollinating


Allium
R
1.4
1.2
1.0
0.8
0.6
outbreeder
versatile
inbreeder


Limmanthes
R
0.9
0.8
0.7
0.6
outcrossing
partially autog.
highly autogamous


Senecio
R
0.80
0.75
0.70
0.65
0.60
0.55
predom. outcross.
intermediate
predom. selfing


Plantago
R
0.75
0.70
0.65
0.60
0.55
SI
largely outcross.
mixed
cleistogamous

Appendix S2

Supplementary figures and tables from Chapter 3

Expressing the composite linkage disequilibrium in terms of allele frequencies and indicative variables

Following the multilocus formalism from Barton and Turelli (1991) and Kirkpatrick et al. (2002), the two-locus moments D_{ab} and $D_{a,b}$ can be expressed as :

$$D_{ab} = \mathbb{E}[\frac{\tilde{D}_{ab,\varnothing} + \tilde{D}_{\varnothing,ab}}{2}] = \frac{1}{2}\mathbb{E}[(X_{a,\varnothing} - p_a)(X_{b,\varnothing} - p_b) + (X_{\varnothing,a} - p_a)(X_{\varnothing,b} - p_b)] \quad \text{(S2.1)}$$

$$D_{a,b} = \mathbb{E}[\frac{\tilde{D}_{a,b} + \tilde{D}_{b,a}}{2}] = \frac{1}{2}\mathbb{E}[(X_{a,\varnothing} - p_a)(X_{\varnothing,b} - p_b) + (X_{b,\varnothing} - p_b)(X_{\varnothing,a} - p_a)] \quad \text{(S2.2)}$$

with $\tilde{D}_{\mathbb{U},\mathbb{V}}$ the association among the set of loci $\mathbb{U}$ and $\mathbb{V}$, accounting for the origin of the alleles (maternally or paternally inherited) and $\mathbb{E}$ denoting the average over all individuals in the population. Given that $\Delta_{ab} = D_{ab} + D_{a,b}$ and after developing and factoring we obtain:

$$\Delta_{ab} = \frac{1}{2}\mathbb{E}[(X_{a,\varnothing} + X_{\varnothing,a} - 2p_a)(X_{b,\varnothing} + X_{\varnothing,b} - 2p_b)], \quad \text{(S2.3)}$$

with $\mathbb{E}$ denoting the average over all individuals, the indicative variable $X_{x,\varnothing}$ that equals 1 if allele x is present on the maternally inherited chromosome (and 0 otherwise) and $X_{x,\varnothing}$ that equals 1 if allele x is present on the paternally inherited chromosome (and 0 otherwise). The sum $X_{x,\varnothing}+X_{x,\varnothing}$ can be equal to 0,1 or 2 if the site is homozygous for the ancestral allele, heterozygous or homozygous for the derived allele x, respectively. Equation S2.3 can also be retrieved from the expression in terms of genotypes (Equation 3.1; Weir 1996) using the fact that:

$$D_{ab} = p_{ab} - p_a p_b, \quad \text{(S2.4)}$$

$$D_{a,b} = p_{a,b} - p_a p_b, \quad \text{(S2.5)}$$

$$p_{ab} = \frac{1}{2}\mathbb{E}[X_{a,\varnothing}X_{b,\varnothing} + X_{\varnothing,a}X_{\varnothing,b}] = \frac{1}{n}(n_1 + \frac{1}{2}n_2 + \frac{1}{2}n_3 + \frac{1}{2}n_{4a}), \quad \text{(S2.6)}$$

$$p_{a,b} = \frac{1}{2}\mathbb{E}[X_{a,\varnothing}X_{\varnothing,b} + X_{b,\varnothing}X_{\varnothing,a}] = \frac{1}{n}(n_1 + \frac{1}{2}n_2 + \frac{1}{2}n_3 + \frac{1}{2}n_{4b}), \tag{S2.7}$$

with p_{ab} the frequency of ab configurations, $p_{a,b}$ the frequency of a, b configurations, n the total number of genotypes, n_1 the number of double homozygotes for alleles a and b (ab/ab), n_2 the number of homozygotes for allele a and heterozygotes for allele b (aB/ab), n_3 the number of heterozygotes for allele a and homozygotes for allele b (Ab/ab), n_4a the number of double heterozygotes with alleles a and b in coupling (ab/AB) and n_4b the number of double heterozygotes with alleles a and b in repulsion (Ab/aB). Note that the total number of double heterozygotes is $n_4 = n_{4a} + n_{4b}$.

Table S2.1: List of accessions of *Capsella grandiflora* with their ID from Josephs et al. (2015), SRA code, number used in the current paper, number of mapped reads and mean coverage after mapping and mean coverage of SNPs after filtering. Individuals shaded in grey correspond to diverged individuals that were removed from LD computing.

ID	SRA	Number	Number of mapped reads	Mean coverage after mapping (X)	Mean coverage of SNPs after filtering (X)
100O	SRR2065171	1	67,778,253	39.6265	38.9136
101S	SRR2065172	2	71,652,964	43.9521	38.8951
102K	SRR2065173	3	55,040,504	33.4317	29.9501
103K	SRR2065174	4	71,096,939	43.0686	40.5397
105G	SRR2065175	5	40,515,048	24.4133	23.1359
106A	SRR2065176	6	74,503,090	43.7571	41.785
107B	SRR2065177	7	72,820,349	43.9293	41.7436
108B	SRR2065178 SRR2065179	8	67,768,983	39.0761	35.4101
109E	SRR2065180	9	67,546,014	40.9459	37.6479
10M	SRR2065181	10	64,989,506	37.0686	35.6434
110M	SRR2065182	11	52,643,683	30.5764	30.3862
111C	SRR2065183 SRR2065184	12	143,035,188	83.04	68.3479
112H	SRR2065185	13	71,583,533	43.2113	40.8899
113A	SRR2065186	14	80,500,398	48.9141	43.5456
115L	SRR2065188	15	57,505,563	34.4798	33.3936
116L	SRR2070888	16	50,457,336	29.4382	29.5714
117R	SRR2065189	17	42,772,970	25.6085	24.8085
118W	SRR2065190	18	74,411,924	44.1095	43.7323
119M	SRR2070889	19	76,444,706	46.4393	41.4909
11G	SRR2065191	20	74,806,894	45.0182	40.4033
120C	SRR2070890	21	56,412,844	33.8994	32.5833
121B	SRR2065192	22	47,003,390	27.9812	6.70313
122	SRR2070891	23	43,741,144	25.9645	19.5828
123D	SRR2065193	24	67,396,367	40.0744	34.8186
124Y	SRR2065194 SRR2065195	25	63,814,643	36.8642	37.6905
125X	SRR2065196	26	59,939,496	34.8289	33.5542
126Z	SRR2065197	27	60,874,245	35.3933	35.679
128P	SRR2065198	28	50,330,242	29.8525	28.9122
129Y	SRR2065199	29	59,748,060	35.8645	35.1082
130	SRR2070892	30	71,681,229	41.7704	40.5615
131X	SRR2065200	31	42,373,447	25.33	22.7073
132L	SRR2065201	32	70,204,351	42.5192	41.3182
133V	SRR2065202	33	82,680,837	47.9079	47.9725
135F	SRR2065203	34	68,212,063	41.3666	39.9496
136R	SRR2065204	35	56,751,296	34.4575	34.9401
137Q	SRR2065205	36	54,998,156	31.8002	29.5052
138Q	SRR2065206	37	58,822,419	35.5121	28.9459
139G	SRR2065207	38	50,305,503	29.3703	29.2626

ID	SRA	Number	Number of mapped reads	Mean coverage after mapping (X)	Mean coverage of SNPs after filtering (X)
13L	SRR2065208	39	66,729,805	40.3389	41.1133
140S	SRR2065209	40	76,818,928	46.5473	47.3456
141P	SRR2065210	41	83,122,625	50.5074	49.5207
142B	SRR2065211	42	59,492,174	36.0179	34.982
143Y	SRR2065212	43	79,023,415	47.3287	43.0101
144Q	SRR2065213	44	59,298,700	34.2662	33.1329
145A	SRR2070893	45	65,754,737	38.299	35.6665
146J	SRR2065214	46	49,366,218	28.6793	25.9689
147Y	SRR2065215	47	69,417,464	42.0787	42.2985
148Z	SRR2065216	48	67,968,689	41.2659	41.5584
149	SRR2070894	49	73,208,253	42.3707	40.701
14R	SRR2065217	50	63,299,238	36.9454	35.5008
151D	SRR2065218	51	77,462,657	46.6692	47.6571
152P	SRR2065219	52	55,738,604	33.0613	9.32303
153G	SRR2065220	53	42,178,619	25.0774	5.67162
154P	SRR2065221	54	68,051,515	40.8469	40.9878
155T	SRR2065222	55	22,682,911	13.5619	13.016
156R	SRR2065223	56	80,085,198	47.2397	48.0434
157x	SRR2065224	57	83,113,379	49.9481	52.9682
158J	SRR2065225	58	88,667,257	52.3153	50.3465
15U	SRR2065226	59	65,930,250	40.0612	37.384
160R	SRR2065227	60	53,656,225	31.8291	28.2538
161A	SRR2065228	61	81,200,362	48.6511	46.0047
162R	SRR2065229	62	76,273,152	45.8311	43.5795
163J	SRR2065230 SRR2065231	63	41,555,559	24.857	11.8835
165I	SRR2065232 SRR2065233	64	86,240,512	49.9489	25.7551
166	SRR2070895	65	77,213,914	45.9567	44.6707
167Q	SRR2065234	66	55,064,212	32.8007	32.9024
168	SRR2070896	67	64,985,455	38.7091	37.5054
16A	SRR2065235 SRR2065236	68	128,526,954	75.3131	34.4697
170V	SRR2065237 SRR2065238	69	101,435,487	58.9531	28.5943
172	SRR2070897	70	83,336,374	50.4784	47.1442
173Z	SRR2065239 SRR2065240	71	130,863,213	76.0802	36.1616
174I	SRR2065241	72	86,513,634	50.3249	47.6696
175D	SRR2065242 SRR2065243	73	107,408,534	64.3758	32.1062
176X	SRR2065244	74	42,929,153	25.6369	21.7436
177K	SRR2065245	75	69,027,380	41.2525	40.3573
178M	SRR2065246	76	92,120,628	55.7699	54.9876
179E	SRR2070898	77	45,270,540	27.4644	25.3947
17Q	SRR2065247	78	79,781,641	47.1432	48.2977

ID	SRA	Number	Number of mapped reads	Mean coverage after mapping (X)	Mean coverage of SNPs after filtering (X)
180T	SRR2070899 SRR2070900	79	135,713,967	81.2489	36.0529
181Y	SRR2065248	80	68,588,317	40.8242	39.8429
182Q	SRR2065249 SRR2065250	81	65,624,998	37.856	36.7498
183E	SRR2065251	82	45,199,313	26.4439	27.0142
184M	SRR2065252 SRR2065253	83	64,417,831	37.2486	36.6804
186W	SRR2065254	84	82,085,974	49.7373	52.0428
187K	SRR2065255	85	57,803,629	34.4179	28.187
189B	SRR2070901	86	60,620,390	35.7028	34.3665
18C	SRR2065256	87	70,251,238	41.5032	40.9202
190U	SRR2065257	88	45,087,744	26.9096	4.76871
192H	SRR2065258	89	53,042,382	31.2101	30.3332
193N	SRR2065259	90	40,187,845	23.9157	10.3354
195L	SRR2065260	91	38,309,552	22.344	22.5537
196X	SRR2070903	92	45,285,316	26.469	27.0444
197L	SRR2065261	93	83,198,562	47.3261	47.3051
198A	SRR2065262	94	76,537,570	45.1522	44.7034
199	SRR2070904	95	64,112,895	37.8811	36.712
19Y	SRR2065263	96	63,152,265	38.2398	40.4914
1N	SRR2065264 SRR2065265	97	123,532,118	72.4085	36.0577
2	SRR2070905	98	79,710,259	46.6854	45.512
200U	SRR2065266	99	69,154,039	40.9279	39.8612
202I	SRR2065267	100	63,451,416	38.1595	35.4487
203A	SRR2065268	101	46,773,953	28.3813	25.1764
204O	SRR2065269	102	122,273,541	74.2437	62.6971
207D	SRR2070906	103	64,959,938	37.8733	36.9108
208	SRR2070907	104	70,009,972	41.3305	41.5862
209P	SRR2070908	105	81,473,140	49.291	51.0841
20B	SRR2065270	106	73,916,835	44.7042	44.918
23Z	SRR2065271	107	58,846,003	34.3074	34.7468
24F	SRR2065272	108	70,901,904	42.1701	40.0409
25I	SRR2065273	109	30,966,308	18.4625	3.72872
26F	SRR2065274	110	81,127,053	49.0621	49.8611
27K	SRR2065275	111	70,841,232	41.9656	38.0147
28C	SRR2065276	112	67,652,969	41.0223	38.2379
29A	SRR2065277	113	52,360,150	30.168	31.2883
3	SRR2070909	114	70,868,057	42.4081	40.2711
30Y	SRR2065278	115	57,257,392	33.935	34.9868
31T	SRR2065279	116	66,956,787	40.5396	39.6854
32	SRR2070910	117	63,974,853	36.8332	36.3046
33E	SRR2065280	118	66,269,515	40.136	39.7893
34D	SRR2065281	119	65,381,523	38.8589	37.5032
35F	SRR2065282	120	78,697,770	46.7425	43.332

ID	SRA	Number	Number of mapped reads	Mean coverage after mapping (X)	Mean coverage of SNPs after filtering (X)
36H	SRR2065283	121	63,378,792	38.2639	39.0714
37	SRR2070911	122	51,697,048	29.7453	29.7603
38B	SRR2065284	123	74,775,822	44.6129	46.4256
39N	SRR2065285	124	43,221,649	24.772	25.1637
41J	SRR2065286	125	73,532,236	43.6266	41.1315
42	SRR2070912	126	64,804,156	39.1906	38.811
43F	SRR2065287	127	61,837,342	35.0002	28.9142
44W	SRR2065288	128	69,967,806	41.3651	23.2308
45	SRR2070913	129	68,586,777	40.6128	37.9552
46	SRR2070914	130	63,373,419	36.4663	35.6593
470x19	SRR2070915	131	73,794,006	43.3242	28.2146
47K	SRR2065289	132	77,879,463	47.0856	45.219
48	SRR2070916	133	61,847,174	35.8292	34.3142
49Z	SRR2065290	134	68,981,492	41.5254	38.9901
4S	SRR2065291 SRR2070917	135	86,021,503	48.9558	43.2052
50S	SRR2065292	136	53,003,268	31.8066	32.5276
51N	SRR2065293	137	62,167,457	35.9878	35.1954
52A	SRR2065294	138	69,974,028	41.4568	38.7089
53	SRR2070918	139	58,899,793	34.0147	32.2693
54T	SRR2065295	140	48,983,666	28.1313	29.0827
55Y	SRR2065296	141	74,374,540	42.797	41.1833
58Y	SRR2065297	142	38,233,744	23.1927	23.8191
59N	SRR2070919	143	95,923,790	57.7721	50.6596
5S	SRR2065298	144	56,562,999	33.1306	31.7725
60T	SRR2070920	145	41,103,610	24.3865	22.4176
61G	SRR2065299	146	31,608,046	19.5132	7.15663
63D	SRR2065300	147	89,221,512	53.8929	52.5131
64E	SRR2065301 SRR2065302	148	114,132,152	68.2621	65.943
65T	SRR2065303	149	27,506,803	15.8318	15.8143
66X	SRR2065304	150	52,376,358	30.2621	30.2495
67R	SRR2065305	151	37,006,004	23.291	20.1181
6H	SRR2070921 SRR2070922	152	109,129,547	61.2037	56.0368
70V	SRR2065306	153	65,599,439	38.091	34.9359
71W	SRR2065307	154	58,970,708	34.4069	34.5394
72G	SRR2065308	155	71,155,909	43.0962	38.483
74M	SRR2065309	156	70,228,412	42.526	38.5986
75H	SRR2065310	157	56,901,796	34.2997	32.6267
76J	SRR2065311	158	52,906,046	31.3695	32.3992
78M	SRR2065312	159	65,199,534	37.8668	35.1942
79G	SRR2065313	160	72,256,279	41.602	36.2624
7K	SRR2065314 SRR2070923	161	61,484,995	34.8942	31.5093
80G	SRR2065315	162	24,004,415	13.576	13.5362

ID	SRA	Number	Number of mapped reads	Mean coverage after mapping (X)	Mean coverage of SNPs after filtering (X)
81J	SRR2065316	163	42,228,487	24.708	23.1726
82J	SRR2065317	164	23,140,063	13.1211	12.7753
83Z	SRR2065318	165	51,511,789	30.558	31.7037
85I	SRR2065319	166	79,359,602	47.8465	45.5222
86I	SRR2065320	167	73,299,422	44.4755	44.3322
88	SRR2070924	168	68,610,724	40.2406	38.111
89L	SRR2065321	169	79,240,635	48.0663	47.7996
8U	SRR2065322	170	71,987,770	42.7442	40.3343
90E	SRR2065323	171	56,934,227	33.6398	33.7031
91C	SRR2065324	172	58,279,751	34.4491	34.4027
91x35	SRR2070925	173	75,035,018	45.1561	46.6624
92H	SRR2065325	174	156,190,901	93.1882	86.4507
	SRR2065326				
93M	SRR2065327	175	103,788,836	62.2141	64.942
94U	SRR2065328	176	52,791,072	31.3697	25.3255
95O	SRR2065329	177	76,074,792	46.1104	43.8904
96O	SRR2065330	178	70,086,066	42.3797	40.5477
97C	SRR2065331	179	100,347,463	57.7404	54.8625
	SRR2065332				
98S	SRR2065333	180	55,438,898	32.4998	30.0239
99X	SRR2065334	181	48,525,235	28.3407	28.1156
9C	SRR2065335	182	70,004,164	41.2806	39.1199

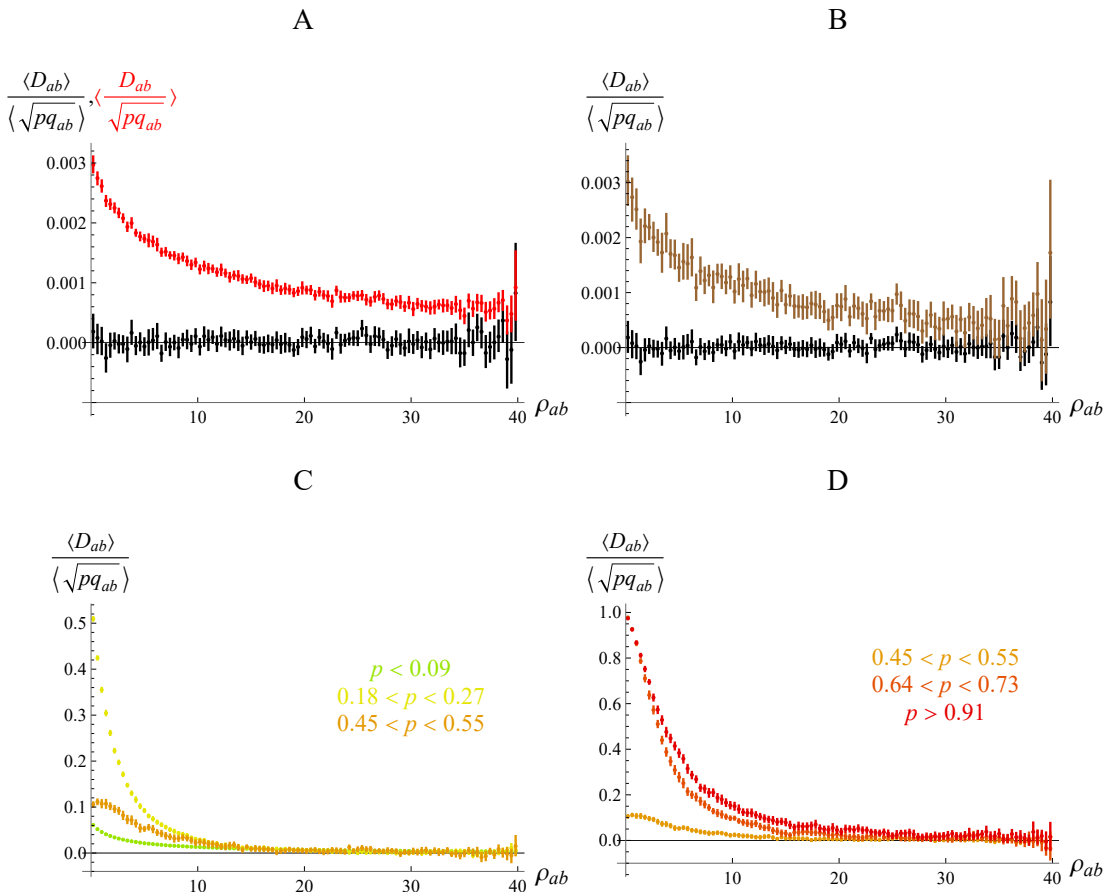


Figure S2.1: Average LD $\langle D_{ab}\rangle$ between derived neutral mutations in the coalescent simulations scaled by the average square root of the product of genetic diversities $\langle\sqrt{pq_{ab}}\rangle = \langle\sqrt{p_a q_a p_b q_b}\rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$. A: Average of D_{ab} (scaled by the average of $\sqrt{pq_{ab}}$, black) and average of $r = D_{ab}/\sqrt{pq_{ab}}$ (red) over all segregating sites for each distance class (no conditioning on frequency). B: Average of D_{ab} (scaled by the average of $\sqrt{pq_{ab}}$) over all segregating sites (black) or excluding singletons (brown). C and D: Average of D_{ab} (scaled by the average of $\sqrt{pq_{ab}}$) over pairs of loci at which the derived allele segregates in a given frequency range. Errors bars correspond to 95% confidence intervals (see Materials & Methods). Parameters values: $\theta = 4N_e u = 100$ (mutation rate); $n = 180$ (sample size).

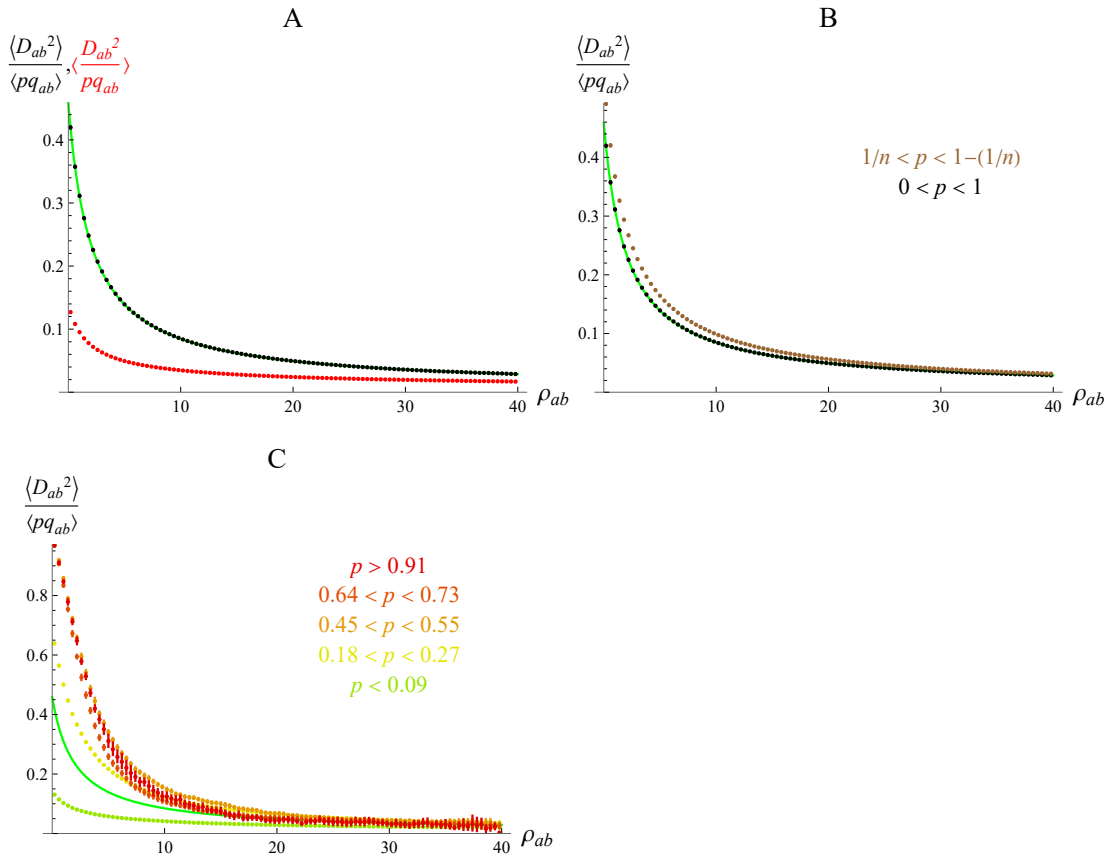


Figure S2.2: Squared LD $\langle {D_{ab}}^2 \rangle$ between derived neutral mutations in the coalescent simulations scaled by the average product of genetic diversities $\langle pq_{ab} \rangle = \langle p_a q_a p_b q_b \rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$. A: Average of ${D_{ab}}^2$ (scaled by the average of pq_{ab}, black) and average of ${D_{ab}}^2/pq_{ab}$ (red) over all segregating sites for each distance class (no conditioning on frequency). B: Average of ${D_{ab}}^2$ (scaled by the average of pq_{ab}) over all segregating sites (black) or excluding singletons (brown). C: Average of ${D_{ab}}^2$ (scaled by the average of pq_{ab}) over pairs of loci at which the derived allele segregates in a given frequency range. The green solid curves correspond to the analytical prediction $(10 + \rho_{ab}) / [(2 + \rho_{ab})(11 + \rho_{ab})]$ from Ohta and Kimura (1969) when averaging over all segregating sites. Parameters values are as in Figure 3.1, S2.1.

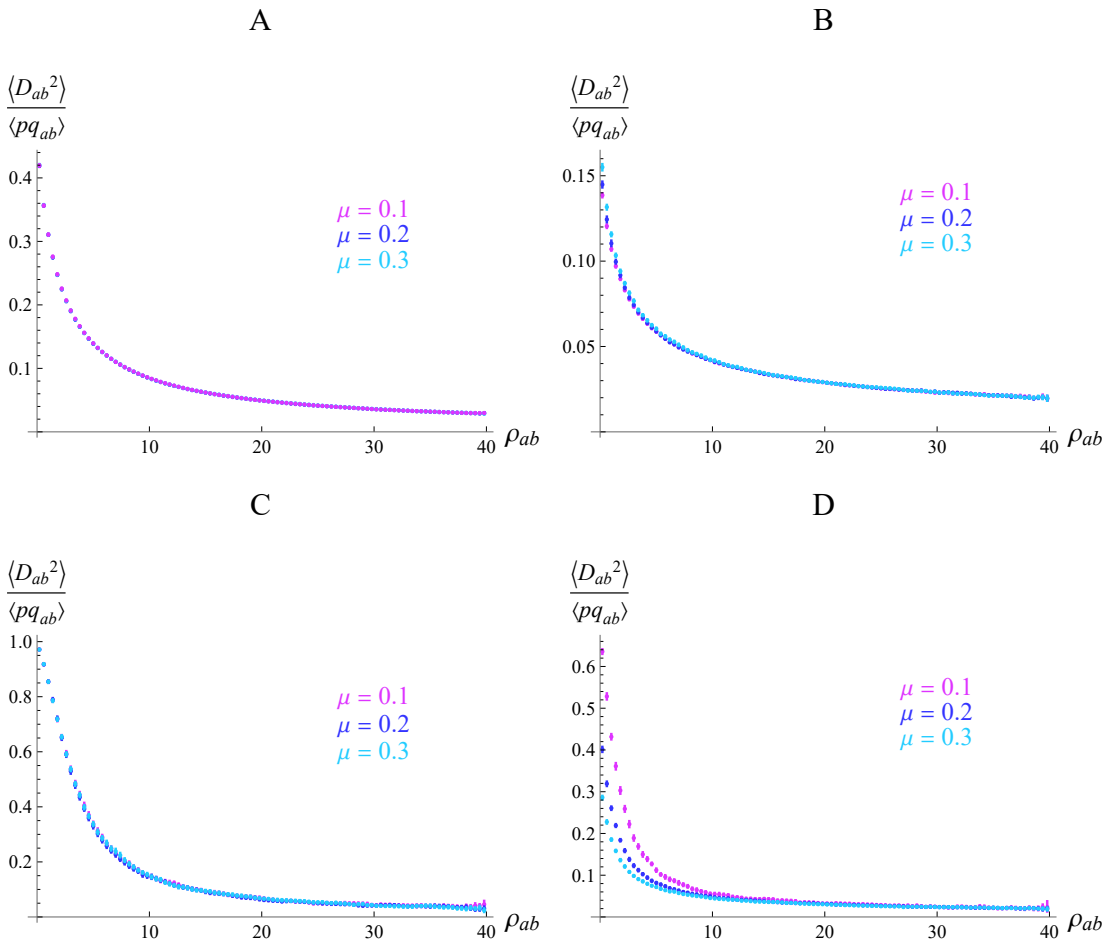


Figure S2.3: Squared LD $\langle D_{ab}{}^2\rangle$ between derived neutral mutations in the coalescent simulations scaled by the average product of genetic diversities $\langle pq_{ab}\rangle = \langle p_a q_a p_b q_b\rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$, and for different values of the rate of misspecification μ of the ancestral/derived state of alleles. A: Averages over all segregating sites for each distance class (no conditioning on frequency); B: only the lowest frequency class ($p < 0.09$, green points in Figure S2.2C); C: only the middle frequency class ($0.45 < p < 0.55$, orange points in Figure S2.2C); D: only the highest frequency class ($p > 0.91$, red points in Figure S2.2C). Parameter values are as in Figure 3.1.

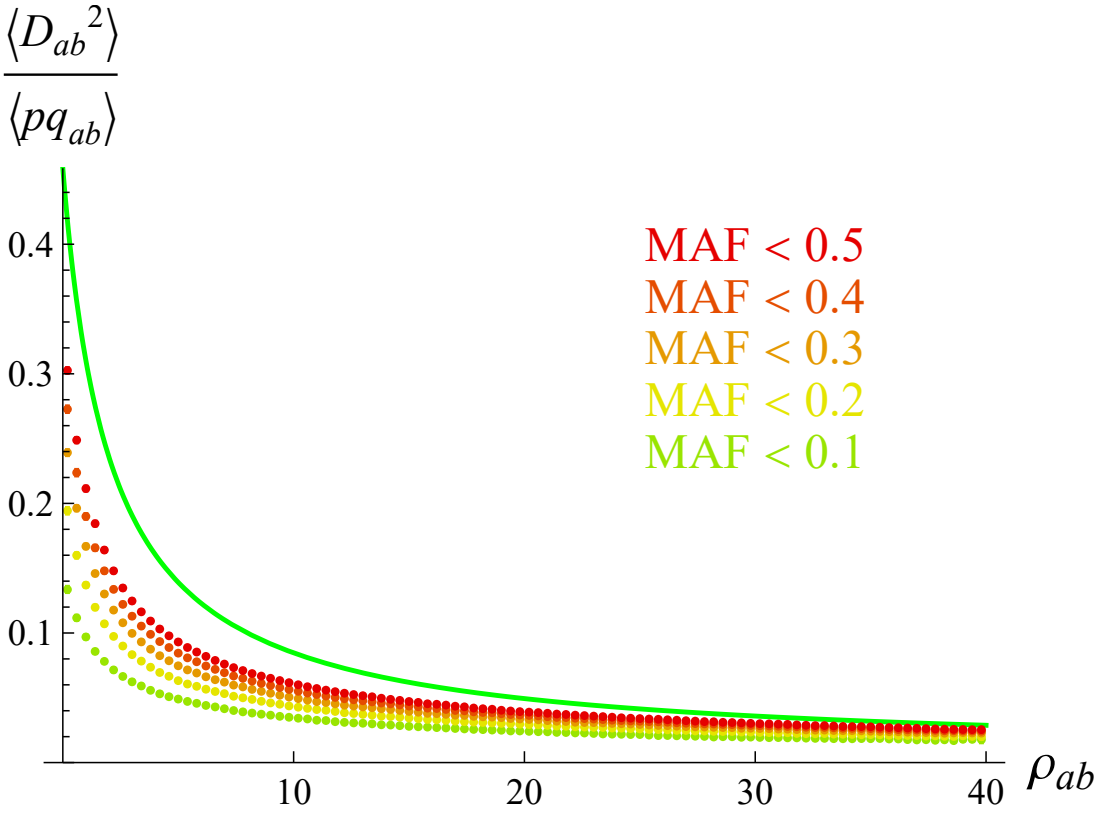


Figure S2.4: Squared LD $\langle {D_{ab}}^2 \rangle$ between derived neutral mutations in the coalescent simulations scaled by the average product of genetic diversities $\langle pq_{ab} \rangle = \langle p_a q_a p_b q_b \rangle$, as a function of the recombination rate $\rho_{ab} = 4N_e r_{ab}$. LD is polarized based on the minor allele frequency (MAF) and averaged over pairs of sites with different MAF thresholds. Parameter values are as in Figure 3.1.

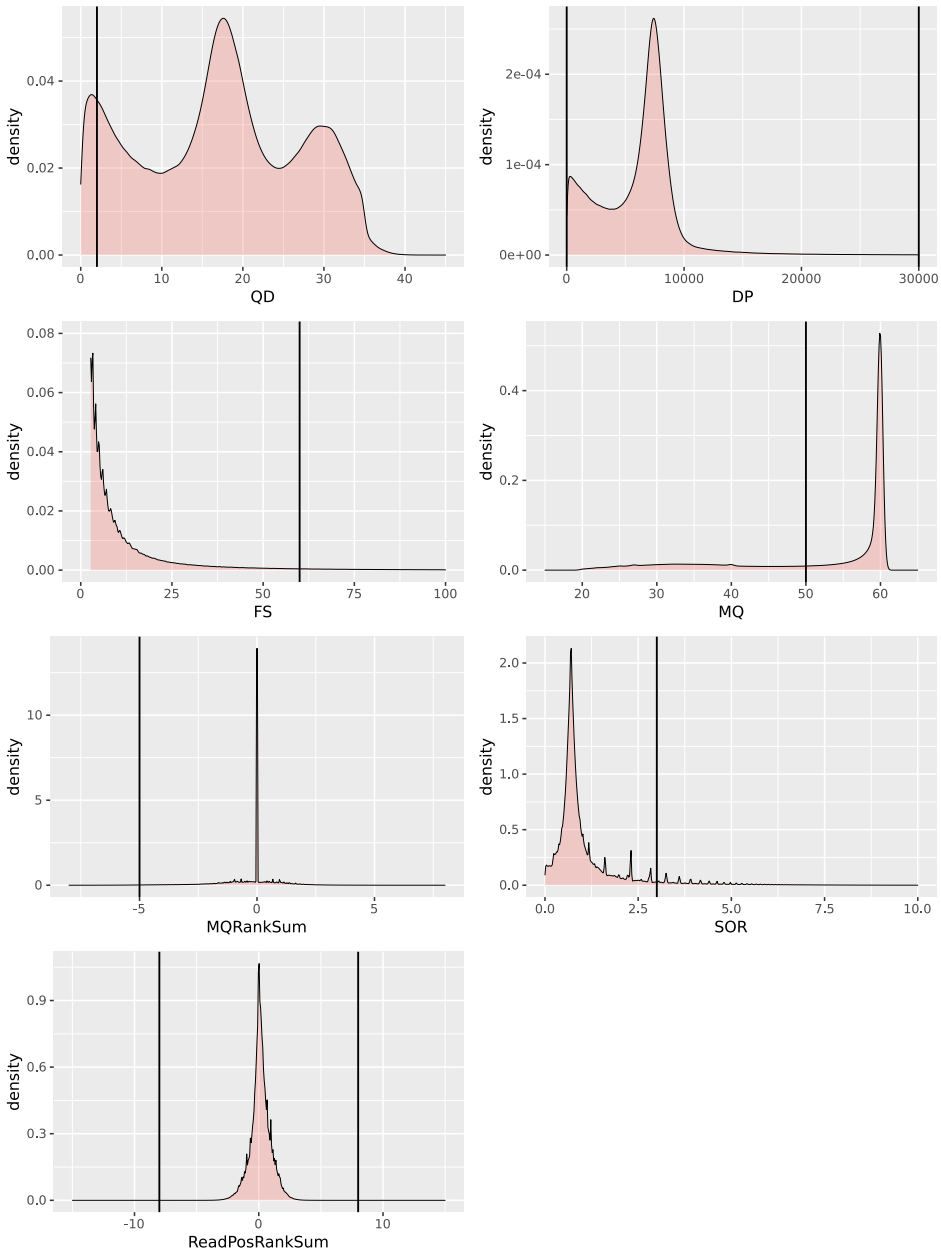


Figure S2.5: Distribution of variant scores and filtration criteria applied (vertical bars) on all SNPs (pooled between *C. grandiflora* and *C. orientalis*). QD: ; DP: Approximate read depth; FS: Phred-scaled p-value using Fisher's exact test to detect strand bias; MQ: RMS Mapping Quality; MQRankSum: Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities; SOR: Symmetric Odds Ratio of 2x2 contingency table to detect strand bias; ReadPosRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias.

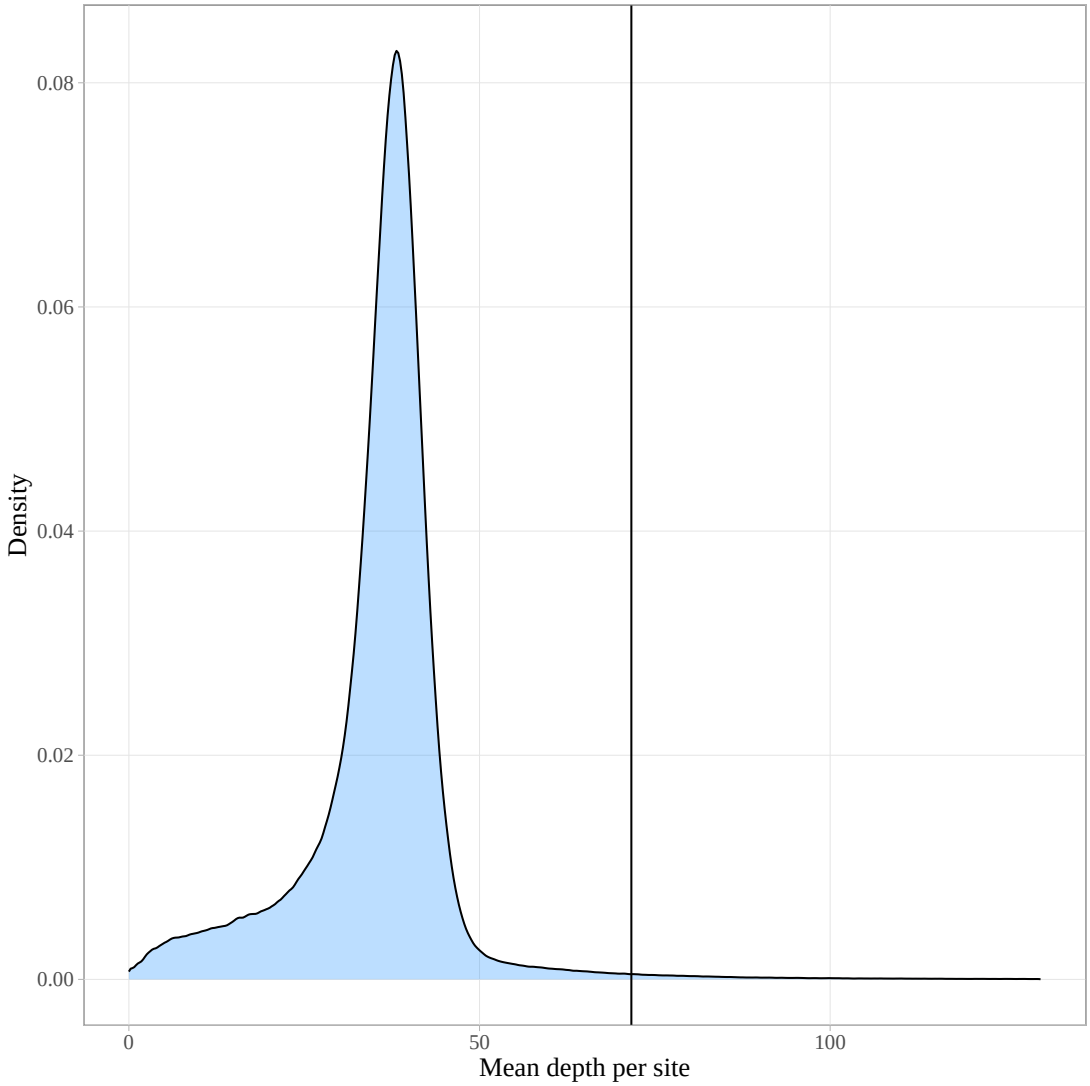


Figure S2.6: Distribution of mean depth per site and filtration criteria applied (vertical bars) on SNPs in *C. grandiflora*.

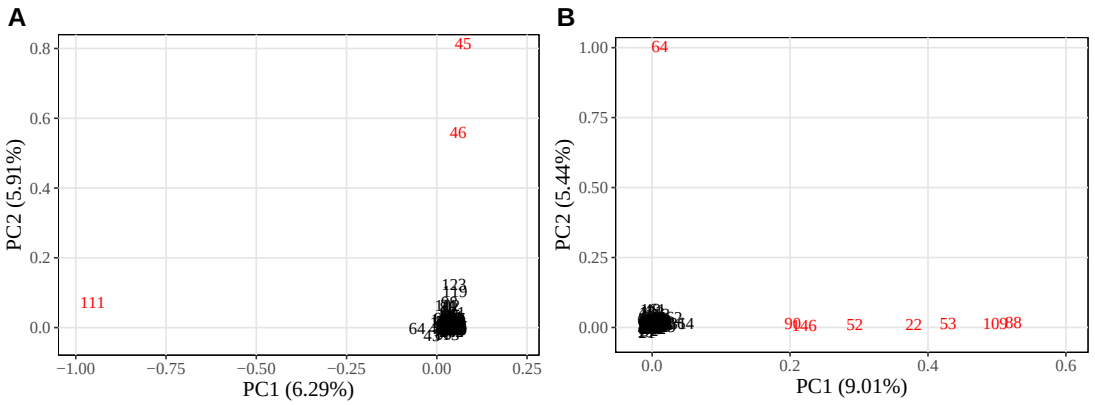


Figure S2.7: The two first components of the PCA on filtered SNPs of *C. grandiflora*. Red individuals in plot A where removed to draw plot B. All red individuals correspond to diverged individuals that were excluded from LD computing. PCA was performed using Plink (Purcell et al., 2007) on independent SNPs extracted with the function *indep-pairwise* (window size = 50 kb, step size = 100 variants and r^2 threshold = 0.1).

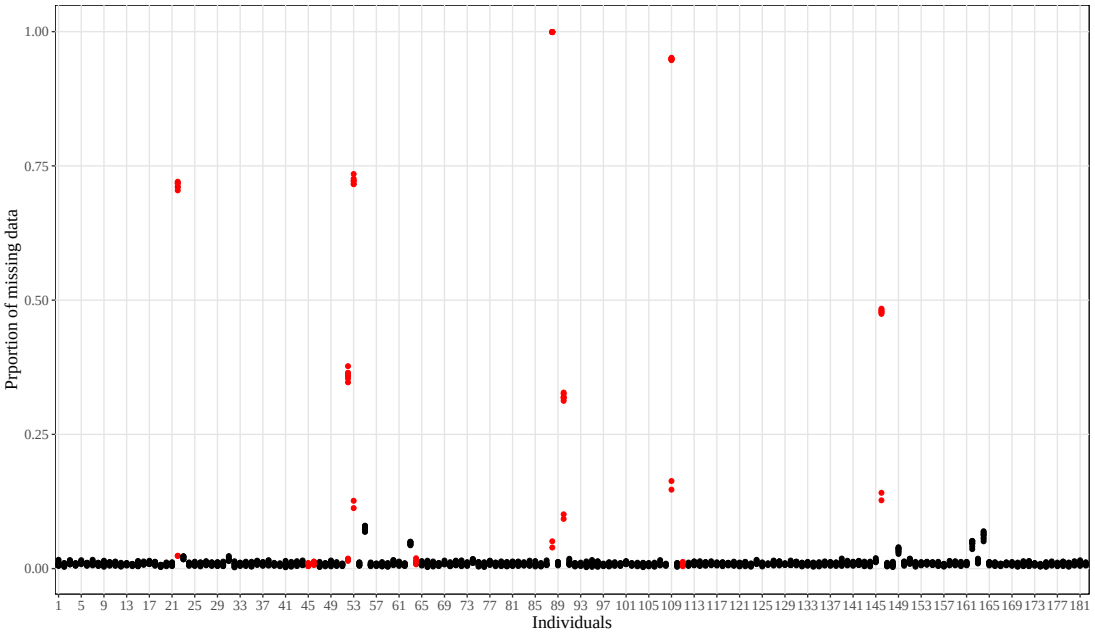


Figure S2.8: Proportion of sites with missing data in SNPs of *C. grandiflora* individuals. The correspondence between individuals' numbers and accession IDs is shown in Table S2.1. Each point represents the proportion computed in each chromosome of each individual. Red points correspond to diverged individuals that were excluded from LD computing.

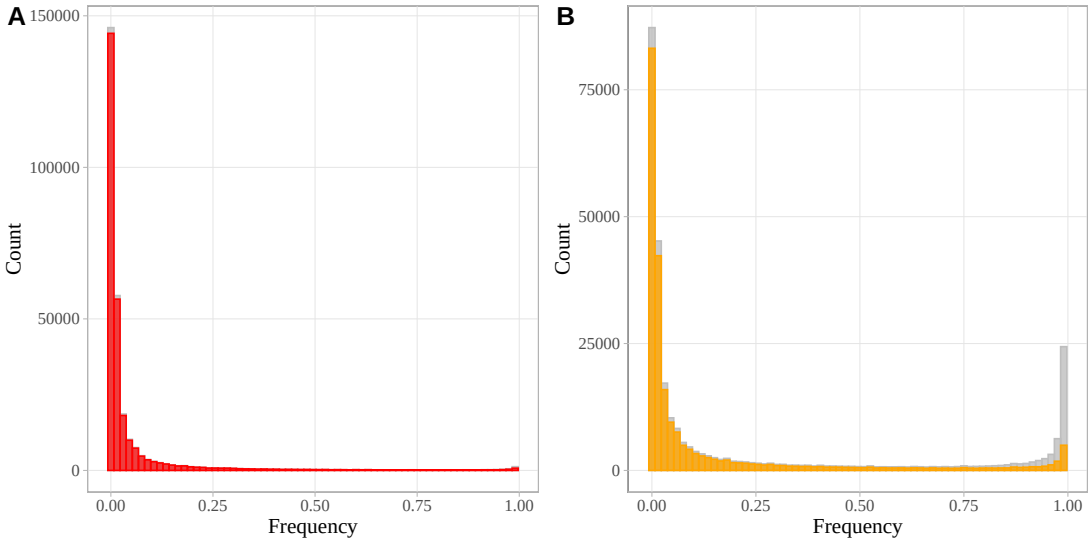


Figure S2.9: Site frequency spectrum for deleterious (A) and mildly deleterious (B) mutations. Colour: derived mutations only (the derived state being assessed from the outgroup sequence of *Neslia paniculata*); Grey: all deleterious and mildly deleterious mutations (including the ancestral ones).

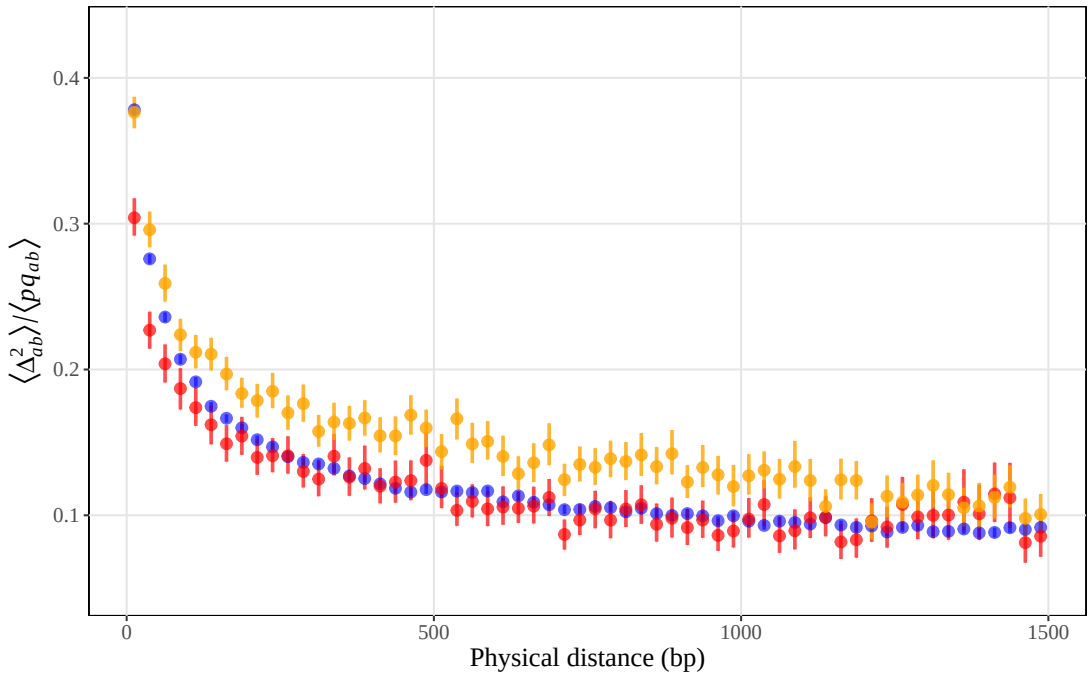


Figure S2.10: Average squared composite linkage disequilibrium Δ_{ab}^2 scaled by the product of allele frequencies between deleterious (red), mildly deleterious (yellow) and neutral (blue) derived mutations according to physical distance.

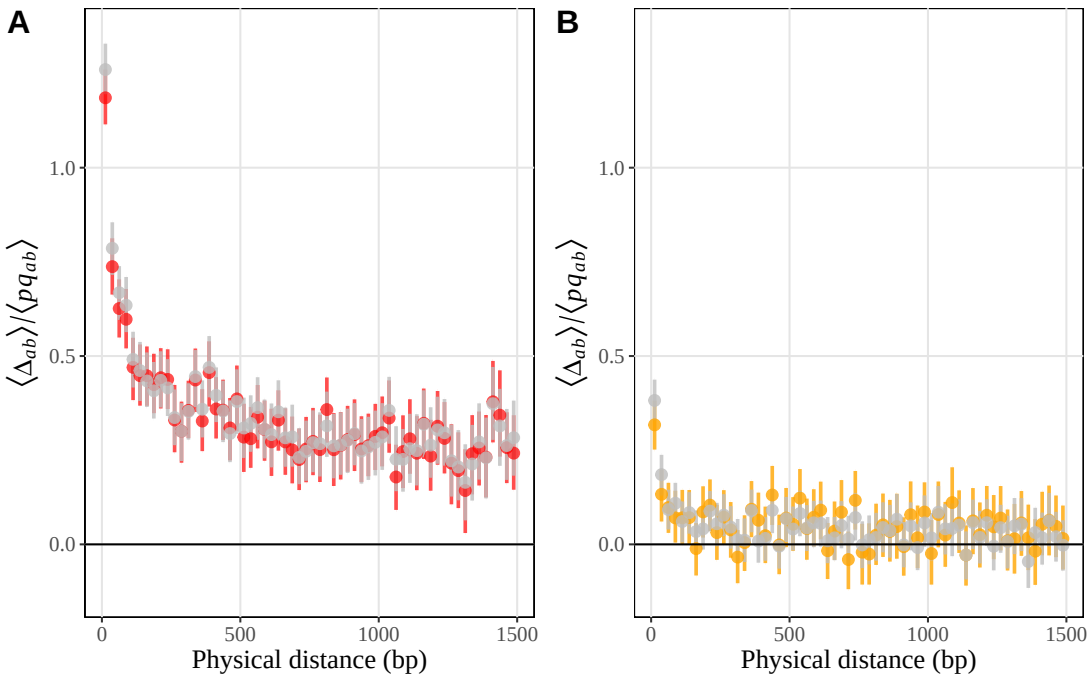


Figure S2.11: Average composite linkage disequilibrium Δ_{ab} scaled by the product of allele frequencies between deleterious (A) or mildly deleterious (B) mutations according to physical distance. Colour: derived deleterious or mildly deleterious mutations only (the derived state being assessed from the outgroup sequence of *Neslia paniculata*); Grey: all deleterious and mildly deleterious mutations (including the ancestral ones).

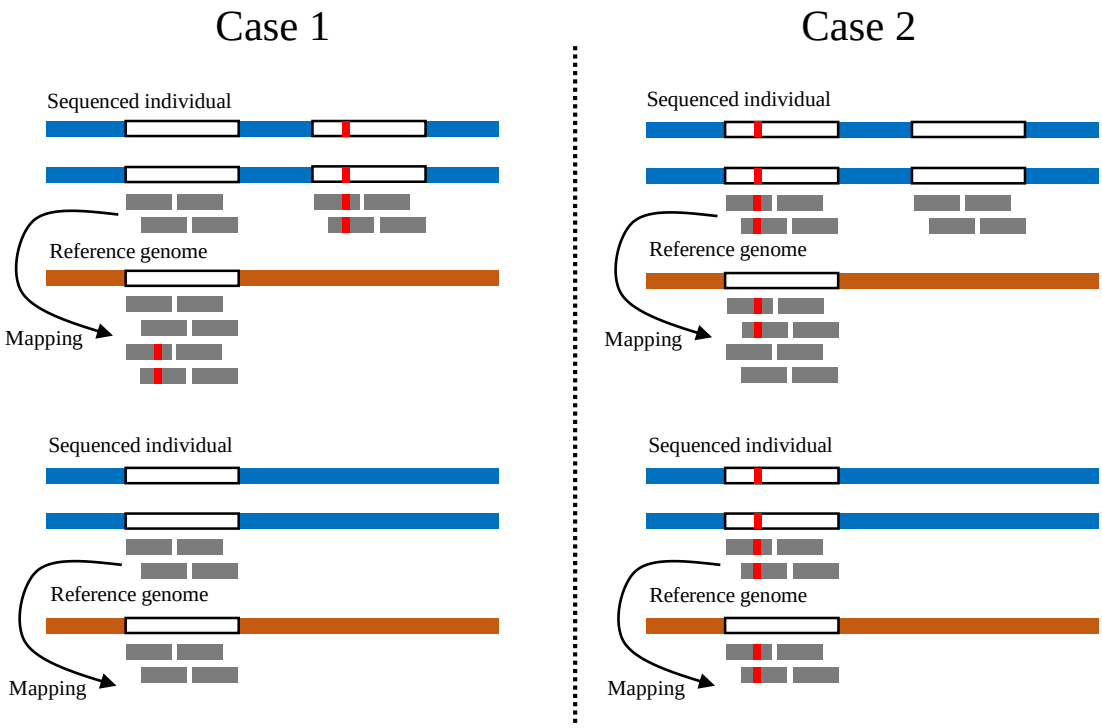


Figure S2.12: Cartoon illustrating the different genotypes found when mapping short reads from individuals carrying or not a duplication which is absent from the reference genome. To simplify the cartoon individuals are assumed to have homozygous genomes. In Case 1 a mutation is fixed in the duplicated region so that all individuals carrying the duplication appear as heterozygotes after mapping against the reference genome while individuals not carrying it appear as homozygous for the ancestral allele. In Case 2 a mutation (not present in the reference genome) is fixed in the original sequence but not the duplicated one, so that all individuals carrying the duplication appear as heterozygotes after mapping against the reference genome while individuals not carrying it appear as homozygous for the derived allele.

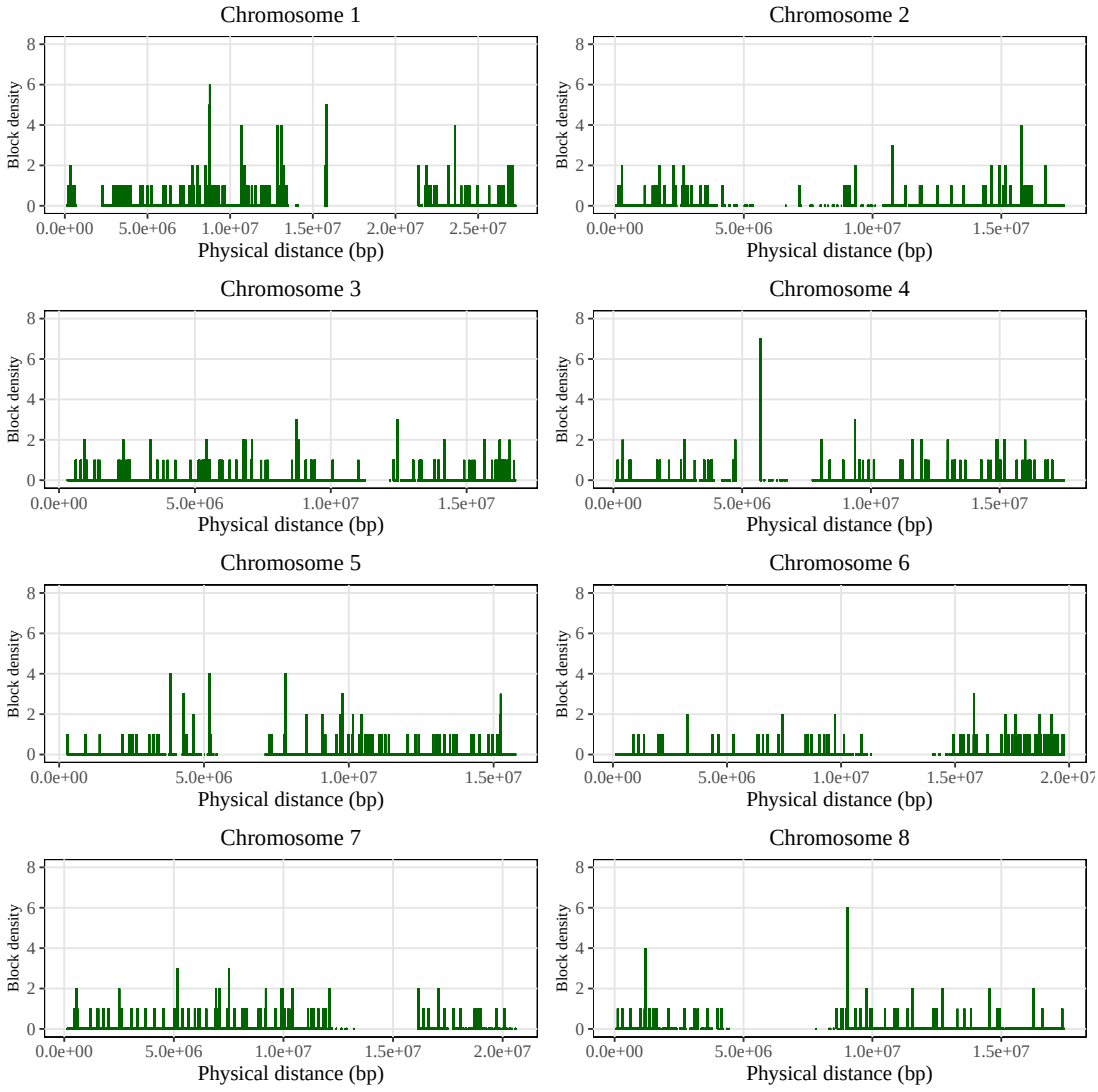


Figure S2.13: Density of blocks corresponding to potential duplications along the genome of *C. granfilora* (having at least 4 SNPs with deleterious mutations). Each vertical bar represents a SNP and its height represents the number of potential duplications it is included in. Note that a single SNPs can be covered by multiple potential duplications, resulting in high block density.

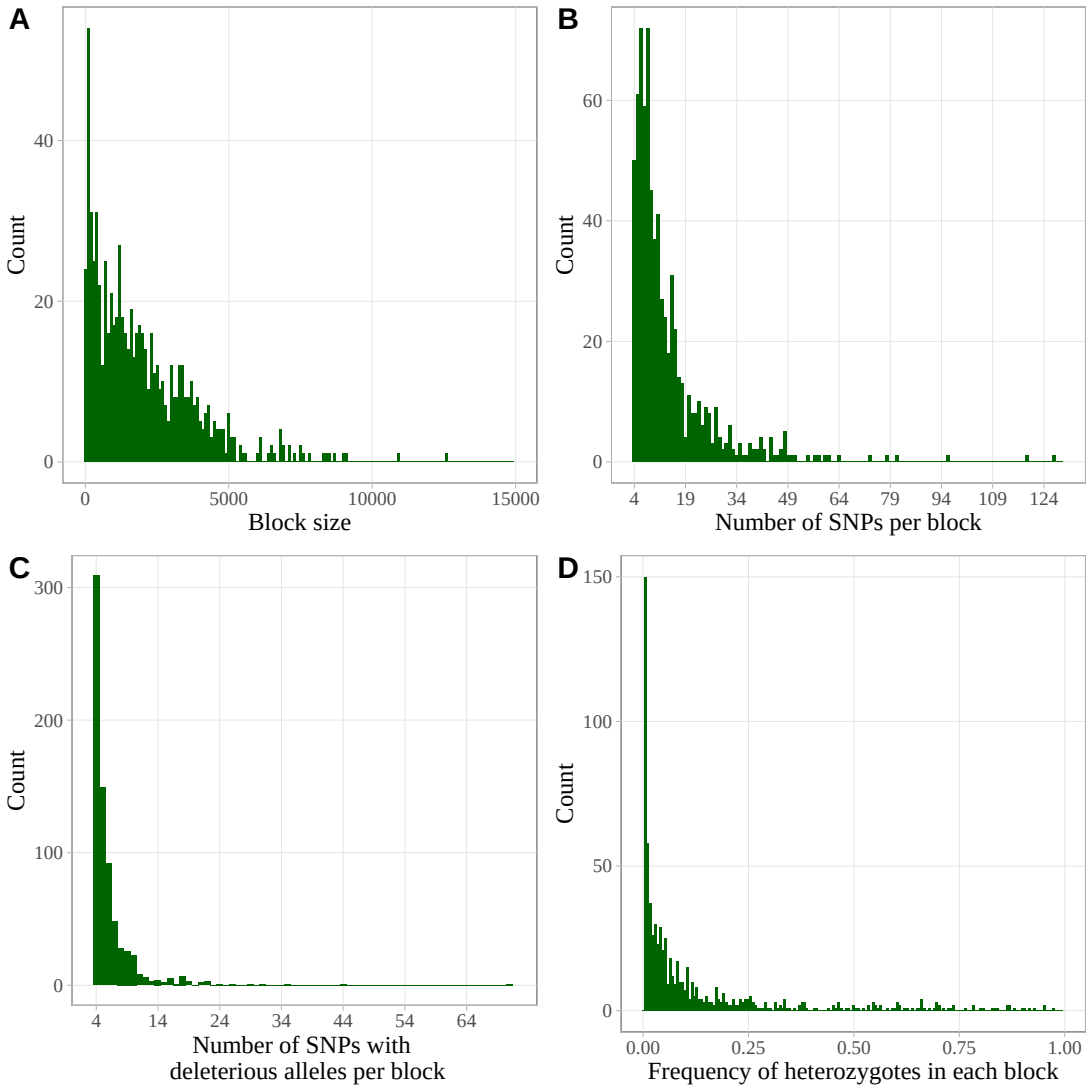


Figure S2.14: Distribution of various quantities related to blocks corresponding to potential duplications in the genome of *C. granfilora*. A. Distribution of block size. B. Distribution of the number of SNPs within each block. C. Distribution of the number of SNPs with deleterious mutations within each block. D. Frequency spectrum of heterozygous individuals (*i.e.*, individuals carrying the duplication) for each block.

Appendix S3

Supplementary figures and tables from Chapter 4

Accession ID	Name	Number
6911	Cvi-0	1
7063	Can-0	2
9533	IP-Cem-0	3
9542	IP-Fun-0	4
9543	IP-Gra-0	5
9545	IP-Her-12	6
9549	IP-Hum-2	7
9550	IP-Iso-4	8
9554	IP-Lso-0	9
9555	IP-Mar-1	10
9574	IP-Rel-0	11
9583	IP-Sne-0	12
9598	IP-Vim-0	13
9600	IP-Vis-0	14
9606	Aitba-1	15
9832	IP-Cat-0	16
9837	IP-Con-0	17
9869	IP-Moj-0	18
9871	IP-Nac-0	19
9879	IP-Per-0	20
9887	IP-Pun-0	21
9905	IP-Ven-0	22
9939	ICE49 / Aitba-2	23
9944	Don-0	24
9947	Ped-0	25

Table S3.1: List of accessions of *Arabidopsis thaliana* labeled as the relict group from The 1001 Genomes Consortium (2016), with their accession ID, name and number used in Chapter 4. Individuals shaded in grey correspond to diverged individuals that were removed from LD computing.

ID	SRA	Number	Number of mapped reads	Mean coverage after mapping (X)	Mean coverage of SNPs after filtering (X)
1719_3	SRR1481617	1	18,390,276	10.0504	10.5592
1719_4	SRR1463025	2	10,574,486	5.6878	7.29812
1979_1	SRR1481500	3	16,451,497	8.8313	11.9095
1979_7	SRR1481499	4	16,292,923	9.0842	4.98391
1981_10	SRR1481618	5	23,333,861	12.6034	16.3639
1981_6	SRR1481625	6	24,627,554	13.3924	16.7625
1985_1	SRR1481626	7	21,260,934	11.473	15.7376
2008_01_01	SRR1481629	8	22,627,341	12.6554	16.0419
2008_07_01	SRR1481628	9	9,791,295	5.1226	6.99365
2008_09_01	SRR1481627	10	15,410,061	8.3294	10.8275
DKCo_10_1982-9	ERR636107	11	69,957,846	42.5355	9.08669
DKCo_11_1983-6	ERR636108	12	63,711,372	38.7837	47.311
DKCo_12_1984-2	ERR636110	13	71,126,519	43.0077	15.1021
DKCo_13_1985-11	ERR636111	14	71,799,581	43.5577	9.11579
DKCo_15_2006-01	ERR636112	15	73,143,441	44.9179	25.0498
DKCo_16_2007-03	ERR636113	16	76,777,610	46.7776	20.7011
DKCo_17_2008-01	ERR636114	17	63,513,398	38.6969	20.4628
DKCo_1_1718-9	ERR636109	18	100,897,399	61.7442	18.0909
DKCo_2_1719-1	ERR636115	19	74,897,711	45.6209	55.1665
DKCo_3_1938-1_1	ERR636116	20	51,805,191	31.5836	15.657
DKCo_4_1939-1_6	ERR636117	21	71,592,836	43.5215	7.73078
DKCo_5_1940-1_1	ERR636118	22	46,732,385	28.6681	15.1894
DKCo_6_1978-6	ERR636119	23	96,368,387	58.7564	11.3313
DKCo_7_1979-02	ERR636120	24	43,256,771	26.343	8.86712
DKCo_8_1980-1	ERR636121	25	108,769,864	66.3334	23.6724
DKCo_9_1981-3	ERR636122	26	61,626,530	37.484	7.37
FY1	SRR6179226	27	30,623,464	18.3567	10.7762
GUB-RUS5	SRR8904471	28	32,968,763	28.7624	12.0383
PAR-RUS	SRR8904459	29	28,937,787	25.2723	8.24175
QH-CHIN4	SRR8904460	30	27,794,870	24.3675	6.91872
QH1	SRR6179228	31	30,817,318	18.4771	5.54056
QH2	SRR6179227	32	27,299,914	16.3607	17.4272
URAL-RUS4	SRR8904461	33	27,249,148	23.8115	29.4385

Table S3.2: List of *Capsella orientalis* individuals with their ID from Ågren et al. (2014); Huang et al. (2018); Kryvokhyzha et al. (2019), SRA code, number used in Chapter 4, number of mapped reads, mean coverage after mapping and mean coverage of SNPs after filtering. Individuals shaded in grey correspond to diverged individuals that were excluded from LD computing.

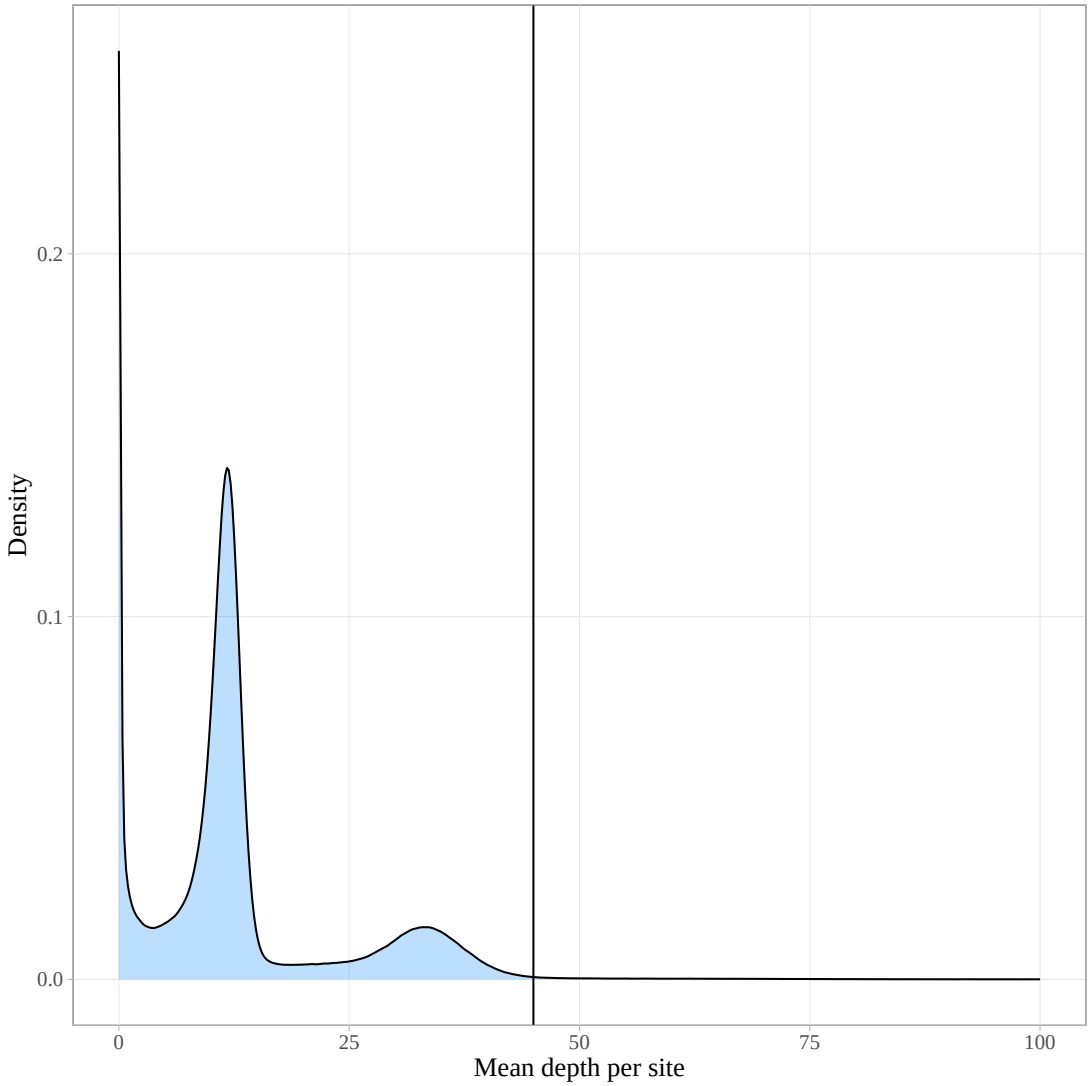


Figure S3.1: Distribution of mean depth per site and filtration criterion applied (vertical bar) on SNPs from *C. orientalis* individuals.

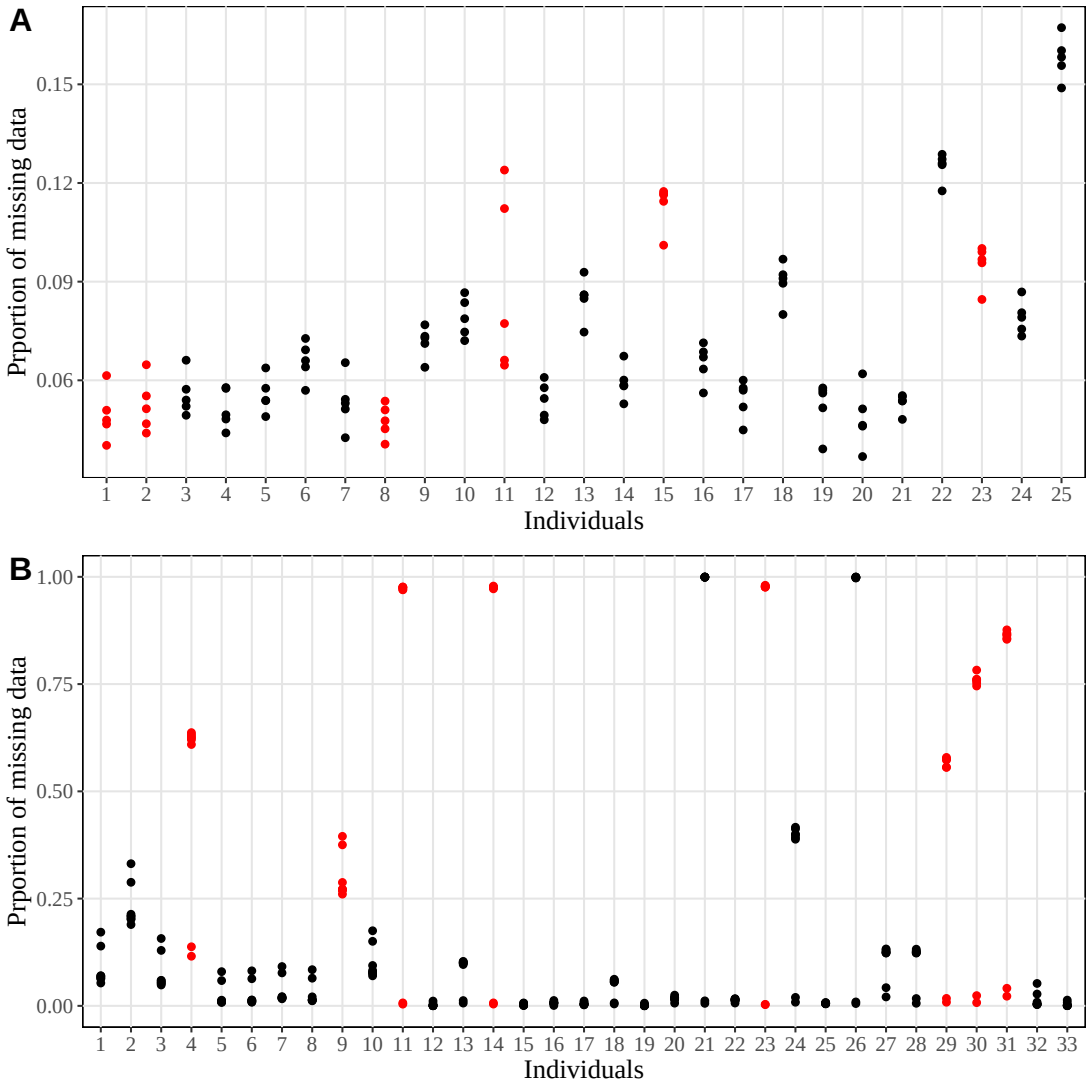


Figure S3.2: Proportion of sites with missing data in SNPs from *A. thaliana* (A) and *C. orientalis* (B) individuals. The correspondence between individuals' numbers and accession IDs is shown in Table S3.1 and Table S3.2. Each point represents the proportion computed for each chromosome of each individual. Red points correspond to diverged individuals that were excluded from LD computing.

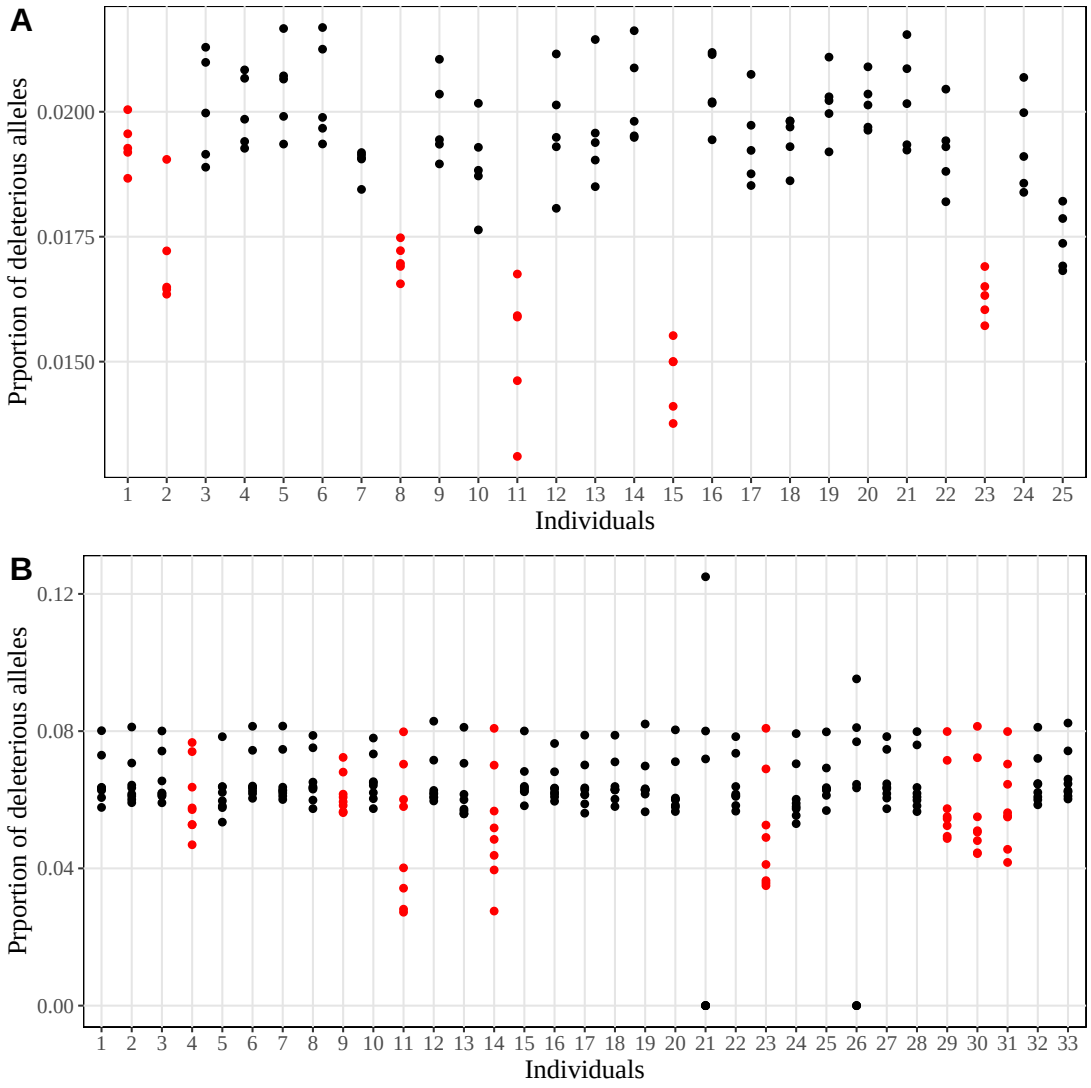


Figure S3.3: Proportion of deleterious alleles in *A. thaliana* (A) and *C. orientalis* (B) individuals. The correspondence between individuals' numbers and accession IDs is shown in Table S3.1 and Table S3.2. The proportion was computed by dividing the number of alleles classified as deleterious by the total number of alleles for each chromosome of each individual. In *C. orientalis* the proportion was computed after removing SNPs with heterozygosity and keeping only common sites with *N. paniculata*. Red points correspond to diverged individuals that were excluded from LD computing.

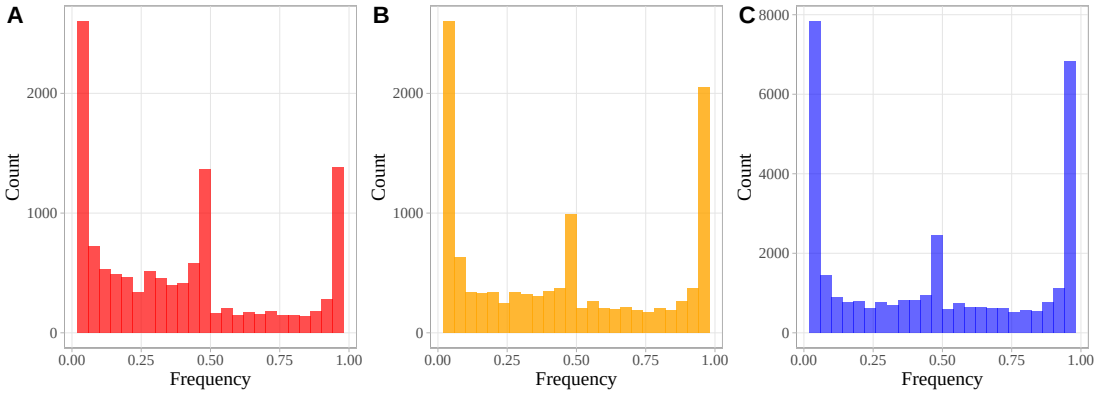


Figure S3.4: Site frequency spectrum for deleterious (A), mildly deleterious (B) and neutral mutations (C) in *C. orientalis* individuals before filtering out SNPs with heterozygosity. Frequencies were computed for all biallelic sites for which at least one neutral mutation segregates. Deleterious and mildly deleterious mutations were polarised using their SIFT score and derived state (using the outgroup sequence of *Neslia paniculata*) while neutral mutations were polarised using their derived state.

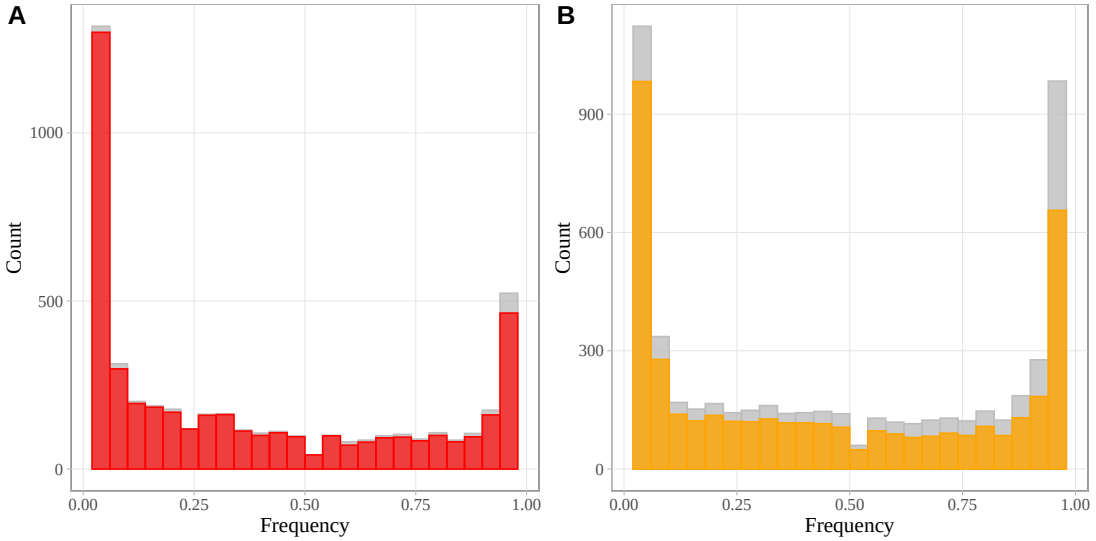


Figure S3.5: Site frequency spectrum for deleterious (A) and mildly deleterious (B) mutations in *C. orientalis* individuals. Frequencies were computed for all biallelic sites for which at least one neutral mutation segregates. Mutations were either polarised using their SIFT score and derived state (colour) or using their SIFT score only (grey).

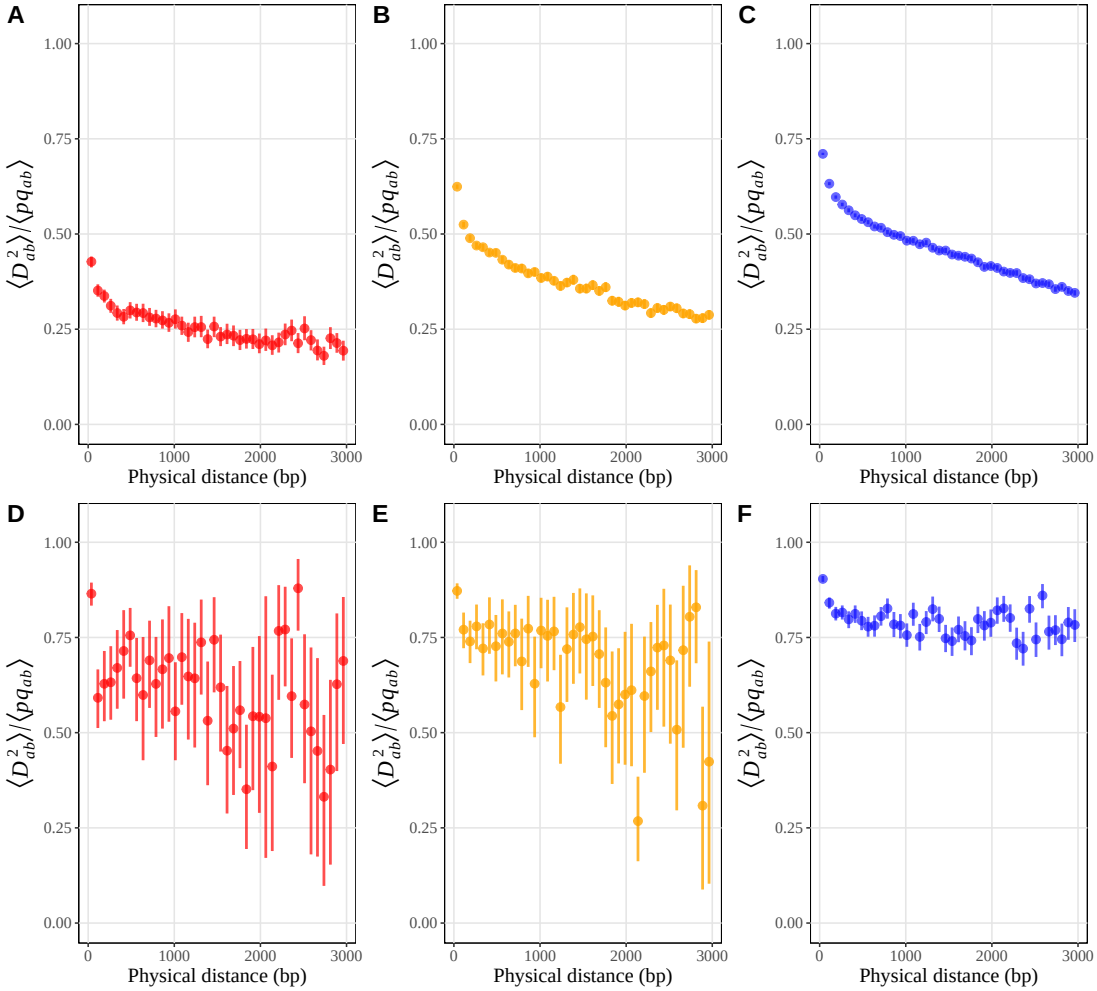


Figure S3.6: Average squared linkage disequilibrium D_{ab}^2 scaled by the average product of allele diversities between deleterious (A, D), mildly deleterious (B, E) and neutral mutations (C, F), for different classes of physical distance between sites (in base pairs), in *A.thaliana* (A, B, C) and *C. orientalis* (D, E, F). Linkage disequilibrium was computed only between fully homozygous SNPs. For *A.thaliana* deleterious and mildly deleterious mutations were polarised based on their SIFT score while neutral mutations were randomly polarised. For *C. orientalis* deleterious and mildly deleterious mutations were polarised using their SIFT score and derived state (using the outgroup sequence of *Neslia paniculata*) while neutral mutations were polarised using their derived state.

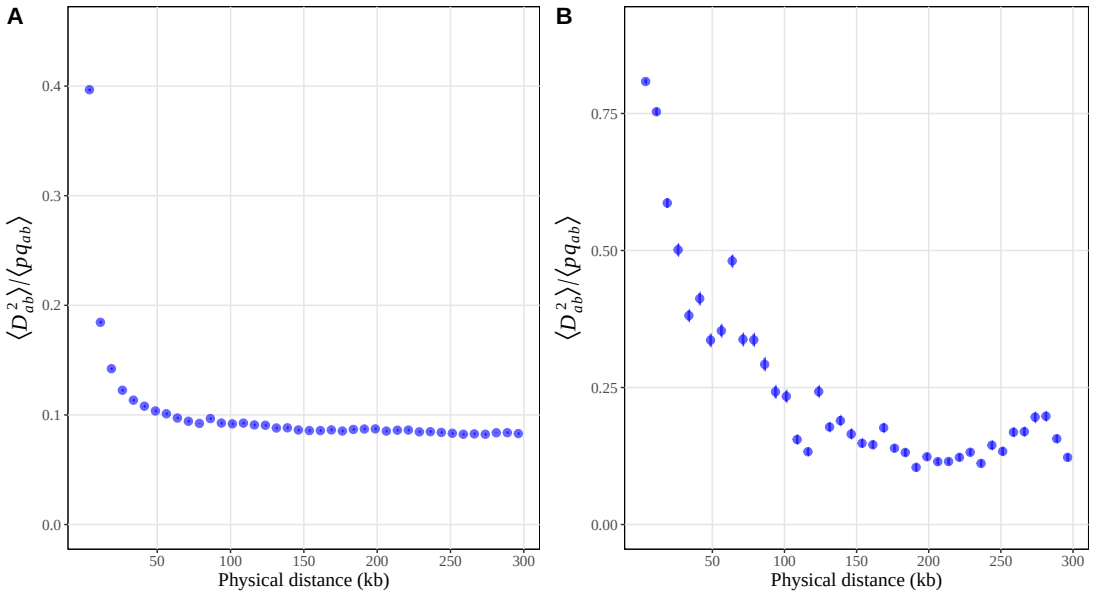


Figure S3.7: Same as Figure S3.6C (A) and Figure S3.6F (B) but for longer physical distances.

References

Ågren, J. A., W. Wang, D. Koenig, B. Neuffer, D. Weigel, and S. I. Wright. 2014. Mating system shifts and transposable element evolution in the plant genus *Capsella*. BMC Genomics 15:602.

Abbott, R. J. and M. F. Gomes. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity 62:411–418.

Abu Awad, D. and D. Roze. 2020. Epistasis, inbreeding depression and the evolution of self-fertilization. Evolution 74:1301–1320.

Agrawal, A. F. 2006a. Evolution of sex: why do organisms shuffle their genotypes? Curr. Biol. 16:R696–R704.

———. 2006b. Similarity selection and the evolution of sex: revisiting the Red Queen. PLoS Biology 4:e265.

Altland, A., A. Fischer, J. Krug, and I. G. Szendro. 2011. Rare events in population genetics: stochastic tunneling in a two-locus model with recombination. Physical Review Letters 106:088101.

Andronic, L. 2012. Viruses as triggers of DNA rearrangements in host plants. Canadian Journal of Plant Science 92:1083–1091.

Auton, A., Y. Rui Li, J. Kidd, K. Oliveira, J. Nadel, J. K. Holloway, J. J. Hayward, P. E. Cohen, J. M. Greally, J. Wang, C. D. Bustamante, and A. R. Boyko. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. PLoS Genetics 9.

Bagowski, C. P., W. Bruins, and A. J. W. te Velthuis. 2010. The nature of protein domain evolution: shaping the Interaction network. Current Genomics 11:368–376.

Barrett, S. C. H. 2002. The evolution of plant sexual diversity. Nature Reviews Genetics 3:274–284.

Barrett, S. C. H., R. Arunkumar, and S. I. Wright. 2014. The demography and population genomics of evolutionary transitions to self-fertilization in plants. Phil. Trans. Roy. Soc. (Lond.) B 369:20130344.

Barton, N. H. 1995. A general model for the evolution of recombination. Genet. Res. 65:123–144.

Barton, N. H. and S. P. Otto. 2005. Evolution of recombination due to random drift. Genetics 169:2353–2370.

Barton, N. H. and M. Turelli. 1991. Natural and sexual selection on many loci. Genetics 127:229–255.

Bauer, E., M. Falque, H. Walter, C. Bauland, C. Camisan, L. Campo, N. Meyer, N. Ranc, R. Rincent, W. Schipprack, T. Altmann, P. Flament, A. E. Melchinger, M. Menz, J. Moreno-González, M. Ouzunova, P. Revilla, A. Charcosset, O. C. Martin, and C. C. Schön. 2013. Intraspecific variation of recombination rate in maize. Genome Biology 14:R103.

Benavente, E. and J. Sybenga. 2004. The relation between pairing preference and chiasma frequency in tetrasomics of rye. Genome 47:122–133.

Billiard, S., M. López-Villavicencio, M. E. Hood, and T. Giraud. 2012. Sex, outcrossing and mating types: unsolved questions in fungi and beyond. J. Evol. Biol. 25:1020–1038.

Blackwell, A. R., J. Dluzewska, M. Szymanska-Lejman, S. Desjardins, A. J. Tock, N. Kbiri, C. Lambing, E. J. Lawrence, T. Bieluszewski, B. Rowan, J. D. Higgins, P. A. Ziolkowski, and I. R. Henderson. 2020. MSH2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. EMBO J. 39:e104858.

Blanquart, F., G. Achaz, T. Bataillon, and O. Tenaillon. 2014. Properties of selected mutations and genotypic landscapes under Fisher's geometric model. Evolution 68:3537–3554.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bomblies, K., L. Yant, R. A. Laitinen, S.-T. Kim, J. D. Hollister, N. Warthmann, J. Fitz, and D. Weigel. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. PLoS Genetics 6:e1000890.

Borts, R. H. and J. E. Haber. 1987. Meiotic recombination in yeast: alteration by multiple heterozygosities. Science 237:1459–1465.

Brachet, E., V. Sommermeyer, and V. Borde. 2012. Interplay between modifications of chromatin and meiotic recombination hotspots. Biology of the Cell 104:51–69.

Brazier, T. and S. Glémin. 2022. Diversity and determinants of recombination landscapes in flowering plants. PLoS Genetics 18:e1010141.

Brekke, C., P. Berg, A. B. Gjuvsland, and S. E. Johnston. 2022. Recombination rates in pigs differ between breeds, sexes and individuals, and are associated with the *RNF212*, *SYCP2*, *PRDM7*, *MEI1* and *MSH4* loci. Genetics Selection Evolution 54:1–14.

Brennan, A. C., B. Méndez-Vigo, A. Haddioui, J. M. Martínez-Zapater, X. F Picó, and C. Alonso-Blanco. 2014. The genetic structure of *Arabidopsis thaliana* in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. BMC Plant Biology 14:17.

Buffalo, V. 2021. Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's paradox. eLife 10:e67509.

Burt, A. 1995. The evolution of fitness. Evolution 49:1–8.

Busch, J. W. and L. F. Delph. 2012. The relative importance of reproductive assurance and automatic selection as hypotheses for the evolution of self-fertilization. Ann. Bot. 109:553–562.

Cavalier-Smith, T. 2002. Origins of the machinery of recombination and sex. Heredity 88:125–141.

Charlesworth, B. 1976. Recombination modification in a fluctuating environment. Genetics 83:181–195.

———. 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet. Res. 55:199–221.

———. 1993. Directional selection and the evolution of sex and recombination. Genet. Res. 61:205–224.

———. 2015. Causes of natural variation in fitness: evidence from studies of *Drosophila* populations. Proc. Natl. Acad. Sci. U. S. A. 112:1662–1669.

Charlesworth, B., M. T. Morgan, and B. Charlesworth. 1990. Inbreeding depression, genetic load, and the evolution of outcrossing rates in a multilocus system with no linkage. Evolution 44:1469–1489.

Charlesworth, D. 2006. Evolution of plant breeding systems. Curr. Biol. 16:R726–R735.

Charlesworth, D., B. Charlesworth, and C. Strobeck. 1977. Effects of selfing on selection for recombination. Genetics 86:213–226.

Cheetham, S. W., G. J. Faulkner, and M. E. Dinger. 2020. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nature Reviews Genetics 21:191–201.

Chen, W. and S. Jinks-Robertson. 1999. The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. Genetics 151:1299–1313.

Cheptou, P. O., O. Carrue, S. Rouifed, and A. Cantarel. 2008. Rapid evolution of seed dispersal in an urban environment in the weed *Crepis sancta*. Proc. Natl. Acad. Sci. U. S. A. 105:3796–3799.

Choi, K. and I. R. Henderson. 2015. Meiotic recombination hotspots – a comparative view. The Plant Journal 83:52–61.

Choi, K., C. Reinhard, H. Serra, P. A. Ziolkowski, C. J. Underwood, X. Zhao, T. J. Hardcastle, N. E. Yelina, C. Griffin, M. Jackson, C. Mézard, G. McVean, G. P. Copenhaver, and I. R. Henderson. 2016. Recombination rate heterogeneity within *Arabidopsis* disease resistance genes. PLoS Genetics 12:e1006179.

Cooper, J. D. and B. Kerr. 2016. Evolution at 'sutures' and 'centers': recombination can aid adaptation of spatially structured populations on rugged fitness landscapes. PLoS Computational Biology 12.

Costanzo, C., E. Kuzmin, J. van Leeuwen, B. Mair, J. Moffat, C. Boone, and B. Andrews. 2019. Global genetic networks and the genotype-to-phenotype relationship. Cell 177:85–100.

Cutter, A. D. 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. Genetics 172:171–184.

Darwin, C. and A. Wallace. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. Journal of the Proceedings of the Linnean Society of London. Zoology 3:45–62.

Davis, B. H., A. F. Y. Poon, and M. C. Whitlock. 2009. Compensatory mutations are repeatable and clustered within proteins. Proc. Roy. Soc. (Lond.) B 276:1823–1827.

De Coster, W., M. H. Weissensteiner, and F. J. Sedlazeck. 2021. Towards population-scale long-read sequencing. Nature Reviews Genetics 22:572–587.

de Massy, B. 2013. Initiation of meiotic recombination: How and where? Conservation and specificities among eukaryotes. Annual Review of Genetics 47:563–599.

de Visser, J. A. G. M. and S. F. Elena. 2007. The evolution of sex: empirical insights into the roles of epistasis and drift. Nat. Rev. Genet. 8:139–149.

de Visser, J. A. G. M. and J. Krug. 2014. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics 15:480–490.

de Visser, J. A. G. M., S.-C. Park, and J. Krug. 2009. Exploring the effect of sex on empirical fitness landscapes. The American Naturalist 174:S15–S30.

Dorant, Y., H. Cayuela, K. Wellband, M. Laporte, Q. Rougemont, C. Mérot, E. Normandeau, R. Rochette, and L. Bernatchez. 2020. Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. Molecular Ecology 29:4765–4782.

Dumont, B. L., K. W. Broman, and B. A. Payseur. 2009. Variation in genomic recombination rates among heterogeneous stock mice. Genetics 182:1345–1349.

Dumont, B. L., A. A. Devlin, D. M. Truempy, J. C. Miller, and N. D. Singh. 2015. No evidence that infection alters global recombination rate in house mice. PLoS ONE 10:e0142266.

Durvasula, A., A. Fulgione, R. M. Gutaker, S. I. Alacakaptan, P. J. Flood, C. Neto, T. Tsuchimatsu, H. A. Burbano, F. X. Picó, C. Alonso-Blanco, and A. M. Hancock. 2017. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U. S. A. 114:5213–5218.

Encinas-Viso, F., A. Young, and J. R. Pannell. 2020. The loss of self-incompatibility in a range expansion. Journal of Evolutionary Biology 33:1235–1244.

Eyre-Walker, A. and P. D. Keightley. 2007. The ditribution of fitness effects of new mutations. Nat. Rev. Genet. 8:610–618.

Feldman, M. W. 1972. Selection for linkage modification: I. Random mating populations. Theoretical Population Biology 3:324–346.

Feldman, M. W., F. B. Christiansen, and L. D. Brooks. 1980. Evolution of recombination in a constant environment. Proc. Natl. Acad. Sci. U. S. A. 77:4838–4841.

Feldman, M. W., S. P. Otto, and F. B. Christiansen. 1997. Population genetic perspectives on the evolution of recombination. Annu. Rev. Genet 30:261–95.

Felsenstein, J. 1965. The effect of linkage on directional selection. Genetics 52:349–363.

———. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

Felsenstein, J. and S. Yokohama. 1976. The evolutionary advantage of recombination. II. Individual selection for recombination. Genetics 83:845–859.

Fernandes, J. B., M. Séguéla-Arnaud, C. Larchevêque, and R. Mercier. 2018. Unleashing meiotic crossovers in hybrid plants. Proc. Natl. Acad. Sci. U. S. A. 115:2431–2436.

Fernandes, J. B., P. Wlodzimierz, and I. R. Henderson. 2019. Meiotic recombination within plant centromeres. Current Opinion in Plant Biology 48:26–35.

Fischer, O. and P. Schmid-Hempel. 2005. Selection by parasites may increase host recombination frequency. Biology Letters 1:193–195.

Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Flatt, T. and L. Partridge. 2018. Horizons in the evolution of aging. BMC Biolgy 16:93.

Frank, S. A. and M. Slatkin. 1992. Fisher's Fundamental Theorem of Natural Selection. Trends Ecol. Evol. 7:92–95.

Fulton, J. E., A. M. McCarron, A. R. Lund, K. N. Pinegar, A. Wolc, O. Chazara, B. Bed'Hom, M. Berres, and M. M. Miller. 2016. A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex *B* region between *BG2* and *CD1A1*. Genetics Selection Evolution 48:1.

Gandon, S. and S. P. Otto. 2007. The evolution of sex and recombination in response to abiotic or coevolutionary fluctuations in epistasis. Genetics 175:1835–1863.

Garcia, J. A. and K. E. Lohmueller. 2021. Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome. PLoS Genetics 17:e1009676.

Genestier, A., L. Duret, and N. Lartillot. 2023. Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis: a model based exploration of the PRDM9 intra-genomic Red Queen. bioRxiv .

Glémin, S. and J. Ronfort. 2013. Adaptation and maladaptation in selfing in outcrossing species: new mutations versus standing variation. Evolution 67:225–240.

Golding, G. B. and C. Strobeck. 1980. Linkage disequilibrium in a finite population that is partially selfing. Genetics 94:777–789.

Goldstein, D. B., A. Bergman, and M. W. Feldman. 1992. The evolution of interference: reduction of recombination among three loci. Theoretical Population Biology 44:246–259.

Good, B. H. 2022. Linkage disequilibrium between rare mutations. Genetics 220:iyac004.

Goodwillie, C., S. Kalisz, and C. G. Eckert. 2005. The evolutionary enigma of mixed mating systems in plants: occurence, theoretical explanations, and empirical evidence. Ann. Rev. Ecol. Evol. Syst. 36:47–79.

Gouyon, P.-H. 1999. Sex: a pluralist approach includes species selection. (One step beyond and it's good.). J. Evol. Biol. 12:1029–1030.

Gray, C., F. Baudat, and B. de Massy. 2018. PRDM9, a driver of the genetic map. PLoS Genetics 14:e1007479.

Gray, D. and P. E. Cohen. 2016. Control of meiotic crossovers: from double-strand break formation to designation. Ann. Rev. Gen. 50:175–210.

Greeff, M. and P. Schmid-Hempel. 2010. Influence of co-evolution with a parasite, *Nosema whitei*, and population size on recombination rates and fitness in the red flour beetle, *Tribolium castaneum*. Genetica 138:737–744.

Gros, P.-A., H. Le Nagard, and O. Tenaillon. 2009. The evolution of epistasis and its links with genetic robustness, complexity and drift in a phenotypic model of adaptation. Genetics 182:277–293.

Guo, Y. L., J. S. Bechsgaard, T. Slotte, B. Neuffer, M. Lascoux, D. Weigel, and M. H. Schierup. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc. Natl. Acad. Sci. U. S. A. 106:5246–5251.

Haenel, Q., T. G. Laurentino, M. Roesti, and D. Berner. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol. Ecol. 27:2477–2497.

Halldorsson, B. V., G. Palsson, O. A. Stefansson, H. Jonsson, M. T. Hardarson, H. P. Eggertsson, B. Gunnarsson, A. Oddsson, G. H. Halldorsson, F. Zink, S. A. Gudjonsson, M. L. Frigge, G. Thorleifsson, A. Sigurdsson, S. N. Stacey, P. Sulem, G. Masson, A. Helgason, D. F. Gudbjartsson, U. Thorsteinsdottir, and K. Stefansson. 2019. Human genetics: characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363:eaau1043.

Hansen, T. F., C. Pélabon, and D. Houle. 2011. Heritability is not Evolvability. Evolutionary Biology 38:258–277.

Hansson, B., A. Kawabe, S. Preuss, H. Kuittinen, and D. Charlesworth. 2006. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 1 and 2 and the corresponding *A. thaliana* chromosome 1: recombination rates, rearrangements and centromere location. Genet. Res. 87:75–85.

Harrison, P. M., H. Hegyi, S. Balasubramanian, N. M. M. Luscombe, P. Bertone, N. Echols, T. Johnson, and M. Gerstein. 2002. Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res. 12:272–280.

Hein, J., S. M. H., and C. Wiuf. 2004. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, USA.

Hill, W. G. and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.

Hill, W. G. and B. S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33:54–78.

Holsinger, K. E. and M. W. Feldman. 1983. Linkage modification with mixed random mating and selfing: a numerical study. Genetics 103:323–333.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics 130:195–204.

Huang, H. R., J. J. Liu, Y. Xu, M. Lascoux, X. J. Ge, and S. I. Wright. 2018. Homeologue-specific expression divergence in the recently formed tetraploid *Capsella bursa-pastoris* (Brassicaceae). New Phytologist 220:624–635.

Huang, W., M. A. Carbone, M. M. Magwire, J. A. Peiffer, R. F. Lyman, E. A. Stone, R. R. H. Anholt, and T. F. C. Mackay. 2015. Genetic basis of transcriptome diversity in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 112:E6010–E6019.

Hurka, H., N. Friesen, D. A. German, A. Franzke, and B. Neuffer. 2012. 'Missing link' species *Capsella orientalis* and *Capsella thracica* elucidate evolution of model plant genus *Capsella* (Brassicaceae). Mol. Ecol. 21:1223–1238.

Jaegle, B., R. Pisupati, L. M. Soto-jiménez, R. Burns, A. R. Fernando, and M. Nordborg. 2023. Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. Genome Biology 24:44.

Jain, K. 2010. Time to fixation in the presence of recombination. Theoretical Population Biology 77:23–31.

Jarne, P. and J. R. Auld. 2006. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution 60:1816–1824.

Jarne, P. and D. Charlesworth. 1993. The evolution of the selfing rate in functionally hermaphrodite plants and animals. Annu. Rev. Ecol. Syst. 24:441–466.

Jiao, K., W. B.and Schneeberger. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nature Communications 11:989.

John R. Pannell, J. R. 2015. Evolution of the mating system in colonizing plants. Molecular Ecology 24:2018–2037.

Johnsson, M., A. Whalen, R. Ros-Freixedes, G. Gorjanc, C. Y. Chen, W. O. Herring, D.-J. de Koning, and J. M. Hickey. 2021. Genetic variation in recombination rate in the pig. Genetics Selection Evolution 53:54.

Johnston, S. E., C. Bérénos, J. Slate, and J. M. Pemberton. 2016. Conserved genetic architecture underlying individual recombination rate variation in a wild population of Soay sheep (*Ovis aries*). Genetics 203:583–598.

Johnston, S. E., J. Huisman, and J. M. Pemberton. 2018. A genomic region containing *REC8* and *RNF212B* is associated with individual recombination rate variation in a wild population of red deer (*Cervus elaphus*). G3: Genes, Genomes, Genetics 8:2265–2276.

Josephs, E. B., Y. W. Lee, J. R. Stinchcombe, and S. I. Wright. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proc. Natl. Acad. Sci. U. S. A. 112:15390–15395.

Kamran-Disfani, A. and A. F. Agrawal. 2014. Selfing, adaptation and background selection in finite populations. J. Evol. Biol. 27:1360–1371.

Kaur, T. and M. V. Rockman. 2014. Crossover heterogeneity in the absence of hotspots in *Caenorhabditis elegans*. Genetics 196:137–148.

Kawabe, A., B. Hansson, A. Forrest, J. Hagenblad, and D. Charlesworth. 2006. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 6 and 7 and *A. thaliana* chromosome IV: evolutionary history, rearrangements and local recombination rates. Genet. Res. 88:45–56.

Kawakami, T., A. Wallberg, A. Olsson, D. Wintermantel, J. R. De Miranda, M. Allsopp, M. Rundlöf, and M. T. Webster. 2019. Substantial heritable variation in recombination rate on multiple scales in honeybees and bumblebees. Genetics 212:1101–1119.

Keightley, P. D. and S. P. Otto. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. Nature 443:89–92.

Kenneth Paigen, K. and P. M. Petkov. 2018. PRDM9 and its role in genetic recombination. Trends in Genetics 34:291–300.

Kerstes, N. A. G., C. Bérénos, P. Schmid-Hempel, and K. M. Wegner. 2012. Antagonistic experimental coevolution with a parasite increases host recombination frequency. BMC Evol. Biol. 12.

Kimura, M. 1956. A model of a genetic system which leads to closer linkage by natural selection. Evolution 10:278–287.

Kirkpatrick, M. 2010. How and why chromosome inversions evolve. PLoS Biology 8:e1000501.

Kirkpatrick, M., T. Johnson, and N. H. Barton. 2002. General models of multilocus evolution. Genetics 161:1727–1750.

Klepikova, A. V., A. S. Kasianov, E. S. Gerasimov, M. D. Logacheva, and A. A. Penin. 2016. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. Plant Journal 88:1058–1070.

Koehler, K. E., R. Scott Hawley, S. Sherman, and T. Hassold. 1996. Recombination and nondisjunction in human and flies. Human Molecular Genetics 5:1495–1504.

Kondrashov, A. S. 1984. Deleterious mutations as an evolutionary factor. I. The advantage of recombination. Genet. Res. 44:199–217.

Kong, A., G. Thorleifsson, M. L. Frigge, G. Masson, D. F. Gudbjartsson, R. Villemoes, E. Magnusdottir, S. B. Olafsdottir, U. Thorsteinsdottir, and K. Stefansson. 2014. Common and low-frequency variants associated with genome-wide recombination rate. Nat. Genet. 46:11–18.

Korol, A. B. and K. G. Iliadi. 1994. Increased recombination frequencies resulting from directional selection for geotaxis in *Drosophila*. Heredity 72:64–68.

Kouyos, R. D., O. K. Silander, and S. Bonhoeffer. 2007. Epistasis between deleterious mutations and the evolution of recombination. Trends Ecol. Evol. 22:308–315.

Kovalchuk, I., O. Kovalchuk, V. Kalck, V. Boyko, J. Filkowski, M. Heinlein, and B. Hohn. 2003. Pathogen-induced systemic plant signal triggers DNA rearrangements. Nature 423:760–762.

Kryvokhyzha, D., P. Milesi, T. Duan, M. Orsucci, S. I. Wright, S. Glémin, and M. Lascoux. 2019. Towards the new normal: transcriptomic convergence and genomic legacy of the two subgenomes of an allopolyploid weed (*Capsella bursa-pastoris*). PLoS Genetics 15:e1008131.

Kuittinen, H., A. A. de Haan, C. Vogl, S. Oikarinen, J. Leppälä, M. Koch, T. Mitchell-Olds, C. H. Langley, and O. Savolainen. 2004. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. Genetics 168:1575–1584.

Lafave, M. C. and J. Sekelsky. 2009. Mitotic recombination: why? when? how? where? PLoS Genetics 5:e1000411.

Lande, R. and D. W. Schemske. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. Evolution 39:24–40.

Latrille, T., L. Duret, and N. Lartillot. 2017. The red queen model of recombination hotspot evolution: a theoretical investigation. Philosophical Transactions of the Royal Society B: Biological Sciences 372:20160463.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall. 2014. LUMPY: A probabilistic framework for structural variant discovery. Genome Biology 15:R84.

Lehtonen, J., M. D. Jennions, and H. Kokko. 2012. The many costs of sex. Trends Ecol. Evol. 27:172–178.

Lenormand, T. 2003. The evolution of sex dimorphism in recombination. Genetics 163:811–822.

Lenski, R. E., M. J. Wiser, N. Ribeck, Z. D. Blount, J. R. Nahum, J. J. Morris, L. Zaman, C. B. Turner, B. D. Wade, R. Maddamsetti, A. R. Burmeister, E. J. Baird, J. Bundy, N. A. Grant, K. J. Card, M. Rowles, K. Weatherspoon, S. E. Papoulis, R. Sullivan, C. Clark, J. S. Mulka, and N. Hajela. 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. Proc. Roy. Soc. (Lond.) B 282:20152292.

Lewontin, R. C. and K.-I. Kojima. 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472.

Li, H. and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Lian, Q., V. Solier, B. Walkemeier, S. Durand, B. Huettel, K. Schneeberger, and R. Mercier. 2022. The megabase-scale crossover landscape is largely independent of sequence divergence. Nature Communications 13:3828.

Libuda, D. E., S. Uzawa, B. J. Meyer, and A. M. Villeneuve. 2013. Meiotic chromosome structures constrain and respond to designation of crossover sites. Nature 502:703–706.

Lynch, M. 2020. The evolutionary scaling of cellular traits imposed by the drift barrier. Proc. Natl. Acad. Sci. U. S. A. 117:10435–10444.

Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. 2016. Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17:704–714.

Lynch, M., J. Blanchard, D. Houle, T. Kibota, S. Schultz, L. Vassilieva, and J. Willis. 1999. Perspective: spontaneous deleterious mutation. Evolution 53:645–663.

Lynch, M. and W. Gabriel. 1990. Mutation load and the survival of small populations. Evolution 44:1725–1737.

Ma, L., J. R. O'Connell, P. M. VanRaden, B. Shen, A. Padhi, C. Sun, D. M. Bickhart, J. B. Cole, D. J. Null, G. E. Liu, Y. Da, and G. R. Wiggans. 2015. Cattle sex-specific recombination and genetic control from a large pedigree analysis. PLoS Genetics 11:e1005387.

Mable, B. K. and S. P. Otto. 1998. The evolution of life cycles with haploid and diploid phases. BioEssays 20:453–462.

MacKay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. MacKey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L. L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y. Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs. 2012. The *Drosophila melanogaster* Genetic Reference Panel. Nature 482:173–178.

Martin, G., S. F. Elena, and T. Lenormand. 2007. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat. Genet. 39:555–560.

Martin, G., S. P. Otto, and T. Lenormand. 2006. Selection for recombination in structured populations. Genetics 172:593–609.

McKinney, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb. 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Molecular Ecology Resources 17:656–669.

McVean, G. A. 2002. A genealogical interpretation of linkage disequilibrium. Genetics 162:987–991.

———. 2007. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 175:1395–1406.

Mérot, C., R. A. Oomen, A. Tigano, and M. Wellenreuther. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends Ecol. Evol. 35:561–572.

Moradigaravand, D. and J. Engelstädter. 2012. The effect of bacterial recombination on adaptation on fitness landscapes with limited peak accessibility. PLoS Computational Biology 8:e1002735.

Narra, H. P. and H. Ochman. 2006. Of what use is sex to bacteria? Current Biology 16.

Nei, M. 1967. Modification of linkage intensity by natural selection. Genetics 57:625–641.

Nei, M., Y. Suzuki, and M. Nozawa. 2010. The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genomics Hum. Genet. 11:265–289.

Nordborg, M. 1997. Structured coalescent processes on different time scales. Genetics 146:1501–1514.

———. 2000a. Coalescent theory. Pp. 1–37 *in* D. Balging and M. Bishop, eds. Handbook of Statistical Genetics. John Wiley and sons.

———. 2000b. Linkage disequilibrium, gene trees and selfing: and ancestral recombination graph with partial self-fertilization. Genetics 154:923–929.

Nordborg, M. and P. Donnelly. 1997. The coalescent process with selfing. Genetics 146:1185–1195.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N. A. Rosenberg, C. Shah, J. D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, and J. Bergelson. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biology 3:1289–1299.

Ohta, T. and M. Kimura. 1969. Linkage disequilibrium due to random genetic drift. Genet. Res. 13:47–55.

———. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68:570–580.

Orsucci, M., P. Milesi, J. Hansen, J. Girodolle, S. Glémin, and M. Lascoux. 2020. Shift in ecological strategy helps marginal populations of shepherd's purse (*Capsella bursa-pastoris*) to overcome a high genetic load: competition avoidance and colonization. Proc. Roy. Soc. (Lond.) B 287:20200463.

Otto, S. P. 2009. The evolutionary enigma of sex. Am. Nat. 174:S1–S14.

———. 2021. Selective interference and the evolution of sex. J. Hered. 112:9–18.

Otto, S. P. and N. H. Barton. 2001. Selection for recombination in small populations. Evolution 55:1921–1931.

Otto, S. P. and M. W. Feldman. 1997. Deleterious mutations, variable epistatic interactions, and the evolution of recombination. Theor. Popul. Biol. 51:134–47.

Otto, S. P. and T. Lenormand. 2002. Resolving the paradox of sex and recombination. Nat. Rev. Genet. 3:252–261.

Otto, S. P. and B. A. Payseur. 2019. Crossover interference: shedding light on the evolution of recombination. Ann. Rev. Gen. 53:19–44.

Payne, B. L. and D. Alvarez-Ponce. 2018. Higher rates of protein evolution in the self-fertilizing plant *Arabidopsis thaliana* than in the out-crossers *Arabidopsis lyrata* and *Arabidopsis halleri*. Genome Biol. Evol. 10:895–900.

Peñalba, J. V. and J. B. W. Wolf. 2020. From molecules to populations: appreciating and estimating recombination rate variation. Nat. Rev. Genet. 21:476–492.

Pérez-Losada, M., M. Arenas, J. C. Galán, F. Palero, and F. González-Candelas. 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. Infection, Genetics and Evolution 30:296–307.

Peters, A. D. and C. M. Lively. 1999. The red queen and fluctuating epistasis: a population genetic analysis of antagonistic coevolution. Am. Nat. 154:393–405.

———. 2007. Short- and long-term benefits and detriments to recombination under antagonistic coevolution. J. Evol. Biol. 20:1206–1217.

Petit, M., J.-M. Astruc, J. Sarry, L. Drouilhet, S. Fabre, C. R. Moreno, and B. Servin. 2017. Variation in recombination rate and its genetic determinism in sheep populations. Genetics 207:767–784.

Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445:383–386.

Pollak, E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117:353–360.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81:559–575.

Ragsdale, A. P. 2022. Local fitness and epistatic effects lead to distinct patterns of linkage disequilibrium in protein-coding genes. Genetics 221:iyac097.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. 2012. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339.

Redfield, R. J. 2001. Do bacteria have sex? Nature Reviews Genetics 2:634–639.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genetics 32:135–142.

Rice, W. R. 2002. Experimental tests of the adaptive significance of sexual recombination. Nat. Rev. Genet. 3:241–251.

Ritz, K. R., M. A. F. Noor, and N. D. Singh. 2017. Variation in recombination rate: adaptive or not? Trends Genet. 33:364–374.

Ross-Ibarra, J. 2007. Genome size and recombination in angiosperms: a second look. J. Evol. Biol. 20:800–806.

Roze, D. 2009. Diploidy, population structure and the evolution of recombination. Am. Nat. 174:S79–S94.

———. 2016. Background selection in partially selfing populations. Genetics 203:937–957.

———. 2021. A simple expression for the strength of selection on recombination generated by interference among mutations. Proc. Natl. Acad. Sci. U. S. A. 118:e2022805118.

Roze, D. and N. H. Barton. 2006. The Hill-Robertson effect and the evolution of recombination. Genetics 173:1793–1811.

Roze, D. and T. Lenormand. 2005. Self-fertilization and the evolution of recombination. Genetics 170:841–857.

Roze, D. and S. P. Otto. 2012. Differential selection between the sexes and selection for sex. Evolution 66:558–574.

Roze, D. and F. Rousset. 2005. Inbreeding depression and the evolution of dispersal rates: a multilocus model. Am. Nat. 166:708–721.

Salathé, M., R. D. Kouyos, and S. Bonhoeffer. 2009. On the causes of selection for recombination underlying the Red Queen hypothesis. Am. Nat. 174:S31–S42.

Salathé, M., R. D. Kouyos, R. R. Regoes, and S. Bonhoeffer. 2008. Rapid parasite adaptation drives selection for high recombination rates. Evolution 62:295–300.

Sandler, G., S. I. Wright, and A. F. Agrawal. 2021. Patterns and causes of signed linkage disequilibria in flies and plants. Mol. Biol. Evol. 38:4310–4321.

Sardell, J. M. and M. Kirkpatrick. 2020. Sex differences in the recombination landscape. American Naturalist 195:361–379.

Satomura, K., N. Osada, and T. Endo. 2019. Achiasmy and sex chromosome evolution. Ecological Genetics and Genomics 13:100046.

Schwander, T., R. Libbrecht, and L. Keller. 2014. Supergenes and complex phenotypes. Current Biology 24:R288–R294.

Sharp, N. P. and A. F. Agrawal. 2016. The decline in fitness with inbreeding: evidence for negative dominance-by-dominance epistasis in *Drosophila melanogaster*. J. Evol. Biol. 29:857–864.

Shipilina, D., S. Stankowski, A. Pal, Y. F. Chan, and N. H. Barton. 2022. On the origin and structure of haplotype blocks. Molecular Ecology 32:1441–1457.

Singh, N. D., D. R. Criscoe, S. Skolfield, K. P. Kohl, E. S. Keebaugh, and T. A. Schlenke. 2015. Fruit flies diversify their offspring in response to parasite infection. Science 349:742–747.

Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. M. Hooper, A. I. Alva I. Strand, Q. Li, B. Raney, C. N. Balakrishnan, S. C. Griffith, G. McVean, and M. Przeworski. 2015. Stable recombination hotspots in birds. Science 350:928–932.

Slatkin, M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9:477–485.

Slotte, T., K. M. Hazzouri, J. A. Ågren, D. Koenig, F. Maumus, Y. L. Guo, K. Steige, A. E. Platts, J. S. Escobar, L. K. Newman, W. Wang, T. Mandáková, E. Vello, L. M. Smith, S. R. Henz, J. Steffen, S. Takuno, Y. Brandvain, G. Coop, P. Andolfatto, T. T. Hu, M. Blanchette, R. M. Clark, H. Quesneville, M. Nordborg, B. S. Gaut, M. A. Lysak, J. Jenkins, J. Grimwood, J. Chapman, S. Prochnik, S. Shu, D. Rokhsar, J. Schmutz, D. Weigel, and S. I. Wright. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nature Genetics 45:831–835.

Smukowski Heil, C. S., C. Ellison, M. Dubin, and M. A. F. Noor. 2015. Recombining without hotspots: a comprehensive evolutionary portrait of recombination in two closely related species of *Drosophila*. Genome Biol. Evol. 7:2829–2842.

Sohail, M., O. A. Vakhrusheva, J. Sul, S. L. Pulit, L. C. Francioli, L. H. Van Den Berg, J. H. Veldink, P. I. De Bakker, G. A. Bazykin, A. S. Kondrashov, and S. R. Sunyaev. 2017. Negative selection in humans and fruit flies involves synergistic epistasis. Science 356:539–542.

Stapley, J., P. G. D. Feulner, S. E. Johnston, A. W. Santure, and C. M. Smadja. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Phil. Trans. Roy. Soc. (Lond.) B 372:20160455.

Stenøien, H. K., C. B. Fenster, A. Tonteri, and O. Savolainen. 2005. Genetic variability in natural populations of *Arabidopsis thaliana* in northern Europe. Molecular Ecology 14:137–148.

Stetsenko, R. and D. Roze. 2022. The evolution of recombination in self-fertilizing organisms. Genetics 222:iyac114.

Stolyarova, A. V., T. V. Neretina, E. A. Zvyagina, A. V. Fedotova, A. S. Kondrashov, and G. A. Bazykin. 2022. Complex fitness landscape shapes variation in a hyperpolymorphic species. eLife 11:e76073.

Sung, W., M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. U. S. A. 109:18488–18492.

Tenaillon, O. 2014. The utility of Fisher's geometric model in evolutionary genetics. Ann. Rev. Ecol. Syst. 45:179–201.

The 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491.

Thompson, M. J. and C. D. Jiggins. 2014. Supergenes and their role in evolution. Heredity 113:1–8.

Valero, M., S. Richerd, V. Perrot, and C. Destombe. 1992. Evolution of alternation of haploid and diploid phases in life cycles. Trends Ecol. Evol. 7:25–29.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. 2013. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics 43:11.10.1–11.10.33.

Vanhoenacker, E., L. Sandell, and D. Roze. 2018. Stabilizing selection, mutational bias, and the evolution of sex. Evolution 72:1740–1758.

Vaser, R., S. Adusumalli, S. N. Leng, M. Sikic, and P. C. C. Ng. 2016. SIFT missense predictions for genomes. Nature Protocols 11:1–9.

Vos, M. 2009. Why do bacteria engage in homologous recombination? Trends in Microbiology 17:226–232.

Žerdoner Čalasan, A., H. Hurka, D. A. A. German, S. Pfanzelt, F. R. Blattner, A. Seidl, and B. Neuffer. 2021. Pleistocene dynamics of the Eurasian steppe as a driving force of evolution: phylogenetic history of the genus *Capsella* (Brassicaceae). Ecology and Evolution 11:12697–12713.

Wang, R. J., M. M. Gray, M. D. Parmenter, K. W. Broman, and B. A. Payseur. 2017. Recombination rate variation in mice from an isolated island. Molecular Ecology 26:457–470.

Wang, Y. and G. P. Copenhaver. 2018. Meiotic recombination: mixing it up in plants. Annu. Rev. Plant Biol 69:577–609.

Weinreich, D. M. and L. Chao. 2005. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. Evolution 59:1175–1182.

Weinreich, D. M., N. F. Delaney, M. A. DePristo, and D. L. Hartl. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114.

Weir, B. S. 1996. Genetic Data Analysis II. Sinauer Associates, Sunderland, MA.

Weir, B. S. and C. C. Cockerham. 1973. Mixed self and random mating at two loci. Genet. Res. 21:247–262.

Weismann, A. 1889. The significance of sexual reproduction in the theory of natural selection. Pp. 251–332 *in* E. B. Poulton, S. Schönland, and A. E. Shipley, eds. Essays upon heredity and kindred biological problems. Clarendon, Oxford.

West, S. A., C. M. Lively, and A. F. Read. 1999. A pluralist approach to sex and recombination. J. Evol. Biol. 12:1003–1012.

Whitehead, M. R., R. Lanfear, R. J. Mitchell, and J. D. Karron. 2018. Plant mating systems often vary widely among populations. Frontiers in Ecology and Evolution 6:1–9.

Willi, Y., M. Fracassetti, S. Zoller, and J. Van Buskirk. 2018. Accumulation of mutational load at the edges of a species range. Mol. Biol. Evol. 35:781–791.

Wright, S. I., R. W. Ness, J. P. Foxe, and S. C. H. Barrett. 2008. Genomic consequences of outcrossing and selfing in plants. Int. J. Plant Sci. 169:105–118.

Xiao, J., M. K. Sekhwal, P. Li, R. Ragupathy, S. Cloutier, X. Wang, and F. M. You. 2016. Pseudogenes and their genome-wide prediction in plants. International Journal of Molecular Sciences 17:1991.

Zelkowski, M., M. A. Olson, M. Wang, and W. Pawlowski. 2019. Diversity and determinants of meiotic recombination landscapes. Trends Genet. 35:359–370.

Zhang, L., S. Wu, X. Chang, X. Wang, Y. Zhao, Y. Xia, R. N. Trigiano, Y. Jiao, and F. Chen. 2020. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. Plant Cell and Environment 43:2847–2856.

Zhao, H. and T. Speed. 1996. On genetic map functions. Genetics 142:1369–1377.

Zickler, D. and N. Kleckner. 2015. Recombination, pairing, and synapsis of homologs during meiosis. Cold Spring Harbor Perspectives in Biology 7:1–28.

Ziolkowski, P. A., L. E. Berchowitz, C. Lambing, N. E. Yelina, X. Zhao, K. A. Kelly, K. Choi, L. Ziolkowska, V. June, E. Sanchez-Moran, C. Franklin, G. P. Copenhaver, and I. R. Henderson. 2015. Juxtaposition of heterozygous and homozygous regions causes reciprocal crossover remodelling via interference during Arabidopsis meiosis. eLife 4:e03708.

Abstract: The evolution of recombination in self-fertilising species: theoretical approach and genomic study of linkage disequilibrium between deleterious mutations in several Angiosperm species

Genetic recombination is a building block of the sexual cycle of eukaryotes, and is often considered as one of the main benefits of sexual reproduction. It is the consequence of crossovers occurring during meiosis, enabling the formation of recombinant gametes. However, the number and position of these crossovers along the genome varies greatly between species, as well as between individuals within the same species. Despite the fact that mechanical constraints related to the segregation of chromosomes during meiosis can impose a minimum and maximum number of crossovers, several empirical results show that recombination rates can evolve rapidly. Interestingly, higher recombination rates are often observed in self-fertilising hermaphroditic species compared with outcrossing species, suggesting that recombination is more strongly favoured in selfing species. The first part of the thesis involved developing theoretical models to better understand the effect of selfing on the evolution of recombination. A first model considered the evolution of a 'modifier' gene affecting the average number of crossovers per chromosome, in the presence of deleterious mutations occurring along the chromosomes. Analytical approximations and simulation results show that, under realistic parameter values, selection for recombination is generally stronger in selfing species; moreover, this selection is mainly due to the negative linkage disequilibrium between deleterious mutations generated by the Hill-Robertson effect (a stochastic effect related to finite population size). A second simulation model explored more realistic scenarios concerning the genetic architecture of recombination rate variation, as well as the distribution of crossovers along the chromosomes. The results show, in particular, that higher recombination rates can be maintained when this distribution is not uniform, or when the effect of recombination modifiers is restricted to a portion of the chromosome. The second part of the thesis focused on an empirical estimation of the linkage disequilibrium between deleterious mutations (an important component of selection for recombination), using genomic data from natural populations of the outcrossing plant *Capsella grandiflora*, as well as the highly selfing plants *Arabidopsis thaliana* and *Capsella orientalis*. This study highlights several methodological biases that can generate positive linkage disequilibrium between deleterious mutations. The first source of bias arises when the analysis is restricted to low-frequency mutations. Indeed, mutations with similar frequencies tend to be in positive linkage disequilibrium. A second source of bias may be generated by the presence of duplications present in the polymorphic state in the sample and absent from the reference genome. The positive linkage disequilibrium between deleterious mutations observed in previous studies (in particular in *C. grandiflora*) is probably largely due to these biases. In the selfing species *A. thaliana* (in which duplications present in the polymorphic state are easier to detect) we nevertheless observed positive linkage disequilibrium between deleterious mutations, which may be caused by epistatic effects or by the strong spatial structure observed in this species. Finally, the implications of these various theoretical and empirical results are discussed, in order to identifying prospects for future works.

Résumé : L'évolution de la recombinaison chez les espèces autogames : approche théorique et étude génomique du déséquilibre de liaison entre mutations délétères chez plusieurs espèces d'Angiospermes.

La recombinaison génétique est un élément constitutif du cycle sexué des Eucaryotes, et est souvent considérée comme l'un des principaux avantages de la reproduction sexuée. Elle est la conséquence de crossing-over se produisant pendant la méiose, permettant la formation de gamètes recombinants. Or, le nombre et la position de ces crossing-over le long du génome est très variable entre espèces, ainsi qu'entre individus au sein d'une même espèce. Malgré le fait que des contraintes mécaniques liées à la ségrégation des chromosomes pendant la méiose peuvent imposer un nombre minimal et maximal de crossing-over, plusieurs résultats empiriques montrent que les taux de recombinaison peuvent évoluer rapidement. De façon intéressante, on observe souvent des taux de recombinaison plus élevés chez les espèces hermaphrodites autogames par rapport aux espèces allogames, suggérant que la recombinaison est plus fortement favorisée chez les autogames. Une première partie de la thèse a consisté à développer des modèles théoriques permettant de mieux comprendre cet effet de l'autofécondation sur l'évolution de la recombinaison. Un premier modèle a considéré l'évolution d'un gène "modificateur" affectant le nombre moyen de crossing-over par chromosome, en présence de mutations délétères se produisant le long des chromosomes. Les approximations analytiques ainsi que les résultats de simulation montrent que, sous des valeurs de paramètres réalistes, la sélection pour la recombinaison est généralement plus forte chez les espèces autogames ; par ailleurs, cette sélection est principalement due au déséquilibre de liaison négatif entre mutations délétères généré par l'effet Hill-Robertson (un effet stochastique lié à la taille finie des populations). Un deuxième modèle de simulation a exploré des scénarios plus réalistes concernant l'architecture génétique de la variation des taux de recombinaison, ainsi que la distribution des crossing-over le long des chromosomes. Les résultats montrent notamment que des taux de recombinaison plus élevés peuvent être maintenus lorsque cette distribution n'est pas uniforme, ou lorsque l'effet des modificateurs de recombinaison est restreint à une portion de chromosome. La deuxième partie de la thèse a porté sur une estimation empirique du déséquilibre de liaison entre mutations délétères (une composante importante de la sélection pour la recombinaison), en utilisant des données génomiques issues de populations naturelles de la plante allogame *Capsella grandiflora*, ainsi que des plantes fortement autogames *Arabidopsis thaliana* et *Capsella orientalis*. Cette étude met en avant plusieurs biais méthodologiques pouvant générer du déséquilibre de liaison positif entre mutations délétères. Une première source de biais se produit lorsque l'analyse est restreinte aux mutations présentes en faible fréquence. En effet, des mutations en fréquences similaires ont tendance à être en déséquilibre de liaison positif. Une deuxième source de biais peut être générée par la présence de duplications présentes à l'état polymorphe dans l'échantillon, et absentes du génome de référence. Ainsi, le déséquilibre de liaison positif entre mutations délétères observé lors de précédents travaux (notamment chez *C. grandiflora*) résulte probablement en grande partie de ces biais. Chez l'espèces autogame *A. thaliana* (chez qui les duplications présentes à l'état polymorphe sont plus faciles à détecter) on observe néanmoins du déséquilibre de liaison positif entre mutations délétères, pouvant être causé par des effets épistatiques ou par la forte structure spatiale observée chez cette espèce. Les implications de ces différents résultats théoriques et empiriques sont enfin discutées afin d'en dégager des perspectives pour des travaux futurs.