

The Recombination Hotspot Paradox: Co-evolution between PRDM9 and its target sites



Francisco Úbeda^{a,*}, Frédéric Fyon^{a,1}, Reinhard Bürger^{b,1}

^a Department of Biology, Royal Holloway University of London, Egham TW20 0EX, UK

^b Faculty of Mathematics, University of Vienna, 1090 Vienna, Austria

ARTICLE INFO

Article history:

Received 21 December 2022

Available online 13 July 2023

Dataset link: [10.5281/zenodo.8098998](https://doi.org/10.5281/zenodo.8098998)

Keywords:

Gene conversion
Mutation
Intragenomic conflict
Limit cycles
Red queen
Chaos

ABSTRACT

Recombination often concentrates in small regions called recombination hotspots where recombination is much higher than the genome's average. In many vertebrates, including humans, gene PRDM9 specifies which DNA motifs will be the target for breaks that initiate recombination, ultimately determining the location of recombination hotspots. Because the sequence that breaks (allowing recombination) is converted into the sequence that does not break (preventing recombination), the latter sequence is over-transmitted to future generations and recombination hotspots are self-destructive. Given their self-destructive nature, recombination hotspots should eventually become extinct in genomes where they are found. While empirical evidence shows that individual hotspots do become inactive over time (die), hotspots are abundant in many vertebrates: a contradiction called the Recombination Hotspot Paradox. What saves recombination hotspots from their foretold extinction? Here we formulate a co-evolutionary model of the interaction among sequence-specific gene conversion, fertility selection, and recurrent mutation. We find that allelic frequencies oscillate leading to stable limit cycles. From a biological perspective this means that when fertility selection is weaker than gene conversion, it cannot stop individual hotspots from dying but can save them from extinction by driving their re-activation (resuscitation). In our model, mutation balances death and resuscitation of hotspots, thus maintaining their number over evolutionary time. Interestingly, we find that multiple alleles result in oscillations that are chaotic and multiple targets in oscillations that are asynchronous between targets thus helping to maintain the average genomic recombination probability constant. Furthermore, we find that the level of expression of PRDM9 should control for the fraction of targets that are hotspots and the overall temperature of the genome. Therefore, our co-evolutionary model improves our understanding of how hotspots may be replaced, thus contributing to solve the Recombination Hotspot Paradox. From a more applied perspective our work provides testable predictions regarding the relation between mutation probability and fertility selection with life expectancy of hotspots.

© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Recombination—the exchange of genetic material between parental chromosomes—is a fundamental biological process that is key in generating DNA variability and trading information across genomes (Lenormand et al., 2016; Alves et al., 2017; Stapley et al., 2017; Peñalba and Wolf, 2020). It has a critical impact in ecology and evolution, e.g., selfish genes (Haig and Grafen, 1991), genomic architecture (Hansen, 2006), sexual reproduction (Otto and Lenormand, 2002), speciation (Butlin, 2005), and conservation (Frankham, 2005), as well as in human

disease, e.g., fertility (Nagaoka et al., 2012; Powers et al., 2020), cancer (Gerton and Hawley, 2005), association between genetic variants and illness (Slatkin, 2008; Rosenberg et al., 2010). The impact of recombination relies not only on its intensity but also on its distribution across the genome. Contrary to prior beliefs, recombination is not uniformly distributed across the genome but concentrated in small chromosomal regions where recombination is ten to a thousand times more frequent than the genome's average; these regions are known as *recombination hotspots* (Petes, 2001; Arnheim et al., 2007; Paigen and Petkov, 2010; Baudat et al., 2013) (although recombination hotspots are absent in some species such as *Drosophila melanogaster* (Heil et al., 2015)). In humans, most mammals and many vertebrates, the location of recombination hotspots is determined by the preference of alleles at the PRDM9 locus for binding specific DNA motifs (Myers et al., 2008, 2010; Baudat et al., 2010; Cavassim et al., 2022). Henceforth

* Corresponding author.

E-mail address: f.ubeda@rhul.ac.uk (F. Úbeda).

¹ All authors contributed equally.

we will use the term recombination hotspots to refer to these PRDM9-directed recombination hotspots.

Empirical work shows that recombination is initiated by a double-strand break (Padmore et al., 1991; Goyon and Lichten, 1993; Petes, 2001) and results in the conversion of the allele that breaks, enabling recombination (*hot allele*), into the allele that does not break, disabling recombination (*cold allele*) (Petes, 2001; Jeffreys and Neumann, 2002, 2005; Baudat et al., 2013). The mechanism that initiates recombination provides a transmission advantage to the cold allele over the hot allele (Petes, 2001; Jeffreys and Neumann, 2002, 2005; Baudat et al., 2013). It is thus a matter of time that recombination hotspots become chromosomal regions where recombination is as frequent as the average across the genome (Boulton et al., 1997) (henceforth *recombination coldspots*). This loss of activity is often referred to as the *death* of individual hotspots (Coop and Myers, 2007). The self-destructive nature of recombination hotspots means that they should become extinct in genomes where they are initially found (Boulton et al., 1997). However, evidence shows that far from being rare, recombination hotspots are abundant (Myers et al., 2005; Arnheim et al., 2007; Paigen and Petkov, 2010; Baudat et al., 2013). The *Recombination Hotspot Paradox* refers to the mismatch between the expected scarcity and observed abundance of recombination hotspots (Boulton et al., 1997): What saves recombination hotspots from their foretold extinction?

Consistent with the transient nature of individual hotspots and their paradoxical long-term persistence, humans and chimpanzees do not share many recombination hotspots and even human subpopulations exhibit some level of recombination hotspot variation (Ptak et al., 2004, 2005; Winckler et al., 2005; Coop et al., 2008; Stevison et al., 2015). This suggests that PRDM9 underlies rapidly changing recombinational landscapes (Ptak et al., 2004, 2005; Winckler et al., 2005; Coop et al., 2008; Stevison et al., 2015). Furthermore, empirical evidence indicates that PRDM9 evolves extraordinarily fast to the point that it has been considered the most rapidly evolving gene in many mammals (Ponting, 2011).

Because recombination underpins fundamental biological processes, the Recombination Hotspot Paradox has received much attention (Pineda-Krch and Redfield, 2005; Calabrese, 2007; Peters, 2008; Úbeda and Wilkins, 2011; Latrille et al., 2017; Úbeda et al., 2019). In spite of the attention, there is not a fully satisfactory solution to the paradox. The initial attempts to solve the paradox explored whether the beneficial effects of recombination on fertility—in particular how recombination favours proper chromosomal segregation during meiosis thus preventing the formation of aneuploid gametes (Fledel-Alon et al., 2009; Nagaoka et al., 2012; Ren et al., 2016; Powers et al., 2020)—can save recombination hotspots from extinction (Boulton et al., 1997; Pineda-Krch and Redfield, 2005; Calabrese, 2007; Peters, 2008). Evolutionary models found that cold alleles spread in the population due to their transmission advantage, even when these alleles reduce the fertility of individual carriers (Boulton et al., 1997; Pineda-Krch and Redfield, 2005; Calabrese, 2007; Peters, 2008). To maintain hot alleles, the benefits of recombination needed to be too strong to be realistic (Boulton et al., 1997; Pineda-Krch and Redfield, 2005; Calabrese, 2007; Peters, 2008). Furthermore, when fertility selection is strong enough, natural selection prevents individual hotspots from dying. This result contradicts empirical observations on the death of individual hotspots and the rapidly changing recombinational landscape over evolutionary time (Ptak et al., 2004, 2005; Winckler et al., 2005; Coop et al., 2008; Stevison et al., 2015; Ponting, 2011).

Advances in our understanding of the mechanisms initiating recombination found that, in many vertebrates, alleles at the PRDM9 locus code for a protein that binds a specific sequence

motif in hotspots (Myers et al., 2008, 2010; Baudat et al., 2010). Binding between protein and sequence causes a double-strand break (DSB) that initiates recombination at the binding site (Myers et al., 2010; Baudat et al., 2010). This observation led to the verbal argument that mutations in PRDM9 could create new recombination hotspots that counteract their individual loss due to gene conversion (Baudat et al., 2010). This verbal argument implicitly assumed that selection for a rare PRDM9 would drive this mutant to fixation in spite of the opposing effects of conversion on its target. The validity of this verbal argument was not supported by a mathematical model and there was no quantification of the strength of selection, if any, that could make this assumption true.

A first attempt to explore whether this argument was valid, used an agent-based model to simulate the co-evolution between PRDM9 and its target sites. This model found that fertility selection, gene conversion, mutation and genetic drift can result in the formation of new recombination hotspots (Úbeda and Wilkins, 2011). These authors argue that fertility selection could drive the birth of hotspots in numbers that compensate their death, thus leading to the long-term co-evolutionary oscillations between PRDM9 and its target motifs characteristic of Red Queen dynamics (Úbeda and Wilkins, 2011; Schenk et al., 2020). However, the model falls short to establish that fluctuating selection acting on PRDM9 and its target sites can result in the long-term oscillations. In particular, the long-term behaviour of the system is unclear. Due to computational limitations, the number of generations explored in the model of Úbeda and Wilkins (2011) is not enough to determine whether in the long-term the birth of hotspots would compensate their death and/or whether the intensity of hotspots will be maintained. Furthermore, the cause of any oscillatory behaviour remains obscure. Due to the number of processes considered (fertility selection, gene conversion, genetic drift, and mutation), it remained unclear whether the birth of hotspots could be attributed mostly to selection or drift. Thus a solution to the paradox remained elusive.

A second attempt to understand the long-term persistence of recombination hotspots, formulated a stochastic model of the evolution of PRDM9 (a one-locus model) under fertility selection, mutation and genetic drift (Latrille et al., 2017). This research also found that fertility selection, mutation and drift can lead to the birth of recombination hotspots. However, in Latrille et al. (2017) the dynamics of alleles at target sites was not modelled explicitly. The model assumed that mutations in PRDM9 result in fully functional recombination hotspots and that conversion at target sites is a function of the frequency of alleles at the PRDM9 locus. That is, Latrille et al. (2017) model the evolution of alleles at PRDM9, but ignore the evolutionary feedback of alleles at target sites. However, the dynamics of alleles at PRDM9 cannot be decoupled from the dynamics of alleles at its target sites—it is their co-evolution that determines whether target sites will be fully functional recombination hotspots or not, and also determines how conversion works as a function of the frequency of motifs at target sites. To show that long-term oscillations between PRDM9 and its target motifs—Red Queen dynamics—can solve the paradox, requires modelling the co-evolution of PRDM9 and its target sites (Schenk et al., 2017, 2020)—as opposed to modelling the evolution of PRDM9 given the assumed oscillatory behaviour of target sites.

The most recent attempt to understand the long-term persistence of recombination hotspots formulated a deterministic model of the co-evolution between PRDM9 and one target site (a two-locus model) under fertility selection and gene conversion (Úbeda et al., 2019). This research found that when a hotspot dies because its binding motif is replaced by a non-binding motif, fertility selection favours the spread of a mutant allele at PRDM9

that binds precisely the non-binding motif at the same recombination site. Thus, natural selection favours the reactivation of recombinational activity at the target site where recombinational activity ceased due to conversion (*resuscitation* of hotspots). The model in Úbeda et al. (2019) demonstrates that in the more realistic case when fertility selection is weak, the co-evolution between PRDM9 and its target sites is more complex than expected from verbal arguments, and results in two different dynamics: (i) when both alleles are frequent at the target site, co-evolutionary oscillations between PRDM9 alleles and its target motifs settle into intermediate recombination levels—midway between the ones in hotspots and coldspots, (ii) when mostly one allele segregates at the target site, co-evolutionary oscillations settle (in practical terms) either into a hotspot or a coldspot. In the long-term, the first dynamics results in target sites with mild and constant recombination whereas the second results in hotspots that do not die. These results are contrary to empirical observations on the death of individual hotspots and the rapidly changing recombinational landscape (Ptak et al., 2004, 2005; Winckler et al., 2005; Coop et al., 2008; Stevison et al., 2015; Ponting, 2011). The model in Úbeda et al. (2019) shows that resuscitation and/or birth of hotspots can either under-compensate or over-compensate their death but does not balance birth and death of hotspots across the genome over evolutionary time.

Here we advance the most recent attempt to model the long-term persistence of recombination hotspots (Úbeda et al., 2019) by incorporating recurrent mutation, both at the PRDM9 locus and its target sites—an evolutionary force neglected in Úbeda et al. (2019). We explore the possibility that the addition of mutation may result in Red Queen dynamics that, in turn, solve the paradox. We thus formulate a deterministic model of the co-evolution between PRDM9 and one target site (a two-locus model) under fertility selection, gene conversion and recurrent mutation at both loci. Furthermore, we explore the effect on the co-evolutionary dynamics of adding realism by having more than two alleles at each locus and more than one target locus (a three-locus model). Finally, we bridge the gap between deterministic predictions and stochastic observations by formulating and exploring a multilocus finite population version of our deterministic model.

From a theoretical perspective we discuss the implications of our model for solving the Recombination Hotspots Paradox. From a more applied perspective, we discuss the insight provided by our model into the mechanism and genomic distribution of PRDM9-directed recombination hotspots. These results allow us to predict how the life-expectancy of hotspots changes with the strength of fertility selection, conversion, mutation (at each of the loci), and the number of targets, thus opening the possibility of calibrating our model with empirical data.

2. Models

We model the interaction between one locus (the PRDM9 locus) coding for a protein that recognises specific motifs at one or more loci (the PRDM9 target loci) where crossover may be initiated. This is the behaviour of gene PRDM9 responsible for the location of recombination hotspots in humans and many vertebrates (Myers et al., 2010; Baudat et al., 2010, 2013; Cavassim et al., 2022).

First we introduce a deterministic model of the co-evolution between the PRDM9 locus and one of its target loci with any number of alleles at each of these two loci (a *two-locus multi-allele model*). This model describes the change in gamete frequencies under fertility selection, gene conversion and recurrent mutation at both loci, when the population is large enough that the effect of genetic drift can be ignored. However, our research focuses

on the special case when the two loci carry two alleles (a *two-locus two-allele model*). We diverge from previous co-evolutionary deterministic models mostly by considering the role of recurrent mutation at PRDM9 and target loci. In addition, we assume that recombination between PRDM9 and its target is independent from crossover at the target site and that there might be fertility costs due to lack of crossover as opposed to double-strand breaks at the target site, assumptions that also diverge from previous work (Úbeda et al., 2019).

To assess the predictive power of this simple model in more realistic scenarios, we increase the number of alleles at each locus and the number of target loci in a chromosome. First we explore the case of a one-target model with three alleles segregating at each locus. This exploration does not require the formulation of a new model but relies on the general one-target multi-allele model already formulated. Second we explore the case of a two-target model with two alleles segregating at each locus. This exploration requires the formulation of a two-target multi-allele model (a *three-locus multi-allele model*).

Finally, we introduce a stochastic model of the co-evolution between PRDM9 and multiple target loci with each of these loci segregating multiple target alleles (a *multi-locus multi-allele stochastic model*). This model adds genetic drift to the dynamics of the deterministic model, thus bridging the gap between the dynamics of infinite and finite populations.

Following previous work, throughout we assume a randomly mating diploid population with equivalent sexes and non-overlapping generations (Boulton et al., 1997; Pineda-Krch and Redfield, 2005; Úbeda and Wilkins, 2011; Latrille et al., 2017; Úbeda et al., 2019). We assume that recombination between PRDM9 and its targets occurs with probability r_m , where $0 \leq r_m \leq \frac{1}{2}$. In all presented numerical results we set $r_m = \frac{1}{2}$, as will be the case if PRDM9 and its targets lie on different chromosomes or are far apart on the same chromosome.

2.1. One target: Deterministic two-locus multi-allele model

In this model, the PRDM9 locus may carry alleles A_1, A_2, \dots, A_l , each encoding a protein that attempts to bind a sequence motif at a target locus B. Locus B may carry alleles B_1, B_2, \dots, B_k , each corresponding to a base pair motif that the protein produced by locus A may attempt to bind. Let $x_{i,k}$ be the frequency of type $A_i B_k$ in gametes. Notice that $0 \leq x_{i,k} \leq 1$ and $\sum_{i,k} x_{i,k} = 1$. Random union of gametes results in a zygote with ordered genotype $\frac{A_i B_k}{A_j B_l}$ with frequency $x_{i,k} x_{j,l}$. The probability that this zygote reaches adulthood is independent of its genotype, but its genotype determines the outcome of meiosis in adults. In each generation, both PRDM9 alleles in a diploid individual show the same level of expression. One protein from this pool is chosen at random (each type of protein having equal probability $\frac{1}{2}$ of being chosen) and attempts to bind one of the two target motifs at random (each target motif having equal probability $\frac{1}{2}$ of experiencing a binding attempt). We assume that there is only one binding attempt per individual per generation during meiosis. Therefore, in an individual with genotype $\frac{A_i B_k}{A_j B_l}$, four potential binding attempts can occur ($A_i \rightarrow B_k, A_i \rightarrow B_l, A_j \rightarrow B_k, A_j \rightarrow B_l$), each with equal probability $\frac{1}{4}$ (Fig. 1).

The attempt of protein A_i to bind motif B_k is successful and results in a double-strand break of allele B_k with probability $b_{i,k}$. The binding attempt is unsuccessful and does not result in a double-strand break with probability $1 - b_{i,k}$ (where $0 \leq b_{i,k} \leq 1$). The probability that during meiosis the protein produced by PRDM9 breaks target B_k is $\bar{b}_{i,k} = \frac{1}{2}(b_{i,k} + b_{j,k})$ because only one of the proteins produced by the two modifier alleles attempts a break. Analogously, the probability that PRDM9 breaks target B_l is $\bar{b}_{j,l} = \frac{1}{2}(b_{i,l} + b_{j,l})$. Therefore, the probability that the protein

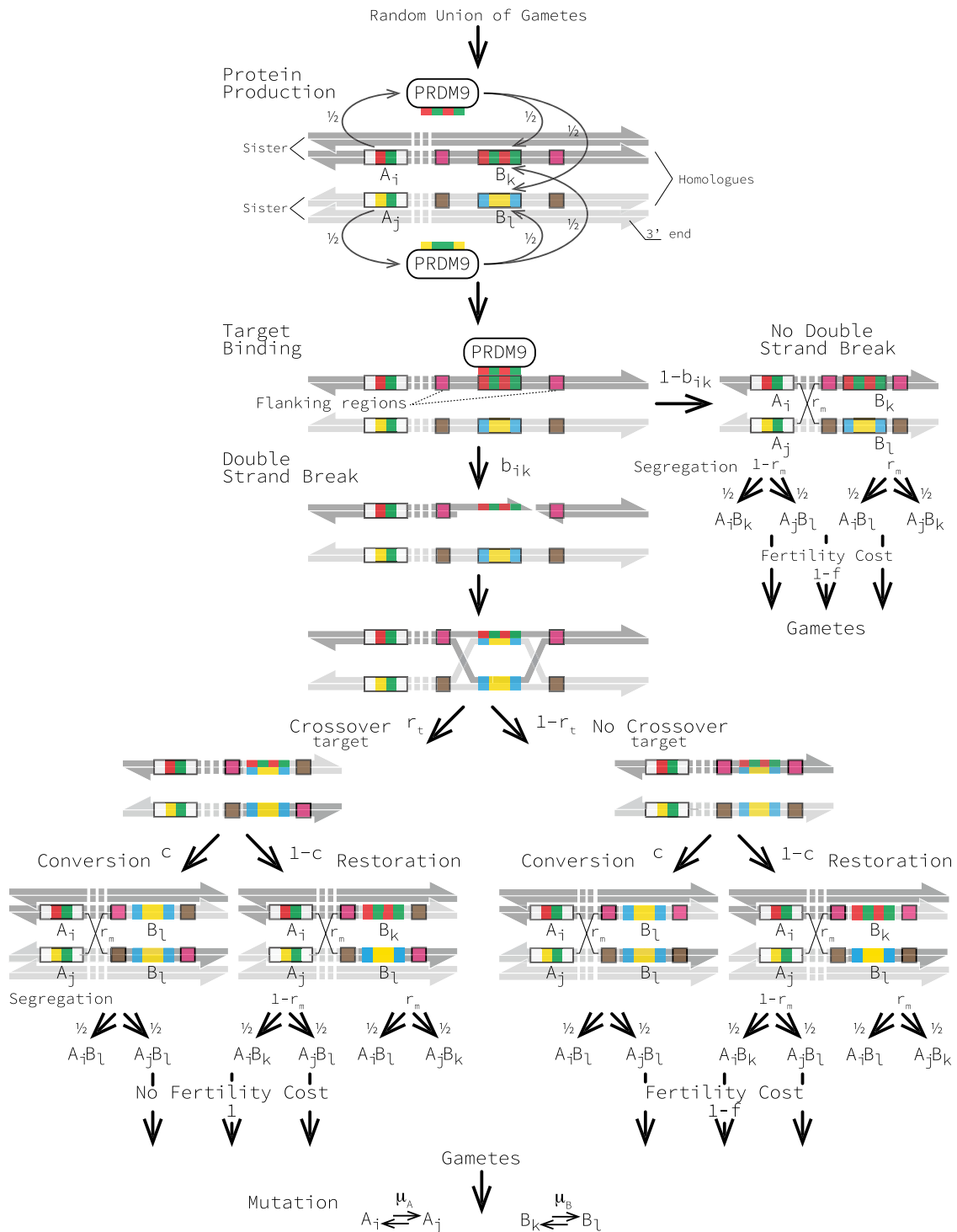


Fig. 1. Model of recombination initiated by a double strand break (DSB) in a specific target sequence. The life cycle starts with the PRDM9 production of a protein that attempts to bind a motif at the target locus of one pair of homologous chromatids. If protein and target bind (represented by a match in colour between recognition sequence and motif), there is a DSB with probability b , and our model follows the canonical DSB repair model for the initiation of recombination (Szostak et al., 1983; Sun et al., 1991). Recombination results in crossover between flanking regions of the target with probability r_t and conversion of the mismatched sequence with probability c . Recombination also results in crossover between alleles at PRDM9 and target locus on the homologous chromatids that experience a DSB attempt with probability r_m . Alleles segregate into gametes fairly. If there is no crossover at the target locus, alleles may not segregate properly resulting in non-viable gametes with probability f . Finally, mutation take place at each locus with probabilities μ_A at PRDM9, and μ_B at its target. For clarity, sister chromatids are represented in the first and last step only.

produced by PRDM9 breaks one of the two targets (either B_k or B_l) is $\bar{b}_{ij,kl} = \frac{1}{2}(\bar{b}_{ij,k} + \bar{b}_{ij,l})$ because only one of the two sequence motifs at the target locus can break (Fig. 1).

A double-strand break initiates recombination, and the chromatid that breaks is repaired using its homologous chromatid as

a template (Lichten and Goldman, 1995; Petes, 2001). During the repair process there might be a crossover event in or near the target locus with probability r_t and none with probability $1 - r_t$, where $0 \leq r_t \leq 1$ (Lichten and Goldman, 1995; Petes, 2001). During the repair process the allelic motif that breaks is converted

into the allelic motif that does not break with probability c and is restored to the allelic motif that breaks with probability $1 - c$, where $0 \leq c \leq 1$ (Szostak et al., 1983; Sun et al., 1991; Lichten and Goldman, 1995; Petes, 2001). Notice that biased gene conversion results in the over-transmission of the allele that is less likely to break (Boulton et al., 1997; Petes, 2001). We assume that recombination between PRDM9 and its target is independent of crossover events at the target site. Consistent with previous work (Úbeda et al., 2019), we assume that recombination affects the pair of homologous chromatids that experience a binding attempt only. With probability r_m there is recombination between PRDM9 and its target site and none with probability $1 - r_m$, where $0 \leq r_m \leq \frac{1}{2}$. Recombination ends up with Mendelian segregation of alleles into gametes (Fig. 1).

Consistent with empirical observations and previous work (Fledel-Alon et al., 2009; Nagaoka et al., 2012; Ren et al., 2016; Powers et al., 2020; Boulton et al., 1997; Pineda-Krch and Redfield, 2005; Peters, 2008; Úbeda et al., 2019), we assume that individuals undergoing crossover at the target locus have proper chromosomal segregation and do not suffer any fitness cost, whereas individuals that do not undergo crossover at the target locus may have defective chromosomal segregation producing aneuploid (non-viable) gametes with probability f (where $0 \leq f \leq 1$). Therefore, the fitness of individuals experiencing crossover at the target locus is 1, and the fitness of individuals not experiencing crossover is $1 - f$. Thus, we require a crossover at the target, which occurs with probability r_t after a double-strand break, to avoid the fitness cost f . Notice that this cost is independent of any recombination event between the PRDM9 and the target locus (Fig. 1).

The frequency of type $A_i B_k$ in gametes after gene conversion and fertility selection is

$$x'_{i,k} = \frac{1}{\bar{w}} \sum_{j,l} \left[\left(F \bar{b}_{ij,kl} + (1 - \bar{b}_{ij,kl})(1 - f) \right) x_{i,k} x_{j,l} - \frac{1}{4} c F \left(\bar{b}_{ij,k} x_{i,k} x_{j,l} - \bar{b}_{ij,l} x_{i,l} x_{j,k} \right) - r_m \left((1 - c) F \bar{b}_{ij,kl} + (1 - \bar{b}_{ij,kl})(1 - f) \right) \left(x_{i,k} x_{j,l} - x_{i,l} x_{j,k} \right) \right], \tag{1}$$

where $F = r_t + (1 - r_t)(1 - f)$ is the expected fitness in the presence of a double-strand break, and

$$\bar{w} = \sum_{i,k} \sum_{j,l} \left[F \bar{b}_{ij,kl} + (1 - \bar{b}_{ij,kl})(1 - f) \right] x_{i,k} x_{j,l} \tag{2}$$

is the population mean fitness. We note that depending on assumptions about the molecular mechanism during meiosis, there are various possibilities to model the interaction of recombination (between PRDM9 and its target) with gene conversion (at the target). They lead to slight variations in the above equation, but the structure and qualitative properties of the model remain unchanged (see Appendix A.1).

Finally, we add mutation to our model, both at PRDM9 and its target locus. We denote the mutation probability from $A_i \rightarrow A_j$ at locus A by $\mu_{A,ij}$ where $i \neq j$. The probability that A_i does not mutate is $\mu_{A,ii} = 1 - \sum_{j:j \neq i} \mu_{A,ij}$. Analogously, we denote the mutation probabilities at locus B by $\mu_{B,kl}$. We assume that the mutation probabilities are sufficiently small, so that the probabilities that no mutation occurs are close to one for all alleles A_i and B_k . Then the probability that the gametic type $A_j B_l$ changes to $A_i B_k$ as a result of a mutation is $\mu_{A,ji} \mu_{B,lk}$ (Fig. 1).

Gametic types segregate following Mendelian rules which brings us back to the beginning of the census. The frequencies of gametes after gene conversion, fertility selection, mutation and

reproduction are

$$x'_{i,k} = \sum_{j,l} x_{j,l}^{(rs)} \mu_{A,ji} \mu_{B,lk}, \tag{3}$$

where $x_{j,l}^{(rs)}$ is given in Eq. (1).

These changes in gametic type frequencies underpin changes in the population mean crossover probability at the target locus,

$$\bar{r}_t = \frac{1}{2} \sum_{i,k} \sum_{j,l} \bar{b}_{ij,kl} r_t x_{i,k} x_{j,l}, \tag{4}$$

which is the phenotype whose evolution we are interested in.

2.2. One target: Deterministic two-locus two-allele model

We specify the general model in the previous section for the particular case when there are two alleles (A_1, A_2) at the PRDM9 locus and two alleles (B_1, B_2) at its only target locus, resulting in four different gametic types ($A_1 B_1, A_1 B_2, A_2 B_1, A_2 B_2$). We assume that a match between the subscripts of the PRDM9 allele and the target allelic sequence results in a double-strand break with probability b , that is $b_{i,k} = b$ if $i = k$ where $0 \leq b \leq 1$ and a mismatch between the subscripts prevents a double-strand break, that is $b_{i,k} = 0$ if $i \neq k$. As such, whether alleles at the target loci are hot or cold depends on the allele at the PRDM9 locus, with alleles B_1 and B_2 being hot and cold respectively on a A_1 genetic background, but the same alleles B_1 and B_2 being cold and hot respectively on a A_2 genetic background. Similarly, two of the gametic types in this model ($A_1 B_1, A_2 B_2$) correspond to types producing a protein that matches its own target sequence (recombination enabling types) and the other two ($A_1 B_2, A_2 B_1$) correspond to gametic types producing a protein that does not match its own target sequence (recombination disabling types).

2.2.1. Gamete frequency dynamics

The dynamic system describing the change in frequency over time of each of these gametic types can be obtained by replacing generic subscripts i and k by specific subscripts 1 and 2 in Eq. (1). In addition, we switch to the notation: $x_1 = x_{1,1}, x_2 = x_{1,2}, x_3 = x_{2,1}, x_4 = x_{2,2}$. Applying straightforward algebra to the multi-allelic system (1) we find that the frequencies of the gametic types after gene conversion and fertility selection satisfy

$$\begin{aligned} \bar{w} x_1^{(rs)} &= w_1 x_1 - \delta_{r_m} D, \\ \bar{w} x_2^{(rs)} &= w_2 x_2 + \delta_{r_m} D, \\ \bar{w} x_3^{(rs)} &= w_3 x_3 + \delta_{r_m} D, \\ \bar{w} x_4^{(rs)} &= w_4 x_4 - \delta_{r_m} D, \end{aligned} \tag{5}$$

where

$$\begin{aligned} w_1 &= 1 - (1 - br_t)fx_1 - \left((1 - \frac{1}{2}br_t)f + \frac{1}{4}bc(1 - f(1 - r_t)) \right) x_2 \\ &\quad - (1 - \frac{1}{2}br_t)fx_3 - (1 - \frac{1}{2}br_t)fx_4, \\ w_2 &= 1 - \left((1 - \frac{1}{2}br_t)f - \frac{1}{4}bc(1 - f(1 - r_t)) \right) x_1 - fx_2 \\ &\quad - (1 - \frac{1}{2}br_t)fx_3 - (1 - \frac{1}{2}br_t)fx_4, \\ w_3 &= 1 - (1 - \frac{1}{2}br_t)fx_1 - (1 - \frac{1}{2}br_t)fx_2 \\ &\quad - fx_3 - \left((1 - \frac{1}{2}br_t)f - \frac{1}{4}bc(1 - f(1 - r_t)) \right) x_4, \\ w_4 &= 1 - (1 - \frac{1}{2}br_t)fx_1 - (1 - \frac{1}{2}br_t)fx_2 \\ &\quad - \left((1 - \frac{1}{2}br_t)f + \frac{1}{4}bc(1 - f(1 - r_t)) \right) x_3 - (1 - br_t)fx_4, \end{aligned} \tag{6}$$

$$\bar{w} = \sum_{i=1}^4 x_i w_i = 1 - f(1 - \frac{1}{2}br_t) + \frac{1}{2}br_t f(x_1^2 + x_4^2 - x_2^2 - x_3^2) \tag{7}$$

is the population mean fitness, which is not affected by conversion,

$$D = x_1x_4 - x_2x_3 \tag{8}$$

is the classical measure of linkage disequilibrium, and

$$\delta_{r_m} = \underbrace{\left(1 - \frac{1}{2}b\right)}_{\text{no break}} \underbrace{(1-f)}_{\text{fertility}} r_m + \underbrace{\frac{1}{2}b}_{\text{break}} \underbrace{(1-f(1-r_t))}_{\text{fertility}} \underbrace{\left(\frac{1}{4}c + (1-c)r_m\right)}_{\text{conversion}} \tag{9}$$

is the *effective recombination rate* (Supplementary *Mathematica* notebook, Section 1). The system (5) has the same structure as a classical two-locus two-allele model with selection and recombination (cf. Bürger, 2000, Chap. II). The effective recombination rate δ_{r_m} invokes not only the physical recombination rate between PRDM9 and its target but also the effects of gene conversion and of fertility selection because, as in classical two-locus models, it has to incorporate the fitnesses of the double heterozygotes. If $r_m = \frac{1}{2}$ and $r_t = 1$ (as we assume in all our numerical analyses), δ_{r_m} simplifies to

$$\delta = \frac{1}{2} \left(1 - \left(1 - \frac{1}{2}b \right) f - \frac{1}{4}bc \right). \tag{10}$$

Finally, mutation modifies the gamete frequencies after gene conversion and fertility selection, $x_i^{(rs)}$, according to (3).

In order to visibly reflect the effects of fertility selection and gene conversion on the selection terms in (5), we define the 4×4 matrix

$$\mathbf{B} = b \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \tag{11}$$

of break probabilities, the 4×4 matrix

$$\mathbf{C} = c \begin{pmatrix} 0 & -\frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & 0 \end{pmatrix}, \tag{12}$$

of conversion probabilities where entries (c_{ij}) measure the expected gain of gamete i caused by conversion to j (which is negative if i is converted to j , and positive if j results from conversion of i), and the 4×4 matrix \mathbf{U} of ones. Then the ‘fitness’ matrix \mathbf{W} underlying the system (5) is given by

$$\mathbf{W} = \mathbf{U} - f(\mathbf{U} - r_t\mathbf{B}) + (1 - f(1 - r_t))\mathbf{B} \circ \mathbf{C}, \tag{13}$$

where \circ denotes the pointwise (Schur) product of matrices. This is easy to check by defining the column vector $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ (T denotes transposition), which contains the four gamete frequencies, and observing that $w_i = \sum_{j=1}^4 \mathbf{W}_{ij}x_j$ and $\sum_{i=1}^4 x_i = 1$. However, because in our model \mathbf{W} combines the effects of fertility selection and gene conversion, it is not symmetric and the entry \mathbf{W}_{ij} is best interpreted in the game-theoretic sense as the payoff of gamete i if ‘playing’ against gamete j .

Because mutations at the two loci occur independently, the mutation matrix \mathbf{M} for gamete types is the Kronecker product

$$\mathbf{M} = \mathbf{M}_A \otimes \mathbf{M}_B \tag{14a}$$

of the two one-locus mutation matrices

$$\mathbf{M}_A = \begin{pmatrix} 1 - \mu_A & \mu_A \\ \mu_A & 1 - \mu_A \end{pmatrix} \text{ and } \mathbf{M}_B = \begin{pmatrix} 1 - \mu_B & \mu_B \\ \mu_B & 1 - \mu_B \end{pmatrix}. \tag{14b}$$

With these ingredients, the vector $\mathbf{x}' = (x'_1, x'_2, x'_3, x'_4)^T$ of gametic frequencies after gene conversion, fertility selection, and mutation is given by

$$\mathbf{x}' = \frac{1}{\bar{w}} \mathbf{M}((\mathbf{W}\mathbf{x}) \circ \mathbf{x} - \mathbf{d}\delta D), \tag{15}$$

where $\mathbf{d} = (1, -1, -1, 1)^T$ determines the signs of the linkage disequilibrium terms. The state space of this dynamical system is the so-called 4-dimensional simplex, which consists of all vectors $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ of (biologically admissible) gamete frequencies (i.e., satisfying $x_i \geq 0$ and $\sum_{i=1}^4 x_i = 1$). It is representable as a tetrahedron.

To simplify the matrix \mathbf{W} and exposit its structure, we define the parameters

$$\begin{aligned} \alpha &= 1 - f(1 - \frac{1}{2}br_t), \\ \beta &= \frac{1}{2}br_t f, \\ \gamma &= \frac{1}{4}bc(1 - f(1 - r_t)). \end{aligned} \tag{16}$$

Then \mathbf{W} can be written as

$$\mathbf{W} = \begin{pmatrix} \alpha + \beta & \alpha - \gamma & \alpha & \alpha \\ \alpha + \gamma & \alpha - \beta & \alpha & \alpha \\ \alpha & \alpha & \alpha - \beta & \alpha + \gamma \\ \alpha & \alpha & \alpha - \gamma & \alpha + \beta \end{pmatrix} \tag{17}$$

and the mean fitness becomes

$$\bar{w} = \alpha + \beta(x_1^2 - x_2^2 - x_3^2 + x_4^2). \tag{18}$$

We note that $0 < \alpha, \beta, \gamma, \delta$ and

$$\beta < \alpha \leq 1, \gamma < \alpha, \text{ and } \frac{1}{2}\gamma \leq \delta < \frac{1}{2}. \tag{19}$$

The matrix \mathbf{W} is reminiscent of the so-called symmetric viability model (Karlin and Feldman, 1970; Bürger, 2020). It differs because of the asymmetry caused by gene conversion via the parameter γ . Thus, conversion induces a frequency-dependent selection component. In addition, it induces negative epistasis because it provides an advantage to A_1B_2 over A_1B_1 and to A_2B_1 over A_2B_2 . In the absence of conversion, when only fertility selection acts ($\gamma = 0, \beta > 0$), the matrix \mathbf{W} becomes diagonal and a special case of the symmetric viability model. As is obvious from the resulting matrix, fertility selection induces positive epistasis and divergent selection on the double homozygotes. Further, \mathbf{W} exhibits also strong dominance.

From (15), (17), and (18), we readily infer that the dynamical equations remain unchanged if we set $\alpha = 1$ by applying the rescaling $\beta \rightarrow \beta/\alpha, \gamma \rightarrow \gamma/\alpha$, and $\delta \rightarrow \delta/\alpha$. For the rest of this paper we assume that $\alpha = 1$ and β, γ , and δ have been rescaled, which does not affect the ratio of c and f .

2.2.2. Allele-frequency and linkage disequilibrium dynamics

Often it yields more insight to perform the analysis of the two-locus two-alleles model in terms of the allele frequencies, $p = x_1 + x_2$ of A_1 and $q = x_1 + x_3$ of B_1 , and the linkage disequilibrium, D , instead of the gamete frequencies. The well known transformation from (p, q, D) coordinates to (x_1, x_2, x_3, x_4) coordinates is given by $x_1 = pq + D, x_2 = p(1 - q) - D, x_3 = (1 - p)q - D$, and $x_4 = (1 - p)(1 - q) + D$. Then the mean fitness can be rewritten as

$$\bar{w} = 1 + \beta[(2p - 1)(2q - 1) + 2D]. \tag{20}$$

Straightforward computations (Supplementary *Mathematica* notebook, Section 2.2) yield the following set of equations for the change in allele frequencies and in LD caused by recombination and selection:

$$\bar{w}\Delta^{(rs)}p = \beta p(1 - p)(2q - 1), \tag{21a}$$

$$\bar{w}\Delta^{(rs)}q = -(\gamma - \beta)q(1 - q)(2p - 1) + \gamma(2q - 1)D, \tag{21b}$$

$$\begin{aligned} \bar{w}^2 \Delta^{(rs)} D &= -(\gamma - \beta)p(1 - p)q(1 - q) \\ &\quad - \left\{ [1 + \beta(2p - 1)(2q - 1)] [\delta + \beta(2p - 1)(2q - 1)] \right. \\ &\quad \left. + \beta p(1 - p)[\gamma - 2(\beta + \gamma)q(1 - q)] \right\} D \\ &\quad + \left\{ \gamma - \beta + \beta(\gamma - 3\beta)(2p - 1)(2q - 1) - 2\beta\delta \right\} D^2 \\ &\quad + 2\beta(\gamma - \beta)D^3. \end{aligned} \tag{21c}$$

The term on the right-hand side of (21a) represents direct selection on locus A. Its strength depends on β (thus on f) and, through epistasis, on the allele frequency at locus B, as signified by the factor $(2q - 1)$. Striking is the absence of a term invoking D . In typical two-locus models, D is multiplied by a factor reflecting the strength of direct (additive and epistatic) selection on the other locus (for a very general treatment see Barton and Turelli (1991)). With epistasis or dominance this factor is typically frequency dependent. In the present model, all contributing terms cancel due to the special structure of the matrix \mathbf{W} , i.e., the symmetry along the diagonal caused by fertility selection, the antisymmetry across the diagonal induced by conversion, and the lack of selection in the right upper and the left lower corners, which is due to the lack of interaction between gametes carrying different alleles at the PRDM9 locus. In (21b), the first additive term on the right-hand side is due to direct selection on locus B, which is again dependent on the allele frequency at the other locus. The factor $(\gamma - \beta)$ shows that fertility selection on the target is counteracted by conversion. The second term arises from indirect selection on B transmitted by linkage disequilibrium D with A. Interestingly, D is multiplied by $\gamma(2q - 1)$ but not by a factor depending on β , which is responsible for direct selection on the other locus. Again, the reason for this deviation from intuition lies in the specific form of epistasis, dominance, and frequency dependence induced by the interaction of fertility selection and biased gene conversion.

Eq. (21c) for the change of linkage disequilibrium is complicated but shows immediately that if $D = 0$ (and all alleles are present), then $\Delta^{(rs)} D < 0$ if $\gamma > \beta$, and $\Delta^{(rs)} D > 0$ if $\gamma < \beta$. The term in the first line is present because fitnesses are not multiplicative (cf. Bürger, 2000, Chap. II). The sign of $-(\gamma - \beta)$ confirms that fertility selection induces positive epistasis, hence positive linkage disequilibrium, and conversion induces negative epistasis and negative linkage disequilibrium.

With gene conversion, fertility selection and mutation, the between-generation changes of p , q , and D can be written as (Supplementary Mathematica notebook, Section 3.3)

$$\Delta p = \Delta^{(rs)} p - 2\mu_A(\Delta^{(rs)} p + p - \frac{1}{2}), \tag{22a}$$

$$\Delta q = \Delta^{(rs)} q - 2\mu_B(\Delta^{(rs)} q + q - \frac{1}{2}), \tag{22b}$$

$$\Delta D = \Delta^{(rs)} D - 2[\mu_A(1 - \mu_B) + (1 - \mu_A)\mu_B](\Delta^{(rs)} D + D). \tag{22c}$$

If mutation is sufficiently much weaker than selection and recombination, terms of order two and higher in the mutation probabilities can be ignored, which slightly simplifies these equations.

The phenotype we are interested in characterising is the population mean crossover probability at the target site:

$$\bar{r}_t = \mathbf{x}^T \mathbf{B} r_t \mathbf{x} = \frac{1}{2} b(x_1 + x_4 - D)r_t, \tag{23}$$

where the latter equality results after some algebra using that $\sum_i x_i = 1$. For simplicity, and because we are mainly interested in the ratio \bar{r}_t/r_t , in this manuscript we assume $r_t = 1$. We note that \bar{r}_t/r_t assumes values between 0 (only recombination disabling types in the population) and $\frac{1}{2}b$ (only recombination enabling types present).

2.3. Two targets: Deterministic three-locus multi-allele model

In this model we consider three loci—one PRDM9 locus and two target locus—with multiple alleles at each target. We assume that both target loci are on the same chromosome.

As in our previous model, the PRDM9 locus segregate alleles A_1, A_2, \dots, A_l , each encoding a protein that attempts to bind a sequence motif at target loci B and C. Locus B and C segregate the same set of alleles T_1, T_2, \dots, T_k , each corresponding to a base pair motif that the protein produced by locus A may attempt to bind. Let $x_{i,k,m}$ be the frequency of type $A_i B_k C_m$ in gametes. Notice that $0 \leq x_{i,k,m} \leq 1$ and $\sum_{i,k,m} x_{i,k,m} = 1$. Random union of gametes results in a zygote with ordered genotype $\frac{A_i B_k C_m}{A_j B_l C_n}$ with frequency $x_{i,k,m} x_{j,l,n}$.

In each generation, both PRDM9 alleles in a diploid individual show the same level of expression. One protein from this pool is chosen at random (each type of protein having equal probability $\frac{1}{2}$ of being chosen) and attempts to bind one of the four target motifs at random (each target motif having equal probability $\frac{1}{4}$ of experiencing a binding attempt). We assume that there is only one binding attempt per individual per generation during meiosis. Therefore, in an individual with genotype $\frac{A_i B_k C_m}{A_j B_l C_n}$, four potential binding attempts can occur ($A_i \rightarrow B_k, A_i \rightarrow B_l, A_i \rightarrow C_m, A_i \rightarrow C_n, A_j \rightarrow B_k, A_j \rightarrow B_l, A_j \rightarrow C_m, A_j \rightarrow C_n$), each with probability $\frac{1}{4}$.

The attempt of protein A_i to bind either motif B_k or C_m is successful and results in a double-strand break with probabilities $b_{i,k}$ or $b_{i,m}$ respectively (where $0 \leq b_{i,k}, b_{i,m} \leq 1$). During meiosis the protein produced by PRDM9 breaks either target B_k or C_m with probabilities $\bar{b}_{ij,k} = \frac{1}{2}(b_{i,k} + b_{j,k})$ or $\bar{b}_{ij,m} = \frac{1}{2}(b_{i,m} + b_{j,m})$ respectively. The protein produced by PRDM9 breaks one of the two targets in either locus B or C with probabilities $\bar{b}_{ij,kl} = \frac{1}{2}(\bar{b}_{ij,k} + \bar{b}_{ij,l})$ or $\bar{b}_{ij,mn} = \frac{1}{2}(\bar{b}_{ij,m} + \bar{b}_{ij,n})$. Finally, the protein produced by PRDM9 breaks one target at one locus—because only one break is possible—with probability $\bar{b}_{ij,kl,mn} = \frac{1}{2}(\bar{b}_{ij,kl} + \bar{b}_{ij,mn})$.

A double-strand break initiates recombination at target B or C. This process results in crossover near one of the target loci with probability r_t , where $0 \leq r_t \leq 1$, and the conversion of the allelic motif that breaks into the allelic motif that does not break with probability c , where $0 \leq c \leq 1$. We assume that recombination between PRDM9 and its targets is independent of crossover at each target site. For simplicity, we assume that recombination affects both pairs of homologous chromatids the pair that experiences the binding attempt and the pair that does not. The probability of recombination between loci A and B, B and C, and A and C is r_m in all cases, where $0 \leq r_m \leq \frac{1}{2}$. Recombination ends up with Mendelian segregation of alleles into gametes.

Finally, we assume that individuals undergoing crossover at one of the two target loci have proper chromosomal segregation and do not suffer any fitness cost, whereas individuals that do not undergo crossover at neither of the two target loci may have defective chromosomal segregation with probability f (where $0 \leq f \leq 1$). This is consistent with the assumption that both target loci reside in the same chromosome. Therefore, the fitness of individuals experiencing crossover at one of the target loci is F , and the fitness of individuals not experiencing crossover at any of the target loci is $1 - f$. Notice that this cost is independent of any recombination event between the PRDM9 and its target loci. For simplicity we assume $r_t = 1$ and thus $F = 1$.

For clarity, we separate the steps of recombination between the three loci modelled (A, B, C), and recombination between the flanking regions of each of the targets (B and C). We start by considering recombination between the three loci first, followed by recombination between the flanking regions of the targets.

Other than clarity, there is no particular reason to chose this order of events and reversing the order does not change the outcome.

The frequency of genotype $\frac{A_i B_k C_m}{A_j B_l C_n}$ following recombination between the loci is

$$(x_{i,k,m} x_{j,l,n})^{(r)} = (1 - r_m)^2 x_{i,k,m} x_{j,l,n} + r_m (1 - r_m) (x_{i,l,n} x_{j,k,m} + x_{i,k,n} x_{j,l,m}) + r_m^2 x_{i,l,m} x_{j,k,n} \tag{24}$$

Then the frequency of the gametic type $A_i B_k C_m$ following selection is

$$x_{i,k,m}^{(rs)} = \frac{1}{\bar{w}} \sum_{j,l,n} \left[\left(1 - f(1 - \bar{b}_{ij,kl,mn}) - \frac{1}{8}c (\bar{b}_{ij,k} + \bar{b}_{ij,m}) \right) (x_{i,k,m} x_{j,l,n})^{(r)} + \frac{1}{8}c (\bar{b}_{ij,l} (x_{i,l,m} x_{j,k,n})^{(r)} + \bar{b}_{ij,n} (x_{i,k,n} x_{j,l,m})^{(r)}) \right],$$

where

$$\bar{w} = \sum_{i,k,m} \sum_{j,l,n} \left[1 - f(1 - \bar{b}_{ij,kl,mn}) \right] x_{i,k,m} x_{j,l,n} \tag{25}$$

Finally, recurrent mutation at the three loci A, B and C takes place. Let $\mu_{C,mn}$ be the mutation probability from $C_m \rightarrow C_n$ where $m \neq n$. For simplicity, we assume that mutation occurs as frequently in one direction as in the opposite, that is $\mu_{A,ij} = \mu_{A,ji} = \mu_A$, $\mu_{B,kl} = \mu_{B,lk} = \mu_B$ and $\mu_{C,mn} = \mu_{C,nm} = \mu_C$. We also assume that mutations are sufficiently rare to ignore multiple mutations. Then the frequency of the gametic type $A_i B_k C_m$ at the beginning of the next generation is

$$x'_{i,k,m} = x_{i,k,m}^{(rs)} \left(1 - \sum_{j:j \neq i} \mu_A - \sum_{l:l \neq k} \mu_B - \sum_{n:n \neq m} \mu_C \right) + \sum_{j:j \neq i} x_{j,k,m}^{(rs)} \mu_A + \sum_{l:l \neq k} x_{i,l,m}^{(rs)} \mu_B + \sum_{n:n \neq m} x_{i,k,n}^{(rs)} \mu_C \tag{26}$$

These changes in frequency over the life cycle underpin changes in the population mean crossover probability in targets B and C, denoted by $\bar{r}_{t(B)}$ and $\bar{r}_{t(C)}$, respectively, and the population mean crossover probability across all targets (*genomic mean crossover probability*), denoted by \bar{r}_t . The expressions for these phenotypes are

$$\bar{r}_{t(B)} = \sum_{i,k,m} \sum_{j,l,n} \bar{b}_{ij,kl} r_t x_{i,k,m} x_{j,l,n} \tag{27a}$$

$$\bar{r}_{t(C)} = \sum_{i,k,m} \sum_{j,l,n} \bar{b}_{ij,mn} r_t x_{i,k,m} x_{j,l,n} \tag{27b}$$

$$\bar{r}_t = \frac{1}{2} \bar{r}_{t(B)} + \frac{1}{2} \bar{r}_{t(C)} \tag{27c}$$

2.4. Stochastic multi-target multi-allele model

To bridge the gap between predictions driven by selection only and observations driven by selection and genetic drift we formulate a stochastic version of our deterministic model. In particular, we formulate a multi-locus multi-allele model.

We consider a population of N diploid individuals that carry one chromosome where one PRDM9 locus and a variable number of target sites are located. We assume that all loci (PRDM9 and its targets) recombine freely. That is, the probability of recombination between any pair of loci is the same r_m and equal to one half, i.e., $r_m = \frac{1}{2}$. We assume that PRDM9 segregates as many alleles as alleles segregate at all target sites, whereas each target site segregates only two alleles. The life cycle of our finite population is equivalent to the life cycle of our infinite population.

We model recombination between flanking regions following recombination between loci. We start by randomly choosing, with equal probability $\frac{1}{2}$, one of the two allelic copies in PRDM9 to

produce the protein that will attempt to bind one of the target loci. In contrast with our deterministic models, we allow for more than one binding attempt per meiotic event per generation. To allow comparisons between models we assume that there is a single break attempt in all cases except one the case explored in Fig. 10. We choose randomly, with equal probability, one or multiple target loci to experience the binding attempt of a PRDM9 protein. We choose randomly, with equal probability $\frac{1}{2}$, one of the two sequence motifs at the target site to be the motif that the protein attempts to bind. If the recognition sequence in the protein and the sequence motif in the target match (signified by the match between subscripts of the allele producing the protein and the allele object of a binding attempt), protein and target bind and result in a double-strand break with probability b . In case of a double-strand break, the broken motif is converted into its unbroken homologue with probability c , and all gametes produced by that individual are fully viable. In contrast, if there is no break, the broken motif is restored to its original sequence with probability $1 - c$ and with probability f the gametes produced by that individual are not viable.

Finally, Mendelian segregation is modelled by choosing randomly, with equal probability $\frac{1}{2}$, one of the two possible homologues to be included in the gamete pool produced by parents. Following segregation we model recurrent mutation at all loci. Mutations occur with probability μ_A at the PRDM9 locus. For simplicity we assume that mutation transforms each allele at PRDM9 into all other possible types with equal probability. Mutations occur with probability μ_B at all target loci. Similarly, we assume that mutations transform each allele at targets into all other possible types within each target with equal probability. We model selection by randomly choosing two gametes from the gamete pool. Each chosen gamete is retained with probability equal to its fertility (1 or $1 - f$). If a gamete is discarded we randomly choose a new one until two gametes are retained to form a new zygote. This process continues until N individuals have been produced.

This life cycle is iterated for a large number of generations. We use the first 2000 generations as a burn-in phase, during which we assume that there is no actual effect of PRDM9 and target alleles, that is no gene conversion ($c = 0$) and no fertility selection ($f = 0$). This allows to shuffle the whole population and reach a crossover–mutation–drift equilibrium. After the burn-in phase, conversion c and fertility f are set to the values we are simulating and we let the system evolve.

Because the population is finite, so that allele frequencies and crossover probabilities vary stochastically, we calculated the periods of the oscillations as follows. The algorithm finds the first generation of a shift from a hot (high crossover probability) to a cold state (low crossover probability), and then calculates the number of generations between each of the following shifts. Eventually, mean and standard deviation are computed.

3. Results

3.1. One target: Deterministic two-locus two-allele model

3.1.1. Analytical results

Much of the subsequent analysis will rely on a perturbation approach that assumes that mutation is much weaker than selection and recombination. In this case, we assume that μ_A and μ_B are sufficiently small such that terms of order μ_A^2 , μ_B^2 , and $\mu_A \mu_B$ can be ignored.

In the absence of mutation there are four corner equilibria $\mathbf{x}^{*1} = (1, 0, 0, 0)$, $\mathbf{x}^{*2} = (0, 1, 0, 0)$, $\mathbf{x}^{*3} = (0, 0, 1, 0)$, and $\mathbf{x}^{*4} = (0, 0, 0, 1)$ which correspond to fixation of the gametic types $A_1 B_1$, $A_1 B_2$, $A_2 B_1$, and $A_2 B_2$ respectively. It was shown in Úbeda

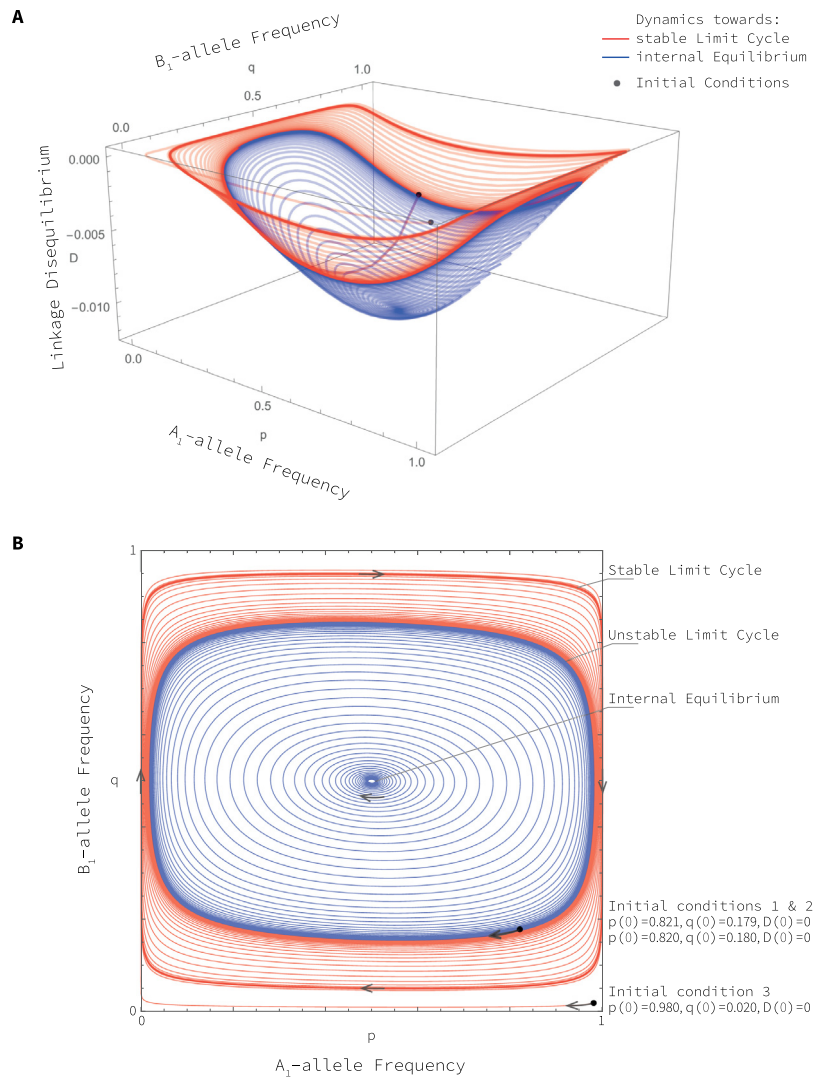


Fig. 2. Examples of dynamics of the allelic frequencies in the parameter region $f < \frac{1}{2}c$. Red and blue are used to represent the trajectories leading to the stable limit cycle and the internal equilibrium respectively. **Panel A** displays the dynamics in the (p, q, D) -space. It shows that the linkage disequilibrium is always negative. If the allele frequencies converge to a cycle, the linkage disequilibrium also converges, but remains weak. If the allele frequencies converge to the symmetric equilibrium, the linkage disequilibrium converges to a fixed negative value that is considerably smaller than the linkage disequilibrium along the stable cycle. **Panel B** displays the dynamics in (p, q) -space. It shows the existence of an unstable limit cycle between the stable limit cycle and the internal equilibrium. The initial conditions are the same in both panels. Note that two trajectories start near the unstable limit cycle (the starting points are indistinguishable). Parameter values are $f = 0.4, b = c = 1, r_m = \frac{1}{2}, \mu = 10^{-5}$.

et al. (2019) that when $\beta > \gamma$, \mathbf{x}^{*1} and \mathbf{x}^{*4} are linearly stable and \mathbf{x}^{*2} and \mathbf{x}^{*3} are unstable, however when $\beta < \gamma$, all four corner equilibria are saddles.

General theory (e.g., Karlin and McGregor, 1972, Theorem 4.4) shows that: corner equilibria that are unstable in the absence of mutation leave the state space, the simplex, when weak mutation is added; and corner equilibria that are stable in the absence of mutation move into the simplex when mutation is added. Therefore, when $\beta > \gamma$, \mathbf{x}^{*2} and \mathbf{x}^{*3} leave the simplex when mutation is introduced, and \mathbf{x}^{*1} and \mathbf{x}^{*4} move into the interior of the simplex. For weak mutation, the coordinates of \mathbf{x}^{*1} and \mathbf{x}^{*4} are given in Appendix A.3. Furthermore, these two equilibria are linearly stable if mutation is sufficiently weak (Karlin and McGregor, 1972, Theorem 4.4). When $\beta < \gamma$, all corner equilibria leave the simplex when mutation is introduced and there are no other equilibria in close vicinity to the corners (again by Theorem 4.4 in Karlin and McGregor, 1972).

In the absence of mutation, a symmetric internal equilibrium, \mathbf{x}^{*5} , exists always. It has all alleles present at frequency $\frac{1}{2}$ and exhibits linkage disequilibrium given by Eq. (A.3) in Appendix. It was shown in Úbeda et al. (2019) that when $\beta > \gamma$ then \mathbf{x}^{*5} is a saddle point, and when $\beta < \gamma$ then \mathbf{x}^{*5} is linearly stable. Except for the degenerate case $\beta = \gamma$, these five equilibria are the only equilibria (Appendix A.2).

General theory (e.g., Karlin and McGregor, 1972, Theorem 4.4) shows that under weak perturbation, an internal equilibrium maintains its stability properties if the unperturbed equilibrium is hyperbolic. Hyperbolic means that there is no eigenvalue of modulus 1. Therefore, upon introducing mutation, when $\beta > \gamma$ the perturbation of \mathbf{x}^{*5} remains a saddle point under weak mutation (and all eigenvalues are real), and when $\beta < \gamma$ the perturbation of \mathbf{x}^{*5} remains linearly stable. At this symmetric internal equilibrium, also denoted \mathbf{x}^{*5} , the allele frequencies of A_1 (p^*) and B_1 (q^*) remain $\frac{1}{2}$ each, independently of mutation. It

exhibits substantial linkage disequilibrium (D^*), which is positive if $\beta > \gamma$ and negative if $\beta < \gamma$. Explicit expressions are given in Eqs. (A.7) and (A.8). The stronger mutation, the closer is D^* to 0, i.e., mutation weakens the linkage disequilibrium—notice however that mutation neither eliminates the linkage disequilibrium nor changes its sign (see Appendix A.3).

If $\beta < \gamma$, then the condition

$$\delta > \frac{(\gamma - 2\beta)^2}{8\gamma} \sqrt{\frac{\gamma - \beta}{\beta}} \tag{28}$$

implies that, in the absence of mutation, the symmetric equilibrium has a pair of conjugate complex (i.e., non-real) eigenvalues (Supplementary *Mathematica* notebook, Section 2.4). Hence, the symmetric equilibrium is a spiral sink in this case. This property is maintained under weak mutation because eigenvalues change continuously. Notice that condition (28) is missing in Úbeda et al. (2019).

In summary, this analysis shows that existence and stability—hence dynamical properties of the system—depend decisively on the relative strength of fertility selection and conversion. If $\beta > \gamma$, which is equivalent to $f > \frac{c}{c+r_t(2-c)}$, the symmetric internal equilibrium is unstable, in fact a saddle point, and the two equilibria \mathbf{x}^1 (close to fixation of A_1B_1) and \mathbf{x}^{*4} (close to fixation of A_2B_2) exist and are linearly stable. If $\beta < \gamma$, i.e., if $f < \frac{c}{c+r_t(2-c)}$, then there is no equilibrium close to the boundary and the symmetric internal equilibrium \mathbf{x}^{*5} is linearly stable and has negative linkage disequilibrium. Thus, crossover enabling gametic types are under-represented at the internal equilibrium.

In addition, solutions starting close enough to the internal equilibrium spiral towards it if condition (28) is satisfied, which is the case if the *effective* recombination probability δ is not too small. Because in the absence of mutation, a stable heteroclinic orbit can exist if $\beta < \gamma$ (Úbeda et al., 2019), with mutation it may be replaced by a stable limit cycle close to the boundary. We have not proved this, but numerical work suggests existence and (local) stability of a limit cycle (see below). In Appendix A.4, we prove that if $\beta < \gamma$, then any attractor (hence the limit cycle when it exists) exhibits negative linkage disequilibrium.

3.1.2. Numerical results

In the previous section we found that $\beta = \gamma$, equivalent to $f = \frac{c}{c+r_t(2-c)}$ determines the change in qualitative behaviour of the dynamic system. Notice that $\frac{c}{c+r_t(2-c)}$ increases from $\frac{1}{2}c$ to 1 when r_t decreases from 1 to 0. Henceforth, for simplicity, we assume that $r_t = 1$ (unless stated otherwise) which implies that the condition for a change in behaviour reduces to $f = \frac{1}{2}c$.

We start by assuming that fertility selection is stronger than gene conversion, that is $f > \frac{1}{2}c$. Recall from above that, in this case, there are three equilibria: two of them, \mathbf{x}^1 and \mathbf{x}^{*4} , near the corners and stable, and one of them, \mathbf{x}^{*5} , internal and unstable. Numerical iteration of the dynamical Eqs. (22) suggests that the trajectories always converge to one of the equilibria where one of the crossover-enabling gametes is close to fixation, \mathbf{x}^1 or \mathbf{x}^{*4} . Our finding recovers previous results indicating that strong fertility selection prevents the death of individual recombination hotspots (Boulton et al., 1997; Pineda-Krch and Redfield, 2005).

In the rest of this section we will focus on the more realistic case when fertility selection is weaker than gene conversion, that is $0 < f < \frac{1}{2}c$. Recall from above that there is only one equilibrium, \mathbf{x}^{*5} , which is internal and stable if the mutation probability is small. Numerical iteration of the dynamical Eqs. (22) suggests that, depending on the initial conditions, the trajectories either approach the internal equilibrium or a limit cycle (Fig. 2).

Using numerical methods, we mapped the set of initial conditions leading to one dynamic behaviour or the other (Fig. 3).

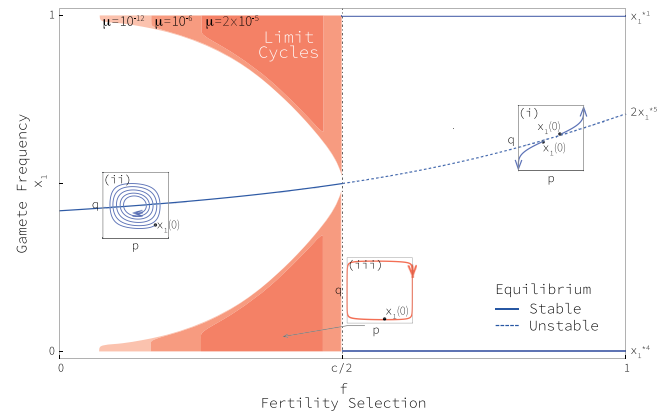


Fig. 3. Basin of attraction of limit cycles and each equilibrium. This figure depicts the frequency of gametic type A_1B_1 (x_1) as a function of the fitness cost f for given values of r_m, b, c, μ . When $f > \frac{1}{2}c$ and $\mu > 0$ equilibria $\mathbf{x}^1, \mathbf{x}^{*4}, \mathbf{x}^{*5}$ are biologically meaningful. Equilibria $\mathbf{x}^1, \mathbf{x}^{*4}$ are stable whereas equilibrium \mathbf{x}^{*5} is unstable. The figure is constructed by starting trajectories at different initial conditions, sampled at equally spaced intervals along the line connecting corners $(1, 0, 0, 0)$ and $(0, 0, 0, 1)$ for which $(x_1(0), x_2(0), x_3(0), x_4(0)) = (x_1(0), 0, 0, 1 - x_1(0))$. Initial conditions above the dotted curve (which represents equilibrium \mathbf{x}^{*5}) lead to equilibrium \mathbf{x}^1 while those below \mathbf{x}^{*5} lead to \mathbf{x}^{*4} . This behaviour is summarised using inset (i), which represents the dynamics in the allelic space (p, q) . When $f < \frac{1}{2}c$ and $\mu > 0$, only equilibrium \mathbf{x}^{*5} is inside the region where x_1 has biological meaning. Equilibrium \mathbf{x}^{*5} is stable. Trajectories starting at initial condition $x_1(0)$ within the white region spiral towards the symmetric internal equilibrium (see inset ii). When initial conditions are within the red region, the solutions oscillate towards a limit cycle (see inset iii). As the mutation probability increases, the set of initial conditions converging to the stable limit cycle shrinks.

We observe that when the initial standing variation at the target site is low, the trajectories converge to a limit cycle where the four gametic types oscillate regularly and permanently over time (Fig. 2). Interestingly, we also found that the higher the mutation probability the lower the standing variation at the target site that can lead to a limit cycle (Fig. 3). Compared to the absence of mutation (Úbeda et al., 2019), we find that mutation qualitatively alters the dynamic behaviour of the system from a heteroclinic cycle (lacking biological meaning (Haig and Grafen, 1991)) to a stable limit cycle. We also find that, mutation reduces the parameter space where oscillations are observed, the greater the mutation rate the smaller the basin of attraction of the limit cycle (Fig. 3). As already noted, the limit cycle emerges from the heteroclinic orbit at the boundary when mutation is introduced. When the mutation rate is increased, it ceases to exist at a critical value (not much higher than $\mu = 2 \times 10^{-5}$ in Fig. 3). Apparently, this occurs when it merges with an unstable cycle, which for smaller μ lies in the manifold separating the basins of attraction of the stable symmetric equilibrium \mathbf{x}^{*5} and the stable limit cycle. Apparently, no (generalised) Hopf bifurcation occurs at \mathbf{x}^{*5} .

To gain intuition on the oscillatory behaviour of our system, we rewrite Eqs. (22a) and (22b) as follows:

$$\Delta p = \underbrace{(1 - 2\mu_A)}_{(+)} \underbrace{\left[\frac{b}{w} fp(1-p)(q - \frac{1}{2}) \right]}_{(+)} - 2\mu_A(p - \frac{1}{2}), \tag{29a}$$

$$\Delta q = \underbrace{(1 - 2\mu_B)}_{(+)} \left[\underbrace{\frac{b}{w} (f - \frac{1}{2}c) q(1-q)(p - \frac{1}{2})}_{(-)} + \underbrace{\frac{b}{w} \frac{1}{2} c D (q - \frac{1}{2})}_{(-)} \right] - \underbrace{2\mu_B (q - \frac{1}{2})}_{(-)}. \tag{29b}$$

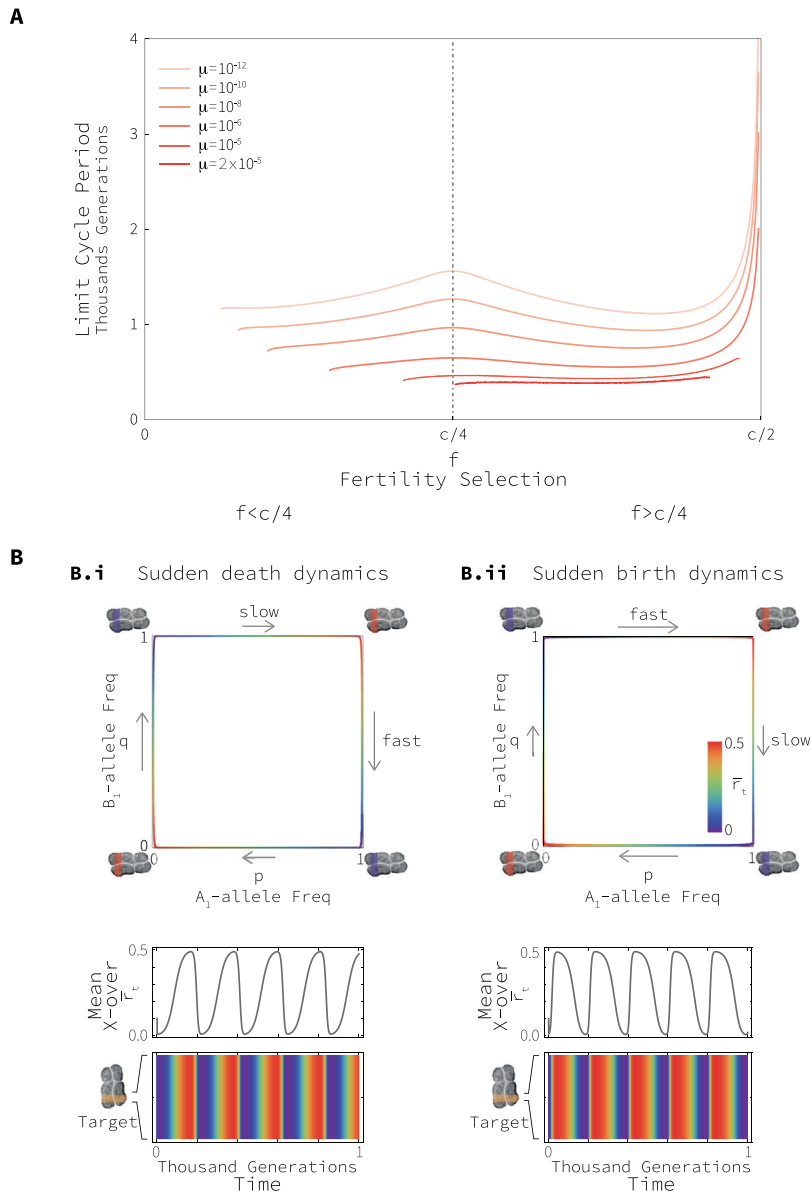


Fig. 4. Periodicity of limit cycles. Panel A displays the period of limit cycles as a function of the fitness cost f within the parameter region $(0, \frac{1}{2}c)$. Results are illustrated for different mutation rates $\mu = \mu_A = \mu_B$. The parameter values used to create this figure are $b = c = 1$, $r_m = \frac{1}{2}$. Initial conditions are $(x_1(0), x_2(0), x_3(0), x_4(0)) = (a, 0, 0, 1 - a)$ where $a = 0.005$. Panel B displays the genetic and phenotypic oscillations over time. Sub-panel B.i displays the case when $0 < f < \frac{1}{4}c$. Here the transition from hot to cold is faster than the transition from cold to hot. We refer to this type of dynamics as “sudden death dynamics”. To create this panel we assumed $f = 0.1$. Sub-panel B.ii displays the case when $\frac{1}{4}c < f < \frac{1}{2}c$. Here the transition from cold to hot is faster than the transition from hot to cold. We refer to this type of dynamics as “sudden birth dynamics”. To create this panel we assumed $f = 0.4$. Within each sub-panel there are three figures: The first one depicts the change in allele frequencies in the limit cycle. The second and third figures depict the change in population mean crossover probability (\bar{r}_t) at the target locus over time. Red colours correspond to high crossover probabilities (hotspot), and blue colours correspond to low crossover probabilities (coldspot). In Panel B we assumed that $b = c = 1$, $r_m = \frac{1}{2}$, $\mu = 10^{-9}$. In both panels $r_t = 1$ and initial conditions are $p(0) = 0.99$, $q(0) = 0.01$, $D(0) = 0$.

From inspecting these equations we conclude that the contribution of fertility, f , to changes in the frequency of A_1 (p) depends on the frequency of B_1 (q). It is positive when B_1 is more abundant than B_2 ($q > \frac{1}{2}$) and negative otherwise; notice that the sign of $(1 - 2\mu_A)\frac{b}{w}fp(1-p)(q - \frac{1}{2})$ is determined by the sign of $q - \frac{1}{2}$. Thus, fertility selection acts in a frequency dependent manner incrementing the frequency of the allele at PRDM9 that matches the most abundant allele at its target locus.

In the parameter region where oscillations are observed ($f < \frac{1}{2}c$), the contribution of fertility relative to conversion ($f - \frac{1}{2}c$) to changes in the frequency of B_1 (q) depends on the frequency of A_1 (p). It is negative when A_1 is more abundant than A_2 ($p > \frac{1}{2}$) and positive otherwise; notice that the sign of $(1 - 2\mu_B)\frac{b}{w}(f -$

$\frac{1}{2}c)q(1-q)(p - \frac{1}{2})$ is determined by the sign of $p - \frac{1}{2}$. Thus fertility (relative to conversion) acts in a frequency dependent manner incrementing the frequency of the target allele that is matched by the less abundant PRDM9 allele.

Finally, mutation has the effect of reducing the strength of selection very slightly (by the factors $(1 - 2\mu_A)$ and $(1 - 2\mu_B)$) and, much more importantly, to push trajectories towards the centre of the state space. In particular, mutation in A reduces the frequency of A_1 when A_1 is the most abundant allele, that is $p > \frac{1}{2}$, and mutation in B reduces the frequency of B_1 when B_1 is the most abundant allele, that is $q > \frac{1}{2}$. By reducing the frequency of the most abundant allele at each locus, mutation prevents the loss of alleles.

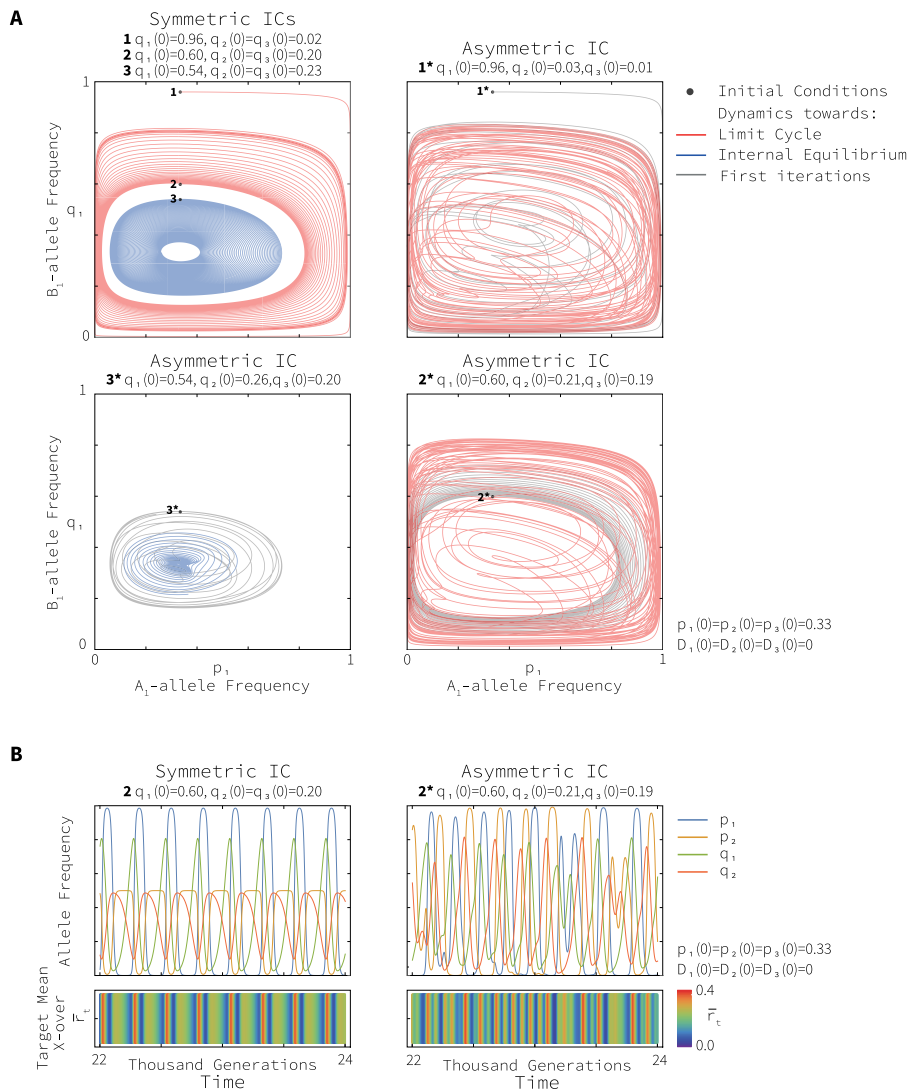


Fig. 5. Genetic and phenotypic dynamics with three alleles in PRDM9 and its target. Panel A. Figures represent the change over time in the frequency of alleles A_1 in PRDM9 and B_1 in its target locus. Alleles A_2 and A_3 and B_2 and B_3 are not represented. The first figure represents the dynamics for three initial conditions (1, 2, 3) where the frequency of alleles B_2 and B_3 is the same, that is $q_2(0) = q_3(0)$. The first and second initial condition, 1 and 2, lead to a limit cycle from outside and inside of the cycle, respectively. The last condition, 3, leads to an internal equilibrium. The other three figures represent the dynamics for the three equivalent initial conditions (1*, 2*, 3*) when the initial frequency of alleles B_2 and B_3 are slightly different. It can be observed that trajectories transition from being regular to being chaotic but the long-term behaviour persists—approaching a limit cycle or interior equilibrium. **Panel B** Figures represent the change over time in allele frequency of alleles A_1 and A_2 in PRDM9 and B_1 and B_2 in its target locus. For clarity alleles A_3 and B_3 have been omitted. In each figure, the change over time in mean crossover rate at the target has been represented. For clarity only the last 2000 generations have been represented. Here we assumed that $f = 0.4$, $b = c = r_t = 1$, $r_m = \frac{1}{2}$. In both panels the initial conditions for the allelic frequency at PRDM9 and the linkage disequilibria remain the same, that is $p_1(0) = p_2(0) = p_3(0) = \frac{1}{3}$ and $D_1(0) = D_2(0) = D_3(0) = 0$ respectively.

The number of generations it takes for a cycle of gametic types to be completed (the *period* of the limit cycle) changes with the strength of fertility selection and the mutation probability. The closer the strength of fertility selection is to $\frac{1}{4}c$, the longer the period of the limit cycle (Fig. 4A). The greater the mutation probability, the shorter the period of the limit cycle (Fig. 4A). For sufficiently strong mutation (slightly in excess of the largest mutation rate shown in the figure), the limit cycle disappears and the symmetric internal equilibrium attracts all solutions. The period of the limit cycle is relevant because it measures, indirectly, the life expectancy of recombination hotspots.

The number of generations required for the replacement of one allele at the target locus (B) relative to such a replacement at the PRDM9 locus (A), depends on the strength of fertility selection and the mutation probability. When $\frac{1}{4}c < f < \frac{1}{2}c$, the replacement of alleles at PRDM9 is faster than that at its

target, which translates into oscillations of the mean crossover probability with sudden birth of recombination hotspots and their gradual death (Fig. 4B.ii). When $0 < f < \frac{1}{4}c$, the replacement of alleles at PRDM9 is slower than that at its target, which translates into sudden death of recombination hotspots and their gradual birth (Fig. 4B.i). The closer f is to $\frac{1}{4}c$, the lower the relative difference between loci in the speed of replacement of alleles. The higher the mutation probability, the lower the relative difference between loci in the speed of replacement of alleles.

The reason why the relation between f and $\frac{1}{4}c$ determines whether replacement at PRDM9 is faster than at its target, or vice versa, can be immediately inferred by comparing the rate-determining factors of selection bf in Eqs. (29a) and $b(\frac{1}{2}c - f)$ in (29b). Obviously, $bf > b(\frac{1}{2}c - f)$ if and only if $f > \frac{1}{4}c$. Therefore, the change in allele frequency at PRDM9 is faster than at its target if $f > \frac{1}{4}c$, and vice versa (Fig. 4B).

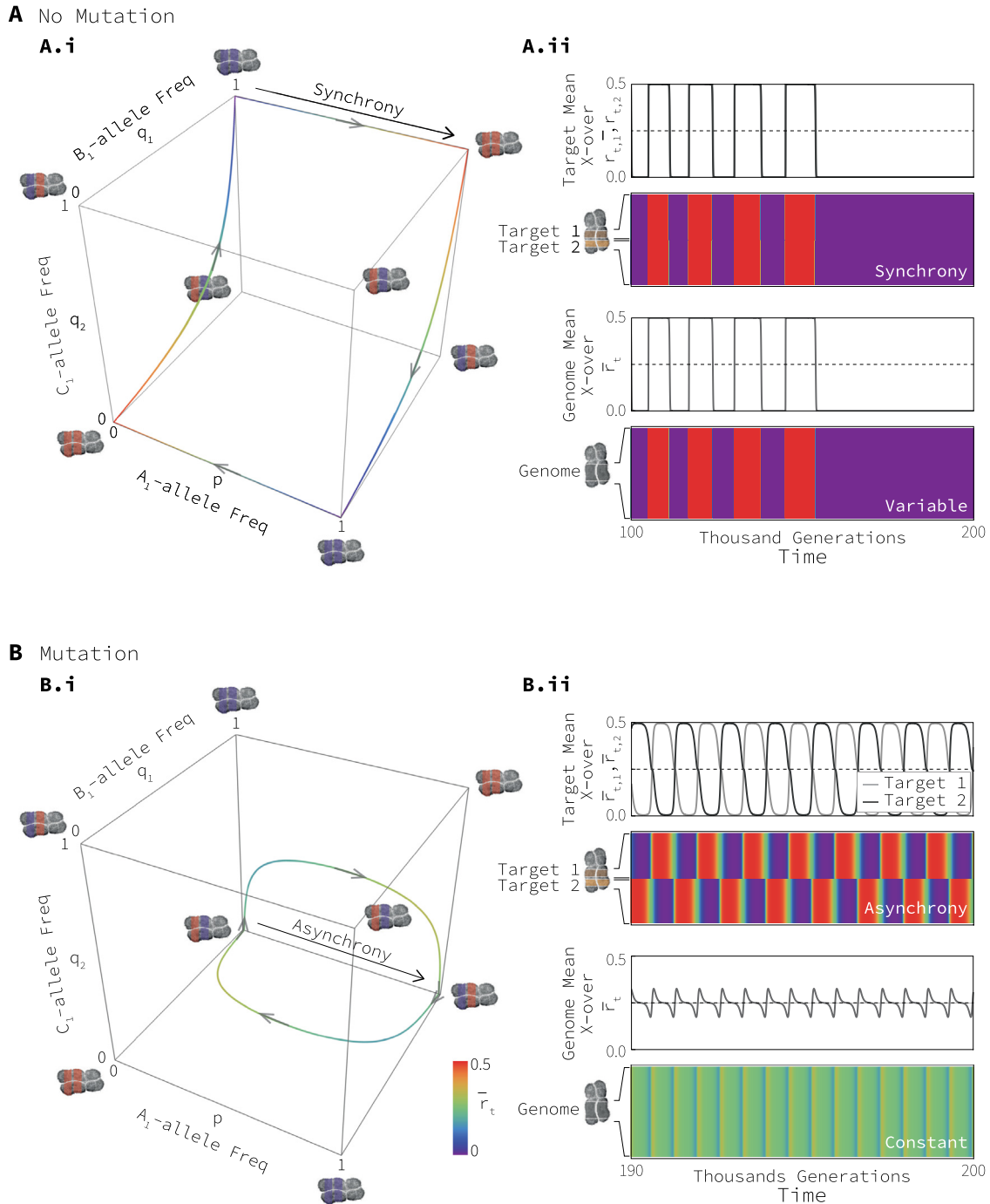


Fig. 6. Genetic and phenotypic dynamics in a system with two targets. Cubes represent the change in allele frequency in PRDM9 and target loci. Figures depict the mean crossover probability in each of the target sites and the genome (meaning the average across all targets). These are represented both as lines and heat maps. **Panel A** corresponds to the case where there are no mutations ($\mu_A = \mu_B = \mu_C = 0$). **Sub-panel A.ii** shows that the dynamics converge to heteroclinic cycles such that both target loci are hot or cold at the same time (synchronous oscillation) and the average genomic crossover rate is variable (oscillating with the average recombination rate in each target). **Panel B** corresponds to the case where there are mutations (in particular $\mu_A = 10^{-6}$, and $\mu_B = \mu_C = 10^{-8}$). **Sub-panel B.ii** shows that the dynamics converge to limit cycles such that one target is hot when the other is cold or vice-versa (asynchronous oscillation) and the average genomic crossover rate is close to being constant (showing minimal oscillation). Here we assumed that $f = 0.18$, $b = r_t = 1$, $c = r_m = \frac{1}{2}$, and $p(0) = q_1(0) = 0.9$, $q_2(0) = 0.8$, $D(0) = 0$.

This difference in the speed of fixation of the alleles in PRDM9 and target loci is relevant because it would allow us to calibrate the model for the strength of fertility selection acting on recombination. In particular, a signature of a stronger selective sweep in the PRDM9 locus (relative to its target loci) would suggest that the fertility loss f is between $\frac{1}{4}c$ and $\frac{1}{2}c$. On the contrary, a signature of a stronger selective sweep at the target loci (relative to the PRDM9 locus) would suggest that the fertility loss f is between 0 and $\frac{1}{4}c$.

3.2. One target: Deterministic two-locus three-allele model

In this section we explore the effect of having more than two alleles—at both PRDM9 and its target—on the observed oscillation of the mean crossover probability in the target site. We investigate this issue by numerically iterating the recursion equations in (3) for the particular case when PRDM9 and its target locus have three alleles. We use these iterations to explore the time

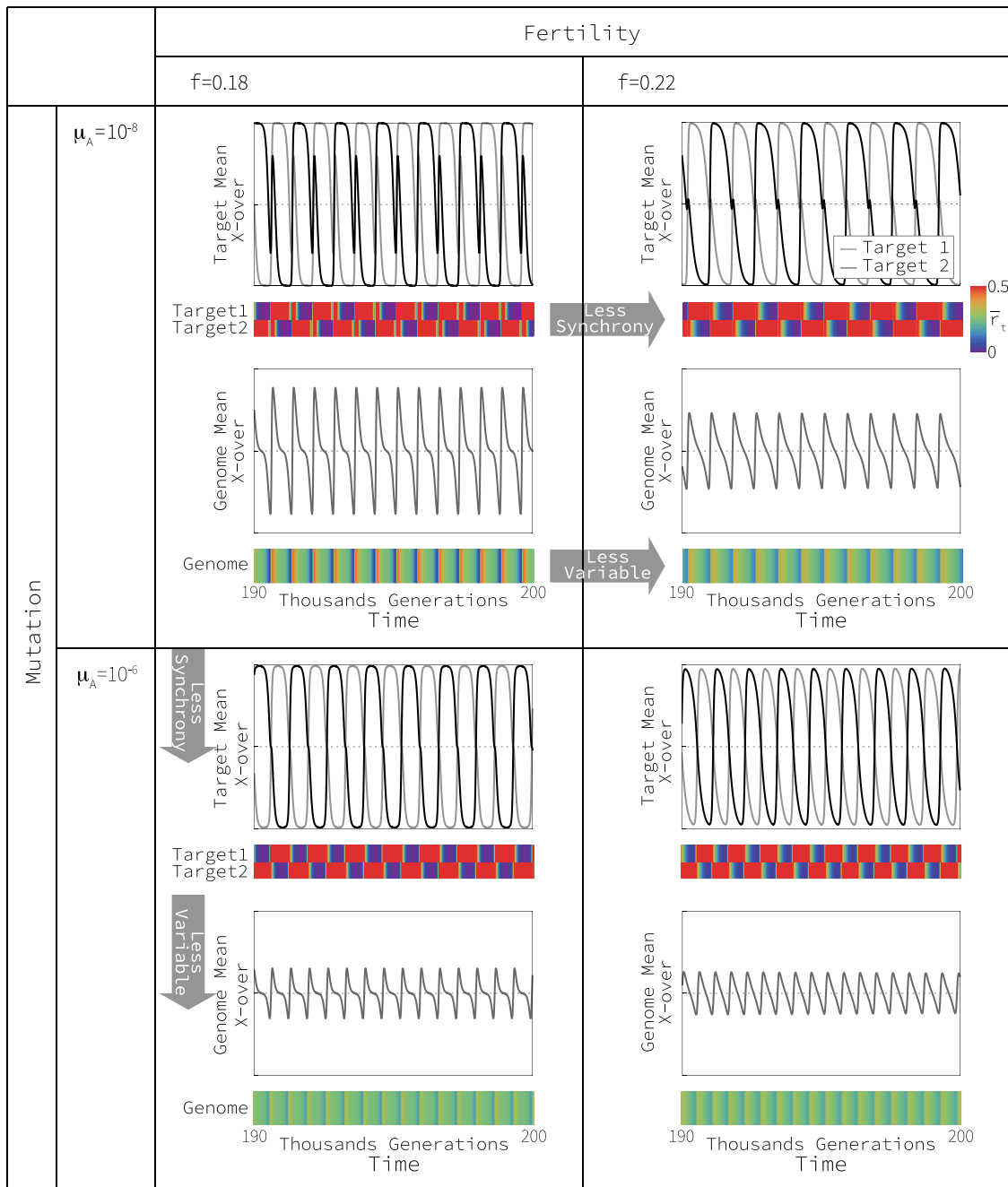


Fig. 7. Effect of mutation probability and selection on the asynchrony of oscillations in each target. This figure depicts the change in crossover probability at each of two target loci for different combinations of the mutation probability and fertility selection, along with the genome-wide mean crossover probability. This figure shows that the greater the mutation probability the lower the overlap of hot and cold phenotypes at different target loci (the greater the asynchrony) and the lower the oscillation of the genomic mean crossover probability. Notice that mutation does not affect the amplitude of the oscillation in each target. Here we assumed $b = r_t = 1$, $c = r_m = \frac{1}{2}$, $\mu_B = \mu_C = 10^{-8}$ and initial frequencies of: $p(0) = q_1(0) = 0.9$, $q_2(0) = 0.8$, $D(0) = 0$.

dependence of the mean crossover probability at the target site given by (4).

As in the previous section we focus on the case when fertility selection is weaker than gene conversion, that is $0 < f < \frac{1}{2}c$. When initially the standing variation at the target site is high enough, the trajectories converge to an internal equilibrium where the six gametic types are present in fixed proportions (Fig. 5). However, when initially the standing variation at the target site is low enough, the trajectories converge to a limit cycle where the six gametic types oscillate permanently over time (Fig. 5). These oscillations can be regular in the very particular case in which the initial standing variation of the less frequent

alleles at the target site is exactly the same. However, even small deviations from equal initial representation of these two alleles results in irregular oscillations with irregular amplitude and period (Fig. 5). This seemingly chaotic dynamics translates into permanent oscillations between hotspots and coldspots that reach variable levels of crossover intensity—maxima and minima—and life-expectancy (Fig. 5).

The intuitive explanation for these chaotic dynamics is that fertility selection favours the PRDM9 allele that matches the most frequent target—as suggested from the analysis of the simpler two-allele model in the previous section. When the PRDM9 allele that matches the most frequent target is close to fixation, the two

mismatching alleles are equivalent from a gene conversion perspective. Small differences in their frequency determines which will be the first to overtake the most frequent target in the population in its decline. However, fertility selection and gene conversion continue to favour oscillations over evolutionary time.

3.3. Two targets: Deterministic three-locus two-allele model

In this section we explore the effect of having more than one target on the observed oscillation of the mean crossover probability in a target site. We investigate this issue by numerically iterating the recursion equations (26). We focus on the time dependence of the mean crossover probabilities at the target sites, given by Eqs. (27a) and (27b), and their genomic average in (27c).

As in previous sections, we focus on the case when fertility selection is weaker than gene conversion, that is $0 < f < \frac{1}{2}c$. When initially the standing variation at the target site is low enough and mutation probabilities are realistic (10^{-6} in PRDM9, $\mu_A = 10^{-6}$, and 10^{-8} in the targets, $\mu_B = \mu_C = 10^{-8}$), we showed numerically that the trajectories converge to a limit cycle (Fig. 6B). In this limit cycle the two gametic types where PRDM9 and only one of its targets match (either $p = q_1 = 1$ and $q_2 = 0$ or $p = q_2 = 0$ and $q_1 = 1$) oscillate in a regular fashion. These dynamics translate into permanent oscillations within each target, with one target being hot when the other is cold and vice-versa (Fig. 6B). That the two targets oscillate asynchronously implies that the mean crossover probability in the genome—across both target sites—remains approximately constant (Fig. 6B). When the mutation probabilities tend to zero, the two targets tend to oscillate synchronously—both targets being hot at the same time—and the mean crossover probability in the genome oscillates with the targets (Fig. 6A).

Interestingly, we find that the mutation probability in PRDM9 and the strength of fertility selection determine the degree of asynchrony between target loci (Fig. 7). In particular, the greater the mutation probability, the greater the asynchrony in mean crossover probability between target sites and the lower the variation in mean crossover probability in the genome (Fig. 7). The intuitive reason why both targets should oscillate asynchronously is that fertility selection drives a selective sweep favouring the PRDM9 allele that matches the most common allele at both target sites. When one target is almost fixed with a cold allele and the other with a hot allele—the target allele that matches the allele almost fixed at PRDM9—and conversion starts turning the hot allele into cold copies, natural selection favours a sweep of a PRDM9 allele that matches the almost fixed cold allele at one target and the incipient number of cold alleles at the second target. The greater the frequency of the rarer allele at PRDM9, the stronger the selective sweep. Because mutation increases the amount of the rarer allele at PRDM9, it is expected that the greater the mutation rate, the earlier fertility selection favours a sweep in PRDM9 that turns cold alleles into hot and vice-versa. In the absence of mutation, fertility selection is not strong enough to result in a sweep until both targets are almost fixed with cold alleles. Then a selective sweep will turn both targets into hotspots resulting in synchronous oscillations.

Similarly, the stronger fertility selection, the greater the asynchrony in mean crossover probability between target sites and the lower the variation in mean crossover probability in the genome (Fig. 7). Our previous intuition extends to the effect of fertility selection, f , on the asynchrony of the oscillations. In particular, stronger fertility selection means that a selective sweep in PRDM9 that switches targets can happen earlier on, before both targets are coldspots.

3.4. Stochastic multi-target multi-allele model

Here we explore the effect of a finite population size on the observed oscillations of the mean crossover probability near a target site. We investigate this issue by exploring simulations of the multi-target multi-allele model described in the models section.

As in previous sections, we focus on the case when fertility selection is weaker than gene conversion, that is $0 < f < \frac{1}{2}c$. We find that the behaviour of the system does not change qualitatively when considering small populations and/or more than two target loci. In particular, we show that the crossover probability in individual target sites oscillates widely while the genome's crossover probability is more stable (Fig. 8B).

For the two-target case, we investigate the dependence on the mutation rate of the period of the oscillations of the mean crossover probabilities at each target. Fig. 9A displays their means and standard deviations. It shows that the period decreases by about two orders of magnitude as the mutation rate increases from 10^{-8} to 10^{-6} . Fig. 9B illustrates that the stochasticity in the model induces ample variation in the number of generations between two successive shifts from hot to cold.

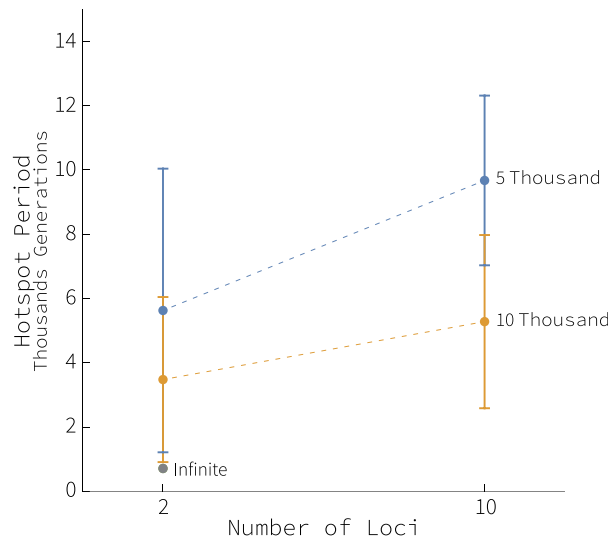
We also explore the dependence on the population size of the period of the oscillations of the mean crossover probabilities at each target. As Fig. 8 shows, the turnover of hotspots accelerates as the population size increases from $N = 5000$ to $N = 10000$ individuals. The reason is that in finite populations genetic drift favours the loss of rare alleles that need to be recreated for a selective sweep to occur. As the recreation of alleles is driven by mutations (which are rare), stronger genetic drift (smaller population size) lengthens hot and cold phases, thus increasing the average period of oscillations. We also explored differences in genomes with two and ten target loci. We find that in both cases the crossover rate in individual target sites oscillates widely, with the turnover of hotspots decelerating as the number of targets increases (Fig. 8).

In our deterministic models, for simplicity, we assumed that there would be one double strand break only. However, it is more realistic to assume that there would be multiple break attempts with the individual experiencing a fertility cost when none of them is successful. In our stochastic model we allow for multiple double strand breaks. We find that compared to the case when there is one break attempt, in the extreme case when there are ten break attempts the targets become cold quicker and selection for the matching PRDM9 allele is not strong enough until almost all targets are cold. While the oscillatory behaviour remains, this oscillation involves less hotspots in a colder genome (Fig. 10). Thus the fraction of targets that are actual recombination hotspots should be affected by the level of expression of PRDM9.

3.5. Choice of parameter values

To illustrate our findings we often use variants of a small set of parameter values. This choice of values does not condition our findings. We have explored a much larger set of parameter values and chosen those parameter values that produce illustrative figures. In particular we choose $r_m = \frac{1}{2}$ because PRDM9 in the majority of cases far from its target or even in another chromosome. Empirical results show that $c \approx \frac{1}{2}$ (Jeffreys and Neumann, 2002, 2005). There are no empirical measurements of b or f . However, what b does primarily is to change the speed of the allele-frequency dynamics by making oscillations faster or slower. Finally we explored a wide range of values of f .

A



B

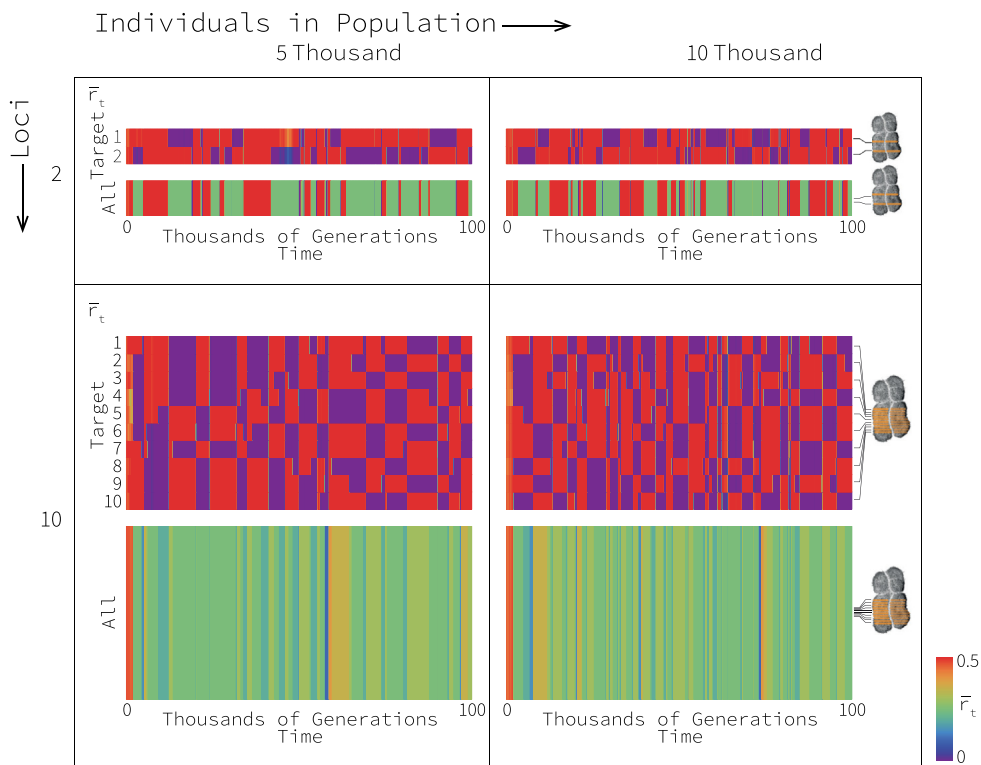


Fig. 8. Effect of population sizes and number of targets on the speed of hotspot oscillations in finite populations. Panel A shows the average period of an oscillation between hot and cold states. Blue and orange dots represent the average value of the period (for two targets in 100 thousand generations) when the population size is five and ten thousand individuals respectively. The grey dot represents the equivalent value for an infinite population. Vertical bars correspond to the standard deviation. We make the conservative assumption that there is only one attempt to produce one double strand break per genome per generation. Panel B shows examples of the dynamics. Examples are arranged in a table with five and ten thousand individuals in the columns and two and ten loci in the rows. This figure shows that the smaller the population size or the greater the number of target loci, the greater is the period of hotspot oscillations. Here we assumed that $f = 0.4$, $b = c = r_t = 1$, $r_m = \frac{1}{2}$ (the recombination rate between any pair of loci is also $\frac{1}{2}$), $\mu_A = 10^{-6}$, $\mu_B = 10^{-7}$, and $p(0) = 0.95$, $q(0) = 0.95$, $D(0) = 0$.

4. Discussion

In this research we focus on a model of the co-evolution between PRDM9, segregating two alleles, and one target site, segregating two motifs. We find that introducing recurrent mutations in PRDM9 and its target site can result in a balance between the resuscitation and death of hotspots that prevents their extinction over evolutionary time. Therefore, the long-term co-evolutionary oscillation between PRDM9 alleles and its target motifs can solve the recombination hotspots paradox by driving the resuscitation

of hotspots over and over again. In the more realistic case when fertility selection is weaker than conversion, we find two possible dynamics. When the standing genetic variation at the target site is high, co-evolutionary oscillations between PRDM9 alleles and its target motifs settle into a polymorphic equilibrium corresponding to a fixed intermediate crossover probability (Fig. 2). This is consistent with previous results in the absence of mutation (Úbeda et al., 2019). However, when the standing variation at the target site is low, co-evolutionary oscillations between PRDM9 alleles and their target motifs are regular and permanent.

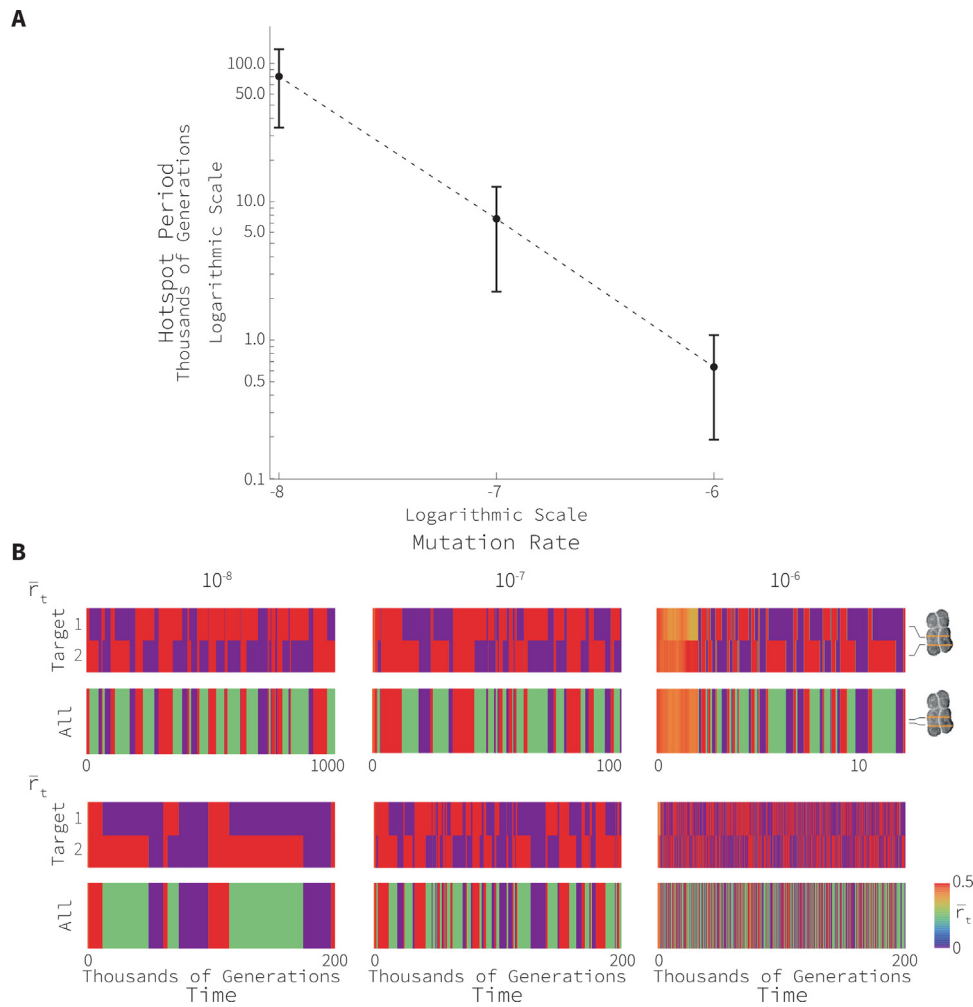


Fig. 9. Effect of the mutation probability on the rate of hotspot oscillations in a finite population. Panel A shows the average period of an oscillation between hot and cold states for different mutation probabilities. Dots represent the average value of the period for a population size of 5000. Vertical bars show the standard deviation. To ensure that the variances are calculated over a similar number of cycles, we performed these computations for 10 000 generations if $\mu_A = \mu_B = 10^{-6}$, for 100 000 generations if $\mu_A = \mu_B = 10^{-7}$, and for 1 000 000 generations if $\mu_A = \mu_B = 10^{-8}$. Panel B shows the change in recombination over time in each target and the average for the genome. We explored three mutation rates that differ in one order of magnitude, namely 10^{-8} , 10^{-7} , and 10^{-6} . The first row shows the hotspots dynamics for the periods of time 1 000 000, 100 000, and 10 000 generations; time changes one order of magnitude with their corresponding mutation probability. This row illustrates how increasing the sampling time one order of magnitude produces similar number of oscillation in the three mutation probabilities considered. In the second row we show the dynamics when the sampling time remains constant. This row illustrates how mutation probabilities increase the rate of oscillations. The parameter values used to create this figure are $f = 0.4$, $b = c = r_t = 1$, $r_m = \frac{1}{2}$, and $p(0) = 0.95$, $q(0) = 0.99$, $D(0) = 0$ and $N = 5000$.

These oscillations result in a limit cycle where hotspots and coldspots alternate, reach the same levels of crossover maxima and minima, and exhibit constant life-expectancy (Fig. 2). This result diverges significantly from previous ones in the absence of mutation where, contrary to empirical evidence, the target site becomes cold or hot permanently (Úbeda et al., 2019). Because this novel dynamics do not settle over evolutionary time they provide a solution to the paradox.

In general, we find low levels of standing genetic variation in the target site can lead to permanent oscillations in crossover rates. However, the amount of standing variation leading to permanent oscillations depends on the strength of fertility selection and the mutation probability. Given the mutation probability, the greater the strength of fertility selection, the greater the range of standing variation leading to oscillation (Fig. 3). Given the strength of fertility selection, the greater the mutation rate, the smaller the range of standing variation leading to oscillation (Fig. 3). The relevance of standing variation for the formation of recombination hotspots had not been appreciated by previous research (Baudat et al., 2010; Úbeda and Wilkins, 2011;

Latrille et al., 2017). Interestingly, we find that whenever fertility selection is weaker than conversion there is negative linkage disequilibrium between alleles in PRDM9 and its target site. This means that there is an excess of combinations of mismatching alleles at PRDM9 and its target sites. This results in a lower recombination rate than would be expected if proteins and targets where to bind randomly (see equation (A.9)). Furthermore, we find that the negative linkage disequilibrium is stronger in the polymorphic equilibrium than in any part of the orbit of the limit cycle (Fig. 2A).

We find that it is Red Queen dynamics between matching PRDM9 alleles and matched target alleles that can solve the paradox. Our results thus substantiate verbal arguments in Úbeda and Wilkins (2011). The term Red Queen dynamics refers specifically to the long-term co-evolutionary dynamics where oscillations in genotype abundance are driven by fluctuating selection (Van Valen, 2019; Schenk et al., 2017, 2020). In our research we did not model selection to be explicitly frequency dependent. However, it can be appreciated from our Eqs. (29) that strength and direction of selection acting on PRDM9 alleles change with the frequency

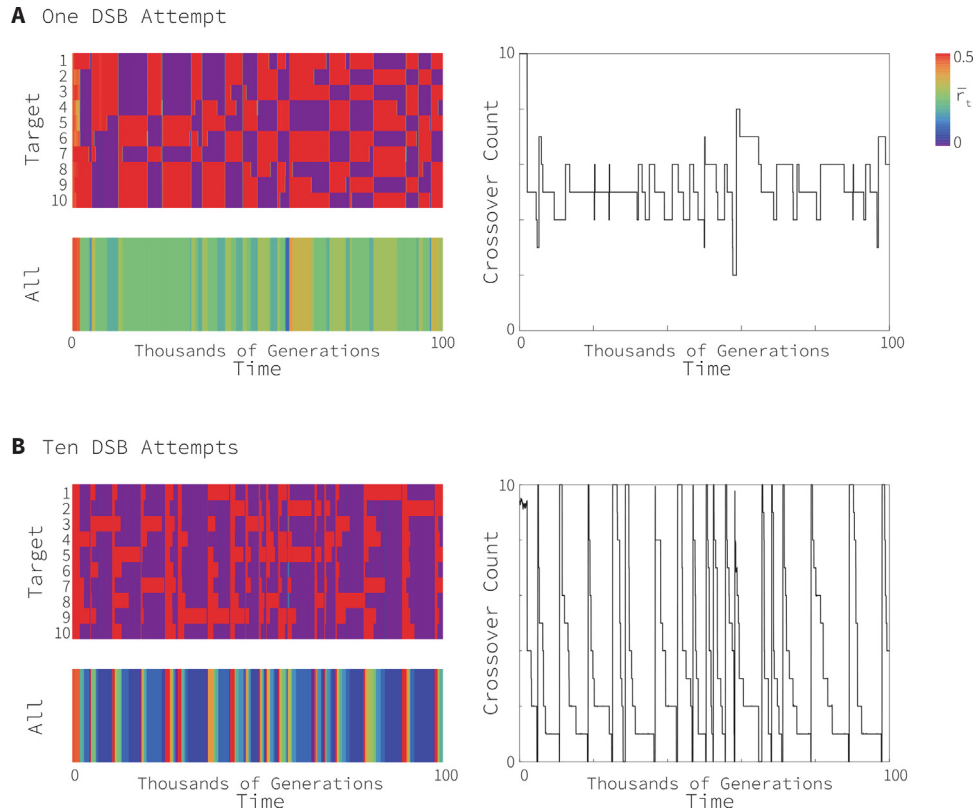


Fig. 10. Comparison of the dynamics when there is one DSB attempt as opposed to multiple DSBs. Panel A Shows the dynamics in a model when there is only one DSB attempt. Panel B Shows the dynamics in a model when there are ten DSB attempts (as many as target loci). In both cases gametes are fully viable when there is at least one crossover event. This figure shows that the greater the number of DSB attempts the colder the genome before a matching PRDM9 allele is driven to fixation. Here we assumed that $f = 0.4$, $b = c = r_t = 1$, $r_m = \frac{1}{2}$ (the recombination rate between any pair of loci is also $\frac{1}{2}$), $\mu_A = 10^{-6}$, $\mu_B = 10^{-7}$, and $p(0) = 0.95$, $q(0) = 0.95$, $D(0) = 0$.

of alleles at the target site, and strength and direction of selection acting on target alleles change with the frequency of alleles at the PRDM9 locus.

We emphasise that, in models of recombination driven by PRDM9, alleles at target sites are not hot or cold per se, instead whether they are hot or cold depends on the genetic background provided by the PRDM9 allele. In our model thus, both alleles at the target site can be hot when matched by the genetic background in PRDM9, and both can be cold when mismatched by PRDM9. When perfectly matched by the genetic background, gene conversion favours the mismatched (cold) allele at the target site—thus initiating the death of the hotspot. Fertility selection favours the matching genetic background while the mismatched (cold) allele is less frequent than the matched (hot) one—thus allowing the death of the hotspot. However, when the mismatched allele becomes more frequent than the matched one, fertility selection favours the alternative genetic background matching the mismatched allele—thus initiating the resuscitation of the hotspot. Our model shows that PRDM9 is selected to matching the most abundant target at any given time. As a result, each of the four combinations of PRDM9 backgrounds and targets can be temporarily best adapted, which explains the oscillations. In our model, fertility selection fluctuates, thus driving (together with conversion) oscillations between hot and cold alleles that do not settle in the long-term thanks to mutation. Therefore, the dynamics that can solve the paradox meets all the requirements of Red Queen dynamics (Van Valen, 2019; Schenk et al., 2017, 2020).

Our model of the co-evolution between PRDM9 segregating two alleles and one target site segregating two motifs (a two-locus two-allele model) does not correspond to the reality of PRDM9 segregating multiple alleles and having many target sites, each segregating multiple motifs. An exhaustive study of the dynamics of all possible multi-allele multi-target extensions of our model is beyond the scope of this research. However, we gain insight into the effect of having multiple alleles by adding to our two-locus two-allele (one-target) model, one allele at PRDM9 and one motif at its target to formulate a two-locus three-allele (one-target) model. We further gain insight into the effect of having multiple targets by adding to our two-locus two-allele (one-target) model, one target segregating the same motifs as the existing target to formulate a three-locus (two-target) model.

We find that adding a third allele at both PRDM9 and its target site still can result in a balance between resuscitation and death of hotspots that prevents their extinction and solves the paradox. When fertility selection is weaker than conversion, co-evolutionary oscillations between alleles at PRDM9 and its target can be permanent, but they often become highly irregular and result in cycles where hotspots and coldspots alternate, with hotspots reaching different levels of crossover maxima in each oscillation and showing variable life-expectancy (Fig. 5). Because these chaotic dynamics do not settle in the long-term, they are chaotic Red Queen dynamics (Schenk et al., 2017, 2020) that can solve the paradox as well. The association between low standing variation with ample irregular oscillations approaching a limit cycle and high standing variation with narrow irregular

oscillation approaching a polymorphic equilibrium is still present in the three-allele model (Fig. 5). An exhaustive exploration of this model would be of interest for future research.

We find that adding a second target can also result in a balance between resuscitation and death of hotspots that prevents their extinction and solves the paradox. However, resuscitation and death alternate between targets (with one target resuscitating when the other is dying and vice versa), thus maintaining the genome's mean crossover probability close to constant. When fertility selection is weaker than conversion and there are low levels of standing variation in target sites, co-evolutionary oscillations between PRDM9 alleles and its two targets can be regular and permanent in each target. These oscillations result in cycles where hotspots and coldspots alternate, with hotspots reaching the same levels of crossover maxima, and having constant life-expectancy (Fig. 6). In principle, the two target sites can be hot and cold at the same time—they oscillate synchronously—or one target site can be hot while the other is cold—they oscillate asynchronously. When targets oscillate synchronously, the genome's crossover probability oscillates with them. When targets oscillate asynchronously, the genome's crossover probability remains nearly constant. An exhaustive exploration of this model would be of interest for future research.

Interestingly, we find that mutations are key in maintaining a constant crossover probability across the genome while permitting the recombination rate in each site to oscillate widely. We find that the greater the mutation rate in PRDM9 (relative to its target sites) and the stronger the fertility selection, the smaller are the fluctuations in the genome's crossover probability (Fig. 7). The intuitive reason for this is that the recombination rate in one target conditions the strength of fertility selection in the other targets. The more coldspots there are, the greater is the strength of fertility selection acting to resuscitate hotspots. The greater the availability of matching PRDM9 mutants, the weaker is the strength of fertility selection necessary to resuscitate hotspots. Therefore, with lower mutation probability and/or fertility cost, the two targets need to be cold before selection favours a switch resulting in all targets becoming hot (synchronous oscillation). With higher mutation probability and/or fertility cost, only one target needs to be cold before selection favours a switch resulting in some of the targets becoming hot while the others become cold (asynchronous oscillation) (Fig. 7).

Because populations are finite, a solution to the Recombination Hotspots Paradox needs to hold in finite populations. Effectively, the dynamics in finite populations result from adding noise (genetic drift) to the underlying dynamics in infinite populations (cf. Schenk et al., 2017, 2020). Therefore having characterised the deterministic co-evolutionary dynamics between PRDM9 and its target sites we are in a good position to better understand its equivalent stochastic dynamics. We find that many of the qualitative results observed in infinite populations extend to finite populations. In particular, similarly to deterministic findings, when fertility selection is weaker than conversion and there are low levels of standing variation in target sites, co-evolutionary oscillations between PRDM9 alleles and its targets can be permanent in each target. As a result there are cycles where hotspots and coldspots alternate, and hotspots reach high levels of crossover probabilities at each target but the crossover probability across the genome remains nearly constant (Fig. 8).

We find that the number of break attempts affects the temperature of the genome without affecting the ability of fertility selection to drive the resuscitation of hotspots. We observe that multiple break attempts reduce the probability that there will be no crossover events at the target loci. Therefore they reduce the strength of fertility selection for hot alleles. However, multiple break attempts accelerate the strength of conversion turning

hotspots into coldspots. As hotspots are turned into coldspots the strength of fertility selections for hot alleles mounts. When the genome is cold enough fertility selection drives the fixation of a hot allele in PRDM9 leading to the resuscitation of hotspots. What is significant is that the oscillatory behaviour is still observed. However, by increasing the number of break attempts reduces the fraction of targets that are hotspots and the temperature of the genome (Fig. 10). This suggests that the level of expression of PRDM9 can control the fraction of targets that are hotspots at any given time.

Oscillations of the same set of alleles correspond to the resuscitation of hotspots whereas the introduction of new alleles can correspond to the birth of new hotspots. We notice that even in the absence of births, resuscitation can prevent the extinction of hotspots thus solving the paradox. We expect that to avoid the potential cost of disrupting a functional gene, eventually, oscillations will be limited to a set of tried and tested PRDM9 alleles and its target motifs, that is mostly resuscitation of hotspots. It is important to notice that even though our model shows that it is the same set of target sites that oscillate between hot and cold states, this does not mean that a two species (or subpopulations) will share the same hotspots. Much on the contrary our model shows that only a fraction of the target sites in the entire genome will be recombination hotspots. It also shows that the set of target sites that are recombination hotspots and their distribution changes rapidly. Therefore, in our model, it is highly unlikely that two snapshots of the distribution of hotspots in two species will show many shared hotspots. In this sense our model not only explains the rapid evolution of hotspots but also the observation that humans and chimpanzees do not share many recombination hotspots and even human subpopulations exhibit some level of recombination hotspot variation (Ptak et al., 2004, 2005; Winckler et al., 2005; Coop et al., 2008; Stevison et al., 2015). The resuscitation of hotspots thus is sufficient to explain the lack of shared recombination hotspots without the need to invoke the birth of hotspots.

In general we find that the long-term co-evolutionary oscillation between PRDM9 alleles and its target motifs can drive the resuscitation of hotspots over and over again. This result holds with multiple alleles at each locus, multiple targets and multiple targets. While each hotspot can reach high and low levels of crossover, the average across the genome remains almost constant due to the asynchrony of oscillations. Furthermore we find that mutation is key in the Red Queen dynamics underlying the permanent and asynchronous oscillation of individual recombination hotspots. Therefore, our findings contribute to solving the Recombination Hotspot Paradox by reconciling the self-destructive nature of PRDM9-directed recombination hotspots with the resuscitation of hotspots to maintain the elevated number of such hotspots observed in various species (Boulton et al., 1997). The findings of our model are consistent with all salient features of recombination hotspots, namely the death of individual hotspots, the preservation of hotspots at the genome level, and the rapid change of the recombinational landscape (with hotspots rarely being shared between closely related species).

Besides offering a plausible solution to the paradox, our research offers the possibility of calibrating our model against genomic data. Such calibration would allow quantitative predictions for the life expectancy of hotspots given different effects of recombination on fertility. This, in turn, can be used to study the signature of oscillatory recombination on linkage disequilibrium in the mammalian genome, a pattern that is relevant to a better understanding of links between alleles and diseases. In addition, our work suggests that signatures of selective sweeps in PRDM9 relative to those in its targets should provide a method for calibrating the strength of fertility selection on recombination.

Data availability

Both Mathematica notebooks used to implement the different deterministic models and the C++ code used to implement the different stochastic models can be found in [10.5281/zenodo.8098998](https://doi.org/10.5281/zenodo.8098998).

Acknowledgements

[FU] and [FF] are supported by a NSFDEB-NERC Research Grant NE/T009322/1. [FU] thanks Carlos Ortega Fernández for lifetime insight and support. We thank the editor, Jeremy Van Cleve, Deborah and Bryan Charlesworth and two anonymous reviewers for helpful comments.

Appendix

A.1. Modelling the interaction of recombination and gene conversion

The precise form of the recursion equation (1) depends on assumptions on interactions between recombination—between PRDM9 and its target—with gene conversion—at the target. In the main text, we assumed that PRDM9 and its target sit on the same chromosome and thus it is most appropriate to consider recombination between the pair of homologous chromatids experiencing a double strand break but not the other pair. If we assume that PRDM9 and its target sit on different chromosomes, then recombination between all four chromatids has to be taken into account. In this case, the term $r_m \left((1 - c) F \bar{b}_{ij,kl} + (1 - \bar{b}_{ij,kl})(1 - f) \right)$ in the last line of Eq. (1) has to be replaced by $r_m \left((1 - \frac{1}{2}c) F \bar{b}_{ij,kl} + (1 - \bar{b}_{ij,kl})(1 - f) \right)$. Notice that the only effect of changing these assumptions on the recursion equations, as given by (5) or the resulting matrix form (15), is that the effective recombination rate δ_{r_m} in (9) becomes

$$\delta_{r_m} = (1 - \frac{1}{2}b)(1 - f)r_m + \frac{1}{2}b(1 - f(1 - r_t)) \left(\frac{1}{4}c + (1 - \frac{1}{2}c)r_m \right). \tag{A.1}$$

Thus, if $r_m = \frac{1}{2}$ and $r_t = 1$, we obtain $\delta = \frac{1}{2} \left(1 - (1 - \frac{1}{2}b)f \right)$, which, as expected, is slightly larger than in the δ in (10). The qualitative properties of the dynamics remain unchanged. Quantitatively, there is slightly less linkage disequilibrium and the basin of attraction of the stable limit cycle is slightly increased.

A.2. Equilibria in the absence of mutation

Here we show that in the absence of mutation and if $\beta \neq \gamma$, no other equilibria than the four corner equilibria and the internal equilibrium \mathbf{x}^{*5} exist. From Eq. (21a) we infer that equilibria must satisfy $p = 0$ or $p = 1$ or $q = \frac{1}{2}$. If $p = 0$ or $p = 1$, then $D = 0$ must hold (otherwise the equilibrium is not in the state space, the simplex). Therefore, (21b) yields $q = 0$ or $q = 1$, whence (21c) yields $\Delta D = 0$.

Now assume $q = \frac{1}{2}$. Then (21b) yields $p = \frac{1}{2}$ at equilibrium. Substituting $p = q = \frac{1}{2}$ into the right-hand side of (21c) yields the cubic equilibrium condition

$$(1 + 2\beta D) \left[(\gamma - \beta)D^2 - \delta D - \frac{1}{16}(\gamma - \beta) \right] = 0. \tag{A.2}$$

The solution $D = -1/(2\beta)$ is not admissible because then $\bar{w} = 0$ by (20). The remaining quadratic has the solutions

$$D_0^* = -\frac{\sqrt{(\gamma - \beta)^2 + 4\delta^2} - 2\delta}{4(\gamma - \beta)} \tag{A.3}$$

and $-\frac{\sqrt{(\gamma - \beta)^2 + 4\delta^2} + 2\delta}{4(\gamma - \beta)}$, but the latter is not admissible because its value is outside the interval $[-\frac{1}{4}, \frac{1}{4}]$ under the parameter constraints (19). Therefore, the only internal equilibrium is \mathbf{x}^{*5} , and it has the coordinates $p = q = \frac{1}{2}$ and $D = D_0^*$. We note that D_0^* is negative if and only if $\gamma > \beta$ (cf. Úbeda et al., 2019). Therefore, there exist precisely five equilibria

The case $\beta = \gamma$ is degenerate (but only in the absence of mutation!). Simple calculations show that then the line segment $q = \frac{1}{2}, D = 0$, and $0 < p < 1$ consists of internal equilibria, and on the boundary $p = 0$ and $p = 1$ (with $D = 0$) are lines of equilibria.

A.3. Equilibria with weak mutation

A weak-mutation approximation or a weak-mutation perturbation of an equilibrium is achieved by assuming $\mu_A = \epsilon m_A$ and $\mu_B = \epsilon m_B$, where $m_A \geq 0$ and $m_B \geq 0$ are fixed, and then performing a series expansion of the relevant expressions to first order in ϵ . Finally, the substitutions $m_A \rightarrow \mu_A/\epsilon$ and $m_B \rightarrow \mu_B/\epsilon$ yield the desired approximation or perturbation in terms of the original parameters. The coordinates of the perturbed corner equilibria can be computed readily (see Supplementary Mathematica notebook, Section 3.2, where also the linear stability analysis can be found). To leading order in μ_A and μ_B , they are given by

$$\begin{aligned} x_1^{*1} &\approx 1 - [\mu_B\beta + \mu_A(\beta - \gamma)] \frac{1 + \beta}{\beta(\beta - \gamma)}, \\ x_2^{*1} &\approx \mu_B \frac{1 + \beta}{\beta - \gamma}, \\ x_3^{*1} &\approx \mu_A \frac{1 + \beta}{\beta}, \\ x_4^{*1} &\approx 0 \end{aligned} \tag{A.4}$$

and those of \mathbf{x}^{*4} are symmetric, i.e.,

$$x_1^{*4} = x_4^{*1}, x_2^{*4} = x_3^{*1}, x_3^{*4} = x_2^{*1}, x_4^{*4} = x_1^{*1}. \tag{A.5}$$

With or without mutation, there always exists a symmetric internal equilibrium, which we denote by \mathbf{x}^{*5} . We start by presenting its coordinates. Because the probability that no mutation occurs is $(1 - 2\mu_A)(1 - 2\mu_B)$, the total mutation probability is

$$\mu_{\text{tot}} = 1 - (1 - 2\mu_A)(1 - 2\mu_B). \tag{A.6}$$

Then the symmetric internal equilibrium is given by (Supplementary Mathematica notebook, Section 3.5)

$$p^* = q^* = \frac{1}{2} \tag{A.7a}$$

and

$$D^* = \frac{2\delta(1 - \mu_{\text{tot}}) + 2\mu_{\text{tot}} - \sqrt{R}}{4(\gamma - \beta) - 4\mu_{\text{tot}}(\beta + \gamma)}, \tag{A.7b}$$

where

$$R = 4[\delta(1 - \mu_{\text{tot}}) + \mu_{\text{tot}}]^2 + \gamma^2(1 - \mu_{\text{tot}})^2 - 2\beta\gamma(1 - \mu_{\text{tot}}) + \beta^2(1 - \mu_{\text{tot}}^2). \tag{A.7c}$$

It is straightforward to show that $D^* < 0$ if and only if $\gamma > \beta$ (also in the degenerate case when $\mu_{\text{tot}} = \frac{\gamma - \beta}{\beta + \gamma}$), $D^* > 0$ if and only if $\gamma < \beta$, and $D^* = 0$ if and only if $\gamma = \beta$.

In the degenerate case $\mu_{\text{tot}} = \frac{\gamma - \beta}{\beta + \gamma}$, one obtains

$$D^* = -\frac{\beta(\gamma - \beta)}{8(\gamma - \beta + 2\beta\delta)}, \tag{A.7d}$$

which is also negative because this case requires $\gamma > \beta$.

Assuming weak mutation, we observe that $\mu_{\text{tot}} \approx 2(\mu_A + \mu_B)$. A simple perturbation analysis, as outlined above, yields the following first-order approximation for D^* :

$$D^* \approx D_0^* + (\mu_A + \mu_B)D_1^*, \tag{A.8a}$$

where

$$D_1^* = \frac{2(\gamma - \beta + 2\beta\delta) \left(\sqrt{(\gamma - \beta)^2 + 4\delta^2} - 2\delta \right) - \beta(\gamma - \beta)^2}{2(\gamma - \beta)^2 \sqrt{(\gamma - \beta)^2 + 4\delta^2}}, \tag{A.8b}$$

and D_0^* is given in (A.3). Recall from above that $D_0^* < 0$ if and only if $\gamma > \beta$. Because the coefficient D_1^* of $\mu_A + \mu_B$ is positive if and only if $\gamma > \beta$, we find that mutation weakens linkage disequilibrium by reducing the absolute value of D^* .

Finally, we derive an explicit expression for the population mean crossover probability \bar{r}_t at the internal equilibrium \mathbf{x}^{*5} . What is of relevance is $\bar{r}_t/r_t = \frac{1}{2}b(x_1 + x_4 - D)$. If $p = q = \frac{1}{2}$, it follows from the transformations between gamete frequencies and allele frequencies and LD (Section 2.2.2) that $x_1 + x_4 - D = \frac{1}{2} + D$. Substituting D by the above expression for D^* in (A.8a), we obtain to leading order in μ ($= \mu_A = \mu_B$):

$$\bar{r}_t/r_t = \frac{1}{4}b \left[1 - \frac{\sqrt{(\gamma - \beta)^2 + 4\delta^2} - 2\delta}{2(\gamma - \beta)} + \mu \frac{2(\gamma - \beta + 2\beta\delta) \left(\sqrt{(\gamma - \beta)^2 + 4\delta^2} - 2\delta \right) - \beta(\gamma - \beta)^2}{2(\gamma - \beta)^2 \sqrt{(\gamma - \beta)^2 + 4\delta^2}} \right]. \tag{A.9}$$

If $\gamma > \beta$, this is always less than $\frac{1}{4}b$, thus less than $\frac{1}{2}$ of its values at an active hotspot (when x_1 or x_4 is close to 1). Note that with stronger mutation, \bar{r}_t/r_t increases (very slightly).

A.4. Attractors exhibit negative linkage disequilibrium if $\gamma > \beta$

Throughout this subsection, we assume $\gamma > \beta$ and $\delta \geq \beta$. However, a simple calculation shows that the latter is satisfied whenever $\gamma > \beta$. As a first step, we prove that $D \geq 0$ implies $\Delta^{(rs)}D \leq 0$ and $\Delta^{(rs)}D = 0$ only on the edges at which one locus is fixed. Let $D \geq 0$. The right-hand side of (21c) can be written as $d_1 + d_2 + d_3 + d_4$, where

$$\begin{aligned} d_1 &= -(\gamma - \beta)p(1-p)q(1-q) + (\gamma - \beta)D^2, \\ d_2 &= -[1 + \beta(2p - 1)(2q - 1)][\delta + \beta(2p - 1)(2q - 1)]D, \\ d_3 &= -\beta p(1-p)[\gamma - 2(\beta + \gamma)q(1-q)]D + 2\beta(\gamma - \beta)D^3, \\ d_4 &= [\beta(\gamma - 3\beta)(2p - 1)(2q - 1) - 2\beta\delta]D^2. \end{aligned}$$

Because $D^2 \leq p(1-p)q(1-q)$ and $\gamma > \beta$, we have $d_1 \leq 0$ and $d_1 = 0$ only if one of the allele frequencies vanishes. Because $\beta < \min\{\delta, 1\}$, each of the two factors in brackets is positive and we conclude that $d_2 \leq 0$ if $D \geq 0$. Next, we rewrite d_3 as

$$d_3 = -\beta D \{ \gamma p(1-p) - 2(\beta + \gamma)p(1-p)q(1-q) - 2(\gamma - \beta)D^2 \}.$$

Then, because $D^2 \leq p(1-p)q(1-q)$, we derive

$$\begin{aligned} \gamma p(1-p) - 2(\beta + \gamma)p(1-p)q(1-q) - 2(\gamma - \beta)D^2 \\ \geq \gamma p(1-p) - 2p(1-p)q(1-q)[(\beta + \gamma) + (\gamma - \beta)] \\ = \gamma p(1-p)[1 - 4q(1-q)] \geq 0. \end{aligned}$$

Therefore, $d_3 \leq 0$. Finally, we have $\beta(\gamma - 3\beta)(2p - 1)(2q - 1) - 2\beta\delta \leq \beta(|\gamma - 3\beta| - 2\delta)$. If $\gamma \leq 3\beta$, then $|\gamma - 3\beta| - 2\delta = 3\beta - \gamma - 2\delta = 2(\beta - \delta) + \beta - \gamma < 0$. If $\gamma > 3\beta$, then $|\gamma - 3\beta| - 2\delta = \gamma - 3\beta - 2\delta < 0$ by using $\delta \geq \frac{1}{2}\gamma$ from (19). This shows that $d_4 \leq 0$. Therefore, $\Delta^{(rs)}D \leq 0$ if $D \geq 0$ and $\Delta^{(rs)}D < 0$ unless at least one of the alleles is absent.

If mutation is weak (but present), so that the higher-order mutation terms in (22c) can be ignored, it follows immediately that $\Delta D < 0$ if $D \geq 0$. In a similar way it can be proved that $D' < 0$ if $D < 0$ (see Supplementary *Mathematica* notebook, Section 2.5). Hence, trajectories not only enter the region $D' < 0$ but also remain there. Therefore, the statement that attractors exhibit negative linkage disequilibrium follows.

References

Alves, I., Houle, A.A., Hussin, J.G., Awadalla, P., 2017. The impact of recombination on human mutation load and disease. *Philos. Trans. R. Soc. B* 372 (1736), 20160465.

Arnheim, N., Calabrese, P., Tiemann-Boege, I., 2007. Mammalian meiotic recombination hot spots. *Annu. Rev. Genet.* 41, 369–399.

Barton, N.H., Turelli, M., 1991. Natural and sexual selection on many loci. *Genetics* 127 (1), 229–255.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., De Massy, B., 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327 (5967), 836–840.

Baudat, F., Imai, Y., De Massy, B., 2013. Meiotic recombination in mammals: localization and regulation. *Nature Rev. Genet.* 14 (11), 794.

Boulton, A., Myers, R.S., Redfield, R.J., 1997. The hotspot conversion paradox and the evolution of meiotic recombination. *Proc. Natl. Acad. Sci.* 94 (15), 8058–8063.

Bürger, R., 2000. *The Mathematical Theory of Selection, Recombination, and Mutation*. John Wiley & Sons.

Bürger, R., 2020. Multilocus population-genetic theory. *Theor. Popul. Biol.* 133, 40–48.

Butlin, R.K., 2005. Recombination and speciation. *Mol. Ecol.* 9 (14), 2621–2635.

Calabrese, P., 2007. A population genetics model with recombination hotspots that are heterogeneous across the population. *Proc. Natl. Acad. Sci.* 104 (11), 4748–4752.

Cavassim, M.I.A., Baker, Z., Hoge, C., Schierup, M.H., Schumer, M., Przeworski, M., 2022. PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2. *Proc. Natl. Acad. Sci.* 119 (9).

Coop, G., Myers, S.R., 2007. Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genetics* 3 (3), e35.

Coop, G., Wen, X., Ober, C., Pritchard, J.K., Przeworski, M., 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319 (5868), 1395–1398.

Fledel-Alon, A., Wilson, D.J., Broman, K., Wen, X., Ober, C., Coop, G., Przeworski, M., 2009. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet.* 5 (9), e1000658.

Frankham, R., 2005. Genetics and extinction. *Biol. Cons.* 2 (126), 131–140.

Gerton, J.L., Hawley, R.S., 2005. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nature Rev. Genet.* 6 (6), 477.

Goyon, C., Lichten, M., 1993. Timing of molecular events in meiosis in *Saccharomyces cerevisiae*: stable heteroduplex DNA is formed late in meiotic prophase. *Mol. Cell. Biol.* 1 (13), 373–382.

Haig, D., Grafen, A., 1991. Genetic scrambling as a defence against meiotic drive. *J. Theoret. Biol.* 153 (4), 531–558.

Hansen, T.F., 2006. The evolution of genetic architecture. *Annu. Rev. Ecol. Evol. Syst.* 37, 123–157.

Heil, C.S., Ellison, C., Dubin, M., Noor, M.A., 2015. Recombining without hotspots: A comprehensive evolutionary portrait of recombination in two closely related species of *Drosophila*. *Genome Biol. Evol.* 7 (10), 2829–2842.

Jeffreys, A.J., Neumann, R., 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genet.* 31, 267–271.

Jeffreys, A.J., Neumann, R., 2005. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genetics* 14 (15), 2277–2287.

Karlin, S., Feldman, M.W., 1970. Linkage and selection: two locus symmetric viability model. *Theor. Popul. Biol.* 1 (1), 39–71.

Karlin, S., McGregor, J., 1972. Polymorphism for genetic and ecological systems with weak coupling. *Theor. Popul. Biol.* 3, 210–238.

Latrille, T., Duret, L., Lartillot, N., 2017. The red queen model of recombination hot-spot evolution: A theoretical investigation. *Philos. Trans. R. Soc. B* 372 (1736), 20160463.

Lenormand, T., Engelstädter, J., Johnston, S.E., Wijnker, E., Haag, C.R., 2016. Evolutionary mysteries in meiosis. *Philos. Trans. R. Soc. B* 371 (1706), 20160001.

Lichten, M., Goldman, A.S., 1995. Meiotic recombination hotspots. *Annu. Rev. Genet.* 29 (1), 423–444.

Myers, S., Bottolo, L., Freeman, C., McVean, G., Donnelly, P., 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310 (5746), 321–324.

- Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., Donnelly, P., 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327 (5967), 876–879.
- Myers, S., Freeman, C., Auton, A., Donnelly, P., McVean, G., 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genet.* 40 (9), 1124–1129.
- Nagaoka, S.I., Hassold, T.J., Hunt, P.A., 2012. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nature Rev.* 13, 493–504.
- Otto, S.P., Lenormand, T., 2002. Resolving the paradox of sex and recombination. *Nature Rev., Genetics* 4 (3), 252–261.
- Padmore, R., Cao, L., Kleckner, N., 1991. Temporal comparison of recombination and synaptonemal complex formation during meiosis in *S. cerevisiae*. *Cell* 6 (66), 1239–1256.
- Paigen, K., Petkov, P., 2010. Mammalian recombination hot spots: properties, control and evolution. *Nature Rev. Genet.* 11 (3), 221.
- Peñalba, J.V., Wolf, J.B., 2020. From molecules to populations: appreciating and estimating recombination rate variation. *Nature Rev. Genet.* 21 (August 2020), 476–492.
- Peters, A., 2008. A combination of cis and trans control can solve the hotspot conversion paradox. *Genetics* 178, 1579–1593.
- Petes, T.D., 2001. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* 2 (5), 360.
- Pineda-Krch, M., Redfield, R., 2005. Persistence and loss of meiotic recombination hotspots. *Genetics* 169, 2319–2333.
- Ponting, C.P., 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Trends Genetics* 5 (27), 165–171.
- Powers, N.R., Dumont, B.L., Emori, C., Lawal, R.A., Brunton, C., Paigen, K., Handel, M.A., Bolcun-Filas, E., Petkov, P.M., Bhattacharyya, T., 2020. Sexual dimorphism in the meiotic requirement for PRDM9: A mammalian evolutionary safeguard. *Sci. Adv.* 6 (43), eabb6606.
- Ptak, S.E., Hinds, D.A., Koehler, K., Nickel, B., Patil, N., Ballinger, D.G., Przeworski, M., Frazer, K.A., Pääbo, S., 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genet.* 37 (4), 429.
- Ptak, S.E., Roeder, A.D., Stephens, M., Gilad, Y., Pääbo, S., Przeworski, M., 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol.* 2 (6), e155.
- Ren, H., Ferguson, K., Kirkpatrick, G., Vinning, T., Chow, V., Ma, S., 2016. Altered crossover distribution and frequency in spermatocytes of infertile men with azoospermia. *PLoS One* 11 (6), e0156817.
- Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., Boehnke, M., 2010. Genome-wide association studies in diverse populations. *Nature Rev. Genet.* 11 (5), 356.
- Schenk, H., Schulerburg, H., Traulsen, A., 2020. How long do red queen dynamics survive under genetic drift? A comparative analysis of evolutionary and eco-evolutionary models. *BMC Evol. Biol.* 9 (20), 1–14.
- Schenk, H., Traulsen, A., Gokhale, C., 2017. Chaotic provinces in the kingdom of the red queen. *J. Theoret. Biol.* (431), 1–10.
- Slatkin, M., 2008. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Rev.* 9, 477–485.
- Stapley, J., Feulner, P., Johnston, S., Santure, A., Smadja, C., 2017. Recombination: the good, the bad and the variable. *Philos. Trans. R. Soc. B* 372 (1736), 20170279.
- Stevenson, L.S., Woerner, A.E., Kidd, J.M., Kelley, J.L., Veeramah, K.R., McManus, K.F., Project, G.A.G., Bustamante, C.D., Hammer, M.F., Wall, J.D., 2015. The time scale of recombination rate evolution in great apes. *Mol. Biol. Evol.* 33 (4), 928–945.
- Sun, H., Treco, D., Szostak, J.W., 1991. Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. *Cell* 64 (6), 1155–1161.
- Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J., Stahl, F.W., 1983. The double-strand-break repair model for recombination. *Cell* 33 (1), 25–35.
- Úbeda, F., Russell, T., Jansen, V., 2019. PRDM9 and the evolution of recombination hotspots. *Theor. Popul. Biol.* 126, 19–32.
- Úbeda, F., Wilkins, J.F., 2011. The red queen theory of recombination hotspots. *J. Evol. Biol.* 24 (3), 541–553.
- Van Valen, L., 2019. A new evolutionary law. *Evol. Theory* 1, 1–30.
- Winckler, W., Myers, S.R., Richter, D.J., Onofrio, R.C., McDonald, G.J., Bontrop, R.E., McVean, G.A., Gabriel, S.B., Reich, D., Donnelly, P., et al., 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308 (5718), 107–111.