The interplay between recombination hotspots, GC-biased gene conversion and natural selection

Julien Joseph

Université Claude Bernard Lyon 1

THESE de DOCTORAT DE
L'UNIVERSITE CLAUDE BERNARD LYON 1

Ecole Doctorale N° 341
Evolution, Ecosystèmes, Microbiologie, Modélisation

Spécialité de doctorat : Biologie
Discipline : Biologie évolutive

Soutenue publiquement le 12/12/2023, par :
Julien JOSEPH

The interplay between recombination hotspots, GC-biased gene conversion and natural selection

---

Devant le jury composé de :

Pr. Cristina VIEIRA-HEDDI	Pr., UCBL Lyon 1, LBBE	Présidente
Dr. Gwenaël PIGANEAU	DR, CNRS, OOB	Rapporteure
Pr. Matthew WEBSTER	Pr., Uppsala University	Rapporteur
Pr. Denis ROZE	DR, CNRS, SBR	Rapporteur
Dr. Aline MUYLE	CR, CNRS, CEFE	Examinatrice
Dr. Laurent DURET	CR, CNRS, LBBE	Directeur de thèse
Dr. Nicolas LARTILLOT	Dr, CNRS, LBBE	Co-directeur de thèse

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de
l'Université Claude Bernard Lyon 1

École doctorale n°341
Évolution, Écosystèmes, Microbiologie, Modélisation

Spécialité de doctorat: Biologie
Discipline: Biologie évolutive

Soutenue publiquement le 12/12/2023, par :

Julien JOSEPH

The interplay between recombination hotspots, GC-biased gene conversion and natural selection

Devant le jury composé de :

Dr. Gwenaël PIGANEAU, DR, CNRS, OOB	Rapportrice
Pr. Matthew WEBSTER, Pr., Uppsala University	Rapporteur
Dr. Denis ROZE, DR, CNRS, SBR	Rapporteur
Pr. Cristina VIEIRA-HEDDI, Pr., UCBL Lyon 1, LBBE	Examinatrice
Dr. Aline MUYLE, CR, CNRS, CEFE	Examinatrice

Dr. Laurent DURET, DR, CNRS, LBBE	Directeur de thèse
Dr. Nicolas LARTILLOT, DR, CNRS, LBBE	Co-directeur de thèse

Abstract

GC-biased gene conversion (gBGC) is a genetic mechanism associated with recombination that skews the segregation ratio of AT:GC polymorphisms in the gametes of heterozygotes. Since its discovery in the 1980s, much effort has been devoted to describing how this process affects base composition, rate of evolution and fitness of many eukaryotes, particularly humans. In particular, gBGC has been shown to be the major driver of variation in GC content along the human genome. In addition, it induces the fixation of GC alleles independently of their effect on fitness and can therefore cause a significant deleterious burden. However, the discovery of gBGC is relatively recent and our understanding of its role in genome evolution is still in its infancy. In this thesis, I first characterised the dynamics of gBGC both across genomes and across lineages. I showed that in mammals, there are some hotspots of gBGC in the genome that are evolutionary stable and therefore largely shared between species. I then showed that these stable hotspots can be subject to positive selection when they disappear, which can confound tests of adaptive evolution. Finally, I showed that the levels of gBGC currently found in eukaryotes can be positively selected for, despite inducing a substantial burden at the population level. Overall, this thesis provides novel insights into the dynamics of recombination landscapes, selection and the drivers of gBGC evolution.

Résumé

La conversion génique biaisée vers GC (gBGC) est un mécanisme génétique associé à la recombinaison qui biaise la répartition des polymorphismes AT:GC dans les gamètes des hétérozygotes. Depuis sa découverte dans les années 1980, de nombreux efforts ont été consacrés à décrire la manière dont ce processus affecte la composition en bases, le taux d'évolution et la fitness de nombreux eucaryotes, en particulier de l'humain. Notamment, il a été démontré que la gBGC est le principal moteur de la variation du contenu en GC le long du génome humain. En outre, il induit la fixation d'allèles G ou C indépendamment de leur effet sur la fitness et peut donc avoir un effet délétère important. Cependant, la découverte de la gBGC est relativement récente et notre compréhension de son rôle dans l'évolution du génome n'en est qu'à ses débuts. Dans cette thèse, j'ai d'abord caractérisé la dynamique de la gBGC à la fois à travers les génomes et à travers les lignées de mammifères. J'ai montré que chez les mammifères, il existe des points chauds de gBGC dans le génome qui sont stables au cours de l'évolution et donc largement partagés entre espèces. J'ai ensuite montré que ces points chauds stables peuvent faire l'objet d'une sélection positive lorsqu'ils disparaissent, ce qui peut fausser les tests d'évolution adaptative. Enfin, j'ai montré que les niveaux de gBGC que l'on observe actuellement

chez les eucaryotes peuvent faire l'objet d'une sélection positive, bien qu'ils induisent un fardeau substantiel au niveau de la population. Dans l'ensemble, cette thèse apporte un éclairage nouveau sur la dynamique des paysages de recombinaison, la sélection et les moteurs de l'évolution de la gBGC.

Résumé étendu

La fécondation conduit au regroupement et la coexistence de deux génomes parentaux distincts et séparés dans la même cellule oeuf. Ces deux génomes vont rester séparés pendant toute la vie de l'individu. La recombinaison méiotique, bien moins connue du grand public, est l'étape pendant laquelle les deux génomes parentaux s'entremêlent dans les cellules sexuelles de leurs descendants, pour être transmis aux petit enfants. La recombiaison méiotique est donc l'essence véritable de la reproduction sexuée, ou deux parties d'individus n'en forment plus qu'un. Les lois qui régissent la façon dont deux génomes s'entremêlent lors de la formation des cellules sexuelles vont donc avoir une importance toute particulière sur quelle partie de l'information génétique va être transmise à la génération suivante, et donc sur le cours de l'évolution.

Une de ces lois décrit la chance qu'une base de l'ADN (A,T,C ou G), à un endroit donné du génome, soit transmise à la génération suivante. Selon les lois de la génétique énoncées par Gregor Mendel à la fin du XIXème siècle, cette probabilité est de 50% en moyenne. En d'autre terme, dans une population, pour chaque base du génome, il y a en moyenne 50% de chance qu'elle vienne du géniteur, et 50% de chance qu'elle vienne de la génitrice. Cette loi est largement vérifiée la plupart du temps. Cependant, lors des évènements de recombinaison, à l'endroit ou les chromosomes fusionnent (qui change à chaque nouvelle cellule sexuelle), les bases guanines et cytosines (G et C) ont plus de chances de passer à la génération suivante que les bases adénine et thymine (A et T). C'est ce qu'on appelle la conversion génique biaisée vers GC (gBGC en anglais). Dans les régions du génome fortement susceptibles de recombiner, ce processus conduit à la transmission biaisée des bases G et C au cours des générations, et eventuellement à leur fixation dans les populations. Ce phénomène n'est pas anectodique : c'est le déterminant majeur de la variation de la composition en base G et C le long du génome humain. Aussi, les bases G et C sont transmises peu importe leur effet sur la capacité reproductive des individus (la fitness en anglais). Ce phénomène interfère donc avec la séléction naturelle en favorisant la transmission de mutations potentiellement délétères.

Depuis sa découverte dans les années 80, de nombreux efforts ont été déployés pour décrire la manière dont la gBGC affecte la composition en bases, le taux d'évolution et la fitness chez de nombreux eucaryotes et en particulier chez l'humain. Cependant, la découverte de la gBGC est relativement récente et notre compréhension de son rôle dans l'évolution des génomes n'en est qu'à ses balbutiements. Pour comprendre l'impact global

de la gBGC sur l'évolution, nous devons d'abord caractériser précisément ses variations le long du génome. En supposant que l'intensité de la distorsion de ségrégation est relativement uniforme à travers le génome, les variations de l'intensité de la gBGC sont principalement le résultat des variations du taux de recombinaison. Dans cette thèse, nous avons développé une méthode pour estimer les taux de recombinaison à partir des signatures de gBGC afin de mieux comprendre la distribution à fine échelle des événements de recombinaison le long du génome.

En particulier, des travaux précédents ont démontré que la plupart des évènements de recombinaiosns étaient concentrées dans un petit nombre de séquences : les points chauds de recombinaison. Chez l'humain et la souris, la position des points chauds de recombinaison est déterminée par une protéine appelée Prdm9 qui se lie à un motif spécifique de l'ADN (e.g. AATTCATACTT). Lorsque cette protéine est absente, les points chauds de recombinaison sont relocalisés vers des zones de l'ADN portant des marques épigénétiques caractéristiques du début de la plupart des gènes chez l'humain. Ce schéma est observé chez la grande majorité des Eucaryotes dépourvus de cette protéine. Il a donc été proposé que le rôle de cette protéine était de diriger les points chauds de recombinaison loin de ces séquences appelées points chauds par défaut. En utilisant notre méthode basée sur la gBGC sur 52 espèces de mammifères placentaires, nous avons montré que l'humain et la souris ne sont pas représentatifs des autres espèces. En effet, pour certaines espèces, même si le gène Prdm9 semble être actif, les points chauds par défaut reçoivent autant d'évènements de recombinaisons que ceux des espèces dépourvues de Prdm9.

Deuxièmement, nous avons établi un cadre théorique pour comprendre l'impact du gBGC sur la dynamique de la sélection naturelle. Nous montrons qu'en favorisant la fixation de mutations délétères vers GC, la gBGC éloigne les gènes de leur optimum de fitness. Par conséquent, ces gènes ont plus d'opportunités pour des mutations bénéfiques. Si le taux de recombinaison diminue, ces mutations sont efficacement sélectionnées. Ainsi, paradoxalement, même si la gBGC est principalement délétère, ses fluctuations induisent de la sélection positive. Dans certaines conditions, nous montrons que ce phénomène peut conduire à une sélection positive plus élevée dans les gènes à forte recombinaison, et donc imite un effet bénéfique de la recombinaison sur l'efficacité de la sélection.

Finalement, à l'aide de modèles théoriques, nous montrons qu'en dépit du fait que le gBGC augmente le fardeau de mutations délétères moyen dans la population, en convertissant des mutations fortement délétères fraîchement arrivées vers AT, il donne toujours un avantage net positif en termes de fitness au niveau individuel. Cela peut expliquer pourquoi le gBGC n'est pas contre-sélectionné et universellement présent chez les eukaryotes. L'existence du gBGC pourrait donc résulter d'une tragédie des communs,

où l'avantage individuel de biaiser la transmission des allèles GC conduit à une réduction globale de la capacité reproductive de la population.

C'est un bon rappel que la séléction naturelle n'optimise pas la reproduction et la survie des espèces dans leur environement, mais des individus, ce qui peut paradoxalement décroître celle de l'espèce dans son ensemble.

Contents

Part I

General introduction

1

The laws of inheritance

1.1 Inheritance and reproductive system

In its broad sense, inheritance, can be defined as the process through which anything passes from one generation to the next. In several human and non human societies, social status, culture, traditions or habitat can be passed on from one generation to the next. From a biological perspective, we also inherit DNA, the cytoplasmic content of our parents, DNA or histone methylation, and many other epigenetic marks. The link between two generations of individuals is a reproduction event, which is more or less a copy/paste of heritable information. A mutation, in its evolutionary sense, can be defined as an error in this copy/paste event. Without heritable information, there cannot be mutations, only damages. Similarly, if a trait cannot be transmitted, it cannot be selected for. Without heredity, there cannot be natural selection. Inheritance is therefore at the core of Darwinian evolution, and the different processes that govern reproduction and heritability deeply affect the course of evolution.

In human societies, there are different political systems, in which sets of laws describe more or less precisely what wealth or goods can or cannot be inherited from parents to children, and in what proportions. Similarly, in biology, there are different reproductive systems in which the laws of inheritance dictate how genetic information is passed from

one generation to the next. A reproductive system can therefore be defined as a set of laws of inheritance. There can be a reproductive system for any type of heritable information, but throughout this manuscript we will focus on the inheritance of DNA.

The reproductive system which by far underlies most reproductive events throughout the tree of life is DNA replication of the whole genome with a small number of errors. Offspring thus get the integrity of the genome of a single parent, except for a usually small number of mutations. This is also called clonal, or asexual reproduction. It is the major reproductive system of many unicellular organisms including eubacteria and archaea, and the reproductive system of the cells of many multicellular Eukaryotes (i.e. mitosis). However, for most eukaryotes, clonal reproduction of cells by mitosis is an evolutionary dead-end. Instead, the long-term persistence of genetic information through time is ensured by sexual reproduction.

1.2 Sexual reproduction

Sexual reproduction regroups several reproductive systems that all share two essential steps in eukaryotes. It starts with the fusion of two genomes within one cell (syngamy), and ends with the distribution of those two genomes in several cells with a single genome (meiosis). It is thus a cycle with a haploid phase (one genome) and a diploid phase (two genomes). Of note, between syngamy and meiosis, or between meiosis and syngamy, there can be a long period of time with a large number of mitotic clonal reproduction events. It is still not entirely clear why sexual reproduction is so prevalent compared to completely asexual or clonal reproduction. Indeed, it comes with several substantial costs, such as finding and securing a partner for syngamy, exposure to predators during mating, or the transmission of diseases and genetic parasites (Otto and Lenormand, 2002).

One convincing hypothesis for the prevalence of sex is a substantial increase in fitness owed to an efficient masking of deleterious mutations, called the sheltering or the masking hypothesis (Kondrashov and Crow, 1991; Otto and Lenormand, 2002). As explained before, mutations between two reproduction events are virtually inevitable. Even if some mutations provide a selective advantage, most new mutations are deleterious. However, many deleterious mutations are also recessive, meaning that when only one genome carries the deleterious mutation, the fitter phenotype can be maintained by the wild-type allele on the other genome. The presence of two genotypes in a cell that do not carry the same mutations can therefore provide increased robustness to the fitness cost induced by deleterious mutations (the mutation load, or genetic load). Interestingly, some species spend most their time with one genome, with the diploid phase being only transient,

meaning that meiosis immediately follows syngamy. In these species, the sheltering hypothesis cannot explain the long-term maintenance of sex, because species spend most their time unsheltered. Instead, the crucial advantage of sexual reproduction could come from the elimination of deleterious mutations through the shuffling of genetic material during reproduction, and particularly through a process called meiotic recombination.

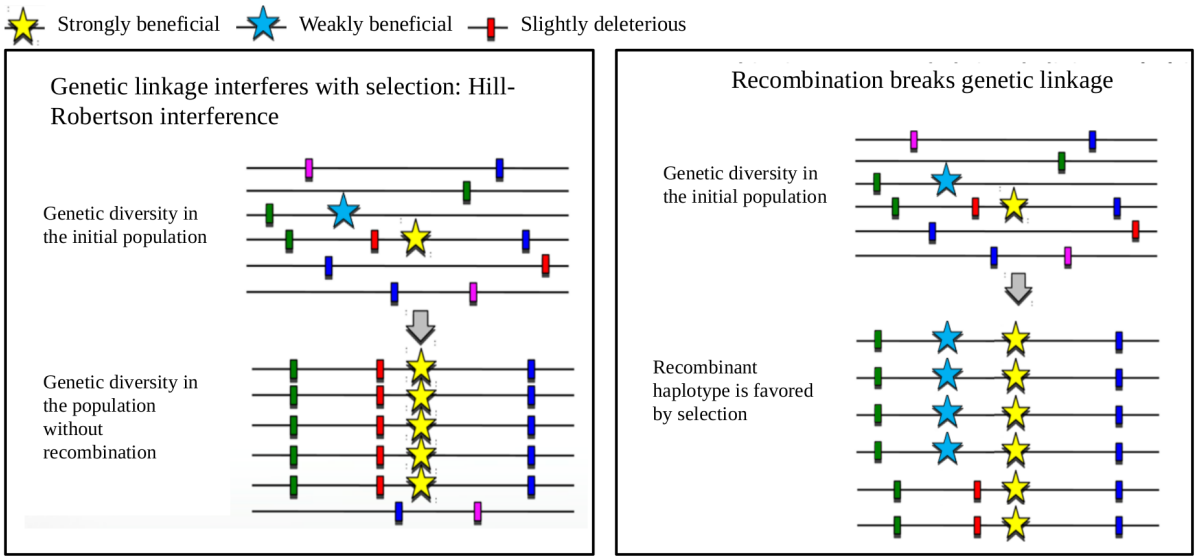


Figure 1.1: *Schematic view of the impact of recombination on the efficiency of selection. Left: Without recombination, the most fit existing haplotype rises in frequency and reach fixation. Right: Recombination allows for the creation of an even fitter haplotype, increasing the frequency of beneficial mutations, and decreasing that of deleterious mutations.*

In its broad sense, meiotic recombination is the exchange of genetic material between the two parental genomes during meiosis. This exchange can be very local (at the scale of few kilobases) and is called gene conversion or non-crossover (NCO) (Winkler, 1930; Roman, 1985), or instead, chromosomes can exchange large pieces of DNA through crossovers (COs). A major consequence of recombination is that offsprings receive genetic information from both parents. This also creates new combination of alleles on which natural selection can act. Some combinations will carry the most deleterious alleles of the two parental genome, and will therefore be less fit, but others will have the most beneficial alleles, and will be more fit (Figure 1.1). Importantly, this phenomenon does not increase the mean fitness of offsprings, but only the variance. The fitness advantage of recombination is therefore not direct. Because non-recombining lineages cannot explore higher levels in the fitness landscape (because of the small variance in offsprings' fitness), at one point they will be outcompeted by the fittest individuals from the pool of offsprings that have recombined (Hill and Robertson, 1966). It is generally accepted that in a finite population with mostly deleterious mutations, meiotic recombination provides an indirect fitness advantage that is sufficiently high to be efficiently selected for (Felsenstein, 1974; Hickey and

Golding, 2018; Roze, 2021).

Even if sexual reproduction is widespread in eukaryotes, it can take several forms, and many laws dictate more specifically how DNA is inherited. Importantly, these laws are not written in marble, and can evolve. Notably, the choice of the mating partners varies widely across organisms. In some systems, not all individuals can mate with each other. There are mating types (often called sexes), which are mostly DNA encoded, and only gametes with different mating type/sex can fuse into one. In some species, the likelihood of two individuals mating may depend on phenotypic similarity, with individuals preferring to mate with their relatives, while in others incest is actively avoided.

During meiosis, a particular law of inheritance that is central to evolutionary theory describes the chances that a given allele will be passed on to the gametes (haploid phase).

1.3 Mendel's law of inheritance

Mendelian genetics is based on the principle that if a diploid individual is heterozygous for an allele, the probability of that allele being transmitted to a gamete is 0.5 (Carothers, 1913). With a very large population size, this maintains genetic variation on which selection can act for a very long time (Hardy, 1908; Weinberg, 1908). Under Mendel's law, the evolution of the frequency of a perfectly neutral allele in a population is totally random and is called genetic drift. And thus, natural selection can be detected as a deviation from this random evolution. In the end, almost all evolutionary theory is based on Mendel's 50:50 law (Lyttle, 1991; Hurst, 2019).

Under Mendel's law, a given genome will only transmit half of itself to the next generation. It appears clearly that if somehow an allele provided to this genome a "cheating" advantage that would allow it to be transmitted more often, let's say $0.5 + \epsilon$, this allele and its neighbours should establish easily in the population (Lyttle, 1991). Several evidences for such loci exist in several eukaryotic organisms, and are called segregation distorters (reviewed in Lyttle (1991); Taylor and Ingvarsson (2003)). The phenomenon by which segregation distorters find themselves in more than half of the gametes is called meiotic drive. In this manuscript, we will adhere to the definition of Bengtsson and Uyenoyama (1990), which distinguishes segregation distorters that can induce non-Mendelian segregation at any site from meiotic drivers that cause their own non-Mendelian segregation.

Some segregation distorters operate at the chromosome scale. In the mammalian female meiosis, only one of the meiotic products will become the egg, and therefore homologous chromosomes that are addressed more often to the egg than to the polar cells will be transmitted to the next generation more often. This can happen when a chromosome of a given pair is more mobile, because of its smaller size or its more compacted chromatin state (reviewed in Clark and Akera (2021)). One major category of segregation distorters are sex ratio distorters. They bias the sex ratio by producing more gametes with either sex-determining alleles/chromosomes. Such sex-linked chromosome drive have been reported in white campion, lemmings, mosquitoes, butterflies, and fruit flies (reviewed in Lyttle (1991)). Another extreme example of segregation distortion occurs in wild mice, where chromosome 17 can carry the t-haplotype that is transmitted 99% of the time instead of 50% in males (Bruck, 1957). This t-haplotype is lethal when homozygous, and despite this, it is maintained in the population (Bruck, 1957).

In this thesis, I will first study how two central recombination-associated laws of inheritance affect natural selection, and conversely, I will assess to what extent these two laws of inheritance evolve through natural selection. The first law of inheritance describes how recombination events are distributed along the genome at fine scale (fine-scale recombination landscape). The second law of inheritance describes the extent to which AT:GC polymorphism segregation in gametes deviate from the 50:50 Mendel's law at the location of recombination events; a special segregation distortion known as GC-biased gene conversion.

2

Recombination hotspots

2.1 Recombination and meiosis success

The most commonly cited hypothesis for the long-term maintenance of recombination is its advantage in increasing the fitness variance of offsprings. However, meiotic recombination is also a fundamental step of sexual reproduction because it is mandatory for meiosis success and fertility for a vast majority of eukaryotes.

2.1.1 The obligatory cross-over

In meiosis, two genomes in one cells are dispatched in two cells with one genome. The different chromosomes that constitute these two genomes are mixed in the cell's nucleus, and the first requirement to be able to pass on complete individual genomes to

the gametes is to properly segregate homologous chromosomes. An incapacity to properly segregate homologous chromosome often leads to meiotic arrest and infertility (Baker *et al.*, 1976; Hassold *et al.*, 2007; Brick *et al.*, 2012; Davies *et al.*, 2016; Mihola *et al.*, 2019, 2021). In most eukaryotes, the initiation of meiotic recombination is crucial for this chromosome pairing (reviewed in Baker *et al.* (1976)). In the early stages of meiosis, chromosomes are organised into loop arrays around a proteinaceous chromosome axis (Moses, 1956; Fawcett, 1956). Meiotic recombination starts with several programmed double-strand breaks (DSBs) performed by SPO11, an evolutionary conserved proteins in eukaryotes, on the chromosomal axis (Keeney *et al.*, 1997; Bergerat *et al.*, 1997; Blat *et al.*, 2002). Instead of a clear cut, it erases part of the sequence on both strands symmetrically around the cut, from 5′ to 3′ (Szostak *et al.*, 1983). The DNA is thus locally single-stranded at this location. With the help of mediator proteins such as RAD51 and DMC1, the single-stranded DNA (called the nucleoprotein filament) will find its match on the homologous chromosome and hybridise (Camerini-Otero and Hsieh, 1995). They will then be repaired either as a CO, or a NCO. On each chromosome, a subset of the DSBs repaired as a CO will attract the so-called synaptonemal complex which stabilizes the interaction between homologous chromosome and coordinates their placement on the polar axis, for further segregation in the second division of meiosis (Baker *et al.*, 1976). COs are therefore crucial for the production of offsprings. The requirement of at least one CO per chromosome arm put strong constraints on the genome-wide recombination rate, and is largely responsible for the differences of recombination rate (per bp) between chromosomes of different sizes (Pardo-Manuel de Villena and Sapienza, 2001; Stapley *et al.*, 2017; Brazier and Glémin, 2022). However some species (mostly insects) do not require COs for correct chromosome segregation (reviewed in Gerton and Hawley (2005)). These so-called achiasmate meiosis are often sex-specific, and are likely due to the peculiar structure of chromosomes in those species (Nokkala *et al.*, 2004; Cabral *et al.*, 2014).

2.1.2 The intractable problem of homology search

For a DSB to be repaired, and chromosomes to correctly segregate, the nucleoprotein filament needs to find its homologous sequence among sometimes billions of base pairs. It has been shown that the nucleoprotein filament takes a lot of time to probe for homology, sliding on the target DNA. As a consequence, if the entire genome had to be searched, meiosis would take unreasonable amount of time (Barzel and Kupiec, 2008; Weiner *et al.*, 2009). This search of homology is therefore facilitated, for instance by premeiotic pairing between homologs (Chikashige *et al.*, 1994; Fung *et al.*, 1998), or by restricting the zones of the genome that are susceptible to receive DSBs (e.g. euchromatin) (Renkawitz

et al., 2014; Peñalba and Wolf, 2020).

In this context, the concentration of recombination events into a small fraction of the genome appears to be a good way to restrict homology search, and therefore to increase the success of meiosis. Interestingly, in many species, some short sequence (1kb) experience a much higher recombination rate than their regional background (Lichten and Goldman, 1995; Petes, 2001; Choi and Henderson, 2015). We call these loci recombination hotspots. Recombination hotspots have been observed in a wide range of eukaryotes, with the notable exception of *Drosophila melanogaster* *Caenorhabditis elegans* and *Apis mellifera* (Chan *et al.*, 2012; Kaur and Rockman, 2014; Wallberg *et al.*, 2015). Whether the raison d'être of these recombination hotspots is solely to accelerate meiosis is still not clear, and many other hypothesis have been formulated (Webster and Hurst, 2012; Brick *et al.*, 2012). To answer this question, considerable insights can be gained by studying the molecular mechanisms that determines them.

2.2 Prdm9-dependent recombination hotspots

2.2.1 Basic principle

In humans, mice and chimps, the location of recombination hotspots is determined by the protein Prdm9 (Myers *et al.*, 2010; Parvanov *et al.*, 2010; Baudat *et al.*, 2010). This protein contains a zinc finger array that directly binds DNA. When Prdm9 binds DNA, it modifies the methylation states of histones, which allows the bound sequence to be tethered to the chromosomal axis where it will receive a DSB by SPO11 (Brick *et al.*, 2012; Eram *et al.*, 2014; Powers *et al.*, 2016; Davies *et al.*, 2016; Grey *et al.*, 2017; Hinch *et al.*, 2019; Li *et al.*, 2019). The zinc finger usually targets a motif of a dozen of base pair, which depends on the amino-acid residues at key sites of each zinc finger (Myers *et al.*, 2005; Smagulova *et al.*, 2016). Therefore, two individuals of a population carrying different Prdm9 alleles for the zinc finger domain will have different recombination hotspots (Pratto *et al.*, 2014; Smagulova *et al.*, 2016; Alleva *et al.*, 2021). The zinc finger array is extremely polymorphic in mice and humans (Buard *et al.*, 2014; Kono *et al.*, 2014; Alleva *et al.*, 2021), due to its high mutation rate (Jeffreys *et al.*, 2013). Indeed the zinc finger array is a mini-satellite sequence that experiences high duplication, deletion and conversion rate among zinc fingers (Jeffreys *et al.*, 2013). For the same number of individual zinc finger sequences, this induces a high copy number variation, which increases the diversity of zinc finger arrays. This high mutation rate is allowed by a very low divergence between zinc fingers inside an array, which is quite

unusual compared to the other zinc finger domains of other proteins (Baker *et al.*, 2017). Therefore Prdm9-directed hotspots are fast-evolving and largely differ between species (Auton *et al.*, 2012; Munch *et al.*, 2014). The maintenance of this high allelic diversity is usually explained by the so-called intragenomic red-queen dynamic of Prdm9.

2.2.2 The red queen hypothesis

If at a given site of the DNA, an individual is heterozygous for the DNA motif targeted by its Prdm9 allele, the DSB will always be initiated on the chromosome that carries the recognized DNA motif (hot allele) (Nicolas *et al.*, 1989; Schultes and Szostak, 1991; Detloff *et al.*, 1992). This DSB will be repaired using the other chromosome where the motif is absent as a template (cold allele). This will therefore favour the transmission of the cold allele to the gametes (Nicolas *et al.*, 1989; Schultes and Szostak, 1991; Detloff *et al.*, 1992; Boulton *et al.*, 1997). Consequently, cold alleles segregate above the 50:50 of Mendel's law. With this process, any mutation that alters a Prdm9 binding motif will spread rapidly in the population as a meiotic driver (Boulton *et al.*, 1997). Through this process, Prdm9 alleles progressively lose their targets (Baker *et al.*, 2015; Smagulova *et al.*, 2016), which should lead to a decrease in fertility due to an impaired capacity of correctly creating COs. The fitness of a Prdm9 allele therefore decreases with its age and its historical activity (Úbeda and Wilkins, 2011). New Prdm9 alleles with new targets are therefore positively selected (Úbeda and Wilkins, 2011) because their capacity to generate COs is higher. This leads to a red-queen dynamic between Prdm9 alleles and their targets where new alleles are systematically favoured by natural selection (Úbeda and Wilkins, 2011). This summarises the basic assumptions of Prdm9's red-queen hypothesis, but several models exist which mainly differ in the causes of the fertility loss following the decrease in the number of targets.

The model of Latrille et al. 2017

In the model of Latrille *et al.* (2017), there is an arbitrary function that connects the fitness of a given individual to the number of targets of its Prdm9 alleles, either a power law or an exponential. This model is therefore phenomenological, not modelling explicitly the mechanisms through which fertility is decreased with the decrease in the number of targets. This model already allowed to characterize two different regimes depending on Prdm9's mutation rate, and the speed of target erosion. The succession regime is characterized by a low population-scaled mutation rate of Prdm9 ($2N\mu$). In this regime, the population is dominated by one major Prdm9 alleles, while others are at low frequency (Figure 2.1A). As the major allele reaches high frequency, it erodes its

targets quickly, leading to positive selection on more recent low frequency alleles. One of these alleles increases in frequency and becomes the new major allele. This regime reflects the diversity of humans Prdm9 alleles, where one allele is major in all human populations, with many lower frequency alleles (Alleva *et al.*, 2021). On the other hand, the polymorphic regime is characterized by a high population-scaled mutation rate of Prdm9. In this regime, several alleles segregate at intermediate frequency (Figure 2.1B). This regime reflects the diversity of mice Prdm9 alleles, where several alleles segregate at intermediate frequency (Buard *et al.*, 2014; Kono *et al.*, 2014).

The model of Baker et al. 2022

In the past few years, experimental evidences revealed the mechanism through which DSB repair as CO is impaired when the number of Prdm9 targets decreases. In fact instead of the raw number of targets, it is the number of homozygous targets that is critical for fitness. Several recent studies showed that when sequences are symmetrically bound by Prdm9, the probability that the induced DSB will be repaired as a CO increases (Hinch *et al.*, 2019; Li *et al.*, 2019). Indeed, when Prdm9 binds symmetrically on both homologs, it tethers both homologous sequence to the chromosomal axis (Hinch *et al.*, 2019; Li *et al.*, 2019). As spatial proximity is crucial for efficient homology search (Renkawitz *et al.*, 2014), one can imagine that this symmetrical binding will accelerate DSB repair and increase the probability of making a CO. Two new factors will therefore be key to understand the fertility decrease following target erosion. The first one is the distribution of target affinity, and the second one is the concentration of Prdm9 in the meiotic cell.

The model of Baker *et al.* (2022) assumes that the concentration of Prdm9 is limiting in the cell, which leads to the competition between two kinds of hotspots, a small number of hotspots of high affinity, and a large number of hotspots of low affinity. Low affinity targets will titrate Prdm9 but without symmetrical binding and will provide the main explanation for which symmetrical binding is limited in the few high affinity targets. In this context, the key factor that decreases the probability of symmetrical binding and therefore fitness is the presence of low affinity targets and the limited concentration of Prdm9 in the cell. The model of Baker *et al.* (2022) allows to reconcile the red queen hypothesis with empirical evidence that clearly demonstrated that the number of target is not limiting (Kauppi *et al.*, 2013; Baker *et al.*, 2014; Diagouraga *et al.*, 2018), and therefore cannot explain the decrease in fitness of old Prdm9 alleles. Baker *et al.* (2022) argue that limiting the number of hotspots (low affinity ones) could provide a selective advantage by increasing the probability of symmetrical binding and therefore increase the success of meiosis.

The model of Genestier et al. 2023

The model of Genestier *et al.* (2023) extends the previous one by relaxing simplifying assumptions, notably regarding the distribution of affinity of Prdm9 targets, the concentration of Prdm9 in meiotic cells, and the expected behaviour of Prdm9 in heterozygotes through the impact of genetic dosage. The results reflect the previous one but with some important new insights. The first one is that Prdm9 need not to be limiting in the cell to explain a red-queen dynamic. The red-queen only emerges as a function of the number of those targets that are affine enough to be symmetrically bound, regardless of the number of low affinity targets.

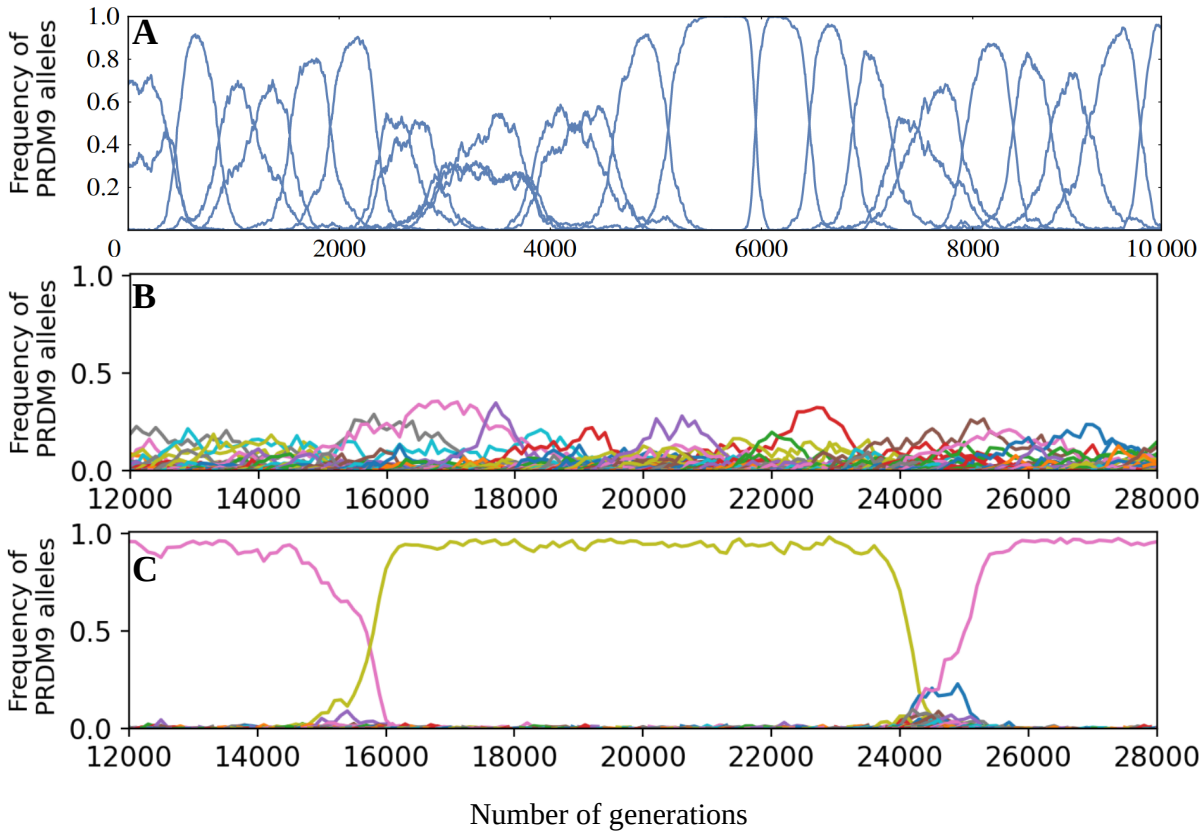


Figure 2.1: *A) Illustration of the succession regime (panel taken from Latrille et al. (2017)). At each time only one allele segregate at high frequency and is being replaced by new alleles when they arrive. B) Illustration of the polymorphic regime (panel taken from Genestier et al. (2023)). A lot of alleles segregate at intermediate frequency. C) Illustration of the eviction regime (panel taken from Genestier et al. (2023)). New alleles are under negative selection until the homozygous targets of the major allele become too scarce, and are therefore positively selected.*

In this view, there is no selection to limit the total number of hotspots but rather a selection to increase the number of high affinity ones.

On top of the succession regime and the polymorphic regime observed by Latrille *et al.* (2017), both models highlight a third regime for the evolution of Prdm9. The eviction regime is characterized by a difference in the probability of symmetrical binding between homozygotes and heterozygotes (Figure 2.1C). This effect is amplified when the genetic dosage of Prdm9 is taken into account, namely that the concentration of a given Prdm9 allele in the cell is divided by two in heterozygotes. In the eviction regime, new alleles that appears in a heterozygous states are less likely to have symmetrical binding, and new alleles are negatively selected. Positive selection occurs only when the old major allele has eroded enough of its high affinity target to be less fit than heterozygotes. This is in bright contrast with the model of Latrille *et al.* (2017) where every new allele was immediately positively selected. In particular, in the eviction regime, both Prdm9's diversity and the speed of the turnover drastically decrease.

Empirical evidences

The critical aspect that is necessary for the red queen to operate is the decrease of fitness of old alleles. Indeed, if the mutation rate of the zinc finger is high enough such that old alleles mutate before being old enough to decrease in fitness, therefore there is no red queen. There is evidence that hybrids carrying two different alleles of Prdm9 can be infertile (Gregorová and Forejt, 2000; Mihola *et al.*, 2019). This fertility can be rescued by introducing a new Prdm9 allele (Davies *et al.*, 2016; Gregorova *et al.*, 2018). This supports the prediction of purifying selection on heterozygotes. However, no decrease in fertility has been observed in Prdm9-homozygous mice so far for any allele tested. The only evidence of the red-queen hypothesis is the high amino-acid diversity observed at the positions of the zinc fingers that interacts with DNA (Oliver *et al.*, 2009; Thomas *et al.*, 2009; Ponting, 2011; Baker *et al.*, 2017; Damm *et al.*, 2022). This suggests that selection in favour of new zinc fingers may have occurred in the history of Prdm9-carrying animals.

2.2.3 Prdm9 phylogenetic distribution

Prdm9 is found in a wide range of animals such as sponges and mammals (Ponting, 2011), but absent in other kingdoms. It is therefore probable that it arose at the base of the tree of animals. Interestingly, it has been lost many times, notably in the ancestor of nematodes, holometabolous insects (Laurent Duret personal communications), *Neoteleostei* which contain most fishes, *Amphibia* and archosaurs (dinosaurs, pterosaurs and crocodiles) (Cavassim *et al.*, 2022). Finally in mammals, it has been lost in canids (dogs, wolves and foxes), which corresponds to the most recent loss of Prdm9 reported so far (Oliver *et al.*, 2009; Axelsson *et al.*, 2012). Of note,

assessing the absence of Prdm9 in one species is difficult because the absence of the gene in the assembly is not an evidence for its absence in the species (Baker *et al.*, 2017; Cavassim *et al.*, 2022). In turn the presence of Prdm9 does not guarantee that the gene has the same functions as in humans and mice.

2.3 Prdm9-independent recombination hotspots

In birds and canids, which both lost Prdm9, recombination hotspots are found at the subset of gene promoters that contain a CpG island (Auton *et al.*, 2013; Singhal *et al.*, 2015; Kawakami *et al.*, 2017). These sequences are characterized by very low rates of CpG methylation, high CpG and GC content, H3K4Me3 histone marks and open chromatin (Bird, 1980; Cooper *et al.*, 1983; Shin Voo *et al.*, 2000; Klose and Bird, 2006). Interestingly, outside of animals, in plants and fungi, recombination are also associated to gene promoters, H3K4Me3 marks and DNA hypomethylation (reviewed in Lichten and Goldman (1995) and Choi and Henderson (2015)). However, the mechanism that is responsible for the location of hotspots in gene promoter when Prdm9 is absent is still controversial, and might not be the same between species (Borde *et al.*, 2009; Sommermeyer *et al.*, 2013; He *et al.*, 2017; Choi *et al.*, 2018). Contrarily to Prdm9-dependent hotspots, Prdm9-independent hotspots seem to be conserved between closely related species (Axelsson *et al.*, 2012; Singhal *et al.*, 2015; Lam and Keeney, 2015; Schield *et al.*, 2020; Hoge *et al.*, 2023). A Knock-out experiment of the gene Prdm9 in mice and rats also leads to the redirection of DSB hotspots in gene promoters that exhibit hypomethetylated cytosines and H3K4Me3 histone marks (Brick *et al.*, 2012; Mihola *et al.*, 2021). This strongly suggest that Prdm9-independent hotspots are the "default" recombination hotspots shared and conserved in Eukaryotes and that Prdm9 hijacks the recombination machinery away from those default hotspots (Brick *et al.*, 2012).

2.4 Hotspot detection

There are several methods to detect recombination hotspots at different temporal scale, but in this section, I will mainly focus on two of the most precise methods to detect recombination hotspots, which allowed to infer the positions of recombination hotspots of mice and dogs used in chapter 6.

2.4.1 Linkage disequilibrium

Let us consider a locus with an ancestral allele A and a derived allele a located on the same chromosome as another locus with an ancestral allele B and a derived allele b. If the mutation $B \mapsto b$ appears on the genotype containing allele A, without recombination, b will always be co-transmitted with A. Therefore, the haplotype ab will never be observed. We say that b is genetically linked to A. Now, if there is very large number of recombination events between A and b, and the population is not structured, we say that the two loci are independent. The frequency of haplotypes Ab and ab will directly be the product of the frequency of allele A times the frequency of allele b and the frequency of allele a times the frequency of allele b respectively. In between these two extreme scenarios lies a gradient of linkage disequilibrium (LD) which depends on the local recombination rate. By examining the amount of linkage observed between pairs of bi-allelic polymorphisms in natural populations, one can infer the distribution of the sex-averaged recombination rate in the history of the population. By using sliding windows, one can detect short sequences that display a recombination rate significantly higher than the regional background: recombination hotspots (Auton and McVean, 2007; Chan *et al.*, 2012; Spence and Song, 2019).

These methods offer a great spatial resolution because both the density of markers between which one can infer a recombination rate and the number of recombination events that occurred between these markers are high. However, this method suffers several drawbacks. The first type of drawback is due to the model's simplifying assumptions of constant population size, no selection, no population structure and no gene flow. (Auton and McVean, 2007; Chan *et al.*, 2012; Spence and Song, 2019; Samuk and Noor, 2022). The second type is methodological, for instance, two regions that are in close physical distance in the genome assembly but in different chromosomes in reality (e.g. because of poor assembly quality) will show unusually low levels of linkage, which will be interpreted as a high recombination rate. Overall, for many reasons, these methods can produce a high level of false positive (see Raynaud *et al.* (2023) for details). Moreover, once loci are totally independent, these methods struggle to make a difference between high recombination rate and very high recombination rate (Chan *et al.*, 2012). In species for which the background recombination rate is already very high relative to the mutation rate, LD-based methods are underpowered to detect hotspots (Raynaud *et al.*, 2023).

2.4.2 DMC1 ChIP-Seq

Another category of method to map recombination hotspots consists in isolating sequences that are physically associated to proteins involved in meiotic recombination. The most efficient method to my knowledge consists in marking DMC1, a protein that coats the single stranded DNA posterior to DSB, with an antibody in male meiotic cells, then isolate the complex DMC1-antibody using Chromatin-ImunoPrecipitation and finally sequence the single-stranded DNA that is associated to this complex (Smagulova *et al.*, 2011; Brick *et al.*, 2012). This procedure allows a very precise mapping of the hotspots. This method detects hotpots at the individual level, therefore, it allows to capture variation in hotspot location between individuals. Naturally, this approach is not sensitive to the violation of the assumptions of the LD approach regarding demography, selection or gene flow. However, this approach requires a large number of meiotic cell to have enough statistical power to detect hotspots (Smagulova *et al.*, 2011; Brick *et al.*, 2012). In mammals it can only be applied to males, and there are no guarantee that recombination hotspots inferred by ChIP-seq are used in females. While recombination hotspots detected with LD approaches mainly correspond to CO hotspots, the hotspots identified by ChIP-seq of DMC1 correspond DSB hotspots which can be repaired either as a CO, NCO, or with the sister chromatid, leaving no trace of the DSB in offsprings.

3

GC-biased gene conversion

3.1 Basic principle

During a recombination event, at the precise location where the two parental single stranded DNA hybridise (the so-called heteroduplex), if the two parents had different alleles (i.e. the individual is heterozygous), it will lead to a non Watson and Crick pairing that can be detected as a mismatch by the DNA repair machinery (Winkler, 1930; Roman, 1985). In many eukaryotes, when this mismatch involves an A/T allele paired to a G/C allele, it is repaired more often as a G paired with a C than as an A paired to a T (discovered by Brown and Jiricny (1987), reviewed in Duret and Galtier (2009) and illustrated in Figure 3.1). This phenomenon is therefore called GC-biased

gene conversion (gBGC).

This mechanism increases the transmission of GC alleles to the gametes. Thus, the recombination and DNA repair machineries act as segregation distorters (Nagylaki, 1983; Bengtsson, 1986; Hurst, 2019). Without gBGC, the probability for a GC allele to be transmitted to the gametes is 0.5 (following Mendel's law). When there is gBGC, this becomes $0.5 + b$. Where b is called the gBGC coefficient (Nagylaki, 1983). This gBGC coefficient depends on two parameters. The first one is the probability for a given heterozygous site to be involved in a recombination event which is r, the recombination rate per base pair, times l, the length of the conversion tract in base pair (Duret and Galtier, 2009). The second one is the repair bias towards GC (b_0). If a mismatch has a 70% chance of being repaired towards GC, then $b_0 = 2(0.7 - 0.5) = 0.4$ (Nagylaki, 1983). This deviation from the Mendelian segregation ratio can be quantified at different scales.

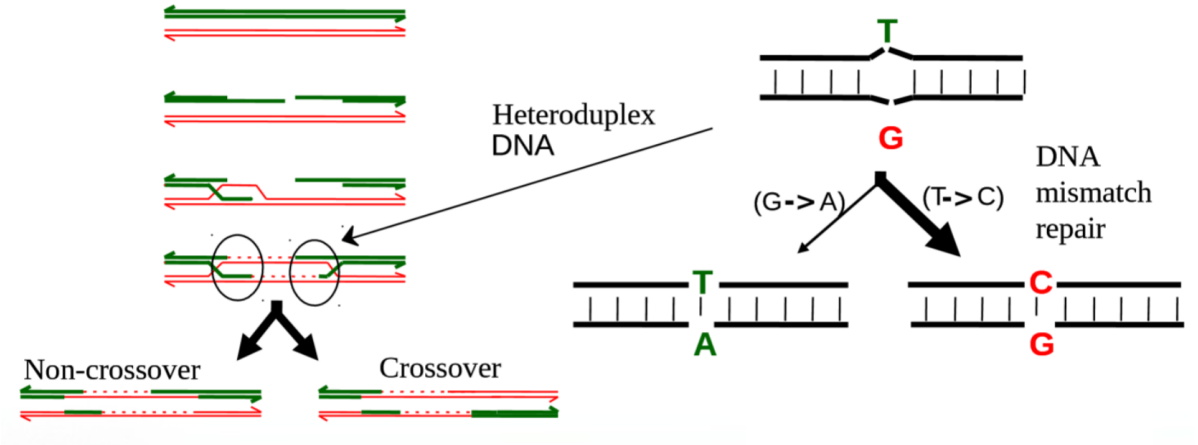


Figure 3.1: *Schematic representation of the mechanism of GC-biased gene conversion in meiosis. © Laurent Duret*

3.2 Direct quantification from meiotic products

The most direct quantification of the repair bias measures the deviation from the 50:50 ratio in meiotic products. By sequencing at least three generations of individuals, one can detect COs and non-COs respectively as a global and local change in ancestry along the chromosome of the last generation. Inside the conversion tracts, one can directly measure how often the GC allele is transmitted compared to the AT one. This approach successfully detected a GC-bias in gene conversion events in humans (Williams *et al.*, 2015; Halldorsson *et al.*, 2016), mice (Li *et al.*, 2019), yeasts (Mancera *et al.*, 2008; Lesecque *et al.*, 2013), and flycatchers (birds) (Smeds *et al.*, 2016). This kind of measure provides the most direct evidence that gene conversion is biased towards GC. However,

when b_0 is small, it requires the sequencing of many individuals to obtain sufficient statistical power. But in turn very small biases can be due to methodological artifacts such as mapping or sequencing errors (Liu *et al.*, 2018; Hurst, 2019).

3.3 Indirect quantification from genetic variation

Instead of focusing on direct observation of transmission bias in meiotic product, population geneticists have derived expectations regarding how the transmission bias towards GC affects changes in allelic frequency and the probability of fixation of mutants in populations.

3.3.1 The Wright-Fisher model with gBGC

The Wright-Fisher model describes the evolution of the frequency of two alleles (here W and S) at a given locus, in a panmictic populations of infinite size with non-overlapping generations. The inclusion of gBGC in the Wright-fisher model that will be presented in this section has been derived by Nagylaki (1983). The frequency of the derived allele S is noted x, and thus the frequency of W is $1 - x$. We further assume that this mutation from W to S is neutral (i.e. a change from W to S does not affect fitness). Without gBGC, at the Hardy Weinberg equilibrium (under panmixia), the frequency of each genotype can be written as follows:

$$\begin{cases} f(SS) = x^2 \\ f(WS) = 2x(1-x) \\ f(WW) = (1-x)^2 \end{cases} \tag{3.1}$$

3.3.2 Change in allelic frequency

Let us assume that W is a A or T allele and S is a G or C allele. Under gBGC, let us consider how the frequency of S changes at the next generation. Homozygotes SS will produce only gametes carrying S. Heterozygotes WS will produce $(1 + b)/2$ gametes carrying S, and $(1 - b)/2$ gametes carrying W. Finally, homozygotes WW won't produce any gamete carriyng S. If we neglect mutations from S to W between two generations, we can thus write:

$$x' = x^2 + 2x(1-x)(1+b)/2 \tag{3.2}$$

which simplifies to

$$x' = x^2 + x(1-x)(1+b) \tag{3.3}$$

We can now compute the expected change in allelic frequency from one generation to another $\Delta x = x' - x$ by simplifying the previous equation to:

$$\Delta x = bx(1-x) \tag{3.4}$$

In practice, population size is finite, and the evolution of allelic frequency is best described as a stochastic process. Let us consider a population of size N, with gametes being drawn randomly from the previous generation. In this setting, from a given generation t, the probability to obtain i gametes carrying the S allele at generation $t+1$ follows a Binomial law with parameters x' and N.

Under the assumption that the conversion bias is weak ($b << 1$), the expected change of allelic frequency is the same as equation 3.4. And thus:

$$\mathbb{E}[\Delta x] = bx(1-x) \tag{3.5}$$

The variance in the change of allelic frequency is given by:

$$\mathbb{V}[\Delta x] = \frac{x'(1-x')}{2N} \tag{3.6}$$

Again, under the assumption of a weak conversion bias, and thus small changes in allelic frequency between two generations, this is well approximated by

$$\mathbb{V}[\Delta x] \simeq \frac{x(1-x)}{2N} \tag{3.7}$$

3.3.3 Fixation probability

When the population size is very large, and the conversion bias very weak, we can approximate the discrete changes in allelic frequencies by a continuous diffusion process. Using Kolmogorov backward equations, we can thus compute the fixation probability of a mutant allele starting in the population at frequency $1/2N$ (Nagylaki, 1983):

$$\mathbb{P}_{fix} = \frac{1 - e^{-2b}}{1 - e^{-4Nb}} \tag{3.8}$$

Conversely, if the mutant allele is A/T and the ancestral allele is G/C, the fixation probability of the mutant allele becomes

$$\mathbb{P}_{fix} = \frac{1 - e^{2b}}{1 - e^{4Nb}} \tag{3.9}$$

In practice, populations are not panmictic, such that not all individual have the same probability of contributing to the next generation. We can define an effective population size N_e such that the number of derived alleles in the next generation can be approximated by a binomial sampling among N_e individuals from the previous generation. This does not affect the expected change in allelic frequency, but the variance changes to $\frac{x(1-x)}{2N_e}$, with N_e being usually smaller that N.

The fixation probability of an A/T to G/C mutation then becomes:

$$\mathbb{P}_{fix} = \frac{1 - e^{-2bN_e/N}}{1 - e^{-4N_e b}} \tag{3.10}$$

And the fixation probability of a G/C to A/T mutation then becomes:

$$\mathbb{P}_{fix} = \frac{1 - e^{2bN_e/N}}{1 - e^{4N_e b}} \tag{3.11}$$

3.3.4 Quantification of gBGC from substitutions patterns

Let us consider again a locus with two state W and S, W corresponding to an A/T allele and S to a G/C allele. In an origination-fixation framework, the substitution rate per

generation can be defined as the product between the mutation rate and the probability of fixation of new mutations (Kimura, 1962; Nagylaki, 1983).

$$Q_{W\mapsto S} = 2N\mu_{W\mapsto S}\frac{1-e^{-2bN_e/N}}{1-e^{-4N_eb}} \tag{3.12}$$

With $2N\mu_{W\mapsto S}$ the number of new mutations per generation in a diploid population of size N. Under the assumption that the conversion bias is weak, this simplifies to

$$Q_{W\mapsto S} = 2N\mu_{W\mapsto S}\frac{2bN_e/N}{1-e^{-4N_eb}} \tag{3.13}$$

which gives:

$$Q_{W\mapsto S} = \mu_{W\mapsto S}\frac{4N_eb}{1-e^{-4N_eb}} \tag{3.14}$$

From this equation it appears clearly that the substitution rate depends on the product of the effective population size N_e, and the transmission bias b. We thus define a population-scaled gBGC coefficient which writes $B = 4N_eb$. The substitution rate from $W \mapsto S$ can therefore be written:

$$Q_{W\mapsto S} = \mu_{W\mapsto S}\frac{B}{1-e^{-B}} \tag{3.15}$$

Similarly, the substitution rate from $S \mapsto W$ can be written:

$$Q_{S\mapsto W} = \mu_{S\mapsto W}\frac{B}{e^{B}-1} \tag{3.16}$$

From these theoretical expectations, one can fit a substitution model to alignments of DNA between different species to estimate the magnitude of the population-scaled gBGC coefficient (B). This approach has been applied to mammals, with estimations being made either in a Maximum Likelihood (Galtier, 2021) or a Bayesian framework (Lartillot, 2013).

These empirical studies revealed an ample variation of the population-scaled gBGC coefficient (B) in mammals (Lartillot, 2013; Galtier, 2021). Of note, the

population-scaled gBGC coefficient correlates positively both with other estimates of effective population size, and with estimates of genome-wide recombination rates, which is expected given that $B = 4b_0 l\mathbf{N_e r}$.

However, from this indirect measure only, it is not possible evaluate the magnitude of the repair bias as in direct measures from meiotic products. The intensity of gBGC always appears as a product between the repair bias, the recombination rate and the effective population size.

3.3.5 Quantification of gBGC from the site frequency spectrum

Similarly to the fixation probability, for a given WS polymorphic site we can derive the probability of observing a derived W or S allele at frequency x (Muyle *et al.*, 2011):

$$H_S(x) = \frac{1 - e^{B(1-x)}}{(1 - e^{-B})(1 - x)} \tag{3.17}$$

And

$$H_W(x) = \frac{1 - e^{-B(1-x)}}{(1 - e^{B})(1 - x)} \tag{3.18}$$

In a given population, the number of polymorphism observed at each frequency represents the site frequency spectrum (SFS). Instead of focusing on fixed changes, it is thus possible to fit the model to a SFS (Muyle *et al.*, 2011; Glémin *et al.*, 2015). This approach has been applied to a wide range of animals and several plants (Muyle *et al.*, 2011; Glémin *et al.*, 2015; Clément *et al.*, 2017; Galtier *et al.*, 2018; Barton and Zeng, 2021; Boman *et al.*, 2021). Interestingly, neither in animals nor in angiosperms has the relationship between B and N_e been recovered (Clément *et al.*, 2017; Galtier *et al.*, 2018; Boman *et al.*, 2021). Despite a wide range of variation of N_e, only a small range of variation of B is observed. The potential reasons for this small range of variation are discussed in chapter 11 and 12.

3.4 A widespread phenomenon in eukaryotes

Even without a precise quantification, it is still possible to assess qualitatively whether gene conversion is biased towards GC in different species. A convincing observation is an intragenomic correlation between recombination rate and GC content (Meunier and Duret, 2004; Webster *et al.*, 2006; Duret and Arndt, 2008). When the recombination landscape is not available, one can still use the premise that shorter chromosome experience higher recombination rates, and therefore should be more GC rich (Pessia *et al.*, 2012; Figuet *et al.*, 2015). This qualitative assessment suggests that gBGC is present in all major clades of eukaryotes (Pessia *et al.*, 2012), and even suggests that it is present in bacteria (Lassalle *et al.*, 2015).

3.5 The molecular mechanism of gBGC

Despite great efforts to quantify gBGC and unravel its impact on genome evolution, the molecular mechanism of gBGC is still poorly understood. In particular, the proteins that are responsible for the GC bias in heteroduplexes are still unknown. In somatic cells, in the presence of mismatches, two repair pathway can intervene, depending on the nature of the mismatches.

3.5.1 DNA repair pathways

The short patch repair pathway is associated with base excision repair (BER), the primary function of which is to correct DNA lesions resulting from oxidation, deamination or alkylation (reviewed in Krokan and Bjørås (2013)). These lesions cause slight distortions in the three-dimensional structure of DNA, which are recognised and repaired by specific enzymes called DNA glycosylases. These glycosylases cut one of the bases of the mismatch, leaving an abasic site (a single-stranded lesion or nick), which is repaired by DNA polymerases and ligated by ligases. These BER pathway therefore acts on a short stretch of DNA, affecting only the excised base, or sometimes a few base pair around it.

The second pathway corrects mismatches that affect larger stretches of DNA and is mostly represented by the Mismatch repair (MMR) (reviewed in Spies and Fishel (2015)). Mismatches in the double helix, caused by errors in DNA polymerisation during replication or spontaneous lesions, are detected by the MMR, which erases one of the two strands and re-synthesises it.

3.5.2 The role of BER in gBGC

Interestingly, among the DNA glycosilases that belong to the BER pathway, some can excise thymine and adenine but none can excise guanine or cytosine (Krokan and Bjørås, 2013). Moreover, it has been demonstrated in bacteria that the absence of the BER pathway is strongly associated to low GC content (Teng *et al.*, 2022). Finally, in mice, it appears that only mismatches that are alone in the heteroduplex are repaired in a biased way (Li *et al.*, 2019). This is consistent with the action of an enzyme that targets single mismatches.

In yeasts, the mechanism seems different, with only those mismatches found at the extremities of the conversion tract being repaired in a biased way (Lesecque *et al.*, 2013). This has been interpreted as a biased action of the MMR (Lesecque *et al.*, 2013), but some evidence suggest that even in the case of multiple mismatches, the erase of the DNA can be initiated by the BER (Grin and Ishchenko, 2016). One can imagine that if the strand that will be erased corresponds to the one that contains the bases that can be excised by the BER (A or T), therefore the first base is repaired in a biased way, using the G or C as a template while the other mismatches are not. Overall, the action of DNA repair enzymes strongly suggest that the BER could be responsible for gBGC, but no direct evidence has been provided yet.

3.6 The Achilles' heel of genomes

By many aspects, gBGC is very different from other kinds of meiotic drive. Most segregation distorters only bias the transmission of a specific locus or at maximum of a chromosome (Lyttle, 1991; Taylor and Ingvarsson, 2003; Lindholm *et al.*, 2016). On the other hand, the transmission bias produced by gBGC affects the whole genome. Moreover, it is the only form of transmission bias (natural selection included) that never generates hitchhiking of neighbor loci. Indeed, gBGC promotes the transmission of GC alleles at the resolution of a single base-pair (Lesecque *et al.*, 2013; Halldorsson *et al.*, 2016; Li *et al.*, 2019).

Most segregation distorters observed in nature are thought to be deleterious (Lyttle, 1991). The main reason for that is probably that segregation distorters that provide a fitness advantage are fixed very quickly, and are rarely observed at the heterozygous state. On the other hand, segregation distorters that are deleterious struggle to fix, and can be maintained at intermediate frequency for a very long time. Overall, segregation distorters that we can observe are a biased sample of those that might exist. Whether

gBGC is advantageous or deleterious is still not clear, and has been the focus of several empirical and theoretical studies (Bengtsson, 1986, 1990; Berglund *et al.*, 2009; Galtier *et al.*, 2009; Glémin, 2010; Necşulea *et al.*, 2011).

3.6.1 Deleterious or advantageous?

Biased gene conversion promotes the fixation of GC alleles no matter their effect on fitness. In section 3.3.4, I explained how gBGC increases the substitution rate from Weak to Strong (AT to GC) and decreases the substitution rate from Strong to Weak (GC to AT), but I assumed that these alleles were selectively neutral. When the Strong allele is deleterious, under the assumption of weak bias and weak selection ($(b - hs) << 1$), the expected change in allele frequency becomes:

$$\mathbb{E}[\Delta x] = (b - hs)x(1 - x) + sx(1 - x)(1 - 2h) \tag{3.19}$$

where s is the difference in fitness between allele Strong and allele Weak, and h the dominance coefficient: the fitness of WW is 1, the fitness of SS is $1 - s$, and the fitness of WS is $1 - hs$ (Nagylaki, 1983; Glémin, 2010).

When $h = 0.5$ (i.e. the fitness of the heterozygote is perfectly intermediate between the two homozygotes), the expected change in allele frequency simplifies to:

$$\mathbb{E}[\Delta x] = (b - \frac{s}{2})x(1 - x) \tag{3.20}$$

The variance of the change in allelic frequency ($\mathbb{V}[\Delta x]$) remains however unchanged.

With selection, and in the co-dominant scenario ($h = 0.5$), the fixation probability of a deleterious Strong derived allele, derived by Nagylaki (1983) is given by:

$$\mathbb{P}_{fix} = \frac{1 - e^{(-2b+s)N_e/N}}{1 - e^{-2N_e(2b-s)}} \tag{3.21}$$

For a simpler writing of the equations, in the co-dominant case, the fitness of the heterozygote is usually noted $1 - s$ and the homozygote $1 - 2s$. This allows us to re-write equation 3.21:

$$\mathbb{P}_{fix} = \frac{1 - e^{-2(b-s)N_e/N}}{1 - e^{-4N_e(b-s)}} \tag{3.22}$$

This way it is clearer that the fate of the new mutation will depend on the difference of its transmission bias induced by gBGC b, and the transmission bias induced by selection s.

Even before the discovery of gBGC, it had been predicted that if gene conversion could be biased towards a class of mutations that is more numerous, it should reduce the mutation load (Bengtsson, 1986). However, if the bias is too strong, it can substantially increase it (Bengtsson, 1990; Glémin, 2010). Overall, it is not clear whether gBGC is more deleterious than advantageous at the intensity observed in living species, and it will be the object of chapter 11.

3.6.2 Empirical evidences

For now, empirical evidence tend to point toward a deleterious effect of gBGC. However, the number of studies that evaluate the effect of gBGC on the genetic load are scarce and are mainly centered on human. Galtier *et al.* (2009) and Berglund *et al.* (2009) showed that bursts of gBGC episodes could lead to the rapid fixation of GC alleles in human exons. Based on the conservation of these exons in other mammals, these substitutions are likely to be deleterious. However those fixation events correspond to the most extreme levels of gBGC, inside recombination hotspots. Necşulea *et al.* (2011) and Lachance and Tishkoff (2014) showed that gBGC maintains disease-causing polymorphism at high frequency in human populations, particularly in regions of high recombination rate, unravelling a significant impact of gBGC on public health. However, diseases that are observed in humans are likely to belong to a specific part of the distribution of fitness effects: those mutations that are deleterious enough to be noticed, but not deleterious enough to be lethal. Some of them might also affect health but not fitness, for instance mutations that have a negative impact on health only at old ages. Overall, a complete analysis of the impact of gBGC on fitness both across the genome, and across fitness effects is still missing.

4

Inferring the strength, sign and direction of selection in protein-coding genes

Detecting the footprints of past selection in genomes is a challenging task. When a new mutation occurs in a population that brings a selective advantage/disadvantage to its bearer, it has several characteristics that distinguish it from a neutral one.

First, if a new mutation is beneficial, it will rise in frequency more often than if it was neutral in a population, and be observed at higher frequency on average. Second, beneficial mutations will fix in the population more often than neutral ones, and therefore will have a higher substitution rate. Finally, when reaching fixation, beneficial mutations will bring their neighbour with them, creating a signal of local depletion of diversity (Smith and Haigh, 1973). On the other hand, if a mutation is deleterious, it will segregate at lower frequency, be fixed less often and will also create a local decrease in diversity (Charlesworth *et al.*, 1993) whose pattern differs from that of beneficial mutations (Elyashiv *et al.*, 2016).

To be able to detect signatures of selection from the genome, one needs to have a neutral expectation, such that the departure from this expectation can be interpreted

as an effect of natural selection. To detect selection from intragenomic variations in diversity, population geneticists usually use the rest of the genome, or the regional value of diversity as a neutral expectation without selection. In protein coding genes, the redundancy of the genetic code provides mutations that do not affect the sequence of a protein (synonymous mutations). By considering that these mutations are selectively neutral, one can therefore compare the frequency, substitution rate and diversity depletion between non-synonymous mutations (that affect the sequence of the protein) and synonymous mutations (Miyata and Yasunaga, 1980; Nei and Gojobori, 1986).

As natural selection is the object of a vast majority of evolutionary studies, a very large number of methods have been designed to detect their hallmarks in protein-coding genes. In this chapter, I will give a brief introduction of the methods that have been used and discussed in this thesis.

4.1 Evolutionary rates

The dN/dS ratio, first introduced by Miyata and Yasunaga (1980), measures the substitution rate of non-synonymous mutations over synonymous mutations. In an origination fixation model, where we consider that the substitution rate is the product of the mutation rate and the probability of fixation, the dN can be written:

$$dN = 2N_c\mu_N\mathbb{P}^N_{fix} \tag{4.1}$$

where N_c is the census population size, μ_N the mutation rate of non-synonymous mutations and $\mathbb{P}^N_{fix}$ is the fixation probability of the derived amino-acid (Spielman and Wilke, 2015). Assuming that the fitness difference between the heterozygote for the derived and ancestral allele and the homozygote for the ancestral allele is s and that the mutation is co-dominant:

$$\mathbb{P}^N_{fix} = \frac{1 - e^{-2sN_e/N_c}}{1 - e^{-4N_e s}} \tag{4.2}$$

where N_e is the effective population size (see section 3.3.3 for a definition in this context). In turn, considering that synonymous changes are neutral, we can write:

$$dS = \mu_S N_c / N_e \tag{4.3}$$

Therefore the dN/dS ratio can be written:

$$dN/dS = 2N_e \frac{\mu_N}{\mu_S} \mathbb{P}^N_{fix} \tag{4.4}$$

The standard neutrality test is to compare the dN/dS ratio to one. If new non-synonymous mutations in a gene are deleterious, dN/dS should be inferior to one and if they are beneficial, the dN/dS ratio should be superior to one, and if perfectly neutral it should be equal to one. From equation 4.4, it appears clearly that this is only true if we assume that the mutation rate of synonymous mutations is equal to the mutation rate of non-synonymous ones. In practice, it is not the case as more mutations are non-synonymous compared to synonymous. Therefore, for the dN/dS to be a correct estimator of the selection exerted on proteins, we need to multiply the raw dN/dS ratio by a correction factor equal to μ_S/μ_N. Moreover in practice, most new mutations in protein-coding genes are deleterious. Therefore, when comparing the average dN/dS between different sites and/or different genes, the dN/dS ratio is most often inferior to one even if some beneficial mutations also occur. The test of $dN/dS > 1$ is therefore underpowered to detect positive selection (Latrille *et al.*, 2023a). Finally, when comparing the dN/dS ratio between two genes, one cannot know if one is higher because it experiences more beneficial mutations or more deleterious ones.

4.2 SFS-based methods

At the population level, the strength of selection in protein-coding genes can be inferred from the SFS. By assuming a certain distribution of fitness effects (DFE) of new mutations, one can derive the expected SFS. By fitting this SFS to empirical ones, one can therefore estimate the parameters of the DFE. This approach has first been developed by Eyre-Walker *et al.* (2006) considering only deleterious and neutral mutations. It was then improved to correct for the distortion induced by the demography (Eyre-Walker and Keightley, 2009) using the SFS of synonymous mutations, and to account for weakly advantageous mutations and errors when inferring the ancestral states of mutations (Galtier, 2016; Tataru *et al.*, 2017).

4.3 Phylogenetic conservation score

Both dN/dS and SFS-based methods are able to capture the strength and sometimes the sign of selection, but they are unable to estimate the direction of selection, meaning they cannot tell at a given site which amino-acids are favoured by natural selection and which ones are disfavoured. This inference can however be made by looking at the occurrence of a given amino-acid at a given site in a multi-species protein alignment. Indeed, if at a given position of a gene, one amino-acid is present in all eukaryotes, this is a strong indication that this amino-acid is more fit than the others. By looking at the frequency of an amino-acid at large time scales, we can infer the preference of selection for any given amino-acid. This hint gave rise to methods that attribute a score for each given amino-acid at each site of a protein based on its frequency and some prior about the biochemical properties of proteins (e.g. Kumar *et al.* (2009)). This score is supposed to reflect fitness.

While dN/dS across sites can reflect selection on a specific branch and SFS-based methods infer selection at the population level, conservation scores are supposed to detect selective constraints at a broader phylogenetic scale. However, phylogenetic scores do not use neutral expectation for the frequency of different amino-acids in an alignment. Therefore, what they measure is any phenomenon that can favour an amino-acid over another, which can be the number of codons it is coded by, mutation biases, or gBGC. In addition, it assumes that all sequences are independent realisation of the same process, and do not account for phylogenetic inertia.

4.4 Mutation-selection models

Just like phylogenetic conservation scores, mutation-selection models infer fitnesses at the level of phylogenies, leveraging the signal of site-wise amino-acid frequencies. However, mutation-selection models are rooted in population genetic theory and offer a flexible framework to disentangle the relative contributions of mutation, selection and drift to molecular evolution. It also takes into account the phylogenetic relationships between species.

The mutation-selection formalism uses a dN/dS-like approach to disentangle mutation from selection by assuming that synonymous mutations are selectively neutral (Halpern and Bruno, 1998). The substitution rate from codon i to codon j at a given site l ($q^{(l)}_{i \mapsto j}$) can be written:

$$\begin{cases} q^{(l)}_{i\mapsto j} & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ q^{(l)}_{i\mapsto j} & = \mu_{i\mapsto j} \text{ if } i \mapsto j \text{ is synonymous} \\ q^{(l)}_{i\mapsto j} & = \mu_{i\mapsto j} \dfrac{4N_e(f^{(l)}_i - f^{(l)}_j)}{1 - e^{4N_e(f^{(l)}_i - f^{(l)}_j)}} \text{ if } i \mapsto j \text{ is non-synonymous} \end{cases} \tag{4.5}$$

where $\mu_{i\mapsto j}$ is the mutation rate from codon $i \mapsto j$ and $f^{(l)}_i$ the fitness of codon i.

By fitting this model to a protein coding-gene alignment and a gene tree, one can infer the mutation matrix per gene, and the fitness of every amino-acid per site. Mutation-selection models allow us to have access to the fitness landscape of a given clade. However they are still limited by several simplifying assumptions: population size is usually assumed constant (but see Latrille *et al.* (2021)), the fitness landscape is site specific (no epistasis), and the fitness landscape is stable (no adaptation).

Therefore, mutation-selection models represent a null model without epistasis and adaptation while accounting for nearly neutral processes on a fixed fitness landscape (Latrille *et al.*, 2023a).

5

Thesis objectives

The evolution of genomes results from a balance between different evolutionary forces such as mutation, natural selection, drift, recombination, migration etc... The fields of molecular evolution and population genetics seek to investigate how those forces affect genome evolution and affect each other to form the patterns of genetic diversity present in living organisms. Mutation, recombination and migration most often create genetic variation, while drift and selection erode this variation. In the 80's, a new potential recombination-associated evolutionary force was acknowledged: biased gene conversion (BGC) (Nagylaki, 1983; Bengtsson, 1986, 1990; Bengtsson and Uyenoyama, 1990). BGC constitute an exception to Mendel's law of inheritance, by distorting the segregation ratio of heterozygous sites during a recombination-induced gene conversion event. Around the same time, experimental evidence for it was found in the form of GC-biased gene conversion (gBGC) (Brown and Jiricny, 1987, 1988). Since then, many researchers have shown that this phenomenon has a profound effect on base composition (Meunier and Duret, 2004; Webster *et al.*, 2006; Duret and Arndt, 2008), rate of evolution (Berglund *et al.*, 2009; Galtier *et al.*, 2009; Ratnakumar *et al.*, 2010; Bolívar *et al.*, 2016, 2019) and fitness (Necşulea *et al.*, 2011; Lachance and Tishkoff, 2014), confirming its legitimacy at the level of evolutionary force. When acknowledging gBGC, the distribution of recombination events along the genome appears to be crucial, as recombination hotspots will likely strongly affect both local base composition and selection dynamics (Lartillot, 2013; Glémin *et al.*, 2015; Galtier,

2021). However, as the discovery of gBGC is relatively recent, our understanding of its influence on evolution is still in its infancy.

5.1 The evolution of fine-scale recombination landscapes

Fine-scale recombination landscapes have been generated for very few species, and their diversity is still poorly understood. In the first part of this thesis, we leverage the effect of gBGC on base composition to gain insight into the determinants and the dynamics of recombination hotspots in placental mammals. We develop a method to measure relative recombination rates along terminal branches based on neutral substitution patterns. We apply this method to gain insights into the determinants of the position and the dynamics of recombination hotspots in the genome of 52 placental mammals. I then discuss more generally the evolutionary origins and reasons for the existence of recombination hotspots in eukaryotes in the light of their impact on both the success of meiosis and on the genetic load they imply.

5.2 The role of beneficial back-mutations in molecular evolution

If the discovery of gBGC is relatively recent, natural selection has been described much earlier (Darwin, 1859), and in fact, biologists did not wait for the discovery of genomes to unravel its impact on the evolution of populations and species. Despite this, there is still considerable disputes and controversies about its contribution to genome evolution. Most of these disputes stem from misunderstandings on the definitions of neutrality, nearly-neutrality, different forms of selection and adaptation, which vary across sub-fields of evolutionary biology, and the philosophical perception of linked selection. In the second part of this manuscript, we empirically show how we could gain in interpretive capability on the processes affecting genome evolution by distinguishing positive selection from adaptive evolution. I then show that this distinction is fundamental if we are to understand the joint action of natural selection, recombination and gBGC on genome evolution. Finally I try to (re)establish a conceptual framework that will hopefully help us to distinguish non-adaptive from adaptive evolution in genomes, to improve our understanding of the interplay between molecular and ecological adaptation.

5.3 The evolutionary origin of GC-biased gene conversion

Even if a lot of attention has been drawn to the consequences of gBGC on genome evolution, much less has been given to the evolution of gBGC itself. In the last part of this manuscript, using a modeling approach, we estimated the strength of selection that acts on a modifier of the intensity of gBGC to identify the factors that may influence its evolution. Finally, I discuss how the molecular mechanisms of gBGC might explain the variation of its intensity in eukaryotes, without involving natural selection.

Part II

The evolution of fine-scale recombination landscapes

6

High prevalence of Prdm9-independent recombination hotspots in placental mammals

Context

As presented in the introduction, intragenomic variations of gBGC intensity often reflect variation of recombination rate (Pessia *et al.*, 2012; Auton *et al.*, 2013; Clément and Arndt, 2013; Lartillot, 2013; Munch *et al.*, 2014; Glémin *et al.*, 2015; Figuet *et al.*, 2015; Bolívar *et al.*, 2016; Charlesworth *et al.*, 2020). We therefore used substitution patterns to explore the diversity of fine-scale recombination landscape in mammals, and in particular the use of stable Prdm9-independent hotspots. To this day, they always have been presented as inactive in mammals, solely based on what has been observed in few species (mostly humans and mice) (Brick *et al.*, 2012; de Massy, 2013; Smagulova *et al.*, 2016; Schield *et al.*, 2020; Hoge *et al.*, 2023). However, in a clade comprising no less that $\sim$ 6,000 species (Burgin *et al.*, 2018), one primate and one rodent is not what one would call a very dense phylogenetic sampling. I must confess that our first intention was not to gain further insight into the use of Prdm9-independent recombination hotspots, as I too had blind faith in mice and humans. In the middle of my last year of PhD, for a different project, the friends that co-author this manuscript and I found ourselves assessing the stability of recombination hotspots in canids. As a sanity check, we looked at signatures of gBGC in regions orthologous to dog hotspots in the southern elephant seal. Since seals have Prdm9, they should not display any gBGC activity in loci orthologous to the dog hotspots. What a surprise to see that instead, equilibrium GC content was through the roof. I then decided to use signatures of gBGC to estimate recombination activity in Prdm9-independent recombination hotspots in 52 species of boreoeutherian mammals. The idea was to check if the southern elephant seal was one very curious animal, or if human and mice were misleading us since the beginning. The title speaks for itself.

Detailed contributions

I designed the study, selected the species, developed the code for calling substitutions in peer-coding with Djivan and developed the method to quantify relative recombination rates from substitution patterns. Alexandre generated the whole genome alignments for 19 carnivore species. Théo annotated protein-coding genes for all 19 species and generated a phylogeny (not included in the paper). Laurent annotated the Prdm9 allele present in the reference genome of all 52 mammals of this study, and investigated signatures of positive selection on the zinc finger of two of them. All authors helped with ideas and discussions. I wrote the first draft of the paper, which was then intensively revised by all co-authors. This manuscript has not been submitted to any journal yet.

High prevalence of Prdm9-independent recombination hotspots in placental mammals

J. Joseph[1], **D. Prentout**[2], **A. Laverré**[3,4], **T. Tricou**[1], **L. Duret**[1]

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

[2]Department of Biological Sciences, Columbia University, New York, NY 10027, USA

[3]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[4]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

julien.joseph@ens-lyon.fr

October 16, 2023

Abstract

In many mammals, recombination events are concentrated into hotspots directed by a sequence specific DNA-binding protein named Prdm9. This protein facilitates chromosome pairing and its inactivation has been shown to induce fertility losses in mice and rats. Intriguingly, *Prdm9* has been lost several times in vertebrates, and notably among mammals, it has been pseudogenized in the ancestor of canids (dogs, wolves foxes). When this gene is inactive, either naturally in dogs, or through knock-out experiments in mice, recombination hotspots still exist, but they tend to occur in promoter-like features such as CpG islands. It has thus been proposed that one role of *Prdm9* could be to direct recombination away from those Prdm9-independent hotspots. However, the ability of Prdm9 to direct recombination hotspots has been assessed only in a handful of species, and a clear picture of how much recombination occurs outside of Prdm9-directed hotspots in mammals is still lacking. In this study, we derived an estimator of past recombination activity based on signatures of GC-biased gene conversion in substitution patterns. We applied it to quantify recombination activity in Prdm9-independent hotspots in 52 species of boreoeutherian mammals. We observed a wide range of recombination rate at these loci: several species (such as mice, humans, some felids or cetaceans) show a deficit of recombination, while a majority of mammals display a clear peak of recombination. Our results demonstrate that Prdm9-directed and Prdm9-independent hotspots can co-exist in mammals, and that their co-existence seem to be the rule rather than an exception.

***K*eywords** Prdm9 · Recombination landscape · Hotspots · Mammals · gBGC

Introduction

Meiotic recombination is a crucial step in the production of gametes for a large majority of eukaryotes. It is initiated by programmed double-strand breaks (DSBs), that can be resolved in two types of recombination events, using the homolog as a template: crossovers (COs), where there is a reciprocal exchange of chromosome arms, and non crossovers (NCOs), where DSB repair leads to gene conversion around the DSB site without reciprocal exchange. In most eukaryotes, at least one CO per chromosome is necessary to ensure correct segregation between homologs and is therefore mandatory for meiosis success (Page and Hawley, 2003; Gerton and Hawley, 2005). In many vertebrates, recombination events are not uniformly distributed along the genome (reviewed in Stapley et al. (2017); Zelkowski et al. (2019)). Instead, they tend to be concentrated in so-called recombination hotspots (Lichten and Goldman, 1995; Tock and Henderson, 2018).

In many mammals, the position of recombination hotspots is determined by the zinc-finger protein Prdm9, which binds specific DNA motifs and recruits the DSB machinery through histone methylation (Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010; Diagouraga et al., 2018). This gene is highly polymorphic and hundreds of alleles have been reported in mice and humans (Buard et al., 2014; Kono et al., 2014; Alleva et al., 2021). Most allelic diversity is concentrated on residues of the zinc fingers that interact with the DNA, leading to changes in DNA sequence specificity. Therefore, the position of recombination hotspots varies within a population and between species (Auton et al., 2012; Smagulova et al., 2016; Alleva et al., 2021). Additionally, Prdm9 tends to erode its targets through gene conversion (Baker et al., 2015; Smagulova et al., 2016). As the available targets for a given Prdm9 allele become scarce, its ability to generate enough COs for meiosis to succeed is compromised. A new allele with more targets will then be positively selected leading to a red-queen dynamic accelerating the turnover of recombination hotspots (Úbeda and Wilkins, 2011; Latrille et al., 2017; Baker et al., 2022; Genestier et al., 2023). It has also been proposed that when Prdm9 binds symmetrically both homologs, it facilitates the repair of DSBs as COs and thereby contributes to the success of meiosis (Davies et al., 2016; Li et al., 2019; Hinch et al., 2019). Indeed, experiments in mice and rats showed that an inactivation of this protein drastically reduces fertility (Mihola et al., 2019, 2021; Brick et al., 2012).

Despite its central role in recombination, *Prdm9* has been repeatedly lost in vertebrates. (Baker et al., 2017; Cavassim et al., 2022). Among amniotes, one loss occurred in the ancestor of archosaurs (crocodiles and birds) (Singhal et al., 2015; Cavassim et al., 2022), and another one in the ancestor of canids (dogs, wolves and foxes) (Oliver et al., 2009; Axelsson et al., 2012; Auton et al., 2013). In dogs and several passerines, studies based on linkage disequilibrium (LD) showed that recombination hotspots tend to occur in CpG islands, and more specifically in hypomethylated ones in dogs (Auton et al., 2013; Berglund et al., 2015; Singhal et al., 2015; Kawakami et al., 2017). Likewise, the DSB hotspots of a mouse whose *Prdm9* has been inactivated through a knock-out ($PRDM9^{-/-}$) also occur in promoter-like features (Brick et al., 2012). Interestingly, an increase of recombination rate near promoters has also been observed in plants and yeasts, which lack *Prdm9* (Petes, 2001; Marand et al., 2017).

Those observations led to the conclusion that there exist two types of recombination landscapes in vertebrates. The first one is Prdm9-dependent, fast evolving with recombination targeted away from functional

elements, while the second is relatively stable, with recombination occurring in promoter-like feature such as CpG islands. However, a recent finding in snakes challenged this binary view. In rattle snakes and corn snakes, which still possess a functional *Prdm9*, some hotspots are directed by Prdm9, but others are directed toward CpG islands (Schield et al., 2020; Hoge et al., 2023). It was first proposed that Prdm9 could have a different role in rattle snake and direct recombination towards CpG islands (Schield et al., 2020), but other observations suggest that the recombination landscapes in these snakes rather reflects the inefficiency of their Prdm9-dependent pathway to direct DSBs away from CpG islands (Hoge et al., 2023). Altogether, the two studies concur on the fact that the recombination landscape in snakes differs from what is observed in mammals, despite the presence of *Prdm9* (Schield et al., 2020; Hoge et al., 2023).

Most of our knowledge about *Prdm9* function and evolution has been acquired in a handful of mammals (mostly human and mice). In mice, Smagulova and colleagues analyzed the position of DSB hotspots in strains carrying different *Prdm9* alleles (Smagulova et al., 2016). Those strains showed differences in their capacity to target DSBs away from the hotspots of the Prdm9$^{-/-}$ mouse (hereafter referred to as 'MDH', for 'Mouse Default Hotspots'). Some strains show a significant deficit of DSB hotspots at MDH loci, while others, carrying less dominant Prdm9 alleles, show up to a 6-fold DSB hotspot enrichment at these loci, even though they represent a small proportion of all recombination hotspots ($\sim$7%) (Smagulova et al., 2016). This shows that even when *Prdm9* is present, Prdm9-independent hotspots can be active in mammals. However, this activity could just be the reflection of a specific Prdm9 deficiency in some mice strains and overall, we still have no clear idea on how prevalent the usage of Prdm9-independent hotspots is in mammals. A comprehensive understanding would require measures of fine-scale variations in recombination rate for a wide range of mammalian species, and ideally across long periods of time.

Recombination hotspots can be mapped directly in meiotic cells (e.g. by chromatin immunoprecipitation with antibodies to DMC1 (Brick et al., 2012; Pratto et al., 2014; Smagulova et al., 2016; Alleva et al., 2021)), but these molecular approaches are tedious and only amenable for a few model organisms. High-resolution recombination maps can also be inferred from patterns of linkage disequilibrium (LD). This approach is more scalable and provides information on sex-averaged historical recombination activity at the population scale. However, this approach remains laborious and expensive (it requires the sequencing of at least 10 individuals per species (Auton and McVean, 2007; Chan et al., 2012)), and it is sensitive to various sources of errors (Spence and Song, 2019; Samuk and Noor, 2022; Raynaud et al., 2023). Hence, for now, such LD-based recombination maps are available only for a very limited number of species.

Alternatively, substitution patterns have been found to be informative about past recombination rates. In particular, it has been shown in mammals that recombination induces a transmission bias of GC alleles through the process of GC-biased gene conversion (gBGC). This eventually leads to an elevation of the WS substitution rate (AT to GC), and a decrease of the SW substitution rate (GC to AT) (Nagylaki, 1983; Duret and Arndt, 2008; Glémin, 2010). The substitution rate matrix can be conveniently summarized by a single parameter, the equilibrium GC-content (hereafter noted GC^*), which corresponds to the GC-content that sequences would reach if the pattern of substitution observed in that branch remained constant over time (Duret and Arndt, 2008). GC^* correlates well with the strength of DSB hotspots in mice (Clément and

Arndt, 2013), with the LD-based recombination rates in humans (Munch et al., 2014; Glémin et al., 2015), and with the LD-based strength of recombination hotspots in dogs (Axelsson et al., 2012; Auton et al., 2013). Moreover, GC^* reflects the recombination activity along the entire branch where substitution patterns are analyzed, and hence can inform about past recombination events that are no longer detectable with methods measuring recombination in individuals or populations (Lesecque et al., 2014; Munch et al., 2014). This provides insights on the long term use of Prdm9-independent hotspots, integrated over long periods of time, probably encompassing the rise and fall of several *Prdm9* alleles. However, variation in GC^* between species cannot be directly interpreted as variations in recombination rates alone, since GC^* also depends on the mutation bias towards AT, the repair bias towards GC, the effective population size and the mean length of the conversion tracts (Eyre-Walker, 1999; Glémin, 2010).

In this study, we present an estimator of relative recombination rates based on substitution patterns that allows us to directly compare fine-scale recombination rate variations in a wide range of species using only 3 genomes (one focal genome, a sister species and an outgroup). We then use it to assess the recombination activity at MDH loci in 52 species spanning the diversity of boreoeutherians. We reveal a high heterogeneity in the use of these Prdm9-independent hotspots. We show that *Prdm9* alleles in humans and mice have been particularly efficient at directing DSBs away from Prdm9-independent hotspots but that these two species are not representative of all mammals. Finally, we show that three species, namely the southern elephant seal, the ring-tailed lemur and the daurian ground squirrel, have used Prdm9-independent hotspots as much as Prdm9-deficient canids. This shows that the two kinds of hotspots-regulation mechanisms that have been described so far in vertebrates are not mutually exclusive and that the fine-scale recombination landscapes of many mammals are much closer to those of birds and other Prdm9-lacking amniotes than previously thought. We further show that the recombination activity observed at MDH loci in Prdm9-containing mammals depends on the conservation of their DNA methylation pattern, which suggests a link between the evolution of DNA methylation and of Prdm9-independent recombination landscapes.

Results

Conservation of recombination hotspots between Prdm9-deficient mammals

In finches and flycatchers, the fine-scale recombination landscape has been shown to be stable through time, as a large proportion of hotspots are shared between closely related species (Singhal et al., 2015; Kawakami et al., 2017). To test whether loci corresponding to Prdm9-independent hotspots are also evolutionary stable in mammals, we analyzed the overlap between recombination hotspots detected in dogs, which naturally lack Prdm9, and those identified in the Prdm9$^{-/-}$ mutant mouse (MDH). Among the 30,929 MDH, 15,009 (49%) could be assigned to one-to-one orthologous loci in the dog genome. Among the 7008 dog hotspots identified in the LD-based recombination map of dogs (Auton et al., 2013), 34% overlap with MDH loci (Suppfig. S1)(compared to 0.06% expected by chance, given that dog hotspots and MDH loci cover respectively 2% and 3% of the dog genome). Although this enrichment is very strong, it should be noted that 42% of the dog hotspots that could be mapped on the mouse genome occur outside of MDH loci (1809/3109) (Suppfig. S1). This number is difficult to interpret because it has been shown that LD-based methods can produce

a high number of false positives (Raynaud et al., 2023). Moreover, DSB hotspots are obtained on males only, while LD maps are sex-averaged. It is also possible that some MDH loci have not been identified as hotspots in dogs simply because they did not meet the threshold criteria to be defined as such. To avoid the problem of the arbitrary threshold, we computed the LD-based recombination rate in dogs (from Auton et al. (2013)) as a function of the distance to the closest MDH loci (Fig. 1A&B) (Auton et al., 2013). We divided the 30,929 mouse Prdm9$^{-/-}$ hotspots into three equally sized categories of strength, based on DMC1 ChIP-seq read counts (Smagulova et al., 2016). Respectively 5,266 strong MDH, 4,961 medium MDH and 4,781 weak MDH could be mapped on the dog genome. There is a sharp peak of recombination centered on MDH loci in dogs. This peak is higher for strong MDH and weaker for medium and weak ones (Fig. 1A). This confirms that many recombination hotspots are conserved between Prdm9$^{-/-}$ mice and dogs, but it does not rule out the existence of species-specific recombination hotspots. To explore factors that might drive the evolution of Prdm9-independent hotspots, we analyzed their DNA methylation level in the germline. Indeed, indirect evidences suggested that recombination hotspots in dogs are associated to germline hypomethylated regions (HMRs) (Berglund et al., 2015). Using HMRs identified by bi-sulfite sequencing in dog sperm (Qu et al., 2018), we observed that 74% of dog hotspots are located inside HMRs, which represent only 3.7% of the dog genome. In mice, the overlap is even stronger: using HMRs identified in mouse sperm (Hammoud et al., 2014), we observed that out of the 30,929 hotspots found in the Prdm9$^{-/-}$ mutant, 93% are located within HMRs, which cover only 4.6% of the mouse genome (see methods for details). This indicates that Prdm9-independent hotspots are associated with DNA hypomethylation both in Prdm9$^{-/-}$ mutant mice and in canids. Interestingly, 48% of MDH loci are methylated in dog sperm if we restrict the definition of hotspots to their midpoint (7,186/15,009). This shows that many MDH loci are specifically hypomethylated in mice but not in dogs. This is consistent with previous observations showing that murid genomes have accumulated many new HMRs compared to other mammals (Qu et al., 2018). To test whether these shifts in methylation levels are associated with changes in recombination activity, we computed the LD-based recombination rate in dogs as a function of the distance to the closest MDH locus, separating those whose midpoints overlap a HMR in dogs (7,186), and those that do not (7,283)(Fig. 1B). There is a high and pronounced recombination peak at MDH loci that are hypomethylated in dog sperm (Fig. 1B). In contrast there is almost no elevation of recombination at MDH loci that are methylated in dogs (Fig. 1B). This confirms that methylation is clearly associated to recombination hotspots in the absence of Prdm9, and that many Prdm9-independent recombination hotspots are species-specific.

Equilibrium GC content is a good predictor of past recombination activity

To assess whether equilibrium GC content could be used as a proxy of recombination rate in other species, we repeated the above analyses using GC^* instead of LD-based recombination rate (Fig. 1C&D). Strikingly, the profile of GC^* perfectly mirrors the LD-based recombination rate profile. It should however be noticed that the GC^* peaks are slightly sharper than the LD-based ones (Fig. 1C&D), as are the peaks of DMC1 ChIP-seq read coverage in the Prdm9$^{-/-}$ mouse (Suppfig. S2). This suggests that GC^* is able to capture signals of past recombination with a higher spatial resolution than LD.

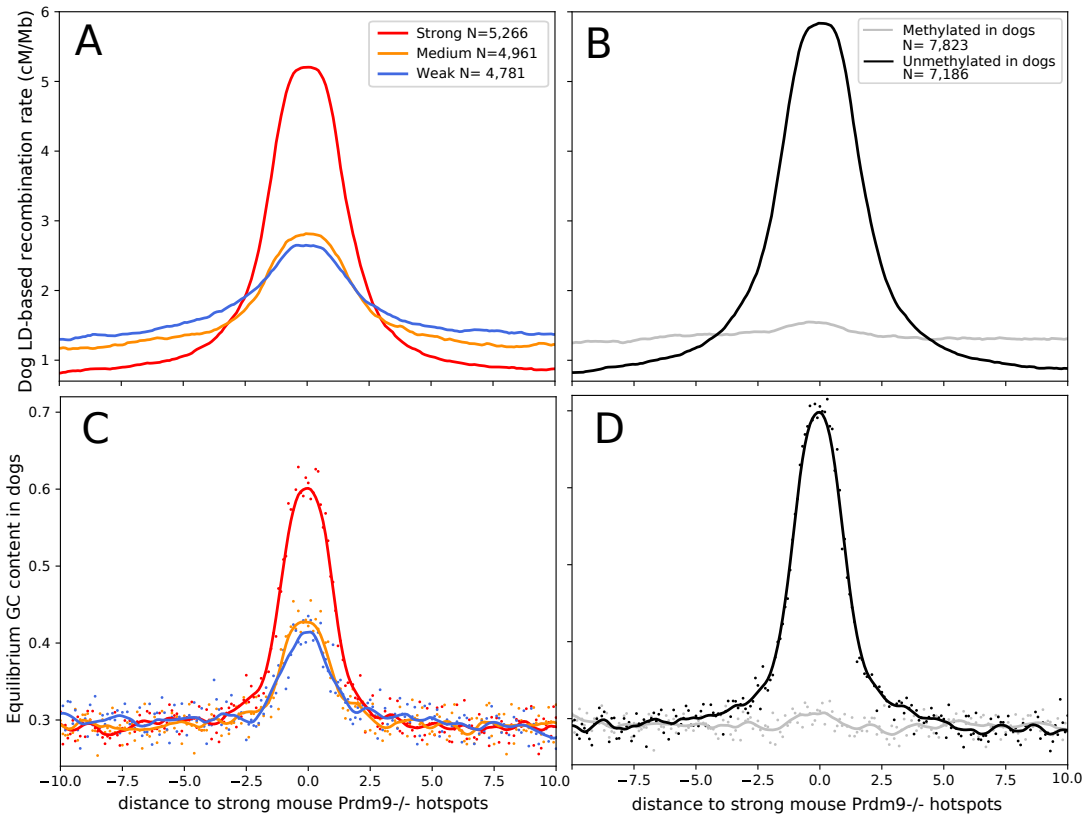


Figure 1: A) Dog LD-based recombination rate as a function of the distance to the closest MDH loci. MDH were divided in three equally sized categories of strength: strong hotspots in red (10-190 FPKM), medium hotspots in orange (5-10 FPKM) and weak hotspots in blue (0-5 FPKM). The line directly correspond to the mean value of LD-based recombination rate in a 100 bp window. B) Same as A but MDH loci were divided in two categories depending on their methylation level in dog sperm. MDH loci that are found outside hypomethylated regions in dogs are in grey and those found inside hypomethylated regions are in black. C) & D) Equilibrium GC content in dogs as a function of the distance to the closest MDH loci. Using the same partitions of hotspots as A) and B). Points correspond to the mean value of GC^* in a 100 bp window. The line correspond to a smoothing of the data with a loess function.

Estimation of the relative recombination rate in Prdm9-independent hotspots

Following a large body of literature, we showed that GC^* can be very informative on intra-genomic recombination rate variations (Pessia et al., 2012; Auton et al., 2013; Clément and Arndt, 2013; Lartillot, 2013; Munch et al., 2014; Glémin et al., 2015; Singhal et al., 2015; Figuet et al., 2015; Bolívar et al., 2016; Galtier et al., 2018; Charlesworth et al., 2020). However, the height of the GC^* peak in hotspots is difficult to interpret in term of recombination rate because it is also affected by other parameters (the length of gene conversion tracts, the mutation bias towards AT, the mismatch repair bias towards GC and the effective population size), which can vary between species (Lartillot, 2013; Galtier et al., 2018; Galtier, 2021). We thus derived an estimator that can capture the relative recombination rate at Prdm9-independent hotspots

that controls for those parameters and is therefore comparable between species. Using the probability of fixation of AT and GC alleles in presence of gBGC derived by Nagylaki (1983), we obtained an expression of the ratio of the recombination rate within hotspots relative to their flanking regions (see details in the Methods). This relative recombination rate only depends on GC^* inside hotspots and in flanking regions, and on the mutation bias (GC^μ).

$$\frac{r_{hot}}{r_{flank}} = \frac{logit(GC^*_{hot}) - logit(GC^\mu)}{logit(GC^*_{flank}) - logit(GC^\mu)} \quad (1)$$

where r_{hot} is the recombination rate in hotspots, r_{flank} the recombination rate in flanking regions and GC^μ the GC content expected under mutation only. It should be noted that r_{hot} and r_{flank} encompass all recombination events that can lead to gBGC (potentially COs and/or NCOs).

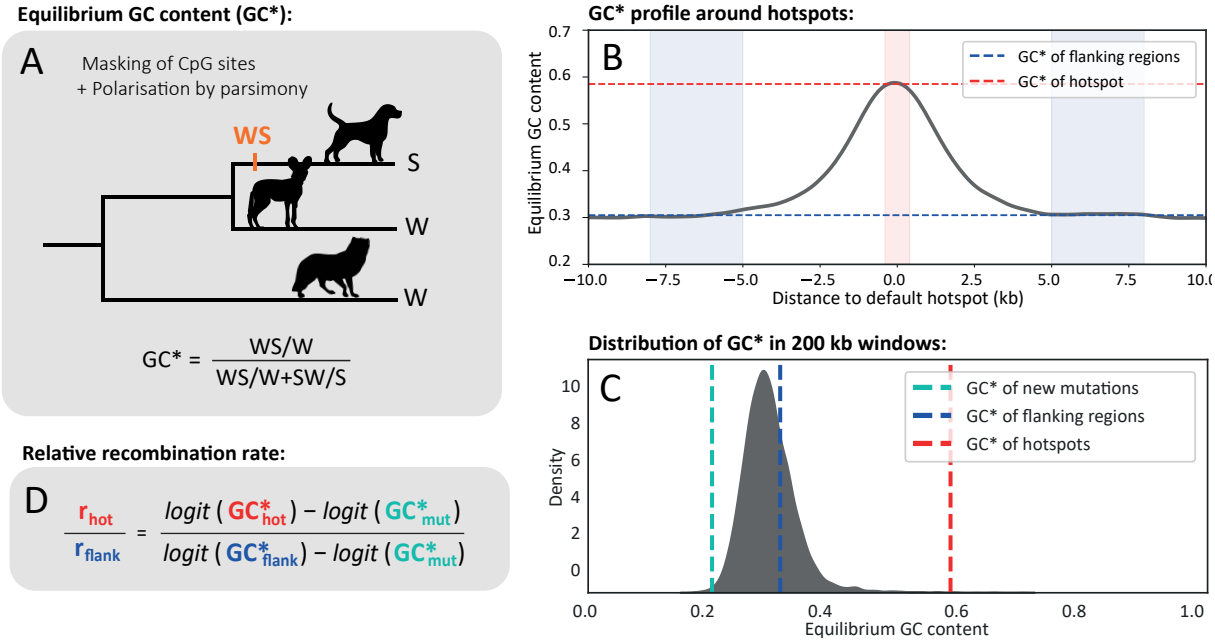


Figure 2: Overview of the method for inferring relative recombination rates in Prdm9-independent hotspots. A) We call substitutions using parsimony on trios of closely related species after having masked CpG dinucleotides. B) We compute GC^* in 400 bp windows centered on the midpoint of the Prdm9-independent hotspots, and GC^* in the flanking regions (from 5 to 8 kb upstream and downstream of the center of the Prdm9-independent hotspot. C) We compute the distribution of GC^* in 200 kb windows and take the 1^{st} percentile as the GC^* of new mutations. D) Using the probability of fixation of AT and GC alleles in presence of gBGC derived by Nagylaki (1983), we compute the relative recombination rate as a function of the three values of GC^* (see methods).

For the rest of the study, hotspots are defined as the 400 bp regions centered on their midpoint, and flanking regions as those spanning from 5 to 8 kb upstream and downstream of the hotspots (Fig. 2A).

Using the three values GC^*_{hot}, GC^*_{flank} and GC^{μ}, it is possible to compute a measure of the relative recombination rate within hotspots compared to their flanking regions, which can then be compared between species. Equation 1 holds true under the assumption that the other four parameters affecting the strength of gBGC (effective population size, mismatch repair bias, length of the conversion tract and the mutation bias) do not differ between the hotspots and their flanking regions (see Methods). As shown in the previous section, Prdm9-independent hotspots are often hypomethylated. This implies that the mutation rate from CpG to TpG or CpA in hotspots is lower than in the flanking regions, which violates our assumption of a constant mutation bias. To avoid this problem, we excluded CpG sites from all the analyses.

GC^*_{hot} and GC^*_{flank} can easily be computed from the substitutions in Prdm9-independent hotspots and their flanking region but GC^{μ} is more difficult to estimate. Interestingly, it has been shown that genome-wide variations of GC^* are mainly the result of gBGC, and that GC^{μ} is quite constant along chromosomes, despite large-scale variation in mutation rates (Smith et al., 2018). An approach to estimate GC^{μ} consists in measuring variation in substitution patterns along the genome, and to consider the regions of the genome with the lowest GC^* as a proxy for GC^{μ} (Lartillot, 2013). Following this logic, we divided the genome of each species in windows of 200kb, and defined GC^{μ} as the value of the first percentile of the distribution of GC^* to avoid outliers (Fig. 2B) (see methods for detailed justifications). This method allows us to estimate GC^{μ} for a wide range of species, simply based on substitution patterns in the terminal branch.

Prdm9-independent recombination hotspots are active in most mammals

Using this estimator of relative recombination rate, we assessed whether MDH loci showed an enrichment of recombination in other mammals. We identified MDH orthologous loci in the genome of 51 other mammals and estimated the relative recombination rates at these loci using the method described above (gBGC-based relative recombination rates). Around 75% of the species (39/52) show a significant enrichment of recombination in MDH loci compared to flanking regions (Fig. 3). In 77% of those species (30/39), the recombination activity at MDH loci is conserved, with strong MDH loci showing a significantly higher recombination enrichment than weak ones (Fig. 3). The remaining 23% of species (9/39) show a lower recombination activity in Prdm9-independent hotspots. Therefore, we might be lacking statistical power to confirm a conservation of hotspot strength in those species (Fig. 3). Interestingly, in 10% of species (5/52), including mice, there is less recombination in strong Prdm9-independent hotspots compared to weak ones, which is consistent with the active deviation of recombination away from those sites observed in those mice having the most dominant Prdm9 alleles (Smagulova et al., 2016). The causes for this active deviation are however not clear (Smagulova et al., 2016).

To test whether as in dogs, methylation plays a role in determining hotspots in other mammals, we separated MDH loci in two subsets: loci for which the hypomethylation pattern is conserved between mice and dogs (and thus likely to be consistently hypomethylated in other mammals as shown in Qu et al. (2018)), and loci that are hypomethylated in mouse but not dogs (a majority of which corresponding to mouse-specific HMRs (Qu et al., 2018)). For MDH loci whose hypomethylation pattern is not conserved, we observed a very weak increase in recombination in all species (Suppfig. S4).

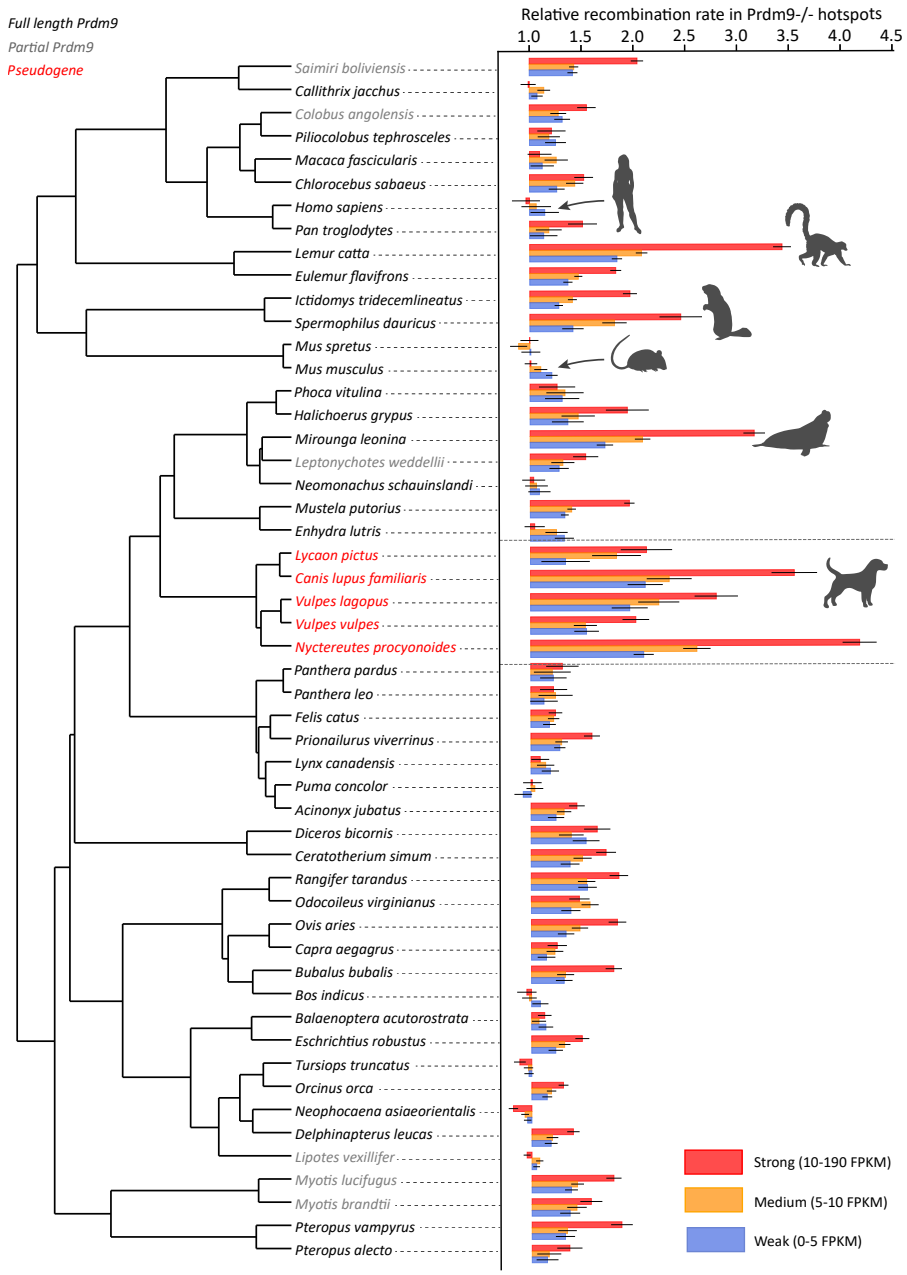


Figure 3: Relative recombination rates at at loci orthologous to mouse Prdm9$^{-/-}$ DSB hotspots (MDH loci) in 52 mammals. MDH loci were binned in 3 equally sized categories of strength based on the number of DMC1 Chip-seq reads of Prdm9$^{-/-}$ DSB hotspots. The number of MDH loci for each category varied from ∼4,000 in *Myotis brandtii* to ∼9,000 in *Mus spretus* (see details in Supplementary Table 1). The tree has been retrieved from TimeTree5 (Kumar et al., 2022). Species with a complete Prdm9 are written in black, in grey species for which we failed to find a complete Prdm9 in the reference genome assembly, and in red the 5 canids (where Prdm9 is a pseudogene). Error bars correspond to a 95% confidence interval obtained by bootstrapping the substitutions for computing GC^*_{flank} and GC^*_{hot}. We considered the recombination enrichment to be significant if the confidence interval is above one.

Conversely, MDH loci with a conserved hypomethylation pattern, show contrasting levels of recombination activity: 25% of species (13/52, including mice and humans) show a deficit of recombination at these loci, 12% (6/52) show no elevation of recombination, whereas 63% (33/52) show a strong recombination activity. This indicates that DNA hypomethylation is associated with a deficit of recombination in some species (such as mice and humans), while it is associated with recombination hotspot in others. This latter group includes the 5 canids (which all lack Prdm9), but also many other mammals with an intact Prdm9. This shows that DNA hypomethylation can be associated with recombination hotspots even in the presence of Prdm9.

To get further insight into the evolution of Prdm9-independent hotspots in mammals, we measured the relative activity at loci orthologous to dog LD-based recombination hotspots (DRH for 'Dog Recombination Hotspots') in the 51 other mammals. We observed a strong correlation between the recombination activity at MDH loci and DRH loci (Suppfig S5), which is expected since they largely overlap. Nevertheless, species that are phylogenetically closer to dogs show higher recombination activity in DRH loci whereas those phylogenetically closer to mouse show higher recombination activity in MDH loci (Suppfig S5). This confirms that despite a general conservation, Prdm9-independent hotspots are still evolving in mammals.

Canids, which all lack Prdm9, dominate the list of species that exhibit the highest recombination levels at MDH loci, holding top ranks out of 52 (Fig. 4A). This finding both shows that the recombination landscape is stable in canids as it is in passerines, and validates our relative recombination rate estimator (Fig. 4A). Interestingly, there are several other mammals that show a similar enrichment of recombination activity at MDH loci. Notably, three species show a recombination activity at MDH loci significantly higher than some of the canids: ring-tailed lemurs (*Lemur catta*), southern elephant seals (*Mirounga leonina*) and daurian ground squirrels (*Spermophilus dauricus*) (Fig. 4A).

It should be noted that these three species encode a full-length Prdm9, encompassing the four protein domains (KRAB, SSXRD, SET and the zinc finger array). The Prdm9 allele represented in the reference genome assembly contains 11 zinc fingers in daurian ground squirrels, and 14 zinc fingers in ring-tailed lemurs. For the southern elephant seal, there is only one complete zinc finger represented in the reference genome because of an assembly gap within the array. In ring-tailed lemurs and daurian ground squirrels, the comparison of zinc finger sequences showed an excess of amino-acid changes relative to the neutral expectation, particularly at positions -1, 3, and 6 that are involved in DNA binding, (SuppFig. S6).

This suggests that Prdm9 has been subject to positive selection in these two lineages, which is suggestive of the red-queen dynamic that is expected when Prdm9 determines the location of recombination hotspots. Thus, Prdm9 does not show any sign of pseudogenization or functional change in these three species where Prdm9-independent hotspots appear to have been particularly active.

As an illustration of our substitution-based approach, we plotted GC^* as a function of the distance to the closest strong MDH loci in dogs, humans, mice and the three outlier species (Fig. 4B). We observed no elevation of GC^* in humans and mice in strong MDH loci, confirming that very little recombination occurred in their lineage (Fig. 4B). In dogs, ring-tailed lemurs, southern elephant seals and daurian ground squirrels, there is a pronounced peak of GC^* at strong MDH loci (Fig. 4B).

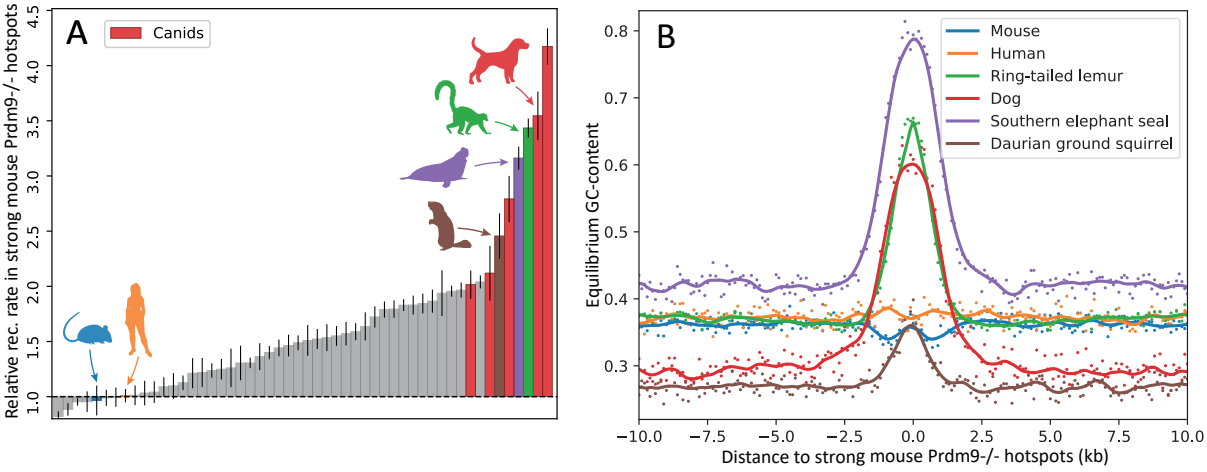


Figure 4: (A): Sorted relative recombination rate in strong DSB hotspots (>10 FPKM) of a Prdm9 -/- mouse in 52 mammals. Error bars correspond to a 95% confidence interval obtained by bootstrapping the substitutions for computing GC^*_{flank} and GC^*_{hot}. (B): GC^* as a function of the distance to the center of the closest MDH locus for humans, mice, dogs, and the three outlier species. Each point corresponds to an estimation of GC^* in a 100 bp window.

Discussion

The current paradigm is that vertebrates either possess a full-length functional Prdm9, recombine away from promoter-like features, and display a fast evolving recombination landscape, or they lack a functional Prdm9 in which case they consistently recombine in CpG islands (e.g. in canids, birds or the swordtail fish) (Baker et al., 2017). Our results revealed that there is a continuum between these two types of recombination landscapes and that despite the presence of Prdm9, some species use Prdm9-independent hotspots as much as canids. We showed that the activity of those Prdm9-independent hotspots is highly dependent on DNA hypomethylation. This implies that despite a general conservation, Prdm9-independent recombination hotspots are evolving slowly, in concert with germline DNA hypomethylation (Qu et al., 2018; Berglund et al., 2015).

From a methodological perspective, while signatures of gBGC have been commonly used in studies of recombination landscapes (Axelsson et al., 2012; Auton et al., 2013; Munch et al., 2014; Lesecque et al., 2014; Singhal et al., 2015; Charlesworth et al., 2020; Hoge et al., 2023), the approach presented here allows to quantify gBGC-based relative recombination rates along a branch that are comparable between species. The analysis of gBGC signatures only requires the genome of three closely related species, and is able to detect recombination activity at a given set of loci with a very high spatial resolution, even better than LD-based methods (compare Fig. 2D and 2B). It thus offers the possibility for large-scale comparative studies of fine-scale recombination landscapes. However, this approach requires a large number of substitutions to estimate GC^* precisely and is therefore not appropriate to measure recombination at a single locus.

Moreover, it should be noted that our estimation of recombination activity using gBGC does not allow one to conclude on the nature of recombination events (COs or NCOs) that we detect at Prdm9-independent hotspots. In humans, there is evidence that both COs and NCOs induce gBGC but in mice, only NCOs appear to do so (Williams et al., 2015; Arbeithuber et al., 2015; Halldorsson et al., 2016; Li et al., 2019). This suggests that the type of recombination events triggering gBGC can vary among mammals. Thus, it is possible that while a large number of DSB occur at Prdm9-independent hotspots in our three outlier species, COs still tend to be associated to Prdm9-directed hotspots. Altogether, while the enrichment of recombination events in Prdm9-independent hotspots is very clear in numerous mammals, the way they are repaired remains to be explored.

On the maintenance of a default hotspot regulation mechanism

It had been previously demonstrated that both Prdm9-dependent and Prdm9-independent pathways coexist in two snakes genera (Hoge et al., 2023). The authors suggested that a change in the binding affinity of a gene which operates downstream of Prdm9 could explain this coexistence (Hoge et al., 2023), and that selection may be operating to fine-tune the usage of Prdm9-independent hotspots in vertebrates (Hoge et al., 2023). However, these explanations were given based on observations of four closely related species of colubroids (a subgroup of snakes). Variations observed in a wider range of taxonomic levels is needed to prove the existence of a continuum and to give insight into the determinant of Prdm9-independent hotspot usage in species possessing Prdm9.

In placental mammals we revealed that the coexistence of both Prdm9-dependent and Prdm9-independent pathways to direct DSBs is pervasive. We showed that those pathways determine recombination hotspots with varying proportions across species. Interestingly, these variations do not show a strong phylogenetic structure, suggesting that this evolution can be very rapid. Furthermore, the species with the highest levels of Prdm9-independent hotspot usage have quite contrasted life history traits, which they mostly share with sister species with lower Prdm9-independent hotspot usage. Thus, selective reasons for which their Prdm9-independent hotspot usage is that high are difficult to imagine. Moreover, in mice, even if the usage of Prdm9-independent hotspots is quite low overall, there exist substantial variations that seem to depend only on Prdm9 alleles (Smagulova et al., 2016).

These observations rather suggest an alternative explanation for the variations in the usage of Prdm9-independent hotspots in boreoeutherians. Overall, proper chromosome pairing can be achieved through the two different pathways mentioned above (Prdm9-dependent or Prdm9-independent). When the efficiency of one pathway is altered, better chromosome pairing can be restored either by a mutation restoring the efficiency of the altered pathway, or by a mutation increasing the efficiency of the other pathway. For the Prdm9 pathway, we know that alleles inevitably decrease in efficiency due to the erosion of their high affinity targets, which reduces the probability of symmetrical binding, and thus impairs efficient chromosome pairing (Baker et al., 2015, 2022; Latrille et al., 2017; Genestier et al., 2023). This efficiency can be restored either by a new Prdm9 allele inducing a red-queen dynamic (Úbeda and Wilkins, 2011; Latrille et al., 2017; Baker et al., 2022; Genestier et al., 2023), but also by a mutation increasing the efficiency of the Prdm9-independent pathway. Every mutation increasing the efficiency of the Prdm9-independent pathway lessens the deleterious

effect of a mutation that reduces Prdm9 efficiency. Conversely, new efficient Prdm9 alleles will lessen the deleterious effect of a mutation that decreases the efficiency of the Prdm9-independent pathway.

Of note, if this dynamic reaches a point where the Prdm9-independent pathway becomes sufficient for correct chromosome pairing, *Prdm9* can be lost without strong fitness consequences. Under this model, *Prdm9* is lost through the accumulation of small effect mutations which reduce its utility, rather than a sudden loss that would imply very inefficient selection. The continuum in the use of Prdm9-independent hotspots we observe in mammals could reflect different stages along this path, and despite still having a fully functional *Prdm9*, our three outlier species could be on their way of losing it. It is also possible that their *Prdm9* have only been going through a temporary inefficient phase, which have been compensated by the Prdm9-independent pathway, but has now been rescued by a new efficient Prdm9 allele.

The recombination landscape of amniotes

Overall, our results suggest that in addition to Prdm9-directed hotspots, many mammals share some of their recombination hotspots with other amniotes, (Axelsson et al., 2012; Singhal et al., 2015; Schield et al., 2020; Hoge et al., 2023), and therefore the fine-scale recombination landscapes of mammals, birds and snakes is probably more similar than previously thought (Baker et al., 2017). However, the determinants of Prdm9-independent hotspots usage remain unclear. Interestingly DNA methylation has also been found to be a suppressor of recombination in a fungi hotspot (Maloisel and Rossignol, 1998), in plants (He et al., 2017; Choi et al., 2018), and in honey bees (Wallberg et al., 2015), which suggests that local hypomethylation is a common determinant of recombination hotspots in eukaryotes. However, the association between hypomethylation and recombination has not been formally established in non-mammalian amniotes and remains to be tested. Moreover, this association need not be causal, as the potentially diverse molecular mechanisms of Prdm9-independent recombination in amniotes remain largely unknown. In particular, the results presented here suggest that despite having lost Prdm9, red foxes (*Vulpes vulpes*) and african wild dogs (*Lycaon pictus*) have only a mild recombination enrichment in Prdm9-independent hotspots compared to other canids. It has been previously noted that the number of recombination hotspots varies between Prdm9-deficient amniotes. Notably in finches and flycatchers, only few LD-based hotspots have been reported compared to dogs (Auton et al., 2013; Singhal et al., 2015; Kawakami et al., 2017). Altogether, it is still not clear what drives the concentration of recombination events in absence of Prdm9, and why some species have numerous hotspots and others less.

Finally, the widespread use of Prdm9-independent recombination hotspots demonstrated in the present study is likely to have important consequences for genome evolution. In particular, the fact that gBGC is stronger in hypomethylated regions in numerous mammals and in several passerines provides a convincing explanation for the widespread GC-richness of CpG islands in amniotes.

Material & Methods

Prdm9-independent hotspots and DNA methylation datasets

Data on the location of DSB hotspots in Prdm9$^{-/-}$ knock-out mice , detected by DMC1 ChIP-seq experiment, were retrieved from the study by Smagulova et al. (2016). The dog LD-based recombination map and the position of hotspots were retrieved from the study of Auton et al. (2013). We excluded hotspots that were larger than 20 kb as they do not fit the definition of hotspots. Data sets of hypomethylated regions in mouse and dog sperm (identified using Bi-sulfite sequencing) were retrieved from the literature (Hammoud et al., 2014; Qu et al., 2018). Even though methylation data on spermatocytes would have been more fit for the task at hand, only sperm is available in the literature for dogs.

Whole genome alignments

For most mammals except canids, felids and phocids, whole genome alignments (WGAs) were obtained from Genereux et al. (2020). In order to get further phylogenetic resolution in canids, and in closely related outgroups, we generated a WGA of high quality genomes for 19 carnivores downloaded from NCBI (Supplementary Table 3), using the Progressive Cactus aligner (v1.3.0) (Armstrong et al., 2020). We first defined a "guide" species tree using the topology obtained from TimeTree5 (Kumar et al., 2022). To streamline the computational process, we ran Progressive Cactus separately for canids, felids and phocids species, using different "–root" options on the same guide tree. We created a root alignment by running Progressive Cactus with the inferred ancestral genome of each of the three clades. We obtained the final WGA using the "halAppendSubtree" command to iteratively include the three sub-alignments at the corresponding ancestral nodes (Hickey et al., 2013).

Defining orthologous regions

To find the orthologous regions of the Prdm9-independent hotspots in the genomes of other mammals we used halLiftover (Hickey et al., 2013). We first made a liftover from the mouse/dog genome to the target genome using the midpoint of each feature and removed multi-mapping features. Then we lifted back the single-mapping features from the target genome to the dog/mouse genome and again removed multi-mapping features. This approach ensures that all orthologous loci were one-to-one.

Hotspots overlap

We considered hotspots to be overlapping if their midpoint was at less than 5 kb one from another. This is equivalent to a strict overlap for hotspots defined as 5 kb windows centered on their midpoint. Using this approach, we calculated the percentage of the genome covered by the hotspots by multiplying the number of hotspots by 5,000, and dividing by the assembly size. For the overlap with HMRs we defined hotspots as the 5 kb windows centered on their midpoint and kept the size of HMRs defined in the study of Qu et al. 2018 (Qu et al., 2018). The percentage of the genome covered by HMRs was computed as the sum of all HMR sizes divided by the assembly size.

Substitution mapping

We selected trios of closely related species such that the divergence between the three species was low enough to avoid double substitutions, but with an outgroup distant enough to avoid incomplete lineage sorting based on the guide tree used in the Zoonomia WGA (Genereux et al., 2020). We tried to take trios spanning the diversity of boreoeutherians, avoiding over-sampling of disproportionately represented groups (primates and artiodactyles). A complete list of the trios used are available in Supplementary table 4. *A posteriori*, it appeared that there were substantial variations between the branch lengths used to map substitutions, but we showed that the divergence was still low enough ($< 2.5\%$) and did not influence our result (Suppfig. S7). Genome quality was very variable. We thus controlled that genome quality (approximated by the N50 statistics) did not influence our results (Suppfig. S8). To call substitutions, we retrieved multispecies alignment using hal2maf (Hickey et al., 2013). We excluded alignment blocks which size was inferior to 50 bp to avoid poorly aligned regions, and duplicated regions. We then excluded CpG sites (sites for which at least one of the three species has a CpG) to avoid convergent mutations. Finally, we called substitutions using parsimony as depicted in Fig. 2.

Measures of equilibrium GC content

We compute the equilibrium GC content as follows:

$$GC^* = \frac{WS/W}{WS/W + SW/S} \tag{2}$$

With W the AT content of the region (CpG masked), S the GC content of the region (CpG masked), WS the number of Weak to Strong substitutions and SW the number of Strong to Weak substitutions.

LD-based recombination rate and GC* profiles around hotspots

We cut the genome in windows of 100 bp. We extracted the LD-based recombination rate for each window. We also computed the distance between the midpoint of the window and the closest midpoint of a hotspot using bedtools closest (Quinlan and Hall, 2010). We then made bins of distances to the closest hotspot every 100 bp. For GC^*, we repeated the same procedure, but we computed the counts of WS and SW substitution for each window, and then computed GC^* using the total substitutions count for all windows in each distance bin as described in the previous section.

Estimation of the mutation bias

Germline mutations rates have been measured by sequencing parent–offspring trios in 36 mammalian species (Bergeron et al., 2023). However, for most species, the number of detected mutations is too limited (typically less than 100 de novo mutations) to estimate GC^μ accurately. Thus, measures of GC^μ based on empirical data are associated with very large confidence intervals (Wong et al., 2016; Milholland et al., 2017; Wang et al., 2020). Even when there is enough statistical power, the results vary substantially between different datasets (Wong et al., 2016; Milholland et al., 2017). In addition to the issue of reproducibility, it has been

demonstrated that the mutation spectra can vary rapidly within human populations (Harris and Pritchard, 2017). Therefore, GC^{μ} estimated from living individuals may not necessarily reflect the average GC^{μ} in the terminal branch. We therefore took a similar approach to (Lartillot, 2013). We divided the genomes in windows of 200 kb and took the value of the first percentile of GC^{*} as an estimate of GC^{μ}. To ensure that our results were not sensitive to our estimation of GC^{μ}, we used different thresholds to compute it (Suppfig. S3B&C), and recovered the same results. We also controlled that our gBGC-based relative recombination rates were not correlated to our estimations of GC^{μ} (Suppfig. S3A).

Estimation of relative recombination rates from GC*

Using a Wright-Fischer diffusion approximation and assuming that mutations are selectively neutral, the rate of Weak-to-Strong substitution in a given branch can be written as follows (Nagylaki, 1983)

$$q_{WS} = 2Ne\mu_{WS}\frac{2b}{1-e^{-4Neb}}T \tag{3}$$

where μ_{SW} is the mutation rate per generation from W to S, b the gBGC coefficient, T the divergence time from the ancestral node in generations, and Ne the effective population size.

The gBGC coefficient is directly linked to the recombination rate, with $b = b_0 r l$ where r is the recombination rate per base pair per meiosis, b_0 the repair bias, and l the length of the conversion tract in base pair. It should be noted that r encompasses all recombination events that can lead to gBGC (CO and/or NCO)

Similarly, the rate of Strong-to-Weak substitutions can be written as follows:

$$q_{SW} = 2Ne\mu_{SW}\frac{2b}{e^{4Neb}-1}T \tag{4}$$

The equilibrium GC content can be written as follows:

$$GC^{*} = \frac{q_{WS}}{q_{WS}+q_{SW}} \tag{5}$$

Thus, we can write:

$$\frac{GC^{*}}{1-GC^{*}} = \frac{q_{WS}}{q_{SW}} \tag{6}$$

Simplifying the previous equations we obtain:

$$\frac{GC^{*}}{1-GC^{*}} = \frac{\mu_{WS}}{\mu_{SW}}e^{4Neb} \tag{7}$$

Thus, the population-scaled gBGC coefficient. ($B = 4Neb$) can be written as:

$$B = log(\frac{GC^{*}}{1-GC^{*}}) - log(\frac{\mu_{WS}}{\mu_{SW}}) \tag{8}$$

or:

$$B = logit(GC^*) - logit(GC^\mu) \tag{9}$$

where GC^μ is the equilibrium GC content under the mutational bias only ($GC^\mu = \frac{\mu_{WS}}{\mu_{SW}+\mu_{WS}}$).

Let us note B_{hot} the population-scaled gBGC coefficient within hotspots:

$$B_{hot} = logit(GC^*_{hot}) - logit(GC^\mu_{hot}) \tag{10}$$

And B_{flank} the population-scaled gBGC coefficient in their flanking regions (defined here as 3kb-long segments, located at 5-kb of the hotspot center, Fig. 2A)

$$B_{flank} = logit(GC^*_{flank}) - logit(GC^\mu_{flank}) \tag{11}$$

B depends on b_0, N_e, $l$and r ($B = 4Nerb_0l$). The first three parameters (b_0, N_e, l) are not expected to differ between the hotspot and their flanking regions. Thus, the ratio between the recombination rate within hotspot (r_{hot}) over the recombination rate in their flanking regions (r_{flank}) can be written as:

$$\frac{r_{hot}}{r_{flank}} = \frac{B_{hot}}{B_{flank}} \tag{12}$$

And thus, under the assumption that the mutational bias does not differ between hotposts and their flanking regions (i.e. $GC^\mu_{hot} = GC^\mu_{flank} = GC^\mu$)

$$\frac{r_{hot}}{r_{flank}} = \frac{logit(GC^*_{hot}) - logit(GC^\mu)}{logit(GC^*_{flank}) - logit(GC^\mu)} \tag{13}$$

Annotation of Prdm9 in mammals

We investigated the presence of Prdm9 homologs in each of the 52 species analyzed. The full-length Prdm9 isoform encompasses four domains (KRAB, SSXRD, SET and the zinc finger array). It is encoded by 10 exons (corresponding to exons 2 to 11 of human Prdm9, exon 1 being within the 5'UTR): exons 3 and 4 encode the KRAB domain, exon 7 encodes the SSXRD domain, exons 8-10 encode the SET domain, and exon 11 encodes the zinc finger array. We first searched for Prdm9 homologs by sequence similarity (Camacho et al., 2009) against mammalian proteins annotated in RefSeq (https://www.ncbi.nlm.nih.gov/blast/), using the human protein (NP_001363829.1) as a query. We performed a multiple alignment of the strongest hits, to assess their completeness: homologs where considered as complete if they encompassed the 10 protein-coding exons, from the start codon, up to the beginning of the zinc finger array. By this approach, we identified complete Prdm9 homologs in 16 species (Supplementary Table 2). For the 36 other species, we further analyzed the corresponding reference genome to identify potential Prdm9 homologs. We performed a TBLASTN search against reference genomes, using the 16 previously identified Prdm9 as queries, and extracted loci presenting hits with the zinc finger domain and with at least 2 other Prdm9 exons, within less than 100 kb. Then, for each candidate genomic fragment, we used GeneWise (Birney et al., 2004) to annotate protein-coding regions by similarity with a representative complete Prdm9 protein taken from a closely related species. Of

note, GeneWise does take frameshifts into account, and is therefore appropriate to annotate both genes and pseudogenes. In most species we identified one single candidate locus per genome. When several loci were found, we retained the one(s) encoding the most complete protein. By this approach, we identified Prdm9 loci in all 36 species, 25 of which encode a complete Prdm9 protein. Thus, in total, we identified complete Prdm9 proteins in 41 of the 52 species analyzed. In agreement with previous reports ((Axelsson et al., 2012; Auton et al., 2013)), we found Prdm9 to be pseudogenized in the 5 canids (*Lycaon pictus*, *Canis lupus familiaris*, *Vulpes lagopus*, *Vulpes vulpes*, *Nyctereutes procyonoides*). In the 6 remaining cases, we failed to annotate a complete Prdm9 protein: either several exons were missing (*Myotis lucifugus*, *Myotis brandtii*, *Leptonychotes weddellii*, *Colobus angolensis*), or the gene contained one exon with a frameshifting mutation (respectively in exon 8 in *Saimiri boliviensis* and in exon 10 in *Lipotes vexillifer*). In absence of data from more individuals, it is difficult to state whether these cases result from sequencing errors or assembly artefacts or if they correspond to bona fide pseudogenes. We therefore tentatively annotate these 6 cases as 'partial' Prdm9. The number of zinc fingers in the annotated proteins vary from 0 to 14. This number should however be considered with caution because the zinc finger array is encoded by a highly polymorphic minisatellite repeat, which is prone to errors during genome assembly. The detailed list of Prdm9 sequences is given in Supplementary Table 2 and the corresponding protein multiple alignment is given in Supplementary Data 1.

Acknowledgments

We wish to thank Nicolas Lartillot, Carina Farah Mugal, and Bernard de Massy for very useful reviews on a previous version of this manuscript, and Anaïs Duhamel for help with the figures. This work was performed using the computing facilities of the CC LBBE/PRABI. **Funding:** Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec. **Author contributions:** Original idea: J.J.; Model conception: J.J.; Code: J.J., D.P, A.L, T.T and L.D; Data analyses: J.J., D.P, A.L, T.T and L.D; Interpretation: J.J., D.P, A.L, T.T and L.D; First draft: J.J.; Editing and revisions: J.J., D.P, A.L, T.T and L.D; Funding: L.D. **Competing interests:** The authors declare no conflicts of interest. **Data and materials availability:** Analysis scripts and documentation will be available upon deposit of this manuscript on a preprint server.

References

Alleva, B., Brick, K., Pratto, F., Huang, M., and Camerini-Otero, R. D. (2021). Cataloging Human PRDM9 Allelic Variation Using Long-Read Sequencing Reveals PRDM9 Population Specificity and Two Distinct Groupings of Related Alleles. *Frontiers in Cell and Developmental Biology*, 9.

Arbeithuber, B., Betancourt, A. J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences*, 112(7):2109–2114.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., Zhang, G., and Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251.

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., Myers, S., Donnelly, P., Przeworski, M., and McVean, G. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*, 336:7.

Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. *Genome Res.*, 17(8):1219–1227.

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., Hayward, J. J., Cohen, P. E., Greally, J. M., Wang, J., Bustamante, C. D., and Boyko, A. R. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genet*, 9(12):e1003984.

Axelsson, E., Webster, M. T., Ratnakumar, A., Consortium, T. L., Ponting, C. P., and Lindblad-Toh, K. (2012). Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.*, 22(1):51–63.

Baker, C. L., Kajita, S., Walker, M., Saxl, R. L., Raghupathy, N., Choi, K., Petkov, P. M., and Paigen, K. (2015). PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. *PLOS Genetics*, 11(1):e1004916.

Baker, Z., Przeworski, M., and Sella, G. (2022). Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence.

Baker, Z., Schumer, M., Haba, Y., Bashkirova, L., Holland, C., Rosenthal, G. G., and Przeworski, M. (2017). Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *eLife*, 6:e24133.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science*, 327(5967):836–840.

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., and Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, pages 1–7.

Berglund, J., Quilez, J., Arndt, P. F., and Webster, M. T. (2015). Germline Methylation Patterns Determine the Distribution of Recombination Events in the Dog Genome. *Genome Biology and Evolution*, 7(2):522–530.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.*, 14(5):988–995.

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. (2016). Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an Avian System. *Mol Biol Evol*, 33(1):216–227.

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400):642–645.

Buard, J., Rivals, E., Segonzac, D. D. d., Garres, C., Caminade, P., Massy, B. d., and Boursot, P. (2014). Diversity of Prdm9 Zinc Finger Array in Wild Mice Unravels New Facets of the Evolutionary Turnover of this Coding Minisatellite. *PLOS ONE*, 9(1):e85021.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.

Cavassim, M. I. A., Baker, Z., Hoge, C., Schierup, M. H., Schumer, M., and Przeworski, M. (2022). PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2. *Proceedings of the National Academy of Sciences*, 119(9):e2114401119.

Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLOS Genetics*, 8(12):e1003090.

Charlesworth, D., Zhang, Y., Bergero, R., Graham, C., Gardner, J., and Yong, L. (2020). Using GC Content to Compare Recombination Patterns on the Sex Chromosomes and Autosomes of the Guppy, Poecilia reticulata, and Its Close Outgroup Species. *Molecular Biology and Evolution*, 37(12):3550–3562.

Choi, K., Zhao, X., Tock, A. J., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J., Ziolkowski, P. A., Yelina, N. E., Hwang, I., Martienssen, R. A., and Henderson, I. R. (2018). Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. *Genome Res.*, 28(4):532–546.

Clément, Y. and Arndt, P. F. (2013). Meiotic Recombination Strongly Influences GC-Content Evolution in Short Regions in the Mouse Genome. *Molecular Biology and Evolution*, 30(12):2612–2618.

Davies, B., Hatton, E., Altemose, N., Hussin, J. G., Pratto, F., Zhang, G., Hinch, A. G., Moralli, D., Biggs, D., Diaz, R., Preece, C., Li, R., Bitoun, E., Brick, K., Green, C. M., Camerini-Otero, R. D., Myers, S. R., and Donnelly, P. (2016). Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature*, 530(7589):171–176.

Diagouraga, B., Clément, J. A. J., Duret, L., Kadlec, J., de Massy, B., and Baudat, F. (2018). PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites. *Molecular Cell*, 69(5):853–865.e6.

Duret, L. and Arndt, P. F. (2008). The Impact of Recombination on Nucleotide Substitutions in the Human Genome. *PLOS Genetics*, 4(5):e1000071.

Eyre-Walker, A. (1999). Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. *Genetics*, 152(2):675–683.

Figuet, E., Ballenghien, M., Romiguier, J., and Galtier, N. (2015). Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates. *Genome Biology and Evolution*, 7(1):240–250.

Galtier, N. (2021). Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer Community Journal*, 1.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5):1092–1103.

Genereux, D. P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V. D., Murén, E., Juan, D., Bejerano, G., Casewell, N. R., Chemnick, L. G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I. T., Garber, M., Gladyshev, V. N., Goodman, L., Haerty, W., Houck, M. L., Hubley, R., Kivioja, T., Koepfli, K.-P., Kuderna, L. F. K., Lander, E. S., Meadows, J. R. S., Murphy, W. J., Nash, W., Noh, H. J., Nweeia, M., Pfenning, A. R., Pollard, K. S., Ray, D. A., Shapiro, B., Smit, A. F. A., Springer, M. S., Steiner, C. C., Swofford, R., Taipale, J., Teeling, E. C., Turner-Maier, J., Alfoldi, J., Birren, B., Ryder, O. A., Lewin, H. A., Paten, B., Marques-Bonet, T., Lindblad-Toh, K., Karlsson, E. K., and Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833):240–245.

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

Gerton, J. L. and Hawley, R. S. (2005). Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nat Rev Genet*, 6(6):477–487.

Glémin, S. (2010). Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics*, 185(3):939–959.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Res.*, 25(8):1215–1228.

Halldorsson, B. V., Hardarson, M. T., Kehr, B., Styrkarsdottir, U., Gylfason, A., Thorleifsson, G., Zink, F., Jonasdottir, A., Jonasdottir, A., Sulem, P., Masson, G., Thorsteinsdottir, U., Helgason, A., Kong, A., Gudbjartsson, D. F., and Stefansson, K. (2016). The rate of meiotic gene conversion varies by sex and age. *Nat Genet*, 48(11):1377–1384.

Hammoud, S. S., Low, D. H. P., Yi, C., Carrell, D. T., Guccione, E., and Cairns, B. R. (2014). Chromatin and Transcription Transitions of Mammalian Adult Germline Stem Cells and Spermatogenesis. *Cell Stem Cell*, 15(2):239–253.

Harris, K. and Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, 6:e24284.

He, Y., Wang, M., Dukowic-Schulze, S., Zhou, A., Tiang, C.-L., Shilo, S., Sidhu, G. K., Eichten, S., Bradbury, P., Springer, N. M., Buckler, E. S., Levy, A. A., Sun, Q., Pillardy, J., Kianian, P. M. A., Kianian, S. F., Chen, C., and Pawlowski, W. P. (2017). Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proceedings of the National Academy of Sciences*, 114(46):12231–12236.

Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10):1341–1342.

Hinch, A. G., Zhang, G., Becker, P. W., Moralli, D., Hinch, R., Davies, B., Bowden, R., and Donnelly, P. (2019). Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science*, 363(6433):eaau8861.

Hoge, C. R., Manuel, M. d., Mahgoub, M., Okami, N., Fuller, Z. L., Banerjee, S., Baker, Z., Mcnulty, M., Andolfatto, P., Macfarlan, T. S., Schumer, M., Tzika, A. C., and Przeworski, M. (2023). Patterns of recombination in snakes reveal a tug of war between PRDM9 and promoter-like features.

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L., and Ellegren, H. (2017). Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Mol Ecol*, 26(16):4158–4172.

Kono, H., Tamura, M., Osada, N., Suzuki, H., Abe, K., Moriwaki, K., Ohta, K., and Shiroishi, T. (2014). Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. *DNA Research*, 21(3):315–326.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., and Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, 39(8):msac174.

Lartillot, N. (2013). Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. *Molecular Biology and Evolution*, 30(3):489–502.

Latrille, T., Duret, L., and Lartillot, N. (2017). The Red Queen model of recombination hot-spot evolution: a theoretical investigation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736):20160463.

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., and Duret, L. (2014). The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. *PLOS Genetics*, 10(11):e1004790.

Li, R., Bitoun, E., Altemose, N., Davies, R. W., Davies, B., and Myers, S. R. (2019). A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat Commun*, 10(1):3900.

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

Maloisel, L. and Rossignol, J.-L. (1998). Suppression of crossing-over by DNA methylation in Ascobolus. *Genes & Development*, 12(9):1381–1389.

Marand, A. P., Jansky, S. H., Zhao, H., Leisner, C. P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A. J., Veilleux, R. E., Buell, C. R., and Jiang, J. (2017). Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biol*, 18(1):1–16.

Mihola, O., Landa, V., Pratto, F., Brick, K., Kobets, T., Kusari, F., Gasic, S., Smagulova, F., Grey, C., Flachs, P., Gergelits, V., Tresnak, K., Silhavy, J., Mlejnek, P., Camerini-Otero, R. D., Pravenec, M., Petukhova, G. V., and Trachtulec, Z. (2021). Rat PRDM9 shapes recombination landscapes, duration of meiosis, gametogenesis, and age of fertility. *BMC Biol*, 19(1):1–20.

Mihola, O., Pratto, F., Brick, K., Linhartova, E., Kobets, T., Flachs, P., Baker, C. L., Sedlacek, R., Paigen, K., Petkov, P. M., Camerini-Otero, R. D., and Trachtulec, Z. (2019). Histone methyltransferase PRDM9 is not essential for meiosis in male mice. *Genome Res.*, 29(7):1078–1086.

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nat Commun*, 8(1):15183.

Munch, K., Mailund, T., Dutheil, J. Y., and Schierup, M. H. (2014). A fine-scale recombination map of the human–chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Res.*, 24(3):467–474.

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science*, 327(5967):876–879.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(20):6278–6281.

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., Beatson, S. A., Lunter, G., Malik, H. S., and Ponting, C. P. (2009). Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. *PLOS Genetics*, 5(12):e1000753.

Page, S. L. and Hawley, R. S. (2003). Chromosome Choreography: The Meiotic Ballet. *Science*, 301(5634):785–789.

Parvanov, E. D., Petkov, P. M., and Paigen, K. (2010). Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, 327(5967):835–835.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome biology and evolution*, 4(7):675–682.

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. *Nat Rev Genet*, 2(5):360–369.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., and Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science*, 346(6211):1256442–1256442.

Qu, J., Hodges, E., Molaro, A., Gagneux, P., Dean, M. D., Hannon, G. J., and Smith, A. D. (2018). Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. *Genome Res.*, 28(2):145–158.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Raynaud, M., Gagnaire, P.-A., and Galtier, N. (2023). Performance and limitations of linkage-disequilibrium-based methods for inferring the genomic landscape of recombination and detecting hotspots: a simulation study. *Peer Community Journal*, 3.

Samuk, K. and Noor, M. A. F. (2022). Gene flow biases population genetic inference of recombination rate. *G3 Genes|Genomes|Genetics*, 12(11):jkac236.

Schield, D. R., Pasquesi, G. I. M., Perry, B. W., Adams, R. H., Nikolakis, Z. L., Westfall, A. K., Orton, R. W., Meik, J. M., Mackessy, S. P., and Castoe, T. A. (2020). Snake Recombination Landscapes Are Concentrated in Functional Regions despite PRDM9. *Molecular Biology and Evolution*, 37(5):1272–1294.

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., McVean, G., and Przeworski, M. (2015). Stable recombination hotspots in birds. *Science*, page 6.

Smagulova, F., Brick, K., Pu, Y., Camerini-Otero, R. D., and Petukhova, G. V. (2016). The evolutionary turnover of recombination hot spots contributes to speciation in mice. *Genes Dev.*, 30(3):266–280.

Smith, T. C. A., Arndt, P. F., and Eyre-Walker, A. (2018). Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet*, 14(3):e1007254.

Spence, J. P. and Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10):eaaw9206.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., and Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil. Trans. R. Soc. B*, 372(1736):20160455.

Tock, A. J. and Henderson, I. R. (2018). Hotspots for Initiation of Meiotic Recombination. *Front. Genet.*, 9:521.

Wallberg, A., Glémin, S., and Webster, M. T. (2015). Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, Apis mellifera. *PLoS Genet*, 11(4):e1005189.

Wang, R. J., Thomas, G. W. C., Raveendran, M., Harris, R. A., Doddapaneni, H., Muzny, D. M., Capitanio, J. P., Radivojac, P., Rogers, J., and Hahn, M. W. (2020). Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. *Genome Res.*, 30(6):826–834.

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., and Przeworski, M. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4:e04637.

Wong, W. S. W., Solomon, B. D., Bodian, D. L., Kothiyal, P., Eley, G., Huddleston, K. C., Baker, R., Thach, D. C., Iyer, R. K., Vockley, J. G., and Niederhuber, J. E. (2016). New observations on maternal age effect on germline de novo mutations. *Nat Commun*, 7(1):10486.

Zelkowski, M., Olson, M. A., Wang, M., and Pawlowski, W. (2019). Diversity and Determinants of Meiotic Recombination Landscapes. *Trends in Genetics*, 35(5):359–370.

Úbeda, F. and Wilkins, J. F. (2011). The Red Queen theory of recombination hotspots. *Journal of Evolutionary Biology*, 24(3):541–553.

7

On the origin and maintenance of recombination hotspots

7.1 Two types of recombination hotspots in eukaryotes

In the previous chapter, we showed that two types of recombination hotspots coexist in placental mammals. They correspond to the only two types of hotspots that have been described in eukaryotes to my knowledge. The first ones were first discovered in yeast, where a clustering of gene conversion events was observed in 5' of genes such as ARG4 and HIS4 (reviewed in Lichten and Goldman (1995)). This observation was extended later to crossovers in many other promoters of protein-coding genes, and is thought to

be linked to open chromatin (again reviewed in Lichten and Goldman (1995)). Studies in angiosperms, passerines, snakes and some mammals revealed a very similar pattern, with recombination hotspots occurring in promoters of protein-coding genes (Brick *et al.*, 2012; Auton *et al.*, 2013; Choi and Henderson, 2015; Singhal *et al.*, 2015; Kawakami *et al.*, 2017). Those hotspots appear to be shared by closely related species and are thus evolutionary stable (Axelsson *et al.*, 2012; Singhal *et al.*, 2015; Lam and Keeney, 2015). They have been reported in all major clades of eukaryotes, and are therefore thought to be ancestral to all eukaryotes. In this section, they will be referred to as ancestral hotspots.

The second type of hotspots, which are active in almost all mammals with the notable exception of canids, are directed by the protein Prdm9 (Parvanov *et al.*, 2010; Baudat *et al.*, 2010; Myers *et al.*, 2010). This protein first binds a specific DNA motif with its C2H2 zinc finger array (Parvanov *et al.*, 2010; Baudat *et al.*, 2010; Myers *et al.*, 2010). Second, it marks histones with H3K4Me3 and H3K6Me36 marks (Brick *et al.*, 2012; Eram *et al.*, 2014; Powers *et al.*, 2016; Davies *et al.*, 2016; Grey *et al.*, 2017). This signal is thought to be recognized by a protein (unknown so far) that brings back the motif to the chromosomal axis, where it will receive a DSB (Baker *et al.*, 2015; Davies *et al.*, 2016; Li *et al.*, 2019; Hinch *et al.*, 2019). Importantly, when Prdm9 binds its targets on both homologs, the sequence that do not receive the DSB is also brought back to the chromosomal axis, thus close to its broken homolog (Baker *et al.*, 2015; Davies *et al.*, 2016; Li *et al.*, 2019; Hinch *et al.*, 2019). This symmetrical binding enhances the repair of DSBs as COs, potentially bringing a significant advantage for meiosis success (Baker *et al.*, 2015; Davies *et al.*, 2016; Li *et al.*, 2019; Hinch *et al.*, 2019; Baker *et al.*, 2022). The C2H2 zinc finger of Prdm9 having one of the highest mutation rate in the genome, the protein is extremely polymorphic (Buard *et al.*, 2014; Kono *et al.*, 2014; Alleva *et al.*, 2021). As different zinc finger arrays usually have different binding motifs, the position of Prdm9-directed recombination hotspots are usually not conserved between closely related species (Auton *et al.*, 2012; Lartillot, 2013; Smagulova *et al.*, 2016; Galtier, 2021).

In this section, I will discuss the evolutionary scenarios that may have given rise (or not) to these different types of hotspots, and what evolutionary forces contributed to their long-term maintenance.

7.2 The hotspot paradox

In yeasts, when deleting the ARG4 or HIS4 recombination hotspot on one homologous chromosome only, several studies reported that DSBs systematically occurred on the intact copy which was systematically converted by the other (Nicolas *et al.*, 1989; Schultes

and Szostak, 1991; Detloff *et al.*, 1992). This observation revealed whats is called now the hotspot paradox: if a mutation were to inactivate the functioning of a hotspot, it will systematically convert the intact copy, and spread quickly in the population (Boulton *et al.*, 1997). This is another type of biased gene conversion, driven by biased initiation of DSB (dBGC). Using simulations, Boulton and colleagues showed that this drive was so strong that neither indirect selection on selection efficiency, nor direct selection for correct chromosome pairing were sufficient to counteract it (Boulton *et al.*, 1997). In summary, in the face of mutations, recombination hotspots should not exist. Two different processes have or may have helped to maintain hotspots despite the hotspot paradox.

As previously said, in many vertebrate species, recombination hotspots are directed by the protein Prdm9 in specific DNA motifs (Parvanov *et al.*, 2010; Baudat *et al.*, 2010; Myers *et al.*, 2010). In mice, several studies demonstrate that Prdm9 erodes its targets through the mechanism of dBGC, leading to hotspot extinction (Baker *et al.*, 2015; Smagulova *et al.*, 2016). However, mutations on the DNA-binding zinc-finger array will often change the binding motifs, and therefore provide new targets (Buard *et al.*, 2014; Smagulova *et al.*, 2016; Alleva *et al.*, 2021). As long as the mutation rate of Prdm9's zinc finger is sufficiently high such that new hotpots appear faster than hotspot erosion, the system stays afloat (Genestier *et al.*, 2023). If the mutation rate is not high enough, individuals carrying old Prdm9 alleles suffer the fertility cost of this system, leading to positive selection on new alleles (Oliver *et al.*, 2009; Ponting, 2011; Latrille *et al.*, 2017). This phenomenon induces a Red Queen-like dynamic, where new alleles are under pervasive positive selection (Úbeda and Wilkins, 2011; Latrille *et al.*, 2017).

The second option would be to target an epigenetic mark which is more robust to mutations, and more dependent on context than on the underlying DNA sequence (Lichten and Goldman, 1995). As we will see in the following section, this appears to be the case for ancestral hotspots.

7.3 The evolutionary origin of ancestral recombination hotspots

7.3.1 Selection to maximize Hill-Robertson interference dissipation

The position of recombination hotspots close to selectively constrained sequences (protein-coding genes) is a nice way to optimize genetic shuffling between regions that

have strong effects on fitness. Interestingly, when considering distance as gene counts rather than physical distance, Brazier and Glémin (2022) showed that the recombination landscape becomes strikingly uniform in several plant species . It is thus tempting to interpret this pattern as an evidence for an indirect selective pressure on the recombination landscape to maximize Hill-Robertson interference dissipation where it matters most. But if we take a closer look at the mechanisms that determine the positions of these hotspots, several other interpretations are possible.

7.3.2 Opportunist recombination hotspots in open chromatin regions

In yeasts, almost all DSB hotspots are sensitive to DNase I and micrococcal nuclease (Wu and Lichten, 1994, 1995; Ohta *et al.*, 1994). Therefore, it seems that a primary condition for being a recombination hotspot is nucleosome depletion and open chromatin (Lichten and Goldman, 1995). This is also true for Prdm9-directed hotspots outside of gene promoters. Prdm9 has indeed been shown to modify chromatin state through histone methylation, further stabilized by the histone modification reader ZCWPW1 (Brick *et al.*, 2012; Davies *et al.*, 2016; Eram *et al.*, 2014; Powers *et al.*, 2016; Diagouraga *et al.*, 2018; Yuan *et al.*, 2022). For both categories of hotspots, open chromatin is required for DSB formation. It seems very plausible that in the absence of a Prdm9-like system that actively opens chromatin prior to DSB formation, hotspots will naturally occur where the chromatin is already open: in gene promoters. The location of DSB hotspots would not be the result of an indirect selection pressure, but a natural consequence of the requirements of protein complexes that need access to DNA.

7.3.3 Selection to reduce ectopic recombination events

It is worth noting that not all open chromatin regions correspond to recombination hotspots (Lichten and Goldman, 1995). In dogs and passerines, which both lack Prdm9, only gene promoters that contain a CpG island are recombination hotspots, suggesting an association between recombination and DNA methylation (Singhal *et al.*, 2015; Auton *et al.*, 2013). Despite a general conservation of recombination hotspots between dogs and a mutant mouse with an inactivated Prdm9, we showed in chapter 6 that the mutant mouse hotspots that are methylated in dog sperm are inactive in dogs. In *Ascobolus* fungi, experimentally induced methylation of the b2 hotspot also leads to its inactivation (Maloisel and Rossignol, 1998). Finally, in honey bees, genes that are methylated display a lower rate of cross-overs (Wallberg *et al.*, 2015).

Methylation of CpG dinucleotides is a widespread mechanism for transposable elements (TEs) repression in eukaryotes (reviewed in Goodier (2016)). In the plant *Arabidopsis thaliana*, but also in the fungi *Neurospora crassa*, methylation of TEs in meiosis is thought to be mediated by an interfering double-stranded RNA which recruits a methyltransferase that methylates CpG dinucleotides (Okamoto and Hirochika, 2001; Shiu *et al.*, 2001). In vertebrates, a well-diversified family of C2H2 zinc finger proteins (the KRAB-ZNFs) recognize specific motifs associated to TEs, and eventually recruit a protein that methylates CpG dinucleotides (reviewed in Yang *et al.* (2017)). It has been hypothesized that this methylation decreases the risk of ectopic recombination event by preventing the recombination machinery from accessing repeated elements (Langley *et al.*, 1988; Maloisel and Rossignol, 1998). It is also very clear that DNA methylation not only prevents DSBs from occuring in TEs, but represses TE expression and contributes to limiting the transposition-induced mutation load in the germline (Hancks and Kazazian, 2016). Altogether both consequences of DNA methylation are thought to have a positive effect on fitness, and could explain its persistence (Langley *et al.*, 1988; Maloisel and Rossignol, 1998; Charlesworth *et al.*, 1997; Roze, 2023). Moreover, even if the main driver of selection for TE methylation is the minimization of the mutational load through transcription repression, a recombination machinery that avoids the hallmarks of DNA methylation should in principle be selected for, as it should drastically reduce non homologous pairing and ectopic recombination events (Roze, 2023).

The exact nature of the mechanism that keeps recombination away from single copy hypomethylated regions is stil debated, and can differ between species. It has been demonstrated that the histone mark H3K4Me3, which is largely localized in hypomethylated regions, is clearly associated to recombination hotspots (Brick *et al.*, 2012; Auton *et al.*, 2013; Petes, 2001). In yeast, an inactivation of the H3K4 methylase Set1 leads to a severe reduction of DSB formation at recombination hotspots (Borde *et al.*, 2009; Sommermeyer *et al.*, 2013). However, a causal role of H3K4Me3 in the positioning of hotspots is much more debated in plants (Choi *et al.*, 2018; He *et al.*, 2017). In fact, in vertebrates, it is hard to dissociate the role of H3K4Me3 marks and DNA hypomethylation since they virtually always co-localize (Klose and Bird, 2006). However, in dogs, H3K4Me3 peaks that do not overlap a CpG island show little increase of recombination, whereas CpG islands devoid of H3K4Me3 marks still correspond to recombination hotspots (Auton *et al.*, 2013).

If the absence of DNA methylation is the major driver of recombination hotspots, and not H3K4Me3 marks, species that exhibit low or no DNA methylation should recombine everywhere, and largely lack recombination hotspots (Zamudio *et al.*, 2015). Intriguingly,

D.melanogaster and *C.elegans*, which show negligible levels of methylation display little fine scale variation of recombination rate (Chan *et al.*, 2012; Kaur and Rockman, 2014). Second, in mice with a DNA methylation deficient background, the concentration of DMC1 (indicating DSB activity) increases in TEs, outside of canonical recombination hotspots. Mugal *et al.* (2015) also showed that birds' genome were much less methylated than those of mammals or fishes. It is thus interesting to note that passerines show much less recombination hotspots than dogs, based on LD recombination maps (Auton *et al.*, 2013; Singhal *et al.*, 2015; Kawakami *et al.*, 2017). Finally, in many mammals, oocytes are much less methylated than the male germline (Bourc'his and Proudhon, 2008; Zamudio *et al.*, 2015). It is again interesting to note that female dogs use recombination hotspots significantly less than males (Campbell *et al.*, 2016). All these observations could suggest that the usage of ancestral recombination hotspots in many vertebrates is a direct consequence of the genome-wide methylation level. A strong testable prediction of this hypothesis is that hotspots density in Prdm9-lacking species should correlate positively with genome-wide levels of DNA methylation.

Overall, the position of recombination hotspots in single copy open chromatin regions in many eukaryotes could result from selection to avoid methylated repeated regions, leading to a reduction of harmful ectopic recombination events.

7.3.4 Opportunist recombination hotspots in replication origins

Another promoter-associated feature that imposes constraints on DSB formation is replication origins. In several eukaryotes, it has been shown that DNA replication and DSB initiation are coupled (Borde *et al.*, 2000; Martin *et al.*, 2011; Murakami and Nurse, 2001; Pratto *et al.*, 2021). In yeasts, Borde *et al.* (2000) showed that there is a tight timing of 1.5 to 2 hours between the passage of the replication fork and the formation of DSB (Borde *et al.*, 2000). When replication is delayed, so is DSB formation, keeping a constant time interval between the two (Borde *et al.*, 2000). In humans and mice, it has been shown that DSB formation was favoured in early replicated regions, no matter the hotspot density (Pratto *et al.*, 2021). In plants, humans, mice and chicken, replication origins are enriched in promoters, but also slightly enriched in transcription ending site, consistent with the recombination patterns observed in several plants and yeasts (Lichten and Goldman, 1995; Choi and Henderson, 2015; Massip *et al.*, 2019; Pratto *et al.*, 2021).

Overall, the existence of ancestral recombination hotpots could only reflect the

concentration of replication initiation in promoters.

7.4 The evolutionary advantage of Prdm9-directed recombination hotspots

An intact version of Prdm9 has been found in all major animal clades (Oliver *et al.*, 2009; Ponting, 2011). It is therefore very likely that it was present in the ancestor of all animals. Whether it kept the same function during the 600 Mys that separates us from this ancestor is still unknown. Evidence from salmonids' CO and DSB maps strongly suggest that Prdm9 is also involved in determining the position of recombination hotspots in this family (Raynaud et al. personal communications). Its function would therefore be traced back to the ancestor of Euteleostomi (∼420 Mya) (Kumar *et al.*, 2022).

The inactivation of this gene by KO experiments in male mice and rats leads to a drastic decrease in fertility due to an inability to efficiently repair DSBs generated at the ancestral location of hotspots (near gene promoters) (Brick *et al.*, 2012; Mihola *et al.*, 2019, 2021). Both its persistence through time and lineages, and the direct evidence in murids suggest that this gene is under strong selective constraints and should thus provide a critical fitness advantage. In this section, I will discuss several hypothesis that have been formulated for the evolutionary advantage of Prdm9 in regards of the recent literature on model and non-model species, and new insights from the present PhD work.

7.4.1 Selection to limit the deleterious effects of recombination in functional elements

Even before the role of Prdm9 in meiosis was understood, it was noticed that in humans and mice, unlike in plants and fungi, recombination hotspots occurred mainly outside functional elements(Lichten and Goldman, 1995). In regard of the damage that recombination can induce via its mutagenic effect, or biased gene conversion, it has been naturally formulated that one evolutionary advantage of Prdm9 could be to deviate recombination away from these functional elements, hence reducing the genome-wide genetic load (Brick *et al.*, 2012; Webster and Hurst, 2012). In this sense, it has been demonstrated that COs were depleted in the body of meiotically transcribed genes both in humans and mice (Necsulea *et al.*, 2009; McVicker and Green, 2010; Pouyet *et al.*, 2017; Jin *et al.*, 2021; Schwarzkopf and Cornejo, 2022). It has been further proposed that the depletion of Prdm9-directed COs in the body of genes could be an advantage to reduce conflicts between the transcription and the

recombination machinery which both need access to DNA in meiosis (Schwarzkopf and Cornejo, 2022). However, in mice, while the CO rate is lower in highly meiotically expressed genes, the DSB rate is higher (Jin *et al.*, 2021). Since the role of Prdm9 is to determine the position of DSBs, it suggests that Prdm9 is not involved in the deficit of CO observed in mice highly expressed genes. Moreover, we can observe the same deficit in dogs that have lost Prdm9 (Figure 7.1). It is thus very unlikely that Prdm9 plays any role in an avoidance mechanism of the transcription machinery, at least not directly. However, it is still possible that Prdm9 reduces the load in potentially functional hypomethylated regions, which are often promoters of highly expressed genes, a role that may have contributed to its maintenance.

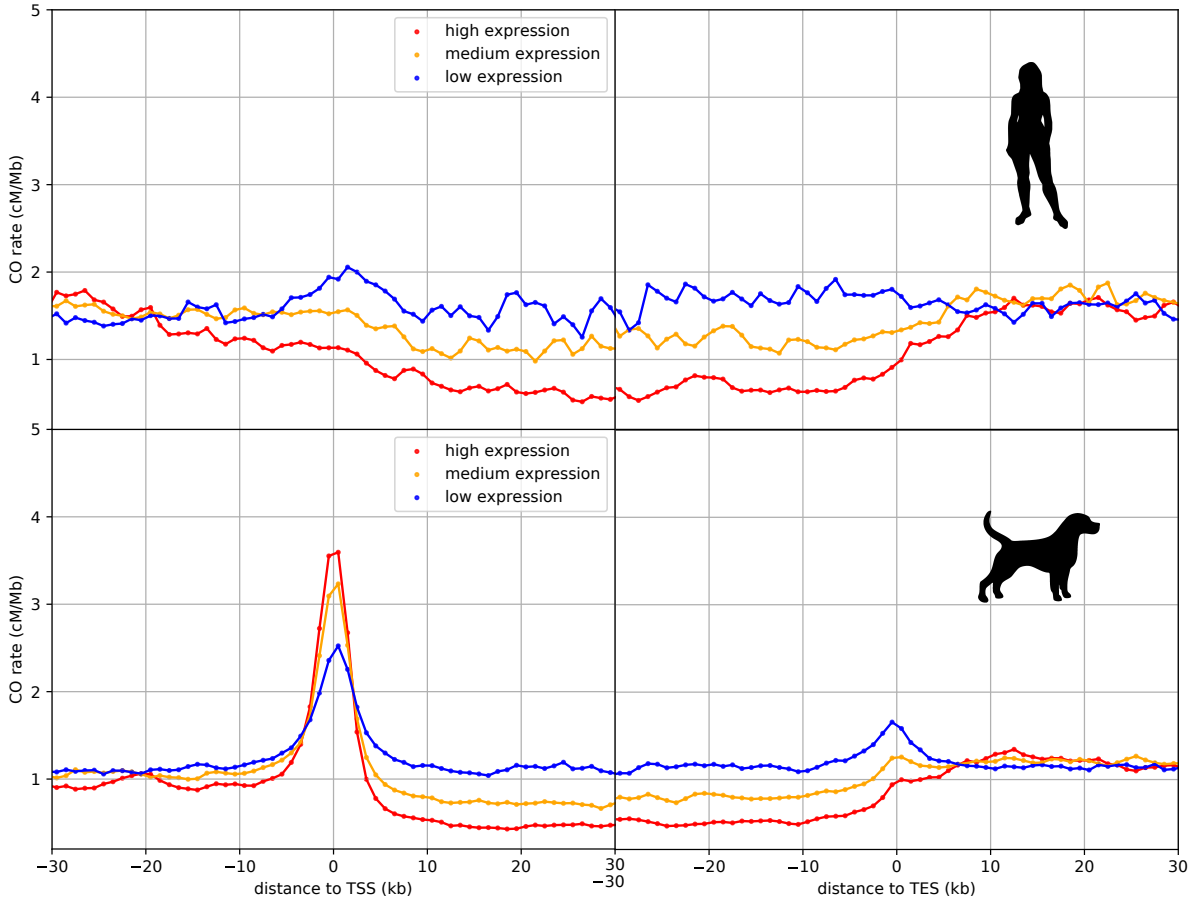


Figure 7.1: *Distribution of CO rate inside genes. Each dot represents a 1kb window. All dot between the TSS and TES are inside a protein coding gene. We distinguished 3 equal sized categories of genes regarding their expression in testis for dog and in ovary for humans. We kept only genes that are expressed in dog testis or human ovary (Dog : 23,107 Humans: 22,074)*

7.4.2 Prdm9 and transposable elements

While ancestral hotspots might increase the genetic load in selectively constrained hypomethylated regions, they at least provide a way to avoid ectopic recombination

events in repeated elements. Because Prdm9 targets short, specific DNA motifs throughout the entire genome, it cannot offer this guarantee. On the contrary, in humans, there is an enrichment of recombination hotspots in transposable elements. (Myers *et al.*, 2005). In mice, Buard *et al.* (2014) reported that predicted binding motifs of several Prdm9 alleles were enriched in transposable elements. Here, using the DMC1 ChIP-seq data for 6 different Prdm9 alleles, I show that this enrichment also concerns true DSB hotspots.

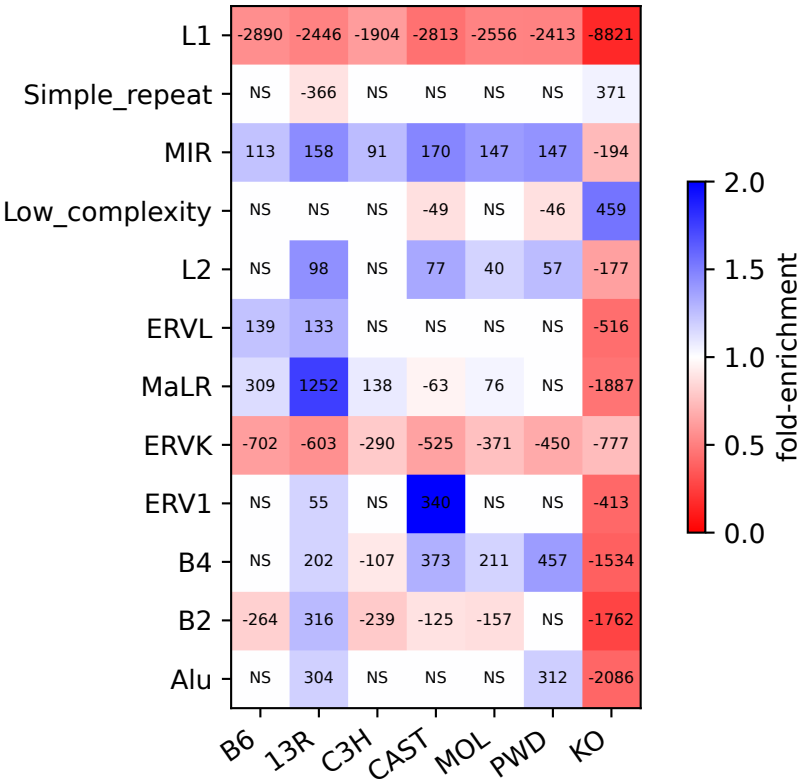


Figure 7.2: *Enrichment of DSB hotspots in repeated elements. Random expectation was computed as the mean of the overlap of repeated elements and 40 sets of random hotspots (see methods). If the value of the true overlap lies between those of the 40 sets of random hotspots, the enrichment is not significant (NS). Numbers correspond to the difference between the true number of overlap and the expected number of overlap. Fold-enrichment correspond to the ratio between the true number of overlap and the expected number of overlap.*

I computed the overlap of hotspots with repeated elements for each of the six alleles of the study of Smagulova *et al.* (2016). I then randomised the position of the hotspots to estimate whether they were more likely to occur near some TE families than by chance. We can see that hotspots of all alleles are depleted around L1 and ERVK TEs (Figure 7.2). This could reflect the fact that those TEs are in regions difficult to access for the DSB machinery. On the other hand, the hotspots of all alleles are enriched in the MIR family. Interestingly, we can see that the hotpots of 13R are enriched in 8 out of 12 families of TEs, and 2 to 4 out of 12 for other alleles (Figure 7.2). It is

particularly enriched in the MaLR family which are responsible for around 1,200 additional hotspots for 13R. We can also see that the hotspots of CAST are strongly enriched in the ERV1 TE family (twice more than by chance) (Figure 7.2). In bright contrast, ancestral hotspots (KO) are clearly depleted in all repeat families except for low complexity repeats for which there is a clear enrichment (Figure 7.2). Pratto and colleagues showed that half of the disease-associated non-allelic homologous structural variants reported in humans occur in the hotspots of the Prdm9^{A} allele, and that most of them occur within low copy repeats Pratto *et al.* (2014). We therefore have direct evidence that Prdm9-directed hotspots increase the load when targeting repeated elements. The fitness advantage of deviating recombination away from single copy functional elements is therefore not very obvious.

One other major shortcoming of Prdm9-directed hotspots is their short lifespan. Indeed, because of dBGC, a new Prdm9 allele progressively loses its target, potentially negatively impacting the fertility of its bearers. It has thus been proposed that targeting actively duplicating elements could be a way for an allele to never run short of targets (Yamada *et al.*, 2017). Additionally, eroding active transposable elements could provide a way to control their expansion (Yamada *et al.*, 2017). However, it is not clear how those weak indirect advantages scale with the direct deleterious effect of ectopic recombination events in repeated sequences (Charlesworth *et al.*, 1997; Roze, 2023).

7.4.3 The advantage of coupling DSB formation and repair

In yeast, the PHD finger module of Spp1 is able to read H3K4Me3 epigenetic marks, and tethers the DNA to the chromosomal axis where it will receive a DSB (Sommermeyer *et al.*, 2013). In mice, it would seem that Prdm9-bound sites are also brought back to the chromosomal axis. The proteins involved in this process remain however unknown (Baker *et al.*, 2015; Jin *et al.*, 2021). Prdm9 often binds its target symmetrically on both homologs (Baker *et al.*, 2015; Smagulova *et al.*, 2016; Li *et al.*, 2019; Hinch *et al.*, 2019). By doing so, Prdm9 brings both the sequence that will receive the DSB, and the sequence it will be repaired with on the chromosomal axis. Prdm9 would therefore confer a strong advantage in coupling DSB formation to its repair as a cross-over (Renkawitz *et al.*, 2014; Davies *et al.*, 2016; Li *et al.*, 2019; Hinch *et al.*, 2019). A strong binding of Prdm9 might be key for this process, as the probability of symmetrical binding directly depends on the affinity of the DSB-directing protein for its target, as well as its concentration in the cell (Baker *et al.*, 2022; Genestier *et al.*, 2023).

Interestingly, ancestral hotspots still exist in wild mice, but they are very weak

compared to the Prdm9-directed ones (Smagulova *et al.*, 2016). Importantly, they also correspond to the strongest hotspots found in the Prdm9-/- mouse (Smagulova *et al.*, 2016). This suggests that the protein involved in the direction of ancestral hotspots in mice has a much lower affinity for its targets. Whether this low affinity is inherent to targeting epigenetic marks rather than DNA motifs, or it is the result of the degeneration of this protein in mice could be tested with *in vitro* kinetics, provided that the protein involved is identified, and compared to a fully functioning protein such as yeast Spp1. Overall, Prdm9 might have acted as a booster of the DSB repair efficiency, allowing faster meiosis, with potentially less resources invested in DSB formation (Hinch *et al.*, 2019; Mihola *et al.*, 2019, 2021; Baker *et al.*, 2022; Genestier *et al.*, 2023).

It is interesting to note that after 600 Mys, the ancestral pathway has not disappeared, and still plays a significant role in determining recombination hotspots in species with a fully functional Prdm9 (Schield *et al.*, 2020; Hoge *et al.*, 2023). If this ancestral pathway is completely opportunistic and does not require any additional steps other than those of the Prdm9 system, its persistence could simply be passive. However, as highlighted before, the direct link to DNA methylation and H3K4Me3 marks, and the evidence for an active mechanism in yeasts rather suggest an active maintenance (Borde *et al.*, 2009; Sommermeyer *et al.*, 2013). In the previous chapter, we proposed an explanation for the maintenance of the ancestral pathway: the decrease in fitness of old alleles that fuels the Red Queen dynamic also maintains an alternative mechanism (detailed in the previous chapter). In any individual who struggles to make enough CO because of an old, inefficient allele, a mutation that increases the efficiency of the ancestral pathway will be positively selected. This might partly explain why Dapper and Payseur (2019) found pervasive positive selection on multiple recombination genes in mammals .

Overall, Prdm9 may have been selected for as a DSB repair facilitator, and not so much for deviating recombination away from functional elements. The fluctuating position of Prdm9-directed hotspots may therefore be a byproduct of the system, rather than a selected trait conferring an independent fitness advantage.

7.5 Conclusions and future directions

The location of hotspots near gene promoters could simply be the result of opportunistic behaviour of the recombination machinery, which needs open chromatin to access DNA (Lichten and Goldman, 1995). This location could also be constrained by the location of replication origins (Borde *et al.*, 2000; Murakami and Nurse, 2001; Martin *et al.*, 2011; Pratto *et al.*, 2021). In either case, the positioning of hotspots is

unrelated to the costs and benefits of recombination. Alternatively, the strong association between recombination hotspots and hypomethylation could suggest that positive selection to avoid repeated elements may have played a role in their maintenance (Berglund *et al.*, 2015; Wallberg *et al.*, 2015; Choi *et al.*, 2018). The observation of many disease-associated ectopic recombination events in low-copy repeats in humans supports this hypothesis (Pratto *et al.*, 2014).

In turn, the main advantage of Prdm9 would be to promote efficient coupling between DSB formation and repair, thanks to the high affinity of Prdm9 for its targets (Hinch *et al.*, 2019; Baker *et al.*, 2022). As this gene's first discovered role was to determine the position of recombination hotspots mainly outside of functional elements, it has been hypothesized that it could reduce the recombination-induced genetic load (Webster and Hurst, 2012; Brick *et al.*, 2012; de Massy, 2013). Here, I first confirmed that PRDM9 was not involved in the deviation of recombination away from gene body, and obviously could not have been selected for that, as previously thought. In addition, it seems unclear whether the load induced by recombining within hypomethylated regions is higher than that of recombining outside, risking increased ectopic recombination events (Pratto *et al.*, 2014). My personal take would be that Prdm9 is more of a facilitator of DSB repair, and this could be the reason for its prevalence in animals.

This belief relies on the assumption that Prdm9 has intrinsically more affinity for its target that any protein that binds histone marks or unmethylated CpG dinucleotides. This could be tested *in vitro* by approaches such as the ones of Patel and colleagues for Prdm9 in humans (Patel *et al.*, 2016), and He and colleagues for Spp1 in yeasts (He *et al.*, 2019). In addition, if it exists, the molecular system involved in the positioning of the ancestral hotspots in species with a functional Prdm9 remains to be identified. The load induced by recombining inside or outside functional elements also remains to be investigated. To answer this question, we need theoretical and empirical studies that can compare the burden of recombination-induced mutagenic effects with the fitness cost of ectopic recombination. In addition to the mutagenic effect, GC-biased gene conversion (gBGC) is another often cited source of load induced by recombination. Nevertheless, as we will see in chapter 11, a significant gBGC load of slightly deleterious mutations in the population does not mean that gBGC is negatively selected. It is thus not clear whether gBGC could play a role in the evolution of recombination landscapes.

Material and methods

Expression datasets and recombination maps

The expression data in humans were obtained from the RNA-seq experiment from (EBI accession number E-MTAB-1733) on human ovaries Fagerberg *et al.* (2014); Kryuchkova-Mostacci and Robinson-Rechavi (2015), extracted from the supplementary data of (Pouyet *et al.*, 2017). The dog expression data is from the RNA-seq experiment of Chen et al. 2018 on male testis Chen *et al.* (2019) (GEO accession number GSE106077). The expression were not measured particularly on meiotic cells but it is the closest transcription profile to a meiotic one which is available on GEO for dogs. The recombination rates in humans were estimated with linkage disequilibrium in Frazer and Consortium (2007) (RRID: SCR_002846). The dog recombination map was also estimated from linkage disequilibrium in Auton *et al.* (2013).

Enrichment in repeated elements

DSB hotspots were retrieved from the study of Smagulova *et al.* (2016) (GEO accession number GSE75419). The dataset contains the positions of hotspots called on DMC1 Chip-seq experiments in meiotic cells on seven individuals, six homozygous for a given Prdm9 allele, and one whose Prdm9 has been inactivated with a knock-out. The positions of repeated elements on the mm10 reference genome were downloaded from the UCSC Table Browser. For each allele, we computed the overlap between each family of repeated elements and the center of hotspots $\pm$ 100 bp. We then created 40 instances of control sequences by shuffling the hotspots in the genome. If a region of the genome of more than 2kb had zero SSDS reads in the ChIP-seq reads, this region was excluded from the potential random hotspots because it is likely that hotspots would not be callable even if they existed. We also excluded the ChIP-seq blacklisted regions from the ENCODE project (Landt *et al.*, 2012), which had been excluded in the original study because it resulted in spurious ChIP-seq peaks. The overlap expected by chance was taken as the mean of the overlap between the 40 sets of random hotspots and repeated elements. If the true value of the overlap was within the values of the 40 sets of random hotspots, the enrichment was not considered as significant.

Part III

The role of beneficial back-mutations in molecular evolution

8

Mammalian protein-coding genes exhibit widespread beneficial mutations that are not adaptive

Context

After having demonstrated the existence of the two kinds of recombination hotspots in mammals, we will be able to address their consequences on selection dynamics via gBGC. But first, we will demonstrate the substantial role of non-adaptive beneficial mutations in molecular evolution, which will be critical for understanding the interplay between recombination hotspots, gBGC and natural selection. We started this project in collaboration with Thibault Latrille, a former PhD student of the lab and then received help from Diego Hartasánchez, a former postdoc of the lab and their postdoc advisor Nicolas Salamin, all working in the University of Lausanne. Thibault came to me in the middle of my second year of PhD with an unused fitness landscape he estimated on mammals for one of his postdoc paper. He wanted to know if I had some ideas regarding questions about natural selection we could answer with this fitness landscape. This came out to be a great coincidence as I was in search of such a fitness landscape to assess the benefits and costs of recombination in species with contrasted recombination landscapes. As a first step, we evaluated whether the fitness landscape was correct using orthogonal validations. The results turned out to be very surprising and by themselves required to be communicated in an independent article. At first, I was uncomfortable including a paper that did not contain the words meiotic recombination or gBGC in a thesis on meiotic recombination and gBGC. But with the months passing by, and the number of projects that resulted in unsatisfactory outcomes increasing, Laurent righteously suggested me to find a way to include it anyway, if I wanted anything to be written in my PhD. In the end, the findings articulated quite well with the other chapters, and I am quite satisfied of its contribution to the scientific narrative of this PhD. I hope the reader gets as excited by the results as I was.

Detailed contributions

This study was co-designed by Thibault and I. We jointly analysed the data and interpreted the results. However, codes were written by Thibault as his coding skills were (and still are) way greater than mine. We benefited from the advices of Nicolas Salamin, and Diego Hartasánchez. We jointly wrote the first draft with Thibault, and benefited from the substantial help of Nicolas and Diego for improving the writing. The manuscript is currently in the second phase of review in *PNAS*.

Mammalian protein-coding genes exhibit widespread beneficial mutations that are not adaptive

T. Latrille[1†], **J. Joseph**[2†], **D. A. Hartasánchez**[1], **N. Salamin**[1]

[1]Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland

[2]Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Lyon 1, Villeurbanne, France

[†]These authors contributed equally to this work

thibault.latrille@ens-lyon.org

October 16, 2023

Abstract

Mutations can be beneficial by bringing innovation to their bearer, allowing them to adapt to environmental change. These mutations are typically unpredictable since they respond to an unforeseen change in the environment. However, mutations can also be beneficial because they are simply restoring an ancestral sequence of higher fitness that was lost due to genetic drift. In contrast to adaptive mutations, these beneficial back-mutations can be predicted if the underlying fitness landscape is stable and known. The contribution of such beneficial back-mutations to molecular evolution has been widely neglected mainly because their detection is very challenging. We have here reconstructed protein-coding-gene fitness landscapes shared between mammals, using mutation-selection models and a multi-species alignments across 87 mammals. These fitness landscapes have allowed us to predict the fitness effect of polymorphisms found in 28 mammalian populations. Using methods that quantify selection at the population level, we have confirmed that beneficial back-mutations are indeed positively selected in extant populations. Our work confirms that deleterious substitutions are accumulating in mammals and are being reverted, generating a balance in which genomes are damaged and restored simultaneously at different loci. We observe that beneficial back-mutations represent between 15% and 45% of all beneficial mutations in 24 of 28 populations analyzed, showing that a substantial part of ongoing positive selection is not driven by adaptation to environmental change in mammals.

***K*eywords** Adaptation · Back-mutations · Phylogenetics · Population-genetics · Codon models

Significance statement

The extent to which adaptation to changing environments is shaping genomes is a central question in molecular evolution. To quantify the rate of adaptation, population geneticists have typically used signatures of positive selection. However, mutations restoring an ancestral sequence of higher fitness lost by genetic drift are also positively selected, but they do not respond to a change in the environment. In this study, we have managed to distinguish beneficial mutations that are due to changing environments and those that are restoring pre-existing functions in mammals. We show that a substantial proportion of beneficial mutations cannot be interpreted as adaptive.

Introduction

Adaptation is one of the main processes shaping the diversity of forms and functions across the tree of life (Darwin, 1859). Evolutionary adaptation is tightly linked to environmental change and species responding to this change (Merrell, 1994; Gavrilets and Losos, 2009). For adaptation to occur, there must be variation within populations, which mostly appears via mutations in the DNA sequence. While neutral mutations will not impact an individual's fitness, deleterious mutations have a negative effect, and beneficial mutations improve their bearer's fitness. A beneficial mutation is thus more likely than a neutral mutation to invade the population and reach fixation, resulting in a substitution at the species level. Upon environmental change, because adaptive beneficial mutations toward new fitness optima are more likely, the number of substitutions also increases (fig. 1A). An increased substitution rate is thus commonly interpreted as a sign of adaptation (McDonald and Kreitman, 1991; Smith and Eyre-Walker, 2002; Welch, 2006). The availability of large-scale genomic data and the development of theoretical models have enabled the detection and quantification of substitution rate changes across genes and lineages (Yang and Bielawski, 2000; Eyre-Walker, 2006; Moutinho et al., 2019). These approaches, now common practice in evolutionary biology, have helped better understand the processes underpinning the rates of molecular evolution, contributing to disentangling the effects of mutation, selection and drift in evolution (Lynch, 2023). However, a collateral effect has been conflating beneficial mutations with adaptive evolution when adaptive evolution is not the only process that can lead to beneficial mutations (Charlesworth and Eyre-Walker, 2007; Mustonen and Lässig, 2009; Jones et al., 2017).

In a constant environment, a deleterious mutation can reach fixation by genetic drift (Ohta, 1992). A new mutation restoring the ancestral fitness will thus be beneficial (fig. 1B), even though the environment has not changed (Gillespie, 1995; Hartl and Taubes, 1996; Sella and Hirsh, 2005; Mustonen and Lässig, 2009; Cvijović et al., 2015). The restoration of the ancestral fitness can either happen through a mutation at a different locus – called a compensatory mutation (Hartl and Taubes, 1996; Mustonen and Lässig, 2009), or at the locus of the initial mutation – called a beneficial back-mutation (Piganeau and Eyre-Walker, 2003; Charlesworth and Eyre-Walker, 2007). While compensatory mutations change the sequence and thus induce genetic diversification, beneficial back-mutations reduce genetic diversity and do not contribute to genetic innovation. Although Tomoko Ohta considered beneficial back-mutations negligible in her nearly-neutral theory (Ohta, 1992), their importance has now been acknowledged for expanding populations (Charlesworth and

Eyre-Walker, 2007). However, differentiating between an adaptive mutation and a beneficial back-mutation remains challenging (Chi et al., 2020). Indeed, an adaptive mutation responding to a change in the environment and a beneficial back-mutation have equivalent fitness consequences for their bearer (Charlesworth and Eyre-Walker, 2007). Similarly, at the population level, both types of mutations will result in a positive transmission bias of the beneficial allele. However, at the macro-evolutionary scale, the consequences of these two types of mutations are fundamentally different.

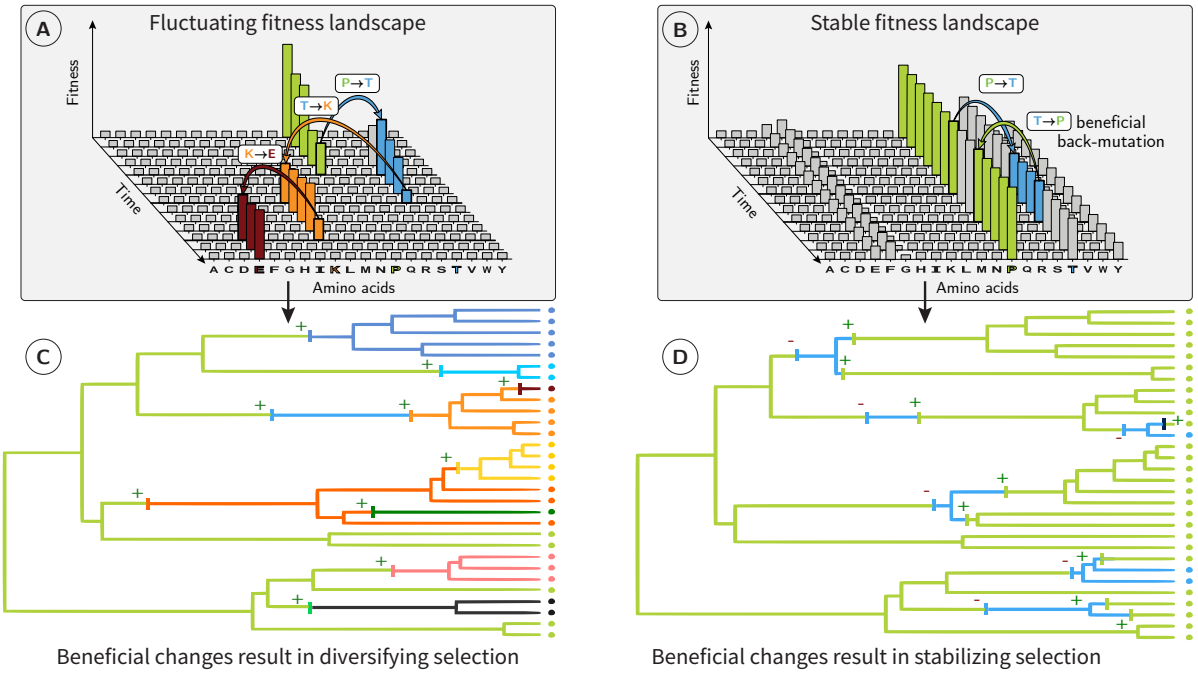


Figure 1: (A & B) For a given codon position of a protein-coding DNA sequence, amino acids (x-axis) have different fitness values (y-axis). Under a fluctuating fitness landscape (A), these fitnesses change with time. The protein sequence follows the moving target defined by the amino-acid fitnesses. Since substitutions are preferentially accepted if they are in the direction of this target, substitutions are, on average, adaptive. At the phylogenetic scale (C), beneficial substitutions are common (positive signs), promoting phenotype diversification across species. Under a stable fitness landscape (B), most mutations reaching fixation are either slightly deleterious reaching fixation due to drift or are beneficial back-mutations restoring a more optimal amino acid. At the phylogenetic scale (D), deleterious substitutions (negative signs) are often reverted via beneficial back-mutations (positive signs), promoting phenotype stability and preserving well-established biological systems. Even though, individually, any back-mutation might have a small beneficial effect on its bearer, we expect beneficial back-mutations to be scattered across the genome and the genome-wide signature of beneficial back-mutations to be detectable and quantifiable.

While adaptive mutations promote phenotype diversification (fig. 1C), beneficial back-mutations promote phenotype stability and may help preserve well-established biological systems (fig. 1D). Additionally, the direction of adaptive evolution is unpredictable because it is caused by an unforeseen change in the environment and, hence, in the underlying fitness landscape (Bazykin, 2015). On the other hand, benefi-

cial back-mutations are predictable because, under a stable fitness landscape, any change from non-optimal to optimal amino acids will move back the site towards the equilibrium expected under the fitness landscape (Moses and Durbin, 2009; Fischer et al., 2011; Chen et al., 2021). It can then be distinguished from truly novel beneficial mutations because the latter are not expected to mutate mutate towards the amino acids of higher fitnesses defined by the stable fitness landscape but rather mutate to amino acids showing a diversified pattern (fig. 1).

Fitness landscape reconstruction

The mutation-selection framework permits to link the patterns of substitution along a phylogenetic tree with the underlying fitness landscape (Halpern and Bruno, 1998; McCandlish and Stoltzfus, 2014). Such mutation-selection models applied to protein-coding DNA sequences at the codon level allow us to estimate relative fitnesses for all amino acids for each site of the sequence, explicitly assuming that the underlying fitness landscape is stable along the phylogenetic tree (Rodrigue and Philippe, 2010; Tamuri and Goldstein, 2012; Rodrigue and Lartillot, 2017). Moreover, effective population size (N_e) is considered constant along the phylogenetic tree precisely because of the fixed fitness landscape assumption, the consequences of which are detailed in the Discussion. Importantly, because mutation-selection codon models at the phylogenetic scale are based on population-genetics equations, their estimates of selection coefficients are directly interpretable as fitness effects at the population scale; and because they work at the DNA level, we are able to account for mutational bias in DNA and structure of the genetic code. The model further integrates the shared evolutionary history between samples and their divergence, which, together, allow us to estimate fitness effects in mammalian phylogenetic trees even though sequences are not independent samples and might not represent the equilibrium distribution of amino acids (see section 1.2 in Materials & Methods, model parameters in section S1).

Theoretically, we can thus use large-scale genomic data to assess whether the fitness estimated at the phylogenetic scale predict the fitness effect at the population scale for both deleterious and beneficial back-mutations. The placental mammals represent an excellent study system to perform such analysis. Having originated ~102 million years ago, they diversified quickly (Foley et al., 2023). Additionally, polymorphism data are available for many species (Howe et al., 2021), as are high quality protein-coding DNA alignments across the genome (Ranwez et al., 2007; Scornavacca et al., 2019). By performing our analysis on 14,509 orthologous protein-coding genes across 87 species, we focus on genes shared across all mammals in our dataset and not newly functionalized genes in a lineage.

For each gene, fitting the mutation-selection model to the multi-species sequence alignment, assuming that the underlying fitness landscape is stable along the phylogenetic tree, allows us to obtain relative fitnesses for all amino acids for each site of the alignment (fig. 2A). Given these relative fitnesses (F) for each amino acid, the difference in fitness between a pair of amino acids is the scaled selection coefficient ($S_0 = \Delta F$), which is formally the product of the selection coefficient at the individual level (s) and the effective population size (N_e), as $S_0 = 4N_e \times s$. The value of S_0 informs us on the strength of selection exerted on amino acids changes. Thus, according to its S_0 value, we can classify any mutation as either a deleterious mutation toward a less fit

amino acid ($\mathcal{D}_0 \coloneqq S_0 < -1$), a nearly-neutral mutation ($\mathcal{N}_0 \coloneqq -1 < S_0 < 1$) or a beneficial back-mutation toward a known fitter amino acid ($\mathcal{B}_0 \coloneqq S_0 > 1$).

Having identified which potential DNA changes represent beneficial back-mutations (fig. 2B), we retrieved polymorphism data from 28 wild and domesticated populations belonging to 6 genera (*Equus*, *Bos*, *Capra*, *Ovis*, *Chlorocebus*, and *Homo*) to assess the presence of beneficial back-mutations at the population scale. We focused on mutations currently segregating within populations and substitutions in the terminal branches, and checked if any of these observed changes were beneficial back-mutations (fig. 2C-E). A similar approach demonstrated the presence of beneficial back-mutations in humans (Moses and Durbin, 2009; Fischer et al., 2011) and in plants (Chen et al., 2021). However, the model used to reconstruct the static fitness landscape in these studies can only be applied to deeply conserved protein domains in the tree of life, which corresponds to a subpart of the proteome that evolves slowly. The mutation-selection model used in the present work integrates phylogenetic relationships, and thus allows us to estimate the fitness landscape in shallower phylogenetic trees, and therefore can be applied almost genome-wide (Rodrigue et al., 2010).

We first quantified the likelihood of any DNA mutation to be a beneficial back-mutation, that is, whenever a DNA mutation increases fitness under a stable fitness landscape ($\mathcal{B}_0$). Subsequently, by quantifying the total amount of beneficial mutations in the current population across all types of DNA mutations ($\mathcal{D}_0$, $\mathcal{D}_0$ and $\mathcal{B}_0$), we could tease apart beneficial back-mutations from beneficial adaptive mutations resulting from a change in the fitness landscape. Altogether, in this study, by integrating large-scale genomic datasets at both phylogenetic and population scales, we propose a way to explicitly quantify the contribution of beneficial back-mutations to positive selection across the entire exome of the six genera (figs. 2F-G).

Results

Selection along the terminal branches

First, we assessed whether fitness effects derived from the mutation-selection model at the phylogenetic scale predict selection occurring in terminal branches. We recovered the mutations that reached fixation in the terminal branches of the six genera. We only considered mutations fixed in a population as substitutions in the corresponding branch by discarding mutations segregating in our population samples. We classified each substitution identified in the terminal branches as either $\mathcal{D}_0$, $\mathcal{N}_0$, or $\mathcal{B}_0$ depending on its S_0 value obtained at the phylogenetic scale (fig. 2A-C). Importantly, the mammalian alignment used to estimate the amino-acid fitness landscape did not include the genera for which we estimated selection at the population scale as this would bias the estimate of the fitness landscape and generate circular reasoning. Because S_0 was based on estimating fitness effects at the phylogenetic scale for all mammals, substitutions with $S_0 > 1$ ($\mathcal{B}_0$) bring the population toward an amino acid predicted to be fitter in mammals. These changes are thus considered as beneficial back-substitutions rather than a species-specific adaptive change. Among all the substitutions found in each terminal branch, between 10 and 13% were $\mathcal{B}_0$, while instead beneficial back-mutations represent between 0.9 and 1.2% of all non-synonymous mutations (table S1 and fig. 3A-B for humans). In principle, beneficial back-mutations are bound to reach fixation more often than neutral mutations. Hence, we calculated the d_N/d_S ratio of non-synonymous over synonymous divergence for all terminal lineages,

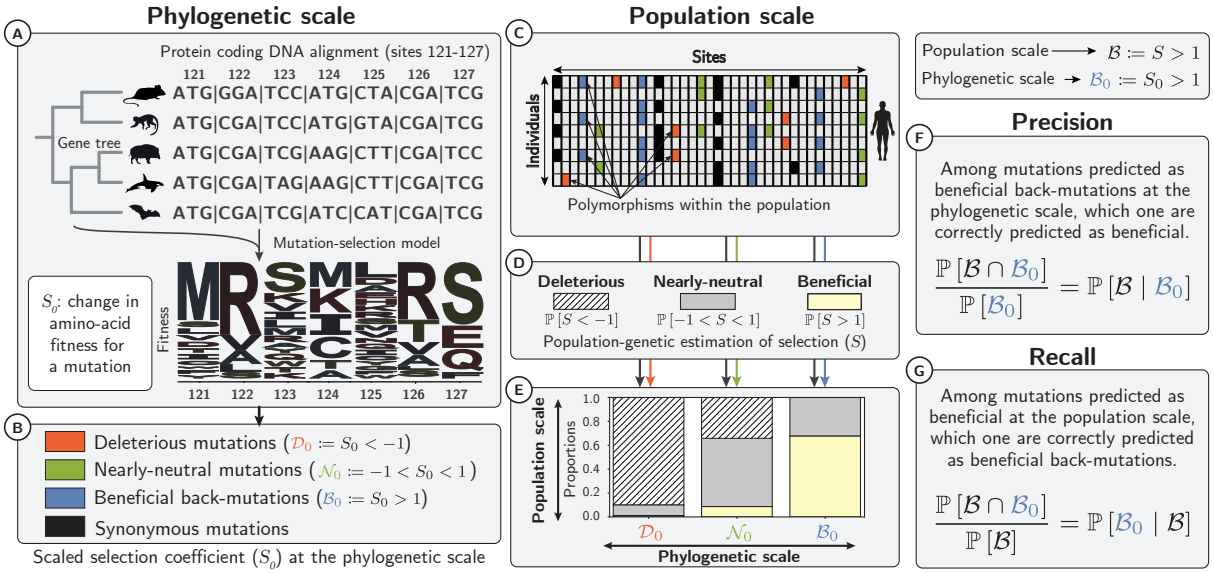


Figure 2: Selection coefficients at the phylogenetic and population scales. At the phylogenetic scale (A), we estimated the amino-acid fitness for each site from protein-coding DNA alignments using mutation-selection codon models. For every possible mutation, the difference in amino-acid fitness before and after the mutation allows us to compute the selection coefficient at the phylogenetic scale (S_0). Depending on S_0 (B) mutations can be predicted as deleterious ($\mathcal{D}_0$), nearly-neutral ($\mathcal{N}_0$) or beneficial back-mutations ($\mathcal{B}_0$) toward a fitter amino acid and repairing existing functions. At the population scale, each observed single nucleotide polymorphism (SNP) segregating in the population can also be classified according to its S_0 value (C). Occurrence and frequency in the population of non-synonymous polymorphisms, contrasted to synonymous polymorphisms (deemed neutral) is used to estimate selection coefficients (D-E) at the population scale (S), for each class of selection ($\mathcal{D}_0$, $\mathcal{N}_0$, $\mathcal{B}_0$). We can thus assess whether S_0 predicts S and compute *precision* (F) and *recall* (G) for each class. The *recall* value for class $\mathcal{B}_0$ is the probability of back-mutations among all beneficial ones (G).

focusing on the non-synonymous changes predicted as beneficial back-mutations ($d_N(\mathcal{B}_0)/d_S$). We obtained values between 1.17 and 1.75 in the different lineages (table S2), implying i) that $\mathcal{B}_0$ mutations reached fixation more frequently than synonymous mutations that are supposed to be neutral, and ii) that these back-mutations are effectively beneficial.

This result further indicated that using d_N/d_S as an estimate of adaptation is biased due to the presence of beneficial back-mutations among the non-synonymous substitutions. By discarding all beneficial back-mutations we can obtain an estimate of d_N/d_S which is not inflated. By comparing these two ways of calculating d_N/d_S (see section 1.5 in Materials & Methods), we calculated that beneficial back-mutations inflate d_N/d_S values by between 9 and 12% across genera while only representing between 0.9 and 1.2% of non-synonymous mutations (table S3).

Selection in populations

Second, we assessed whether our calculated S_0 values predicted at the phylogenetic scale were also indicative of the selective forces exerted at the population level. We retrieved single nucleotide polymorphisms (SNPs) segregating in 28 mammalian populations. To determine if SNPs were ancestral of derived, we reconstructed the ancestral exome of each population. We then classified every non-synonymous SNP as either $\mathcal{D}_0$, $\mathcal{N}_0$, or $\mathcal{B}_0$ according its S_0 value (fig. 2B-C).

In humans, some SNPs have been associated with specific clinical prognosis terms obtained by clinical evaluation of the impact of variants on human Mendelian disorders (Landrum et al., 2018). Although this classification also relies on deep protein alignments and therefore cannot be considered an independent result from our own, it does provide a consistency check if the effect of a mutation on human health is in line with its fitness effect predicted by our method. Therefore, we investigated whether the non-synonymous SNPs classified as $\mathcal{D}_0$ or $\mathcal{B}_0$ showed enrichment in specific clinical terms compared to SNPs classified as $\mathcal{N}_0$. Our results show that SNPs predicted as deleterious are associated with clinical terms such as *Likely Pathogenic* and *Pathogenic*, implying that, in general, the selective pressure of a mutation exerted across mammals is also predictive of its clinical effect in humans (table S4) (Sullivan et al., 2023). Conversely, back-mutations are associated with clinical terms such as *Benign* and *Likely Benign*, which shows that back-mutations are less likely to be functionally damaging (table S5).

In addition to clinical prognosis, frequencies at which SNPs are segregating within populations provide information on their selective effects. For instance, deleterious SNPs usually segregate at lower frequencies because of purifying selection, which tends to remove them from the population (fig. 3C for humans). By gathering information across many SNPs, it is possible to estimate the distribution of fitness effects (DFE) at the population scale, taking synonymous SNPs as a neutral expectation (Eyre-Walker et al., 2006; Eyre-Walker and Keightley, 2009; Galtier, 2016; Tataru et al., 2017). From the estimated DFE, we can derive the proportion of beneficial mutations ($\mathbb{P}[\mathcal{B}]$), nearly-neutral mutations ($\mathbb{P}[\mathcal{N}]$) and deleterious mutations ($\mathbb{P}[\mathcal{D}]$) at the population scale (see section 1.6 in Materials & Methods). These approaches offer a unique opportunity to contrast selection coefficients estimated at the population scale (S) and at the phylogenetic scale (S_0).

Across our selection classes ($\mathcal{D}_0$, $\mathcal{N}_0$ and $\mathcal{B}_0$), one can ultimately estimate the proportion of correct and incorrect predictions, leading to an estimation of *precision* and *recall* (fig. 2F-G and section 1.7 in Materials & Methods). Across 28 populations of different mammal species, mutations predicted to be deleterious at the phylogenetic scale ($\mathbb{P}[\mathcal{D}]$) were indeed purged at the population scale, with a *precision* in the range of 90–97% (table 1 and fig. 3D for humans). Conversely, a *recall* in the range of 96–100% implied that mutations found to be deleterious at the population scale were most likely also predicted to be deleterious at the phylogenetic scale (table 1). Altogether, purifying selection is largely predictable and amino acids with negative fitness across mammals have been effectively purged away in each population.

Mutations predicted as $\mathcal{N}_0$ were effectively composed of a mix of neutral and selected mutations with varying *precision* (36–63%) and *recall* (32–45%) across the different populations (table 1, fig. 3D for humans). The variable proportions between populations can be explained by the effective number of individuals in the

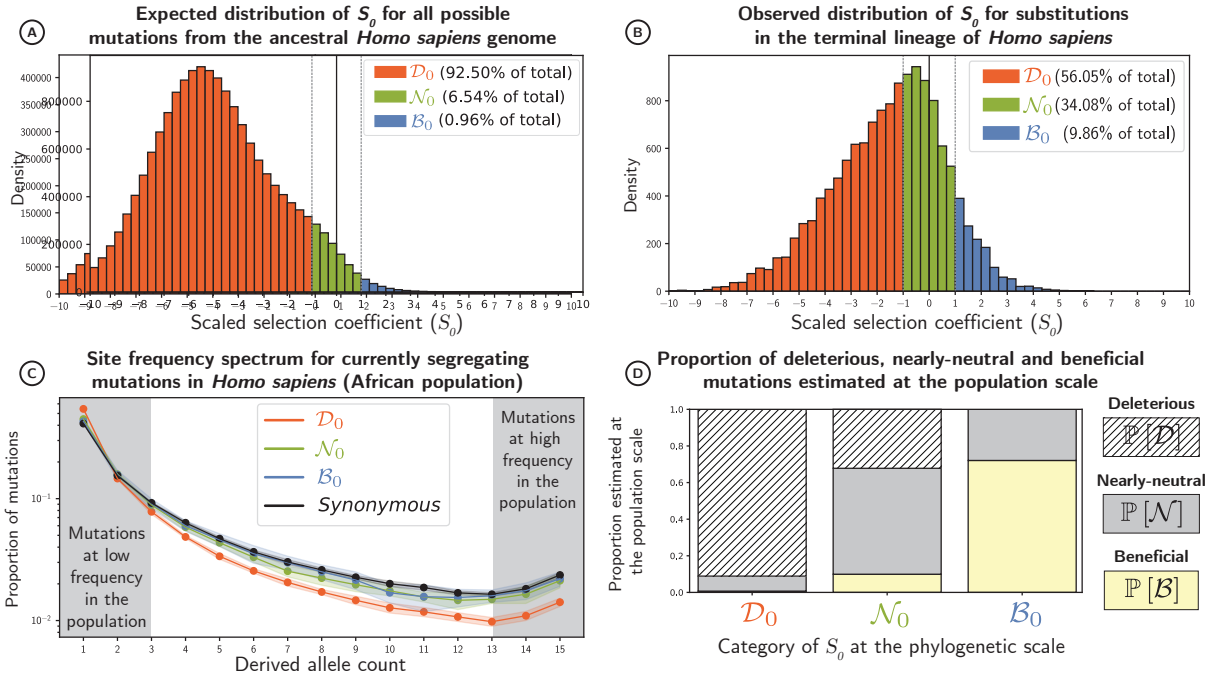


Figure 3: (A) Distribution of scaled selection coefficients (S_0), predicted for all possible non-synonymous DNA mutations away from the ancestral human exome (section 1.4). Mutations are divided into three classes of selection: deleterious ($\mathcal{D}_0$), nearly-neutral ($\mathcal{N}_0$) and beneficial ($\mathcal{B}_0$, supposedly beneficial back-mutations) (B) Distribution of scaled selection coefficients (S_0) for all observed substitutions along the *Homo* branch after the *Homo-Pan* split (section 1.5). If there are fewer substitutions than expected, this class is thus undergoing purifying selection, as is the case for $\mathcal{D}_0$. (C) The site-frequency spectrum (SFS) in humans of African descent for a random sample of 16 alleles (means in solid lines and standard deviations in color shades) for each class of selection and for synonymous mutations, supposedly neutral (black). The SFS represents the proportion of mutations (y-axis) with a given number of derived alleles in the population (x-axis). At high frequencies, deleterious mutations are underrepresented. (D) Proportion of beneficial $\mathbb{P}[\mathcal{D}]$, nearly-neutral $\mathbb{P}[\mathcal{N}]$, and deleterious mutations $\mathbb{P}[\mathcal{B}]$ estimated at the population scale for each class of selection at the phylogenetic scale (section 1.6). Proportions depicted here are not weighted by their mutational opportunities.

population (N_e), a major driver of selection efficacy. Moreover, estimates of mutation rate per generation (u), from Bergeron et al. (2023) and Orlando et al. (2013), and Watterson's θ obtained from the synonymous SFS as in Achaz (2009) allow us to obtain N_e through $N_e = \theta/(4 \times u)$. Using correlation analyses that accounted for phylogenetic relationship (see section 1.8 in Materials & Methods), we found that higher N_e was associated with a smaller proportion of nearly-neutral mutations ($r^2 = 0.31$, $p = 0.001$, fig. 4A). This result follows the prediction of the nearly-neutral theory and suggests that in populations with higher diversity (e.g., *Bos* or *Ovis*), discrimination between beneficial and deleterious mutations is more likely to occur (figs. S1-S3). Conversely, many more mutations are effectively neutral in populations with lower diversity (e.g., *Homo*).

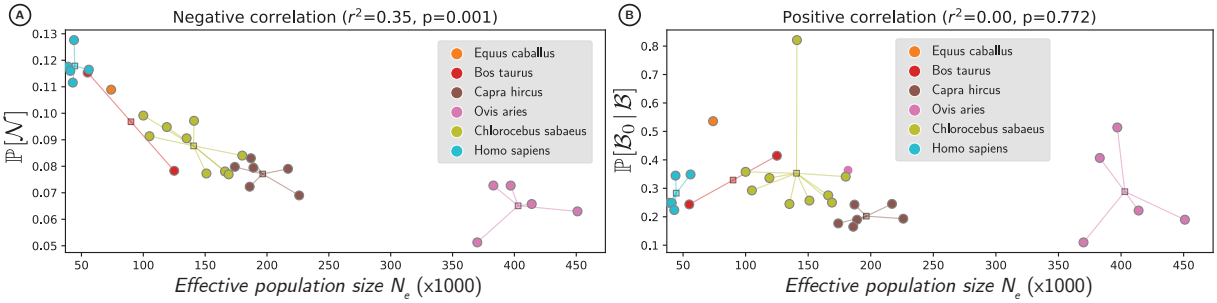


Figure 4: Populations in circles, mean of the species across the populations as squares. (A) Proportion of nearly-neutral mutations at the population scale ($\mathbb{P}[\mathcal{N}]$ in the y-axis), shown as a function of estimated effective population size (N_e in the x-axis). (B) Proportion of beneficial back-mutations among all beneficial mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ in the y-axis), shown as a function of N_e in x-axis. Correlations account for phylogenetic relationship and non-independence of samples, through the fit of a Phylogenetic Generalized Linear Model (see section 1.6 in Materials & Methods).

Finally, mutations predicted to be beneficial back-mutations based on the selection coefficients at the phylogenetic scale ($\mathcal{B}_0$) were indeed beneficial for individuals bearing them, with a *precision* (fig. 2F) in the range of 19–87% (table 1 and fig. 3D for humans). This result confirms that selection towards amino acids restoring existing functions is ongoing in these populations. Importantly, the *recall* value in this case, computed as $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$, is the probability for a beneficial mutation at the population scale to be a beneficial back-mutation, i.e., going toward a fitter amino acid (fig. 2G, table S6). In other words, the *recall* value quantifies the number of beneficial mutations restoring damaged genomes instead of creating adaptive innovations. Across the 28 populations, this proportion is in the range of 11–82% (table 1).

Because the phylogenetic mutation-selection codon model should fit better for genes with uniformly conserved functions, as a control, we filtered out genes under pervasive adaptation (Latrille et al., 2023), and indeed we found an increase in the proportion of beneficial back-mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$), consistent with our expectation (Wilcoxon signed-rank, $s = 80$, $p = 0.002$, table S7).

Discussion

Beneficial mutations are not necessarily adaptive

This study represents an essential step toward integrating the different evolutionary scales necessary to understand the combined effects of mutation, selection, and drift on genome evolution. In particular, we have been able to quantify the proportion of beneficial back-mutations among all beneficial mutations, which has only been achievable by combining genome-wide data from both phylogenetic and population scales. At the phylogenetic scale, codon diversity at each site of a protein-coding DNA alignment allows for reconstructing an amino-acid fitness landscape, assuming that this landscape is stable along the phylogenetic tree. These amino-acid fitness landscapes allow us to predict any mutation's selection coefficient (S_0) along a protein-coding sequence. We can compare these selective effects to observations at the population level (figs. S4-6).

			Deleterious mutations $\mathcal{D} \coloneqq S < -1$ $\mathcal{D}_0 \coloneqq S_0 < -1$		Nearly-neutral mutations $\mathcal{N} \coloneqq -1 < S < 1$ $\mathcal{N}_0 \coloneqq -1 < S_0 < 1$		Beneficial mutations $\mathcal{B} \coloneqq S > 1$ $\mathcal{B}_0 \coloneqq S_0 > 1$	
Population	**Species**	N_e	**Precision** $\mathbb{P}[\mathcal{D} \mid \mathcal{D}_0]$	**Recall** $\mathbb{P}[\mathcal{D}_0 \mid \mathcal{D}]$	**Precision** $\mathbb{P}[\mathcal{N} \mid \mathcal{N}_0]$	**Recall** $\mathbb{P}[\mathcal{N}_0 \mid \mathcal{N}]$	**Precision** $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$	**Recall** $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$
Equus c.	Equus caballus	7.5×10^4	0.923	0.972	0.570	0.341	0.648	0.536
Iran	Bos taurus	5.6×10^4	0.915	1.000	0.632	0.358	0.873	0.243
Uganda	Bos taurus	1.3×10^5	0.951	0.969	0.495	0.414	0.576	0.415
Australia	Capra hircus	1.7×10^5	0.944	0.971	0.527	0.437	0.368	0.177
France	Capra hircus	1.9×10^5	0.946	0.971	0.508	0.423	0.368	0.190
Iran (C. aegagrus)	Capra hircus	1.9×10^5	0.948	0.969	0.486	0.444	0.368	0.165
Iran	Capra hircus	2.3×10^5	0.953	0.966	0.425	0.407	0.368	0.193
Italy	Capra hircus	1.9×10^5	0.947	0.971	0.551	0.439	0.368	0.243
Morocco	Capra hircus	2.2×10^5	0.950	0.970	0.527	0.440	0.368	0.245
Iran	Ovis aries	3.8×10^5	0.961	0.961	0.452	0.415	0.205	0.407
Iran (O. orientalis)	Ovis aries	4.5×10^5	0.964	0.960	0.420	0.445	0.193	0.190
Iran (O. vignei)	Ovis aries	3.7×10^5	0.967	0.959	0.361	0.470	0.190	0.110
Various	Ovis aries	4.1×10^5	0.962	0.962	0.433	0.440	0.229	0.222
Morocco	Ovis aries	4×10^5	0.962	0.961	0.462	0.424	0.211	0.514
Barbados	Chlorocebus sabaeus	1.1×10^5	0.935	0.975	0.565	0.402	0.648	0.293
Central Afr. Rep.	Chlorocebus sabaeus	1.7×10^5	0.948	0.971	0.508	0.423	0.535	0.275
Ethiopia	Chlorocebus sabaeus	1.4×10^5	0.935	0.975	0.580	0.416	0.552	0.245
Gambia	Chlorocebus sabaeus	1.4×10^5	0.944	0.975	0.654	0.437	0.577	0.821
Kenya	Chlorocebus sabaeus	1.5×10^5	0.946	0.972	0.538	0.453	0.588	0.257
Nevis	Chlorocebus sabaeus	1×10^5	0.933	0.976	0.629	0.412	0.599	0.358
South Africa	Chlorocebus sabaeus	1.8×10^5	0.944	0.971	0.548	0.423	0.574	0.341
Saint Kitts	Chlorocebus sabaeus	1.2×10^5	0.936	0.975	0.586	0.402	0.598	0.336
Zambia	Chlorocebus sabaeus	1.7×10^5	0.945	0.971	0.512	0.432	0.585	0.250
African	Homo sapiens	5.6×10^4	0.911	0.976	0.579	0.325	0.721	0.349
Admixed American	Homo sapiens	4.5×10^4	0.902	0.978	0.584	0.299	0.690	0.345
East Asian	Homo sapiens	4×10^4	0.905	0.978	0.585	0.325	0.688	0.249
European	Homo sapiens	4.2×10^4	0.906	0.978	0.584	0.329	0.688	0.248
South Asian	Homo sapiens	4.4×10^4	0.908	0.978	0.584	0.342	0.691	0.224

Table 1: *Precision* and *recall* for estimated selection coefficient of mutations given by mutation-selection models (S_0). *Precision* is the estimation of the selection coefficient at population scale (S) given that S_0 is known. Conversely, *recall* is the estimation of S_0 given selection coefficient at the population scale (S) is known. *Recall* for beneficial mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$) is thus the proportion of beneficial back-mutations among all beneficial mutations. N_e is the estimated effective population size for each population.

By doing so, we confirmed that mutations predicted to be deleterious ($\mathcal{D}_0 \coloneqq S_0 < -1$) are purified away in extant populations. Our results concur with previous studies showing that SIFT scores (Ng and Henikoff, 2003; Vaser et al., 2016), based on amino acid alignments across species, also inform on the deleterious fitness effects exerted at the population scale (Chen et al., 2021). However – contrary to SIFT scores – our mutation-selection model is parameterized by a fitness function such that changes are directly interpretable as fitness effects. In this regard, an interesting prediction of our model is that back-mutations are beneficial because they revert previous deleterious changes. We have tested this hypothesis and found that these back-mutations ($\mathcal{B}_0 \coloneqq S_0 > 1$) are indeed beneficial in extant populations. We estimated that between 11 and 82% of all beneficial mutations in mammalian populations have not been driven by adaptation but instead have reversed deleterious substitutions which have accumulated along the phylogenetic tree. More specifically, in 24 out of 28 populations analyzed, the percentage of beneficial mutations estimated to be back-mutations falls between 15 and 45%. These results suggest that many beneficial mutations are non-adaptive but rather restore ancestral states of higher fitness. Hence, we can correctly estimate the extent of adaptive evolution only if we account for the number of beneficial back-mutations (Keightley and Eyre-Walker, 2010; Rice et al., 2015). Altogether, we argue that we should dissociate positive selection from adaptive evolution and

limit the use of adaptive mutations to those that are associated with adaptation to environmental change as such (Charlesworth and Eyre-Walker, 2007; Mustonen and Lässig, 2009).

Interpreting the proportion of beneficial back-mutations

Across the genome, beneficial-back mutations and deleterious mutations reaching fixation create a balance in which genomes are constantly damaged and restored simultaneously at different loci. Since the probability of fixation of mutations depends on the effective population size (N_e), the history of N_e plays a crucial role in setting the number of beneficial back-mutations compensating for deleterious mutations (Latrille et al., 2021). For example, a population size expansion will increase the efficacy of selection, and a larger proportion of mutations will be beneficial (otherwise effectively neutral), thus increasing the number of beneficial back-mutations. On the other hand, a population that has experienced a high N_e throughout its history should be closer to an optimal state under a stable fitness landscape, having suffered fewer fixations of deleterious mutations and therefore decreasing the probability of beneficial back-mutations (Huber et al., 2017). Overall, we expect the proportion of beneficial back-mutations to be more dependent on N_e's long-term expansions and contractions than on the short-term ones (Charlesworth and Eyre-Walker, 2007; Huber et al., 2017). Accounting for phylogenetic relationships, we found no correlation between the proportion of back-mutations and N_e ($r^2 = 0.00$, $p = 0.772$, fig. 4B)

The exact estimation of the contribution of beneficial back-mutations to positive selection relies on some hypotheses at both the phylogenetic and population scales and is sensitive to methodological limitations. Indeed, to be conservative, we considered that any mutation under positive selection at the population scale ($\mathcal{B}$) but not predicted as such at the phylogenetic scale (not $\mathcal{B}_0$) is potentially an adaptation. However, adaptation is not the only factor hindering the detection of beneficial back-mutations; data quality and potentially inadequate modeling choices of both the fitness landscape and the DFE might also lead to missed predictions.

Our model assumes that amino-acid fitness landscapes are site-specific and also independent of one another, whereas under pervasive epistasis, the fitness effect of any mutation at a particular site would depend on the amino acids present at other sites. Epistasis has been shown to play a role in the evolution of protein-coding genes, with amino-acid residues in contact within a protein or between proteins tending to co-evolve (Morcos et al., 2011; Marks et al., 2012; Starr and Thornton, 2016). Particularly, the residues in contact co-evolve to become more compatible with each other generating an entrenchment (Goldstein et al., 2015; Goldstein and Pollock, 2017; Park et al., 2022). Epistasis therefore allows for compensatory mutations, which restore fitness through mutations at loci different from where deleterious mutations took place. It therefore, represents another case case of non-adaptive beneficial mutations, but one which is not accounted for by our method. Therefore, the beneficial mutations that we classify as putatively adaptive might in fact be compensatory mutations, making our estimation of the rate of non-adaptive beneficial mutations conservative.

Despite epistasis being an important factor in protein evolution, several deep mutational scanning laboratory experiments have revealed that a site-specific fitness landscape is a reasonable approximation of the

fitness landscape of proteins, and even predicts the evolution of sequences in nature (Ashenberg et al., 2013; Doud et al., 2015; Bloom, 2017). Additionally, the fact that we observe such a high proportion of beneficial back-mutations suggests that the underlying assumptions of our model, namely site-independence, implying no epistasis, and a static fitness landscape, are a reasonable approximation for the underlying fitness landscape of proteins. Our results imply that the fitness effects of new mutations are mostly conserved across mammalian orthologs, in agreement with other studies showing that for conserved orthologs with similar structures and functions, models without epistasis provide a reasonable estimate of fitness effects in protein coding genes (Youssef et al., 2020; Vigué et al., 2022).

Moreover, because our model assumes a fixed landscape, it implicitly assumes that N_e is constant along the phylogenetic tree. Fluctuations due to changes in the fitness landscape or in N_e will be averaged out by the assumption of the current model that N_e is constant across lineages. It was recently shown (Latrille et al., 2021), using highly computer intensive mutation-selection models with fluctuating N_e, that relaxing the assumption of a constant N_e results in more extreme estimates of amino-acid fitnesses than with the standard model used in this study. In other words, by assuming a constant N_e, we are underpowered to detect beneficial back-mutations since amino acids will have more similar fitnesses. As a consequence, mutations that would otherwise fall into $\mathcal{B}$ or $\mathcal{D}$ will be classified as $\mathcal{N}$, resulting ultimately in lower estimates of the proportion of beneficial back-mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$). Given this inflation of missed predictions due to change in population sizes (Lanfear et al., 2014; Jones et al., 2017; Platt et al., 2018), our estimated proportion of beneficial back-mutations among adaptive ones is likely to therefore be an underestimation. In practice, advantageous mutations can be depleted in the $\mathcal{N}_0$ and $\mathcal{D}_0$ class in some species but pervasive in others, depending on the model used to infer the DFE at the population scale (tables S8-12). It appears that our estimation can be quite sensitive to model misspecification and overall, while we provide an order of magnitude for the contribution of beneficial back-mutations to positive selection, further methodological development on the estimation of the distribution of fitness effects is needed to increase the confidence on this value.

Detecting adaptation above the nearly-neutral background

A long-standing debate in molecular evolution is whether the variations we observe between species in protein-coding genes are primarily due to nearly-neutral mutations reaching fixation by drift or primarily due to adaptation (Kimura, 1968; Jensen et al., 2019; Gillespie, 1994; Ohta, 1992). Here we provide evidence that in mammalian orthologs, many substitutions occur through fixation of both deleterious mutations and beneficial back-mutations. However, detecting adaptation above this background of nearly-neutral substitutions remains a central question (Kimura, 1968; Ohta and Gillespie, 1996).

One first strategy is precisely to use a nearly-neutral substitution model as a null model of evolution. Under a strictly neutral evolution of protein-coding sequence, we expect the ratio of non-synonymous over synonymous substitutions (d_N/d_S) to be equal to one. Deviations from this neutral expectation, such as $d_N/d_S > 1$, which can be generated by an excess of non-synonymous substitutions, is generally interpreted as a sign of adaptation. However, as shown in this study, a $d_N/d_S > 1$ is not necessarily a signature of adaptation but can be due to beneficial back-mutations. So, by relaxing the strict neutrality and assuming

a stable fitness landscape instead, one can predict the expected rate of evolution, called ω_0 (Spielman and Wilke, 2015; Dos Reis, 2015). Adaptation can thus be considered as evolution under a changing fitness landscape and tested as such by searching for the signature of $d_N/d_S > \omega_0$ (Cvijović et al., 2015; Rodrigue and Lartillot, 2017; Rodrigue et al., 2021). Using a stable fitness landscape as a null model of evolution, thus accounting for selective constraints exerted on the different amino acids, increased the statistical power in testing for adaptation (Latrille et al., 2023). Instead of relying solely on summary statistics (such as d_N/d_S or ω_0), another strategy to detect adaptation is to include changes in the fitness landscapes inherently within the mutation-selection framework (Tamuri and dos Reis, 2021). Such mechanistic models could be more general than site-specific fitness landscapes, including epistasis and changing fitness landscapes (Goldstein and Pollock, 2017; Stolyarova et al., 2020).

Here, we have provided empirical evidence that an evolutionary model assuming a stable fitness landscape at the mammalian scale allows us to predict the fitness effects of mutations in extant populations and individuals, acknowledging the balance between deleterious and beneficial back-mutations. We argue that such a model represents a null expectation for the evolution of protein-coding genes in the absence of adaptation. In that sense, to avoid conflating beneficial mutations with adaptive evolution, the term "adaptation" should retain its original meaning associated with a change in the underlying fitness landscape and be modelled as such.

1 Materials & Methods

1.1 Phylogenetic dataset

Protein-coding DNA sequence alignments in placental mammals and their corresponding gene trees come from the OrthoMaM database and were processed as in Latrille et al. (2023). OrthoMaM contains a total of 116 mammalian reference sequences in v10c (Ranwez et al., 2007; Douzery et al., 2014; Scornavacca et al., 2019).

Genes located on the X and Y chromosomes and on the mitochondrial genome were discarded from the analysis because the level of polymorphism – which is necessary for population-based analyses – is expected to be different in these three regions compared to the autosomal genome. Sequences of species for which we used population-level polymorphism (see section 1.3) and their sister species, were removed from the analysis to ensure independence between the data used in the phylogenetic and population scales. Sites in the alignment containing more than 10% of gaps across the species were discarded. Altogether, our genome-wide dataset contains $14,509$ protein-coding DNA sequences in 87 placental mammals.

1.2 Selection coefficient (S_0) in a phylogeny-based method

We analyzed the phylogenetic-level data using mutation-selection models. These models assume the protein-coding sequences are at mutation-selection balance under a fixed fitness landscape characterized by a fitness vector over the 20 amino acids at each site (Yang and Nielsen, 2008; Halpern and Bruno, 1998; Rodrigue and Philippe, 2010). Mathematically, the rate of non-synonymous substitution from codon a to codon b

($q^{(i)}_{a\mapsto b}$) at site i of the sequence is equal to the rate of mutation of the underlying nucleotide change ($\mu_{a\mapsto b}$) multiplied by the scaled probability of mutation fixation ($\mathbb{P}^{(i)}_{a\mapsto b}$). The probability of fixation depends on the difference between the scaled fitness of the amino acid encoded by the mutated codon ($F^{(i)}_b$) and the amino acid encoded by the original codon ($F^{(i)}_a$) at site i (Wright, 1931; Fisher, 1930).

The rate of substitution from codon a to b at a site i is thus:

$$\begin{cases} q^{(i)}_{a\mapsto b} & = 0 \text{ if codons } a \text{ and } b \text{ are more than one mutation away,} \\ q^{(i)}_{a\mapsto b} & = \mu_{a\mapsto b} \text{ if codons } a \text{ and } b \text{ are synonymous, and} \\ q^{(i)}_{a\mapsto b} & = \mu_{a\mapsto b} \dfrac{F^{(i)}_b - F^{(i)}_a}{1 - \mathrm{e}^{F^{(i)}_a - F^{(i)}_b}} \text{ if codons } a \text{ and } b \text{ are non-synonymous.} \end{cases} \tag{1}$$

Fitting the mutation-selection model on a multi-species sequence alignment leads to an estimation of the gene-wide 4×4 nucleotide mutation rate matrix ($\boldsymbol{\mu}$) as well as the 20 amino-acid fitness landscape ($\boldsymbol{F^{(i)}}$) at each site i. The priors and full configuration of the model are given in section S1. From a technical perspective, the Bayesian estimation is a two-step procedure (Rodrigue et al., 2008). The first step is a data augmentation of the alignment, consisting in sampling a detailed substitution history along the phylogenetic tree for each site, given the current value of the model parameters. In the second step, the parameters of the model can then be directly updated by a Gibbs sampling procedure, conditional on the current substitution history. Alternating between these two sampling steps yields a Markov chain Monte-Carlo (MCMC) procedure whose equilibrium distribution is the posterior probability density of interest (Lartillot, 2004; Rodrigue et al., 2008). Additionally, across-site heterogeneities in amino-acid fitness profiles are captured by a Dirichlet process. More precisely, the number of amino-acid fitness profiles estimated is lower than the number of sites in the alignment. Consequently each profile has several sites assigned to it, resulting in a particular configuration of the Dirichlet process. Conversely, sites with similar signatures are assigned to the same fitness profile. This configuration of the Dirichlet process is resampled through the MCMC to estimate a posterior distribution of amino acid profiles for each site specifically (Rodrigue et al., 2010; Lartillot, 2013). From a more mechanistic perspective, even though not all amino acids occur at every single codon site of the DNA alignment, we can nevertheless estimate the distribution of amino-acid fitnesses by generalizing the information recovered across sites and across amino acids based on the phylogenetic relationship among samples. In particular, synonymous substitutions along the tree contain the signal to estimate branch lengths and the nucleotide transition matrix, while non-synonymous substitutions contain information on fitness difference between codons connected by single nucleotide changes (Rodrigue et al., 2010).

The selection coefficient for a mutation from codon a to codon b at site i is defined as:

$$S^{(i)}_0(a \mapsto b) = \Delta F^{(i)} = F^{(i)}_b - F^{(i)}_a. \tag{2}$$

In our subsequent derivation the source (a) and target (b) codons as well as the site (i) are implicit and thus never explicitly written.

We used the Bayesian software *BayesCode* (`https://github.com/ThibaultLatrille/bayescode`, v1.3.1) to estimate the selection coefficients for each protein-coding gene in the mammalian dataset. We ran the MCMC algorithm implemented in BayesCode for $2,000$ generations as described in Latrille et al.

(2023). For each gene, after discarding a burn-in period of 1,000 generations of MCMC, we obtained posterior mean estimates (over the 1,000 generations left of MCMC) of the mutation rate matrix ($\boldsymbol{\mu}$) as well as the 20 amino-acid fitness landscape ($\boldsymbol{F^{(i)}}$) at each site i.

1.3 Polymorphism dataset

The genetic variants representing the population level polymorphisms were obtained from the following species and their available datasets: *Equus caballus* (EquCab2 assembly in the EVA study PRJEB9799 (Al Abri et al., 2020)), *Bos taurus* (UMD3.1 assembly in the NextGen project: `https://projects.ensembl.org/nextgen/`), *Ovis aries* (Oar_v3.1 assembly in the NextGen project), *Capra hircus* (CHIR1 assembly in the NextGen project), converted to ARS1 assembly with dbSNP identifiers (Sherry et al., 2001)), *Chlorocebus sabaeus* (ChlSab1.1 assembly in the EVA project PRJEB22989 (Svardal et al., 2017)), *Homo sapiens* (GRCh38 assembly in the 1000 Genomes Project (Zheng-Bradley et al., 2017)). In total, we analyzed 28 populations across the 6 different species with polymorphism data. The data was processed as described in Latrille et al. (2023).

Only bi-allelic single nucleotide polymorphisms (SNPs) found within a gene were in our polymorphism dataset, while nonsense variants and indels were discarded. To construct the dataset, we first recovered the location of each SNP (represented by its chromosome, position, and strand) in the focal species and matched it to its corresponding position in the coding sequence (CDS) using gene annotation files (GTF format) downloaded from Ensembl (`ensembl.org`). We then verified that the SNP downloaded from Ensembl matched the reference in the CDS in FASTA format. Next, the position in the CDS was converted to the corresponding position in the multi-species sequence alignment (containing gaps) from the OrthoMaM database (see section 1.2) for the corresponding gene by doing a global pairwise alignment (Biopython function pairwise2). This conversion from genomic position to alignment position was only possible when the assembly used for SNP-calling was the same as the one used in the OrthoMaM alignment, the GTF annotations, and the FASTA sequences. SNPs were polarized using the three closest outgroups found in the OrthoMaM alignment with est-usfs v2.04 (Keightley and Jackson, 2018), and alleles with a probability of being derived lower than 0.99 were discarded.

1.4 Mutational opportunities

The mutational opportunities of any new mutation refer to its likelihood of falling into a specific category (synonymous, deleterious, nearly-neutral, or beneficial). Deriving such opportunities is necessary to estimate the strength of selection exerted at the population scale since different categories might have different mutational opportunities, and thus polymorphism and divergence need to be corrected accordingly (see sections 1.5, 1.6, and 1.7). To calculate mutational opportunities, we reconstructed the ancestral exome of each of the 28 populations by using the most likely ancestral state from est-usfs (see section 1.3), which differs from the corresponding species reference exome since it accounts for the variability present in the specific population.

From the reconstructed ancestral exome, all possible mutations were computed, weighted by the instantaneous rate of change between nucleotides obtained from the mutation rate matrix ($\boldsymbol{\mu}$, see section 1.2), summing to μ_{tot} across the whole exome, and to μ_{syn} when restricted to synonymous mutations. Finally, the mutational opportunities for synonymous mutations were computed as the total number of sites across the exome (L_{tot}) weighted by the proportion of synonymous mutations among all possible mutations as:

$$L_{\text{syn}} = L_{\text{tot}} \frac{\mu_{\text{syn}}}{\mu_{\text{tot}}}. \tag{3}$$

Similarly, for non-synonymous mutations, the total mutation rate for each class of selection $x \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}$, called $\mu(x)$, was estimated as the sum across all non-synonymous mutations if their selection coefficient at the phylogenetic scale is in the class $S_0 \in x$. Accordingly, the mutational opportunities ($L(x)$) for each class of selection coefficient (x) was finally computed as the total number of sites across the exome (L_{tot}) weighted by the ratio of the aggregated mutations rates falling in the class $\mu(x)$:

$$L(x) = L_{\text{tot}} \frac{\mu(x)}{\mu_{\text{tot}}}. \tag{4}$$

Finally, $\mathbb{P}[x]$ is the probability for a non-synonymous mutation to be in the class x, thus computed as:

$$\mathbb{P}[x] = \frac{L(x)}{\sum_{y \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}} L(y)}. \tag{5}$$

1.5 Substitution mapping and d_N/d_S in the terminal branch

We inferred the protein-coding DNA sequences for each node of the 4-taxa tree containing the focal species and the three closest outgroups species found in the OrthoMaM alignment by applying the M5 codon model (gamma site rate variation) as implemented in FastML.v3.11 (Ashkenazy et al., 2012). Consequently, for each focal species we reconstructed the protein coding DNA sequence of the whole exome at the base of the terminal branch before the split from the sister species. We considered *Ceratotherium simum simum* as *Equus caballus'* sister species; *Bison bison bison* as *Bos taurus'* sister species; *Pantholops hodgsonii* as *Ovis aries'* sister species; *Pantholops hodgsonii* as *Capra hircus'* sister species; *Macaca mulatta* as *Chlorocebus sabaeus'* sister species and finally, we considered *Pan troglodytes* as *Homo sapiens'* sister species. From this reconstructed exome, we determined the direction of the substitution occurring along the terminal branch of the phylogenetic tree toward each extant population. SNPs segregating in the population were discarded, and the most likely ancestral state from est-usfs (see section 1.3) was used as the reference for each extant population. For each substitution, we recovered its S_0 value as calculated through the phylogeny-based method (see section 1.2). Finally, the rate of non-synonymous over synonymous substitutions for a given class of selection coefficient ($x \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}$) was computed as:

$$\begin{cases} d_N(x) & = \dfrac{D(x)}{L(x)}, \\ d_S & = \dfrac{D_{\text{syn}}}{L_{\text{syn}}}, \end{cases} \tag{6}$$

where $D(x)$ was the number of non-synonymous substitutions in class x, D_{syn} was the number of synonymous substitutions across the exome, while $L(x)$ and L_{syn} were the numbers of non-synonymous and synonymous

mutational opportunities, respectively, as defined in section 1.4. $\delta(d_N/d_S)$ was computed as the difference between d_N/d_S computed over all substitutions and d_N/d_S when we removed beneficial back-mutations $d_N(S_0 < 1)/d_S$, normalized by d_N/d_S. Note that the quantities $\delta(d_N/d_S)$ and $\delta(d_N)$ are equivalent due to the simplification of the factor d_S:

$$\delta(d_N/d_S) = \frac{d_N/d_S - d_N(S_0 < 1)/d_S}{d_N/d_S} = \frac{d_N - d_N(S_0 < 1)}{d_N} = \delta(d_N). \tag{7}$$

1.6 Scaled selection coefficients (S) in a population-based method

To obtain a quantitative estimate of the distribution of selection coefficients for each category of SNPs, we used the polyDFE model (Tataru et al., 2017; Tataru and Bataillon, 2020). This model uses the count of derived alleles to infer the distribution of fitness effects (DFE). The probability of sampling an allele at a given frequency (before fixation or extinction) is informative of its scaled selection coefficient at the population scale (S). Therefore, pooled across many sites, the site-frequency spectrum (SFS) provides information on the underlying S of mutations. However, estimating a single S for all sampled mutations is biologically unrealistic, and a DFE of mutations is usually assumed (Eyre-Walker et al., 2006; Eyre-Walker and Keightley, 2009). The polyDFE(Tataru et al., 2017; Tataru and Bataillon, 2020) software implements a mixture of a Γ and exponential distributions to model the DFE of non-synonymous mutations, while synonymous mutations are considered neutral. The model estimates the parameters β_d, b, p_b and β_b for non-synonymous mutations as:

$$\phi\left(S; \beta_d, b, p_b, \beta_b\right) = \begin{cases} \left(1 - p_b\right) f_\Gamma(-S; -\beta_d, b) & \text{if } S \leq 0, \\ p_b f_e(S; \beta_b) & \text{if } S > 0, \end{cases} \tag{8}$$

where $\beta_d \leq -1$ is the estimated mean of the DFE for $S \leq 0$; $b \geq 0.2$ is the estimated shape of the Γ distribution; $0 \leq p_b \leq 1$ is the estimated probability that $S > 0$; $\beta_b \geq 1$ is the estimated mean of the DFE for $S > 0$; and $f_\Gamma(S; m, b)$ is the density of the Γ distribution with mean m and shape b, while $f_e(S; m)$ is the density of the exponential distribution with mean m.

PolyDFE requires one SFS for non-synonymous mutations and one for synonymous mutations (neutral expectation), as well as the number of sites on which each SFS was sampled. For populations containing more than 8 individuals, the SFS was subsampled down to 16 chromosomes (8 diploid individuals) without replacement (hyper-geometric distribution) to alleviate the effect of different sampling depths in the 28 populations. Altogether, for each class of selection ($x \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}$) of non-synonymous SNPs, we aggregated all the SNPs in the selection class x as an SFS. The number of sites on which each SFS was sampled is given by $L(x)$ for the non-synonymous SFS and L_{syn} for the synonymous SFS respectively. For each class of selection x, once fitted to the data using maximum likelihood with polyDFE, the parameters of the DFE $(\beta_d, b, p_b, \beta_b)$ were used to compute $\mathbb{P}[\mathcal{D} \mid x]$, $\mathbb{P}[\mathcal{N} \mid x]$, and $\mathbb{P}[\mathcal{B} \mid x]$ as:

$$\mathbb{P}[\mathcal{D} \mid x] = \mathbb{P}[S < -1 \mid x] = (1 - p_b) \int_1^{+\infty} f_\Gamma(S; -\beta_d, b) \mathrm{d}S, \tag{9}$$

$$\mathbb{P}[\mathcal{N} \mid x] = \mathbb{P}[-1 < S < 1 \mid x] = p_b \int_0^1 f_e(S; \beta_b) \mathrm{d}S + (1 - p_b) \int_0^1 f_\Gamma(S; -\beta_d, b) \mathrm{d}S, \tag{10}$$

$$\mathbb{P}[\mathcal{B} \mid x] = \mathbb{P}[S > 1 \mid x] = p_b \int_1^{+\infty} f_e(S; \beta_b) \mathrm{d}S. \tag{11}$$

1.7 *Precision* and *recall*

For readability, we give here *precision* and *recall* for beneficial mutations ($\mathcal{B}_0$ and $\mathcal{B}$), but it can be obtained using the same derivation for the deleterious mutations ($\mathcal{D}_0$ and $\mathcal{D}$) and nearly-neutral mutations ($\mathcal{N}_0$ and $\mathcal{N}$).

Precision is the proportion of mutations correctly predicted as beneficial ($\mathbb{P}[\mathcal{B} \cap \mathcal{B}_0]$) out of all predicted as beneficial back-mutations ($\mathbb{P}[\mathcal{B}_0]$), which can be written as a conditional probability:

$$\frac{\mathbb{P}[\mathcal{B} \cap \mathcal{B}_0]}{\mathbb{P}[\mathcal{B}_0]} = \mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]. \tag{12}$$

Namely, *precision* corresponds to the probability for a back-mutation ($\mathcal{B}_0$) to be effectively beneficial at the population level ($\mathcal{B}$). This probability, computed from eq. 11, is obtained by restricting our analysis to SNPs that are predicted to be beneficial back-mutations (yellow fill for the category $\mathcal{B}_0$ in fig. 3D).

Recall is the proportion of mutations correctly predicted as beneficial ($\mathbb{P}[\mathcal{B} \cap \mathcal{B}_0]$) out of all beneficial mutations ($\mathbb{P}[\mathcal{B}]$), which can be written as a conditional probability:

$$\frac{\mathbb{P}[\mathcal{B} \cap \mathcal{B}_0]}{\mathbb{P}[\mathcal{B}]} = \mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]. \tag{13}$$

Namely, *recall* corresponds to the probability for a beneficial mutation at the population level ($\mathcal{B}$) to be a beneficial back-mutation ($\mathcal{B}_0$). Using Bayes theorem, *recall* can be re-written as:

$$\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}] = \frac{\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0] \times \mathbb{P}[\mathcal{B}_0]}{\mathbb{P}[\mathcal{B}]}, \tag{14}$$

where $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$ and $\mathbb{P}[\mathcal{B}_0]$ can be calculated using equations 12 and 5, respectively, and $\mathbb{P}[\mathcal{B}]$ is the probability of a mutation to be beneficial at the level of the population, which can be computed from the law of total probabilities as:

$$\mathbb{P}[\mathcal{B}] = \sum_{x \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}} \mathbb{P}[\mathcal{B} \mid x] \times \mathbb{P}[x]. \tag{15}$$

1.8 Correlation with effective population size (N_e)

Genetic diversity estimator Watterson's θ_S was obtained for each population from the synonymous SFS as in Achaz (2009). For each popuation, N_e was estimated from the equation $N_\mathrm{e} = \theta_\mathrm{S}/(4 \times u)$, where u is the mutation rate per generation. Estimates for u were averaged per species across the pedigree-based estimation in Bergeron et al. (2023) for *Homo*, *Bos*, *Capra* and *Chlorocebus*. For *Ovis* we used the estimated u of *Capra*.

For *Equus*, we used u as estimated in Orlando et al. (2013) ($u = 7.24 \times 10^{-9}$). Because a correlation must account for phylogenetic relationship and non-independence of samples, we fitted a Phylogenetic Generalized Linear Model in *R* with the package `caper` (Orme et al., 2013). The mammalian dated tree was obtained from TimeTree (Kumar et al., 2017) and pruned to include only the species analysed in this study, with multi-furcation of the different populations from each species placed at the same divergence time as the species (section S4.1).

Data availability

The data underlying this article are available at `https://doi.org/10.5281/zenodo.7878954`. Snakemake pipeline, analysis scripts and documentation are available at github.com/ThibaultLatrille/SelCoeff.

Acknowledgements

We gratefully acknowledge the help of Mélodie Bastian, Nicolas Lartillot, Carina Farah Mugal, Laurent Duret, Alexandre Reymond, Daniele Silvestro and Nicolas Gambardella for their advice and reviews concerning this manuscript. This work was performed using the computing facilities of the CC LBBE/PRABI. This study makes use of data generated by the NextGen Consortium. Funding: Université de Lausanne; Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec. The European Union's Seventh Framework Programme (FP7/2010-2014) provided funding for the project under grant agreement no 244356 - "NextGen".

Competing interests

The authors declare no conflicts of interest.

Author information

Original idea: T.L. and J.J.; Model conception: T.L., J.J. and N.S.; Code: T.L.; Data analyses: T.L. and J.J.; Interpretation: T.L., J.J., D.A.H. and N.S.; First draft: T.L. and J.J.; Editing and revisions: T.L., J.J., D.A.H. and N.S. Project management and funding: N.S.

References

Achaz, G. (2009). Frequency Spectrum Neutrality Tests: One for All and All for One. *Genetics*, 183(1):249–258.

Al Abri, M. A., Holl, H. M., Kalla, S. E., Sutter, N. B., and Brooks, S. A. (2020). Whole genome detection of sequence and structural polymorphism in six diverse horses. *PLoS ONE*, 15(4):e0230899.

Ashenberg, O., Gong, L. I., and Bloom, J. D. (2013). Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 110(52):21071–21076.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., and Pupko, T. (2012). FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40(W1):W580–W584.

Bazykin, G. A. (2015). Changing preferences: Deformation of single position amino acid fitness landscapes and evolution of proteins. *Biology letters*, 11(10):20150315.

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., and Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, pages 1–7.

Bloom, J. D. (2017). Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*, 12(1):1–24.

Charlesworth, J. and Eyre-Walker, A. (2007). The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proceedings of the National Academy of Sciences*, 104(43):16992–16997.

Chen, J., Bataillon, T., Glémin, S., and Lascoux, M. (2021). Hunting for beneficial mutations: Conditioning on SIFT scores when estimating the distribution of fitness effect of new mutations. *Genome Biology and Evolution*.

Chi, P. B., Kosater, W. M., and Liberles, D. A. (2020). Detecting Signatures of Positive Selection against a Backdrop of Compensatory Processes. *Molecular Biology and Evolution*, 37(11):3353–3362.

Cvijović, I., Good, B. H., Jerison, E. R., and Desai, M. M. (2015). Fate of a mutation in a fluctuating environment. *Proceedings of the National Academy of Sciences*, 112(36):E5021–E5028.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, volume 220. John Murray.

Dos Reis, M. (2015). How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the fisher-wright mutation-selection framework. *Biology Letters*, 11(4):20141031.

Doud, M. B., Ashenberg, O., and Bloom, J. D. (2015). Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Molecular Biology and Evolution*, 32(11):2944–2960.

Douzery, E. J., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., and Ranwez, V. (2014). OrthoMaM v8: A database of orthologous exons and coding sequences for comparative genomics in mammals. *Molecular Biology and Evolution*, 31(7):1923–1928.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in Ecology & Evolution*, 21(10):569–575.

Eyre-Walker, A. and Keightley, P. D. (2009). Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9):2097–2108.

Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2):891–900.

Fischer, A., Greenman, C., and Mustonen, V. (2011). Germline Fitness-Based Scoring of Cancer Mutations. *Genetics*, 188(2):383–393.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. The Clarendon Press.

Foley, N. M., Mason, V. C., Harris, A. J., Bredemeyer, K. R., Damas, J., Lewin, H. A., Eizirik, E., Gatesy, J., Karlsson, E. K., Lindblad-Toh, K., Zoonomia Consortium, Springer, M. S., and Murphy, W. J. (2023). A genomic timescale for placental mammal evolution. *Science*, 380(6643):eabl8189.

Galtier, N. (2016). Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genetics*, 12(1):e1005774.

Gavrilets, S. and Losos, J. B. (2009). Adaptive Radiation: Contrasting Theory with Data. *Science*, 323(5915):732–737.

Gillespie, J. H. (1994). Substitution processes in molecular evolution. III. Deleterious alleles. *Genetics*, 138(3):943–952.

Gillespie, J. H. (1995). On Ohta's hypothesis: Most amino acid substitutions are deleterious. *Journal of Molecular Evolution*, 40(1):64–69.

Goldstein, R. A., Pollard, S. T., Shah, S. D., and Pollock, D. D. (2015). Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Molecular Biology and Evolution*, 32(6):1373–1381.

Goldstein, R. A. and Pollock, D. D. (2017). Sequence entropy of folding and the absolute rate of amino acid substitutions. *Nature Ecology & Evolution*, 1(12):1923–1930.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917.

Hartl, D. L. and Taubes, C. H. (1996). Compensatory nearly neutral mutations: Selection without adaptation. *Journal of Theoretical Biology*, 182(3):303–309.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., Garcia Giron, C., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O. G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Gonzalez Martinez, J., Marugán, J. C., Maurel, T., McMahon, A. C., Mohanan, S., Moore, B., Muffato, M., Oheh, D. N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M. P., Salam, A. I. A., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., De Silva, N., Flint, B., Frankish, A., Hunt, S. E., IIsley, G. R., Langridge, N., Loveland, J. E., Martin, F. J., Mudge, J. M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S. J., Cunningham, F., Yates, A. D., Zerbino, D. R., and Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891.

Huber, C. D., Kim, B. Y., Marsden, C. D., and Lohmueller, K. E. (2017). Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences*, 114(17):4465–4470.

Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2019). The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*, 73(1):111–114.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. (2017). Shifting balance on a static mutation–selection landscape: A novel scenario of positive selection. *Molecular Biology and Evolution*, 34(2):391–407.

Keightley, P. D. and Eyre-Walker, A. (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1187–1193.

Keightley, P. D. and Jackson, B. C. (2018). Inferring the probability of the derived vs the ancestral allelic state at a polymorphic site. *Genetics*, 209(3):897–906.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34(7):1812–1819.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L., and Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067.

Lanfear, R., Kokko, H., and Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology and Evolution*, 29(1):33–41.

Lartillot, N. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109.

Lartillot, N. (2013). Inférence probabiliste pour la phylogénie, la génomique comparative et les sciences de la macro-évolution. Technical report.

Latrille, T., Lanore, V., and Lartillot, N. (2021). Inferring long-term effective population size with mutation–selection models. *Molecular Biology and Evolution*, 38(10):4573–4587.

Latrille, T., Rodrigue, N., and Lartillot, N. (2023). Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proceedings of the National Academy of Sciences of the United States of America*, 120(11):e2214977120.

Lynch, M. (2023). Mutation pressure, drift, and the pace of molecular coevolution. *Proceedings of the National Academy of Sciences*, 120(27):e2306741120.

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080.

McCandlish, D. M. and Stoltzfus, A. (2014). Modeling evolution using the probability of fixation: History and implications. *Quarterly Review of Biology*, 89(3):225–252.

McDonald, J. H. and Kreitman, M. (1991). Adaptative protein evolution at Adh locus in Drosophila. *Nature*, 351(6328):652–654.

Merrell, D. J. (1994). *The Adaptive Seascape: The Mechanism of Evolution*. U of Minnesota Press.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301.

Moses, A. M. and Durbin, R. (2009). Inferring Selection on Amino Acid Preference in Protein Domains. *Molecular Biology and Evolution*, 26(3):527–536.

Moutinho, A. F., Bataillon, T., and Dutheil, J. Y. (2019). Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, 34(3):315–338.

Mustonen, V. and Lässig, M. (2009). From fitness landscapes to seascapes: Non-equilibrium dynamics of selection and adaptation. *Trends in genetics*, 25(3):111–119.

Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23(1992):263–286.

Ohta, T. and Gillespie, J. H. (1996). Development of Neutral and Nearly Neutral Theories. *Theoretical Population Biology*, 49(2):128–142.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A. M. V., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G. D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J. F., Weinstock, J., Gregersen, K., Røed, K. H., Eisenmann, V., Rubin, C. J., Miller, D. C., Antczak, D. F., Bertelsen, M. F., Brunak, S., Al-Rasheid, K. A. S., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M. T. P., Kjær, K., Sicheritz-Ponten, T., Jensen, L. J., Olsen, J. V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., and Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74–78.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., and Pearse, W. (2013). The caper package: Comparative analysis of phylogenetics and evolution in R. *R package version*, 5(2):1–36.

Park, Y., Metzger, B. P. H., and Thornton, J. W. (2022). Epistatic drift causes gradual decay of predictability in protein evolution. *Science*, 376(6595):823–830.

Piganeau, G. and Eyre-Walker, A. (2003). Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18):10335–10340.

Platt, A., Weber, C. C., and Liberles, D. A. (2018). Protein evolution depends on multiple distinct population size parameters. *BMC Evolutionary Biology*, 18(1):1–9.

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M. K., and Douzery, E. J. (2007). OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, 7(1):1–12.

Rice, D. P., Good, B. H., and Desai, M. M. (2015). The Evolutionarily Stable Distribution of Fitness Effects. *Genetics*, 200(1):321–329.

Rodrigue, N. and Lartillot, N. (2017). Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. *Molecular Biology and Evolution*, 34(1):204–214.

Rodrigue, N., Lartillot, N., and Philippe, H. (2008). Bayesian comparisons of codon substitution models. *Genetics*, 180(3):1579–1591.

Rodrigue, N., Latrille, T., and Lartillot, N. (2021). A Bayesian mutation-selection framework for detecting site-specific adaptive evolution in protein-coding genes. *Molecular Biology and Evolution*, 38(3):1199–1208.

Rodrigue, N. and Philippe, H. (2010). Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends in Genetics*, 26(6):248–252.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10):4629–34.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J., and Ranwez, V. (2019). OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4):861–862.

Sella, G. and Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9541–9546.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Smith, N. G. and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–1024.

Spielman, S. J. and Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Molecular biology and evolution*, 32(4):1097–1108.

Starr, T. N. and Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218.

Stolyarova, A. V., Nabieva, E., Ptushenko, V. V., Favorov, A. V., Popova, A. V., Neverov, A. D., and Bazykin, G. A. (2020). Senescence and entrenchment in evolution of amino acid sites. *Nature Communications*, 11(1):4603.

Sullivan, P. F., Meadows, J. R. S., Gazal, S., Phan, B. N., Li, X., Genereux, D. P., Dong, M. X., Bianchi, M., Andrews, G., Sakthikumar, S., Nordin, J., Roy, A., Christmas, M. J., Marinescu, V. D., Wang, C., Wallerman, O., Xue, J., Yao, S., Sun, Q., Szatkiewicz, J., Wen, J., Huckins, L. M., Lawler, A., Keough, K. C., Zheng, Z., Zeng, J., Wray, N. R., Li, Y., Johnson, J., Chen, J., Zoonomia Consortium, Paten, B., Reilly, S. K., Hughes, G. M., Weng, Z., Pollard, K. S., Pfenning, A. R., Forsberg-Nilsson, K., Karlsson, E. K., and Lindblad-Toh, K. (2023). Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*, 380(6643):eabn2937.

Svardal, H., Jasinska, A. J., Apetrei, C., Coppola, G., Huang, Y., Schmitt, C. A., Jacquelin, B., Ramensky, V., Müller-Trutwin, M., Antonio, M., Weinstock, G., Grobler, J. P., Dewar, K., Wilson, R. K., Turner, T. R., Warren, W. C., Freimer, N. B., and Nordborg, M. (2017). Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nature Genetics*, 49(12):1705–1713.

Tamuri, A. U. and dos Reis, M. (2021). A mutation-selection model of protein evolution under persistent positive selection. *Molecular Biology and Evolution*.

Tamuri, A. U. and Goldstein, R. A. (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(3):1101–1115.

Tataru, P. and Bataillon, T. (2020). polyDFE: Inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data. In *Methods in Molecular Biology*, volume 2090, pages 125–146. Humana Press Inc.

Tataru, P., Mollion, M., Glémin, S., and Bataillon, T. (2017). Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, 207(3):1103–1119.

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

Vigué, L., Croce, G., Petitjean, M., Ruppé, E., Tenaillon, O., and Weigt, M. (2022). Deciphering polymorphism in 61,157 Escherichia coli genomes via epistatic sequence landscapes. *Nature Communications*, 13(1):4030.

Welch, J. J. (2006). Estimating the Genomewide Rate of Adaptive Protein Evolution in Drosophila. *Genetics*, 173(2):821–837.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2):97–159.

Yang, Z. and Bielawski, J. R. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12):496–503.

Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3):568–579.

Youssef, N., Susko, E., and Bielawski, J. P. (2020). Consequences of stability-induced epistasis for substitution rates. *Molecular Biology and Evolution*.

Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., and the 1000 Genomes Project Consortium (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, 6(7):gix038.

9

Increased positive selection in highly recombining genes is not an evidence for a beneficial effect of recombination

Context

Although it fits best as a third chapter, this work was actually done at the end of my last year of PhD. During the whole PhD, I found it unsatisfactory to use the dN/dS ratio to assess the costs and benefits of recombination. Personally, I found that an approach based on fitness was more straightforward, but I did not take the time to fully explore my distrust of the dN/dS. As the pile of new projects was growing taller and taller, I did not take the time to look at it more properly. But when working on another project with Thibault, using the fitness landscape described in chapter 8, a contradiction was puzzling me: while highly recombining proteins appeared to be under more positive selection in humans, they were less fit than others, but had a lower dN/dS. I realized that after more than two years of PhD on evolution, I had still a poor understanding of natural selection. Shortly afterwards, at a conference, I heard Nicolas say something to a colleague on an unrelated matter: "Why not actually model the biological processes we are assuming for a change?". I therefore decided to do just that. I have to say I pursued this work alone not because of a profound aversion for the company of my peers, but if I had told my supervisors and collaborators that I was going to start a new project without having finished the others at the end of the last year of my PhD, they would have been (righteously) very annoyed (out of concern, of course!). I hope the reader will enjoy the results of this risky endeavour.

Detailed contributions

As the short list of authors suggests, I did most of the work, but I benefited from advices and reviews of Nicolas Lartillot, Carina Farah Mugal and Laurent Duret on a first version of this manuscript. This manuscript has not been submitted to any journal yet.

Increased positive selection in highly recombining genes is not an evidence for a beneficial effect of recombination

J. Joseph[1]

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

julien.joseph@ens-lyon.fr

October 16, 2023

Abstract

It is commonly thought that the long-term advantage of meiotic recombination is to dissipate genetic linkage, allowing natural selection to act independently on different loci. It is thus theoretically expected that genes with higher recombination rates evolve under more effective selection, and can adapt more easily to environmental change. On the other hand, recombination is often associated with GC-biased gene conversion (gBGC), which theoretically interferes with selection by promoting the fixation of deleterious GC alleles. To test these predictions, several empirical studies assessed whether highly recombining genes experienced higher rates of positive selection (due to dissipation of genetic linkage) or higher fixation of deleterious GC alleles (due to gBGC), assuming a fixed distribution of fitness effects (DFE) for all genes. In this study, I directly derive the expected shape of the DFE from the evolutionary history of a gene (shaped by mutation, selection, drift and gBGC) under empirical fitness landscapes. I show that genes that have known high levels of gBGC have a DFE shifted towards positive values. Only a slight decrease in the genome-wide intensity of gBGC therefore leads to true positive selection specifically in highly recombining genes, but with AT-biased substitutions. This shows that even truly increased positive selection in highly recombining genes is not necessarily an evidence for a beneficial effect of recombination. Additionally, I show that the death of a long-lived recombination hotspots can lead to a higher dN/dS than its birth, but with substitutions patterns biased towards AT, and only at selected position. This shows that controlling for a substitution bias towards GC is therefore not sufficient to rule out the contribution of gBGC to signatures of accelerated evolution.

***K*eywords** Recombination · gBGC · DFE · back-mutations · positive selection · dN/dS · fitness landscape

1 Introduction

Meiotic recombination is a key cellular process with major consequences for evolution. In the vast majority of sexually reproducing species, the formation of a crossing over is required for the proper segregation of homologs during meiosis (Baker et al., 1976; Davis and Smith, 2001; Pardo-Manuel de Villena and Sapienza, 2001; Gerton and Hawley, 2005). Failure of this process often leads to meiotic arrest or the formation of aneuploid gametes with missing or extra chromosomes (Hassold et al., 2007; Handel and Schimenti, 2010; Székvölgyi and Nicolas, 2010; Brick et al., 2012; Mihola et al., 2019).

From an evolutionary perspective, the most commonly cited hypothesis for the long-term maintenance of recombination is its effect on genetic linkage (Felsenstein, 1974; Otto and Barton, 1997; Otto and Lenormand, 2002; Keightley and Otto, 2006). By creating new combinations of alleles, recombination events dissociate the selective pressures exerted on different loci, dissipating the so-called Hill-Robertson interference (Hill and Robertson, 1966; Smith and Haigh, 1973; Charlesworth et al., 1993; Roze and Barton, 2006). A large number of theoretical studies have explored conditions under which recombination can have a positive, negative or no effect on the efficacy of natural selection. As a result, it is now generally accepted that in populations of finite size and where new mutations are mostly deleterious, increased recombination rates enhance the efficacy of both purifying and positive selection (Felsenstein, 1974; Hickey and Golding, 2018; Roze, 2021). To test this theoretical result, several empirical studies tried to quantify signatures of selection across regions of different recombination rate, with mixed results (Bullaughey et al., 2008; Gossmann et al., 2014; Hussin et al., 2015; Bolívar et al., 2016; Castellano et al., 2016; Corcoran et al., 2017; Rousselle et al., 2019; Castellano et al., 2020; Murga-Moreno et al., 2019).

One of the reasons for these mixed results is that recombination has another effect on genomes that has been largely overlooked in theoretical studies: GC-biased gene conversion (gBGC) (Brown and Jiricny, 1987; Duret and Galtier, 2009). The very mechanism of recombination requires hybridisation between the two single-stranded parental DNAs. If an individual is heterozygous at this position, this will cause a mismatch in the resulting double-stranded DNA (heteroduplex) that can be repaired in the direction of either allele. One allele thus converts the other and this phenomenon is therefore called gene conversion (Winkler, 1930) (reviewed in Roman (1985)). In many Eukaryotes including vertebrates, plants and fungi, recombination-associated gene conversion is biased towards GC alleles (Pessia et al., 2012; Galtier et al., 2018). The evolutionary consequences of this mechanism is the rapid spread of GC alleles in regions of high recombination rates, eventually leading to their rapid fixation in the population.

As the main methods of inference of positive selection rely on the detection of rapid fixation events (McDonald and Kreitman, 1991), the pervasive fixation of GC alleles acts as a confounding factor (Galtier et al., 2009; Ratnakumar et al., 2010). Ratnakumar et al. (2010) estimated that ∼20% of protein-coding genes showing a rapidly increasing dN/dS ratio in the human branch were likely to be the result of the birth of a gBGC hotspot, and did not correspond to adaptation to the new human environment as previously hypothesized (Kosiol et al., 2008). This proportion was estimated using the fact that increased gBGC skews substitution patterns towards GC, even in non coding sequence, which is not expected under positive selection only (Ratnakumar et al., 2010).

However, it is not clear whether contrasting positive selection across genes with different recombination rates is a suitable way to test for the beneficial effect of recombination, even when controlling for the rapid fixation of GC alleles. Indeed, even true positive selection in highly recombining genes is not necessarily the result of a higher selection efficacy. In fact, the rate of positive selection is the product of the number of beneficial mutations per generation, times their fixation probability (Kimura, 1962). By interpreting increased positive selection as a sign of increased selection efficacy, a strong hypothesis is implicitly made: genes that evolved under different recombination rates have the same number of opportunities for beneficial mutations. Alternatively stated, the above-mentioned empirical studies implicitly (or sometimes explicitly) assume an invariant distribution of fitness effects (DFE) of new mutations across recombination rate categories (Bullaughey et al., 2008; Gossmann et al., 2014; Hussin et al., 2015; Bolívar et al., 2016; Castellano et al., 2016; Corcoran et al., 2017; Rousselle et al., 2019; Castellano et al., 2020; Murga-Moreno et al., 2019).

The DFE of new mutation is a key parameter in evolution, and largely depends on the fitness landscape. Indeed, in a gene that is strongly selectively constrained, most new mutations will be deleterious. Conversely in a non-functional pseudogene, all mutations will be neutral. Moreover, the DFE will be affected by the position of a sequence in its fitness landscape. Indeed, in a gene that is close to its fitness optimum, most mutations will be deleterious. Conversely, in a gene that is far from its fitness optimum, many mutations will be beneficial. In a more formal phrasing, the DFE can be seen as the local derivative of the fitness landscape at the position of a sequence in it, and can be computed as such (Martin and Lenormand, 2006).

As the fitness landscape is usually difficult to access, it is a common practice in population genetics to consider the DFE as a fixed parameter (e.g. Bolívar et al. (2016) and Corcoran et al. (2017) for theoretical models of gBGC), even though the DFE is supposed to depend on the other parameters of the model. In particular, the position of the sequence in a fitness landscape is driven by natural selection, mutation, drift and gBGC.

In this paper, I used a simple mutation-selection model to compute the expected DFE in genes that have evolved under different recombination rates, under a collection of experimental and empirical fitness landscapes. I show that by promoting the fixation of many deleterious mutations, gBGC drives sequences away from their optimal fitness, which directly creates more opportunities for beneficial back-mutations and skews the DFE towards positive values in highly recombining genes. Upon a decrease of the repair bias towards GC, or a decrease of the genome-wide recombination rate (keeping the recombination landscape constant), those beneficial back-mutations eventually get fixed by positive selection. This creates a true signal of positive selection in highly recombining genes that does not imply either higher selection efficacy or increased adaptation. Finally, I show that the death of a long-lived recombination hotspot can generate an even higher dN/dS than its birth. Of note, in this case, substitution patterns are biased towards AT, and only at selected positions. Importantly, this increase in dN/dS is due to positive selection, but is not a signature of an adaptation to changing environments. Therefore, these genes potentially add to the list of false positives when using accelerated evolution or signatures of positive selection to test for adaptive evolution (Hartl and Taubes, 1996; Galtier and Duret, 2007; Mustonen and Lässig, 2009; Jones et al., 2017).

2 Results

2.1 Impact of gBGC on the equilibrium DFE of new mutations

As stated in the introduction, the DFE of new mutations depends both on the fitness landscape, and on the position of a given individual in it. A total of six fitness landscapes were used in this study: one mammalian fitness landscape obtained by fitting a mutation-selection model to multi-species alignments of 14,509 protein-codon genes in (Latrille et al., 2023), four fitness landscapes obtained from deep mutational scanning (DMS) experiments in Influenza (Thyagarajan and Bloom, 2014; Doud et al., 2015), *E.coli* (Stiffler et al., 2015), and *S.cervisae* (Kitzman et al., 2015), and a concatenate of the four previous DMS fitness landscapes. As the fitness landscapes are fixed, I therefore study selection dynamics in the absence of adaptation. Moreover, because the fitness landscapes are site-specific, I thus neglect epistatic interactions. The consequences of this latter assumption are discussed below. The position of a sequence in these fitness landscapes depends on its evolutionary history. In particular, this position will result from an equilibrium between mutation, selection, drift and gBGC (Nagylaki, 1983; Hartl and Taubes, 1998; Sella and Hirsh, 2005; Mustonen and Lässig, 2009). Using a mutation-selection model, one can compute the equilibrium frequency for each codon at each site for all the fitness landscapes (details in methods).

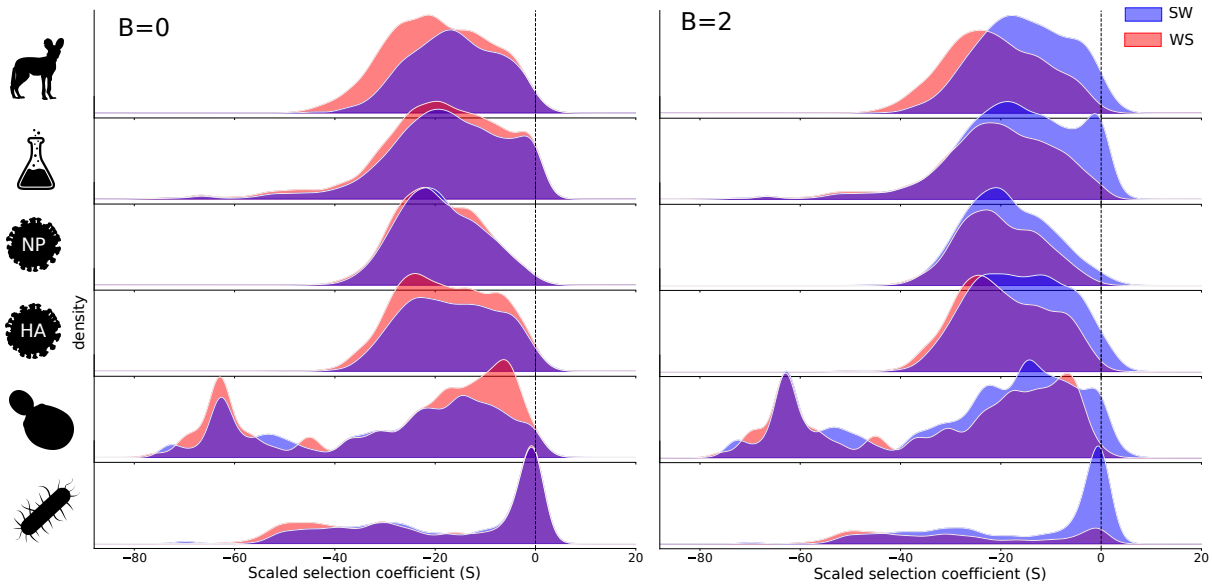


Figure 1: Distribution of fitness effects of new mutations at equilibrium separately for WS (red) and SW (blue) mutations. Equilibrium frequencies were computed for each fitness landscape with $B = 0$ left and $B = 2$ right. From top to bottom: 2000 sites randomly sampled from the mammalian fitness landscapes, the concatenate of the DMS fitness landscapes (1389 sites), the fitness landscape of the influenza protein NP (498 sites), the influenza protein HA (564 sites), the *S.cervisae* protein Gal4 (64 sites) and the *E.coli* protein β-lactamase (263 sites).

At mutation-selection-drift-gBGC equilibrium, by sampling all possible mutations, weighted by the equilibrium frequency of the ancestral codon and their mutation rate, and associating them with their fitness

effect, one directly obtains a DFE of new mutations (see details in methods). I computed the DFE separately for WS (from AT $\mapsto$ GC) and SW (from GC $\mapsto$ AT) mutations, using a population-scaled gBGC coefficient ($B = 4N_e b$) of 0 (no gBGC) or 2. For the sake of clarity, I computed equilibrium frequencies with a Jukes Cantor mutation matrix (which assumes that all bases have equal mutation rates), such as to avoid the confounding effect of mutation biases on the distribution of fitness effects studied in Latrille and Lartillot (2022) and to focus on the effect of gBGC. For all the DFEs presented in this study, I considered only non-synonymous mutations. To be able to compare fitness landscapes, I rescaled them such that the mean population-scaled selection coefficient from the fittest codon ($S = 4N_e s$) is 20. This ensures that the selective pressure on protein is high enough such that most sites are little influenced by gBGC, as expected in reality (Duret and Galtier, 2009).

	B = 0			**B = 2**		
Fitness landscape	**All mutations**	**WS**	**SW**	**All mutations**	**WS**	**SW**
Mammalian	1.25%	1.11%	1.30%	1.64%	0.35%	2.97%
Concatenated DMS	3.04%	2.68%	3.33%	3.22%	0.79%	5.39%
NP (Influenza)	0.61%	0.68%	0.54%	0.71%	0.15%	1.27%
HA (Influenza)	1.42%	1.15%	1.74%	1.73%	0.29%	3.35%
Gal4 (*S.cervisae*)	1.57%	1.33%	1.83%	1.93%	0.43%	3.84%
β-lactamase (*E.coli*)	11.9%	10.8%	12.4%	12.11%	4.47%	16.2%

Table 1: Proportion of beneficial mutations under the two equilibrium conditions $B = 0$ and $B = 2$ for all the fitness landscapes of the study.

In the absence of gBGC, the positive part of the DFE is almost identical between WS and SW mutations (Fig.1 and Table 1). However, even in the absence of gBGC or mutational bias, there appears to be slightly more opportunity for deleterious mutations towards GC than towards AT (Fig.1). This can be partly explained by the structure of the genetic code: a majority of amino-acids have an obligatory A or T in first or second position of codons. Indeed, when randomly switching amino-acids in the fitness landscape of a given site, there are still more deleterious mutations towards GC (Fig.S1).

When the sequence has evolved under gBGC ($B = 2$), the DFE of WS mutations is skewed towards negative values, while the DFE of SW mutations is skewed towards positive values (Fig.1). Indeed, at equilibrium, GC alleles that were slightly deleterious have been fixed because of gBGC, and SW mutations are therefore more beneficial on average (Table 1). Conversely, most WS mutations that were slightly advantageous have already been fixed by both positive selection and gBGC. Therefore, only WS mutations that cannot fix because they are deleterious enough to be wiped out by purifying selection even when helped by gBGC remain available as mutational opportunities (Table 1). It is also important to note that when $B = 2$, one always expects more beneficial back-mutations in total than without gBGC (Table 1).

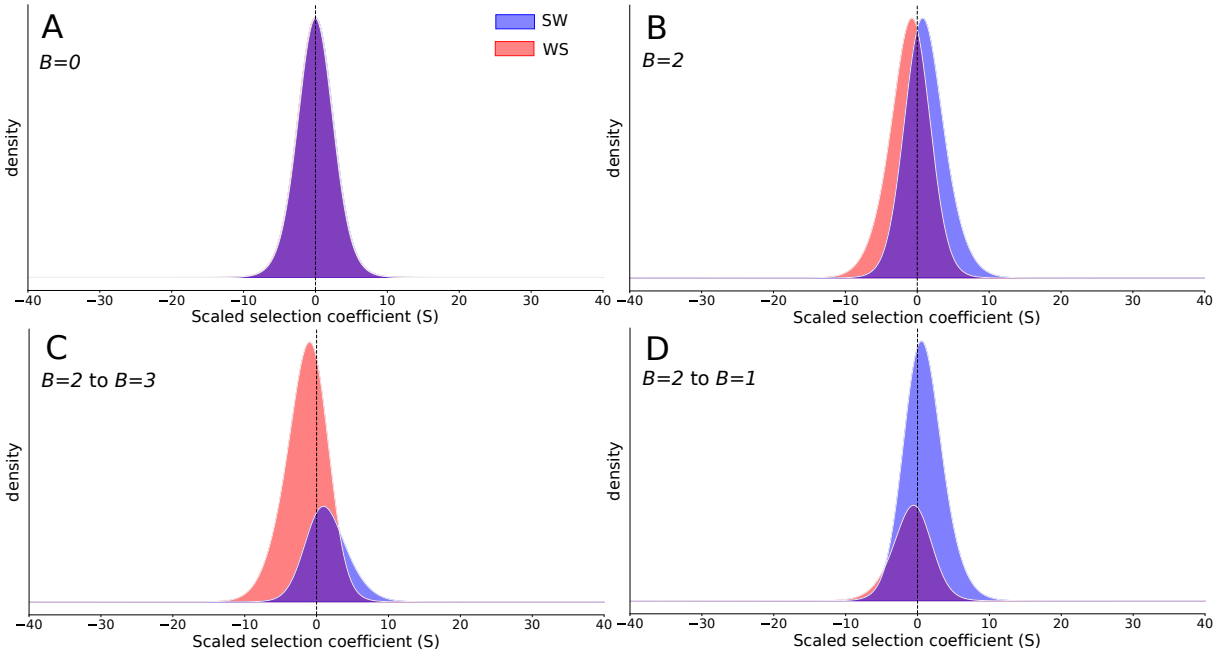


Figure 2: Distribution of fitness effects of new substitutions separately for WS (red) and SW (blue) substitutions under the concatenated DMS fitness landscape. A) Equilibrium frequencies were computed with $B = 0$ and the substitution rate also with $B = 0$ (no gBGC). B) Equilibrium frequencies were computed with $B = 2$ and the substitution rate also with $B = 2$. C) Equilibrium frequencies were computed with $B = 2$ and the substitution rate also with $B = 3$ (increase in gBGC). D) Equilibrium frequencies were computed with $B = 2$ and the substitution rate also with $B = 1$ (decrease in gBGC). The panel A&B therefore represent substitution rates at equilibrium, while panels C&D represent substitution rates out of equilibrium.

2.2 Impact of gBGC fluctuations on selection dynamics

Similarly to the DFE of new mutations, by sampling all possible mutations weighted by the equilibrium frequency of the ancestral codon and their substitution rate, and by associating them to their fitness effect, one directly obtains a DFE of substitutions (see details in methods). I computed this DFE of substitutions at equilibrium for $B = 0$ and $B = 2$. As the results are almost identical for all fitness landscapes, I only represented results obtained on the concatenate of DMS fitness landscapes in Fig.2, and presented the other ones in Supplementary Fig.S2. For $B = 0$, there are as many beneficial substitutions as deleterious ones (Fig.2A and Fig.S2). Additionally, there are as many WS substitutions as SW ones (Fig.2A and Fig.S2). This is expected since the sequence is at equilibrium both for fitness and base composition. For $B = 2$, the exact same conditions are met: there are as many beneficial substitutions as deleterious ones and there are as many WS substitutions as SW ones, again, the sequence is at equilibrium (Fig.2B and Fig.S2). However, as hinted from the DFE of new mutations, WS substitutions are more often deleterious, while SW substitutions are more often beneficial (Fig.2B and Fig.S2). To gain insight on selection dynamic out of equilibrium, I computed the equilibrium frequencies of codons under $B = 2$, and computed the substitution rates from

this equilibrium sequence either with $B = 1$ or $B = 3$ (Fig.2C&D and Fig.S2). When gBGC increases ($B = 2 \mapsto B = 3$), there is an increase in the fixation of deleterious GC alleles (Fig.2D and Fig.S2), as predicted by theoretical models, and observed empirically in diverse organisms (Nagylaki, 1983; Bengtsson, 1990; Galtier et al., 2009; Glémin, 2010; Necşulea et al., 2011; Bolívar et al., 2016; Rousselle et al., 2019). Conversely, when gBGC decreases ($B = 2 \mapsto B = 1$), there is an equivalent increase in the fixation of beneficial AT alleles (Fig.2C and Fig.S2).

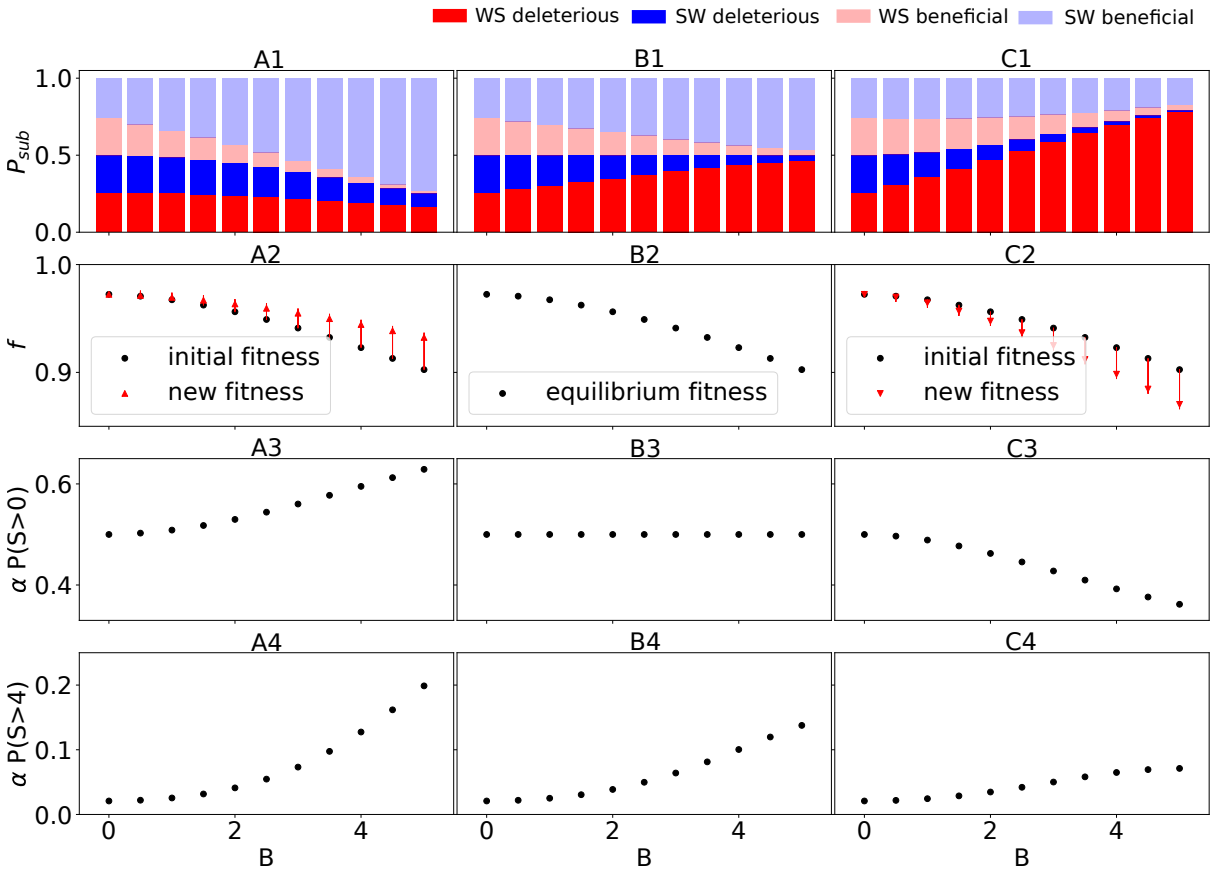


Figure 3: Row 1: Proportion of the substitutions (P_{sub}) contributed by WS deleterious (bright red), SW deleterious (bright blue), WS beneficial (light red) ans SW beneficial (light blue) mutations as a function of the population-scaled gBGC coefficient under the concatenated DMS fitness landscape in three scenarios: Equilibrium frequencies are computed with B, and substitutions from this equilibrium sequence are computed with $0.7 \times B$ (A), B (B) and $1.3 \times B$ (C). Row 2: Relative fitness (f) of sequences as a function of the B they are evolving under (black dots). Red arrows correspond to the fitness evolution of the sequences after A) a decrease of B by 30%, or C) after an increase of B by 30%. Row 3: Proportion of positively selected substitutions (P(S¿0) in the three scenarios described above. Row 4: Proportion of strongly beneficial substitutions (P(S¿4) in the three scenarios described above.

I then computed equilibrium codon frequencies under a wide range of gBGC intensity, mimicking genes evolving under different recombination rates. I then computed the DFE of substitutions in two scenarios.

The first one corresponds to an increase of gBGC by 30% for all genes, the second to a decrease by the same amount. These scenario mimic either a change in the genome-wide recombination rate, or a change in the meiotic repair bias. To visualize the displacement of the sequence in the fitness landscape, I computed the relative fitness of the sequence for all equilibrium states (see details in methods). I represented the proportions of beneficial, deleterious, WS, and SW substitutions in (Fig.3 row 1 and Fig.S3). When gBGC increases, genes that evolved under stronger gBGC show the higher proportion of deleterious substitutions towards GC (Fig.3C1 and Fig.S3). Conversely, when gBGC decreases, genes that previously evolved under gBGC show higher levels of beneficial substitutions towards AT than others (Fig.3A1 and Fig.S3).

Positive selection can be quantified by the estimator α, which measures the proportion of positively selected substitutions (for all mutation types) among all substitutions. At equilibrium, without adaptation, we can see that $\alpha = 0.5$. This is indeed expected at equilibrium, where the fitness is static. However, when gBGC increases, α decreases, substitutions are more deleterious, and the sequences falls down the fitness landscape (Fig.3C2&3 and Fig.S4). Conversely, when gBGC decreases α increases. This is due to the fact that a sequence that has previously experienced higher levels of gBGC climbs larger distances in the fitness landscape when gBGC is relaxed (Fig.3A2&3 and Fig.S4).

Usually, mutations with $|N_e s| < 1$ are not included in α, since they are considered effectively neutral (Eyre-Walker, 2002). This is because α is usually intended to capture the proportion of adaptive substitutions, assuming that only adaptation (i.e. a change in the fitness landscape) can lead to strongly beneficial mutations (McDonald and Kreitman, 1991). Indeed, on a fixed fitness landscape and without gBGC, when effectively neutral mutations are discarded, $\alpha \simeq 0$ (Figure 3 row 4). However, at equilibrium, α increases with the strength of gBGC because gBGC induces the fixation of substitutions that are increasingly deleterious, and therefore compensated by substitution that become strongly beneficial (Figure 3B4). When gBGC decreases, α increases even more with gBGC, because it compensates the fixation of strongly deleterious mutation previously fixed under higher levels of gBGC while current levels of gBGC do not fix as many deleterious mutations (Figure 3A4). Finally, when gBGC increases, we still observe an increase of α with the strength of gBGC (lower than at equilibrium though) (Figure 3C4). While gBGC fixes more deleterious mutations than at equilibrium, the proportion of compensating strongly beneficial mutations is still higher than without gBGC, where almost every mutation is effectively neutral. Therefore, paradoxically, genes that experience the highest α show the steeper decrease in fitness at the same time (Figure 3C2). The results presented above represent the expectation of the true impact of gBGC on positive selection and coding sequence evolution.

2.3 Impact of gBGC fluctuations on the rate of coding sequence evolution

It has also been shown that gBGC can alter our ability to detect selection in protein-coding genes, mainly through incorrect interpretation of the dN/dS ratio (Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010; Bolívar et al., 2016). Originally, the dN/dS ratio was designed to quantify the selective constraints exerted on the sequence of a protein (Miyata and Yasunaga, 1980; Nei and Gojobori, 1986). It represents the probability of fixation of a mutation that alters the sequence of a protein, relative to a mutation that do not alter it, supposedly neutral (Spielman and Wilke, 2015). There are two key assumptions

under which the dN/dS ratio serves its purpose: 1) allelic frequency changes of synonymous mutations are supposed to be due to drift only and 2) mutation rates of synonymous and non-synonymous changes within a gene are supposed to be identical, such that the ratio of substitution rates represent only differences in fixation probability (Spielman and Wilke, 2015). In the presence of gBGC, both those assumptions are violated: 1) synonymous mutations do not evolve only by drift but under a directional force that affects their substitution rate (Nagylaki, 1983), and 2) as the base composition differs at the three nucleotides of a codon (typically with the third codon showing higher GC-content in highly recombining genes), and SW mutation rates being usually higher that WS, mutation rates at synonymous and non-synonymous sites within a gene are different (Bolívar et al., 2016; Latrille and Lartillot, 2022). It appears therefore natural that the intensity of gBGC influences the dN/dS ratio (Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010; Bolívar et al., 2016).

To tackle this question in a fitness landscape framework, I used a realistic mutation matrix for humans to compute both equilibrium frequencies and substitution rates instead of the Jukes-Cantor. This allows me to capture the effect of the second violation of the assumptions of the dN/dS ratio: mutation rates between synonymous and non-synonymous changes are different, due to differences in base composition and base-specific mutation rates. Then, I computed the dN, dS and dN/dS for sequences that reached equilibrium under $B = 0$ and that subsequently accumulate substitutions under $B = 0 \mapsto 10$, mimicking the sudden birth of a recombination hotspot (Fig.4 and Fig.S5). Similarly, I computed dN, dS and dN/dS for sequences that reached equilibrium under $B = 0 \mapsto 10$ and that subsequently accumulate substitutions under $B = 0$, mimicking the sudden death of a recombination hotspot (Fig.4 and Fig.S5).

The ratio dN/dS is higher at equilibrium than during the death of a hotspots (from $B = 10 \mapsto B = 0$), than during its birth (from $B = 0 \mapsto B = 10$) (Fig.4C). In both cases dN is high, either because of the fixation of deleterious GC alleles or because of the fixation of beneficial back-mutations towards AT (Fig.4A). The difference of dN/dS mainly stems from the different behaviour of the dS (Fig.4B). During the birth of a hotspot, dS is very high because of the rapid fixation of neutral GC alleles under strong gBGC. But when it dies, the increase of dS is small, and only due to the higher mutation rates of GC nucleotides, which are more abundant after a strong gBGC episode. Because the behaviour of the dN/dS are mainly the result of the impact of gBGC on the dS, the results presented here do not differ much with those of Bolívar et al. (2016) where the dS is modelled in the same way. However, Bolívar et al. (2016) predict a decrease of the dN when a sequence goes from a high GC content to a lower one (mimicking a decrease in gBGC), while the present model predicts an increase because of beneficial back-mutations. Overall, under all fitness landscapes, beneficial back-mutations affect the dN/dS ratio, and the death of a recombination hotspot induces a higher dN/dS than its birth.

3 Discussion

In this study, I estimated the impact of beneficial back-mutations on the dynamics of natural selection in the presence of gBGC. I showed that gBGC induces a shift towards positive values in the DFE of new mutations. When the intensity of gBGC decreases, this leads to an increase of beneficial substitutions, and a

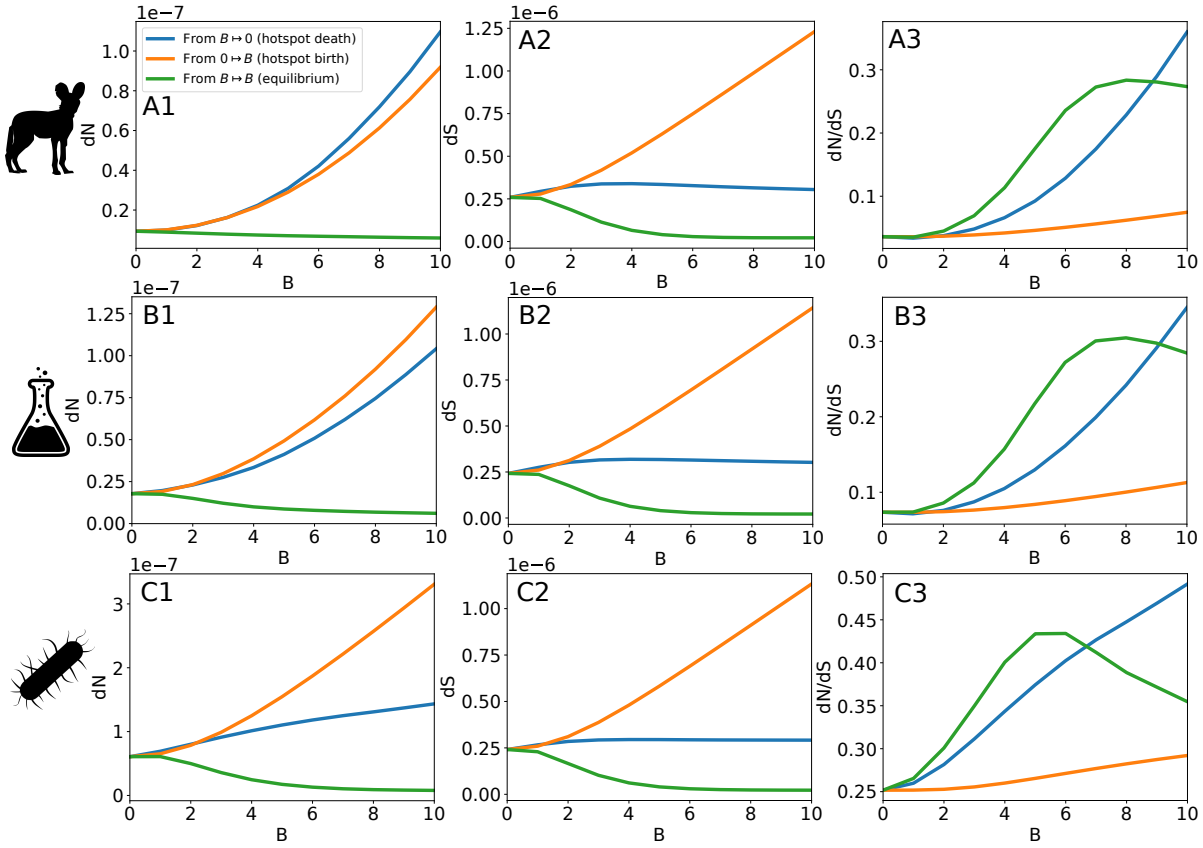


Figure 4: dN (A), dS (B), and dN/dS (C) as a function of the population scaled gBGC coefficient B in three scenarios: Equilibrium codon frequencies are computed without gBGC, and substitutions subsequently accumulate under a population-scaled gBGC coefficient of B (orange line), mimicking the birth of a recombination hotspot. Equilibrium codon frequencies are computed under a population-scaled gBGC coefficient of B, and substitutions subsequently accumulate without gBGC (orange line), mimicking the death of a recombination hotspot. And finally, equilibrium codon frequencies are computed under a population-scaled gBGC coefficient of B, and substitutions subsequently accumulate at equilibrium, under B (green line). From top to bottom: the concatenate of the DMS fitness landscapes (1389 sites), 2000 sites randomly sampled from the mammalian fitness landscapes and the *E.coli* protein β-lactamase (263 sites).

decrease of deleterious ones. If not properly investigated, this selection dynamic can be misinterpreted as a beneficial effect of recombination, when paradoxically it is fuelled by its deleterious consequences. Moreover, I showed that one can also expect an increase of the dN/dS due to these beneficial-back mutations. My results strengthens those of previous studies demonstrating that in the presence of gBGC, the dN/dS ratio is not a correct estimator of the selective pressure exerted on proteins (Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010; Bolívar et al., 2016, 2019). Importantly, the present result highlights that the absence of a substitution pattern skewed towards GC, affecting both coding and non-coding sequences does not allow the interpretation of local accelerated evolution as an adaptation to changing environments. It appears therefore unwise to infer adaptive evolution from the dN/dS ratio of a given gene without prior

information on gBGC, the fitness landscape, the local recombination dynamic or an external corroboration (e.g. correlation with phenotypic evolution).

3.1 Comparison with empirical results

In several studies, it was observed that genes richer in GC3 (supposedly having experienced higher recombination rates) have a lower dN/dS. Some of these studies interpreted this pattern as higher selection efficacy due to the dissipation of Hill-Robertson interference (Gossmann et al., 2014; Rousselle et al., 2019), while another argued that it was a consequence of gBGC (Bolívar et al., 2016). The rationale behind the latter claim is that when a sequence is strongly constrained, the dN will be little affected by gBGC while the dS will rise quickly, leading to a lower dN/dS (Bolívar et al., 2016). Here, we do not observe a regime when the dN/dS decreases, because the dS decreases always faster than the dN, leading to an increase of the dN/dS. However, both the present model and that of Bolívar et al. (2016) predict a decrease of the dS at equilibrium, while empirical data suggest an increase (Bolívar et al., 2016; Corcoran et al., 2017; Rousselle et al., 2019). This increase is usually attributed to a mutagenic effect of recombination (Bolívar et al., 2016; Rousselle et al., 2019; Castellano et al., 2020), not modelled here, although there are many properties that are correlated with GC3 that can explain an increase of the synonymous divergence. It is indeed difficult to interpret correlations between dN/dS and GC3 in empirical data, because genes with different GC3 might also evolve under different fitness landscapes and different mutational processes that are not directly linked to gBGC or recombination.

3.2 The equilibrium assumption in regards of hotspot dynamics

In this study, for mathematical convenience, I chose to model out of equilibrium dynamics by first computing codon frequencies under equilibrium conditions and then computing substitution rates under new conditions. In mammals, for instance, as most of the genome has a rather low recombination rate, it is reasonable to say that most genes are around equilibrium for a rather low strength of gBGC. In this case, the birth of a strong hotspot will indeed leave a clear signature of gBGC in substitution patterns (Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010). However, in humans and mice, the vast majority of hotspots are short-lived (Auton et al., 2012; Smagulova et al., 2016; Pratto et al., 2014; Alleva et al., 2021). Therefore, they do not have time to reach equilibrium, and the signatures of the death of a hotspot in the dN/dS might not be strong enough to be observed in these species. On the other hand, many species including most placental mammals, passerine birds, colubroid snakes, budding yeasts and many angiosperm plants exhibit recombination hotspots in 5' of genes that are relatively long-lived (Axelsson et al., 2012; Choi and Henderson, 2015; Singhal et al., 2015; Lam and Keeney, 2015; Kawakami et al., 2017; Schield et al., 2020; Hoge et al., 2023). In a previous study, we showed that those hotspots are vastly shared in placental mammals (chapter 6). Interestingly we also showed that they co-evolve slowly with DNA methylation (chapter 6). When those long-lived hotspots stop to be hypomethylated, they usually die (Maloisel and Rossignol (1998) chapter 6). Thus, they should leave a strong signal of positive selection, that do not correspond to an adaptive response of the gene to changing environments.

Altogether, both fast and slow hotspot dynamics should lead to pervasive fixation of deleterious mutations because of gBGC, compensated by positive selection. Even if this positive selection can be weak at a given gene when hotspots are short-lived, it should still have a significant impact throughout the genome.

3.3 Beneficial back-mutations versus compensatory mutations

The fitness landscapes used in this study are simplistic, in the sense that they imply that each amino-acid evolves independently, and does not interact with others (no epistasis). However, numerous studies found that epistasis plays an important role in protein evolution (Bonhoeffer et al., 2004; Breen et al., 2012; Starr and Thornton, 2016; Miton and Tokuriki, 2016). Let us consider the simple example of two residues interacting such that the optimal state of the protein requires either an alanine at one site and a valine at the other, or a lysine at one site and an arginine at the other. If the first residue mutates from alanine to lysine, optimal protein state can either be restored by a beneficial back-mutation from lysine to alanine, or a compensatory mutation at the other site from valine to arginine. In this simple case, it is easy to see that the fixation of a deleterious amino-acid does not always lead to an opportunity for positive selection at the same site. In the end, with epistasis, the deleterious effect of gBGC at one site/gene can be compensated by positive selection at another site/gene. Of note, if compensatory mutations occur preferentially in the same gene, one still expects a decrease of the genome-wide strength of gBGC to increase positive selection in highly recombining genes.

Interestingly, deep mutational scanning experiments showed that in some proteins, the site-specific fitness landscape can remain unchanged even in divergent lineages and predicts quite accurately the evolution of sequences *in natura* (Ashenberg et al., 2013; Doud et al., 2015; Bloom, 2017). This suggests that contrarily to the previous example, fitness at one site does not always depends on the amino-acids present at other sites. In this sense, there is also accumulating evidence for convergent adaptation at the molecular level (Christin et al., 2007; Zhen et al., 2012; Davies et al., 2012; Wu et al., 2020; Duchemin et al., 2023), suggesting that there is sometimes a limited number of mutations that can lead to a given phenotype, and thus limited opportunities for compensatory mutations. Moreover, an important role of beneficial back mutations have been reported in plants (Chen et al., 2021) and mammals (Moses and Durbin, 2009; Latrille et al., 2023). Recently, we demonstrated that beneficial-back mutations constitute an important fraction of beneficial mutations in several mammals (between 20% and 40%) (Latrille et al., 2023). Finally, several human accelerated genes show a remarkably strong AT-biased substitution pattern (Galtier et al., 2009), but they have not been more deeply investigated, and have been left unchecked in other studies (Ratnakumar et al., 2010; Kostka et al., 2012).

Altogether, while epistasis can spread throughout the genome the compensatory response to a strong gBGC episode, this compensation might not always be possible, and beneficial back-mutations are therefore expected to leave a signature of positive selection upon a decrease of gBGC. Still, the relative contribution of beneficial back-mutations and compensatory mutations to stabilizing selection in general is an open question that remains to be investigated, and the burst of fixation of deleterious mutations induced by gBGC could in fact be an interesting case study to investigate it.

3.4 Incorporating beneficial back-mutations into evolutionary thinking

In population genetics, the infinite site model approximation inherently excludes the possibility of beneficial back-mutations (Kimura, 1969; Ohta, 1992). This approximation can be very reasonable mathematically at a very short timescale, and is widely used in population genetics. Nevertheless, all sequences are the product of a very long evolutionary history. With a finite genome and longer evolutionary time, beneficial back-mutations are expected to occur (Gillespie, 1995; Hartl and Taubes, 1996; Piganeau and Eyre-Walker, 2003; Sella and Hirsh, 2005; Charlesworth and Eyre-Walker, 2007; Mustonen and Lässig, 2009). Under a slowly evolving fitness landscape, they largely contribute to molecular evolution (Chen et al., 2021; Latrille et al., 2023). Modelling evolution as the displacement of a sequence on a fitness landscape, determined by natural selection and nonadaptive evolutionary forces, instead of modelling selection with a constant DFE is therefore more relevant to study selection dynamics (Halpern and Bruno, 1998; Rodrigue et al., 2010; Tamuri et al., 2012; Rodrigue and Lartillot, 2014; Jones et al., 2017; Tamuri and dos Reis, 2022; Latrille et al., 2023).

Using this framework, I showed that signatures of positive and negative selection cannot be used to conclude on a beneficial or negative effect of recombination. For species where meiotic gene conversion is biased towards GC, this seriously questions all the previous interpretations of this beneficial effect (or its absence) that are based on the dN/dS or other methods contrasting non-synonymous and synonymous changes (Bullaughey et al., 2008; Gossmann et al., 2014; Castellano et al., 2016; Bolívar et al., 2016; Rousselle et al., 2019; Murga-Moreno et al., 2019; Hämälä and Tiffin, 2020). In presence of gBGC, it is thus important to account for beneficial back-mutations.

Even without gBGC, using the same framework, previous theoretical studies demonstrated that the nearly universal mutation bias towards AT necessarily induces a fixation bias towards GC in selectively constrained sequence, also because of beneficial back-mutations (Latrille and Lartillot, 2022; Kaj et al., 2023). This calls for a re-interpretation of results from several studies that interpreted the fact that GC alleles segregate at higher frequency at non-synonymous sites as an evidence for the negative impact of gBGC on adaptation (Hämälä and Tiffin, 2020; Liang et al., 2022). For the same reasons that increased positive selection in highly recombining genes is not an evidence for a beneficial effect of recombination, a fixation bias towards GC at non-synonymous sites is not an evidence for the presence of gBGC.

Incorporating beneficial back-mutations into evolutionary thinking is therefore essential for understanding selection dynamics in the presence of biases, and should prevent the spread of many misinterpretations and erroneous conclusions.

4 Material and methods

4.1 Fitness landscapes

The mammalian fitness landscape was reconstructed by fitting a mutation-selection model to a multispecies alignment of 14,509 protein-coding genes of 87 mammalian species in Latrille et al. (2023). This fitness landscape is site-specific: the fitness of amino-acids at one site does not depend on amino-acids at other

sites (epistasis is neglected). Importantly, mutation-selection models cannot disentangle selection coefficients from effective population sizes (Rodrigue et al., 2010). The model can only estimate scaled fitness differences, assuming a constant effective population size throughout the mammalian evolutionary history. The experimental fitness landscapes were taken from Deep mutational scanning studies on the virus Influenza (Thyagarajan and Bloom, 2014; Doud et al., 2015), on the bacteria *E.coli* (Stiffler et al., 2015) and on the yeast *S.cervisae* (Kitzman et al., 2015). Except for yeasts, the experimental fitness landscapes have therefore not been obtained in organisms that are particularly subject to gBGC (Mancera et al., 2008), but still reflect selective pressures exerted on proteins in general. These experimental fitness landscapes were retrieved from the study of Bloom (2017).

4.2 Codon equilibrium frequencies

Using a Wright-Fisher diffusion approximation, for a given site l, one can compute the substitution rates from codon i to codon j ($q^{(l)}_{i\mapsto j}$) (Nagylaki, 1983; Halpern and Bruno, 1998).

$$\begin{cases} q^{(l)}_{i\mapsto j} & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ q^{(l)}_{i\mapsto j} & = \mu_{i\mapsto j} \text{ if } s^{(l)}_{i,j} + b = 0 \\ q^{(l)}_{i\mapsto j} & = \mu_{i\mapsto j} \dfrac{4N_e(s^{(l)}_{i,j} + b)}{1 - e^{4N_e(s^{(l)}_{i,j}+b)}} \text{ if } s^{(l)}_{i,j} + b \neq 0. \end{cases} \tag{1}$$

Here, $\mu_{i\mapsto j}$ is the mutation rate from codon i to codon j, N_e the effective population size, $s^{(l)}_{i,j}$ the difference of fitness between codon i and codon j at site l, and b the gBGC coefficient. When codon i and j are separated by a synonymous mutation, I consider that $s_{i,j} = 0$. The gBGC coefficient is the product of the repair bias towards GC, the recombination rate per base pair and the length of the conversion tract. A local increase of recombination rate thus directly increases the intensity of gBGC. If the mutation that separates codon i from codon j is from AT to GC (WS), b is positive, if it is from GC to AT (SW), b is negative, and if it is GC conservative (A $\leftrightarrow$ T or C $\leftrightarrow$ G), $b = 0$. Two mutation matrices have been used: a Jukes-Cantor mutation matrix where all mutation rates are equal, and an empirical mutation matrix estimated from singletons of the human's chromosome 1 extracted from the human 1000 genomes project (Gazal et al., 2015).

One can then build a 61x61 transition matrix for each pair of codons $\boldsymbol{Q^{(l)}}$. The diagonal of this matrix is the negative sum of the $q_{i,j}$:

$$q^{(l)}_{i,i} = -\sum_{j\neq i} q^{(l)}_{i,j} \tag{2}$$

From this transition matrix, the vector of equilibrium frequencies of each codon $\boldsymbol{X^{(l)}}$ can be calculated by solving the following system:

$$\boldsymbol{Q^{(l)}X^{(l)}} = \boldsymbol{0} \tag{3}$$

4.3 DFE of new mutations

At equilibrium, the probability for observing a mutation from codon i to codon j is $x_i^{(l)}\mu_{i\mapsto j}$ where $x_i^{(l)}$ is the equilibrium frequency of codon i at site l. For the whole sequence, the DFE of new mutations represented in Fig.1 has been computed as:

$$\phi_{nm}(s) = \sum_l \sum_i \sum_j x_i^{(l)} \mu_{i\mapsto j} K(s - s_{i,j}^{(l)}) \tag{4}$$

where K is a Gaussian kernel function.

4.4 DFE of substitutions

Similarly to the DFE of new mutations, at equilibrium, the probability of observing a substitution from codon i to codon j is $x_i^{(l)} q_{i\mapsto j}^{(l)}$. For the whole sequence, the DFE of substitutions represented in Fig.2 has been computed as:

$$\phi_{sub}(s) = \sum_l \sum_i \sum_j x_i^{(l)} q_{i\mapsto j}^{(l)} K(s - s_{i,j}^{(l)}) \tag{5}$$

4.5 Relative fitness

The relative fitness at a given site is defined as the average fitness of all codons weighted by their equilibrium frequency, divided by the fitness of the fittest codon. The relative fitness of the sequence f is just the average of the relative fitnesses of all sites:

$$f = \frac{1}{L} \sum_l \frac{\sum_i x_i^{(l)} f_i^{(l)}}{max(f_i^{(l)})} \tag{6}$$

where $f_i^{(l)}$ is the fitness of codon i at site l, and L the total number of sites.

4.6 dN/dS

I computed dN/dS as the ratio between the rates of non-synonymous over synonymous substitutions given by:

$$\begin{cases} dN^{(l)} = \sum_i \sum_j x_i^{(l)} q_{i\mapsto j}^{(l)} \{i \mapsto j \text{ non-synonymous}\} \\ dS^{(l)} = \sum_i \sum_j x_i^{(l)} q_{i\mapsto j}^{(l)} \{i \mapsto j \text{ synonymous}\} \\ dN/dS = \kappa \sum_l \frac{dN^{(l)}}{dS^{(l)}} \end{cases} \tag{7}$$

where κ is a constant normalisation factor corresponding to the ratio of the number of possible synonymous mutations over the number of possible non-synonymous mutations from the equilibrium sequence.

Acknowledgments

I wish to thank Philippe Veber for insightful discussions on mutation-selection models, and Nicolas Lartillot, Carina Farah Mugal and Laurent Duret for their very helpful reviews on a first version of this manuscript. **Funding:** Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec. **Competing interests:** I declare no conflicts of interest. **Data and materials availability:** Analysis scripts and documentation will be available upon deposit of this manuscript on a preprint server.

References

Alleva, B., Brick, K., Pratto, F., Huang, M., and Camerini-Otero, R. D. (2021). Cataloging Human PRDM9 Allelic Variation Using Long-Read Sequencing Reveals PRDM9 Population Specificity and Two Distinct Groupings of Related Alleles. *Frontiers in Cell and Developmental Biology*, 9.

Ashenberg, O., Gong, L. I., and Bloom, J. D. (2013). Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences*, 110(52):21071–21076.

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., Myers, S., Donnelly, P., Przeworski, M., and McVean, G. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*, 336:7.

Axelsson, E., Webster, M. T., Ratnakumar, A., Consortium, T. L., Ponting, C. P., and Lindblad-Toh, K. (2012). Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.*, 22(1):51–63.

Baker, B. S., Carpenter, A. T. C., Esposito, M. S., Esposito, R. E., and Sandler, L. (1976). The genetic control of meiosis. *Annu. Rev. Genet.*, 10(1):53–134.

Bengtsson, B. O. (1990). The effect of biased conversion on the mutation load. *Genetics Research*, 55(3):183–187.

Berglund, J., Pollard, K. S., and Webster, M. T. (2009). Hotspots of Biased Nucleotide Substitutions in Human Genes. *PLOS Biology*, 7(1):e1000026.

Bloom, J. D. (2017). Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol Direct*, 12(1):1–24.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. (2019). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol*, 20(1):5.

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. (2016). Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an Avian System. *Mol Biol Evol*, 33(1):216–227.

Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M., and Petropoulos, C. J. (2004). Evidence for Positive Epistasis in HIV-1. *Science*, 306(5701):1547–1550.

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538.

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400):642–645.

Brown, T. C. and Jiricny, J. (1987). A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, 50(6):945–950.

Bullaughey, K., Przeworski, M., and Coop, G. (2008). No effect of recombination on the efficacy of natural selection in primates. *Genome Res.*, 18(4):544–554.

Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., and Eyre-Walker, A. (2016). Adaptive Evolution Is Substantially Impeded by Hill–Robertson Interference in *Drosophila*. *Mol Biol Evol*, 33(2):442–455.

Castellano, D., Eyre-Walker, A., and Munch, K. (2020). Impact of Mutation Rate and Selection at Linked Sites on DNA Variation across the Genomes of Humans and Other Homininae. *Genome Biology and Evolution*, 12(1):3550–3561.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics*, page 15.

Charlesworth, J. and Eyre-Walker, A. (2007). The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proceedings of the National Academy of Sciences*, 104(43):16992–16997.

Chen, J., Bataillon, T., Glémin, S., and Lascoux, M. (2021). Hunting for beneficial mutations: conditioning on SIFT scores when estimating the distribution of fitness effect of new mutations. *Genome Biology and Evolution*, (evab151).

Choi, K. and Henderson, I. R. (2015). Meiotic recombination hotspots – a comparative view. *The Plant Journal*, 83(1):52–61.

Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M. R., and Besnard, G. (2007). C4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes. *Current Biology*, 17(14):1241–1247.

Corcoran, P., Gossmann, T. I., Barton, H. J., The Great Tit HapMap Consortium, Slate, J., and Zeng, K. (2017). Determinants of the Efficacy of Natural Selection on Coding and Noncoding Variability in Two Passerine Species. *Genome Biology and Evolution*, 9(11):2987–3007.

Davies, K. T. J., Cotton, J. A., Kirwan, J. D., Teeling, E. C., and Rossiter, S. J. (2012). Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity*, 108(5):480–489.

Davis, L. and Smith, G. R. (2001). Meiotic recombination and chromosome segregation in Schizosaccharomyces pombe. *Proceedings of the National Academy of Sciences*, 98(15):8395–8402.

Doud, M. B., Ashenberg, O., and Bloom, J. D. (2015). Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Molecular Biology and Evolution*, 32(11):2944–2960.

Duchemin, L., Lanore, V., Veber, P., and Boussau, B. (2023). Evaluation of Methods to Detect Shifts in Directional Selection at the Genome Scale. *Molecular Biology and Evolution*, 40(2):msac247.

Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genom. Hum. Genet.*, 10(1):285–311.

Eyre-Walker, A. (2002). Changing Effective Population Size and the McDonald-Kreitman Test. *Genetics*, 162(4):2017–2024.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, page 20.

Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6):273–277.

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

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5):1092–1103.

Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E., and Leutenegger, A.-L. (2015). High level of inbreeding in final phase of 1000 Genomes Project. *Sci Rep*, 5(1):17453.

Gerton, J. L. and Hawley, R. S. (2005). Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nat Rev Genet*, 6(6):477–487.

Gillespie, J. H. (1995). On Ohta's hypothesis: Most amino acid substitutions are deleterious. *J Mol Evol*, 40(1):64–69.

Glémin, S. (2010). Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics*, 185(3):939–959.

Gossmann, T. I., Santure, A. W., Sheldon, B. C., Slate, J., and Zeng, K. (2014). Highly Variable Recombinational Landscape Modulates Efficacy of Natural Selection in Birds. *Genome Biology and Evolution*, 6(8):2061–2075.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917.

Handel, M. A. and Schimenti, J. C. (2010). Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat Rev Genet*, 11(2):124–136.

Hartl, D. L. and Taubes, C. H. (1996). Compensatory Nearly Neutral Mutations: Selection without Adaptation. *Journal of Theoretical Biology*, 182(3):303–309.

Hartl, D. L. and Taubes, C. H. (1998). Towards a theory of evolutionary adaptation. *Genetica*, 102(0):525–533.

Hassold, T., Hall, H., and Hunt, P. (2007). The origin of human aneuploidy: where we have been, where we are going. *Human Molecular Genetics*, 16(R2):R203–R208.

Hickey, D. A. and Golding, G. B. (2018). The advantage of recombination when selection is acting at many genetic Loci. *Journal of Theoretical Biology*, 442:123–128.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet Res.*, page 26.

Hoge, C. R., Manuel, M. d., Mahgoub, M., Okami, N., Fuller, Z. L., Banerjee, S., Baker, Z., Mcnulty, M., Andolfatto, P., Macfarlan, T. S., Schumer, M., Tzika, A. C., and Przeworski, M. (2023). Patterns of recombination in snakes reveal a tug of war between PRDM9 and promoter-like features.

Hussin, J. G., Hodgkinson, A., Idaghdour, Y., Grenier, J.-C., Goulet, J.-P., Gbeha, E., Hip-Ki, E., and Awadalla, P. (2015). Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet*, 47(4):400–404.

Hämälä, T. and Tiffin, P. (2020). Biased Gene Conversion Constrains Adaptation in *Arabidopsis thaliana*. *Genetics*, 215(3):831–846.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. (2017). Shifting Balance on a Static Mutation–Selection Landscape: A Novel Scenario of Positive Selection. *Molecular Biology and Evolution*, 34(2):391–407.

Kaj, I., Mugal, C. F., and Müller, R. (2023). A Wright-Fisher graph model and the impact of directional selection on genetic variation.

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L., and Ellegren, H. (2017). Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Mol Ecol*, 26(16):4158–4172.

Keightley, P. D. and Otto, S. P. (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92.

Kimura, M. (1962). ON THE PROBABILITY OF FIXATION OF MUTANT GENES IN A POPULATION. *Genetics*, 47(6):713–719.

Kimura, M. (1969). THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS. *Genetics*, 61(4):893–903.

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. *Nat Methods*, 12(3):203–206.

Kosiol, C., Vinař, T., Fonseca, R. R. d., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. (2008). Patterns of Positive Selection in Six Mammalian Genomes. *PLOS Genetics*, 4(8):e1000144.

Kostka, D., Hubisz, M. J., Siepel, A., and Pollard, K. S. (2012). The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Molecular Biology and Evolution*, 29(3):1047–1057.

Lam, I. and Keeney, S. (2015). Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science*, 350(6263):932–937.

Latrille, T., Joseph, J., Hartasánchez, D. A., and Salamin, N. (2023). Mammalian protein-coding genes exhibit widespread beneficial mutations that are not adaptive.

Latrille, T. and Lartillot, N. (2022). An Improved Codon Modeling Approach for Accurate Estimation of the Mutation Bias. *Molecular Biology and Evolution*, 39(2):msac005.

Liang, Y.-Y., Chen, X.-Y., Zhou, B.-F., Mitchell-Olds, T., and Wang, B. (2022). Globally Relaxed Selection and Local Adaptation in Boechera stricta. *Genome Biology and Evolution*, 14(4):evac043.

Maloisel, L. and Rossignol, J.-L. (1998). Suppression of crossing-over by DNA methylation in Ascobolus. *Genes & Development*, 12(9):1381–1389.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485.

Martin, G. and Lenormand, T. (2006). A General Multivariate Extension of Fisher's Geometrical Model and the Distribution of Mutation Fitness Effects Across Species. *Evolution*, 60(5):893–907.

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654.

Mihola, O., Pratto, F., Brick, K., Linhartova, E., Kobets, T., Flachs, P., Baker, C. L., Sedlacek, R., Paigen, K., Petkov, P. M., Camerini-Otero, R. D., and Trachtulec, Z. (2019). Histone methyltransferase PRDM9 is not essential for meiosis in male mice. *Genome Res.*, 29(7):1078–1086.

Miton, C. M. and Tokuriki, N. (2016). How mutational epistasis impairs predictability in protein evolution and design. *Protein Science*, 25(7):1260–1272.

Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*, 16(1):23–36.

Moses, A. M. and Durbin, R. (2009). Inferring Selection on Amino Acid Preference in Protein Domains. *Molecular Biology and Evolution*, 26(3):527–536.

Murga-Moreno, J., Coronado-Zamora, M., Hervas, S., Casillas, S., and Barbadilla, A. (2019). iMKT: the integrative McDonald and Kreitman test. *Nucleic Acids Research*, 47(W1):W283–W288.

Mustonen, V. and Lässig, M. (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3):111–119.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(20):6278–6281.

Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., and Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mutation*, 32(2):198–206.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–426.

Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, 23:263–286.

Otto, P. and Barton, N. H. (1997). The Evolution of Recombination: Removing the Limits to Natural Selection. page 28.

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

Pardo-Manuel de Villena, F. and Sapienza, C. (2001). Recombination is proportional to the number of chromosome arms in mammals. *Incorporating Mouse Genome*, 12(4):318–322.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome biology and evolution*, 4(7):675–682.

Piganeau, G. and Eyre-Walker, A. (2003). Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. *Proceedings of the National Academy of Sciences*, 100(18):10335–10340.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., and Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science*, 346(6211):1256442–1256442.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552):2571–2580.

Rodrigue, N. and Lartillot, N. (2014). Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7):1020–1021.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10):4629–4634.

Roman, H. (1985). Gene conversion and crossing-over. *Environmental Mutagenesis*, 7(6):923–932.

Rousselle, M., Laverré, A., Figuet, E., Nabholz, B., and Galtier, N. (2019). Influence of Recombination and GC-biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals versus Birds. *Molecular Biology and Evolution*, 36(3):458–471.

Roze, D. (2021). A simple expression for the strength of selection on recombination generated by interference among mutations. *PNAS*, 118(19).

Roze, D. and Barton, N. H. (2006). The Hill–Robertson Effect and the Evolution of Recombination. *Genetics*, 173(3):1793–1811.

Schield, D. R., Pasquesi, G. I. M., Perry, B. W., Adams, R. H., Nikolakis, Z. L., Westfall, A. K., Orton, R. W., Meik, J. M., Mackessy, S. P., and Castoe, T. A. (2020). Snake Recombination Landscapes Are Concentrated in Functional Regions despite PRDM9. *Molecular Biology and Evolution*, 37(5):1272–1294.

Sella, G. and Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27):9541–9546.

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., McVean, G., and Przeworski, M. (2015). Stable recombination hotspots in birds. *Science*, page 6.

Smagulova, F., Brick, K., Pu, Y., Camerini-Otero, R. D., and Petukhova, G. V. (2016). The evolutionary turnover of recombination hot spots contributes to speciation in mice. *Genes Dev.*, 30(3):266–280.

Smith, J. M. and Haigh, J. (1973). The hitch-hiking effect of a favourable gene. *Genetics Research*, page 13.

Spielman, S. J. and Wilke, C. O. (2015). The Relationship between dN/dS and Scaled Selection Coefficients. *Molecular Biology and Evolution*, 32(4):1097–1108.

Starr, T. N. and Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218.

Stiffler, M. A., Hekstra, D. R., and Ranganathan, R. (2015). Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell*, 160(5):882–892.

Székvölgyi, L. and Nicolas, A. (2010). From meiosis to postmeiotic events: Homologous recombination is obligatory but flexible. *The FEBS Journal*, 277(3):571–589.

Tamuri, A. U. and dos Reis, M. (2022). A Mutation–Selection Model of Protein Evolution under Persistent Positive Selection. *Molecular Biology and Evolution*, 39(1):msab309.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. (2012). Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. *Genetics*, 190(3):1101–1115.

Thyagarajan, B. and Bloom, J. D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300.

Winkler, H. (1930). Die Konversion der Gene : eine vererbungstheoretische Untersuchung. *G Fischer*.

Wu, D.-D., Yang, C.-P., Wang, M.-S., Dong, K.-Z., Yan, D.-W., Hao, Z.-Q., Fan, S.-Q., Chu, S.-Z., Shen, Q.-S., Jiang, L.-P., Li, Y., Zeng, L., Liu, H.-Q., Xie, H.-B., Ma, Y.-F., Kong, X.-Y., Yang, S.-L., Dong, X.-X., Esmailizadeh, A., Irwin, D. M., Xiao, X., Li, M., Dong, Y., Wang, W., Shi, P., Li, H.-P., Ma, Y.-H., Gou, X., Chen, Y.-B., and Zhang, Y.-P. (2020). Convergent genomic signatures of high-altitude adaptation among domestic mammals. *National Science Review*, 7(6):952–963.

Zhen, Y., Aardema, M. L., Medina, E. M., Schumer, M., and Andolfatto, P. (2012). Parallel Molecular Evolution in an Herbivore Community. *Science*, 337(6102):1634–1637.

10

Reconciling molecular and ecological adaptation

10.1 The definition of adaptation

The adaptation of species to their environment is a fascinating phenomenon for both evolutionary biologists and the general public. After centuries of natural history discoveries, we are still amazed at how life has colonised and adapted to almost every environment on our planet. Moreover, understanding how it will continue to do so under the drastic environmental changes we are facing is currently of paramount importance. In the XIXth century, it was proposed that heritable adaptations are made possible by mutations (Lamarck, 1815) and natural selection (Darwin, 1859). When the environment changes, if there is functional variation within a population (provided by mutations), individuals that can reproduce more will pass on their heritable traits, causing a shift towards those traits that best fit the new environment at the population

level (Darwin, 1859). While I anticipate that this definition of evolutionary adaptation will be agreeable to readers from all fields of biology, there is actually a second one that is used almost exclusively in the literature of molecular evolution and population genetics. In this second definition, any mutation advantageous for its bearer is called adaptive, regardless of whether the reason for this mutation being beneficial is changing environments. However, these definitions are not equivalent, and the second one disconnects studies evaluating the rate of adaptation at the DNA level with the biological processes driving adaptation at the ecological level. The earliest use of the second definition that I could find dates back to the pioneering work of Sewall Wright on the concept of the adaptive landscape (Wright, 1932).

10.1.1 Adaptive landscape

The adaptive landscape, also referred to as the fitness landscape, describes all the combinations of genotypes possible and the possibility of evolution from one to another by means of mutation and natural selection. The fitness landscape can be represented as a graph where each genotype is a node defined by a fitness in a given environment. For a given genotype, neighbours are any other genotype that can be accessed with one mutation. For a given individual, any mutation that goes towards a genotype of higher fitness is more likely to reach fixation in the population. By means of mutations and natural selection, at the population level, the population climbs this adaptive landscape through an "adaptive walk" towards a fitness peak (Figure 10.1).

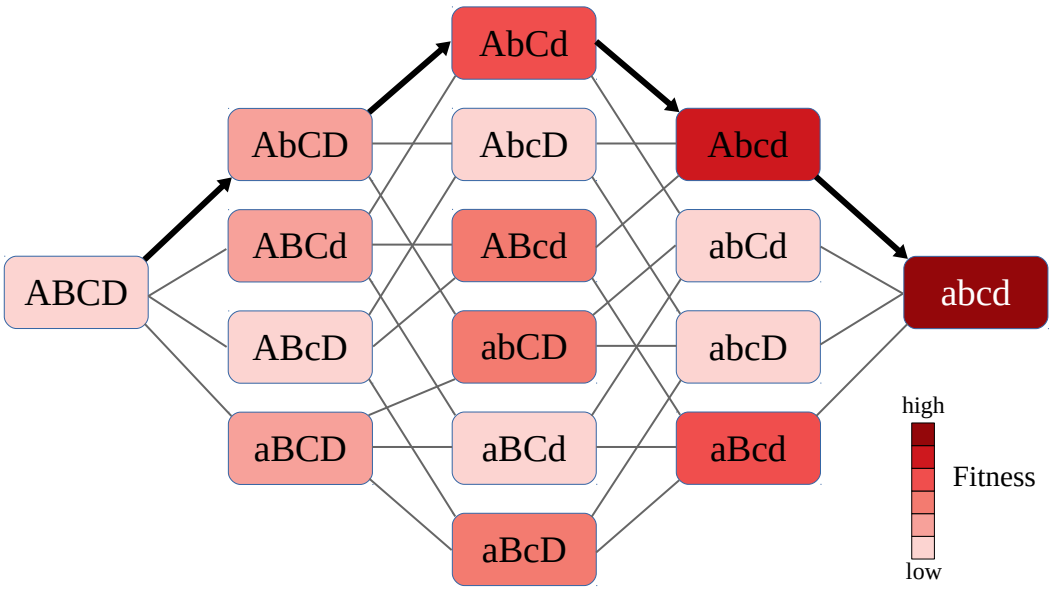


Figure 10.1: *Schematic representation of a 4 sites bi-allelic fitness landscape. The black arrows represent an example of a so-called adaptive walk.*

In the adaptive landscape model, "adaptive" is merely used as a synonym of "positively selected", that only means that the allele favoured by selection is the one produced by a new mutation, which allows the individual to "climb" the fitness landscape. Under certain assumptions, the equality between adaptation to changing environments and positive selection is justified: if species are always at the top of a fitness peak in a stable fitness landscape, therefore the only reason for a species to find itself below a peak and climb up is a change in the fitness landscape. However, when acknowledging the influence of genetic drift or other evolutionary forces on genome evolution, this assumption is misleading. In fact, populations or individuals find themselves at an equilibrium between mutation, selection, drift and other non-adaptive evolutionary forces such as biased gene conversion (BGC) (Hartl and Taubes (1998); Sella and Hirsh (2005) chapter 9). This equilibrium is necessarily below the fitness peak. Climbing up towards this peak can either be due to a change in the fitness landscape (resulting from a change in the environment), or a change in the non-adaptive forces that define this equilibrium (Charlesworth and Eyre-Walker, 2007; Jones *et al.*, 2017).

10.1.2 Distinction between adaptation and positive selection

The vocabulary that evolutionary biologists can rely on to describe evolutionary processes is rich but usually not very well defined. This is a natural consequence of the emergence of new concepts that appeared independently at different times and in different fields. There is no need to have a clear definition of a concept until we reach the point where we need to refine our understanding of that concept to understand how it articulates with others.

With the dramatic increase in genomic and phenotypic data, in addition to quantifying the footprints of natural selection, we can now investigate the ecological, demographic or genetic processes that affected the evolution of species and left those footprints of natural selection in their genomes. We have therefore reached a point where we can interpret the signatures of positive selection found in substitutions, polymorphisms or even time series in terms of the biological generating process, and this process is not necessarily a response to changing environments. It is therefore more important than ever to clarify what we call adaptive and non-adaptive evolution to facilitate cross-talk between ecology and molecular evolution.

Distinguishing adaptation-driven from non-adaptation-driven positive selection is of particular importance for understanding the impact of positive selection on evolution

at macro-evolutionary scales. Positive and negative selection, at the micro-evolutionary scale, describe the fate of a mutant in a population.

The new mutant is either favoured (positively selected) or disfavoured (negatively selected). In fact, when there is natural selection, there is always one allele that is favoured and another one that is disfavoured, whether we call it positive or negative selection only depends on the ancestral state. In the fitness landscape framework, while positive selection elevates the population in the fitness landscape, negative selection theoretically eliminates deleterious mutants and maintains the population at the same position. In practice, the efficiency of negative selection is limited by drift and non-Mendelian segregation, leading to the fixation of deleterious mutants, and the descent of individuals in the fitness landscape (Nagylaki, 1983). As individuals do not indefinitely fall down in the fitness landscape, there is necessarily positive selection at the same time but potentially at other sites, that maintains the individuals at a given equilibrium, in a constant environment (Gillespie, 1995; Hartl and Taubes, 1996; Sella and Hirsh, 2005; Charlesworth and Eyre-Walker, 2007; Mustonen and Lässig, 2009; Razeto-Barry *et al.*, 2012; Jones *et al.*, 2017).

Adaptive evolution is the phenomenon that generated (and still generates) the great diversity of form and functions in the tree of life. Indeed, even if functional variation is expected to occur under a stable fitness landscape, it is only transient, as selection will eventually bring back the phenotype at equilibrium (Martin and Lenormand, 2006; Silander *et al.*, 2007; Tenaillon *et al.*, 2007; Amado and Bank, 2023). For functional and phenotypic diversity to be created and maintained over macro-evolutionary timescales, fitness landscapes are necessarily variable, which is induced by variations in the species' environment (Hietpas *et al.*, 2013; Bajić *et al.*, 2018; Amado and Bank, 2023). Of note, genes under adaption during environmental changes are necessarily functionally linked to the new environment of the species. On the other hand, I propose to define purifying selection, at the macro-evolutionary scale (to be distinguished from negative selection at the population scale) as the selection that continuously purifies away misfit genotypes in a fixed fitness landscape whether they are ancestral or derived, leading to either positive or negative selection in a population (Hartl and Taubes, 1996; Mustonen and Lässig, 2009; Jones *et al.*, 2017). In this view, purifying selection is the molecular equivalent of stabilizing selection on a phenotype. Of note, in a fixed fitness landscape, genes under positive selection can have any function that is not necessarily linked to the species-specific environment.

At the molecular level, by using the term adaptation for both positive selection under a changing environment and under a non-changing environment, we put under the same

name processes that generates phenotypic and functional diversity (adaptive evolution) with processes that deplete it (purifying positive selection).

In fact, in the macro-evolution literature, this distinction is already clearly made: a classical test of adaptive evolution on a trait consists in assessing whether species that evolve in different environments evolve towards different traits optima (Hansen, 1997; Grabowski *et al.*, 2023). If all species seem to evolve under a unique optima, the adaptive hypothesis is rejected, and it is usually concluded that this trait is under stabilizing selection. Calling beneficial back-mutations adaptive would be equivalent to say that, under purely stabilizing selection on a trait, the part of the Brownian noise that brings the trait closer to the selective optima is a signature of adaptive evolution. If we are to understand how natural selection has shaped the diversity of forms and function of living organisms, we also need to clarify the distinction between adaptive evolution and stabilizing selection at the molecular level.

10.2 The detection of adaptation to changing environments at the molecular level

In practice, it is already difficult to estimate correctly the prevalence of positive selection from molecular data. Estimating the contribution of adaptation to positive selection is therefore even more challenging (Latrille *et al.*, 2023b). A common approach, named "the forward genetics of adaptation" in Bomblies and Peichel (2022), consists in detecting adaptation at the phenotypic level, and then identifying genomic changes that are associated with this adaptive evolution of the trait: the genetic architecture of adaptation.

It is indeed easier to detect adaptive evolution on a macroscopic trait rather than DNA. In the best case scenarios, one has direct observations that the value or state of the trait in population 1 increases fitness in its environment, while the value or state of the trait in population 2 increases fitness in another environment. However, these approaches also face several important challenges, mainly because population structure and heterogeneity in recombination rates can artefactually increase the number of regions/loci associated with the adaptive phenotype. Theses approaches have been extensively reviewed (e.g. Bomblies and Peichel (2022)), and are not the main focus of this discussion. Indeed, we do not seek to detect adaptation at the phenotypic level, but at the molecular level. The main advantage of this latter approach called the "reverse genetics of adaptation" is that it does not look at adaptation on a particular phenotype, but instead assesses the role of adaptation to changing environments in

shaping our genomes more generally (Bomblies and Peichel, 2022). For this, one can use the different signatures left by the rise of a beneficial allele on genomes, while making sure that they cannot be explained by the action of non-adaptive forces in a fixed environment.

10.2.1 Selective sweeps

When an allele is strongly beneficial, it can increase in frequency so rapidly that recombination do not have time to break the haplotype leading to a local depletion in diversity around the allele after fixation; which is called a selective sweep (reviewed in Stephan (2019)). For a sweep to be detected, mutations need to be beneficial enough to deplete diversity in a large region. Beneficial back-mutations revert previous deleterious fixed changes. Deleterious fixed changes are usually only slightly deleterious, otherwise they would not have fixed (Hartl and Taubes, 1996; Charlesworth and Eyre-Walker, 2007), and thus, beneficial back-mutations are usually weakly beneficial. Therefore, I believe a strong selective sweep still suggests an adaptive fixation rather than one induced by a beneficial back-mutation. Soft sweeps, on the other hand is not an evidence of polygenic adaptation to changing environment. However, local depletion of diversity could also result from population size contraction or population structure (Moinet *et al.*, 2022; Schlichta *et al.*, 2022). Thus, one might want to be careful when interpreting a local diversity drop as evidence for adaptation, without information on the population demographic history and structure.

10.2.2 Signatures of accelerated evolution

Measuring the rate of adaptation has been central to inform the neutralist-selectionist debate with empirical data. In particular, the estimator α was designed to quantify the proportion of non-synonymous substitutions that have been fixed because they provided a fitness advantage, in contrast to those that have been fixed by drift, whether they were neutral or slightly deleterious (McDonald and Kreitman, 1991). α has been shown to vary between species, but mainly because of differences in the substitution rate of slightly deleterious mutations which decreases in species with high effective population sizes, and mechanically increase the proportion of adaptive ones (Galtier, 2016). Another parameter that could be more informative on the role of adaptation in molecular evolution is ω_a, which represent the rate of fixation of beneficial non-synonymous mutations compared to synonymous ones supposedly neutral. This estimator was first introduced by McDonald and Kreitman (1991) in the so-called MK-test, and was then progressively extended (Eyre-Walker and Keightley,

2007; Galtier, 2016; Tataru *et al.*, 2017).

The general idea behind the test is to compare the dN/dS ratio (substitutions) to the $\pi N/\pi S$ ratio (polymorphisms) McDonald and Kreitman (1991). The $\pi N/\pi S$ is supposed to provide a nearly-neutral expectation of the rate of evolution under negative selection. Adaptive substitutions are assumed to be fixed very rapidly and therefore not observed in polymorphism. The difference between the dN/dS and the $\pi N/\pi S$ should therefore be explained by adaptive substitutions. One major issue with this approach is the time lapse between the dN/dS estimated on a branch, and the $\pi N/\pi S$ estimated at the tip. If the mutation-selection-drift-BGC equilibrium has shifted, because of an increase/decrease of population size or genome-wide recombination rate, then, $\pi N/\pi S$ is no longer a correct estimate of the strength of purifying selection, which can lead to weirdly negative α or ω_a (Eyre-Walker, 2002; Latrille *et al.*, 2023a). Moreover, as previously said and shown in chapter 5, dN/dS (and probably $\pi N/\pi S$) are poor measures of the efficiency of selection in the presence of gBGC.

Overall, a genome-wide estimate of ω_a is difficult to compare between species, because it is influenced by long-term fluctuations of effective population size (Eyre-Walker, 2002; Charlesworth and Eyre-Walker, 2007; Soni *et al.*, 2022; Latrille *et al.*, 2023a). But this "rate of adaptation" can still prove to be useful when compared across genes of a same genome. Indeed, as a first approximation, changes in effective population size will affect equally all genes. Instead of focusing on the parameter ω_a, one can compute the deviation of ω_a from the genome average ($\Delta\omega_a$) (Latrille *et al.*, 2023a). However, to have enough statistical power, this needs to be performed on several genes (for instance genes involved in a function suspected to have changed along the branch of a given species) for which we want to test if they were accelerated.

Altogether, in the face of changes in mutation-selection-drift-gBGC equilibrium, signatures of accelerated evolution do not allow us to evaluate the genome-wide rate of adaptation of a species in a given branch, because they cannot disentangle the signatures of adaptive evolution with those of non-adaptive processes. However, considering that non-adaptive forces affect the genome equally, comparing ω_a across genes could still provide useful information on potential adaptations (Latrille *et al.* (2023a) but see Soni *et al.* (2022) and chapter 5).

10.2.3 Modelling the changes in fitness landscapes

Alternatively, understanding the impact of environmental change to molecular evolution can be achieved by estimating the rate at which fitness landscapes evolve. When looking at the genome-wide scale, the fitness landscape is probably changing all the time because of all kinds of fluctuating selection. Therefore, the question of how fast the genome-wide fitness landscape changes is of very limited interest. A more interesting question concerns how fast the fitness landscape of a given site, gene or set of genes evolve. For instance, in a given clade, what proportion of genes/sites have conserved the same fitness landscape? Are there genes that change fitness landscape all the time? Can we associate the change in the fitness landscape of a given gene/site to a phenotypic innovation?

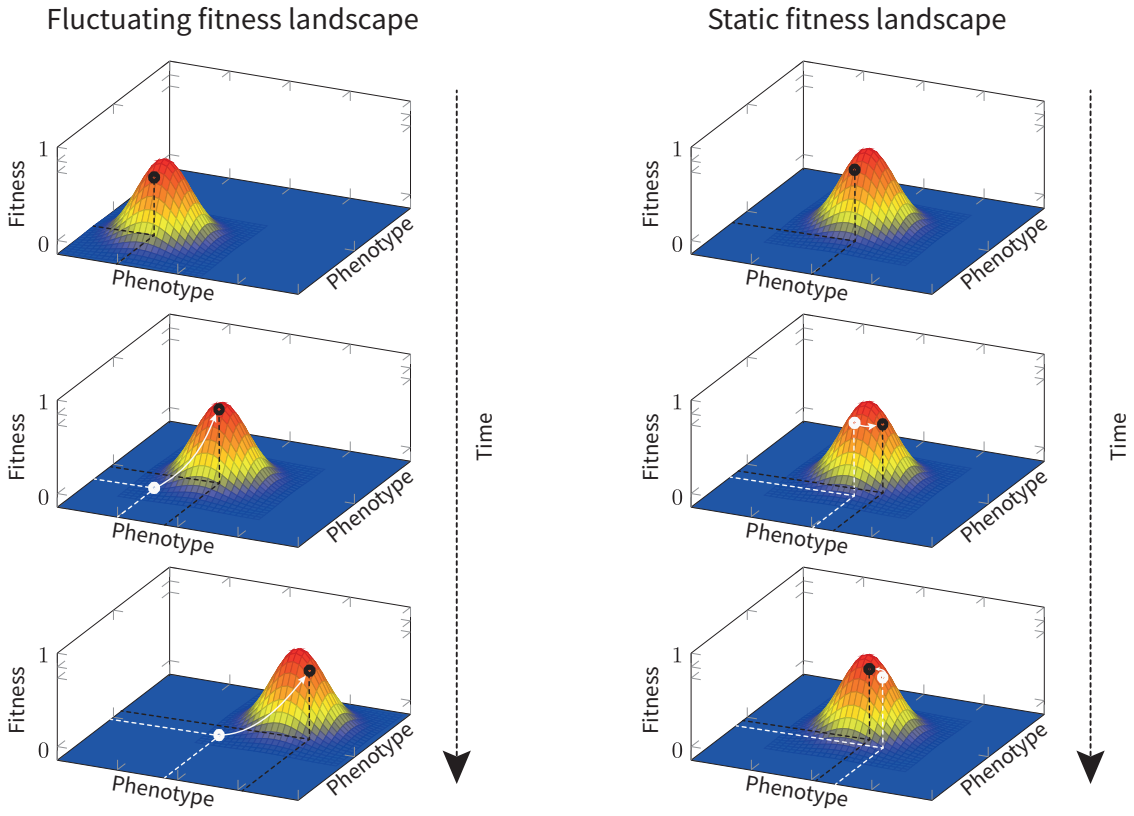


Figure 10.2: *Schematic representation of a 2 phenotype Fisher's fitness landscape. © Thibault Latrille*

By explicitly modelling the fitness landscape, and integrating the contribution of both fitness differences and non-adaptive forces to the probability of a switch between states, mutation-selection models provide an appropriate modelling framework to answer these questions at large time scales (Halpern and Bruno, 1998; Rodrigue *et al.*, 2010; Tamuri *et al.*, 2012; Rodrigue and Lartillot, 2014; Latrille *et al.*, 2021; Latrille and Lartillot, 2022; Tamuri and dos Reis, 2022). These models represent an approximation of Wright's fitness landscape in the case where the amino-acid fitness landscapes at each site of a protein

are independent. The fitness of the sequence is therefore described by the additive effects of the fitness at all sites. This site-independent model approximation is quite convenient mathematically, and proved to be a reasonable approximation of the fitness landscape of protein-coding genes (Ashenberg *et al.*, 2013; Doud *et al.*, 2015; Bloom, 2017; Latrille *et al.*, 2023b). However, where possible, I think it is always better to model selection explicitly at the level of phenotypes, where one expects it to act.

Fisher's geometric model provides a valuable framework for achieving this goal (Fischer, 1930). In Fisher's geometric model, instead of associating fitness to genotype, fitness is defined at the phenotypic level. The traits of an individual are projected on a n-dimensional continuous phenotypic space. This space is characterized by an optimum and a function that describes the decrease in fitness with the distance from this optimum. Mutations have a given distribution of effect size on the phenotypes through which they affect fitness. On top of providing a modelling framework closer to the representation we have of natural selection, this framework genuinely accounts for fitness interactions between sites: epistasis (Tenaillon, 2014).

10.2.4 Consequences of epistasis

Epistasis can arise when the fitness of a given mutation of the genome depends on the genotype at other loci. To account for this, some studies have used the statistical framework of direct coupling analysis (DCA) to not only estimate the site-specific fitness landscape, but also to estimate the fitness interactions between sites to some extent (Weigt *et al.*, 2009; Baldassi *et al.*, 2014; Bisardi *et al.*, 2022; Vigué *et al.*, 2022). However, these models assume that sites are at their equilibrium frequencies and do not account for the influence of phylogenetic inertia or biased gene conversion. Moreover, the computational cost of evaluating interaction only allows for testing interaction within a gene, or between two genes but is not applicable to detect interactions genome-wide. Despite these limitations, I think there is room for improvement, and the features of the model we used in chapter 4 could be included in DCA. Without an explicit phenotype-genotype-fitness map, this framework could be valuable to measure the evolution of fitness landscapes, while accounting for interdependence between sites.

From a conceptual perspective, under a gene-centered view of evolution, one can consider that epistasis is the adaptation of genes to changes elsewhere in the genome by considering the states at other interacting genes as part of the gene's environment. The same reasoning can be applied to different interacting sites in a given protein. In turn, under an individual-centered view, even if a change at one site impacts the

fitnesses at other sites, the overall fitness landscape has not changed. In the end, whether we can call a compensatory change due to epistasis adaptive depends on the scale at which we define the fitness landscape. This ambivalent status of epistasis takes it roots in the more fundamental difficulty to define the object of selection and therefore its environment. This definition might depend on the biological question of the study. For instance, if the insertion of transposable elements (TEs) generates a selective response at other genes to limit their spread in the genome, one might want to call this response adaptive by considering the TE as a pathogen which is part of the environment. Now, if the same insertions slightly disrupt the function of genes, which is compensated by substitutions at other sites of these genes, one might consider that the selective response arises from epistasis rather than adaptation.

Altogether, whether we include compensatory responses in adaptation depends on the context, and this problem is much less of a problem if the environment that generates the adaptive response is clearly defined. In contrast, beneficial back-mutations do not occur because of the change elsewhere in the genome, but because of a deleterious substitution at the same site. Because the environment is defined as everything outside of a system, a nucleotide cannot be part of its own environment. Therefore, contrarily to compensatory mutations, beneficial back-mutations cannot be considered as adaptive. Of note, epistatic compensatory mutations still induce inter-specific genetic diversity by encoding the same phenotype under different genotypes, while beneficial back-mutations reduce genetic diversity by replacing rare deleterious nucleotides by common advantageous ones.

10.3 Conclusion

Altogether, to describe the rise in frequency of a new beneficial allele, allowing populations to climb in the fitness landscape, I think it is best to stick to the term that is perfectly adapted: positive selection. And I suggest we restrain our use of adaptation or adaptive to the cases when we have some evidence that the fitness advantage of the new allele is induced by changing environments (which have to be defined). With this definition, we will be able to understand how positive selection at the DNA level contributes to between species trait diversification in diverse environments. I am aware that the use of adaptation as positive selection has a long history in molecular evolution and population genetics. I hope that if after reading this piece, population geneticists are still not convinced that the benefits of distinguishing adaptation from positive selection outweigh the costs of breaking with this long-lasting tradition, they will at least clearly define adaptation or adaptive in their study to avoid

misinterpretation or erroneous conclusions across different fields of evolutionary biology.

Part IV

The evolutionary origin of GC-biased gene conversion

11

The evolution of GC-biased gene conversion by means of natural selection

Context

While reading this manuscript, at some point, I hope that the reader has begun to think, as I have, about how much better we could be without gBGC and to wonder why we have to suffer this burden anyway. During my last follow-up comity, at the beginning of my final year of PhD, we put on the table all the potential adaptive and nonadaptive arguments we could think of to explain the existence of gBGC. Unfortunately, it appeared that nonadaptive arguments were quite difficult to test. On the other hand, modelling the effect of natural selection on a modifier of gBGC to see what value it evolved to, why, and what burden it implied, seemed a much more achievable goal. Therefore, Nicolas and I decided to propose the subject to a very enthusiastic intern from the ENS de Lyon (Augustin Clessin). In only four months, he had time to learn to code in C++, code a simulator and derive semi-analytical approximations for the action of selection on a modifier of gBGC (which is a great amount of work). As he did not have enough time to do more, Nicolas and I performed the subsequent analysis and wrote a manuscript, which I hope the reader will enjoy.

Detailed contributions

This study was co-designed by Nicolas, Augustin and myself. Augustin developed the simulator and derived equations to compute the selective pressure exerted on a modifier of gBGC. Most analysis were then performed by Nicolas, and I performed the analysis on the genetic load. Nicolas and I wrote a first draft of the results and materials and methods of the analysis we each carried out. I wrote the first draft of introduction, abstract, and nearly all the discussion. We jointly revised all the parts of the manuscript. This manuscript has not been submitted to any journal yet.

The evolution of GC-biased gene conversion by means of natural selection

A. Clessin[1], **J. Joseph**[1], **N. Lartillot**[1]

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

nicolas.lartillot@univ-lyon1.fr

October 17, 2023

Abstract

GC-biased gene conversion (gBGC) is a recombination-associated evolutionary force that biases the segregation ratio of AT:GC polymorphisms in the gametes of heterozygotes, in favour of GC alleles. This non-adaptive process is the major determinant of base composition in humans and can be the cause of a substantial burden of GC deleterious alleles. While the importance of GC-biased gene conversion in molecular evolution is increasingly recognised, the reasons for its existence and its variation between species remain largely unknown. Using simulations and semi-analytical approximations, we investigated the evolution of gBGC as a quantitative trait evolving by mutation, drift and natural selection. We show that in a finite population where most mutations are deleterious, gBGC is under weak stabilising selection around a positive value that depends on the intensity of the mutation bias and on the selective constraints exerted on the genome. Importantly, the levels of gBGC that evolve by natural selection do not minimize the load in the population, and even increase it substantially in regions of high recombination rate. Therefore, despite reducing the population's fitness, levels of gBGC that are currently observed in humans could in fact be positively selected: a genetic tragedy of the commons.

***K*eywords** gBGC · Recombination · Modifier · Genetic load · Mutation bias · Natural selection

1 Introduction

In meiosis, during the repair of double strand breaks (DSBs), the single stranded DNA from the broken chromosome invades the other such that the two form a double stranded DNA chimera (heteroduplex) of the two parental chromosomes. At this location, if the individual is heterozygous, there will be a mismatch (non Waston and Crick pairing). This mismatch can be resolved by repairing either parental allele with the other. This phenomenon therefore induces gene conversion (Winkler, 1930; Roman, 1985). In the late 80's Brown

and Jiricny (1987) found that in human and green monkey cells, gene conversion was biased towards GC alleles. Since then, direct and indirect evidence for this bias have been found in a wide range of eukaryotes (Mancera et al., 2008; Duret and Galtier, 2009; Pessia et al., 2012; Smeds et al., 2016; Clément et al., 2017; Galtier et al., 2018; Boman et al., 2021). GC-biased gene conversion (gBGC) is therefore a special case of non-Mendelian segregation where recombination and DNA repair machineries act as segregation distorters (Nagylaki, 1983; Bengtsson and Uyenoyama, 1990). Most methods that detect selection or infer demography from genetic data are based on the assumption of Mendelian segregation, and gBGC therefore confounds both selection and demography inference (Galtier and Duret, 2007; Ratnakumar et al., 2010; Kostka et al., 2012; Pouyet et al., 2017, 2018; Bolívar et al., 2019). Moreover, it has been demonstrated, notably in humans and birds, that gBGC is the major determinant of GC content variations along the genome (Galtier et al., 2001; Meunier and Duret, 2004; Webster et al., 2006). Despite its major impact on genome evolution, the evolutionary origins of gBGC and the reasons for its maintenance remain quite uncertain. Bengtsson (1986) made the prediction that if gene conversion could be biased against the most common class of mutations, it could provide an advantage by reducing the genetic load. GC $\mapsto$ AT mutations being the most common type in most species (Long et al., 2018), it has thus been naturally hypothesized that gBGC could have been selected as a correction mechanism that counteracts the almost universal mutational bias towards AT (Glémin, 2010). However, Glémin (2010) demonstrated that the levels of gBGC that should minimize the load are very weak compared to empirical values observed in regions of high recombination rate.

In fact, empirical studies so far are quite unanimous on a mostly deleterious effect (Berglund et al., 2009; Galtier et al., 2009; Necşulea et al., 2011; Lachance and Tishkoff, 2014; Bolívar et al., 2016). Having a mechanism that seems to be mostly deleterious being so widespread in eukaryotes is therefore quite paradoxical. Interestingly, both in angiosperms and animals, studies observed a negative correlation between the transmission bias b, and effective population size (Clément et al., 2017; Galtier et al., 2018). Galtier et al. (2018) proposed that this pattern could be explained by a drift barrier hypothesis, whereby gBGC is a deleterious process which can be efficiently counter-selected only in species whose effective population size is high. However, as the way mutation, drift and selection affect the evolution of gBGC lacks theoretical expectations, this argument is verbal and requires theoretical validation. Bengtsson and Uyenoyama (1990) investigated the evolution of a modifier of biased gene conversion (BGC) under different scenarios, and recovered that a positive value of BGC evolves naturally when mutation is biased. However, this study was conducted under the approximation of infinite population sizes and at a single strongly selected locus. Therfore, this study do not allow to explain the variations of gBGC between species of different population sizes.

To tackle this question in a more realistic setting, we developed a model in which the intensity of gBGC evolves freely as a quantitative trait that affects the whole genome in a finite population. We confirm that in the presence of a mutational bias towards AT, gBGC naturally evolves towards positive values (Bengtsson and Uyenoyama, 1990). As expected, the equilibrium value of the transmission bias towards GC depends both on the intensity of the mutational bias towards AT and on the magnitude of selective constraints exerted on the genome. Interestingly, we predict that the equilibrium value of the transmission bias correlates negatively with effective population size, suggesting that high levels of gBGC are not only more efficiently counter

selected in high-N_e species, but also more deleterious. Importantly, we show that even if gBGC leads to a higher deleterious burden at the population level, this does not mean that it is negatively selected, even in high N_e species. In the present model, high gBGC intensity results from a tragedy of the commons where the short-term advantage of converting AT deleterious alleles in heterozygotes leads to a higher deleterious burden in the population. Overall, by capturing the selective pressures acting on gBGC under empirically realistic conditions, this model provides insight into the role of natural selection in shaping the evolution of gBGC in eukaryotes.

2 Results

Model summary

A model for the evolution of biased gene conversion was designed and implemented as a simulation program. The model is meant to represent a population of randomly-mating diploid individuals, of fixed size N, evolving under a typical nearly-neutral regime, that is, under purifying selection against deleterious mutations susceptible to occur over a broad (gamma-distributed) range of selective effects, from very weak to very strong (Ohta, 1992; Eyre-Walker and Keightley, 2007). Those mutations occur over a set of bi-allelic loci, with allelic states W, or Weak (corresponding to AT), and S, or Strong (corresponding to GC). The model assumes a mutation bias λ, which will be typically in favour of Weak alleles (so as to mimic the mutational bias in favour of AT seen across many eukaryotic species (Long et al., 2018)). Selection, on the other hand, is statistically balanced with respect to either W or S, in the sense that, for each locus, either W or S is randomly chosen to be the deleterious allele with probability $1/2$.

On top of this nearly-neutral background, the model invokes a modifier locus, encoding an additive quantitative trait modulating biased gene conversion. Specifically, the locus determines the value of the conversion bias parameter β, which will play during meiotic recombination as follows: in addition to a unique cross-over uniformly chosen along the chromosome, a certain fraction of the genome undergoes gene conversion at rate α per nucleotide position. If a position somewhere in the genome is heterozygous and happens to undergo gene conversion, then the W allele is converted into the S allele with probability $(1+\beta)/2$, and conversely, the S allele is converted into the W allele with probability $(1-\beta)/2$. As a result, the net strength of biased gene conversion, defined as the net excess of transmission of S alleles, relative to W, at a WS heterozygous position, is $b = \alpha\beta$.

The basal rate of gene conversion, α, is assumed to be fixed, possibly because of specific constraints related to the molecular mechanisms of meiosis. The conversion bias β, on the other hand, is allowed to evolve, by introducing mutant alleles at the modifier locus (at rate w) contributing a small shift, either positive or negative, in the value of β. As a result of this mutational input, biased gene conversion is susceptible to show variation among individuals. The whole question is then whether this genetically-encoded variation in gBGC is in turn subject to indirect selection, and whether this results in predictable patterns of evolution of gBGC in the long run.

Biased gene conversion is under stabilizing selection

Typical trajectories of the population-mean of biased gene conversion (b) under the model are shown in Figure 1. Here, a mutational bias of $\lambda = 3$ is considered (bias in favour of Weak, or AT alleles), with a basal mutation rate of $u = 10^{-4}$ (for W to S mutations), a population size of $N = 1000$, a genome consisting of $L = 10000$ selected loci, with a gamma distribution of selective effects of mean $\bar{h}s = 0.01$ and shape 0.2. Two alternative settings are considered for the dominance effect of those mutations: either co-dominant ($h = 0.5$) or partially recessive ($h = 0.1$). In both cases, the modifier locus undergoes mutations at a rate of $w = 10^{-3}$ per generation, with effect sizes of mean 0.1 on β. Finally, the basal gene conversion rate is equal to $\alpha = 0.1$. Of note, these parameter values are not meant, at that stage, to match any specific empirical situation. Instead, the aim is to reveal the inner workings of the model, and how its output relates to the input parameter values.

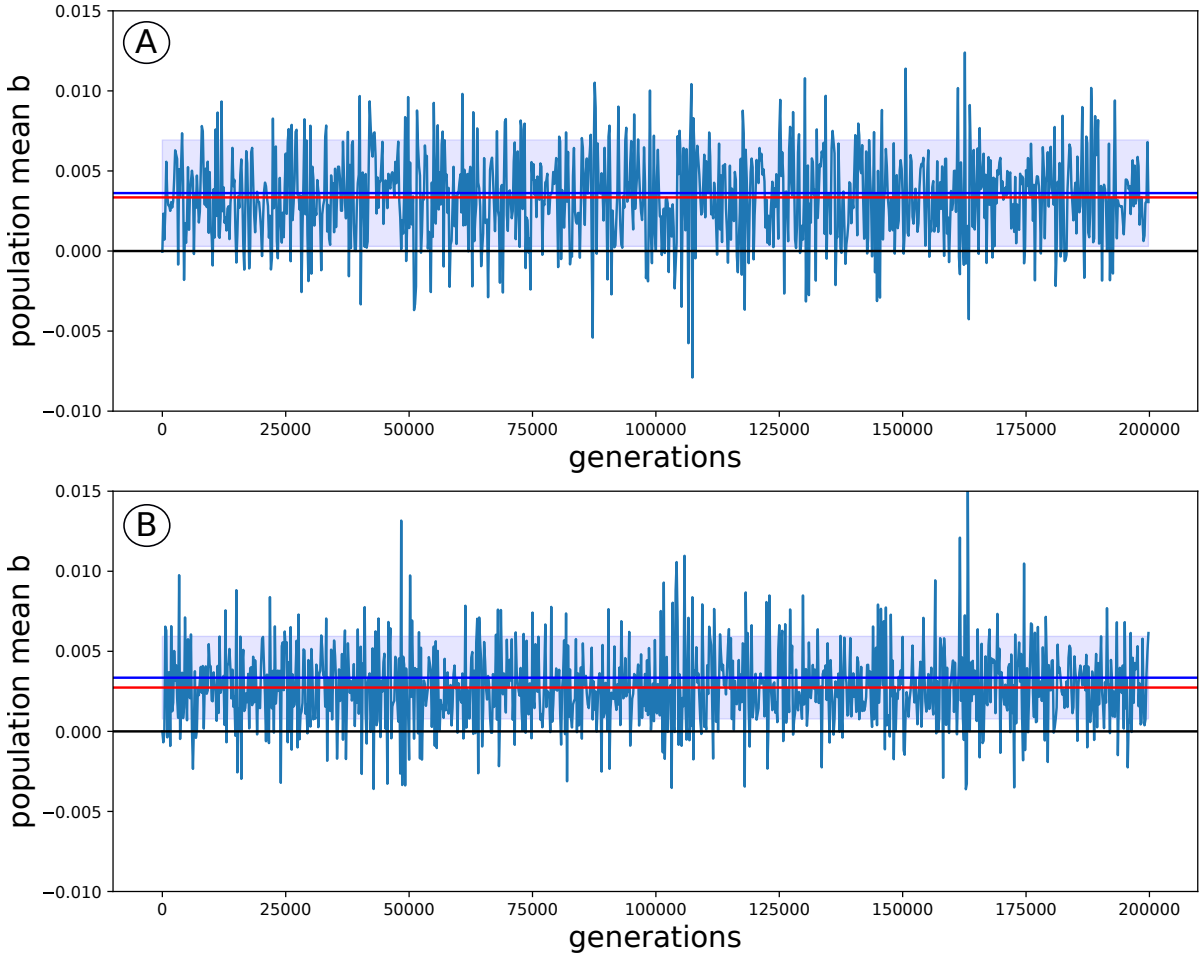


Figure 1: Evolution of the strength of gBGC (mean of b over the population) over the generations, for the co-dominant (A) and recessive (B) cases. Red horizontal line: mean over the entire run; blue horizontal line: equilibrium value predicted by the analytical approximation; light-blue shaded area: predicted equilibrium variance.

Running the model under these parameter values results in a population-level gBGC evolving towards positive value of b, reaching an evolutionary equilibrium with a long-term mean of the order of $b \simeq 0.003$ (Figure 1). There is a substantial evolutionary variance, such that the population still spends about 10% of the time with negative values of b. Nevertheless, this variance remains much smaller than what is observed when inactivating all selection (i.e. by setting $\bar{s} = 0$), in which case the population visits, over evolutionary times, the entire range of values of b that can be realized by the underlying genetic architecture, that is, from $-\alpha$ to α, with $\alpha = 0.1$ (not shown). These experiments show that gBGC is susceptible to spontaneously evolve in favour of Strong (GC) alleles. They also more specifically suggest the existence of some form of stabilizing selection acting on gBGC, driving the population towards, and maintaining it around, an evolutionary equilibrium.

The mutation-segregation tradeoff between AT- and GC-deleterious mutations

The observations gathered in the last section call for a deeper understanding of what drives the equilibrium value of b, and its variance. Given a mutation bias towards W, it seems relatively straightforward that a conversion mechanism playing blindly against W alleles during meiosis should be error-correcting on average and could therefore be selected (Bengtsson, 1986). What is perhaps less obvious is why selection induced on gBGC modifiers is stabilizing rather than extremal, resulting in an optimal value of b. The fundamental reason for this lies in the feedback of the evolution of gBGC on the segregation frequencies at the selected loci across the genome.

Consider a population initially devoid of gBGC. In this context, modifiers increasing gBGC are selectively favoured due to their error-correcting effect on deleterious polymorphisms, which are primarily towards W. Such modifiers will therefore invade. As a consequence, however, the population starts to live and reproduce under increasingly high levels of gBGC. This in turn changes the frequency at which S and W alleles segregate, increasing the frequency of S and decreasing the frequency of W alleles in the population. This shift in the segregation frequencies of deleterious alleles in favour of S tends to compensate for the mutation bias in favour of W. The balance between these two opposing effects, mutation versus segregation bias, is reached for an intermediate value of b.

This mutation-segregation tradeoff can be mathematically formalized under the assumption that gBGC evolves slowly and that most of the selection induced on gBGC is fundamentally contributed by selected loci that are not strongly linked to the modifier locus (these assumptions are discussed below). The detailed derivation is given in the Methods. Here, the main intuitions are presented and graphically illustrated.

The key is to express the mean selective effect induced on a modifier increasing the value of gBGC by an amount δb, in a population at equilibrium under a strength of gBGC equal to b. This induced selection is here more precisely defined as the difference between the mean fitness of the offspring of an individual bearing the modifier (and thus implementing a gBGC of strength $b + \delta b$ in its meiosis) and the mean fitness of the offspring of an individual not bearing the modifier. For small δb, this difference is proportional to δb and can be written:

$$\delta ln f \quad = \quad G(b)\delta b. \tag{1}$$

If $G(b)$ is positive, then modifiers increasing b will be favoured, and conversely if $G(b)$ is negative. Considering W and S alleles separately, $G(b)$ can be expressed as the difference between the net gain upon converting W alleles $G_W(b)$, and the cost of converting S-deleterious alleles, $G_S(b)$:

$$G(b) \quad = \quad G_W(b) - G_S(b). \tag{2}$$

Both terms are positive, and the sign of $G(b)$ will thus be determined by which of these two contributions, gain or cost, is largest.

Since the selection induced on gBGC is contributed by the entire genome, both $G_W(b)$ and $G_S(b)$ can be expressed as averages over the distribution of selective effects of the mean selective impact of gene conversion events, scaled by the number of positions under selection, which is $L/2$ for both cases:

$$G_W(b) \quad = \quad \frac{L}{2}\langle H_W(s,b)\rangle, \tag{3}$$

$$G_S(b) \quad = \quad \frac{L}{2}\langle H_S(s,b)\rangle, \tag{4}$$

Here, $H_W(s,b)$ and $H_S(s,b)$ denote the mean selective impact of gene conversion events at loci with selection coefficient s, at equilibrium under a gBGC equal to b. The angle brackets stand for an expectation over the gamma distribution of selective effects.

Finally, in order to account for the stochastic fluctuations in the segregation frequencies of selected loci, the functions $H_S(s,b)$ and $H_W(s,b)$ are themselves expectations over the frequency distribution for W and S alleles, of the expected selective differences contributed in the offspring by conversion events occurring during meiosis on the selected positions that happen to be heterozygous in a typical individual. Thus, taking the case of W-deleterious alleles, let x denote the frequency at which the allele segregates in the population. Under random mating, an individual will be heterozgous for this allele with probability

$$P \quad = \quad 2x(1-x), \tag{5}$$

in which case, accounting for all possible genotypes for the other parent, the mean gain induced by a conversion event at that position in the offspring will be equal to (see methods):

$$C \quad = \quad \frac{s}{2}\left[h + x(1-2h)\right]. \tag{6}$$

At mutation-selection-conversion balance, x is a random variable drawn from an equilibrium frequency distributions noted $\phi^W_{s,b}(x)$, and thus, overall, the mean gain will amount to:

$$H_W(s,b) \quad = \quad \mathbb{E}_{\phi^W_{s,b}}\left[P \times C\right] \tag{7}$$

The same derivation can be conducted in the case of S-deleterious loci. For bi-allelic loci, the equilibrium distributions $\phi^W_{s,b}(x)$ and $\phi^S_{s,b}(x)$, for both W and S loci, can be explicitly written, up to a normalization constant, such that expectations over these distributions can be computed numerically (see methods).

Predicting the equilibrium mean and variance strength of gBGC

The value of $G(b) = G_W(b) - G_S(b)$ can be plotted as a function of the population-level b (Figure 2). This function is decreasing, crossing 0 at an intermediate, positive value of b^*. Numerically solving for the value

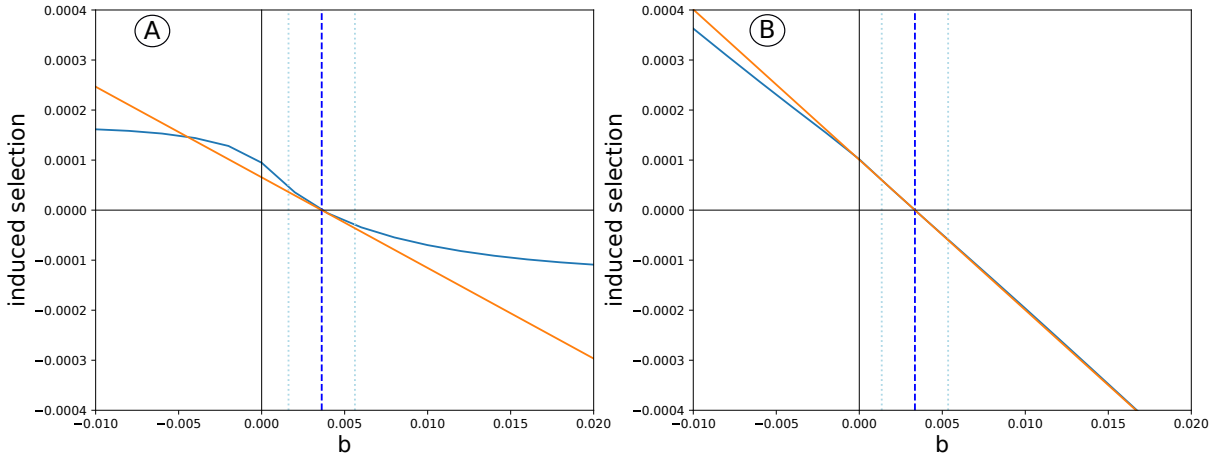


Figure 2: Strength of selection induced on gBGC modifiers, $G(b)$, as a function of b (blue curve), under the co-dominant (A) and recessive (B) settings. Dark blue dotted vertical line: numerically estimated value of b^*, for which $G(b) = 0$; orange line: numerically estimated tangent at b^*; light blue dotted vertical lines: predicted standard deviation around b^*.

b^* such that $G(b^*) = 0$ gives $b^* = 0.0036$ in the co-dominant case, and $b^* = 0.0031$, which is close to the mean value observed in the simulation ($\bar{b} = 0.0033$ and $\bar{b} = 0.0027$, respectively). The numerical and simulation-based estimates are both represented as a blue and red lines, respectively, in Figure 1.

A rough quantitative estimate of the evolutionary variance can also be obtained, based on the slope γ of the tangent to the curve at b^* (Figure 2A). Specifically, the equilibrium evolutionary variance is predicted to be approximately equal to $v_{eq} \simeq \frac{1}{2NL\gamma}$(reported as a shaded area on Figure 1). Of note, the selective response shows a steeper slope at the equilibrium point in the recessive case, resulting in a smaller predicted evolutionary variance than in the co-dominant case.

The drivers of gBGC

The behaviour of the simulation model, along with the analytical approximation just introduced, were further investigated by plotting the predicted equilibrium value of the strength of gBGC, b^*, as a function of several key parameters (mutation bias, mean strength of purifying selection, number of positions under selection and mutation rate). The case of the response of b^* to changes in effective population size is examined further below.

Not surprisingly, the mean equilibrium strength of gBGC is directly related to the strength of the mutational bias (Figure 3A). Owing the symmetry of the problem, running the model with $\lambda < 1$, i.e. under a mutational bias in favour of the Strong alleles results in a population evolving towards a mean conversion bias in favour of Weak (left side of Figure 3A). The mean equilibrium strength of biased gene conversion is also directly influenced by the mean strength of the purifying selection acting over the genome (Figure 3B),thus clearly indicating that its evolutionary dynamics is a direct consequence of the selective effects induced by converting non-neutral polymorphisms in the germ-line. The mean equilibrium value is insensitive

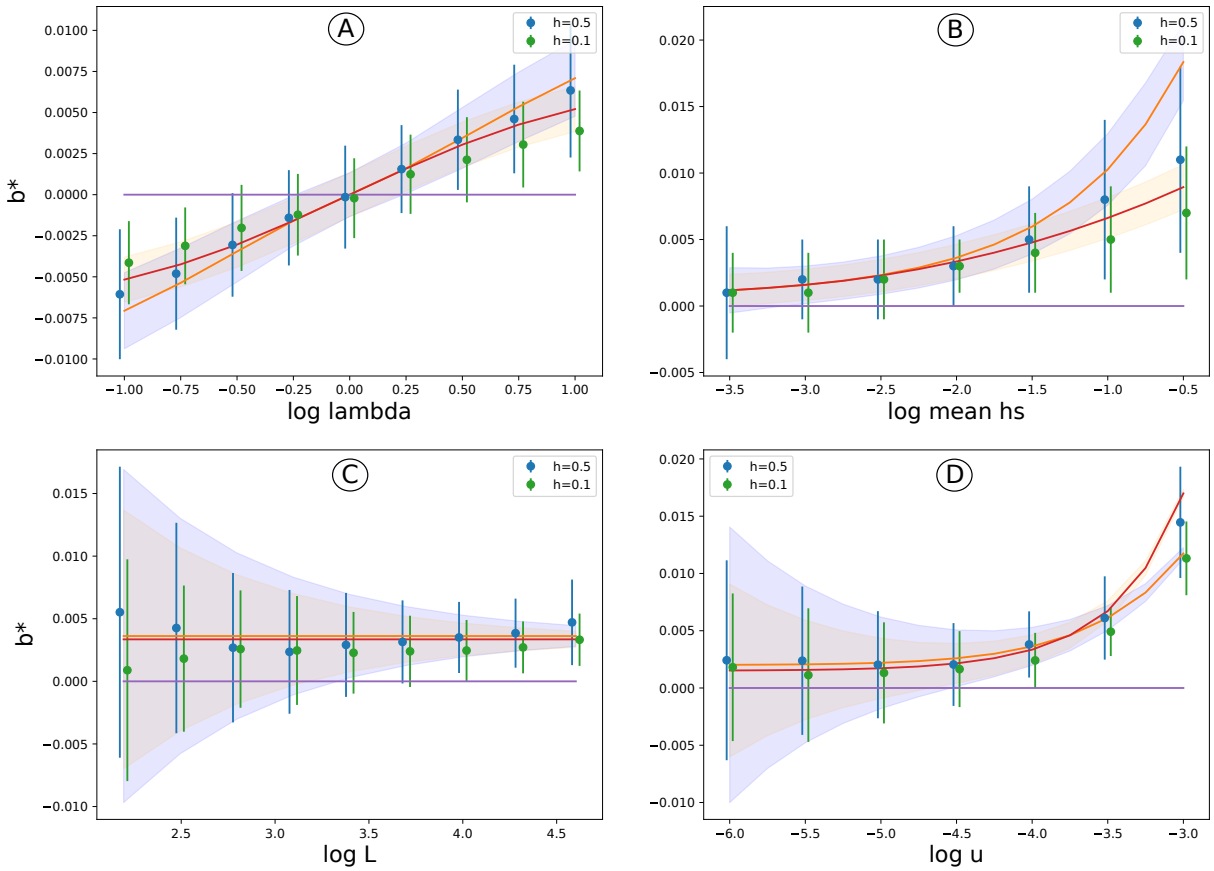


Figure 3: Mean equilibrium b^* and standard deviation, as a function of λ (A), $h\bar{s}$ (B), L (C) and u (D), under the co-dominant (blue) and the recessive (orange) case, obtained by simulations (dots and associated vertical bars) and predicted by the analytical approximation (curve and associated shaded area).

to the number L of selected loci, but its evolutionary variance, on the other hand, is affected, showing a clear decreasing trend with L, which corresponds to the scaling in $1/L$ predicted by the analytical approximation (Figure 3C).

Finally, the strength of gBGC responds very weakly to the mutation rate, except for very high mutation rates ($4Nu >> 1$), in which case it shows a sharp increase (Figure 3D). For low $4Nu$, not so much the mean than the evolutionary variance of gBGC is impacted by the mutation rate, with larger variances being observed under lower mutation rates. In this respect, the response of the model to variation in u is not unlike its response to variation in L (Figure 3C). This similar behaviour can be understood by noting that any indirect selective effect acting on the modifier locus can only be mediated by heterozygous positions. Thus, the strength of induced selection will be directly determined, not just to L, but more fundamentally, by the mean number of selected positions at which a typical individual is heterozygous. The mean heterozygosity in the population is in turn directly impacted by the mutation rate, and this, under most selective regimes.

In the limit of very low mutation rates, gene conversion just does not have much on which to act across the genome, and as a consequence, is not experiencing any meaningful induced stabilizing selection.

The analytical approximation (plain lines in Figure 3) are globally in good agreement with the simulation results (filled circles), except for large $\bar{s}$ or large u, where the analytical prediction appears to be an underestimate. These corresponds to regimes where the diffusive approximation used for deriving the analytical predictions is not valid, owing to a large variance in genome-wide log-fitness between individuals. In practice, these regimes are far from empirical reasonable conditions. As for the analytical variance estimate (shared areas in Figure 3), they are correct for low u or low L but appear to be underestimates in all other cases, when compared to simulation-based variance estimates (bar plots). This could be a consequence of the linear approximation implicit in estimating the variance based on the slope of the G function at the equilibrium set point (see above). Alternatively, this could be due to the fact that the analytical approximation assumes an instantaneous equilibration of the genome to the current population mean value of b. In practice, the selected loci show some inertia in their response to the variation in b at the level of the population. This in turn is expected to result in a delay in how the strength of the balancing selection acting on b responds to the fluctuations of this variable, ultimately resulting in an increased variance.

Finally, across all scaling experiments shown in Figure 3, the stabilizing selection induced on b appears to be globally tighter in the partially recessive case, for which both the response of the equilibrium value of b to changes in parameter values and the equilibrium variance are less pronounced than in the co-dominant case.

Which class of mutations contribute to stabilizing selection on gBGC ?

As the mean fitness effect of deleterious mutations is a key parameter for the evolution of intermediate levels of gBGC, it appears probable that under a DFE, not all mutations contribute equally to it. To further investigate this point, the analytical approximation was recruited to examine how the mean frequency at which W-deleterious alleles (Figure 4A&B) and S-deleterious alleles (Figure 4C&D) segregate in a population at equilibrium as a function of their selection coefficient, and how this segregation is modulated by slight variations in b (the dotted, plain, and dashed lines correspond to increasingly larger values of b). The bottom panels show the corresponding expected fitness gain $H_W(s, b)$ incurred by converting W-deleterious alleles (blue curves, above 0), and the expected fitness cost $H_S(s, b)$ incurred by converting S-deleterious alleles (red curves, below 0), both weighted by the distribution of selective effects (DFE). These are plotted as functions of s, for 3 different values of b. Weighting H by the DFE gives a better sense of the relative contributions of mutations with different selection coefficients to the total cost and gain. Also, with this weighting, averaging H_W and H_S over the DFE simply amounts to computing the area under the two curves, which thus directly correspond to $G_W(b)$ and $G_S(b)$, respectively. The parameter values used for Figure 4 correspond to the simulation trajectory displayed in Figure 1, for the co-dominant and recessive cases.

As b increases, W alleles segregate at a lower frequency (Figure 4A&B) and S alleles at a higher frequency (Figure 4C&D). Correlatively, the expected gain contributed by converting W alleles (Figure 4E&F, blue curves) decreases, and the cost contributed by converting S alleles (red curves) increases with population-

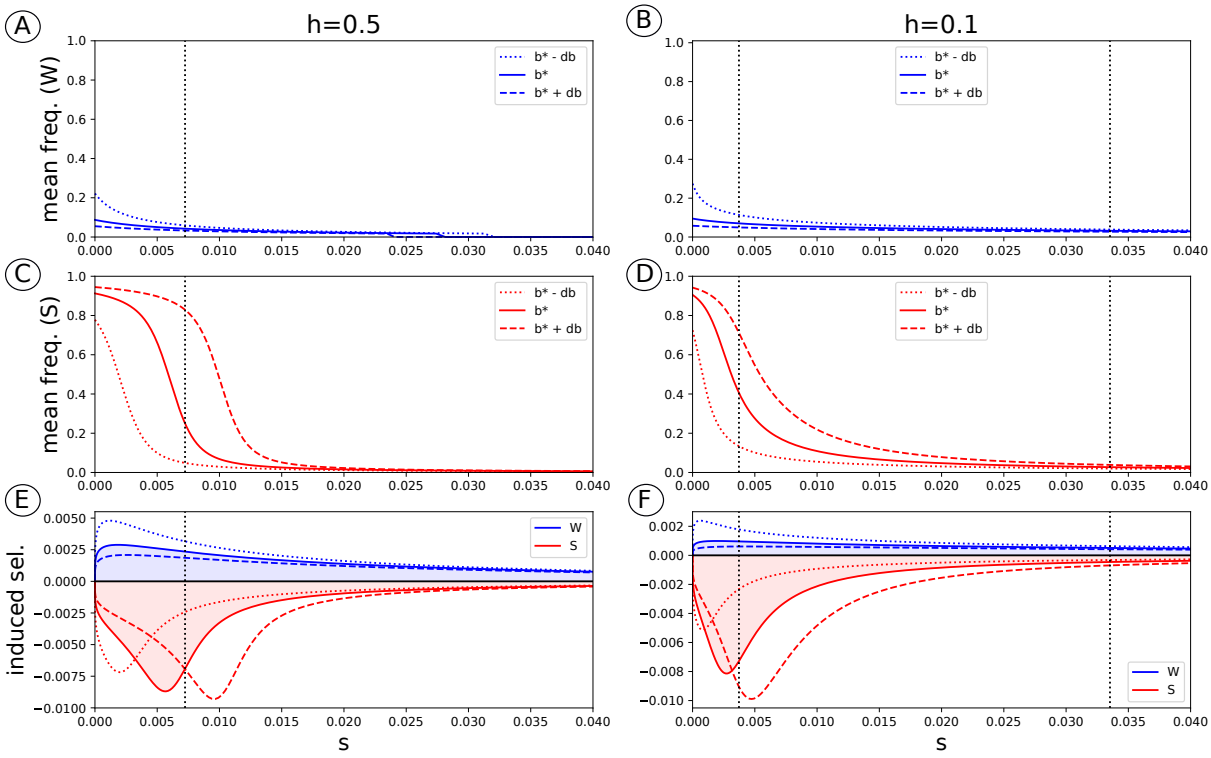


Figure 4: Mean segregation frequency of W alleles (A&B), S alleles (C&D), and induced selection (E&F), as a function of s, under the co-dominant (A,C&E) and recessive (B,D&F) settings, for different values of b: plain lines correspond to the equilibrium value of b, long stretches correspond to a slightly increased b, and dots correspond to a slightly decreased b. Blue lines correspond to W alleles, while red line correspond to S alleles.

level b. The intermediate value of b used in Figure 4 is precisely the one for which the areas under the two curves in panel C are equal (shaded areas in blue and red) – it is thus the predicted evolutionary optimum (of note, the areas under the two curves may not look equal to each other on the figure, in particular in the recessive case, but this is only because the two curves extend much further to the right than is shown).

Importantly, the way the two compartments, Weak and Strong, react to changes in population-level b is very different. On one side, W-deleterious polymorphisms are only moderately affected, and this, mostly in the range of small selective effects. In contrast S-deleterious polymorphisms are strongly affected. More specifically, increasing b leads to a surge in the segregation of S-deleterious mutations of intermediate strength, for which gBGC and selection are of the same order of magnitude. This surge translates into a peak in the expected cost (red curve, bottom panel), whose area increases with b. Translating these observations in terms of the net selection acting on gBGC, the fitness advantage is mostly contributed by converting W strongly-deleterious mutations, and is essentially a constant. The fitness cost, on the other hand, is mostly contributed by S mutations of selective effects of the order of b. This fitness cost varies strongly with b and is the main factor responsible for modulating the selection induced on gBGC modifiers, as a function of the population b.

Of note, the exact patterns differ between the co-dominant and the recessive cases. In the co-dominant case, the peak in the conversion cost is around $s \simeq b/h$, the value for which gBGC and selection exactly compensate each other. In the recessive case, the peak is broader and covers most of the range $b/(1-h) \leq s \leq b/h$, or equivalently $hs \leq b \leq (1-h)s$. This is the range for which the strength of gBGC is stronger than selection against the heterozygote but weaker than selection against the homozygote for the deleterious mutant. As a result, these GC-deleterious polymorphisms tend to segregate at intermediate frequencies, as if they were over-dominant (i.e. advantageous when in one copy, deleterious when in two copies), resulting in a higher fraction of heterozygotes in the population, and thus a substantial cost against gBGC (Glémin, 2010). This can explain why the strength of selection around the equilibrium value of b is higher in the partly recessive case.

gBGC is partially buffered against changes in population size

A somewhat paradoxical consequence of gBGC is the extreme sensitivity of equilibrium base composition to even mild variation in its population-scaled intensity $B = 4Nb$ (Eyre-Walker, 1999). Quantitatively, the neutral equilibrium GC/AT composition ratio scaling exponentially with B, which can quickly lead to very large GC content even for moderate increase in N. For instance, based on the current estimate of b in humans, increasing effective population size by a factor 10 would imply a long-term neutral equilibrium GC content greater than 99% in the 10% most highly recombining fraction of the genome. How to explain, then, that gBGC does not more often lead to diverging base composition across species?

Implicit in the argument just exposed is that the strength of gBGC is fixed, while population size varies, or at least, that there is no internal mechanism for tuning the raw intensity of gBGC (b) depending on effective population size (N), so as to somehow guarantee that $B = 4Nb$ never becomes too large. Yet, if gBGC is under stabilizing selection, this raises the possibility for such an internal mechanism to spontaneously emerge. This fundamentally depends on how the evolutionary optimum b^* scales with population size.

To examine this point, the optimal value b^* predicted by the model was computed (using the semi-analytical approximation) over a broad range of values of N between 10^2 and 10^6. For this experiment, a mutation rate of $u = 10^{-8}$ was assumed (for $S \to W$ mutations), and a bias of $\lambda = 2$. Both the co-dominant case ($h = 0.5$) and the partially recessive case ($h = 0.1$ and $h = 0.01$) were considered.

In all cases (Figure 5), whether co-dominant or recessive, b^* decreases with N. The trend is moderate in the co-dominant case but more pronounced in the recessive case. In both cases, the decrease is less than linear, such that $B = 4Nb$ still increases as a function of N. This increase is quite substantial in the co-dominant case, with B reaching values above 10 for population sizes of $N = 10^4$ and above 100 for $N > 3.10^5$. In the recessive case, on the other hand, B is much less responsive to changes in population size, ranging from $B \simeq 3$ for $N = 10^2$ up to $B \simeq 15$ for $N = 10^6$ – barely a 5-fold increase over 4 orders of magnitude for N.

Interestingly, a mixture of 50% co-dominant and 50% partially recessive ($h = 0.1$) essentially behaves like the pure partially recessive case (all mutations with $h = 0.1$). Even a small proportion of 10% of partially recessive positions, mixed with 90% of co-dominant positions, shows substantially more stable levels of gBGC

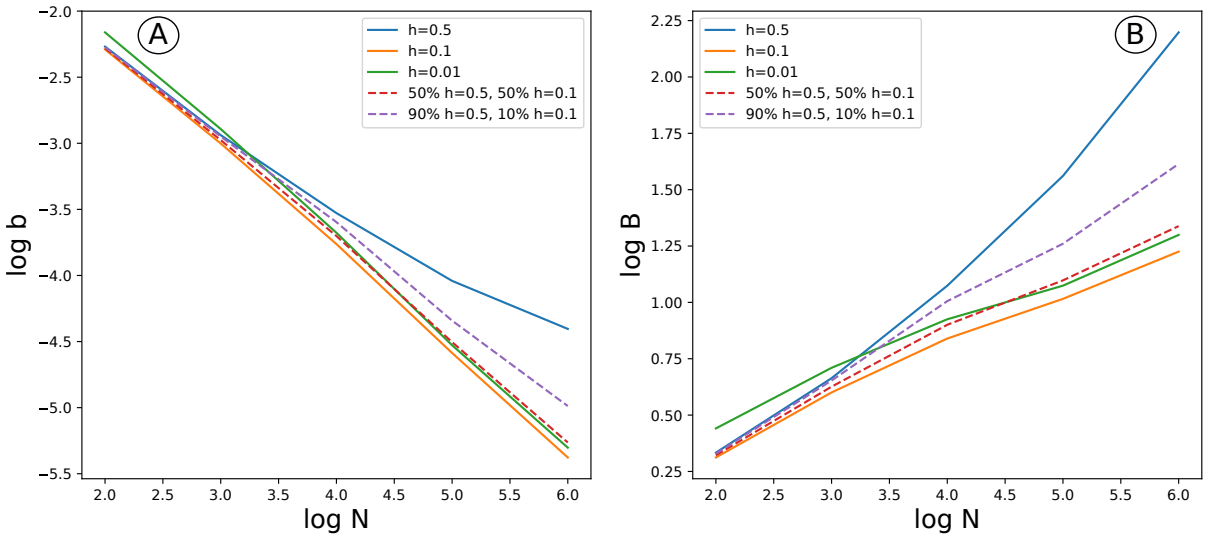


Figure 5: Scaling of b^* (A) and $B^* = 4Nb^*$ (B) as a function of N, under the co-dominant ($h = 0.5$) and recessive ($h = 0.1$ and $h = 0.01$) settings (plain curves), or assuming a mixture of co-dominant and recessive mutations (dashed curves).

as a function of N (Figure 5, dashed lines). Recessive mutations thus appear to represent an efficient buffer against changes in population-scaled gBGC induced by changes in population size.

The fundamental reason why b^* decreases with N can be understood by examining the structure of the induced selective response (Figure 6). As mentioned above, the mutation-segregation balance essentially takes the form of a tradeoff between, on one side, a net error-correcting effect on strongly deleterious mutations (more often deleterious towards W than towards S) and, on the other side, a conversion load mostly contributed by S-deleterious mutations with selection coefficients of the order of b. The first component, being in the strong selection regime, is essentially insensitive to N (Figure 6E&F, blue curves). The second component, on the other hand, precisely because of the compensation between gBGC and selection, is effectively in a regime dominated by drift, and thus, in many respects, has an evolutionary dynamics resembling nearly-neutral evolution. As such, its mean heterozygosity is strongly influenced by changes in effective population size, and more precisely, will tend to increase with N. Since biased gene conversion is in direct proportion to the amount of heterozygosity, the conversion cost itself will also increase with N (Figure 6E&F, red curves). Altogether, GC-deleterious mutations with selective effects of the order of b are efficiently mobilized (i.e. contribute more to standing variation) upon an increase in N and thus represent a key force buffering B^* against changes in population size.

Of note, and as already explored above (Figure 4), in the co-dominant case (Figure 6, A,C&E), the range of GC-deleterious mutations that are mobilized consists of a relatively narrow peak around b/h. In contrast, in the recessive case, a good fraction of the range comprised between $b/(1-h)$ and b/h (the two dotted vertical lines on Figure 6B,D&F), corresponding to the co-dominant regime, is mobilized, thus contributing a much more responsive buffer against changes in N – which can easily dominate the overall response even if

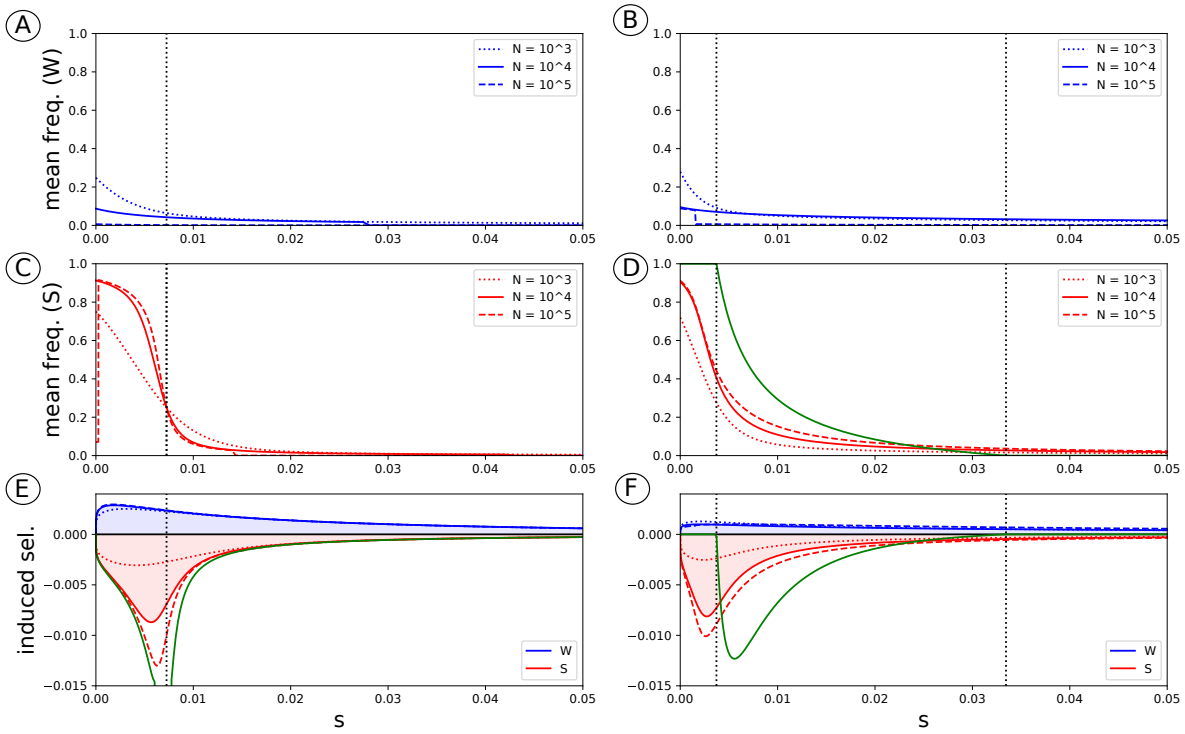


Figure 6: Mean segregation frequency of W alleles (A&B), S alleles (C&D), and induced selection (E&F), as a function of s, under the co-dominant (A,C&E) and recessive (B,D&F) settings, for different values of N: plain lines correspond to the equilibrium value of b, long stretches correspond to a slightly increased N, and dots correspond to a slightly decreased N. Blue lines correspond to W alleles, while red line correspond to S alleles. Green lines correspond to the deterministic approximation (Ns very large).

recessive mutations represent a minority of the total standing variation, as observed above (Figure 6, dashed lines).

gBGC and the genetic load

gBGC is often depicted as a force that interferes with selection, and that causes a significant deleterious burden (Galtier and Duret, 2007; Berglund et al., 2009; Galtier et al., 2009; Necşulea et al., 2011). On the other hand, our results reflect those of previous studies showing that BGC confers a significant fitness advantage by correcting the most common class of mutations (here $S \mapsto W$) (Bengtsson and Uyenoyama, 1990). But as pointed out by Glémin (2010), the levels of gBGC that evolve naturally are not necessarily the ones that minimize the average genetic load of a population. The average genetic load of a population can be decomposed into the load of W deleterious alleles:

$$L_W = \langle \mathbb{E}_{\phi^W_{s,b}} \left[2hsx_W(1-x_W) + sx_W^2 \right] \rangle_\gamma \tag{8}$$

Where x is a random variable drawn from the equilibrium frequency distribution of W alleles $\phi^W_{s,b}(x)$, and that of S deleterious alleles:

$$L_S = \langle \mathbb{E}_{\phi^S_{s,b}} \left[2hsx(1-x) + sx^2 \right] \rangle_\gamma \tag{9}$$

Where x is a random variable drawn from the equilibrium frequency distribution of S alleles $\phi^S_{s,b}(x)$.

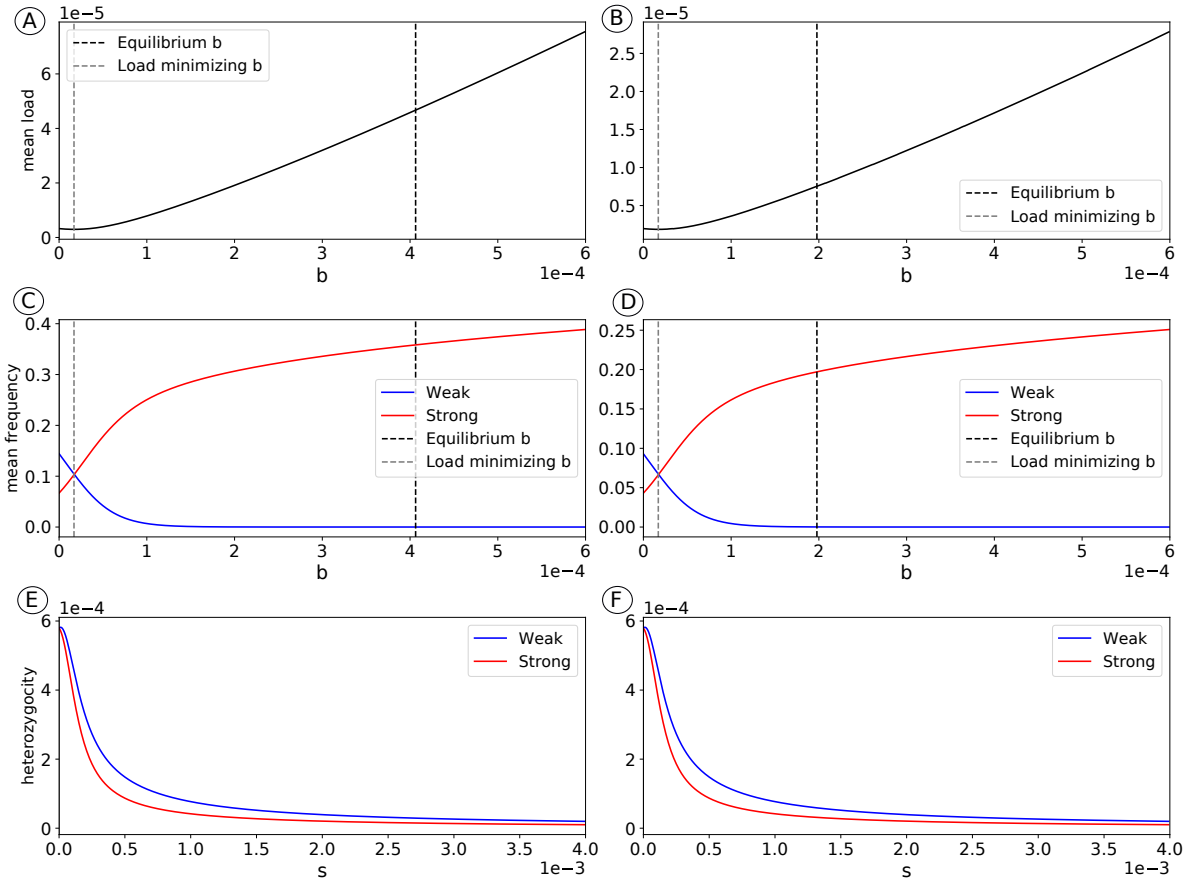


Figure 7: A&B: average deleterious load in a population as a function of b. The black lines shows the value equilibrium value of b and the grey line shows the value of b that minimizes the average load. C&D: frequency of W and S alleles as a function of b. E&F: heterozygosity for W ans S alleles as a function of their deleterious effect s under the value of b that minimizes the load. A,C&E: $h = 0.5$. B,D&F: $h = 0.1$

Using a population size of $N = 10,000$, a mutation rate of $u = 10^{-8}$ and a mean selection coefficient $hs = 0.01$, we computed the average genetic load as a function of b for $h = 0.5$ and $h = 0.1$ (Figure 7A&B). The load is minimized for a very small value of b compared to the one that naturally evolves. The level of gBGC that minimizes the average genetic load corresponds to the level that equalizes the frequencies of W and S alleles, leading to a average GC content of 0.5 (Figure 7C&D). It is worth noting that when the average frequencies of W and S alleles are equal, it does not mean that they are distributed evenly in heterozygotes. In fact, W deleterious alleles are more often heterozygous (Figure 7 E&F), because they are numerous due to a high $S \mapsto W$ mutation rate, but at low frequency because of gBGC. Conversely, S deleterious alleles are less numerous due to a high $S \mapsto W$ mutation rate, but more often at high frequency because of gBGC and thus more often homozygous. Therefore, when the mean load in the population is minimal, there still is an individual advantage to convert W deleterious alleles more often for heterozygotes.

Empirical calibration

Overall, the modeling work presented thus far suggests that biased gene conversion in favour of GC can in principle evolve as a consequence of the mutation bias towards AT and that its intensity can also be modulated in an adaptive manner as a function of key parameters, in particular effective population size. An important question that remains is whether it provides quantitatively reasonable predictions when confronted to current empirical knowledge about the strength of gBGC in various species.

Humans and the mouse represent good cases to consider. One important problem to address, however, is that the model considered thus far assumes a uniform recombination landscape. Yet most mammals, and many other eukaryotes, have recombination hotspots, such that 90% of the recombination is typically concentrated in about 10% of the genome (Smagulova et al., 2011; Pratto et al., 2014). In the presence of hot spots, most of the selection induced on gBGC will be contributed by the highly recombining regions. This can be reasonably well approximated by considering that only the highly recombining regions are being explicitly modeled.

Assuming that ∼10% of the genome is under selection (Rands et al., 2014), and 10% of those selected loci are in highly recombining regions, for a genome of total size 3 Gb, this gives $L = 3.10^6$. In humans, the mutation rate is $u = 3.10^{-8}$ per base pair and per generation. In the mouse, the mutation rate is a bit lower $u = 10^{-8}$. Here, only $u = 3.10^{-8}$ is considered. The mutation bias is in both cases of the order of $\lambda = 2$, the value used here. Current estimates of the DFE suggest a shape parameter a between 0.2 and 0.3. Here, $a = 0.2$ is used. The mean selection coefficient under this DFE is difficult to estimate. In humans, recent estimates are of the order of $h\bar{s} = 0.01$ to 0.05, both of which were tried in what follows. Finally, the co-dominant and partially recessive cases are considered, as well as the 50:50 and 90:10 mixtures of these two dominance regimes, with population sizes varying from $N = 10^4$ to $N = 10^6$, so as to cover most of the range of what can be expected more generally in mammals.

The estimates of $B^* = 4Nb^*$ returned by the numerical approach under these parameter values are reported in Table 1. Of note, since only highly recombining regions are modeled, this should be interpreted as the value of B prevailing in those regions, which is thus about 100 times higher than the strength of gBGC in the remaining 90% of the genome. Under co-dominant selection, the predicted values for B^* range from 11 to 300, showing quite some sensitivity to both effective population size and mean selection strength across the genome. In contrast, assuming partially recessive mutations return a much narrower range of estimates, from 7 to 30. Fitting the model assuming a mix of co-dominant and recessive mutations (last rows of Table 1) suggests that a moderate fraction of recessive mutations is sufficient to make B less responsive to changes in N.

Empirical estimates of gBGC in Humans are of the order of $B = 0.3$ for the genome-wide mean, and around $B = 5.2$ to 6.5 in the top 20% regions of high recombination (Duret and Arndt, 2008). The theoretical predictions (Table 1) are globally higher than these empirical estimates, although they get reasonably close to them (predicted $B < 10$) for the lower values of $\bar{s}$ or assuming the presence of recessive mutations. This, together with the rather extreme results obtained for the largest population size under the co-dominant case, suggest that recessive mutations may play a role in buffering gBGC.

		$h\bar{s} = 0.01$			$h\bar{s} = 0.05$		
h	N	B_h^*	stdev	$p(B < 0)$	B_h^*	stdev	$p(B < 0)$
0.5	10^4	11	8	0.07	25	12	0.02
	10^5	36	15	0.009	89	14	0.01
	10^6	157	11	< 0.001	297	6	< 0.001
0.1	10^4	7	4	0.04	9	4	0.01
	10^5	11	4	0.005	13	4	< 0.001
	10^6	24	7	0.002	31	8	< 0.001
50% 0.1 : 50% 0.5	10^4	8	5	0.05	11	5	0.009
	10^5	13	5	0.004	16	5	< 0.001
	10^6	32	9	< 0.001	45	10	< 0.001
10% 0.1 : 90% 0.5	10^4	10	6	0.05	16	6	0.005
	10^5	20	7	0.001	26	7	< 0.001
	10^6	65	13	< 0.001	104	12	< 0.001

Table 1: Numerical estimates of B_h^* (scaled conversion strength in hot spots), equilibrium standard deviation and probability of a negative gBGC, for different parameter values for N, h, $h\bar{s}$.

Finally, the predicted evolutionary variance is substantial for low population size, such that b^* can take on values very close to 0 or even negative, around 1 to 5% of the time. This suggests that induced selection on gBGC may not be sufficiently powerful to guarantee a bias towards Strong over the whole range of molecular evolutionary regimes susceptible to be observed across mammals – although, even then, it may still represent a sufficiently strong selective force preventing gBGC to become unreasonably large.

3 Discussion

In this study, we developed a model to characterize the evolution of gBGC by means of natural selection. We first showed that in the presence of a mutation bias towards AT, gBGC was under relatively weak stabilizing selection around a positive value of the transmission bias, in agreement with a previous study (Bengtsson and Uyenoyama, 1990). The equilibrium value of the transmission bias (b^*) corresponds to the one that equalizes the fitness gain of converting strongly deleterious AT mutations in heterozygotes and the fitness cost of transmitting slightly deleterious GC mutations to offsprings. This balance depends both of the strength of the mutation bias towards AT, but also on the mean fitness and dominance effects of deleterious mutations. When even few deleterious mutations are recessive, the cost of transmitting slightly deleterious GC alleles becomes quickly higher, and b^* decreases. Importantly, b^* is negatively correlated to effective population size. In fact, the fitness gain of correcting strongly deleterious AT mutations is essentially independent of effective population size, while the cost of transmitting slightly deleterious GC alleles increases quickly with it. This could contribute to the absence, or the weak positive correlation between the population-scaled gBGC coefficient ($B = 4N_e b$) and effective population size (N_e) reported in several clades of eukaryotes (Lartillot, 2013; Clément et al., 2017; Galtier et al., 2018; Galtier, 2021; Boman et al., 2021).

A tragedy of the commons

gBGC is often described as an evolutionary force that antagonizes natural selection. It has even earned the nickname of "Achilles' heel of genomes" (Galtier and Duret, 2007). When the impact of an intrinsically deleterious biological process still can be limited to a minimum by natural selection, it leads to a drift-barrier behaviour, where the effect of this biological process is limited by increasing effective population size. This is thought to be the case for the evolution of mutation rates (Sung et al., 2012), or for splicing errors (Bénitière et al., 2023). Galtier et al. (2018) proposed that the negative relationship between gBGC and effective population size observed in angiosperms and animals could also arise from a drift-barrier hypothesis. Here we show that despite a deleterious effect at the population level, gBGC is still (weakly) positively selected. Therefore, the levels of gBGC observed in animals may not be counter-selected at all. In this view, the pervasive existence of gBGC in eukaryotes is not explained by a limited efficiency of negative selection due to drift, but by a tragedy of the commons where the short-term advantage of biasing gene conversion towards GC limits the long-term reproductive capacity of the population as a whole.

The strength of selection acting on gBGC

Of note, the strength of stabilizing selection on gBGC according to our model is weak and the expected variance of b around its equilibrium value depends on several assumptions. First, we only considered the direct conversion gain/cost in fitness in one generation, while in fact, a gBGC modifier will be statistically linked to half of the deleterious AT alleles it corrects in the next generation, and again half of it in the next etc... By considering only the direct gain/cost at one generation, we might be underestimating the strength of selection acting on a modifier, and thus overestimating the evolutionary variance. However, when linkage is taken into account (in the simulations), the evolutionary variance is higher than under the semi-analytical approximations. This can be explained by a strong assumption of the semi-analytical model: the population has time to reach mutation-selection-drift-gBGC equilibrium between each modification of b (low mutation rate at modifier loci). When the number or the effects of loci that can influence the strength of gBGC is high enough such that the population does not have time to reach mutation-selection-drift-gBGC equilibrium between two consecutive modifications of b, the short term benefit/cost of converting with a bias b is not coupled to its long term benefit/cost. In this case, we might observe increased oscillations around $b*$, and thus increased evolutionary variance.

Essentially, the results obtained here give the best-case scenario among all possible genetic architectures for gBGC, i.e. the scenario for which stabilizing selection on b is tightest.

The penetrance of a somatic repair bias in meiosis

It is very likely that the mechanisms that bias DNA repair towards GC in meiosis are the same as those that operate in somatic cells. Most single nucleotide DNA damages involve wrongly incorporated As, Ts, or even Us. Repair enzymes that minimize the somatic mutation rate should therefore be GC-biased. In this sense, in mammals, the base excision repair pathway has DNA glycosilases for excising adenines and thymines, but none for guanine or cytosine (Krokan and Bjørås, 2013). This leads to a re-interpretation of the results

presented here: fundamentally suggesting that there is enough selection for limiting the penetrance of the somatic repair bias in meiosis, if this ever leads to overly strong gBGC (deleterious at the individual level). In this view, the expected strength of gBGC should lie between the somatic repair bias (strongly GC biased) and b^*. This could explain why meiotic gene conversion appears to be universally GC-biased, despite the weak selection preventing it from being AT-biased (Table 1). Of note, even if selection to limit the penetrance of the somatic repair bias is maximally effective, the expected value of b still induces a substantial load at the population level.

gBGC and effective population size

In mammals, there is a (weak) correlation between effective population size N_e and the population-scaled gBGC coefficient ($B = 4N_e b$) (Lartillot, 2013; Galtier, 2021). This correlation is also observed among human populations (Glémin et al., 2015; Subramanian, 2019), and effective population size seems to explain the difference in B between two passerine species (Barton and Zeng, 2021). However, in *Leptidea* butterflies there is no relationship between B and genetic diversity, suggesting that the transmission bias b is lower in species/populations of higher effective population size (Boman et al., 2021). Finally, no correlation has been observed between B and N_e in 29 species (Galtier et al., 2018), or in 11 species of angiosperms (Clément et al., 2017).

The most probable hypothesis so far is that in animals and plants, there is a negative correlation between the repair bias b_0 and effective population size (Galtier et al., 2018). Several arguments have been proposed to explain this negative relationship. As previously said, Galtier et al. (2018) proposed a drift-barrier hypothesis: assuming that gBGC is deleterious, it can be more efficiently counter-selected in species with higher N_e, and thus b_0 should be negatively correlated with N_e.

On the other hand, it has been shown that depending on the repair mechanisms, the intensity of the bias could be negatively correlated with heterozygosity (Lesecque et al., 2013; Li et al., 2019). As heterozygosity is supposed to be proportional to N_e, this can also explain why we observe no correlation between B and N_e (Clément et al., 2017; Galtier et al., 2018; Boman et al., 2021), while we still expect one under selection only. However, the switch to such heterozygosity-dependent mechanisms could also be an adaptive response to the increasing cost of gBGC, and the two hypotheses are not mutually exclusive. Nevertheless, these hypothesis remain verbal, and a proper modelling of the molecular mechanisms of gBGC and their selective advantage is needed to put them to the test.

Empirical relevance

We highlighted that gBGC being deleterious at the population level is not an indicator that it is negatively selected. It is therefore unclear whether the levels of gBGC currently found in eukaryotes are actually negatively selected. Here, we computed the expected b^* under empirically realistic parameters, and recover a rather high b^*. It is important to note that this estimation is sensitive to parameters that are very difficult to estimate reliably. Notably, the size of the genome that is under selection(Rands et al., 2014), the DFE and more specifically the size of the compartment of strongly deleterious mutations. Moreover, b^* is strongly

sensitive to the distribution of dominance effects, about which little is known (Billiard et al., 2021). Finally, it relies on the assumption that half the genome under selection has a GC allele as optimal and the other half an AT allele. This assumption is intuitive and is made in almost all models of gBGC (Bengtsson, 1990; Glémin, 2010; Bolívar et al., 2016; Corcoran et al., 2017) However, when using empirical fitness landscapes in protein coding genes instead of an arbitrary distribution of selective effects, it appears that AT encoded amino-acids are more often optimal than GC-ones, which is due to the structure of the genetic code (chapter 9). Under this scenario, a slight mutational bias is actually beneficial, and thus b^* should be lower.

Overall, while the present model significantly improves our understanding of the selective pressures exerted on gBGC, it is by no mean an attempt to accurately predict the strength of gBGC *in vivo*.

4 Methods

4.1 Model

The model assumes a population of fixed size N diploid individuals, randomly mating and with non-overlapping generations. The genome is composed of a single chromosome. Since neutral loci don't have any impact on the evolution of gBGC, they are not explicitly modeled. As a result, the chromosome is assumed to consist of L bi-allelic positions, with two alternative alleles, W (weak) or S (strong), that are all under selection with locus-specific selective strengths. The model also invokes a modifier locus placed somewhere along the chromosome (in the experiments conducted here, at one third of the total length of the chromosome).

For a given selected position i, $1 \leq i \leq L$, either the W allele or the S allele is considered deleterious with probability $1/2$, in which case the selection strength s_i acting on the deleterious allele is randomly drawn from a gamma distribution of mean $\bar{s}$ and shape parameter a. All selected loci share the same dominance coefficient h. In the following, the co-dominant case $h = 1/2$ and the recessive case $0 < h < 1/2$ are both considered. The selective effects are assumed additive over loci. Thus, assuming locus i is such that W is the deleterious allele, then the log-fitness contribution is 0 for genotype SS, hs_i for genotype SW and s_i for genotype WW (and conversely for loci for which S is the deleterious allele). Letting $Q^1_{ij}, Q^2_{ij} \in \{0,1\}^2$ stand for the genotype of diploid individual j at position i, with the convention that 1 stands for the deleterious allele (which can be either S or W depending on the locus), the total Malthusian (log) fitness of individual j is then given by:

$$\ln W_j \quad = \quad -\sum_{i=1}^{L} \left(Q^1_{ij}(1-Q^2_{ij}) + Q^2_{ij}(1-Q^1_{ij})\right) hs_i \; + \; Q^1_{ij}Q^2_{ij}s_i.$$

The selected positions undergo recurrent mutations between W and S. Allele W mutates towards S at rate u, and allele S mutates towards W at rate λu per generation.

The modifier locus encodes an additive quantitative trait controlling the bias of gene conversion. For individual j, with genotype $(z^1_j, z^2_j) \in R^2$ at the modifier locus, the bias is then equal to:

$$\beta_j = \frac{1}{2}\left(z^1_j + z^2_j\right).$$

How this bias exactly impacts gene conversion during meiosis is described below. The modifier locus mutates are rate w, in which case the quantitative contribution of the mutant allele is equal to that of its parent, plus a normally distributed increment, of mean 0 and standard deviation Δz:

$$z' \quad \sim \quad N(z, \Delta z^2).$$

A simplified version of meiosis is implemented as follows. Consider individual j. First, each selected position that happens to be heterozygous in this individual undergoes gene conversion with probability α, in which case conversion is towards the S allele with probability $(1+\beta_j)/2$ and towards the W allele with probability $(1-\beta_j)/2$, with β_j such as defined above (equation..). Second, a cross-over point is chosen uniformly at random over the chromosome, and two recombinant chromosomes are produced by swapping the segments on both sides of the cross-over point. Thus, both the rate of cross-over and the rate of gene conversion are considered fixed and invariant across individuals, while the bias of the gene conversion events is allowed to vary between individuals, based on the genotype at the modifying locus. Of note, a positive (resp. negative) value for β_j results in biased gene conversion towards S (resp. towards W). Quantitatively, at a given selected position at which individual j is heterozygous, the net proportions of gametes produced by this individual bearing the S allele is:

$$q_S \quad = \quad (1-\alpha)\frac{1}{2} + \alpha\frac{1+\beta}{2} \; = \; \frac{1+\alpha\beta}{2} \; = \; \frac{1+b}{2},$$

with $b = \alpha\beta$. Similarly, the proportion of gametes with the W allele is $q_W = \frac{1-b}{2}$.

The overall life cycle runs as follows. First, all individuals of the current generation undergo mutations both at the modifying and at the selected loci, with mutation rates such as given above. Next, each individual of the next generation is produced by first randomly choosing two parents in the current generation, each with a probability proportional to is fitness W (such as given by eq. above). Each of the two chosen individuals then undergoes a meiosis, producing a pair of gametes, one of which is randomly picked out and paired with the gamete produced by applying the same procedure to the other individual. Of note, only one gamete per meiosis is kept for the next generation, the other one being discarded.

Altogether, the parameters of the model are:

N**:** population size

L**:** number of loci under selection

$\bar{s}$**:** mean selection strength at the selected loci

a**:** shape of the distribution of selection strengths across loci

h**:** dominance coefficient

u**:** basal mutation rate at the selected loci

λ**:** mutation bias ($S \to W$ relative to $W \to S$)

w**:** mutation rate at the modifier locus

Δz**:** mean effect size of the mutations at the modifier locus

α**:** gene conversion rate (per generation and per selected locus)

4.2 Theory / Analytical approximation

Here, a semi-analytical approximation is derived for determining the equilibrium value of the net strength of biased gene conversion b as well as its evolutionary variance. This derivation assumes a low mutation rate at the modifier locus (low w), such that the population is, at any time, approximately monomorphic for the strength of gBGC, and all selected loci are at mutation-selection-drift-conversion equilibrium under this value of gBGC. The derivation also assumes that linkage both among selected loci and between the modifier and the selected loci is negligible. The first condition implies that background selection is weak, and that the mutation-selection-drift-conversion equilibrium can be determined independently at each locus. The second is motivated by the fact that, in practice, most selected loci are sufficiently far from the modifier, such that most of the induced selection is contributed by loci that not tightly linked with the modifier.

Consider a population monomorphic at the modifier locus for an allele of strength β, at equilibrium under a gBGC of strength $b = \alpha\beta$. In this population, a mutant at the modifier locus, of size $2\delta\beta$ appears in an individual. This individual thus has a gBGC strength of $b' = b + \delta b$ in its germline, with $\delta b = \alpha\,\delta\beta$. We want to determine the net selective advantage or disadvantage incurred by this individual, owing to its departure from the population-level gBGC. This selection will be indirectly contributed by the effect of biased gene conversion on the selected loci across the genome. Therefore, in the following, this will be called the selection *induced* on the gBGC modifier, or induced selection for short.

Under efficient linkage dissipation, induced selection is the sum of the contributions of all selected loci. Consider in a first step a single locus at which W is the deleterious allele, with selection s, dominance h and segregating in the population at frequency x. Given x, the probability for the individual to be heterozygous at this position is:

$$P(x) \quad = \quad 2x(1-x),$$

in which case the S and W allele are transmitted in the gametes with probability $\frac{1+b'}{2}$ and $\frac{1-b'}{2}$, respectively. In a random mating population, this will result in an average fitness gain in the offspring of:

$$\begin{aligned}
\ln f_W(x,s,h) &= \frac{1+b'}{2}((1-x)\times 0 + x\times(-hs)) + \frac{1-b'}{2}((1-x)\times(-hs) + x\times(-s)) \\
&= b'\,\frac{s}{2}\,[h + x(1-2h)] \\
&= b'\,C(s,h,x),
\end{aligned}$$

where:

$$C(s,h,x) \quad = \quad \frac{s}{2}\,[h + x(1-2h)].$$

Of note, if $b' > 0$, this is indeed a gain, since on average, S alleles, which have a higher fitness at that position, are over-transmitted. Next, to assess the fate of the gBGC mutant, one should discount the equivalent gain, but under a gBGC equal to b in the population, such that the average selective advantage contributed by the selected position under consideration to the individual bearing the mutant allele for the modifier (now accounting for the probability for this individual to be a heterozygote at the selected locus) is:

$$\delta \ln f_W(x,s,h) \quad = \quad P(x)\,C(s,h,x)\,\delta b.$$

Equation 10 gives the cost conditional on the frequency x of the W allele at the focal selected position and conditional on the selection coefficient s. This needs to be averaged over x at mutation-selection-conversion-drift equilibrium (here noted $\phi^W_{s,b}$) and then summed over the distribution of selective effects across the $L/2$ loci being deleterious towards the W allele:

$$\begin{aligned} G_W(b) &= \frac{\delta \ln f_W}{\delta b} \\ &= \frac{L}{2} \langle H_W(s,b) \rangle, \end{aligned}$$

where the angle brackets stand for the expectation over s under the DFE, and,

$$H_W(s,b) = E_{\phi^W_{s,b}} [P \times C]$$

is the expectation over x under $\phi^W_{b,s}$ of $P(x)\, C(s,h,x)$. In other words, it is the net gain induced by conversion events at loci that are W-deleterious, with selection coefficient s and dominance coefficient h. In turn, the distribution $\phi^W_{b,s}$ is given by (Wright, Glemin):

$$\phi^W_{b,s}(x) = \frac{1}{Z^W_{b,s}} x^{4Nv-1}(1-x)^{4Nu-1} e^{-4Nx(b+s(h(1-x)+x(1-h)))},$$

where $Z^W_{b,s}$ is the normalization constant:

$$Z^W_{b,s} = \int x^{4Nv-1}(1-x)^{4Nu-1} e^{-4Nx(b+s(h(1-x)+x(1-h)))} dx.$$

A similar derivation is done for a locus at which W is the deleterious allele, which, by symmetry, gives:

$$\begin{aligned} G_S(b) &= \frac{\delta \ln f_S}{\delta b} \\ &= -\frac{L}{2} \langle H_S(s,b) \rangle, \end{aligned}$$

where

$$H_S(s,b) = E_{\phi^S_{s,b}} [P \times C]$$

and

$$\phi^S_{b,s}(x) = \frac{1}{Z^S_{b,s}} x^{4Nu-1}(1-x)^{4Nv-1} e^{-4Nx(-b+s(h(1-x)+x(1-h)))},$$

with normalization constant:

$$Z^S_{b,s} = \int x^{4Nvu1}(1-x)^{4Nv-1} e^{-4Nx(-b+s(h(1-x)+x(1-h)))} dx.$$

Of note, $P(x)$ and $C(s,h,x)$ are positive for all x, and thus, increasing biased gene conversion towards the strong alleles results in a net gain over W-deleterious loci, but a net a loss over S-deleterious loci. Whether the mutant for gBGC is favoured by this induced selection will depend on the balance between these two components. In other words, the total selection induced on the modifier is:

$$\begin{aligned} \frac{\delta \ln f}{\delta b} &= G(b) \\ &= G_W(b) - G_s(b). \end{aligned}$$

4.3 Implementation

The model was implemented in C++, using openMP for parallelizing the computations. For all results presented here, it was run under population sizes of size $N = 1000$, with $L = 10000$ selected loci, for a total of 210 000 generations, discarding the first 10 000 generations (burn-in) and subsampling 1 every 100 generations, upon which averages and standard deviations for quantities of interest were computed on the remaining 2000 points.

Numerical integration and solving was done in Python, using the scipy library for numerical integration over the allele frequency distributions. For integrating over the gamma distribution of selective effects, the gamma distribution was discretized into $n = 300$ points, corresponding to the mid-points of the successive $1/n$ quantiles, and then the integral over the distribution was approximated as the equal-weighted average of the integrand over these n values for hs.

Acknowledgments

We gratefully acknowledge the help of Laurent Duret for his reviews on a first version of this manuscript. **Funding:** Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec. **Author contributions:** Original idea: N.L.; Model conception: A.C., J.J. and N.L.; Code: A.C., J.J. and N.L.; Data analyses: A.C., J.J. and N.L.; Interpretation: A.C., J.J. and N.L.; First draft: J.J. and N.L.; Editing and revisions: A.C., J.J. and N.L.; Funding: N.L.; **Competing interests:** The author declare no conflicts of interest. **Data and materials availability:** Analysis scripts and documentation will be available upon deposit of this manuscript on a preprint server.

References

Barton, H. J. and Zeng, K. (2021). The effective population size modulates the strength of GC biased gene conversion in two passerines. Pages: 2021.04.20.440602 Section: New Results.

Bengtsson, B. O. (1986). Biased conversion as the primary function of recombination. *Genetics Research*, 47(1):77–80. Publisher: Cambridge University Press.

Bengtsson, B. O. (1990). The effect of biased conversion on the mutation load. *Genetics Research*, 55(3):183–187. Publisher: Cambridge University Press.

Bengtsson, B. O. and Uyenoyama, M. K. (1990). Evolution of the segregation ratio: Modification of gene conversion and meiotic drive. *Theoretical Population Biology*, 38(2):192–218.

Berglund, J., Pollard, K. S., and Webster, M. T. (2009). Hotspots of Biased Nucleotide Substitutions in Human Genes. *PLOS Biology*, 7(1):e1000026. Publisher: Public Library of Science.

Billiard, S., Castric, V., and Llaurens, V. (2021). The integrative biology of genetic dominance. *Biological Reviews*, 96(6):2925–2942. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/brv.12786.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. (2019). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol*, 20(1):5.

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. (2016). Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an Avian System. *Mol Biol Evol*, 33(1):216–227.

Boman, J., Mugal, C. F., and Backström, N. (2021). The Effects of GC-Biased Gene Conversion on Patterns of Genetic Diversity among and across Butterfly Genomes. *Genome Biology and Evolution*, 13(5).

Brown, T. C. and Jiricny, J. (1987). A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, 50(6):945–950. Publisher: Elsevier.

Bénitière, F., Necsulea, A., and Duret, L. (2023). Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. Pages: 2022.12.09.519597 Section: New Results.

Clément, Y., Sarah, G., Holtz, Y., Homa, F., Pointet, S., Contreras, S., Nabholz, B., Sabot, F., Sauné, L., Ardisson, M., Bacilieri, R., Besnard, G., Berger, A., Cardi, C., Bellis, F. D., Fouet, O., Jourda, C., Khadari, B., Lanaud, C., Leroy, T., Pot, D., Sauvage, C., Scarcelli, N., Tregear, J., Vigouroux, Y., Yahiaoui, N., Ruiz, M., Santoni, S., Labouisse, J.-P., Pham, J.-L., David, J., and Glémin, S. (2017). Evolutionary forces affecting synonymous variations in plant genomes. *PLOS Genetics*, 13(5):e1006799. Publisher: Public Library of Science.

Corcoran, P., Gossmann, T. I., Barton, H. J., The Great Tit HapMap Consortium, Slate, J., and Zeng, K. (2017). Determinants of the Efficacy of Natural Selection on Coding and Noncoding Variability in Two Passerine Species. *Genome Biology and Evolution*, 9(11):2987–3007.

Duret, L. and Arndt, P. F. (2008). The Impact of Recombination on Nucleotide Substitutions in the Human Genome. *PLOS Genetics*, 4(5):e1000071. Publisher: Public Library of Science.

Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genom. Hum. Genet.*, 10(1):285–311.

Eyre-Walker, A. (1999). Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. *Genetics*, 152(2):675–683.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8(8):610–618. Number: 8 Publisher: Nature Publishing Group.

Galtier, N. (2021). Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer Community Journal*, 1.

Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6):273–277. Publisher: Elsevier.

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

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, 159(2):907–911.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5):1092–1103.

Glémin, S. (2010). Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics*, 185(3):939–959.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Res.*, 25(8):1215–1228.

Kostka, D., Hubisz, M. J., Siepel, A., and Pollard, K. S. (2012). The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Molecular Biology and Evolution*, 29(3):1047–1057.

Krokan, H. E. and Bjørås, M. (2013). Base Excision Repair. *Cold Spring Harb Perspect Biol*, 5(4):a012583. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Lachance, J. and Tishkoff, S. A. (2014). Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles. *The American Journal of Human Genetics*, 95(4):408–420. Publisher: Elsevier.

Lartillot, N. (2013). Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. *Molecular Biology and Evolution*, 30(3):489–502.

Lesecque, Y., Mouchiroud, D., and Duret, L. (2013). GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance. *Molecular Biology and Evolution*, 30(6):1409–1419.

Li, R., Bitoun, E., Altemose, N., Davies, R. W., Davies, B., and Myers, S. R. (2019). A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat Commun*, 10(1):3900. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: DNA mismatch repair;DNA recombination;Genetic hybridization;Genetic variation;Homologous recombination Subject_term_id: dna-mismatch-repair;dna-recombination;genetic-hybridization;genetic-variation;homologous-recombination.

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W. R., Muscarella, M. E., Luo, H., Lennon, J. T., Brun, Y. V., and Lynch, M. (2018). Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol*, 2(2):237–240. Number: 2 Publisher: Nature Publishing Group.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485. Number: 7203 Publisher: Nature Publishing Group.

Meunier, J. and Duret, L. (2004). Recombination Drives the Evolution of GC-Content in the Human Genome. *Molecular Biology and Evolution*, 21(6):984–990.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(20):6278–6281. Publisher: Proceedings of the National Academy of Sciences.

Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., and Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mutation*, 32(2):198–206. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.21407.

Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, 23:263–286. Publisher: Annual Reviews.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome biology and evolution*, 4(7):675–682. Publisher: Oxford University Press.

Pouyet, F., Aeschbacher, S., Thiéry, A., and Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7:e36317.

Pouyet, F., Mouchiroud, D., Duret, L., and Sémon, M. (2017). Recombination, meiotic expression and human codon usage. *eLife*, 6:e27344.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., and Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science*, 346(6211):1256442–1256442.

Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLOS Genetics*, 10(7):e1004525. Publisher: Public Library of Science.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552):2571–2580. Publisher: Royal Society.

Roman, H. (1985). Gene conversion and crossing-over. *Environmental Mutagenesis*, 7(6):923–932. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/em.2860070614.

Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343):375–378. Number: 7343 Publisher: Nature Publishing Group.

Smeds, L., Mugal, C. F., Qvarnström, A., and Ellegren, H. (2016). High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS Genetics*, 12(5):e1006044. Publisher: Public Library of Science.

Subramanian, S. (2019). Population size influences the type of nucleotide variations in humans. *BMC Genet*, 20(1):1–12. Number: 1 Publisher: BioMed Central.

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences*, 109(45):18488–18492. Publisher: Proceedings of the National Academy of Sciences.

Webster, M. T., Axelsson, E., and Ellegren, H. (2006). Strong Regional Biases in Nucleotide Substitution in the Chicken Genome. *Molecular Biology and Evolution*, 23(6):1203–1216.

Winkler, H. (1930). Die Konversion der Gene : eine vererbungstheoretische Untersuchung. *G Fischer*.

12

The evolution of GC-biased gene conversion in the light of its molecular mechanisms

12.1 The relationship between gBGC and effective population size

In the previous section, we modelled the influence of natural selection on the evolution of gBGC. We identified parameters that can induce a shift in gBGC optimal value, giving insights into the potential causes for variations in its strength across the tree of life. When investigating the relationship between N_e and the equilibrium value of the transmission bias towards GC, we recovered a negative relationship between b^* and N_e, but still a positive one between B^* and N_e. While this result is consistent with observations in mammals (Lartillot, 2013; Subramanian, 2019; Galtier, 2021), it is not what has been observed either in butterflies, where no correlation has been found for closely related *Leptidea* species (Boman *et al.*, 2021), or in a relatively large sample of 29 animals

spanning a wide range of effective population sizes (Galtier *et al.*, 2018), or in a sample of 11 angiosperms (Clément *et al.*, 2017). As said in chapter 11, this observation might be due to an intrinsic negative correlation between b_0 and N_e that might be explained by a direct dependency between the DNA repair machineries and heterozygosity (Lesecque *et al.*, 2013; Li *et al.*, 2019). Those explanations remain however verbal and it is not clear which patterns to expect exactly if those mechanism were to be modelled (Galtier *et al.*, 2018). In this section, I therefore explicitly model how the repair mechanisms described so far can influence the correlation between B and N_e, and discuss the relevance of investigating the evolution of gBGC through the lens of those repair mechanisms.

12.1.1 Gene conversion models

Let us consider three models of gene conversion with two different kinds of AT/GC mismatches. Some mismatches are repaired in a biased way towards GC with a probability $(1 + b_0)/2$, while others are repaired in a unbiased way towards GC with probability $1/2$.

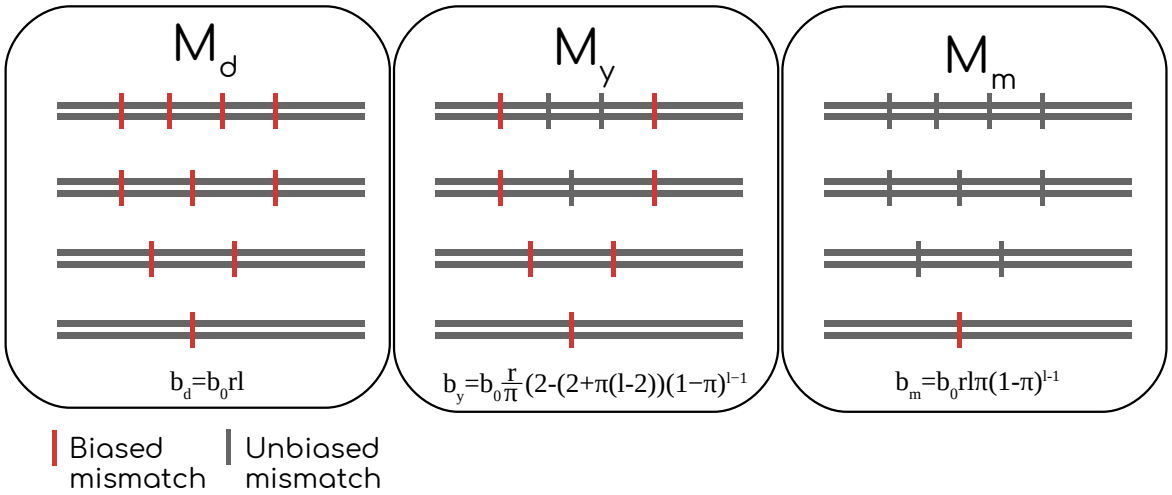


Figure 12.1: *Schematic view of the different gene conversion models. M_d All mismatches are biased. M_y: Only mismatches at the extremity of the conversion tract are biased. M_m: Only mismatches that are alone in the conversion tract are biased.*

The default conversion model

In the first model (M_d), all mismatches found in heteroduplexes are repaired towards GC with probability $(1 + b_0)/2$ (Figure 12.1). In this model, the transmission bias b_d is equal to $b_0 rl$. This is the model we assumed in chapter 11.

The mouse conversion model

In the second model (M_m), only mismatches that are alone in the heteroduplex are repaired with probability $(1 + b_0)/2$ (Figure 12.1). The probability for a mismatch to be alone in the heteroduplex will depend on the density of heterozygous sites ($\pi = 4N_e\mu$ where μ is the mutation rate per generation) and the size of the heteroduplex approximated by the length of the conversion tract l.

Let us call K a random variable that describes the number of mismatches that are expected in a given heteroduplex. Assuming that heterozygous site are evenly distributed in the genome, K follows a binomial law of parameter π and l. The probability for a heterozygous site to be converted with bias can be written:

$$\mathbb{P}_{biased} = \frac{r}{\pi} P(K = 1) \tag{12.1}$$

Where r is the recombination rate per base pair and $P(K = 1) = l\pi(1 - \pi)^{l-1}$. The genome-wide effective transmission bias b_m will be:

$$b_m = b_0 r l (1 - \pi)^{l-1} \tag{12.2}$$

This model describes the mechanism found in mice by Li *et al.* (2019).

The yeast conversion model

In the third model (M_y), only mismatches that are at the extremity of a conversion tract are biased (Figure 12.1). This means that when conversion tracts involve two mismatches or more, only two mismatches are repaired in a biased way, and the rest are repaired without bias. Thus:

$$\mathbb{P}_{biased} = \frac{r}{\pi}(1 \times P(K = 1) + 2 \times P(K \geq 2)) \tag{12.3}$$

Which gives:

$$b_y = b_0 \frac{r}{\pi}(2 - (2 + \pi(l - 2))(1 - \pi)^{l-1} \tag{12.4}$$

This model describes the mechanism found in the budding yeast COs by Lesecque *et al.* (2013).

12.1.2 Empirical calibration

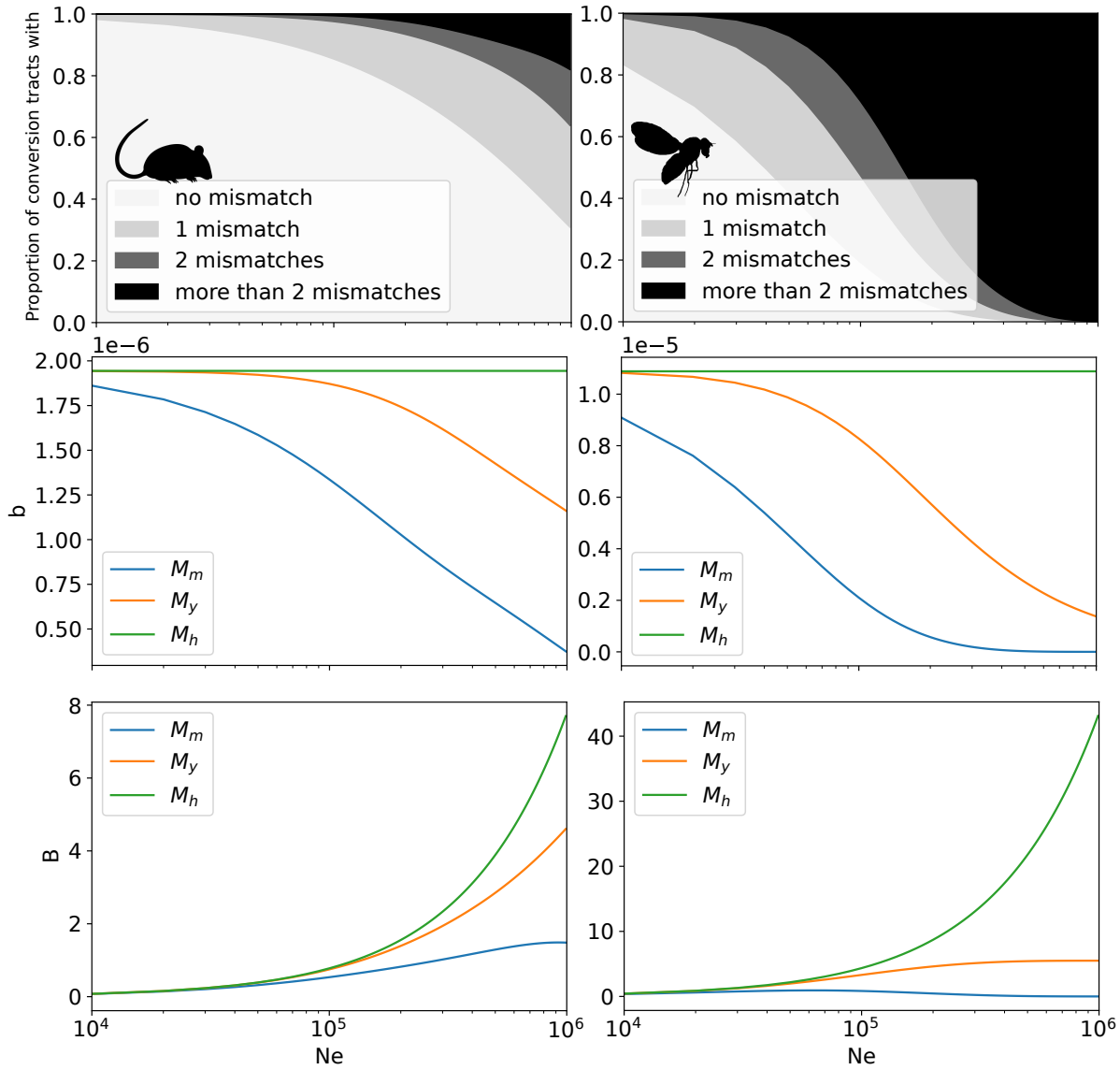


Figure 12.2: *Top: Proportion of conversion tracts with no, one, two or more mismatches as a function of effective population size. Middle: gBGC coefficient (b) as a function of effective population size with the model M_d in green, M_m in blue and M_y in orange. Bottom: Population-scaled gBGC coefficient (B) as a function of effective population size with the model M_d in green, M_m in blue and M_y in orange. Left: Using the mouse set of empirical parameters values. Right: Using the D.melanogaster set of empirical parameters values.*

I assume that the recombination rate (r), the mutation rate (μ) and the length of the conversion tract (l) are fixed parameters. I consider two sets of empirical values for these parameters. The first one corresponds to empirical values estimated in mice: $r_{NCO} = 6 \times 10^{-8}$, $r_{CO} = 6 \times 10^{-9}$, $\mu = 5.4 \times 10^{-9}$, $l_{NCO} = 50$ and $l_{CO} = 400$ (Cox *et al.*,

2009; Cole *et al.*, 2014; Adewoye *et al.*, 2015; Long *et al.*, 2018; Li *et al.*, 2019). The second one corresponds to empirical values estimated in *Drosophila melanogaster*: $r_{NCO} \simeq r_{CO} = 2.2 \times 10^{-7}$, $\mu = 5.5 \times 10^{-9}$, $l_{NCO} = 400$ and $l_{CO} = 1000$ (Miller *et al.*, 2016; Long *et al.*, 2018).

I consider the repair bias of the bias-inducing DNA repair pathway fixed at $b_0 = 0.68$, which correspond to the bias of single mismatches in humans and mice (Halldorsson *et al.*, 2016; Li *et al.*, 2019).

12.1.3 An intrinsic negative correlation between b and N_e

I can then compute the expected value of the transmission bias as a function of the effective population size N_e for all three gene conversion models, in the two sets of empirical parameters (Figure 12.2). It appears that when the effective population size is small, and thus π is small, the three mechanisms lead to the same transmission bias (Figure 12.2). This can be seen analytically: when computing $\lim_{\pi \to 0} b_m$ and $\lim_{\pi \to 0} b_y$, one directly obtains b_d. Using the mouse set of empirical parameters, as N_e increases, increasing numbers of conversion tracts include two mismatches or more and are therefore not biased in the model M_m, and less biased in the model M_y (Figure 12.2). This leads to a decrease of the transmission biases b_m and b_y with N_e, and therefore a saturation of $B = 4N_e b$ with N_e (Figure 12.2). Of note, under this set of empirical parameters and for this range of N_e variation, even if there is a negative correlation between b and N_e, we still expect a positive correlation between B and N_e (Figure 12.2).

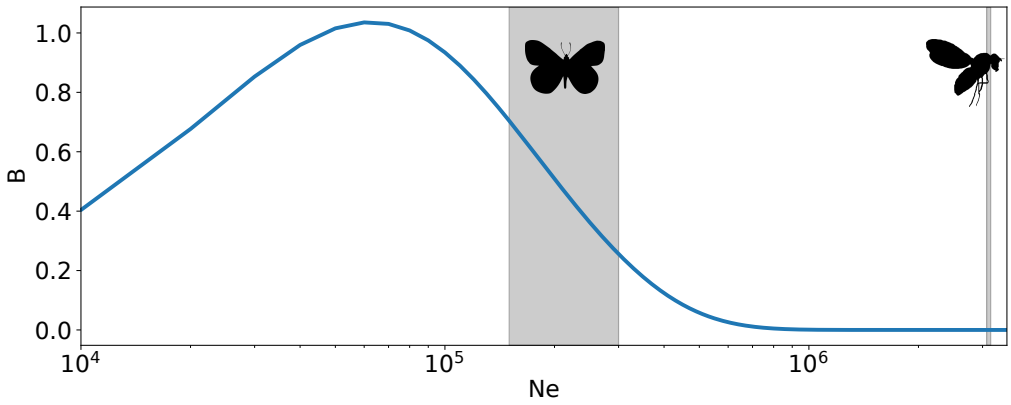


Figure 12.3: *Zoom on the relationship between the population-scaled gBGC coefficient B and effective population size N_e, using the D.melanogaster set of empirical parameters values. The grey areas correspond to the range of N_e variations of D.melanogaster African populations taken from the study of Kapopoulou et al. (2018), and Leptidea butterflies taken from the study of Boman et al. (2021). For butterflies I used heterozygocity that I transformed to effective population size with $N_e = \pi/4\mu$, using the mutation rate of D.melanogaster.*

Using the *D.melanogaster* set of empirical parameters, B_d becomes very large for large values of N_e (Figure 12.2). On the other hand, B_y seems to saturate. This can be explained by the fact that with large π and long conversion tracts, almost every conversion tract includes more than 2 mismatches (Figure 12.2). Even if the number of heterozygous sites increases with increasing effective population size, the number of sites that will be converted with bias stays still, and $b_y \simeq 2b_0 r/\pi$, which gives $B_y \simeq 2b_0 r/\mu$. Therefore, for large N_e and large l, B_y does not depend on N_e anymore. Nevertheless, under the yeast model and for this range of N_e variation, even if there is a negative correlation between b and N_e, we still expect a positive correlation between B and N_e (Figure 12.2).

Finally, under the mouse gene conversion model (M_m) and the *D.melanogaster* set of empirical parameters, as N_e becomes large, there is almost no conversion tracts with only one mismatch and thus $b_m \simeq 0$ (Figure 12.3). Interestingly, there is only weak evidence for gBGC in *D.melanogaster* except for some parts of its genome (Galtier *et al.*, 2006; Robinson *et al.*, 2014; Jackson and Charlesworth, 2021). If in fact *D.melanogaster* shared its gene conversion mechanism with mice and humans, we would not expect any gBGC, only because wild populations of fruit flies have very high effective population size ($\simeq 3 \times 10^6$ for African populations) (Kapopoulou *et al.*, 2018). However, other *Drosophila* species with lower N_e could have gBGC and paradoxically, in the range of $N_e = 10^5 \rightarrow 10^6$, we should even expect a negative correlation between B and N_e (Figure 12.3). Consistent with this prediction, the two Leptidea species/populations with the lowest genetic diversity are also the ones with the highest population-scaled gBGC coefficient (Boman *et al.*, 2021).

12.2 Decoupling somatic vs meiotic repair bias

In the previous chapter, we hypothesised that the meiotic gene conversion bias towards GC could be under negative selection to limit the penetrance of the somatic repair bias in meiosis. If this repair bias is strongly biased towards GC such as to minimize the somatic mutation rate, it could lead to a very strong fitness cost because of gBGC. In this case, there is a selective advantage to decouple the somatic and the meiotic repair bias. One way to achieve this could be to underexpress the enzymes that are responsible for this bias in meiosis. However, meiotic cells are *a priori* just as susceptible to single base pair damage as somatic cells. In *E.Coli*, the inactivation of the adenine DNA glycosilase MutM leads to a multiplication by 80 of the GC $\mapsto$ AT mutation rate (Foster *et al.*, 2015). Inactivating this bias-inducing repair enzyme in meiosis is therefore quite risky given the increase of the mutation rate it implies for meiotic cells and might not

be the optimal way to increase fitness.

Alternatively, the mechanisms presented in the previous section provide a simple way to decouple somatic DNA repair from the ones involved in the repair of mismatches in the context of a heteroduplex. If we consider that at a given time, single-base-pair damages are relatively rare, they should be relatively isolated. If cells treat isolated damages with biased enzymes (such as DNA glycosilases), the mutation rate should be limited. In turn, depending on the heterozygosity, one can expect several mismatches in a single heteroduplex and therefore they should be more often clustered. Treating clustered mismatches with an unbiased machinery (such as the MMR) could therefore provide a way to treat differently somatic DNA damages and mismatches caused by a heteroduplex. Importantly this trick does not require any change in expression of DNA repair enzymes between mitotic and somatic cells and ensures both a heavily biased DNA damage repair and a relatively unbiased meiotic gene conversion. I showed that at equal levels of heterozygosity, this decoupling can be achieved by a change in the size of the heteroduplex. Interestingly, rather than the expression of DNA repair enzymes, the size of the heteroduplex may be the quantitative trait that regulates the strength of gBGC, especially in large populations where both the cost of gBGC and the probability of having clustered mismatches are higher.

Of note, there is an advantage to decouple somatic and meiotic DNA repair bias only if the somatic repair bias is inducing a b higher than b^*. Whether this is true empirically is however difficult to assess, mainly because b^* is quite sensitive to parameters that are themselves difficult to estimate precisely (see chapter 11). Moreover, it can also be that DNA repair models described above (M_m and M_y) correspond to the models that limit the somatic mutation rate better than M_d, and there is no need to invoke any selection on biased gene conversion to explain their existence. The limitation of gBGC with increasing effective population size could just be a coincidence.

12.3 Conclusions and future directions

12.3.1 The diversity of the molecular mechanisms of gBGC

So far, only three molecular mechanisms for GC-biased gene conversion have been described. Two of them imply a direct negative correlation between the strength of gBGC b and genetic diversity. Interestingly, (Li *et al.*, 2019) showed that conditional to having only one mismatch in the conversion tract, the intensity of the repair bias in humans and mice are equal, and the authors suggest that they share the same repair

mechanism. However, these results were found by excluding a subset of crossovers in humans, which show a strong GC-bias (Laurent Duret personal communications). This could suggest that the differences of b_0 we observe between species are strongly shaped by differences in heterozygosity and/or the size of the heteroduplex. However, our hypothesis are as good as the variations we can observe and interpret. Indeed, only three mechanisms have been observed so far, but they have been investigated only in a handful of species. Therefore, behind those three mechanisms may hide plenty of others. However, identifying those mechanisms remains challenging and even with decreased sequencing costs, whole genome sequencing of pedigrees over several generations remain scarce. Moreover for the mouse mechanism to be observed, there is a need for at least some recombination events to have isolated mismatches, which is difficult to observe when heterozygosity is high and heteroduplexes are large. On the other hand, inbred line strains (when viable) could in fact provide a good test on whether b_0 depends on heterozygosity. Despite all the difficulties, I am deeply convinced that a better understanding of the evolution of gBGC will certainly be achieved through a clearer vision of the diversity of repair mechanisms in eukaryotes.

12.3.2 The genetic architecture of gBGC

Even if there are some insight on the type of mismatches that are converted in a biased way, the proteins involved in this bias and its variation have not been identified yet even in model species. Several approaches could allow us to identify those proteins. When one wonders how a phenotypic trait is genetically encoded, several questions arise: how many loci are involved ? what are their effect sizes on the phenotype ? are those effects additive ?

To answer those questions, the most common approach is the so-called Genome-Wide Association Studies (GWAS) (see Uffelmann *et al.* (2021) for a review). The idea is to find statistical associations between the genotype at each locus to a continuous trait. This method can only detect loci involved in variations that exist inside a population, missing a important part of the genetic architecture: large effect loci. Li *et al.* (2019) showed that the difference in the strength of the bias between humans and mice was mainly due to different heterozygocity. There is therefore not much hope for the strength of gBGC to be very variable inside a single populations, but it is not impossible. When the effect sizes are small, this approach requires sequencing and phenotyping a very large cohort of individuals to get enough statistical power. Unfortunately, measuring gBGC in one individual represent a substantial investment in time and money, and for all these reasons this method is for now not applicable to gBGC with reasonable costs and efforts.

Alternatively, several methods use association between genotype and phenotype across a phylogeny of much more distant species (Mayrose and Otto, 2011; Pease *et al.*, 2016; Levy Karin *et al.*, 2017; Kowalczyk *et al.*, 2019; Partha *et al.*, 2019; Duchemin *et al.*, 2023; Fukushima and Pollock, 2023). One of the more robust method uses a shift in the site-specific amino-acid fitness profiles associated to the phenotype to infer association between the evolution of a site and the evolution of the phenotype (Duchemin *et al.*, 2023). This method is for now only applicable to discrete phenotypes, but a continuous version is already in development (Bastien Boussau and Philippe Veber personal communications).

However, given that fitness profiles are estimated at the amino-acid level, they are sensitive to gBGC (Duchemin *et al.*, 2023). And if the trait for which we seek phenotype-genotype association is gBGC, nearly all sites that are found in highly recombining genes will be associated to the trait, which is problematic. There are two options: the first is to estimate fitness profiles at the DNA level, as we did in chapter 8. Those profiles should be robust to gBGC because they control for non-selective substitution patterns, but they are not in practice (ongoing work). This method could be however modified to explicitly account for gBGC, but it is far more costly in computation resources to estimate fitness profiles at the DNA level in a Bayesian framework compared to the method presented above (Duchemin *et al.*, 2023). A fast and integrated approach to identify sites that co-evolve with gBGC would therefore require substantial development. Another trick could be to identify the predictable signature that gBGC leaves on fitness profiles (it increases the fitness of GC-encoded amino-acids) to *a posteriori* exclude sites whose changes in inferred fitness landscape is solely due to gBGC. However, two major problem remain.

First, if gBGC is driven by heterozygosity, we should not expect any particular locus to be associated to the strength of gBGC. The second problem is that the transmission bias b_0 is not available for many species. The only measure that is readily accessible with existing methods is B, but it is confounded with recombination rate and effective population size. When looking for candidate loci, we will not be able to distinguish those that are truly involved in the evolution of gBGC from those that are involved in the evolution of recombination rates or N_e-related traits such as body size or lifespan. Multivariate methods that can account for several traits jointly could be a way forward to resolve this problem and disentangle the part of the variance that is explained by recombination rate or effective population size from that of the repair bias towards GC.

Part V

Conclusion

This PhD was funded by ANR project "HotRec" whose aim was to understand the raison d'être of the diversity in fine-scale recombination landscapes in animals. Within this project, my task was to evaluate the role of natural selection in shaping this diversity. More specifically, I was to investigate whether the beneficial effects of recombination (dissipation of genetic linkage), and its deleterious effects (gBGC and the mutagenic effect) could have played a role in selecting one type of recombination landscape over another.

Even though I did not provide any definitive answer to this question, I believe that chapter 6, 7 and 11 yield some valuable insights. Of note, during this PhD, I did not test or quantify the beneficial effect of recombination on the efficacy of selection. Indeed, in chapter 9, I showed that current methods did not allow to test for a beneficial effect of recombination by comparing signatures of adaptive or purifying selection between genes that have evolved under different recombination rates in presence of gBGC. Overall, given the results of this thesis and the recent literature, I believe that recombination landscapes most likely reflect constraints on the efficiency of chromosome pairing rather than indirect selection to optimise genetic shuffling or to reduce the deleterious load due to gBGC or the mutagenic effect of recombination.

Second, I believe that the present work contributes to a better understanding of positive selection, its causes and its dynamics across the genome in presence of gBGC. I hope that chapter 9 has convinced that modelling the evolution of a sequence on a fitness landscape can reveal surprising patterns that are not accounted for using a fixed DFE.

Finally, I investigated the factors influencing the evolution of gBGC. Chapter 12 revealed that it is not trivial to determine whether gBGC should be positively or negatively selected, even if the overall consequences on the genome appear to be deleterious. Anyway, I believe that constraints linked to the way DNA is repaired might be the primary factors influencing the evolution of gBGC, although selection on the DNA repair machinery cannot be completely excluded either.

Overall, although I have not been able to complete the original task that I was given by the ANR, I believe that this thesis still represents an important step forward in our understanding of several aspects of molecular evolution.

Remerciements

Il est maintenant grand temps de remercier toustes celleux qui ont rendu ce travail possible.

J'aimerais remercier toustes les chercheu.r.se.s du LBBE avec qui j'ai pu intéragir scientifiquement, et de qui j'ai énormément appris, notamment Bastien Boussau, Anouk Necsulea, Carina Mugal, Laure Ségurel, Damien de Vienne, Etienne Rajon, Annabelle Haudry, Sylvain Charlat, Sylvain Mousset, Mélodie Bastian, Florian Bénitière, Alice Genestier, Mark Stoneking, Philippe Veber et toustes celleux que j'ai oublié. J'aimerais aussi remercier tous les membres du pôle info et du pôle admin pour leur bienveillance, leur patience et leur disponibilité. Notamment merci à celleux que j'ai le plus embêté : Philippe Veber, Bruno Spataro, Stéphane Delmotte, Simon Penel, Christophe Blanchet et Nathalie Arbasetti. J'aimerais remercier l'équipe de direction du LBBE, l'animatorium et plus généralement les personnes qui font en sorte que le LBBE soit aussi animé scientifiquement, ce qui a largement participé ma motivation .

J'aimerais remercier les collègues de l'ANR HotRec, les collègues de l'ISEM à Montpellier, de EcoEvoPaleo à Lille et puis toustes les collègues rencontré.e.s en conférences ou de visite à Lyon pour les discussions passionantes qui m'ont permises d'améliorer ma compréhension de l'évolution en général. J'aimerais remercier mes collègues paléontologues: Gilles Escarguel, Antoine Louchart, Julien Clavel, et Anaïs Duhamel pour m'avoir fait regarder l'évolution sous un autre angle et prendre du recul sur mon travail.

Merci aux stagiaires que j'ai co-encadrés pendant ces 3 ans: Aurélie Fischer, Jeanne Pithon, Louis Schroll et Augustin Cléssin pour leur enthousiastme et leur motivation.

Un grand merci à l'équipe de doctorant.e.s et postdocs du LBBE pour les excellents moment passés ensembles : Léa Bariod, Mélodie Bastian, Florian Bénitière, Solène Cambreling, Emilie Fleurot, Alice Genestier, Benjamin Guinet, Chloé Haberkorn, Léa Keurinck, Lucas Lalande, Alexandre Laverré, Camille Mayeux, Marie Morel, Alexia Nguyen Trung, Lisa Nicvert, Djivan Prentout, Théo Tricou, Mary Varoux, Barbara Vuillaume et tous les autres avec qui j'ai un peu moins interagi.

J'aimerais aussi remercier Marie Sémon, Cyril Charles, Stéphane Vincent et les autres professeur.e.s qui m'ont soutenu et m'ont permis de traverser une scolarité difficile à l'ENS de Lyon. Sans leur soutien, je ne serais sans doute pas arriver jusqu'à la thèse.

J'aimerais exprimer toute ma reconnaissance à mon directeur de thèse, Laurent Duret. Après trois ans dans le milieu académique, je me suis rendu compte qu'il y a peu de personnes qui ont une approche aussi saine de l'encadrement. Du début à la fin, il m'a laissé totalement maître de mes directions de recherches, il a toujours respecté mes choix et mes ressentis, il m'a fait confiance, avec une patience, une disponibilité et une bienveillance sans limites. Un grand merci aussi à Nicolas Lartillot mon co-directeur, pour son honnêteté, sa sagesse et ses conseils qui ont profondément marqué ma manière d'appréhender la recherche en évolution.

J'aimerais remercier Thibault Latrille pour son amitié et le plaisir immense que j'ai eu a travailler avec lui pendant ces trois ans ainsi que mes amis, collègues et bientôt co-auteurs Djivan Prentout, Alex Laverré et Théo Tricou pour leur accueil dans ce laboratoire et tous les apéros/D&D/bars/fêtes/vacances.

Enfin j'aimerais remercier ma famille et mes amis pour leur soutien et leur amour, ainsi que ma compagne, Anaïs, pour absolument tout.

Part VI

Annexes

Supplementary materials from: High prevalence of Prdm9-independent recombination hotspots in placental mammals

J. Joseph[1], D. Prentout[2], A. Laverré[3,4], T. Tricou[1], L. Duret[1]

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

[2]Department of Biological Sciences, Columbia University, New York, NY 10027, USA

[3]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[4]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

julien.joseph@ens-lyon.fr

October 4, 2023

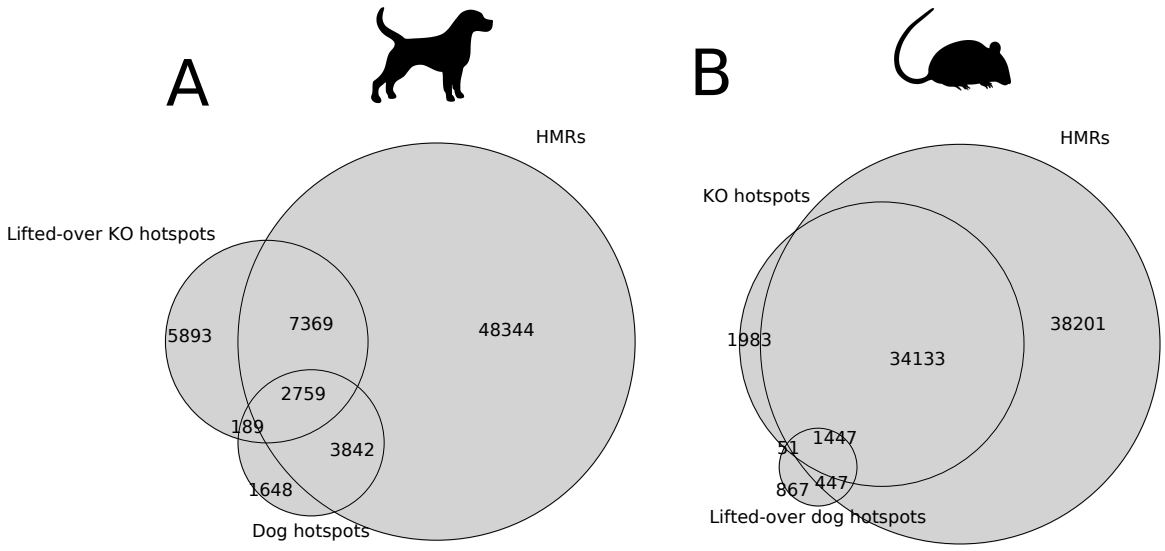


Figure S1: A) Venn diagramm representing the overlap between MDH loci in the dog genome, dog hotspots and dog HMRs in the dog genome. B) Venn diagramm representing the overlap between lifted-over dog hotspots (DRH) on the mouse genome, mouse Prdm9$^{-/-}$ hotspots and mouse HMRs in the mouse genome. Features were considered to overlap if their midpoint were less than 5 kb apart. The numbers indicated in the circles correspond to the number of HMRs that overlap the other features, and dog hotspots in intersections between Prdm9$^{-/-}$ mouse hotspots and dog hotspot.

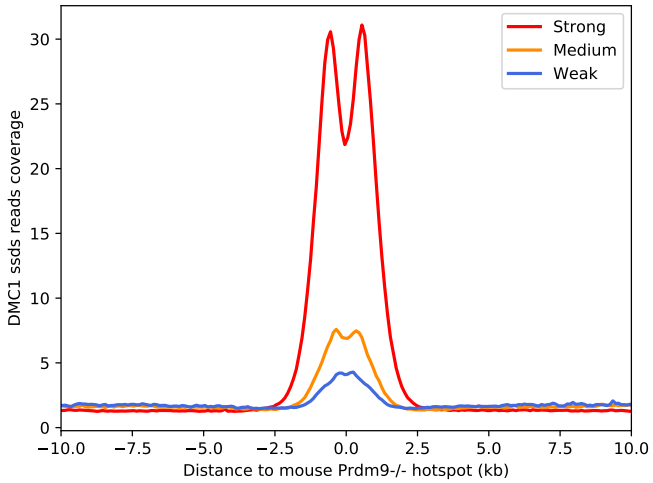


Figure S2: DMC1 ssds read coverage as a function of the distance to the closest MDH in the mouse genome. Hotspots were divided in three equally sized categories of strength: strong hotspots in red (10-190 FPKM), medium hotspots in orange (5-10 FPKM) and weak hotspots in blue (0-5 FPKM). The line directly correspond to the mean value of DMC1 ssds read coverage in a 100 bp window.

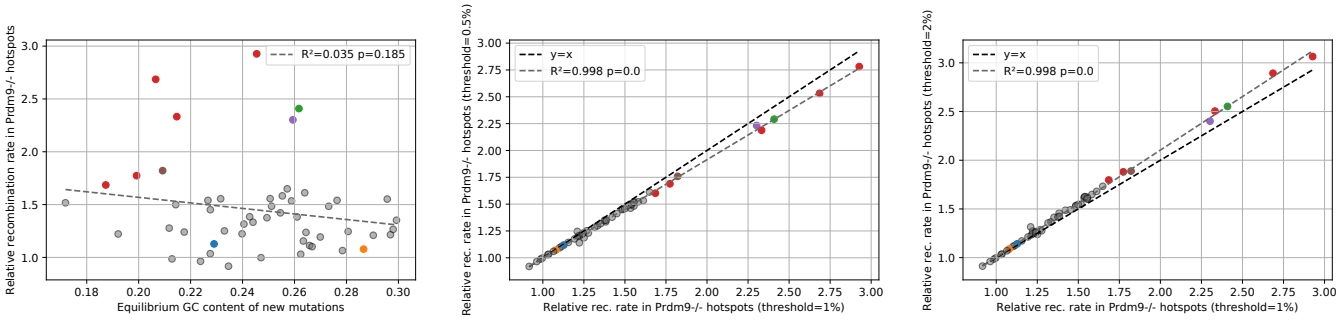


Figure S3: A) Relative recombination rate in lifted-over mouse Prdm9$^{-/-}$ hotspots as a function of the estimated GC^* of new mutations in all 52 species of mammals. B) and C) Relative recombination rate in lifted-over mouse Prdm9$^{-/-}$ hotspots using different thresholds to estimate GC^*_{mut} (0.5% in B and 2% in C) as a function of the relative recombination rate in lifted-over mouse Prdm9$^{-/-}$ hotspots using the 1% threshold used throughout the study in all 52 species of mammals. Red points correspond to canids, the green point indicate ring-tailed lemurs, the purple point northern elephant seals, the blue point mice, the brown point daurian ground squirrels, and the orange point humans.

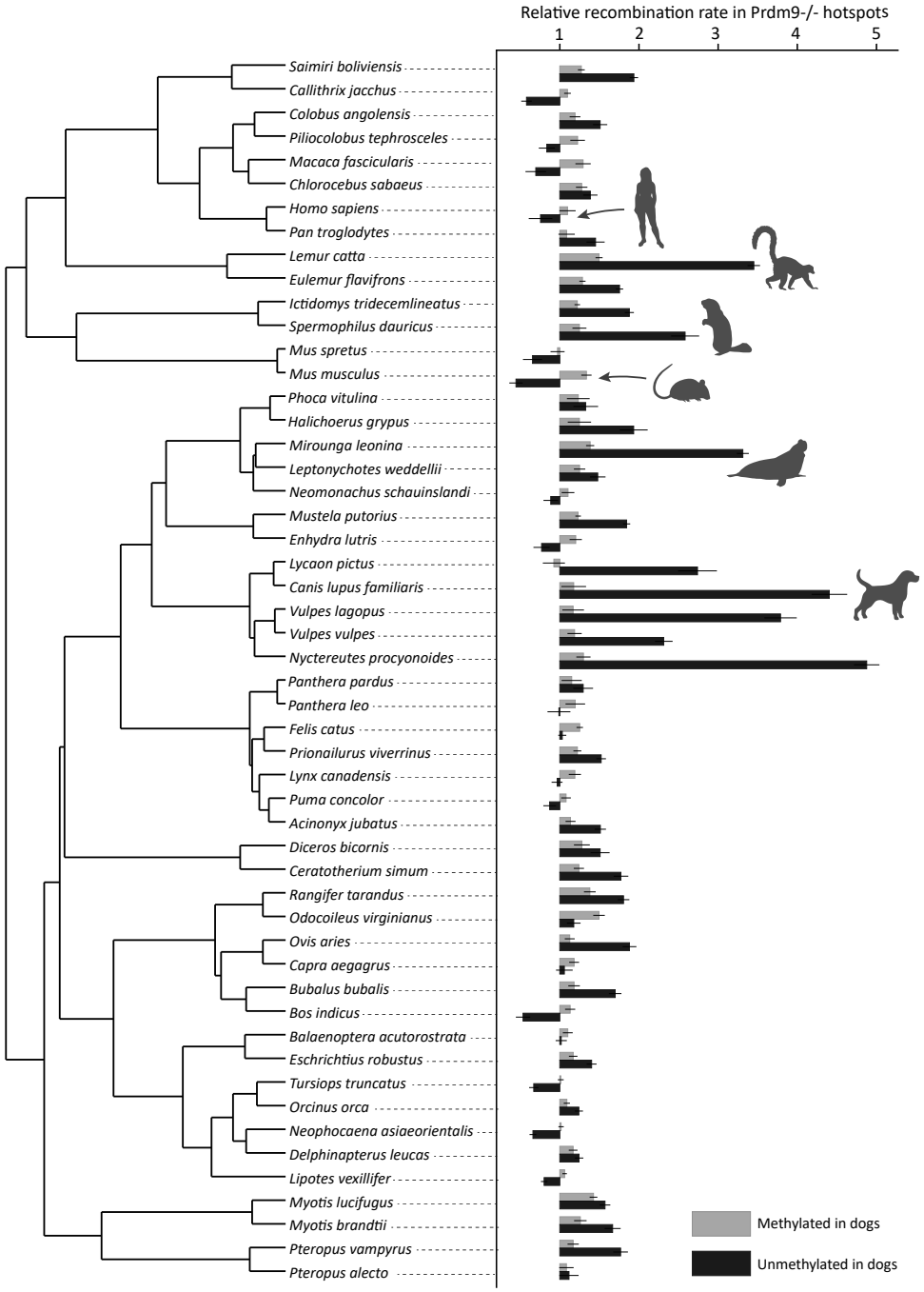


Figure S4: Relative recombination rates at MDH loci in 52 mammals. The hotspots were splited in two categories depending on their methylation level in dog sperm. The number of lifted features for each categories varied from ∼4,000 in *Myotis brandtii* to ∼9,000 in *Mus spretus*. Detailed numbers of mapped features are presented in Supplementary Table XX. The tree have been retrieved from TimeTree5 Kumar et al. (2022). Error bars correspond to a 95% confidence interval obtained by bootstrapping the substitutions for computing GC^*_{flank} and GC^*_{hot}.

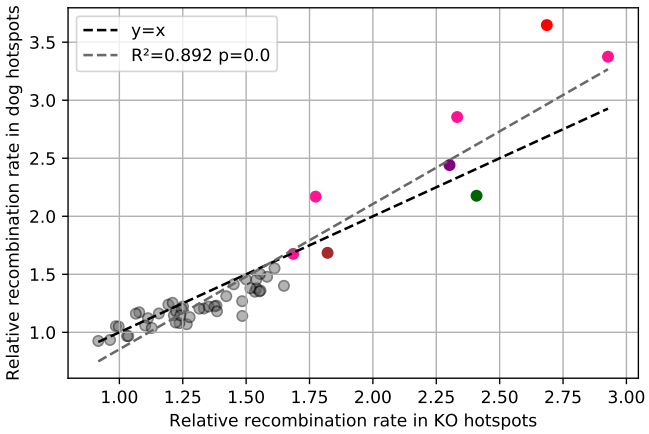


Figure S5: Relative recombination rate in lifted-over dog LD-based hotspots as a function of the relative recombination rate in MDH loci. Red points correspond to canids, the green point indicate ring-tailed lemurs, the purple point northern elephant seals, the blue point mice, the brown point daurian ground squirrels, and the orange point humans.

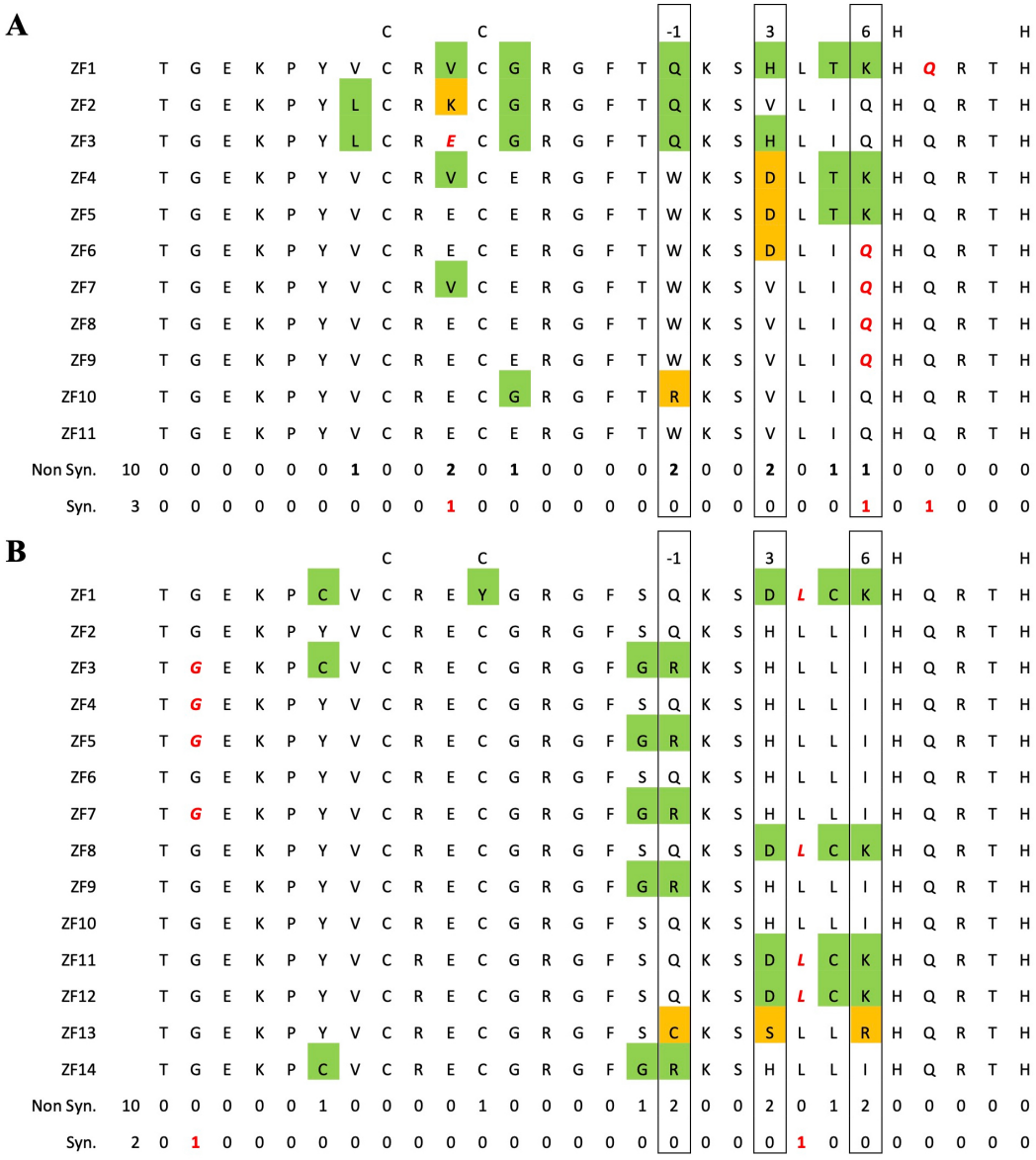


Figure S6: Sequence diversity of zinc finger arrays of the Prdm9 allele present in the reference genome of *Spermophilus dauricus* (A) and *Lemur catta* (B). Amino-acid variants are highlighted by a colored background. Synonymous variants are written in red. The number of non-synonymous and synonymous variants is indicated at each position. In both species, amino-acid variants are enriched at positions -1, 3, and 6 that are involved in DNA binding. Under neutral evolution, one would expect approximatively one third of variants to be synonymous and two thirds to be non-synonymous. In both species, the number of non-synonymous variants exceeds the neutral expectation, which is suggestive of positive selection. *Lemur catta* Prdm9: RefSeq protein accession number: XP_045389275.1_1, genomic coordinates of the exon encoding the zinc finger array: complement(NC_059147.1: 34824301..34826009). *Spermophilus dauricus* Prdm9: the protein is not annotated in RefSeq, genomic coordinates of the exon encoding the zinc finger array: KZ296155.1: 118252..119699.

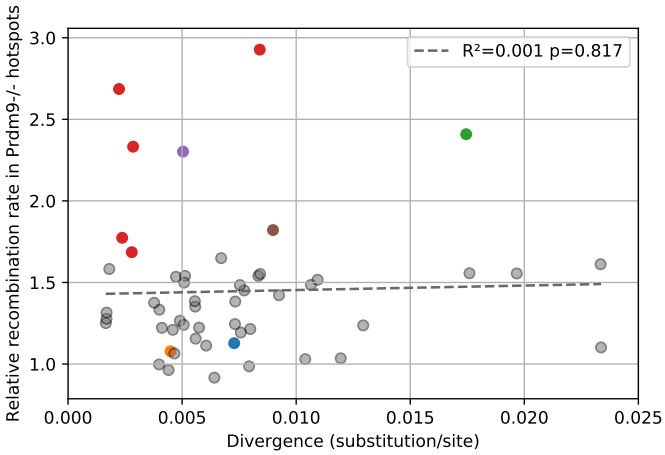


Figure S7: Relative recombination rate in lifted-over mouse Prdm9$^{-/-}$ hotspots as a function of the estimated length of the branch used to compute GC^* in substitutions per/site. Total number of sites were computed with sites for which all threes species (target sister and outgroup) were genotyped and aligned, excluding CpG dinucleotides. Red points correspond to canids, the green point indicate ring-tailed lemurs, the purple point northern elephant seals, the blue point mice, the brown point daurian ground squirrels, and the orange point humans.

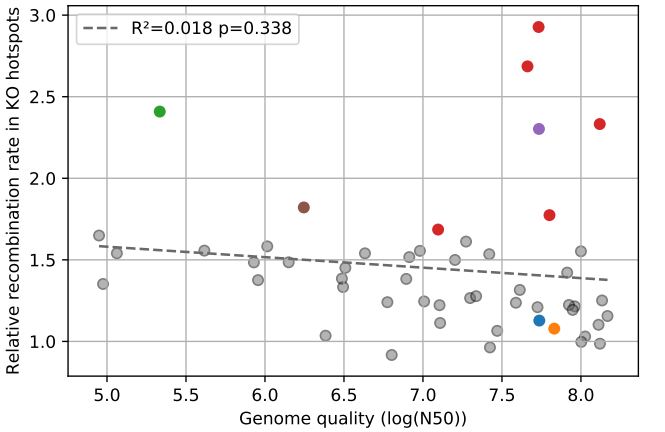


Figure S8: Relative recombination rate in lifted-over mouse Prdm9$^{-/-}$ hotspots as a function of the genome quality summarized by the log of the N50 metric. Red points correspond to canids, the green point indicate ring-tailed lemurs, the purple point northern elephant seals, the blue point mice, the brown point daurian ground squirrels, and the orange point humans.

References

Sudhir Kumar, Michael Suleski, Jack M Craig, Adrienne E Kasprowicz, Maxwell Sanderford, Michael Li, Glen Stecher, and S Blair Hedges. TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, 39(8):msac174, August 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac174.

Mammalian protein-coding genes exhibit widespread beneficial mutations that are not adaptive

T. Latrille[1†], **J. Joseph**[2†], **D. A. Hartasánchez**[1], **N. Salamin**[1]

[1]Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland

[2]Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Lyon 1, Villeurbanne, France

[†]These authors contributed equally to this work

thibault.latrille@ens-lyon.org

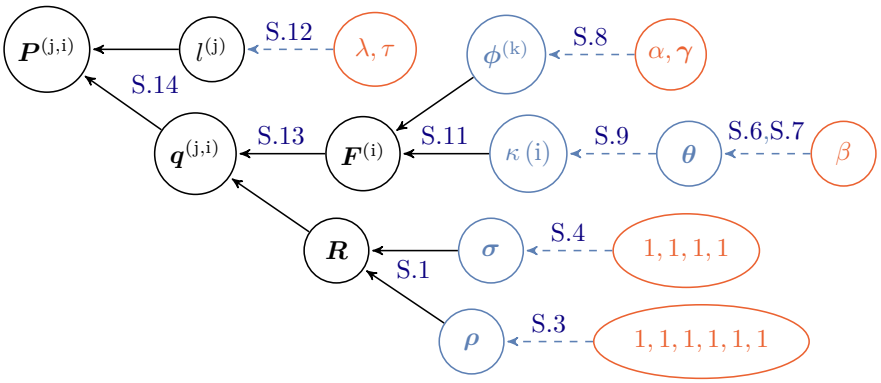
Supplementary materials

Contents

1 Mutation-selection codon models

1.1 Prior distributions and parameters of the model

The parameterization of the models is described as a Bayesian hierarchical model, including the prior distributions and the parameters of the model. This hierarchical model is formally represented as directed acyclic graph of dependencies between variables, depicted below. Nodes of the directed acyclic graph are the variables, and edges are the functions. Hyper-parameters are depicted in red circles, random variables in blue circles, and transformed variables in black. Blue dashed line denotes a drawing from a random distribution, and black solid lines denote a function. All the nodes pointing toward a given node (upstream) are its dependencies which determines its distribution. The other way around, following the arrows in the DAG (downstream), simple prior distributions are combined together to form more complex joint prior distribution which ultimately defines the prior distribution of the model.



1.2 Nucleotide mutation rates

The generalized time-reversible (GTR) nucleotide mutation rate matrix $\boldsymbol{R}$ is a function of the nucleotide frequencies $\boldsymbol{\sigma}$ and the symmetric exchangeability rates $\boldsymbol{\rho}$ [1]. $\boldsymbol{\sigma} = (\sigma_A, \sigma_C, \sigma_G, \sigma_T)$ is the equilibrium base frequency vector, giving the frequency at which each base occurs at each site. $\boldsymbol{\rho} = (\rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT}, \rho_{GT})$ is the vector of exchangeabilities between nucleotides. Altogether, the rate matrix is:

$$\boldsymbol{R} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & \rho_{AC}\sigma_C & \rho_{AG}\sigma_G & \rho_{AT}\sigma_T \\ \rho_{AC}\sigma_A & - & \rho_{CG}\sigma_G & \rho_{CT}\sigma_T \\ \rho_{AG}\sigma_A & \rho_{CG}\sigma_C & - & \rho_{GT}\sigma_T \\ \rho_{AT}\sigma_A & \rho_{CT}\sigma_C & \rho_{GT}\sigma_G & - \end{pmatrix} \end{array} \tag{S.1}$$

By definition, the sum of the entries in each row of the nucleotide rate matrix $\boldsymbol{R}$ is equal to 0, giving the diagonal entries:

$$R_{a,a} = - \sum_{b \neq a, b \in \{A,C,G,T\}} R_{a,b} \tag{S.2}$$

The prior on the exchangeabilities $\boldsymbol{\rho}$ is a uniform Dirichlet distribution of dimension 6:

$$\boldsymbol{\rho} \sim \mathrm{Dir}\,(1, 1, 1, 1, 1, 1)\,. \tag{S.3}$$

The prior on the equilibrium base frequencies $\boldsymbol{\sigma}$ is a uniform Dirichlet distribution of dimension 4:

$$\boldsymbol{\sigma} \sim \mathrm{Dir}\,(1, 1, 1, 1) \tag{S.4}$$

The general time-reversible nucleotide matrix is normalized with a total flow of 1:

$$\sum_{a\in\{A,C,G,T\}} -\sigma_a R_{a,a} = 1, \tag{S.5}$$

such that we expect 1 substitution per unit of branch length.

1.3 Site-specific amino-acid fitness profiles

Site-specific amino-acid fitness profiles are assumed i.i.d. from a mixture model, itself endowed with a truncated Dirichlet process prior. Specifically, the mixture has K components (K = 30 by default). The prior on component weights ($\boldsymbol{\theta}$) is modeled using a stick-breaking process, truncated at K and of parameter $\beta = 1$:

$$\begin{aligned} &\boldsymbol{\theta} \sim \text{StickBreaking}\left(\text{K}, \beta\right) \\ \iff &\theta_{\text{k}} = \psi_{\text{k}} \cdot \prod_{l=1}^{\text{k}-1} \left(1 - \psi_l\right), \ \text{k} \in \{1, \ldots, \text{K}\}, \end{aligned} \tag{S.6}$$

where ψ_{k} are i.i.d. from a beta distribution

$$\psi_{\text{k}} \sim \text{Beta}\left(1, \beta\right), \ \text{k} \in \{1, \ldots, \text{K}\}. \tag{S.7}$$

Of note, the weights decrease geometrically in expectation, at rate β, such that lower values of β induce more heterogeneous distributions of weights.

Each component of the mixture defines a 20-dimensional fitness profile $\boldsymbol{\phi}^{(\text{k})}$ (summing to 1), for $\text{k} \in \{1, \ldots, \text{K}\}$. These fitness profiles are i.i.d. from a Dirichlet of center $\boldsymbol{\gamma} = 1$ and concentration $\alpha = 1$:

$$\boldsymbol{\phi}^{(\text{k})} \sim \text{Dir}\left(\boldsymbol{\gamma}, \ \alpha\right), \ \text{k} \in \{1, \ldots, \text{K}\}. \tag{S.8}$$

Site allocations to the mixture components $\kappa\left(\text{i}\right) \in \{1, \ldots, \text{K}\}$, for $\text{i} \in \{1, \ldots, \text{N}\}$ running over the N sites of the alignment, are i.i.d. multinomial of parameter $\boldsymbol{\theta}$:

$$\boldsymbol{m} \sim \text{Multinomial}\left(\boldsymbol{\theta}\right), \tag{S.9}$$

$$\text{where } m_{\text{k}} = \sum_{\text{i}\in\{1,\ldots,\text{N}\}} \mathbb{1}_{\kappa(\text{i})=\text{k}} \tag{S.10}$$

For a given parameter configuration for the mixture, the scaled fitness $\boldsymbol{F}^{(\text{i})}$ at site i, are obtained by taking the logarithm of fitness assigned to this site:

$$\boldsymbol{F}^{(\text{i})} = \ln\left(\boldsymbol{\phi}^{(\kappa(\text{i}))}\right), \ \text{i} \in \{1, \ldots, \text{N}\}. \tag{S.11}$$

1.4 Branch length

The topology of the rooted phylogenetic tree is supposed to be known and is not estimated by the model. The branch lengths $l^{(\text{j})}$ are defined as the expected number of neutral substitutions per DNA site along a branch, each from a Gamma distribution of mean $\lambda = 0.1$ and scale $\tau = 1$:

$$l^{(\text{j})} \sim \text{Gamma}\left(\lambda, \tau\right). \tag{S.12}$$

1.5 Codon substitution rates

The mutation rate between codons a and b, denoted $\mu_{a\mapsto b}$ depends on the underlying nucleotide change between the codons. First, if codons a and b are not nearest-neighbours, $\mu_{a\mapsto b}$ is equal to 0. Second, if codons a and b are only one mutation away, $\mu_{a\mapsto b}$ is given by the underlying nucleotide relative rate ($R_{a\mapsto b}$).

For a given site i, the codon substitution rate matrix $\boldsymbol{q}^{(\mathrm{i})}$ is given by:

$$\begin{cases} q^{(\mathrm{i})}_{a\mapsto b} = 0 \text{ if codons } a \text{ and } b \text{ are not nearest-neighbors,} \\ q^{(\mathrm{i})}_{a\mapsto b} = \mu_{a\mapsto b} \text{ if codons } a \text{ and } b \text{ are synonymous,} \\ q^{(\mathrm{i})}_{a\mapsto b} = \mu_{a\mapsto b} \dfrac{F^{(\mathrm{i})}_b - F^{(\mathrm{i})}_a}{1 - \mathrm{e}^{F^{(\mathrm{i})}_a - F^{(\mathrm{i})}_b}} \text{ if } a \text{ and } b \text{ are non-synonymous,} \\ q^{(\mathrm{i})}_{a,a} = - \sum\limits_{b\neq a, b=1}^{61} q^{(\mathrm{i})}_{a\mapsto b}. \end{cases} \tag{S.13}$$

Together, the probability of transition between codons for a given branch j and site i is:

$$\boldsymbol{P}^{(\mathrm{j,i})} = \mathrm{e}^{l^{(\mathrm{j})}\boldsymbol{q}^{(\mathrm{i})}}, \tag{S.14}$$

which are the matrices necessary to compute the likelihood of the data (D) given the parameters of the model using the pruning algorithm.

1.6 Bayesian implementation

Bayesian inference was conducted using Markov Chain Monte Carlo (MCMC). Most phylogenetic MCMC samplers target the distribution over the model parameters given the sequence alignment, which means that they have to repeatedly invoke the pruning algorithm to recalculate the likelihood which is most often the limiting step of the MCMC. An alternative, which is used here, is to do the MCMC conditionally on the detailed substitution history $\mathcal{H}$, thus doing the MCMC over the augmented configuration ($\mathcal{H}$, D), under the target distribution obtained by combining the mapping-based likelihood with the prior over model parameters. The key idea that makes this strategy efficient is that the mapping-based likelihood depends on compact summary statistics of $\mathcal{H}$, leading to very fast evaluation of the likelihood. On the other hand, this requires to implement more complex MCMC procedures that have to alternate between:

1. sampling $\mathcal{H}$ conditionally on the data and the current parameter configuration.

2. re-sampling the parameters conditionally on $\mathcal{H}$.

To implement the mapping-based MCMC sampling strategy, we first sample the detailed substitution history $\mathcal{H}$ for all sites along the tree. Several methods exist for doing this [2, 3], which are used here in combination (first trying the accept-reject method of Nielsen, then switching to the uniformization approach of Rodrigue *et al* if the first round has failed).

Then, we write down the probability of $\mathcal{H}$ given the parameters, and finally, we collect all factors that depend on some parameter of interest and make some simplifications. This ultimately leads to relatively compact sufficient statistics allowing for fast numerical evaluation of the likelihood [4, 5].

1.7 BayesCode software

In *BayesCode* (github.com/ThibaultLatrille/BayesCode, v1.3.1), we ran the mutation-selection codon models *mutselomega* for 2000 points of MCMC with the options:

```
mutselomega ---omegashift 0.0 --ncat 30 -a my_alignment.phy -t my_tree.newick -u 2000 my_genename
```

The collection of site-specific fitness profiles ($\boldsymbol{F^{(i)}}, \forall i$) are then obtained by running *readmutselomega*, reading 1000 points of MCMC (first 1000 are considered as burn-in) with the options:

```
readmutselomega --every 1 --until 2000 --burnin 1000 --ss my_genename
```

The gene-specific 4×4 nucleotide mutation rate matrix ($\boldsymbol{\mu}$) is also obtained by running *readmutselomega*, reading 1000 points of MCMC (first 1000 are considered as burn-in) with the options:

```
readmutselomega --every 1 --until 2000 --burnin 1000 --nuc my_genename
```

2 Beneficial mutations in the terminal lineages and populations

2.1 Summary tables

2.1.1 Probability of mutations and substitutions to be $\mathcal{D}_0$, $\mathcal{N}_0$ or $\mathcal{B}_0$

Table S1: Probability of mutations and substitutions to be $\mathcal{D}_0$, $\mathcal{N}_0$ or $\mathcal{B}_0$.

Population	Species	$\mathbb{P}[\mathcal{D}_0]$	$\mathbb{P}[\mathcal{N}_0]$	$\mathbb{P}[\mathcal{B}_0]$	$\mathbb{P}_{div}[\mathcal{D}_0]$	$\mathbb{P}_{div}[\mathcal{N}_0]$	$\mathbb{P}_{div}[\mathcal{B}_0]$
Equus c.	Equus caballus	0.923	0.065	0.012	0.462	0.419	0.118
Iran	Bos taurus	0.924	0.065	0.011	0.515	0.362	0.123
Uganda	Bos taurus	0.924	0.065	0.011	0.514	0.361	0.125
Australia	Capra hircus	0.923	0.066	0.011	0.494	0.386	0.121
France	Capra hircus	0.923	0.066	0.011	0.494	0.386	0.120
Iran (C. aegagrus)	Capra hircus	0.923	0.066	0.011	0.493	0.386	0.120
Iran	Capra hircus	0.923	0.066	0.011	0.492	0.387	0.121
Italy	Capra hircus	0.923	0.066	0.011	0.494	0.386	0.120
Morocco	Capra hircus	0.923	0.066	0.011	0.491	0.387	0.122
Iran	Ovis aries	0.922	0.067	0.012	0.568	0.323	0.109
Iran (O. orientalis)	Ovis aries	0.922	0.067	0.011	0.573	0.320	0.108
Iran (O. vignei)	Ovis aries	0.922	0.067	0.012	0.567	0.325	0.109
Various	Ovis aries	0.922	0.067	0.011	0.572	0.321	0.107
Morocco	Ovis aries	0.922	0.067	0.012	0.570	0.321	0.108
Barbados	Chlorocebus sabaeus	0.926	0.065	0.009	0.485	0.393	0.122
Central Afr. Rep.	Chlorocebus sabaeus	0.926	0.065	0.009	0.485	0.391	0.124
Ethiopia	Chlorocebus sabaeus	0.926	0.065	0.009	0.484	0.393	0.124
Gambia	Chlorocebus sabaeus	0.926	0.065	0.009	0.483	0.394	0.123
Kenya	Chlorocebus sabaeus	0.926	0.065	0.009	0.485	0.392	0.123
Nevis	Chlorocebus sabaeus	0.926	0.065	0.009	0.484	0.393	0.123
South Africa	Chlorocebus sabaeus	0.926	0.065	0.009	0.480	0.394	0.125
Saint Kitts	Chlorocebus sabaeus	0.926	0.065	0.009	0.483	0.394	0.123
Zambia	Chlorocebus sabaeus	0.926	0.065	0.009	0.485	0.393	0.123
African	Homo sapiens	0.925	0.065	0.010	0.561	0.341	0.099
Admixed American	Homo sapiens	0.925	0.065	0.010	0.561	0.340	0.099
East Asian	Homo sapiens	0.925	0.065	0.010	0.560	0.341	0.098
European	Homo sapiens	0.925	0.065	0.010	0.562	0.340	0.098
South Asian	Homo sapiens	0.925	0.065	0.010	0.561	0.341	0.099

- $\mathbb{P}[\mathcal{D}_0]$ (eq. 5) is the probability for a new mutation to be a deleterious. These mutations have a selection coefficient predicted at the phylogenetic-scale lower than -1, thus toward a less fit amino-acid.
- $\mathbb{P}[\mathcal{N}_0]$ (eq. 5) is the probability for a new mutation to be a nearly-neutral. These mutations have a selection coefficient predicted at the phylogenetic-scale between -1 and 1.
- $\mathbb{P}[\mathcal{B}_0]$ (eq. 5) is the probability for a new mutation to be a beneficial back-mutation. These mutations have a selection coefficient predicted at the phylogenetic-scale larger than 1, thus toward a more fit amino-acid.
- $\mathbb{P}_{div}[\mathcal{D}_0]$ is the proportion of substitutions in the terminal branch that are $\mathcal{D}_0$.
- $\mathbb{P}_{div}[\mathcal{N}_0]$ is the proportion of substitutions in the terminal branch that are $\mathcal{N}_0$.
- $\mathbb{P}_{div}[\mathcal{B}_0]$ is the proportion of substitutions in the terminal branch that are $\mathcal{B}_0$.

2.1.2 d_N/d_S for $\mathcal{D}_0$, $\mathcal{N}_0$ or $\mathcal{B}_0$

Table S2: d_N/d_S for $\mathcal{D}_0$, $\mathcal{N}_0$ or $\mathcal{B}_0$.

Population	Species	d_N/d_S	$d_N(\mathcal{D}_0)/d_S$	$d_N(\mathcal{N}_0)/d_S$	$d_N(\mathcal{B}_0)/d_S$
Equus c.	Equus caballus	0.129	0.065	0.832	1.267
Iran	Bos taurus	0.114	0.063	0.629	1.280
Uganda	Bos taurus	0.116	0.064	0.638	1.328
Australia	Capra hircus	0.108	0.058	0.631	1.169
France	Capra hircus	0.109	0.058	0.635	1.172
Iran (C. aegagrus)	Capra hircus	0.108	0.058	0.632	1.170
Iran	Capra hircus	0.108	0.058	0.632	1.178
Italy	Capra hircus	0.109	0.058	0.636	1.171
Morocco	Capra hircus	0.108	0.058	0.633	1.183
Iran	Ovis aries	0.128	0.079	0.619	1.215
Iran (O. orientalis)	Ovis aries	0.129	0.080	0.619	1.212
Iran (O. vignei)	Ovis aries	0.127	0.078	0.619	1.201
Various	Ovis aries	0.129	0.080	0.620	1.201
Morocco	Ovis aries	0.129	0.080	0.621	1.217
Barbados	Chlorocebus sabaeus	0.118	0.062	0.713	1.521
Central Afr. Rep.	Chlorocebus sabaeus	0.119	0.062	0.717	1.559
Ethiopia	Chlorocebus sabaeus	0.118	0.062	0.716	1.549
Gambia	Chlorocebus sabaeus	0.119	0.062	0.719	1.537
Kenya	Chlorocebus sabaeus	0.119	0.062	0.716	1.546
Nevis	Chlorocebus sabaeus	0.118	0.062	0.715	1.532
South Africa	Chlorocebus sabaeus	0.119	0.062	0.720	1.577
Saint Kitts	Chlorocebus sabaeus	0.118	0.062	0.715	1.539
Zambia	Chlorocebus sabaeus	0.118	0.062	0.716	1.537
African	Homo sapiens	0.170	0.103	0.885	1.744
Admixed American	Homo sapiens	0.170	0.103	0.884	1.742
East Asian	Homo sapiens	0.170	0.103	0.888	1.742
European	Homo sapiens	0.170	0.103	0.884	1.733
South Asian	Homo sapiens	0.170	0.103	0.886	1.745

- d_N/d_S (eq. 6) is the ratio of non-synonymous over synonymous substitutions estimated for all the non-synonymous substitutions in the terminal branch.
- $d_N(\mathcal{D}_0)/d_S$ (eq. 6) is the ratio of non-synonymous over synonymous substitutions, when restricted to non-synonymous substitutions in the terminal branch that are $\mathcal{D}_0$.
- $d_N(\mathcal{N}_0)/d_S$ (eq. 6) is the ratio of non-synonymous over synonymous substitutions, when restricted to non-synonymous substitutions in the terminal branch that are $\mathcal{N}_0$.
- $d_N(\mathcal{B}_0)/d_S$ (eq. 6) is the ratio of non-synonymous over synonymous substitutions, when restricted to non-synonymous substitutions in the terminal branch that are $\mathcal{B}_0$.

2.1.3 d_N/d_S over-estimation due to beneficial back-mutations

Table S3: d_N/d_S over-estimation due to beneficial back-mutations.

Population	Species	d_N/d_S	$d_N(S_0 < 1)/d_S$	$\delta(d_N/d_S)$
Equus c.	Equus caballus	0.129	0.115	10.7
Iran	Bos taurus	0.114	0.101	11.3
Uganda	Bos taurus	0.116	0.102	11.5
Australia	Capra hircus	0.108	0.096	11.1
France	Capra hircus	0.109	0.097	11.0
Iran (C. aegagrus)	Capra hircus	0.108	0.096	11.1
Iran	Capra hircus	0.108	0.096	11.2
Italy	Capra hircus	0.109	0.097	11.0
Morocco	Capra hircus	0.108	0.096	11.2
Iran	Ovis aries	0.128	0.115	9.894
Iran (O. orientalis)	Ovis aries	0.129	0.117	9.740
Iran (O. vignei)	Ovis aries	0.127	0.115	9.822
Various	Ovis aries	0.129	0.116	9.663
Morocco	Ovis aries	0.129	0.116	9.805
Barbados	Chlorocebus sabaeus	0.118	0.104	11.4
Central Afr. Rep.	Chlorocebus sabaeus	0.119	0.105	11.5
Ethiopia	Chlorocebus sabaeus	0.118	0.105	11.5
Gambia	Chlorocebus sabaeus	0.119	0.105	11.4
Kenya	Chlorocebus sabaeus	0.119	0.105	11.5
Nevis	Chlorocebus sabaeus	0.118	0.105	11.4
South Africa	Chlorocebus sabaeus	0.119	0.105	11.7
Saint Kitts	Chlorocebus sabaeus	0.118	0.104	11.5
Zambia	Chlorocebus sabaeus	0.118	0.105	11.4
African	Homo sapiens	0.170	0.154	8.996
Admixed American	Homo sapiens	0.170	0.155	8.978
East Asian	Homo sapiens	0.170	0.155	8.966
European	Homo sapiens	0.170	0.155	8.926
South Asian	Homo sapiens	0.170	0.155	8.986

- d_N/d_S (eq. 6) is the ratio of non-synonymous over synonymous substitutions estimated for all the non-synonymous substitutions in the terminal branch.
- $d_N(S_0 < 1)/d_S$ (eq. 6) is the ratio of non-synonymous over synonymous substitutions, when restricted to non-synonymous substitutions in the terminal branch that are not $\mathcal{B}_0$. This is the estimated divergence when we removed beneficial back-mutations.
- $\delta(d_N/d_S)$ (eq. 7) is the fraction of the divergence (d_N/d_S) that is over-estimated, d_N/d_S is compared to the estimated divergence when we removed beneficial back-mutations $d_N(S_0 < 1)/d_S$.

We estimated that between 9 and 11% of d_N/d_S is over-estimated, corresponding to beneficial back-mutation inflating the d_N/d_S statistic.

3 Clinically related terms for mutations

3.1 Terms associated with deleterious mutations $\mathcal{D}_0$

Table S4: Terms associated with deleterious mutations $\mathcal{D}_0$

SNP clinical ontology	n_{Observed}	n_{Expected}	Odds ratio	p_{v}	$p_{\text{v-adjusted}}$
Benign	2969	4043.0	0.734	1.000	1.000
Likely benign	2994	3399.8	0.881	0.999	1.000
Risk factor	102	118.2	0.863	0.798	1.000
Likely pathogenic	221	68.5	3.226	1.7×10^{-8}	$\mathbf{6.7 \times 10^{-8}}$*
Pathogenic	560	193.6	2.893	4.2×10^{-17}	$\mathbf{2.1 \times 10^{-16}}$*

In humans, non-synonymous SNPs in the test group ($\mathcal{D}_0$) are contrasted to SNPs in the control group ($\mathcal{N}_0$). For each clinical term, a 2x2 contingency tables is built by counting the number of SNPs based on their selection coefficient and their clinical terms (whether they have this specific term or not). Fisher's exact tests are then performed for these 2x2 contingency tables. * for $p_{\text{v}}^{\text{adj}}$ corrected for multiple comparison (Holm–Bonferroni correction) lower than the risk $\alpha = 0.05$. SNPs predicted with $\mathcal{D}_0$ are statistically associated to clinical terms such as *Likely Pathogenic* and *Pathogenic*.

3.2 Terms associated with beneficial back-mutations $\mathcal{B}_0$

Table S5: Terms associated with beneficial back-mutations $\mathcal{B}_0$

SNP clinical ontology	n_{Observed}	n_{Expected}	Odds ratio	p_{v}	$p_{\text{v-adjusted}}$
Benign	319	261.7	1.219	0.002	**0.009***
Likely benign	263	222.7	1.181	0.012	**0.049***
Risk factor	5	7.847	0.637	0.879	0.879
Likely pathogenic	7	4.552	1.538	0.227	0.682
Pathogenic	16	12.9	1.241	0.268	0.682

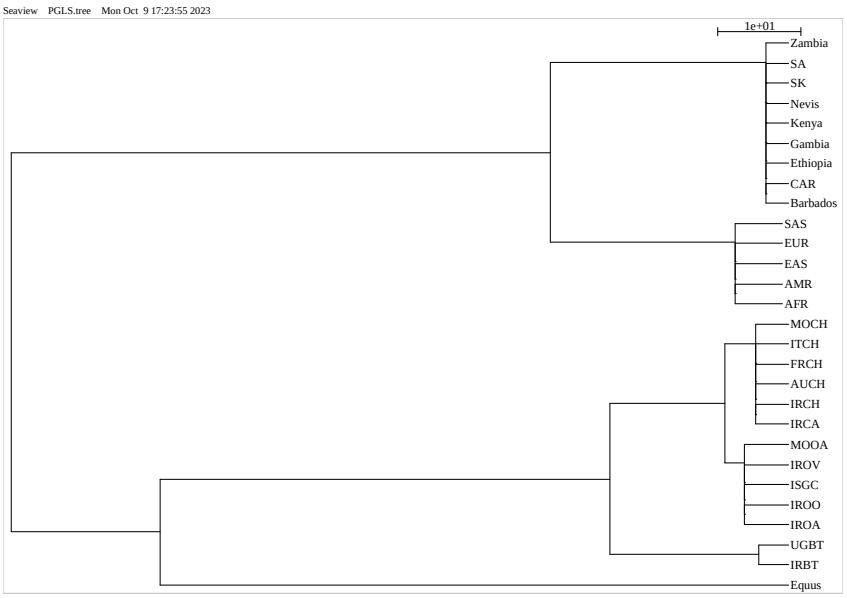
In humans, non-synonymous SNPs in the test group ($\mathcal{B}_0$) are contrasted to SNPs in the control group ($\mathcal{N}_0$). For each clinical term, a 2x2 contingency tables is built by counting the number of SNPs based on their selection coefficient and their clinical terms (whether they have this specific term or not). Fisher's exact tests are then performed for these 2x2 contingency tables. * for $p_{\text{v}}^{\text{adj}}$ corrected for multiple comparison (Holm–Bonferroni correction) lower than the risk $\alpha = 0.05$. Beneficial back-mutations are associated with clinical terms such as *Benign* and *Likely Benign*.

4 Correlation with diversity

4.1 Phylogenetic Generalized Linear Model

Because a correlation must account for phylogenetic relationship and non-independence of samples, we fitted a Phylogenetic Generalized Linear Model (PGLM) in R with the package caper[6], with multi-furcation of the different populations inside each species. The mammalian tree imported from TimeTree[7] and pruned to the species used in this study. Multifurcations of the different populations are placed at the same divergence time as the species.

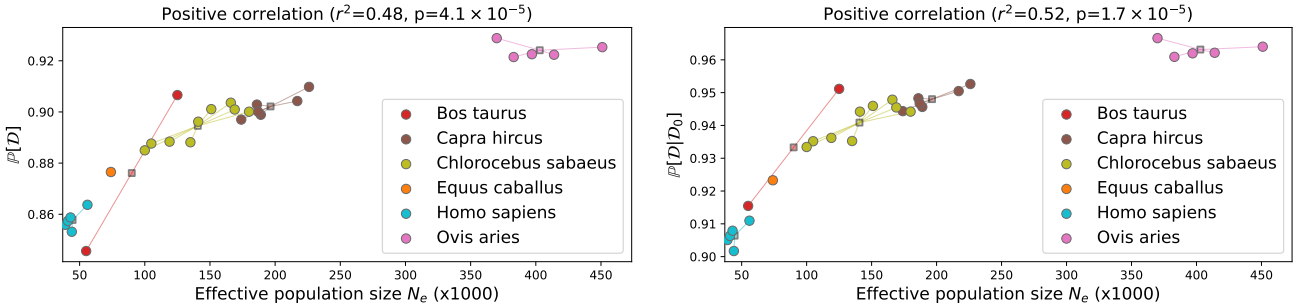
```
((Equus:76,((IRBT:3.55912,UGBT:3.55912)1993:18.0665,((IROA:5.33074,IROO:5.33074,ISGC
    :5.33074,IROV:5.33074,MOOA:5.33074)2219:2.32334,(IRCA:3.96087,IRCH:3.96087,AUCH
    :3.96087,FRCH:3.96087,ITCH:3.96087,MOCH:3.96087)2201:3.69321)2218:13.9716)
    2262:54.3744)2405:18,((AFR:5.59333,AMR:5.59333,EAS:5.59333,EUR:5.59333,SAS
    :5.59333)8454:22.42,(Barbados:2.69204,CAR:2.69204,Ethiopia:2.69204,Gambia:2.69204,
    Kenya:2.69204,Nevis:2.69204,SK:2.69204,SA:2.69204,Zambia:2.69204)8449:26.128)
    8800:65.18);
```



Then, for each population, the proportion of deleterious ($\mathbb{P}[\mathcal{D}]$), nearly-neutral ($\mathbb{P}[\mathcal{N}]$) and of beneficial ($\mathbb{P}[\mathcal{B}]$) mutations estimated at the population-genetic scale is shown as function of N_e. r^2 and p-value are obtained from the PGLM model.

4.1.1 Proportion of deleterious mutations ($\mathcal{D}$)

Figure S1: Proportion of deleterious mutations ($\mathcal{D}$)

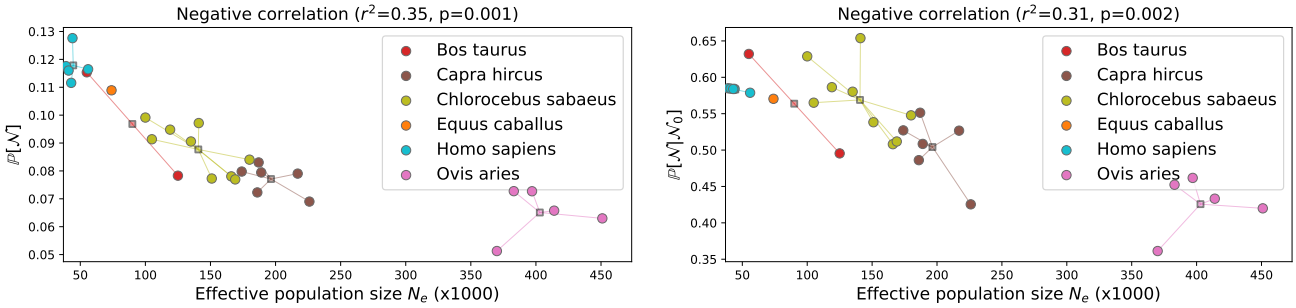


- N_e is the estimated effective population size.
- $\mathbb{P}[\mathcal{D}]$ (eq. 15) is the probability for a mutation to be deleterious. These mutations have a selection coefficient at the population-scale lower than -1.
- $\mathbb{P}[\mathcal{D} \mid \mathcal{D}_0]$ (eq. 12) is the probability for a mutation to be deleterious at the population scale, given it is predicted to be a deleterious at the phylogenetic scale.

We can see that higher effective population size (N_e) is typically accompanied by a higher proportion of effectively deleterious mutations at the population scale ($\mathbb{P}[\mathcal{D}]$). This trend is also confirmed when we restricted the analysis to class of mutations that are supposedly deleterious at the phylogenetic scale ($\mathcal{D}_0$). This result is qualitatively in accordance with the nearly-neutral theory of evolution which argues that very slightly deleterious mutations are more efficiently purified in large populations.

4.1.2 Proportion of nearly-neutral mutations ($\mathcal{N}$)

Figure S2: Proportion of nearly-neutral mutations ($\mathcal{N}$)

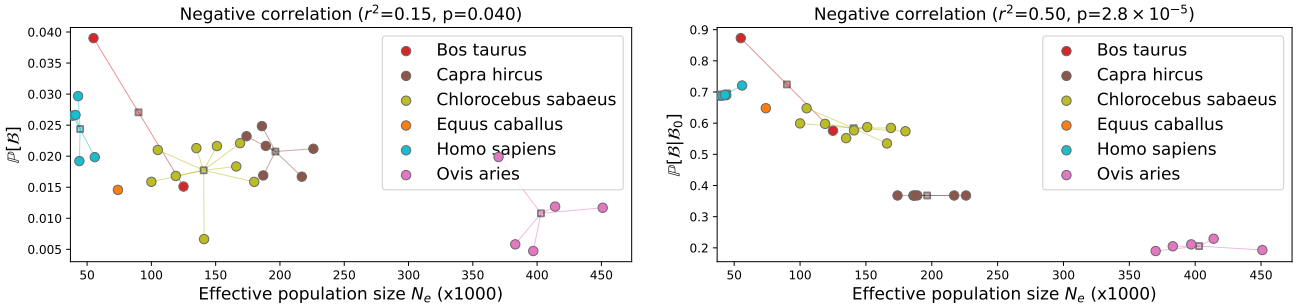


- N_e is the estimated effective population size.
- $\mathbb{P}[\mathcal{N}]$ (eq. 15) is the probability for a mutation to be nearly-neutral. These mutations have a selection coefficient at the population-scale between -1 and 1.
- $\mathbb{P}[\mathcal{N} \mid \mathcal{N}_0]$ (eq. 12) is the probability for a mutation to be nearly-neutral at the population scale, given it is predicted to be a nearly-neutral at the phylogenetic scale.

We confirmed that higher effective population size (N_e) is typically accompanied by a smaller proportion of neutral mutations at the population scale ($\mathbb{P}[\mathcal{N}]$ in the range 0.06-0.18). This result is more pronounced ($\mathbb{P}[\mathcal{N} \mid \mathcal{N}_0]$ in the range 0.36-0.73) when we restricted the analysis to class of mutations that are supposedly nearly-neutral at the phylogenetic scale ($\mathcal{N}_0$). This result suggests that populations with higher diversity (e.g. *Bos* or *Ovis*) are more likely to discriminate whether mutations are beneficial or deleterious. Alternatively stated, mutations in populations with low diversity (e.g. *Homo*) are effectively nearly-neutral and behave as would a neutral mutation. This result is qualitatively in accordance with the nearly-neutral theory of evolution which argues that mutations are less efficiently selected for in small populations.

4.1.3 Proportion of beneficial mutations ($\mathcal{B}$)

Figure S3: Proportion of beneficial mutations ($\mathcal{B}$)



- N_e is the estimated effective population size.
- $\mathbb{P}[\mathcal{B}]$ (eq. 15) is the probability for a mutation to be beneficial. These mutations have a selection coefficient at the population-scale larger than 1.
- $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$ (eq. 12) is the probability for a mutation to be beneficial at the population scale, given it is predicted to be a beneficial back-mutation at the phylogenetic scale (the *precision*).

Higher effective population size (N_e) is accompanied by a smaller proportion of beneficial mutations at the population scale ($\mathbb{P}[\mathcal{B}]$). This trend is also confirmed when we restricted the analysis to class of mutations that are supposedly beneficial back-mutations at the phylogenetic scale ($\mathcal{B}_0$). This result is much more difficult to interpret. The fraction of beneficial back mutations is supposed to depend on long term demographic history, which is not directly accessible. If the long term effective population size is relatively similar to the short term one, we expect little opportunity for beneficial back-mutations. And thus, counter-intuitively we expect a diminution of positively selected mutations. However, we can see that the proportion of beneficial back-mutations among advantageous one does not decrease with N_e. This means that somehow adaptive mutations also decrease with N_e for a reason we fail to explain.

5 Probabilities of beneficial back-mutations among all beneficial ones ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$)

Table S6: Probability of beneficial back-mutations among all beneficial ones ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$).

Population	Species	N_e	$\mathbb{P}[\mathcal{B}_0]$	$\mathbb{P}[\mathcal{B}]$	$\frac{\mathbb{P}[\mathcal{B}_0]}{\mathbb{P}[\mathcal{B}]}$	$\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$	$\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$
Equus c.	Equus caballus	7.5×10^4	0.012	0.015	0.827	0.648	0.536
Iran	Bos taurus	5.6×10^4	0.011	0.039	0.279	0.873	0.243
Uganda	Bos taurus	1.3×10^5	0.011	0.015	0.720	0.576	0.415
Australia	Capra hircus	1.7×10^5	0.011	0.023	0.480	0.368	0.177
France	Capra hircus	1.9×10^5	0.011	0.022	0.515	0.368	0.190
Iran (C. aegagrus)	Capra hircus	1.9×10^5	0.011	0.025	0.448	0.368	0.165
Iran	Capra hircus	2.3×10^5	0.011	0.021	0.525	0.368	0.193
Italy	Capra hircus	1.9×10^5	0.011	0.017	0.660	0.368	0.243
Morocco	Capra hircus	2.2×10^5	0.011	0.017	0.667	0.368	0.245
Iran	Ovis aries	3.8×10^5	0.012	0.006	1.984	0.205	0.407
Iran (O. orientalis)	Ovis aries	4.5×10^5	0.011	0.012	0.983	0.193	0.190
Iran (O. vignei)	Ovis aries	3.7×10^5	0.012	0.020	0.579	0.190	0.110
Various	Ovis aries	4.1×10^5	0.011	0.012	0.967	0.229	0.222
Morocco	Ovis aries	4×10^5	0.012	0.005	2.435	0.211	0.514
Barbados	Chlorocebus sabaeus	1.1×10^5	0.009	0.021	0.452	0.648	0.293
Central Afr. Rep.	Chlorocebus sabaeus	1.7×10^5	0.009	0.018	0.515	0.535	0.275
Ethiopia	Chlorocebus sabaeus	1.4×10^5	0.009	0.021	0.444	0.552	0.245
Gambia	Chlorocebus sabaeus	1.4×10^5	0.009	0.007	1.423	0.577	0.821
Kenya	Chlorocebus sabaeus	1.5×10^5	0.009	0.022	0.437	0.588	0.257
Nevis	Chlorocebus sabaeus	1×10^5	0.009	0.016	0.597	0.599	0.358
South Africa	Chlorocebus sabaeus	1.8×10^5	0.009	0.016	0.594	0.574	0.341
Saint Kitts	Chlorocebus sabaeus	1.2×10^5	0.009	0.017	0.563	0.598	0.336
Zambia	Chlorocebus sabaeus	1.7×10^5	0.009	0.022	0.427	0.585	0.250
African	Homo sapiens	5.6×10^4	0.010	0.020	0.484	0.721	0.349
Admixed American	Homo sapiens	4.5×10^4	0.010	0.019	0.500	0.690	0.345
East Asian	Homo sapiens	4×10^4	0.010	0.027	0.362	0.688	0.249
European	Homo sapiens	4.2×10^4	0.010	0.027	0.361	0.688	0.248
South Asian	Homo sapiens	4.4×10^4	0.010	0.030	0.324	0.691	0.224

- N_e is the estimated effective population size.
- $\mathbb{P}[\mathcal{B}_0]$ (eq. 5) is the probability for a new mutation to be a beneficial back-mutation. These mutations have a selection coefficient predicted at the phylogenetic-scale larger than 1, thus toward a more fit amino-acid.
- $\mathbb{P}[\mathcal{B}]$ (eq. 15) is the probability for a mutation to be beneficial. These mutations have a selection coefficient at the population-scale larger than 1.
- $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$ (eq. 12) is the probability for a mutation to be beneficial at the population scale, given it is predicted to be a beneficial back-mutation at the phylogenetic scale (the *precision*).
- $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ (eq. 14) is the probability for a mutation to a beneficial back-mutations, given it is beneficial at the population scale (the *recall*). This probability is obtained using Bayes' formula.

6 Excluding genes under adaptation

Table S7: $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ for each population when excluding or not genes under adaptation.

Population	Species	Control	Case
Equus c.	Equus caballus	0.536	0.880
Iran	Bos taurus	0.243	0.249
Uganda	Bos taurus	0.415	0.429
Australia	Capra hircus	0.177	0.190
France	Capra hircus	0.190	0.201
Iran (C. aegagrus)	Capra hircus	0.165	0.169
Iran	Capra hircus	0.193	0.176
Italy	Capra hircus	0.243	0.261
Morocco	Capra hircus	0.245	0.283
Iran	Ovis aries	0.407	0.454
Iran (O. orientalis)	Ovis aries	0.190	0.207
Iran (O. vignei)	Ovis aries	0.110	0.135
Various	Ovis aries	0.222	0.246
Morocco	Ovis aries	0.514	0.627
Barbados	Chlorocebus sabaeus	0.293	0.325
Central Afr. Rep.	Chlorocebus sabaeus	0.275	0.273
Ethiopia	Chlorocebus sabaeus	0.245	0.242
Gambia	Chlorocebus sabaeus	0.821	0.873
Kenya	Chlorocebus sabaeus	0.257	0.254
Nevis	Chlorocebus sabaeus	0.358	0.395
South Africa	Chlorocebus sabaeus	0.341	0.375
Saint Kitts	Chlorocebus sabaeus	0.336	0.355
Zambia	Chlorocebus sabaeus	0.250	0.263
African	Homo sapiens	0.349	0.363
Admixed American	Homo sapiens	0.345	0.312
East Asian	Homo sapiens	0.249	0.233
European	Homo sapiens	0.248	0.211
South Asian	Homo sapiens	0.224	0.218

Genes under adaptation retrieved from [8]. Comparison of $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ for the whole genome (control) and when excluding genes under adaptation (case). The non-parametric Wilcoxon signed-rank test tests the null hypothesis that the distribution of the differences (case versus control) is symmetric about zero. Applied to the paired samples (case and control), Wilcoxon signed-rank test results in $s = 80$ with p-value $= 0.002$. for one-sided test (case higher than control). The proportion of beneficial back-mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$) is higher when excluding genes under adaptation, consistent with our expectation that genes with uniformly conserved functions should fit better the back-mutation equilibrium model.

7 Replicability across populations

The proportion of deleterious ($\mathbb{P}[\mathcal{D}]$), nearly-neutral ($\mathbb{P}[\mathcal{N}]$) and of beneficial ($\mathbb{P}[\mathcal{B}]$) mutations estimated at the population-genetic scale across the different populations is shown for each class of selection ($x \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}$).

Figure S4: Estimation of selection at the population scale for $\mathcal{D}_0$ mutations

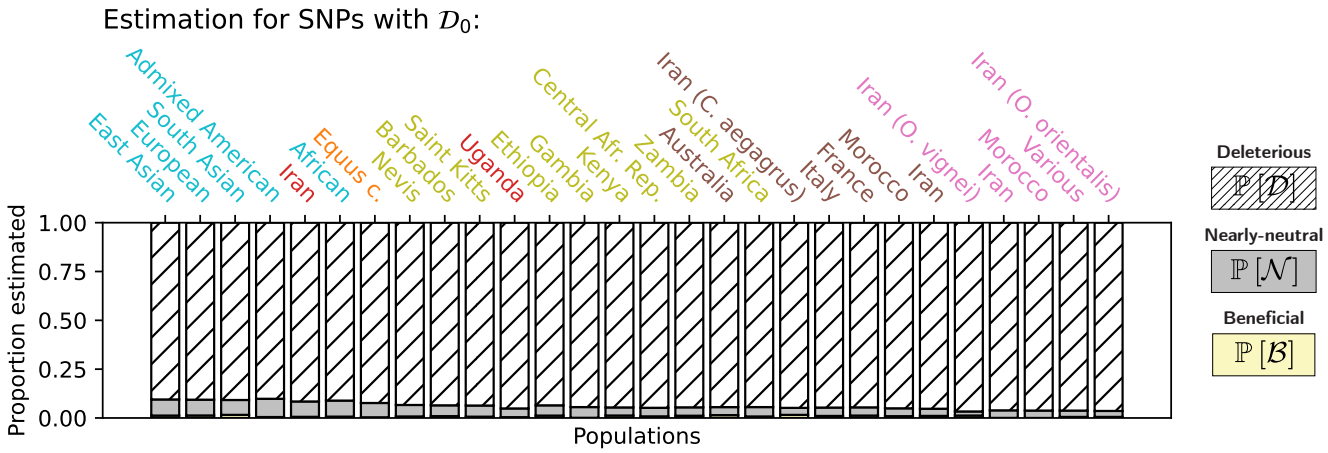


Figure S5: Estimation of selection at the population scale for $\mathcal{N}_0$ mutations

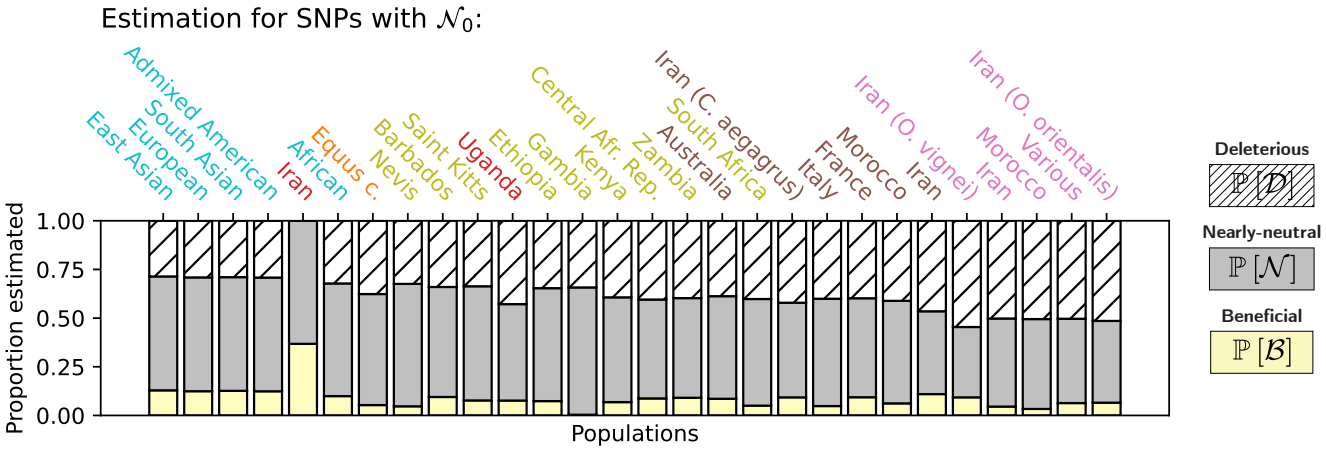
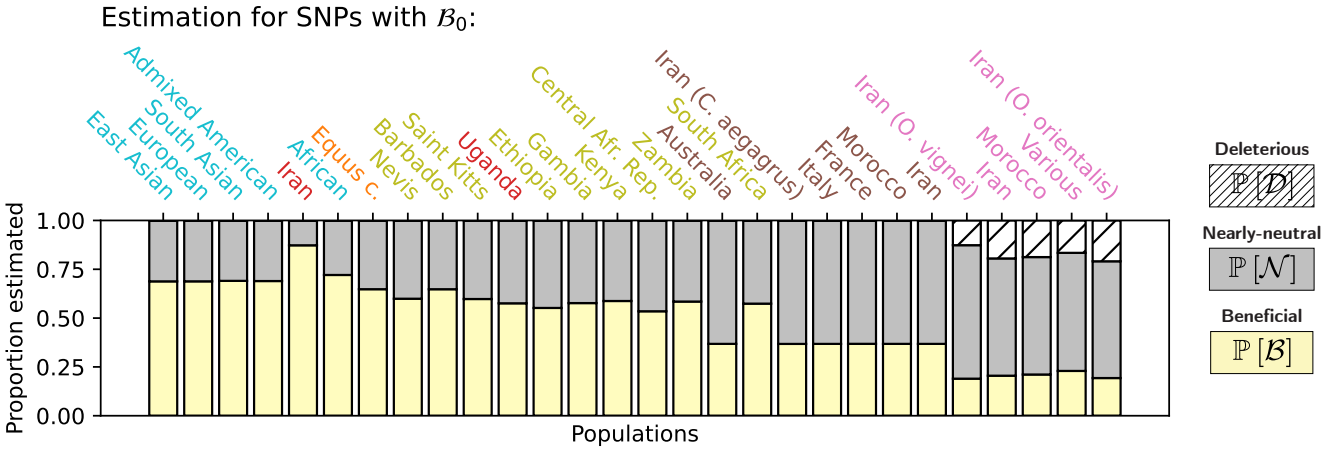


Figure S6: Estimation of selection at the population scale for $\mathcal{B}_0$ mutations



8 Including divergence data for the estimation of S

8.1 Including divergence in polyDFE

Divergence data (number of substitutions per site) can also be included into polyDFE to estimate the DFE. For the class of a given class of selection coefficient ($x \in \{\mathcal{D}_0, \mathcal{N}_0, \mathcal{B}_0\}$), the number of substitutions has already been computed and is given by $D(x)$ (see eq. 6), and the number of sites is given by $L(x)$ (see eq. 4). Otherwise, the procedure is the same as described section *Scaled selection coefficients (S) in a population-based method.*

8.2 Probabilities of beneficial back-mutations

Table S8: Probability of beneficial back-mutations among all beneficial ones - including divergence.

Population	Species	N_e	$\mathbb{P}[\mathcal{B}_0]$	$\mathbb{P}[\mathcal{B}]$	$\frac{\mathbb{P}[\mathcal{B}_0]}{\mathbb{P}[\mathcal{B}]}$	$\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$	$\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$
Equus c.	Equus caballus	7.5×10^4	0.012	0.057	0.213	0.498	0.106
Iran	Bos taurus	5.6×10^4	0.011	0.006	1.860	0.368	0.684
Uganda	Bos taurus	1.3×10^5	0.011	0.012	0.900	0.307	0.276
Australia	Capra hircus	1.7×10^5	0.011	0.007	1.651	0.178	0.294
France	Capra hircus	1.9×10^5	0.011	0.007	1.614	0.183	0.295
Iran (C. aegagrus)	Capra hircus	1.9×10^5	0.011	0.012	0.923	0.177	0.163
Iran	Capra hircus	2.3×10^5	0.011	0.008	1.483	0.184	0.273
Italy	Capra hircus	1.9×10^5	0.011	0.007	1.686	0.178	0.299
Morocco	Capra hircus	2.2×10^5	0.011	0.006	1.892	0.184	0.349
Iran	Ovis aries	3.8×10^5	0.012	0.010	1.132	0.189	0.214
Iran (O. orientalis)	Ovis aries	4.5×10^5	0.011	0.011	1.012	0.213	0.216
Iran (O. vignei)	Ovis aries	3.7×10^5	0.012	0.021	0.561	0.186	0.105
Various	Ovis aries	4.1×10^5	0.011	0.011	1.069	0.215	0.229
Morocco	Ovis aries	4×10^5	0.012	0.012	0.973	0.217	0.211
Barbados	Chlorocebus sabaeus	1.1×10^5	0.009	0.004	2.120	0.341	0.723
Central Afr. Rep.	Chlorocebus sabaeus	1.7×10^5	0.009	0.013	0.720	0.368	0.265
Ethiopia	Chlorocebus sabaeus	1.4×10^5	0.009	0.007	1.352	0.368	0.498
Gambia	Chlorocebus sabaeus	1.4×10^5	0.009	0.010	0.956	0.368	0.352
Kenya	Chlorocebus sabaeus	1.5×10^5	0.009	0.013	0.727	0.368	0.268
Nevis	Chlorocebus sabaeus	1×10^5	0.009	0.004	2.133	0.368	0.784
South Africa	Chlorocebus sabaeus	1.8×10^5	0.009	0.011	0.874	0.368	0.322
Saint Kitts	Chlorocebus sabaeus	1.2×10^5	0.009	0.005	1.905	0.368	0.701
Zambia	Chlorocebus sabaeus	1.7×10^5	0.009	0.014	0.698	0.368	0.257
African	Homo sapiens	5.6×10^4	0.010	0.013	0.726	0.437	0.317
Admixed American	Homo sapiens	4.5×10^4	0.010	0.011	0.886	0.429	0.380
East Asian	Homo sapiens	4×10^4	0.010	0.008	1.213	0.426	0.517
European	Homo sapiens	4.1×10^4	0.010	0.009	1.104	0.422	0.466
South Asian	Homo sapiens	4.4×10^4	0.010	0.009	1.016	0.426	0.433

- N_e is the estimated effective population size.
- $\mathbb{P}[\mathcal{B}_0]$ (eq. 5) is the probability for a new mutation a beneficial back-mutation. These mutations have a selection coefficient predicted at the phylogenetic-scale larger than 1, thus toward a more fit amino-acid.
- $\mathbb{P}[\mathcal{B}]$ (eq. 15) is the probability for a mutation to be beneficial. These mutations have a selection coefficient at the population-scale larger than 1.
- $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$ (eq. 12) is the probability for a mutation to be beneficial at the population scale, given it is predicted to be a beneficial back-mutation at the phylogenetic scale (the *precision*).
- $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ (eq. 14) is the probability for a mutation to a beneficial back-mutations, given it is beneficial (the *recall*). This probability is obtained using Bayes' formula.

8.3 Precision and recall

Table S9: Precision and recall - including divergence

			Deleterious mutations $\mathcal{D} \coloneqq S < -1$ $\mathcal{D}_0 \coloneqq S_0 < -1$		**Nearly-neutral mutations** $\mathcal{N} \coloneqq -1 < S < 1$ $\mathcal{N}_0 \coloneqq -1 < S_0 < 1$		**Beneficial mutations** $\mathcal{B} \coloneqq S > 1$ $\mathcal{B}_0 \coloneqq S_0 > 1$	
Population	**Species**	N_e	**Precision** $\mathbb{P}[\mathcal{D} \mid \mathcal{D}_0]$	**Recall** $\mathbb{P}[\mathcal{D}_0 \mid \mathcal{D}]$	**Precision** $\mathbb{P}[\mathcal{N} \mid \mathcal{N}_0]$	**Recall** $\mathbb{P}[\mathcal{N}_0 \mid \mathcal{N}]$	**Precision** $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$	**Recall** $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$
Equus c.	Equus caballus	7.5×10^4	0.937	0.955	0.111	0.192	0.498	0.106
Iran	Bos taurus	5.6×10^4	0.935	0.970	0.565	0.356	0.368	0.684
Uganda	Bos taurus	1.3×10^5	0.951	0.967	0.516	0.422	0.307	0.276
Australia	Capra hircus	1.7×10^5	0.952	0.970	0.598	0.452	0.178	0.294
France	Capra hircus	1.9×10^5	0.951	0.969	0.574	0.433	0.183	0.295
Iran (C. aegagrus)	Capra hircus	1.9×10^5	0.957	0.966	0.515	0.460	0.177	0.163
Iran	Capra hircus	2.3×10^5	0.951	0.967	0.542	0.422	0.184	0.273
Italy	Capra hircus	1.9×10^5	0.951	0.969	0.582	0.436	0.178	0.299
Morocco	Capra hircus	2.2×10^5	0.949	0.968	0.556	0.412	0.184	0.349
Iran	Ovis aries	3.8×10^5	0.963	0.962	0.421	0.424	0.189	0.214
Iran (O. orientalis)	Ovis aries	4.5×10^5	0.968	0.961	0.413	0.460	0.213	0.216
Iran (O. vignei)	Ovis aries	3.7×10^5	0.972	0.959	0.360	0.528	0.186	0.105
Various	Ovis aries	4.1×10^5	0.966	0.961	0.430	0.454	0.215	0.229
Morocco	Ovis aries	4×10^5	0.966	0.958	0.372	0.416	0.217	0.211
Barbados	Chlorocebus sabaeus	1.1×10^5	0.940	0.976	0.654	0.409	0.341	0.723
Central Afr. Rep.	Chlorocebus sabaeus	1.7×10^5	0.950	0.971	0.526	0.421	0.368	0.265
Ethiopia	Chlorocebus sabaeus	1.4×10^5	0.942	0.974	0.609	0.402	0.368	0.498
Gambia	Chlorocebus sabaeus	1.4×10^5	0.950	0.974	0.611	0.452	0.368	0.352
Kenya	Chlorocebus sabaeus	1.5×10^5	0.950	0.972	0.545	0.434	0.368	0.268
Nevis	Chlorocebus sabaeus	1×10^5	0.940	0.977	0.668	0.412	0.368	0.784
South Africa	Chlorocebus sabaeus	1.8×10^5	0.947	0.972	0.550	0.411	0.368	0.322
Saint Kitts	Chlorocebus sabaeus	1.2×10^5	0.940	0.976	0.646	0.407	0.368	0.701
Zambia	Chlorocebus sabaeus	1.7×10^5	0.950	0.971	0.527	0.426	0.368	0.257
African	Homo sapiens	5.6×10^4	0.911	0.980	0.666	0.341	0.437	0.317
Admixed American	Homo sapiens	4.5×10^4	0.902	0.976	0.580	0.284	0.429	0.380
East Asian	Homo sapiens	4×10^4	0.904	0.984	0.744	0.341	0.426	0.517
European	Homo sapiens	4.1×10^4	0.906	0.984	0.737	0.344	0.422	0.466
South Asian	Homo sapiens	4.4×10^4	0.907	0.984	0.737	0.349	0.426	0.433

- N_e is the estimated effective population size.
- *Precision* is the estimation of the selection coefficient at population scale (S) given that S_0 is known.
- *Recall* is the estimation of S_0 given selection coefficient at the population scale (S) is known.
- *Recall* for beneficial mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$) is thus the proportion of beneficial back-mutations among all beneficial mutations.

Altogether, comparing this table to tables 1 and S3, we acknowledge that the exact proportion of beneficial back-mutations among all beneficial ones is different whether we included or not substitutions in the terminal lineage for the estimation of S. However, we can still see that beneficial back-mutations are positively selected compared to neutral and deleterious ones, and this result is not sensitive to inclusion of substitutions in the terminal lineage.

8.4 Excluding genes under adaptation

Table S10: $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ for each population when excluding or not genes under adaptation - including divergence.

Population	Species	Control	Case
Equus c.	Equus caballus	0.106	0.106
Iran	Bos taurus	0.684	0.755
Uganda	Bos taurus	0.276	0.310
Australia	Capra hircus	0.294	0.319
France	Capra hircus	0.295	0.325
Iran (C. aegagrus)	Capra hircus	0.163	0.174
Iran	Capra hircus	0.273	0.299
Italy	Capra hircus	0.299	0.321
Morocco	Capra hircus	0.349	0.419
Iran	Ovis aries	0.214	0.241
Iran (O. orientalis)	Ovis aries	0.216	0.213
Iran (O. vignei)	Ovis aries	0.105	0.089
Various	Ovis aries	0.229	0.196
Morocco	Ovis aries	0.211	0.206
Barbados	Chlorocebus sabaeus	0.723	0.833
Central Afr. Rep.	Chlorocebus sabaeus	0.265	0.285
Ethiopia	Chlorocebus sabaeus	0.498	0.547
Gambia	Chlorocebus sabaeus	0.352	0.571
Kenya	Chlorocebus sabaeus	0.268	0.265
Nevis	Chlorocebus sabaeus	0.784	0.892
South Africa	Chlorocebus sabaeus	0.322	0.307
Saint Kitts	Chlorocebus sabaeus	0.701	0.825
Zambia	Chlorocebus sabaeus	0.257	0.271
African	Homo sapiens	0.317	0.278
Admixed American	Homo sapiens	0.380	0.363
East Asian	Homo sapiens	0.517	0.499
European	Homo sapiens	0.466	0.456
South Asian	Homo sapiens	0.433	0.304

Genes under adaptation retrieved from [8]. Comparison of $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ while including divergence for the whole genome (control) and when excluding genes under adaptation (case). The non-parametric Wilcoxon signed-rank test tests the null hypothesis that the distribution of the differences (case versus control) is symmetric about zero. Applied to the paired samples (case and control), Wilcoxon signed-rank test results in $s = 120$ with p-value $= 0.027$ for one-sided test (case higher than control). The proportion of beneficial back-mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$) is higher when excluding genes under adaptation, consistent with our expectation that genes with uniformly conserved functions should fit better the back-mutation equilibrium model.

9 Discrete distribution of S at the population scale

9.1 polyDFE model D (discrete distribution)

Additionally to including divergence, with also tested our prediction with polyDFE model D instead of model C. In polyDFE model D, the DFE of non-synonymous mutations is given as a discrete DFE of K categories (instead of a continuous distribution in model C), where the selection coefficients of each category i ($1 \leq i \leq K$) are fixed parameters S_i, and each value S_i has a probability p_i (estimated), with $\sum_{i=1}^{K} p_i = 1$. We used $K = 6$ with $S_1 = -500$, $S_2 = -4$, $S_3 = -1$, $S_4 = 0$, $S_5 = 1$, $S_6 = 4$.

For each class of selection x, the parameters p_i ($i \in \{1 \leq i \leq 6\}$) were used to compute $\mathbb{P}[\mathcal{D} \mid x]$, $\mathbb{P}[\mathcal{N} \mid x]$, and $\mathbb{P}[\mathcal{B} \mid x]$ as:

$$\mathbb{P}[\mathcal{D} \mid x] = \mathbb{P}[S < -1 \mid x] = p_1 + p_2 \tag{S.15}$$

$$\mathbb{P}[\mathcal{N} \mid x] = \mathbb{P}[-1 < S < 1 \mid x] = p_3 + p_4 + p_5, \tag{S.16}$$

$$\mathbb{P}[\mathcal{B} \mid x] = \mathbb{P}[S > 1 \mid x] = p_6. \tag{S.17}$$

9.2 Probabilities of beneficial back-mutations

Table S11: Probability of beneficial back-mutations among all beneficial ones - model D.

Population	Species	N_e	$\mathbb{P}[\mathcal{B}_0]$	$\mathbb{P}[\mathcal{B}]$	$\frac{\mathbb{P}[\mathcal{B}_0]}{\mathbb{P}[\mathcal{B}]}$	$\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$	$\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$
Equus c.	Equus caballus	7.5×10^4	0.012	0.006	2.050	0.163	0.334
Iran	Bos taurus	5.6×10^4	0.011	0.005	2.125	0.133	0.283
Uganda	Bos taurus	1.3×10^5	0.011	0.010	1.139	0.140	0.159
Australia	Capra hircus	1.7×10^5	0.011	0.002	5.637	0.151	0.850
France	Capra hircus	1.9×10^5	0.011	0.002	6.117	0.154	0.940
Iran (C. aegagrus)	Capra hircus	1.9×10^5	0.011	0.002	5.842	0.156	0.911
Iran	Capra hircus	2.3×10^5	0.011	0.002	6.400	0.147	0.943
Italy	Capra hircus	1.9×10^5	0.011	0.002	5.793	0.155	0.895
Morocco	Capra hircus	2.2×10^5	0.011	0.003	4.192	0.150	0.627
Iran	Ovis aries	3.8×10^5	0.012	0.020	0.584	0.154	0.090
Iran (O. orientalis)	Ovis aries	4.5×10^5	0.011	0.021	0.535	0.148	0.079
Iran (O. vignei)	Ovis aries	3.7×10^5	0.012	0.020	0.568	0.151	0.086
Various	Ovis aries	4.1×10^5	0.011	0.021	0.553	0.152	0.084
Morocco	Ovis aries	4×10^5	0.012	0.022	0.529	0.153	0.081
Barbados	Chlorocebus sabaeus	1.1×10^5	0.009	0.002	5.152	0.170	0.877
Central Afr. Rep.	Chlorocebus sabaeus	1.7×10^5	0.009	0.007	1.266	0.157	0.199
Ethiopia	Chlorocebus sabaeus	1.4×10^5	0.009	0.002	3.979	0.161	0.640
Gambia	Chlorocebus sabaeus	1.4×10^5	0.009	0.008	1.201	0.159	0.191
Kenya	Chlorocebus sabaeus	1.5×10^5	0.009	0.003	2.801	0.165	0.461
Nevis	Chlorocebus sabaeus	1×10^5	0.009	0.002	5.361	0.169	0.908
South Africa	Chlorocebus sabaeus	1.8×10^5	0.009	0.005	1.912	0.154	0.294
Saint Kitts	Chlorocebus sabaeus	1.2×10^5	0.009	0.002	4.561	0.165	0.753
Zambia	Chlorocebus sabaeus	1.7×10^5	0.009	0.004	2.661	0.153	0.406
African	Homo sapiens	5.6×10^4	0.010	0.010	0.997	0.154	0.154
Admixed American	Homo sapiens	4.5×10^4	0.010	0.008	1.206	0.158	0.190
East Asian	Homo sapiens	4×10^4	0.010	0.007	1.455	0.163	0.237
European	Homo sapiens	4.2×10^4	0.010	0.007	1.316	0.167	0.219
South Asian	Homo sapiens	4.4×10^4	0.010	0.007	1.340	0.164	0.219

- N_e is the estimated effective population size.
- $\mathbb{P}[\mathcal{B}_0]$ (eq. 5) is the probability for a new mutation a beneficial back-mutation. These mutations have a selection coefficient predicted at the phylogenetic-scale larger than 1, thus toward a more fit amino-acid.

- $\mathbb{P}[\mathcal{B}]$ (eq. 15) is the probability for a mutation to be beneficial. These mutations have a selection coefficient at the population-scale larger than 1.
- $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$ (eq. 12) is the probability for a mutation to be beneficial at the population scale, given it is predicted to be a beneficial back-mutation at the phylogenetic scale (the *precision*).
- $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$ (eq. 14) is the probability for a mutation to a beneficial back-mutations, given it is beneficial (the *recall*). This probability is obtained using Bayes' formula.

9.3 Precision and recall

Table S12: Precision and recall - model D

			Deleterious mutations $\mathcal{D} \coloneqq S < -1$ $\mathcal{D}_0 \coloneqq S_0 < -1$		**Nearly-neutral mutations** $\mathcal{N} \coloneqq -1 < S < 1$ $\mathcal{N}_0 \coloneqq -1 < S_0 < 1$		**Beneficial mutations** $\mathcal{B} \coloneqq S > 1$ $\mathcal{B}_0 \coloneqq S_0 > 1$	
Population	**Species**	N_e	**Precision** $\mathbb{P}[\mathcal{D} \mid \mathcal{D}_0]$	**Recall** $\mathbb{P}[\mathcal{D}_0 \mid \mathcal{D}]$	**Precision** $\mathbb{P}[\mathcal{N} \mid \mathcal{N}_0]$	**Recall** $\mathbb{P}[\mathcal{N}_0 \mid \mathcal{N}]$	**Precision** $\mathbb{P}[\mathcal{B} \mid \mathcal{B}_0]$	**Recall** $\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$
Equus c.	Equus caballus	7.5×10^4	0.911	0.974	0.645	0.319	0.163	0.334
Iran	Bos taurus	5.6×10^4	0.895	0.973	0.639	0.288	0.133	0.283
Uganda	Bos taurus	1.3×10^5	0.951	0.974	0.664	0.492	0.140	0.159
Australia	Capra hircus	1.7×10^5	0.912	0.984	0.834	0.386	0.151	0.850
France	Capra hircus	1.9×10^5	0.912	0.983	0.819	0.381	0.154	0.940
Iran (C. aegagrus)	Capra hircus	1.9×10^5	0.928	0.980	0.767	0.408	0.156	0.911
Iran	Capra hircus	2.3×10^5	0.955	0.969	0.616	0.459	0.147	0.943
Italy	Capra hircus	1.9×10^5	0.911	0.983	0.822	0.379	0.155	0.895
Morocco	Capra hircus	2.2×10^5	0.950	0.976	0.706	0.468	0.150	0.627
Iran	Ovis aries	3.8×10^5	0.983	0.958	0.421	0.796	0.154	0.090
Iran (O. orientalis)	Ovis aries	4.5×10^5	0.982	0.961	0.448	0.806	0.148	0.079
Iran (O. vignei)	Ovis aries	3.7×10^5	0.974	0.963	0.480	0.685	0.151	0.086
Various	Ovis aries	4.1×10^5	0.982	0.964	0.500	0.822	0.152	0.084
Morocco	Ovis aries	4×10^5	0.983	0.958	0.387	0.782	0.153	0.081
Barbados	Chlorocebus sabaeus	1.1×10^5	0.895	0.987	0.860	0.351	0.170	0.877
Central Afr. Rep.	Chlorocebus sabaeus	1.7×10^5	0.925	0.974	0.653	0.373	0.157	0.199
Ethiopia	Chlorocebus sabaeus	1.4×10^5	0.892	0.980	0.763	0.320	0.161	0.640
Gambia	Chlorocebus sabaeus	1.4×10^5	0.939	0.982	0.770	0.469	0.159	0.191
Kenya	Chlorocebus sabaeus	1.5×10^5	0.916	0.976	0.680	0.345	0.165	0.461
Nevis	Chlorocebus sabaeus	1×10^5	0.895	0.989	0.884	0.358	0.169	0.908
South Africa	Chlorocebus sabaeus	1.8×10^5	0.930	0.971	0.604	0.361	0.154	0.294
Saint Kitts	Chlorocebus sabaeus	1.2×10^5	0.898	0.986	0.839	0.353	0.165	0.753
Zambia	Chlorocebus sabaeus	1.7×10^5	0.912	0.975	0.676	0.335	0.153	0.406
African	Homo sapiens	5.6×10^4	0.945	0.976	0.627	0.431	0.154	0.154
Admixed American	Homo sapiens	4.5×10^4	0.936	0.977	0.641	0.397	0.158	0.190
East Asian	Homo sapiens	4×10^4	0.943	0.977	0.648	0.422	0.163	0.237
European	Homo sapiens	4.2×10^4	0.945	0.977	0.644	0.429	0.167	0.219
South Asian	Homo sapiens	4.4×10^4	0.945	0.977	0.644	0.430	0.164	0.219

- N_e is the estimated effective population size.
- *Precision* is the estimation of the selection coefficient at population scale (S) given that S_0 is known.
- *Recall* is the estimation of S_0 given selection coefficient at the population scale (S) is known.
- *Recall* for beneficial mutations ($\mathbb{P}[\mathcal{B}_0 \mid \mathcal{B}]$) is thus the proportion of beneficial back-mutations among all beneficial mutations.

Altogether, comparing this table other estimations, we acknowledge that the exact proportion of beneficial back-mutations among all beneficial ones is dependent on the model used to estimate the DFE. However, we can still see that beneficial back-mutations are positively selected compared to neutral and deleterious ones, and this result is not sensitive to the underlying DFE at the population scale.

References

1. Tavaré, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on mathematics in the life sciences* **17,** 57–86 (1986).

2. Nielsen, R. Mapping Mutations on Phylogenies. *Systematic Biology* **51** (ed Huelsenbeck, J.) 729–739 (2002).

3. Rodrigue, N., Philippe, H. & Lartillot, N. Uniformization for Sampling Realizations of Markov Processes: Applications to Bayesian Implementations of Codon Substitution Models. *Bioinformatics* **24,** 56–62 (2008).

4. Irvahn, J. & Minin, V. N. Phylogenetic Stochastic Mapping without Matrix Exponentiation. *Journal of Computational Biology* **21,** 676–690 (2014).

5. Davydov, I. I., Robinson-Rechavi, M. & Salamin, N. State Aggregation for Fast Likelihood Computations in Molecular Evolution. *Bioinformatics* **33,** 354–362 (2017).

6. Orme, D. *et al.* The Caper Package: Comparative Analysis of Phylogenetics and Evolution in R. *R package version* **5,** 1–36 (2013).

7. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution* **34,** 1812–1819 (2017).

8. Latrille, T., Rodrigue, N. & Lartillot, N. Genes and Sites under Adaptation at the Phylogenetic Scale Also Exhibit Adaptation at the Population-Genetic Scale. *Proceedings of the National Academy of Sciences of the United States of America* **120,** e2214977120 (2023).

Supplementary material from: Increased positive selection in highly recombining genes is not an evidence for a beneficial effect of recombination

J. Joseph[1]

[1]LBBE, Université Lyon 1, CNRS, UMR 5558, Villeurbanne, France

julien.joseph@ens-lyon.fr

October 14, 2023

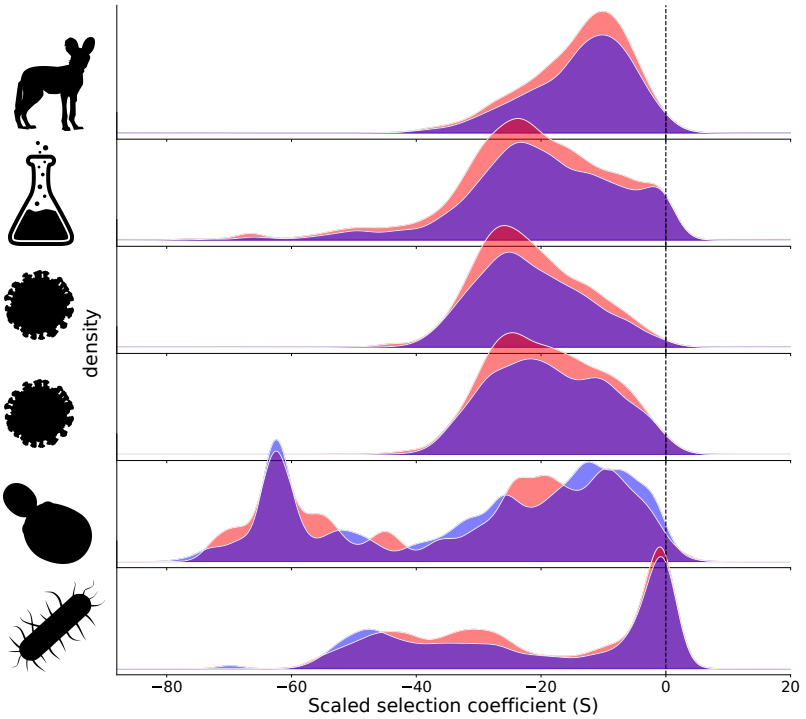


Figure S1: Distribution of fitness effects of new mutations at equilibrium separately for WS (red) and SW (blue) mutations. For each fitness landscape, the relative fitnesses were shuffled among amino-acids. Equilibrium frequencies were computed with $B = 0$. From top to bottom: 2000 sites randomly sampled from the mammalian fitness landscapes, the concatenate of the DMS fitness landscapes (1389 sites), the fitness landscape of the influenza protein NP (498 sites), the influenza protein HA (564 sites), the *S.cervisae* protein Gal4 (64 sites) and the *E.coli* protein β-lactamase (263 sites).

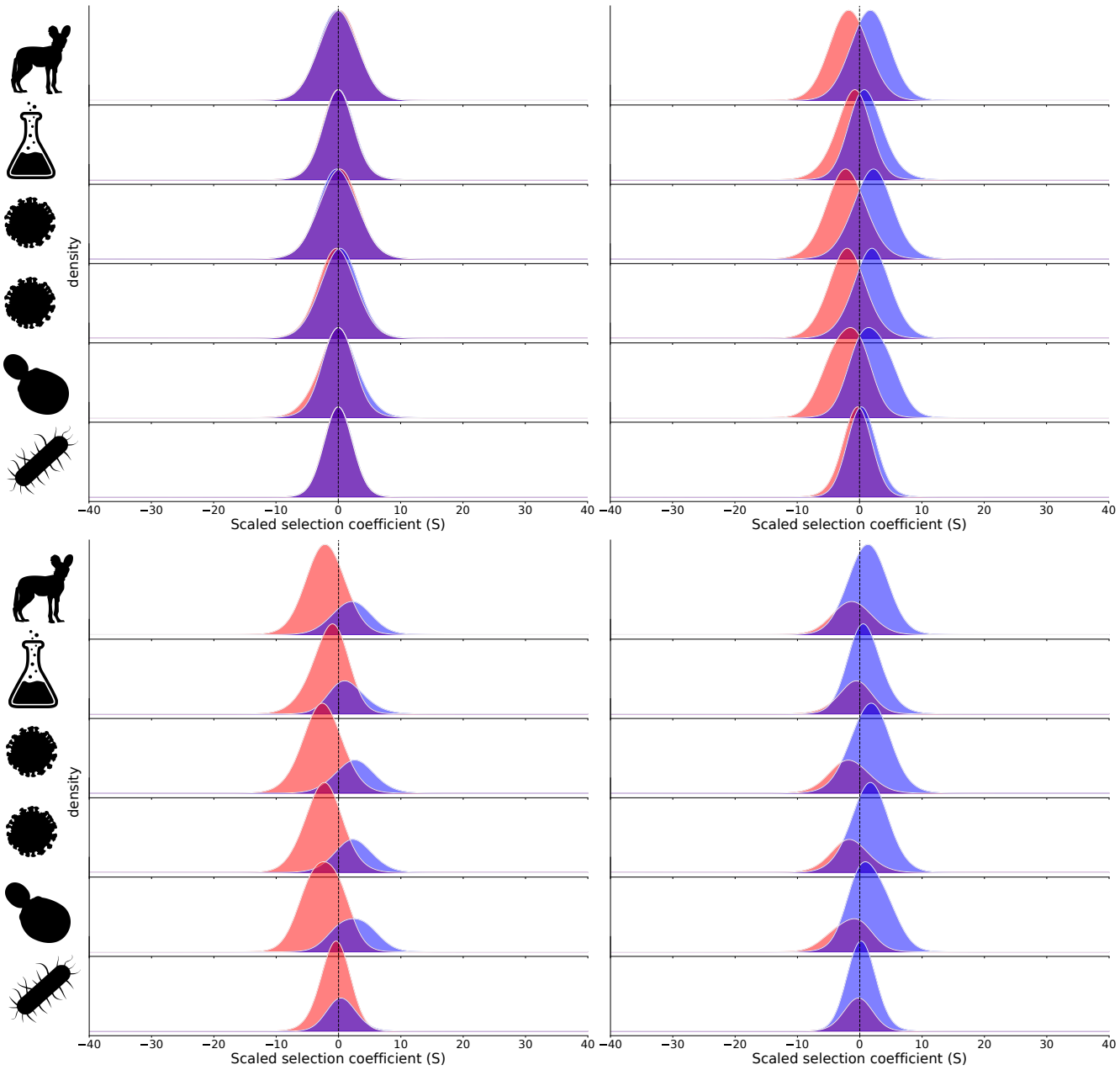


Figure S2: Distribution of fitness effects of new substitutions separately for WS (red) and SW (blue) substitutions. Equilibrium frequencies were computed for the concatenated DMS fitness landscape, with $B = 0$ (top left) and $B = 2$ (top right and bottom). The substitution rate from the equilibrium sequence was subsequently computed with $B = 0$ (top left), $B = 2$ (top right), $B = 3$, (bottom left), and $B = 1$ (bottom right). The top panels therefore represent substitution rates at equilibrium, while the bottom panels represent substitution rates out of equilibrium. From top to bottom: 2000 sites randomly sampled from the mammalian fitness landscapes, the concatenate of the DMS fitness landscapes (1389 sites), the fitness landscape of the influenza protein NP (498 sites), the influenza protein HA (564 sites), the *S.cervisae* protein Gal4 (64 sites) and the *E.coli* protein β-lactamase (263 sites).

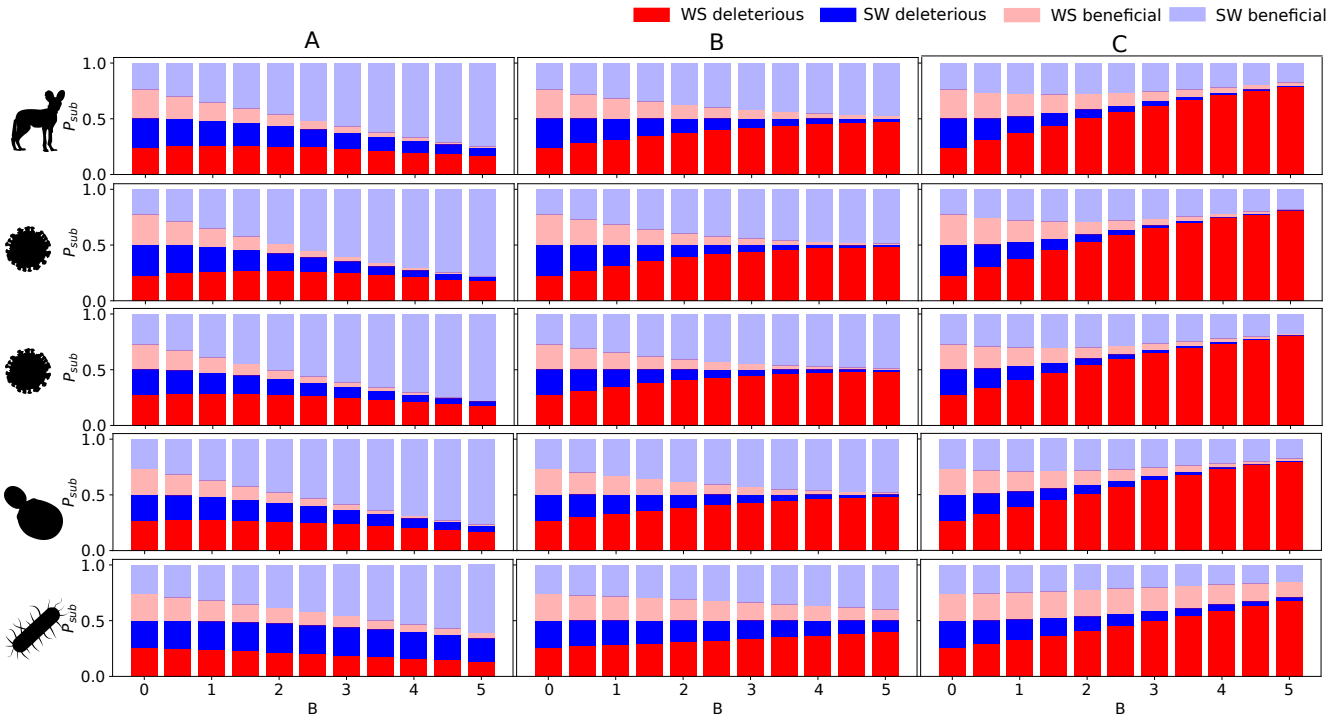


Figure S3: Proportion of the substitutions (P_{sub}) contributed by WS deleterious (bright red), SW deleterious (bright blue), WS beneficial (light red) ans SW beneficial (light blue) mutations as a function of the population-scaled gBGC coefficient under the concatenated DMS fitness landscape in three scenarios: Equilibrium frequencies are computed with B, and substitutions from this equilibrium sequence are computed with $0.7 \times B$ (A), B (B) and $1.3 \times B$ (C). From top to bottom: 2000 sites randomly sampled from the mammalian fitness landscapes, the fitness landscape of the influenza protein NP (498 sites), the influenza protein HA (564 sites), the *S.cervisae* protein Gal4 (64 sites) and the *E.coli* protein β-lactamase (263 sites).

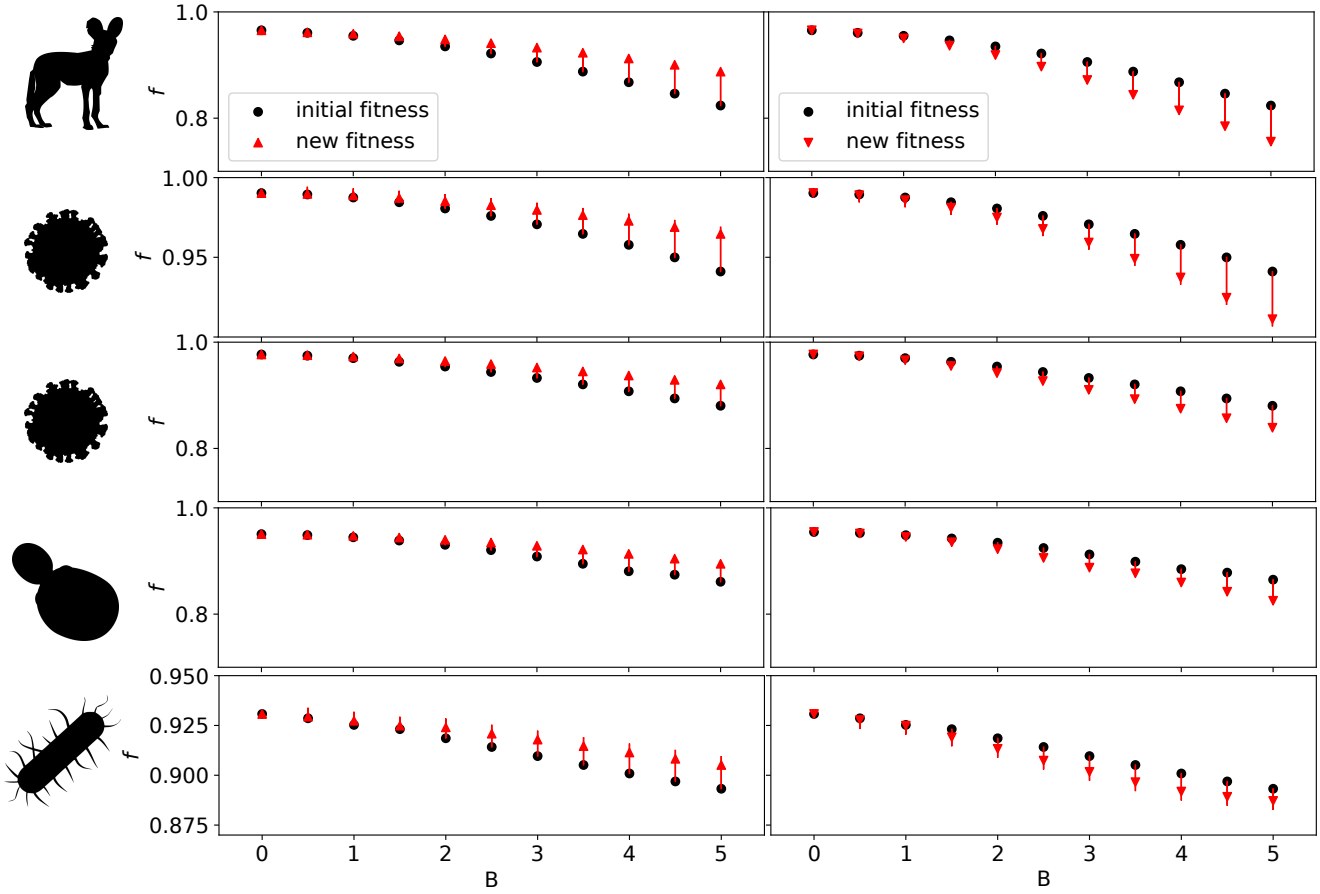


Figure S4: Relative fitness (f) of sequences as a function of the B they are evolving under (black dots). Red arrows correspond to the fitness evolution of the sequences after a decrease of B by 30% (left panels), or after an increase of B by 30% (right panels). From top to bottom: 2000 sites randomly sampled from the mammalian fitness landscapes, the fitness landscape of the influenza protein NP (498 sites), the influenza protein HA (564 sites), the *S.cervisae* protein Gal4 (64 sites) and the *E.coli* protein β-lactamase (263 sites).

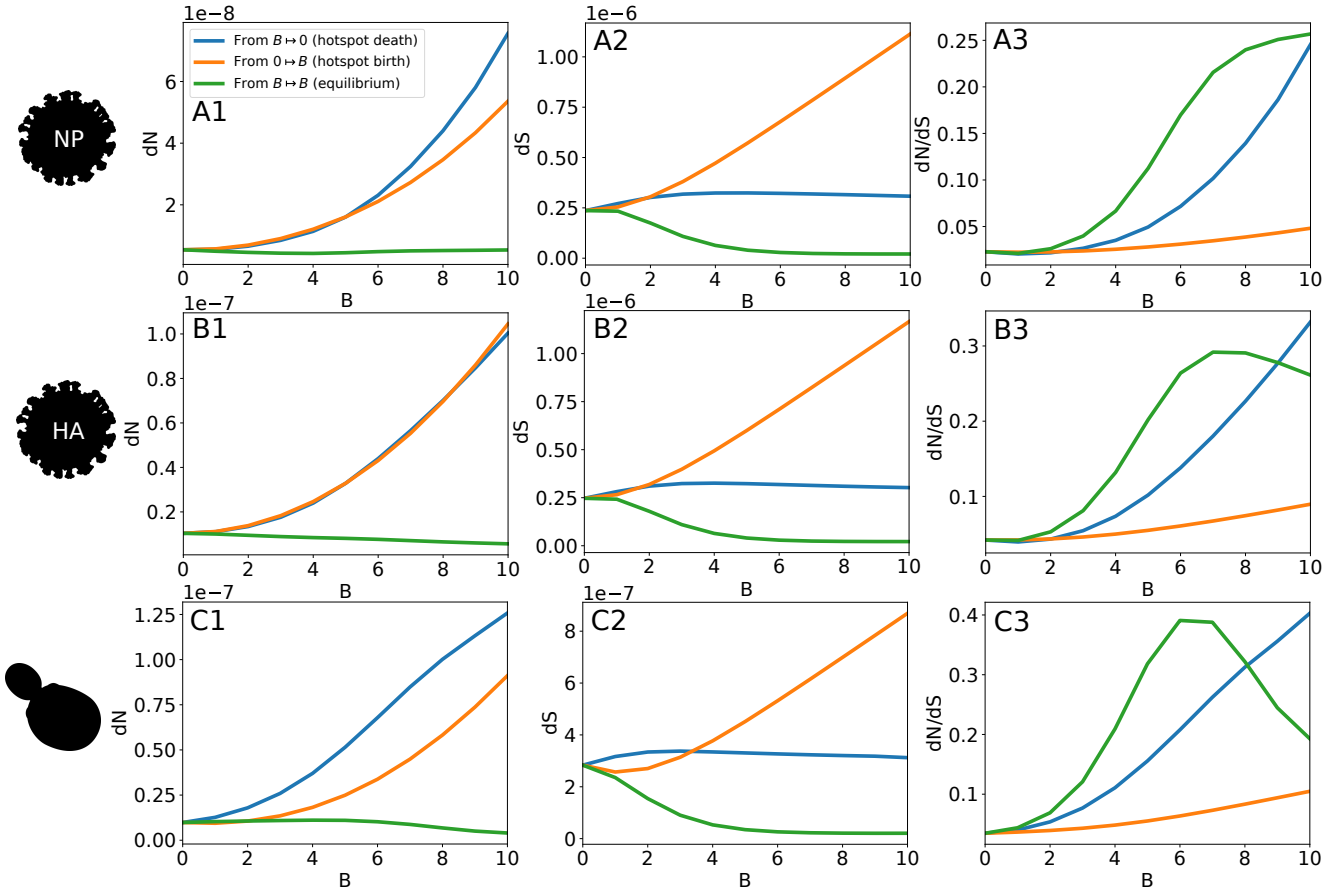


Figure S5: dN (A), dS (B), and dN/dS (C) as a function of the population scaled gBGC coefficient B in three scenarios: Equilibrium codon frequencies are computed without gBGC, and substitutions subsequently accumulate under a population-scaled gBGC coefficient of B (orange line), mimicking the birth of a recombination hotspot. Equilibrium codon frequencies are computed under a population-scaled gBGC coefficient of B, and substitutions subsequently accumulate without gBGC (orange line), mimicking the death of a recombination hotspot. And finally, equilibrium codon frequencies are computed under a population-scaled gBGC coefficient of B, and substitutions subsequently accumulate at equilibrium, under B (green line). From top to bottom: the fitness landscape of the influenza protein NP (498 sites), the influenza protein HA (564 sites) and the *S.cervisae* protein Gal4 (64 sites).

Bibliography

Adewoye, A. B., Lindsay, S. J., Dubrova, Y. E., and Hurles, M. E. 2015. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat Commun*, 6(1): 6684. *Cité page 172*

Alleva, B., Brick, K., Pratto, F., Huang, M., and Camerini-Otero, R. D. 2021. Cataloging Human PRDM9 Allelic Variation Using Long-Read Sequencing Reveals PRDM9 Population Specificity and Two Distinct Groupings of Related Alleles. *Frontiers in Cell and Developmental Biology*, 9. *Cité pages 9, 11, 64, 65*

Amado, A. and Bank, C. 2023. Ecological tradeoffs lead to complex evolutionary trajectories and sustained diversity on dynamic fitness landscapes. Pages: 2023.10.11.561986 Section: New Results. *Cité page 131*

Ashenberg, O., Gong, L. I., and Bloom, J. D. 2013. Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences*, 110(52): 21071–21076. *Cité page 136*

Auton, A. and McVean, G. 2007. Recombination rate estimation in the presence of hotspots. *Genome Res.*, 17(8): 1219–1227. *Cité page 15*

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., Myers, S., Donnelly, P., Przeworski, M., and McVean, G. 2012. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*, 336: 7. *Cité pages 10, 64*

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., Hayward, J. J., Cohen, P. E., Greally, J. M., Wang, J., Bustamante, C. D., and Boyko, A. R. 2013. Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genet*, 9(12): e1003984. *Cité pages 14, 38, 64, 66, 67, 68, 75*

Axelsson, E., Webster, M. T., Ratnakumar, A., Consortium, T. L., Ponting, C. P., and Lindblad-Toh, K. 2012. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.*, 22(1): 51–63. *Cité pages 13, 14, 64*

Bajić, D., Vila, J. C. C., Blount, Z. D., and Sánchez, A. 2018. On the deformability of an empirical fitness landscape by microbial evolution. *Proceedings of the National Academy of Sciences*, 115(44): 11286–11291. *Cité page 131*

Baker, B. S., Carpenter, A. T. C., Esposito, M. S., Esposito, R. E., and Sandler, L. 1976. The genetic control of meiosis. *Annu. Rev. Genet.*, 10(1): 53–134. *Cité page 8*

Baker, C. L., Walker, M., Kajita, S., Petkov, P. M., and Paigen, K. 2014. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. *Genome Res.*, 24(5): 724–732. *Cité page 11*

Baker, C. L., Kajita, S., Walker, M., Saxl, R. L., Raghupathy, N., Choi, K., Petkov, P. M., and Paigen, K. 2015. PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. *PLOS Genetics*, 11(1): e1004916. *Cité pages 10, 64, 65, 72*

Baker, Z., Schumer, M., Haba, Y., Bashkirova, L., Holland, C., Rosenthal, G. G., and Przeworski, M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *eLife*, 6: e24133. *Cité pages 10, 13, 14*

Baker, Z., Przeworski, M., and Sella, G. 2022. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence. Pages: 2022.09.27.509707 Section: New Results. *Cité pages 11, 64, 72, 73, 74*

Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., and Pagnani, A. 2014. Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLOS ONE*, 9(3): e92721. *Cité page 136*

Barton, H. J. and Zeng, K. 2021. The effective population size modulates the strength of GC biased gene conversion in two passerines. Pages: 2021.04.20.440602 Section: New Results. *Cité page 23*

Barzel, A. and Kupiec, M. 2008. Finding a match: how do homologous sequences get together for recombination? *Nat Rev Genet*, 9(1): 27–37. *Cité page 8*

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. 2010. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science*, 327(5967): 836–840. *Cité pages 9, 64, 65*

Bengtsson, B. O. 1986. Biased conversion as the primary function of recombination. *Genetics Research*, 47(1): 77–80. *Cité pages 18, 26, 27, 33*

Bengtsson, B. O. 1990. The effect of biased conversion on the mutation load. *Genetics Research*, 55(3): 183–187. *Cité pages 26, 27, 33*

Bengtsson, B. O. and Uyenoyama, M. K. 1990. Evolution of the segregation ratio: Modification of gene conversion and meiotic drive. *Theoretical Population Biology*, 38(2): 192–218. *Cité pages 5, 33*

Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.-C., Nicolas, A., and Forterre, P. 1997. An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature*, 386(6623): 414–417. *Cité page 8*

Berglund, J., Pollard, K. S., and Webster, M. T. 2009. Hotspots of Biased Nucleotide Substitutions in Human Genes. *PLOS Biology*, 7(1): e1000026. *Cité pages 26, 27, 33*

Berglund, J., Quilez, J., Arndt, P. F., and Webster, M. T. 2015. Germline Methylation Patterns Determine the Distribution of Recombination Events in the Dog Genome. *Genome Biology and Evolution*, 7(2): 522–530. *Cité page 74*

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7): 1499–1504. *Cité page 14*

Bisardi, M., Rodriguez-Rivas, J., Zamponi, F., and Weigt, M. 2022. Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution. *Molecular Biology and Evolution*, 39(1): msab321. *Cité page 136*

Blat, Y., Protacio, R. U., Hunter, N., and Kleckner, N. 2002. Physical and Functional Interactions among Basic Chromosome Organizational Features Govern Early Steps of Meiotic Chiasma Formation. *Cell*, 111(6): 791–802. *Cité page 8*

Bloom, J. D. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol Direct*, 12(1): 1–24. *Cité page 136*

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. 2016. Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an Avian System. *Mol Biol Evol*, 33(1): 216–227. *Cité pages 33, 38*

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol*, 20(1): 5. *Cité page 33*

Boman, J., Mugal, C. F., and Backström, N. 2021. The Effects of GC-Biased Gene Conversion on Patterns of Genetic Diversity among and across Butterfly Genomes. *Genome Biology and Evolution*, 13(5). *Cité pages 23, 168, 172, 173*

Bomblies, K. and Peichel, C. L. 2022. Genetics of adaptation. *Proceedings of the National Academy of Sciences*, 119(30): e2122152119. *Cité pages 132, 133*

Borde, V., Goldman, A. S. H., and Lichten, M. 2000. Direct Coupling Between Meiotic DNA Replication and Recombination Initiation. *Science*, 290(5492): 806–809. *Cité pages 68, 73*

Borde, V., Robine, N., Lin, W., Bonfils, S., Géli, V., and Nicolas, A. 2009. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *The EMBO Journal*, 28(2): 99–111. *Cité pages 14, 67, 73*

Boulton, A., Myers, R. S., and Redfield, R. J. 1997. The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences*, 94(15): 8058–8063. *Cité pages 10, 65*

Bourc'his, D. and Proudhon, C. 2008. Sexual dimorphism in parental imprint ontogeny and contribution to embryonic development. *Molecular and Cellular Endocrinology*, 282(1): 87–94. *Cité page 68*

Brazier, T. and Glémin, S. 2022. Diversity and determinants of recombination landscapes in flowering plants. *PLOS Genetics*, 18(8): e1010141. *Cité pages 8, 66*

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. 2012. Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400): 642–645. *Cité pages 8, 9, 14, 16, 38, 64, 66, 67, 69, 74*

Brown, T. C. and Jiricny, J. 1987. A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, 50(6): 945–950. *Cité pages 17, 33*

Brown, T. C. and Jiricny, J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell*, 54(5): 705–711. *Cité page 33*

Bruck, D. 1957. Male segregation ratio advantage as a factor in maintaining lethal alleles in wild populations of house mice*. *Proceedings of the National Academy of Sciences*, 43(1): 152–158. *Cité page 6*

Buard, J., Rivals, E., Segonzac, D. D. d., Garres, C., Caminade, P., Massy, B. d., and Boursot, P. 2014. Diversity of Prdm9 Zinc Finger Array in Wild Mice Unravels New Facets of the Evolutionary Turnover of this Coding Minisatellite. *PLOS ONE*, 9(1): e85021. *Cité pages 9, 11, 64, 65, 71*

Burgin, C. J., Colella, J. P., Kahn, P. L., and Upham, N. S. 2018. How many species of mammals are there? *Journal of Mammalogy*, 99(1): 1–14. *Cité page 38*

Cabral, G., Marques, A., Schubert, V., Pedrosa-Harand, A., and Schlögelhofer, P. 2014. Chiasmatic and achiasmatic inverted meiosis of plants with holocentric chromosomes. *Nat Commun*, 5(1): 5070. *Cité page 8*

Camerini-Otero, R. D. and Hsieh, P. 1995. Homologous recombination proteins in prokaryotes and eukaryotes. *Annu. Rev. Genet.*, 29(1): 509–552. *Cité page 8*

Campbell, C. L., Bhérer, C., Morrow, B. E., Boyko, A. R., and Auton, A. 2016. A Pedigree-Based Map of Recombination in the Domestic Dog Genome. *G3*, 6(11): 3517–3524. *Cité page 68*

Carothers, E. E. 1913. The mendelian ratio in relation to certain orthopteran chromosomes. *J. Morphol.*, 24(4): 487–511. *Cité page 5*

Cavassim, M. I. A., Baker, Z., Hoge, C., Schierup, M. H., Schumer, M., and Przeworski, M. 2022. PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2. *Proceedings of the National Academy of Sciences*, 119(9): e2114401119. *Cité pages 13, 14*

Chan, A. H., Jenkins, P. A., and Song, Y. S. 2012. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLOS Genetics*, 8(12): e1003090. *Cité pages 9, 15, 68*

Charlesworth, B., Morgan, M. T., and Charlesworth, D. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics*, page 15. *Cité page 28*

Charlesworth, B., Langley, C. H., and Sniegowski, P. D. 1997. Transposable Element Distributions in Drosophila. *Genetics*, 147(4): 1993–1995. *Cité pages 67, 72*

Charlesworth, D., Zhang, Y., Bergero, R., Graham, C., Gardner, J., and Yong, L. 2020. Using GC Content to Compare Recombination Patterns on the Sex Chromosomes and Autosomes of the Guppy, Poecilia reticulata, and Its Close Outgroup Species. *Molecular Biology and Evolution*, 37(12): 3550–3562. *Cité page 38*

Charlesworth, J. and Eyre-Walker, A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proceedings of the National Academy of Sciences*, 104(43): 16992–16997. *Cité pages 130, 131, 133, 134*

Chen, J., Swofford, R., Johnson, J., Cummings, B. B., Rogel, N., Lindblad-Toh, K., Haerty, W., Palma, F. d., and Regev, A. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.*, 29(1): 53–63. *Cité page 75*

Chikashige, Y., Ding, D.-Q., Funabiki, H., Haraguchi, T., Mashiko, S., Yanagida, M., and Hiraoka, Y. 1994. Telomere-Led Premeiotic Chromosome Movement in Fission Yeast. *Science*, 264(5156): 270–273. *Cité page 8*

Choi, K. and Henderson, I. R. 2015. Meiotic recombination hotspots – a comparative view. *The Plant Journal*, 83(1): 52–61. *Cité pages 9, 14, 64, 68*

Choi, K., Zhao, X., Tock, A. J., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J., Ziolkowski, P. A., Yelina, N. E., Hwang, I., Martienssen, R. A., and Henderson, I. R. 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. *Genome Res.*, 28(4): 532–546. *Cité pages 14, 67, 74*

Clark, F. E. and Akera, T. 2021. Unravelling the mystery of female meiotic drive: where we are. *Open Biology*, 11(9): 210074. *Cité page 6*

Clément, Y. and Arndt, P. F. 2013. Meiotic Recombination Strongly Influences GC-Content Evolution in Short Regions in the Mouse Genome. *Molecular Biology and Evolution*, 30(12): 2612–2618. *Cité page 38*

Clément, Y., Sarah, G., Holtz, Y., Homa, F., Pointet, S., Contreras, S., Nabholz, B., Sabot, F., Sauné, L., Ardisson, M., Bacilieri, R., Besnard, G., Berger, A., Cardi, C., Bellis, F. D., Fouet, O., Jourda, C., Khadari, B., Lanaud, C., Leroy, T., Pot, D., Sauvage, C., Scarcelli, N., Tregear, J., Vigouroux, Y., Yahiaoui, N., Ruiz, M., Santoni,

S., Labouisse, J.-P., Pham, J.-L., David, J., and Glémin, S. 2017. Evolutionary forces affecting synonymous variations in plant genomes. *PLOS Genetics*, 13(5): e1006799. *Cité pages 23, 169*

Cole, F., Baudat, F., Grey, C., Keeney, S., de Massy, B., and Jasin, M. 2014. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat Genet*, 46(10): 1072–1080. *Cité page 172*

Cooper, D. N., Taggart, M. H., and Bird, A. P. 1983. Unmethlated domains in vertebrate DNA. *Nucleic Acids Research*, 11(3): 647–658. *Cité page 14*

Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. A., Wergedal, J. E., Bult, C., Paigen, B., Flint, J., Tsaih, S.-W., Churchill, G. A., and Broman, K. W. 2009. A New Standard Genetic Map for the Laboratory Mouse. *Genetics*, 182(4): 1335–1344. *Cité page 171*

Damm, E., Ullrich, K. K., Amos, W. B., and Odenthal-Hesse, L. 2022. Evolution of the recombination regulator PRDM9 in minke whales. *BMC Genomics*, 23(1): 1–16. *Cité page 13*

Dapper, A. L. and Payseur, B. A. 2019. Molecular evolution of the meiotic recombination pathway in mammals. *Evolution*, 73(12): 2368–2389. *Cité page 73*

Darwin, C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray. *Cité pages 34, 128, 129*

Davies, B., Hatton, E., Altemose, N., Hussin, J. G., Pratto, F., Zhang, G., Hinch, A. G., Moralli, D., Biggs, D., Diaz, R., Preece, C., Li, R., Bitoun, E., Brick, K., Green, C. M., Camerini-Otero, R. D., Myers, S. R., and Donnelly, P. 2016. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature*, 530(7589): 171–176. *Cité pages 8, 9, 13, 64, 66, 72*

de Massy, B. 2013. Initiation of Meiotic Recombination: How and Where? Conservation and Specificities Among Eukaryotes. *Annu. Rev. Genet.*, 47(1): 563–599. *Cité pages 38, 74*

Detloff, P., White, M. A., and Petes, T. D. 1992. Analysis of a gene conversion gradient at the HIS4 locus in Saccharomyces cerevisiae. *Genetics*, 132(1): 113–123. *Cité pages 10, 65*

Diagouraga, B., Clément, J. A. J., Duret, L., Kadlec, J., de Massy, B., and Baudat, F. 2018. PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites. *Molecular Cell*, 69(5): 853–865.e6. *Cité pages 11, 66*

Doud, M. B., Ashenberg, O., and Bloom, J. D. 2015. Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Molecular Biology and Evolution*, 32(11): 2944–2960. *Cité page 136*

Duchemin, L., Lanore, V., Veber, P., and Boussau, B. 2023. Evaluation of Methods to Detect Shifts in Directional Selection at the Genome Scale. *Molecular Biology and Evolution*, 40(2): msac247. *Cité page 176*

Duret, L. and Arndt, P. F. 2008. The Impact of Recombination on Nucleotide Substitutions in the Human Genome. *PLOS Genetics*, 4(5): e1000071. *Cité pages 24, 33*

Duret, L. and Galtier, N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genom. Hum. Genet.*, 10(1): 285–311. *Cité pages 17, 18*

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., and Sella, G. 2016. A Genomic Map of the Effects of Linked Selection in Drosophila. *PLoS Genet*, 12(8): e1006130. *Cité page 28*

Eram, M. S., Bustos, S. P., Lima-Fernandes, E., Siarheyeva, A., Senisterra, G., Hajian, T., Chau, I., Duan, S., Wu, H., Dombrovski, L., Schapira, M., Arrowsmith, C. H., and Vedadi, M. 2014. Trimethylation of Histone H3 Lysine 36 by Human Methyltransferase PRDM9 Protein *. *Journal of Biological Chemistry*, 289(17): 12177–12188. *Cité pages 9, 64, 66*

Eyre-Walker, A. 2002. Changing Effective Population Size and the McDonald-Kreitman Test. *Genetics*, 162(4): 2017–2024. *Cité page 134*

Eyre-Walker, A. and Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8(8): 610–618. *Cité page 133*

Eyre-Walker, A. and Keightley, P. D. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9): 2097–2108. *Cité page 30*

Eyre-Walker, A., Woolfit, M., and Phelps, T. 2006. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics*, 173(2): 891–900. *Cité page 30*

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J. M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, , Feilitzen, K. v., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F., and Uhlén, M. 2014. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics *. *Molecular & Cellular Proteomics*, 13(2): 397–406. *Cité page 75*

Fawcett, D. W. 1956. THE FINE STRUCTURE OF CHROMOSOMES IN THE MEIOTIC PROPHASE OF VERTEBRATE SPERMATOCYTES. *The Journal of Biophysical and Biochemical Cytology*, 2(4): 403–406. *Cité page 8*

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics*, page 20. *Cité page 4*

Figuet, E., Ballenghien, M., Romiguier, J., and Galtier, N. 2015. Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates. *Genome Biology and Evolution*, 7(1): 240–250. *Cité pages 24, 38*

Fischer, R. A. 1930. The Genetical Theory of Natural Selection. *Oxford, UK: Oxford Univ. Press*. *Cité page 136*

Foster, P. L., Lee, H., Popodi, E., Townes, J. P., and Tang, H. 2015. Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 112(44): E5990–E5999. *Cité page 173*

Frazer, K. A. and Consortium, T. I. H. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164): 851–861. *Cité page 75*

Fukushima, K. and Pollock, D. D. 2023. Detecting macroevolutionary genotype–phenotype associations using error-corrected rates of protein convergence. *Nat Ecol Evol*, 7(1): 155–170. *Cité page 176*

Fung, J. C., Marshall, W. F., Dernburg, A., Agard, D. A., and Sedat, J. W. 1998. Homologous Chromosome Pairing in Drosophila melanogaster Proceeds through Multiple Independent Initiations. *Journal of Cell Biology*, 141(1): 5–20. *Cité page 8*

Galtier, N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLoS Genet*, 12(1): e1005774. *Cité pages 30, 133, 134*

Galtier, N. 2021. Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer Community Journal*, 1. *Cité pages 22, 33, 64, 168*

Galtier, N., Bazin, E., and Bierne, N. 2006. GC-Biased Segregation of Noncoding Polymorphisms in Drosophila. *Genetics*, 172(1): 221–228. *Cité page 173*

Galtier, N., Duret, L., Glémin, S., and Ranwez, V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*, 25(1): 1–5. *Cité pages 26, 27, 33*

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5): 1092–1103. *Cité pages 23, 169*

Genestier, A., Duret, L., and Lartillot, N. 2023. Bridging the gap between the evolutionary dynamics and the molecular mechanisms of meiosis: a model based exploration of the PRDM9 intra-genomic Red Queen. Pages: 2023.03.08.531712 Section: New Results. *Cité pages 12, 65, 72, 73*

Gerton, J. L. and Hawley, R. S. 2005. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nat Rev Genet*, 6(6): 477–487. *Cité page 8*

Gillespie, J. H. 1995. On Ohta's hypothesis: Most amino acid substitutions are deleterious. *J Mol Evol*, 40(1): 64–69. *Cité page 131*

Glémin, S. 2010. Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics*, 185(3): 939–959. *Cité pages 26, 27*

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res.*, 25(8): 1215–1228. *Cité pages 23, 33, 38*

Goodier, J. L. 2016. Restricting retrotransposons: a review. *Mobile DNA*, 7(1): 1–30. *Cité page 67*

Grabowski, M., Pienaar, J., Voje, K. L., Andersson, S., Fuentes-González, J., Kopperud, B. T., Moen, D. S., Tsuboi, M., Uyeda, J., and Hansen, T. F. 2023. A Cautionary Note on "A Cautionary Note on the Use of Ornstein Uhlenbeck Models in Macroevolutionary Studies". *Systematic Biology*, 72(4): 955–963. *Cité page 132*

Gregorova, S., Gergelits, V., Chvatalova, I., Bhattacharyya, T., Valiskova, B., Fotopulosova, V., Jansa, P., Wiatrowska, D., and Forejt, J. 2018. Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. *eLife*, 7: e34282. *Cité page 13*

Gregorová, S. and Forejt, J. 2000. PWD/Ph and PWK/Ph inbred mouse strains of Mus m. musculus subspecies–a valuable resource of phenotypic variations and genomic polymorphisms. *Folia Biol (Praha)*, 46(1): 31–41. *Cité page 13*

Grey, C., Clément, J. A. J., Buard, J., Leblanc, B., Gut, I., Gut, M., Duret, L., and Massy, B. d. 2017. In vivo binding of PRDM9 reveals interactions with noncanonical genomic sites. *Genome Res.*, 27(4): 580–590. *Cité pages 9, 64*

Grin, I. and Ishchenko, A. A. 2016. An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation. *Nucleic Acids Research*, 44(8): 3713–3727. *Cité page 25*

Halldorsson, B. V., Hardarson, M. T., Kehr, B., Styrkarsdottir, U., Gylfason, A., Thorleifsson, G., Zink, F., Jonasdottir, A., Jonasdottir, A., Sulem, P., Masson, G., Thorsteinsdottir, U., Helgason, A., Kong, A., Gudbjartsson, D. F., and Stefansson, K. 2016. The rate of meiotic gene conversion varies by sex and age. *Nat Genet*, 48(11): 1377–1384. *Cité pages 18, 25, 172*

Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7): 910–917. *Cité pages 31, 135*

Hancks, D. C. and Kazazian, H. H. 2016. Roles for retrotransposon insertions in human disease. *Mobile DNA*, 7(1): 1–28. *Cité page 67*

Hansen, T. F. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution*, 51(5): 1341–1351. *Cité page 132*

Hardy, G. H. 1908. Mendelian Proportions in a Mixed Population. *Science*, 28(706): 49–50. *Cité page 5*

Hartl, D. L. and Taubes, C. H. 1996. Compensatory Nearly Neutral Mutations: Selection without Adaptation. *Journal of Theoretical Biology*, 182(3): 303–309. *Cité pages 131, 133*

Hartl, D. L. and Taubes, C. H. 1998. Towards a theory of evolutionary adaptation. *Genetica*, 102(0): 525–533. *Cité page 130*

Hassold, T., Hall, H., and Hunt, P. 2007. The origin of human aneuploidy: where we have been, where we are going. *Human Molecular Genetics*, 16(R2): R203–R208. *Cité page 8*

He, C., Liu, N., Xie, D., Liu, Y., Xiao, Y., and Li, F. 2019. Structural basis for histone H3K4me3 recognition by the N-terminal domain of the PHD finger protein Spp1. *Biochemical Journal*, 476(13): 1957–1973. *Cité page 74*

He, Y., Wang, M., Dukowic-Schulze, S., Zhou, A., Tiang, C.-L., Shilo, S., Sidhu, G. K., Eichten, S., Bradbury, P., Springer, N. M., Buckler, E. S., Levy, A. A., Sun, Q., Pillardy, J., Kianian, P. M. A., Kianian, S. F., Chen, C., and Pawlowski, W. P. 2017. Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proceedings of the National Academy of Sciences*, 114(46): 12231–12236. *Cité pages 14, 67*

Hickey, D. A. and Golding, G. B. 2018. The advantage of recombination when selection is acting at many genetic Loci. *Journal of Theoretical Biology*, 442: 123–128. *Cité page 4*

Hietpas, R. T., Bank, C., Jensen, J. D., and Bolon, D. N. A. 2013. SHIFTING FITNESS LANDSCAPES IN RESPONSE TO ALTERED ENVIRONMENTS. *Evolution*, 67(12): 3512–3522. *Cité page 131*

Hill, W. G. and Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.*, page 26. *Cité page 4*

Hinch, A. G., Zhang, G., Becker, P. W., Moralli, D., Hinch, R., Davies, B., Bowden, R., and Donnelly, P. 2019. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science*, 363(6433): eaau8861. *Cité pages 9, 11, 64, 72, 73, 74*

Hoge, C. R., Manuel, M. d., Mahgoub, M., Okami, N., Fuller, Z. L., Banerjee, S., Baker, Z., Mcnulty, M., Andolfatto, P., Macfarlan, T. S., Schumer, M., Tzika, A. C., and Przeworski, M. 2023. Patterns of recombination in snakes reveal a tug of war between PRDM9 and promoter-like features. Pages: 2023.07.11.548536 Section: New Results. *Cité pages 14, 38, 73*

Hurst, L. D. 2019. A century of bias in genetics and evolution. *Heredity*, 123(1): 33–43. *Cité pages 5, 18, 19*

Jackson, B. and Charlesworth, B. 2021. Evidence for a force favoring GC over AT at short intronic sites in Drosophila simulans and Drosophila melanogaster. *G3 Genes/Genomes/Genetics*, 11(9). *Cité page 173*

Jeffreys, A. J., Cotton, V. E., Neumann, R., and Lam, K.-W. G. 2013. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. *Proceedings of the National Academy of Sciences*, 110(2): 600–605. *Cité page 9*

Jin, X., Fudenberg, G., and Pollard, K. S. 2021. Genome-wide variability in recombination activity is associated with meiotic chromatin organization. *Genome Res.*, 31(9): 1561–1572. *Cité pages 69, 70, 72*

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2017. Shifting Balance on a Static Mutation–Selection Landscape: A Novel Scenario of Positive Selection. *Molecular Biology and Evolution*, 34(2): 391–407. *Cité pages 130, 131*

Kapopoulou, A., Pfeifer, S. P., Jensen, J. D., and Laurent, S. 2018. The Demographic History of African Drosophila melanogaster. *Genome Biology and Evolution*, 10(9): 2338–2342. *Cité pages 172, 173*

Kauppi, L., Barchi, M., Lange, J., Baudat, F., Jasin, M., and Keeney, S. 2013. Numerical constraints and feedback control of double-strand breaks in mouse meiosis. *Genes Dev.*, 27(8): 873–886. *Cité page 11*

Kaur, T. and Rockman, M. V. 2014. Crossover Heterogeneity in the Absence of Hotspots in Caenorhabditis elegans. *Genetics*, 196(1): 137–148. *Cité pages 9, 68*

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L., and Ellegren, H. 2017. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Mol Ecol*, 26(16): 4158–4172. *Cité pages 14, 64, 68*

Keeney, S., Giroux, C. N., and Kleckner, N. 1997. Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family. *Cell*, 88(3): 375–384. *Cité page 8*

Kimura, M. 1962. ON THE PROBABILITY OF FIXATION OF MUTANT GENES IN A POPULATION. *Genetics*, 47(6): 713–719. *Cité page 22*

Klose, R. J. and Bird, A. P. 2006. Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences*, 31(2): 89–97. *Cité pages 14, 67*

Kondrashov, A. S. and Crow, J. F. 1991. Haploidy or diploidy: which is better? *Nature*, 351(6324): 314–315. *Cité page 3*

Kono, H., Tamura, M., Osada, N., Suzuki, H., Abe, K., Moriwaki, K., Ohta, K., and Shiroishi, T. 2014. Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. *DNA Research*, 21(3): 315–326. *Cité pages 9, 11, 64*

Kowalczyk, A., Meyer, W. K., Partha, R., Mao, W., Clark, N. L., and Chikina, M. 2019. RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics*, 35(22): 4815–4817. *Cité page 176*

Krokan, H. E. and Bjørås, M. 2013. Base Excision Repair. *Cold Spring Harb Perspect Biol*, 5(4): a012583. *Cité pages 24, 25*

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. 2015. Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. *PLOS ONE*, 10(6): e0131673. *Cité page 75*

Kumar, P., Henikoff, S., and Ng, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4(7): 1073–1081. *Cité page 31*

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., and Hedges, S. B. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, 39(8): msac174. *Cité page 69*

Lachance, J. and Tishkoff, S. A. 2014. Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles. *The American Journal of Human Genetics*, 95(4): 408–420. *Cité pages 27, 33*

Lam, I. and Keeney, S. 2015. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science*, 350(6263): 932–937. *Cité pages 14, 64*

Lamarck, J. B. d. M. 1815. *Histoire naturelle des Animaux Sans vertèbres*, volume 1. *Cité page 128*

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9): 1813–1831. *Cité page 75*

Langley, C. H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genetics Research*, 52(3): 223–235. *Cité page 67*

Lartillot, N. 2013. Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. *Molecular Biology and Evolution*, 30(3): 489–502. *Cité pages 22, 33, 38, 64, 168*

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. 2015. GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLOS Genetics*, 11(2): e1004941. *Cité page 24*

Latrille, T. and Lartillot, N. 2022. An Improved Codon Modeling Approach for Accurate Estimation of the Mutation Bias. *Molecular Biology and Evolution*, 39(2): msac005. *Cité page 135*

Latrille, T., Duret, L., and Lartillot, N. 2017. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736): 20160463. *Cité pages 10, 12, 13, 65*

Latrille, T., Lanore, V., and Lartillot, N. 2021. Inferring Long-Term Effective Population Size with Mutation–Selection Models. *Molecular Biology and Evolution*, 38(10): 4573–4587. *Cité pages 32, 135*

Latrille, T., Rodrigue, N., and Lartillot, N. 2023a. Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proceedings of the National Academy of Sciences*, 120(11): e2214977120. *Cité pages 30, 32, 134*

Latrille, T., Joseph, J., Hartasánchez, D. A., and Salamin, N. 2023b. Mammalian protein-coding genes exhibit widespread beneficial mutations that are not adaptive. Pages: 2023.05.03.538864 Section: New Results. *Cité pages 132, 136*

Lesecque, Y., Mouchiroud, D., and Duret, L. 2013. GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance. *Molecular Biology and Evolution*, 30(6): 1409–1419. *Cité pages 18, 25, 169, 171*

Levy Karin, E., Wicke, S., Pupko, T., and Mayrose, I. 2017. An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. *Systematic Biology*, 66(6): 917–933. *Cité page 176*

Li, R., Bitoun, E., Altemose, N., Davies, R. W., Davies, B., and Myers, S. R. 2019. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat Commun*, 10(1): 3900. *Cité pages 9, 11, 18, 25, 64, 72, 169, 170, 172, 174, 175*

Lichten, M. and Goldman, A. S. H. 1995. Meiotic recombination hotspots. *Annu. Rev. Genet.*, 29(1): 423–444. *Cité pages 9, 14, 63, 64, 65, 66, 68, 69, 73*

Lindholm, A. K., Dyer, K. A., Firman, R. C., Fishman, L., Forstmeier, W., Holman, L., Johannesson, H., Knief, U., Kokko, H., Larracuente, A. M., Manser, A., Montchamp-Moreau, C., Petrosyan, V. G., Pomiankowski, A., Presgraves, D. C., Safronova, L. D., Sutter, A., Unckless, R. L., Verspoor, R. L., Wedell, N., Wilkinson, G. S., and Price, T. A. R. 2016. The Ecology and Evolutionary Dynamics of Meiotic Drive. *Trends in Ecology & Evolution*, 31(4): 315–326. *Cité page 25*

Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., Yu, L., Liti, G., Tian, D., Hurst, L. D., and Yang, S. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat Ecol Evol*, 2(1): 164–173. *Cité page 19*

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W. R., Muscarella, M. E., Luo, H., Lennon, J. T., Brun, Y. V., and Lynch, M. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol*, 2(2): 237–240. *Cité page 172*

Lyttle, T. W. 1991. Segregation distorters. *Annu. Rev. Genet.*, 25(1): 511–581. *Cité pages 5, 6, 25*

Maloisel, L. and Rossignol, J.-L. 1998. Suppression of crossing-over by DNA methylation in Ascobolus. *Genes & Development*, 12(9): 1381–1389. *Cité pages 66, 67*

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203): 479–485. *Cité page 18*

Martin, G. and Lenormand, T. 2006. A General Multivariate Extension of Fisher's Geometrical Model and the Distribution of Mutation Fitness Effects Across Species. *Evolution*, 60(5): 893–907. *Cité page 131*

Martin, M. M., Ryan, M., Kim, R., Zakas, A. L., Fu, H., Lin, C. M., Reinhold, W. C., Davis, S. R., Bilke, S., Liu, H., Doroshow, J. H., Reimers, M. A., Valenzuela, M. S., Pommier, Y., Meltzer, P. S., and Aladjem, M. I. 2011. Genome-wide depletion of replication initiation events in highly transcribed regions. *Genome Research*, 21(11): 1822–1832. *Cité pages 68, 73*

Massip, F., Laurent, M., Brossas, C., Fernández-Justel, J., Gómez, M., Prioleau, M.-N., Duret, L., and Picard, F. 2019. Evolution of replication origins in vertebrate genomes: rapid turnover despite selective constraints. *Nucleic Acids Research*, 47(10): 5114–5125. *Cité page 68*

Mayrose, I. and Otto, S. P. 2011. A Likelihood Method for Detecting Trait-Dependent Shifts in the Rate of Molecular Evolution. *Molecular Biology and Evolution*, 28(1): 759–770. *Cité page 176*

McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328): 652–654. *Cité pages 133, 134*

McVicker, G. and Green, P. 2010. Genomic signatures of germline gene expression. *Genome Research*, 20(11): 1503–1511. *Cité page 69*

Meunier, J. and Duret, L. 2004. Recombination Drives the Evolution of GC-Content in the Human Genome. *Molecular Biology and Evolution*, 21(6): 984–990. *Cité pages 24, 33*

Mihola, O., Pratto, F., Brick, K., Linhartova, E., Kobets, T., Flachs, P., Baker, C. L., Sedlacek, R., Paigen, K., Petkov, P. M., Camerini-Otero, R. D., and Trachtulec, Z. 2019. Histone methyltransferase PRDM9 is not essential for meiosis in male mice. *Genome Res.*, 29(7): 1078–1086. *Cité pages 8, 13, 69, 73*

Mihola, O., Landa, V., Pratto, F., Brick, K., Kobets, T., Kusari, F., Gasic, S., Smagulova, F., Grey, C., Flachs, P., Gergelits, V., Tresnak, K., Silhavy, J., Mlejnek, P., Camerini-Otero, R. D., Pravenec, M., Petukhova, G. V., and Trachtulec, Z. 2021. Rat PRDM9 shapes recombination landscapes, duration of meiosis, gametogenesis, and age of fertility. *BMC Biol*, 19(1): 1–20. *Cité pages 8, 14, 69, 73*

Miller, D. E., Smith, C. B., Kazemi, N. Y., Cockrell, A. J., Arvanitakis, A. V., Blumenstiel, J. P., Jaspersen, S. L., and Hawley, R. S. 2016. Whole-Genome Analysis of Individual Meiotic Events in Drosophila melanogaster Reveals That Noncrossover Gene Conversions Are Insensitive to Interference and the Centromere Effect. *Genetics*, 203(1): 159–171. *Cité page 172*

Miyata, T. and Yasunaga, T. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*, 16(1): 23–36. *Cité page 29*

Moinet, A., Schlichta, F., Peischl, S., and Excoffier, L. 2022. Strong neutral sweeps occurring during a population contraction. *Genetics*, 220(4): iyac021. *Cité page 133*

Moses, M. J. 1956. Chromosomal Structures in Crayfish Spermatocytes. *The Journal of Biophysical and Biochemical Cytology*, 2(2): 215–218. *Cité page 8*

Mugal, C. F., Arndt, P. F., Holm, L., and Ellegren, H. 2015. Evolutionary Consequences of DNA Methylation on the GC Content in Vertebrate Genomes. *G3 Genes/Genomes/Genetics*, 5(3): 441–447. *Cité page 68*

Munch, K., Mailund, T., Dutheil, J. Y., and Schierup, M. H. 2014. A fine-scale recombination map of the human–chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Res.*, 24(3): 467–474. *Cité pages 10, 38*

Murakami, H. and Nurse, P. 2001. Regulation of premeiotic S phase and recombination-related double-strand DNA breaks during meiosis in fission yeast. *Nat Genet*, 28(3): 290–293. *Cité pages 68, 73*

Mustonen, V. and Lässig, M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3): 111–119. *Cité page 131*

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. 2011. GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (rice). *Molecular Biology and Evolution*, 28(9): 2695–2706. *Cité page 23*

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. 2005. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science*, 310(5746): 321–324. *Cité pages 9, 71*

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. 2010. Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science*, 327(5967): 876–879. *Cité pages 9, 64, 65*

Nagylaki, T. 1983. Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(20): 6278–6281. *Cité pages 18, 19, 21, 22, 26, 33, 131*

Necsulea, A., Sémon, M., Duret, L., and Hurst, L. D. 2009. Monoallelic expression and tissue specificity are associated with high crossover rates. *Trends in Genetics*, 25(12): 519–522. *Cité page 69*

Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., and Duret, L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mutation*, 32(2): 198–206. *Cité pages 26, 27, 33*

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5): 418–426. *Cité page 29*

Nicolas, A., Treco, D., Schultes, N. P., and Szostak, J. W. 1989. An initiation site for meiotic gene conversion in the yeast Saccharomyces cerevisiae. *Nature*, 338(6210): 35–39. *Cité pages 10, 64*

Nokkala, S., Kuznetsova, V. G., Maryanska-Nadachowska, A., and Nokkala, C. 2004. Holocentric chromosomes in meiosis. I. Restriction of the number of chiasmata in bivalents. *Chromosome Res*, 12(7): 733–739. *Cité page 8*

Ohta, K., Shibata, T., and Nicolas, A. 1994. Changes in chromatin structure at recombination initiation sites during yeast meiosis. *The EMBO Journal*, 13(23): 5754–5763. *Cité page 66*

Okamoto, H. and Hirochika, H. 2001. Silencing of transposable elements in plants. *Trends in Plant Science*, 6(11): 527–534. *Cité page 67*

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., Beatson, S. A., Lunter, G., Malik, H. S., and Ponting, C. P. 2009. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. *PLOS Genetics*, 5(12): e1000753. *Cité pages 13, 65, 69*

Otto, S. P. and Lenormand, T. 2002. Resolving the paradox of sex and recombination. *Nat Rev Genet*, 3(4): 252–261. *Cité page 3*

Pardo-Manuel de Villena, F. and Sapienza, C. 2001. Recombination is proportional to the number of chromosome arms in mammals. *Incorporating Mouse Genome*, 12(4): 318–322. *Cité page 8*

Partha, R., Kowalczyk, A., Clark, N. L., and Chikina, M. 2019. Robust Method for Detecting Convergent Shifts in Evolutionary Rates. *Molecular Biology and Evolution*, 36(8): 1817–1830. *Cité page 176*

Parvanov, E. D., Petkov, P. M., and Paigen, K. 2010. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, 327(5967): 835–835. *Cité pages 9, 64, 65*

Patel, A., Horton, J. R., Wilson, G. G., Zhang, X., and Cheng, X. 2016. Structural basis for human PRDM9 action at recombination hot spots. *Genes Dev.*, 30(3): 257–265. *Cité page 74*

Pease, J. B., Haak, D. C., Hahn, M. W., and Moyle, L. C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biology*, 14(2): e1002379. *Cité page 176*

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome biology and evolution*, 4(7): 675–682. *Cité pages 24, 38*

Petes, T. D. 2001. Meiotic recombination hot spots and cold spots. *Nat Rev Genet*, 2(5): 360–369. *Cité pages 9, 67*

Peñalba, J. V. and Wolf, J. B. W. 2020. From molecules to populations: appreciating and estimating recombination rate variation. *Nat Rev Genet*, 21(8): 476–492. *Cité page 9*

Ponting, C. P. 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Trends in Genetics*, 27(5): 165–171. *Cité pages 13, 65, 69*

Pouyet, F., Mouchiroud, D., Duret, L., and Sémon, M. 2017. Recombination, meiotic expression and human codon usage. *eLife*, 6: e27344. *Cité pages 69, 75*

Powers, N. R., Parvanov, E. D., Baker, C. L., Walker, M., Petkov, P. M., and Paigen, K. 2016. The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. *PLOS Genetics*, 12(6): e1006146. *Cité pages 9, 64, 66*

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., and Camerini-Otero, R. D. 2014. Recombination initiation maps of individual human genomes. *Science*, 346(6211): 1256442–1256442. *Cité pages 9, 72, 74*

Pratto, F., Brick, K., Cheng, G., Lam, K.-W. G., Cloutier, J. M., Dahiya, D., Wellard, S. R., Jordan, P. W., and Camerini-Otero, R. D. 2021. Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell*, 184(16): 4251–4267.e20. *Cité pages 68, 73*

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552): 2571–2580. *Cité page 33*

Raynaud, M., Gagnaire, P.-A., and Galtier, N. 2023. Performance and limitations of linkage-disequilibrium-based methods for inferring the genomic landscape of recombination and detecting hotspots: a simulation study. *Peer Community Journal*, 3. *Cité page 15*

Razeto-Barry, P., Díaz, J., and Vásquez, R. A. 2012. The Nearly Neutral and Selection Theories of Molecular Evolution Under the Fisher Geometrical Framework: Substitution Rate, Population Size, and Complexity. *Genetics*, 191(2): 523–534. *Cité page 131*

Renkawitz, J., Lademann, C. A., and Jentsch, S. 2014. Mechanisms and principles of homology search during recombination. *Nat Rev Mol Cell Biol*, 15(6): 369–383. *Cité pages 8, 11, 72*

Robinson, M. C., Stone, E. A., and Singh, N. D. 2014. Population Genomic Analysis Reveals No Evidence for GC-Biased Gene Conversion in Drosophila melanogaster. *Molecular Biology and Evolution*, 31(2): 425–433. *Cité page 173*

Rodrigue, N. and Lartillot, N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7): 1020–1021. *Cité page 135*

Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10): 4629–4634. *Cité page 135*

Roman, H. 1985. Gene conversion and crossing-over. *Environmental Mutagenesis*, 7(6): 923–932. *Cité pages 4, 17*

Roze, D. 2021. A simple expression for the strength of selection on recombination generated by interference among mutations. *PNAS*, 118(19). *Cité page 5*

Roze, D. 2023. Causes and consequences of linkage disequilibrium among transposable elements within eukaryotic genomes. *Genetics*, 224(2): iyad058. *Cité pages 67, 72*

Samuk, K. and Noor, M. A. F. 2022. Gene flow biases population genetic inference of recombination rate. *G3 Genes|Genomes|Genetics*, 12(11): jkac236. *Cité page 15*

Schield, D. R., Pasquesi, G. I. M., Perry, B. W., Adams, R. H., Nikolakis, Z. L., Westfall, A. K., Orton, R. W., Meik, J. M., Mackessy, S. P., and Castoe, T. A. 2020. Snake Recombination Landscapes Are Concentrated in Functional Regions despite PRDM9. *Molecular Biology and Evolution*, 37(5): 1272–1294. *Cité pages 14, 38, 73*

Schlichta, F., Peischl, S., and Excoffier, L. 2022. The Impact of Genetic Surfing on Neutral Genomic Diversity. *Molecular Biology and Evolution*, 39(11): msac249. *Cité page 133*

Schultes, N. P. and Szostak, J. W. 1991. A Poly(dA · dT) Tract Is a Component of the Recombination Initiation Site at the ARG4 Locus in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 11(1): 322–329. *Cité pages 10, 64*

Schwarzkopf, E. J. and Cornejo, O. E. 2022. PRDM9-directed recombination hotspots depleted near meiotically transcribed genes. *Gene*, 813: 146123. *Cité pages 69, 70*

Sella, G. and Hirsh, A. E. 2005. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27): 9541–9546. *Cité pages 130, 131*

Shin Voo, K., Carlone, D. L., Jacobsen, B. M., Flodin, A., and Skalnik, D. G. 2000. Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1. *Molecular and Cellular Biology*, 20(6): 2108–2121. *Cité page 14*

Shiu, P. K. T., Raju, N. B., Zickler, D., and Metzenberg, R. L. 2001. Meiotic Silencing by Unpaired DNA. *Cell*, 107(7): 905–916. *Cité page 67*

Silander, O. K., Tenaillon, O., and Chao, L. 2007. Understanding the Evolutionary Fate of Finite Populations: The Dynamics of Mutational Effects. *PLOS Biology*, 5(4): e94. *Cité page 131*

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., McVean, G., and Przeworski, M. 2015. Stable recombination hotspots in birds. *Science*, page 6. *Cité pages 14, 64, 66, 68*

Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343): 375–378. *Cité page 16*

Smagulova, F., Brick, K., Pu, Y., Camerini-Otero, R. D., and Petukhova, G. V. 2016. The evolutionary turnover of recombination hot spots contributes to speciation in mice. *Genes Dev.*, 30(3): 266–280. *Cité pages 9, 10, 38, 64, 65, 71, 72, 73, 75*

Smeds, L., Mugal, C. F., Qvarnström, A., and Ellegren, H. 2016. High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS Genetics*, 12(5): e1006044. *Cité page 18*

Smith, J. M. and Haigh, J. 1973. The hitch-hiking effect of a favourable gene. *Genetics Research*, page 13. *Cité page 28*

Sommermeyer, V., Béneut, C., Chaplais, E., Serrentino, M. E., and Borde, V. 2013. Spp1, a Member of the Set1 Complex, Promotes Meiotic DSB Formation in Promoters by Tethering Histone H3K4 Methylation Sites to Chromosome Axes. *Molecular Cell*, 49(1): 43–54. *Cité pages 14, 67, 72, 73*

Soni, V., Moutinho, A. F., and Eyre-Walker, A. 2022. Changing Population Size in McDonald–Kreitman Style Analyses: Artifactual Correlations and Adaptive Evolution between Humans and Chimpanzees. *Genome Biology and Evolution*, 14(2): evac022. *Cité page 134*

Spence, J. P. and Song, Y. S. 2019. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10): eaaw9206. *Cité page 15*

Spielman, S. J. and Wilke, C. O. 2015. The Relationship between dN/dS and Scaled Selection Coefficients. *Molecular Biology and Evolution*, 32(4): 1097–1108. *Cité page 29*

Spies, M. and Fishel, R. 2015. Mismatch Repair during Homologous and Homeologous Recombination. *Cold Spring Harb Perspect Biol*, 7(3): a022657. *Cité page 24*

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., and Smadja, C. M. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil. Trans. R. Soc. B*, 372(1736): 20160455. *Cité page 8*

Stephan, W. 2019. Selective Sweeps. *Genetics*, 211(1): 5–13. *Cité page 133*

Subramanian, S. 2019. Population size influences the type of nucleotide variations in humans. *BMC Genet*, 20(1): 1–12. *Cité page 168*

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. 1983. The double-strand-break repair model for recombination. *Cell*, 33(1): 25–35. *Cité page 8*

Tamuri, A. U. and dos Reis, M. 2022. A Mutation–Selection Model of Protein Evolution under Persistent Positive Selection. *Molecular Biology and Evolution*, 39(1): msab309. *Cité page 135*

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. 2012. Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. *Genetics*, 190(3): 1101–1115. *Cité page 135*

Tataru, P., Mollion, M., Glémin, S., and Bataillon, T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics*, 207(3): 1103–1119. *Cité pages 30, 134*

Taylor, D. R. and Ingvarsson, P. K. 2003. Common Features of Segregation Distortion in Plants and Animals. *Genetica*, 117(1): 27–35. *Cité pages 5, 25*

Tenaillon, O. 2014. The Utility of Fisher's Geometric Model in Evolutionary Genetics. *Annu. Rev. Ecol. Evol. Syst.*, 45(1): 179–201. *Cité page 136*

Tenaillon, O., Silander, O. K., Uzan, J.-P., and Chao, L. 2007. Quantifying Organismal Complexity using a Population Genetic Approach. *PLOS ONE*, 2(2): e217. *Cité page 131*

Teng, W., Liao, B., Chen, M., and Shu, W. 2022. Genomic Legacies of Ancient Adaptation Illuminate GC-Content Evolution in Bacteria. *Microbiology Spectrum*, 11(1): e02145–22. *Cité page 25*

Thomas, J. H., Emerson, R. O., and Shendure, J. 2009. Extraordinary Molecular Evolution in the PRDM9 Fertility Gene. *PLOS ONE*, 4(12): e8505. *Cité page 13*

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. 2021. Genome-wide association studies. *Nat Rev Methods Primers*, 1(1): 1–21. *Cité page 175*

Vigué, L., Croce, G., Petitjean, M., Ruppé, E., Tenaillon, O., and Weigt, M. 2022. Deciphering polymorphism in 61,157 Escherichia coli genomes via epistatic sequence landscapes. *Nat Commun*, 13(1): 4030. *Cité page 136*

Wallberg, A., Glémin, S., and Webster, M. T. 2015. Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, Apis mellifera. *PLoS Genet*, 11(4): e1005189. *Cité pages 9, 66, 74*

Webster, M. T. and Hurst, L. D. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in Genetics*, 28(3): 101–109. *Cité pages 9, 69, 74*

Webster, M. T., Axelsson, E., and Ellegren, H. 2006. Strong Regional Biases in Nucleotide Substitution in the Chicken Genome. *Molecular Biology and Evolution*, 23(6): 1203–1216. *Cité pages 24, 33*

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1): 67–72. *Cité page 136*

Weinberg, W. 1908. Uber den Nachweis der Vererbung beim Menschen. *Jh. Ver. vaterl. Naturk. Wurttemb.*, 64: 369–382. *Cité page 5*

Weiner, A., Zauberman, N., and Minsky, A. 2009. Recombinational DNA repair in a cellular context: a search for the homology search. *Nat Rev Microbiol*, 7(10): 748–755. *Cité page 8*

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., and Przeworski, M. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4: e04637. *Cité page 18*

Winkler, H. 1930. Die Konversion der Gene : eine vererbungstheoretische Untersuchung. *G Fischer*. *Cité pages 4, 17*

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *The sixth international congress of genetics*, vol I: pp 356–366. *Cité page 129*

Wu, T.-C. and Lichten, M. 1994. Meiosis-Induced Double-Strand Break Sites Determined by Yeast Chromatin Structure. *Science*, 263(5146): 515–518. *Cité page 66*

Wu, T. C. and Lichten, M. 1995. Factors that affect the location and frequency of meiosis-induced double-strand breaks in Saccharomyces cerevisiae. *Genetics*, 140(1): 55–66. *Cité page 66*

Yamada, S., Kim, S., Tischfield, S. E., Jasin, M., Lange, J., and Keeney, S. 2017. Genomic and chromatin features shaping meiotic double-strand break formation and repair in mice. *Cell Cycle*, 16(20): 1870–1884. *Cité page 72*

Yang, P., Wang, Y., and Macfarlan, T. S. 2017. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics*, 33(11): 871–881. *Cité page 67*

Yuan, S., Huang, T., Bao, Z., Wang, S., Wu, X., Liu, J., Liu, H., and Chen, Z.-J. 2022. The histone modification reader ZCWPW1 promotes double-strand break repair by regulating cross-talk of histone modifications and chromatin accessibility at meiotic hotspots. *Genome Biol*, 23(1): 187. *Cité page 66*

Zamudio, N., Barau, J., Teissandier, A., Walter, M., Borsos, M., Servant, N., and Bourc'his, D. 2015. DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev.*, 29(12): 1256–1270. *Cité pages 67, 68*

Úbeda, F. and Wilkins, J. F. 2011. The Red Queen theory of recombination hotspots. *Journal of Evolutionary Biology*, 24(3): 541–553. *Cité pages 10, 65*