Increased positive selection in highly recombining genes does not necessarily reflect an evolutionary advantage of recombination

Julien Joseph[1]

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

julien.joseph@ens-lyon.fr

April 8, 2024

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. 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 studies assessed whether selection was more effective in highly recombining genes (due to dissipation of genetic linkage) or less effective (due to gBGC), assuming a fixed distribution of fitness effects (DFE) for all genes. In this study, I directly derive the DFE from a gene's evolutionary history (shaped by mutation, selection, drift and gBGC) under empirical fitness landscapes. I show that genes that have experienced high levels of gBGC are less fit and thus have more opportunities for beneficial mutations. Only a small decrease in the genome-wide intensity of gBGC leads to the fixation of these beneficial mutations, particularly in highly recombining genes. This results in increased positive selection in highly recombining genes that is not caused by more effective selection. Additionally, I show that the death of a recombination hotspot can lead to a higher dN/dS than its birth, but with substitution patterns biased towards AT, and only at selected positions. 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. Finally, although gBGC does not affect the fixation probability of GC-conservative mutations, I show that by altering the DFE, gBGC can also significantly affect non-synonymous GC-conservative substitution patterns.

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

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; Murga-Moreno et al., 2019; Grandaubert et al., 2019; Castellano et al., 2020; Cavassim et al., 2021). Some studies claim a beneficial effect of recombination (Gossmann et al., 2014; Hussin et al., 2015; Castellano et al., 2016; Corcoran et al., 2017; Murga-Moreno et al., 2019; Castellano et al., 2020), some claim no effect of recombination on the efficacy of selection (Bullaughey et al., 2008), and others claim a deleterious effect (Necşulea et al., 2011; Lachance and Tishkoff, 2014; Bolívar et al., 2016).

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; Grandaubert et al., 2019; Murga-Moreno et al., 2019; Castellano et al., 2020; Cavassim et al., 2021).

The DFE of new mutations 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).

Results

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 amino-acid 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-coding 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 (Hartl and Taubes, 1996; Mustonen and Lässig, 2009; Jones et al., 2017). 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).

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 all the DFEs presented in this study, I considered only non-synonymous mutations. To be able to compare the results obtained on different 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).

First, I computed equilibrium frequencies with a Jukes Cantor mutation matrix (which assumes that all bases have equal mutation rates). It appears that 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).

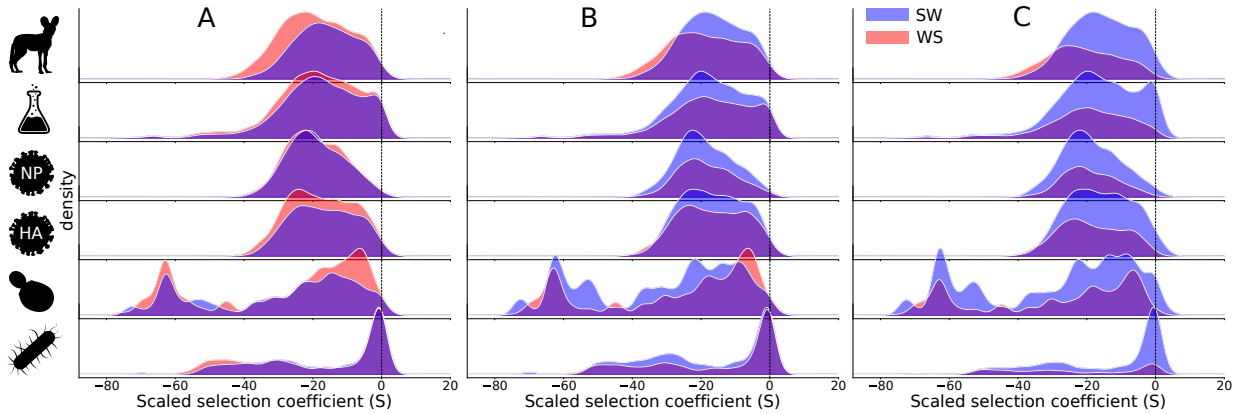

Figure 1: Distribution of fitness effects of new mutations at equilibrium separately for WS (red) and SW (blue) mutations. Equilibrium frequencies were computed for each fitness landscape under A) $B = 0$ and a JC mutation matrix, B) $B = 0$ and a human mutation matrix and C), $B = 2$ and a human mutation matrix. 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).

	B = 0 JC			**B = 0 HM**			**B = 2 HM**		
Fitness landscape	**All**	**WS**	**SW**	**All**	**WS**	**SW**	**All**	**WS**	**SW**
Mammalian	1.33%	1.13%	1.48%	1.37%	1.62%	1.11%	1.64%	0.35%	2.97%
Concatenated DMS	3.04%	2.68%	3.33%	3.02%	3.66%	2.54%	3.22%	0.79%	5.39%
NP (Influenza)	0.61%	0.68%	0.54%	0.72%	1.30%	0.42%	0.71%	0.15%	1.27%
HA (Influenza)	1.42%	1.15%	1.74%	1.51%	1.7%	1.38%	1.73%	0.29%	3.35%
Gal4 (*S.cervisae*)	1.57%	1.33%	1.83%	1.37%	1.81%	0.97%	1.93%	0.43%	3.84%
β-lactamase (*E.coli*)	11.9%	10.8%	12.4%	11.6%	12.7%	10.3%	12.11%	4.47%	16.2%

Table 1: Proportion of beneficial mutations under $B = 0$ and a Jukes Cantor (JC) mutation matrix, $B = 0$ and a human mutation matrix (HM) and $B = 2$ and a human mutation matrix

Using a realistic human mutation matrix, where mutations are typically biased towards AT, the number of SW mutations increases as expected (Fig.1). Importantly, this increase is not uniform along the DFE, and mostly concerns the deleterious compartment (Fig.1). Indeed, beneficial mutation towards AT have already been fixed, and only mutations that are so deleterious that they cannot be fixed remain as mutational opportunities. Although the number of expected SW and WS beneficial back mutations is roughly the same, it is important to note that the proportion of beneficial WS mutations among all WS is higher than the proportion of beneficial SW mutations among all SW (Table 1). This means that WS mutations are on average more beneficial than SW ones. This result has already been found with a similar approach by Latrille and Lartillot (2022) and Kaj et al. (2023).

When the sequence has evolved under gBGC ($B = 2$, corresponding to a region of high recombination rate in the human genome estimated by Glémin et al. (2015)), the SW mutation rate is still higher because of the mutational bias, but given that gBGC has increased GC content, the number of SW mutational opportunities is also higher. This results in even more SW mutations (Fig.1). Moreover, the DFE of WS mutations is skewed towards negative values, while the DFE of SW mutations is skewed towards positive values (Fig.1 and Table 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). Of note even if SW mutations are more beneficial, they do not fix more often because of gBGC, and remain as mutational opportunities. 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). Interestingly, Glémin (2010) showed that gBGC cancels the fixation load induced by the mutational bias towards AT (λ) when $B = ln(\lambda)$, and that this value does not depend on the distribution of selective effects (which is analogous to a fitness landscape in bi-allelic models). For the human mutational bias, it means a B close to 0.7. However, this is assuming that half optimal states are GC and half are AT. As shown here, the genetic code (and potentially other factors) increases the proportion of AT optimal states, which already alleviates the load induced by the mutational bias. In this case, the value of B that perfectly compensate the mutational bias is expected to be even lower than 0.7.

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 $B = 2$ using the realistic human mutation matrix. 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 dynamics 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.2C 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;

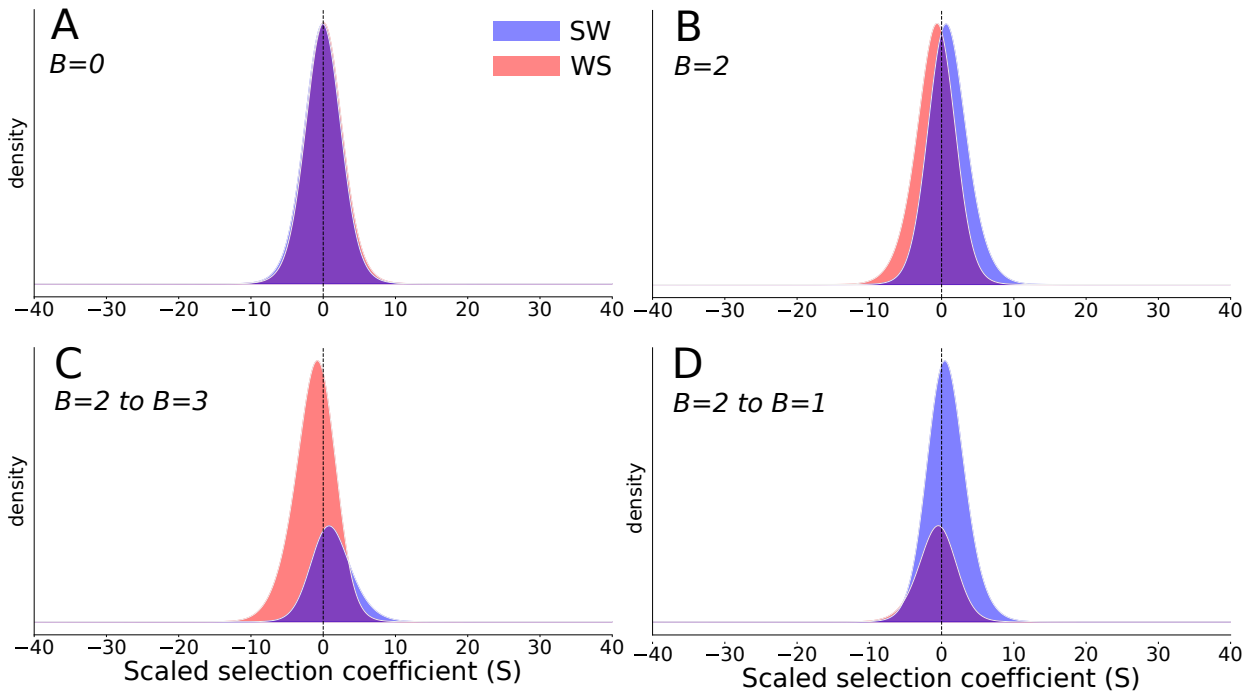

Figure 2: Distribution of fitness effects of new substitutions separately for WS (red) and SW (blue) substitutions under the concatenated DMS fitness landscape. A) Equilibrium frequencies were computed with $B = 0$ and the substitution rate also with $B = 0$ (no gBGC). B) Equilibrium frequencies were computed with $B = 2$ and the substitution rate also with $B = 2$. C) Equilibrium frequencies were computed with $B = 2$ and the substitution rate with $B = 3$ (increase in gBGC). D) Equilibrium frequencies were computed with $B = 2$ and the substitution rate 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.

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.2D and Fig.S2).

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 scenarios 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).

In population genetics, positive selection can be quantified by the parameter α, introduced by McDonald and Kreitman (1991), which is an estimator the proportion of positively selected substitutions (for all muta-

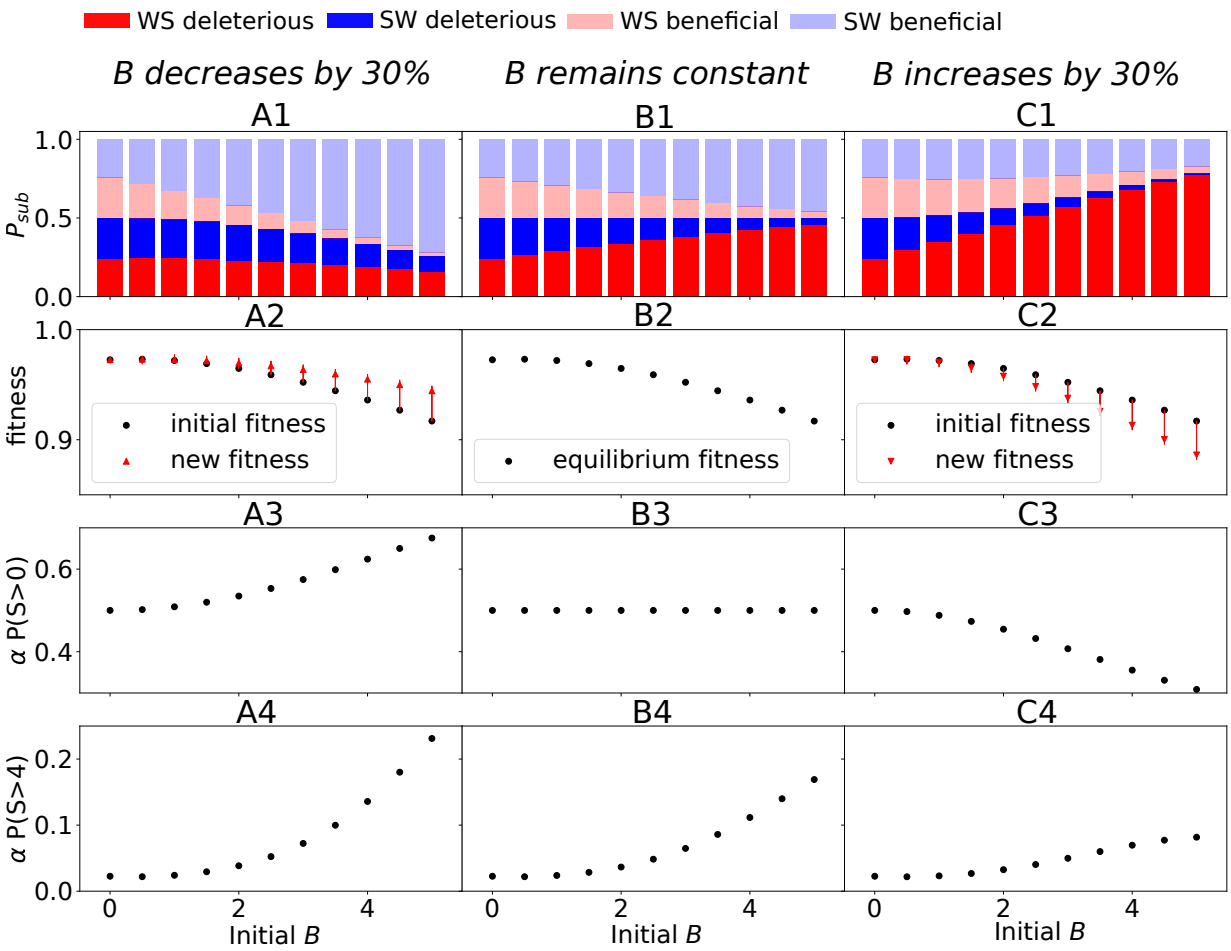

Figure 3: Row 1: Proportion of the substitutions (P_{sub}) contributed by WS deleterious (bright red), SW deleterious (bright blue), WS beneficial (light red) and SW beneficial (light blue) mutations as a function of the population-scaled gBGC coefficient under the concatenated DMS fitness landscape 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.

tion types) among all substitutions. Here, because the fitness landscape is known, the proportion of positively selected substitutions is directly accessible from the DFE of substitutions. We can therefore investigate the effect of gBGC on the true proportion of positively selected substitutions that McDonald-Kreitman based methods aim to estimate. 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 sequence 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 substitutions that become strongly beneficial (Figure 3B4). When gBGC decreases, the increase of α with gBGC is higher than at equilibrium. This is because beneficial substitutions compensate the fixation of strongly deleterious mutations 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 substitutions is still higher than without gBGC, where almost every substitution 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.

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 does 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 represents 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).

First, 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

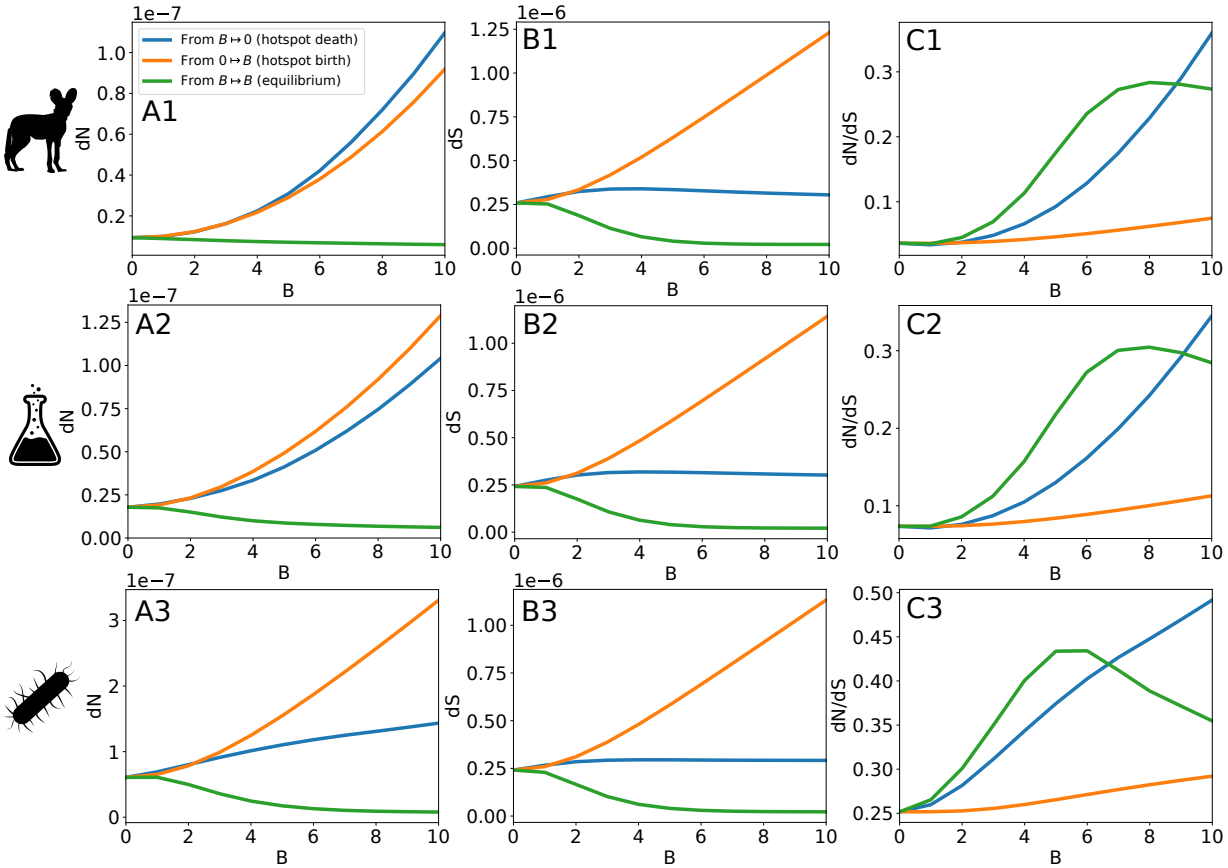

Figure 4: dN (A), dS (B), and dN/dS (C) as a function of the population scaled gBGC coefficient B in three scenarios: Equilibrium codon frequencies are computed without gBGC, and substitutions subsequently accumulate under a population-scaled gBGC coefficient of B (orange line), mimicking the birth of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

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 hotspot (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 is 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.

Non-synonymous GC-conservative substitution patterns are affected by gBGC

To avoid the confounding effect of gBGC on the dN/dS ratio, several studies computed dN/dS and $\pi N/\pi S$ using GC-conservative substitutions/polymorphisms only ($G \leftrightarrow C$ and $A \leftrightarrow T$) (Bolívar et al., 2016; Corcoran et al., 2017; Rousselle et al., 2019; Castellano et al., 2020). The idea being that as gBGC only distorts the segregation ratio of AT:GC heterozygous polymorphisms, the GC-conservative dN/dS should be an gBGC-robust estimator of the efficacy of purifying selection. However, although gBGC does not affect the fixation probability of GC-conservative mutations, by influencing the equilibrium sequence, it might affect mutational opportunities and therefore patterns of GC-conservative substitutions.

To explore this possibility, I also computed the DFE of WW and SS mutations with or without gBGC under all fitness landscapes. It appears that WW mutations are on average more deleterious in the presence of gBGC (B=2) under all fitness landscapes (Table S1). This is due to the fact that under gBGC, weakly constrained sites that are AT optimal are fixed in a GC state. Therefore, the only remaining opportunities for WW mutations are from sites that are sufficiently constrained to have resisted gBGC, and therefore these WW mutations are more deleterious on average. The consequence of this is a decrease of the WW dN/dS with recombination (Fig.5C and Fig.S6C). Importantly, since gBGC does not affect the fixation probability of WW mutations, this decrease depends only on past gBGC activity and not on the current value of gBGC. On the other hand, the impact of gBGC on the DFE of SS mutations is weak and depends on the fitness landscape (Table S1). In AT-optimal sites that have been fixed in a GC state because of gBGC, the DFE of SS mutations will depend on the fitness difference between the least deleterious GC-rich codon (favoured by selection) and the other deleterious GC-rich codons, which is not easily predictable and might differ depending on the fitness landscape. As a consequence the relationship between the SS dN/dS and recombination is weaker and sometimes non-monotonous (Fig.5C and Fig.S6C). As for WW substitutions, SS substitution patterns only depend on past gBGC activity. Overall, under realistic fitness landscapes, a slight decrease of the GC-conservative dN/dS ratio can in principle be caused by gBGC only (especially for WW mutations). Therefore, a negative correlation between recombination rate and the GC-conservative dN/dS does not necessarily reflect an increased efficacy of purifying selection as proposed by Bolívar et al. (2016), Corcoran et al. (2017), Rousselle et al. (2019) and Castellano et al. (2020). Importantly, under the assumptions of the model, although the rates of non-synonymous WW and SS substitutions are affected by gBGC, the proportions of positively selected WW and SS substitutions remain fairly unaffected (Fig.S7).

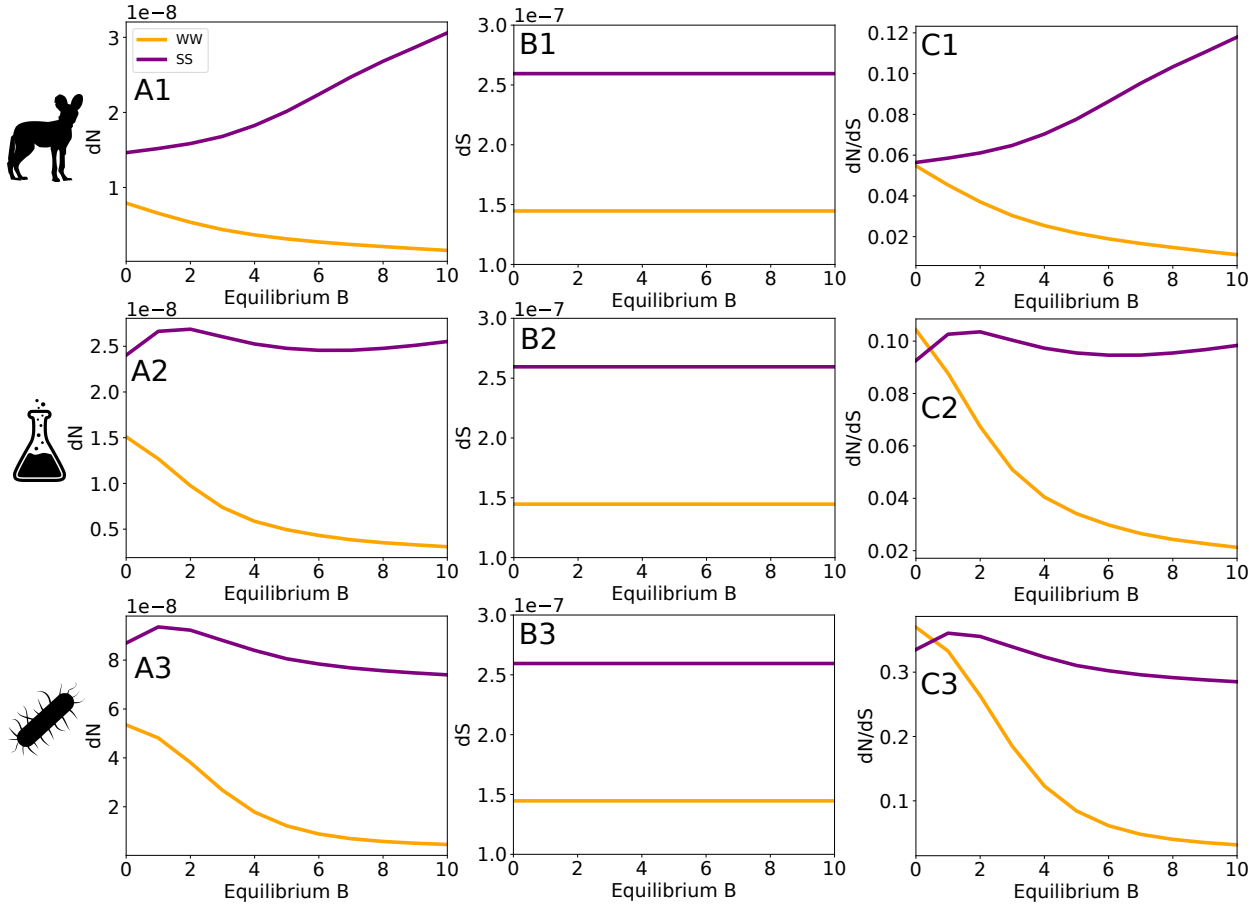

Figure 5: dN (A), dS (B), and dN/dS (C) as a function of the equilibrium population-scaled gBGC coefficient B for WW and SS substitutions. From top to bottom: 2000 sites randomly sampled from the mammalian fitness landscapes, the concatenate of the DMS fitness landscapes (1389 sites) and the *E.coli* protein β-lactamase (263 sites).

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 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. These results strengthen 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; Rousselle et al., 2019). Of note, I show that the GC-conservative dN/dS, previously thought to be a gBGC-robust estimator of the efficacy of natural selection, is in fact also affected by gBGC.

Moreover, 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 or environmental changes).

This being said, it is still very possible that Hill-Robertson interference dissipation in highly recombining genes helps to fix beneficial mutations. In this sense, some studies have reported a positive correlation between measures of positive selection and recombination in *Drosophila melanogaster* (Castellano et al., 2016; Murga-Moreno et al., 2019), where the contrast between lowly recombining genes and highly recombining ones is higher than in humans (Bullaughey et al., 2008), and where the current levels of gBGC seem to be very low (Robinson et al. (2014) but see Jackson and Charlesworth (2021)).

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 in which 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 (they tend to have different functions, see Pouyet et al. (2017)) and different mutational processes (they might be more susceptible to CpG hyper-mutability) that are not directly linked to gBGC or recombination.

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; Clément and Arndt, 2013; Lesecque et al., 2014). 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, it is very likely that these short-lived hotspots do not have time to reach equilibrium, and the signatures of the death of a hotspot in the dN/dS may not be strong enough to be observed in individual hotspots in these species. Of note, those hotspot still remain long enough so that they leave a clear gBGC footprint in substitutions across mammals (Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010; Clément and Arndt, 2013; Lartillot, 2013; Lesecque et al., 2014; Galtier, 2021). 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; Joseph et al., 2023). Interestingly, in mammals these recombination hotspots co-evolve slowly with DNA methylation (Joseph et al., 2023). When those long-lived hotspots stop to be hypomethylated, they usually die (Joseph et al., 2023). 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.

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; Fukushima and Pollock, 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). A recent study suggests 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.

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; Grandaubert et al., 2019; Hämälä and Tiffin, 2020; Castellano et al., 2020; Cavassim et al., 2021). In presence of gBGC, it is thus important to account for beneficial back-mutations.

Even without gBGC, this work is consistent with previous theoretical studies showing 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 does not necessarily reflect a beneficial effect of recombination, a fixation bias towards GC at non-synonymous sites does not necessarily reflect 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.

Material and methods

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). Individual proteins with peculiar functions can have peculiar fitness landscapes which are not necessarily representative of the genome. To have a beginning of an average fitness landscape, I also performed analyses on a concatenate of the four DMS fitness landscape.

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 $q_{i,j}$:

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

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

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

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.

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:

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

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.

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 and Thibault Latrille for insightful discussions on mutation-selection models and Nicolas Lartillot, Carina Farah Mugal, Laurent Duret and Bastien Boussau 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 are available at https://gitlab.in2p3.fr/julien.joseph/backdfe

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.

Cavassim, M. I. A., Andersen, S. U., Bataillon, T., and Schierup, M. H. (2021). Recombination Facilitates Adaptive Evolution in Rhizobial Soil Bacteria. *Molecular Biology and Evolution*, 38(12):5480–5490.

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.

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.

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.

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.

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.

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.

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.

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.

Grandaubert, J., Dutheil, J. Y., and Stukenbrock, E. H. (2019). The genomic determinants of adaptive evolution in a fungal pathogen. *Evolution Letters*, 3(3):299–312.

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.

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).

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.

Joseph, J., Prentout, D., Laverré, A., Tricou, T., and Duret, L. (2023). High prevalence of Prdm9-independent recombination hotspots in placental mammals.

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.

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.

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

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., 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.

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.

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.

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.

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.

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.

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.

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.