



**HAL**  
open science

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

Augustin Clessin, Julien Joseph, Nicolas Lartillot

► **To cite this version:**

Augustin Clessin, Julien Joseph, Nicolas Lartillot. The evolution of GC-biased gene conversion by means of natural selection. 2024. hal-04757128

**HAL Id: hal-04757128**

**<https://hal.science/hal-04757128v1>**

Preprint submitted on 28 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

---

# THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

---

Augustin Clessin<sup>1</sup>, Julien Joseph<sup>1</sup> , Nicolas Lartillot<sup>1</sup> 

<sup>1</sup>Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, UMR 5558, Villeurbanne, France

Authors for Correspondence: [julien.joseph@ens-lyon.fr](mailto:julien.joseph@ens-lyon.fr) ; [nicolas.lartillot@univ-lyon1.fr](mailto:nicolas.lartillot@univ-lyon1.fr)

June 21, 2024

## Abstract

1 GC-biased gene conversion (gBGC) is a recombination-associated evolutionary process that  
2 biases the segregation ratio of AT:GC polymorphisms in the gametes of heterozygotes, in  
3 favour of GC alleles. This process is the major determinant of variation in base composition  
4 across the human genome and can be the cause of a substantial burden of GC deleterious  
5 alleles. While the importance of GC-biased gene conversion in molecular evolution is in-  
6 creasingly recognised, the reasons for its existence and its variation between species remain  
7 largely unknown. Using simulations and semi-analytical approximations, we investigated  
8 the evolution of gBGC as a quantitative trait evolving by mutation, drift and natural se-  
9 lection. We show that in a finite population where most mutations are deleterious, gBGC  
10 is under weak stabilising selection around a positive value that mainly depends on the in-  
11 tensity of the mutation bias and on the intensity of selective constraints exerted on the  
12 genome. Importantly, the levels of gBGC that evolve by natural selection do not minimize  
13 the load in the population, and even increase it substantially in regions of high recombi-  
14 nation rate. Therefore, despite reducing the population's fitness, levels of gBGC that are  
15 currently observed in humans could in fact have been (weakly) positively selected.

16 **Keywords** gBGC · Recombination · Modifier · Genetic load · Mutation bias · Natural selection

## 17 1 Introduction

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

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

late 80's [Brown and Jiricny \(1987\)](#) found that in human and green monkey cells, DNA repair was biased towards GC alleles. Since then, direct and indirect evidence has revealed that repair biases operate on meiotic gene conversion events in a wide range of eukaryotes, leading to a biased transmission of GC alleles to offspring ([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 genome-wide 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](#); [Joseph, 2024](#)). Moreover, it has been demonstrated, notably in humans and chickens, 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  $\leftrightarrow$  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 of gBGC ([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. In this direction, [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. This result gives a stronger theoretical basis to the idea that gBGC could evolve as a consequence of an AT mutation bias. On the other hand, the study was conducted under the approximation of infinite population sizes and at a single strongly selected locus. As such, it does not provide an explanation for the variation in the strength of gBGC between species of different effective 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](#)

61 Uyenoyama, 1990). As expected, the equilibrium value of the transmission bias towards GC depends both on  
62 the intensity of the mutational bias towards AT and on the magnitude of selective constraints exerted on the  
63 genome. Interestingly, we predict that the equilibrium value of the transmission bias correlates negatively  
64 with effective population size. Importantly, we show that even if gBGC leads to a higher deleterious burden  
65 at the population level, this does not mean that it is negatively selected, even in high  $N_e$  species. In the  
66 present model, high gBGC intensity results from the short-term advantage of converting AT deleterious  
67 alleles in heterozygotes, which leads to a higher deleterious burden in the population. Overall, by capturing  
68 the selective pressures acting on gBGC under empirically realistic conditions, this model provides insight  
69 into the role of natural selection in shaping the evolution of gBGC in eukaryotes.

## 70 2 Results

### 71 Model summary

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

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

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

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

98 **Biased gene conversion is under stabilizing selection**

99 Typical trajectories of the population-mean of biased gene conversion ( $b$ ) under the model are shown in  
100 Figure 1. Here, a mutational bias of  $\lambda = 3$  is considered (bias in favour of Weak, or AT alleles), with a basal  
101 mutation rate of  $u = 10^{-4}$  (for  $W$  to  $S$  mutations), a population size of  $N = 1000$ , a genome consisting  
102 of  $L = 10000$  selected loci, with a gamma distribution of selective effects of mean  $\bar{h}s = 0.01$  and shape  
103 0.2. Two alternative settings are considered for the dominance effect of those mutations: either co-dominant  
104 ( $h = 0.5$ ) or partially recessive ( $h = 0.1$ ). In both cases, the modifier locus undergoes mutations at a rate of  
105  $w = 10^{-3}$  per generation, with effect sizes of mean 0.1 on  $\beta$ . Finally, the basal gene conversion rate is equal  
106 to  $\alpha = 0.1$ . Of note, these parameter values are not meant, at that stage, to match any specific empirical  
107 situation. Instead, the aim is to reveal the inner workings of the model, and how its output relates to the  
108 input parameter values.

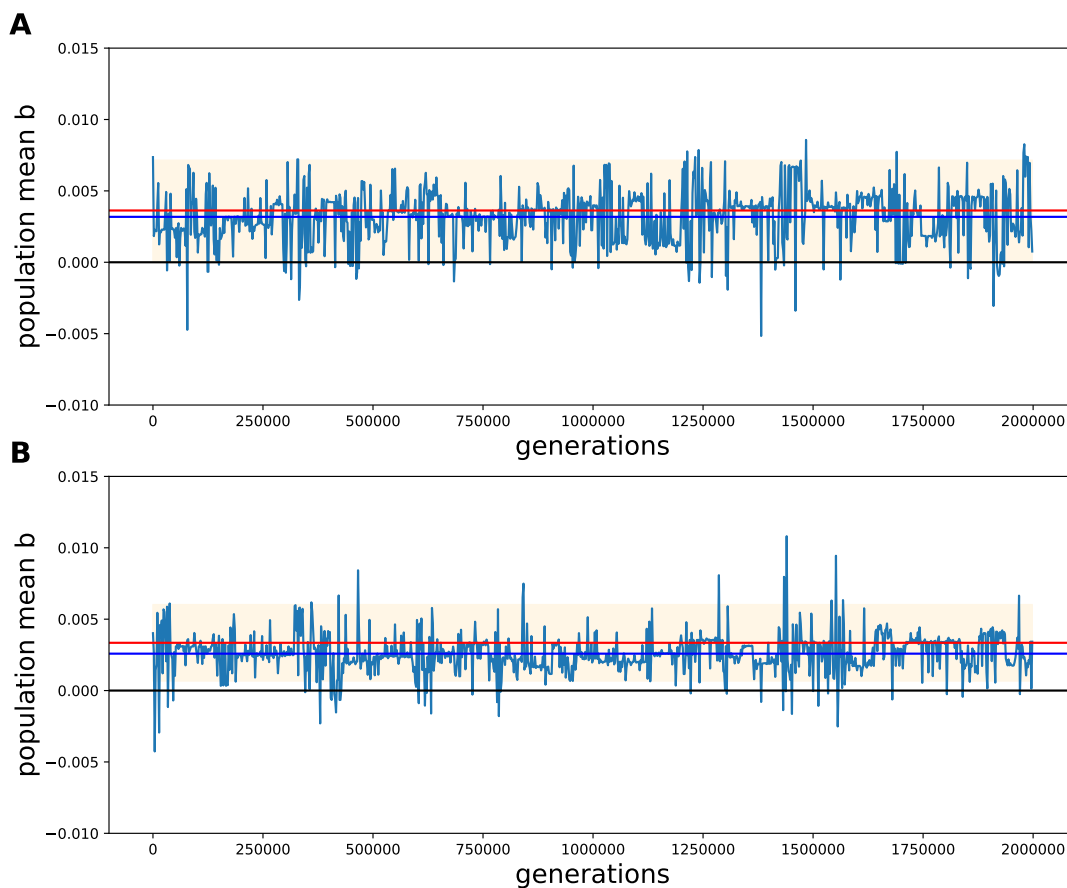


Figure 1: Evolution of the strength of gBGC (mean of  $b$  over the population) over the generations, for the co-dominant (A) and recessive (B) cases. Blue horizontal line: mean over the entire run; red horizontal line: equilibrium value predicted by the analytical approximation; shaded area: predicted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

109 Running the model under these parameter values results in a population-level gBGC evolving towards  
110 positive value of  $b$ , reaching an evolutionary equilibrium with a long-term mean of the order of  $b \simeq 0.003$   
111 (Figure 1). There is a substantial evolutionary variance, such that the population still spends about 5%  
112 of the time with negative values of  $b$ . Nevertheless, these experiments show that gBGC is susceptible to  
113 spontaneously evolve in favour of Strong (GC) alleles. They also more specifically suggest the existence  
114 of some form of stabilizing selection acting on gBGC, driving the population towards, and maintaining it  
115 around, an evolutionary equilibrium.

## 116 **The mutation-segregation tradeoff between AT- and GC-deleterious mutations**

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

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

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

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

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

142 If  $G(b)$  is positive, then modifiers increasing  $b$  will be favoured, and conversely if  $G(b)$  is negative. Considering  
143  $W$  and  $S$  alleles separately,  $G(b)$  can be expressed as the difference between the net gain upon converting

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

144  $W$ -deleterious alleles  $G_W(b)$ , and the cost of converting  $S$ -deleterious alleles,  $G_S(b)$ :

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

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

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

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

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

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

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

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

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

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

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

$$H_W(s, b) = \mathbb{E}_{\phi_{s,b}^W} [P \times C] \quad (7)$$

163 The same derivation can be conducted in the case of  $S$ -deleterious loci. For bi-allelic loci, the equilibrium  
164 distributions  $\phi_{s,b}^W(x)$  and  $\phi_{s,b}^S(x)$ , for both  $W$  and  $S$  loci, can be explicitly written, up to a normalization  
165 constant, such that expectations over these distributions can be computed numerically (see methods).

### 166 Predicting the equilibrium mean and variance strength of gBGC

167 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  
168 function is decreasing, crossing 0 at an intermediate, positive value of  $b^*$ . Numerically solving for the value  
169  $b^*$  such that  $G(b^*) = 0$  gives  $b^* = 0.0036$  in the co-dominant case, and  $b^* = 0.0033$ , which is close to  
170 the mean value observed in the simulation ( $\bar{b} = 0.0032$  and  $\bar{b} = 0.0026$ , respectively). The numerical and  
171 simulation-based estimates are both represented as a blue and red lines, respectively, in Figure 1.

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

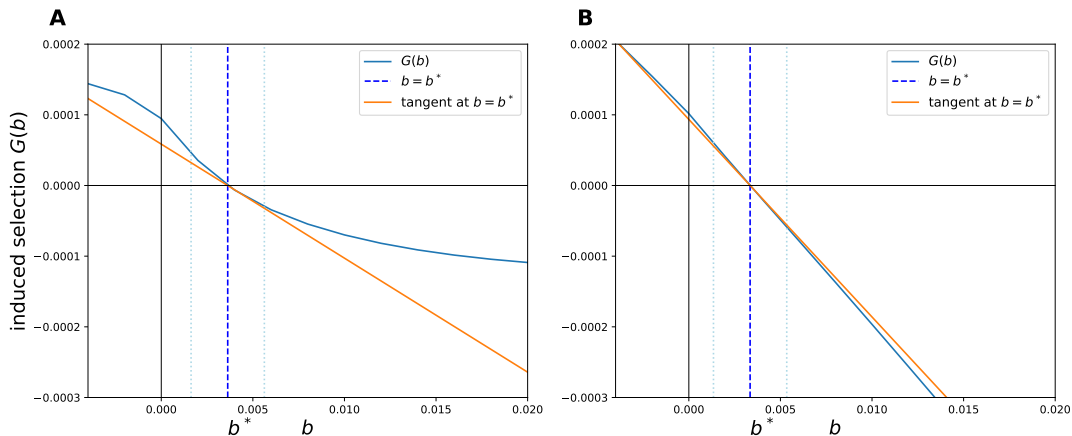


Figure 2: Strength of selection induced on gBGC modifiers,  $G(b)$ , as a function of  $b$  (blue curve), under the co-dominant (A) and recessive (B) settings. Dark blue dotted vertical line: numerically estimated value of  $b^*$ , for which  $G(b) = 0$ ; orange line: numerically estimated tangent at  $b^*$ ; light blue dotted vertical lines: predicted standard deviation around  $b^*$ .

172 A rough quantitative estimate of the evolutionary variance can also be obtained, based on the slope  $\gamma$  of  
 173 the tangent to the curve at  $b^*$  (Figure 2A). Specifically, the equilibrium evolutionary variance is predicted  
 174 to be approximately equal to  $v_{eq} \simeq \frac{1}{2NL\gamma}$  (reported as a shaded area on Figure 1). Of note, the selective  
 175 response shows a steeper slope at the equilibrium point in the recessive case, resulting in a smaller predicted  
 176 evolutionary variance than in the co-dominant case.

177 **The drivers of gBGC**

178 The behaviour of the simulation model, along with the analytical approximation just introduced, were further  
 179 investigated by plotting the predicted equilibrium value of the strength of gBGC,  $b^*$ , as a function of several  
 180 key parameters (mutation bias, mean strength of purifying selection, number of positions under selection  
 181 and mutation rate). The case of the response of  $b^*$  to changes in effective population size is examined further  
 182 below.

183 Not surprisingly, the mean equilibrium strength of gBGC is directly related to the strength of the mu-  
 184 tational bias (Figure 3A). Owing the symmetry of the problem, running the model with  $\lambda < 1$ , i.e. under  
 185 a mutational bias in favour of the Strong alleles results in a population evolving towards a mean conversion  
 186 bias in favour of Weak (left side of Figure 3A). The mean equilibrium strength of biased gene conversion is  
 187 also directly influenced by the mean strength of the purifying selection acting over the genome (Figure 3B),  
 188 thus clearly indicating that its evolutionary dynamics is a direct consequence of the selective effects induced  
 189 by converting non-neutral polymorphisms in the germ-line. The mean equilibrium value is insensitive to  
 190 the number  $L$  of selected loci, but its evolutionary variance, on the other hand, is affected, showing a clear  
 191 decreasing trend with  $L$ , which corresponds to the scaling in  $1/L$  predicted by the analytical approximation  
 192 (Figure 3C).



THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

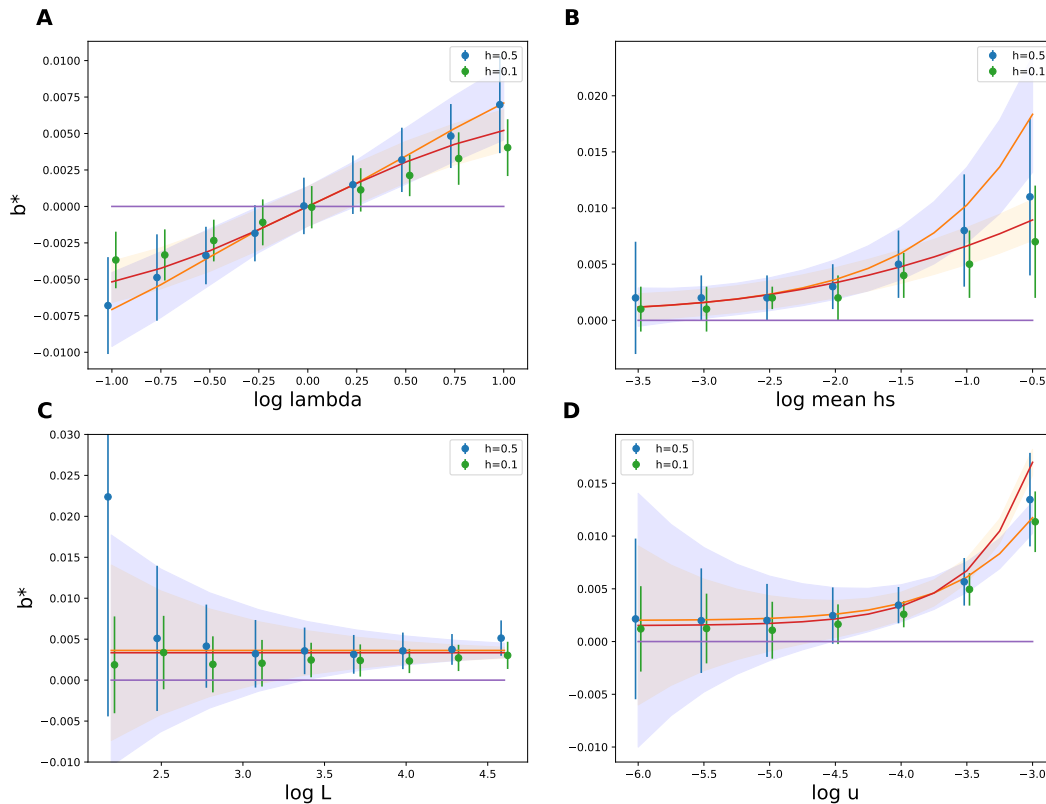


Figure 3: Mean equilibrium  $b^*$  and standard deviation, as a function of mutation bias  $\lambda$  (A), mean selective effect  $h\bar{s}$  (B), number of selected positions in the genome  $L$  (C) and mutation rate  $u$  (D), under the co-dominant (blue) and the recessive (orange) case, obtained by simulations (dots and associated vertical bars) and predicted by the analytical approximation (curve and associated shaded area), under a mutation rate of  $w = 10^{-4}$  ( $Nw = 0.1$ ).

193 Finally, the strength of gBGC responds very weakly to the mutation rate, except for very high mutation  
 194 rates ( $4Nu \gg 1$ ), in which case it shows a sharp increase (Figure 3D). For low  $4Nu$ , not so much the  
 195 mean than the evolutionary variance of gBGC is impacted by the mutation rate, with larger variances being  
 196 observed under lower mutation rates. In this respect, the response of the model to variation in  $u$  is not unlike  
 197 its response to variation in  $L$  (Figure 3C). This similar behaviour can be understood by noting that any  
 198 indirect selective effect acting on the modifier locus can only be mediated by heterozygous positions. Thus,  
 199 the strength of induced selection will be directly determined, not just to  $L$ , but more fundamentally, by the  
 200 mean number of selected positions at which a typical individual is heterozygous. The mean heterozygosity  
 201 in the population is in turn directly impacted by the mutation rate, and this, under most selective regimes.

202 The analytical approximation (plain lines in Figure 3) is globally in good agreement with the simulation  
 203 results (filled circles), except for large  $\bar{s}$  or large  $u$ , where it underestimates the equilibrium strength of gBGC.

204 However, these corresponds to regimes where the diffusive approximation used for deriving the analytical  
205 predictions is not valid, owing to a large variance in genome-wide log-fitness between individuals. In practice,  
206 these regimes are far from empirical reasonable conditions.

207 Finally, across all scaling experiments shown in Figure 3, the stabilizing selection induced on  $b$  appears  
208 to be globally tighter in the partially recessive case, for which both the response of the equilibrium value of  
209  $b$  to changes in parameter values and the equilibrium variance are less pronounced than in the co-dominant  
210 case.

### 211 Which class of mutations contribute to stabilizing selection on gBGC ?

212 As the mean fitness effect of deleterious mutations is a key parameter for the evolution of intermediate  
213 levels of gBGC, it appears probable that under a distribution of fitness effects (DFE), not all mutations  
214 contribute equally to it. To further investigate this point, the analytical approximation was recruited to  
215 examine how the mean frequency at which  $W$ -deleterious alleles (Figure 4A&B) and  $S$ -deleterious alleles  
216 (Figure 4C&D) segregate in a population is modulated by slight variations in  $b$  (the dotted, plain, and dashed  
217 lines correspond to increasingly larger values of  $b$ ). The bottom panels show the corresponding expected  
218 fitness gain  $H_W(s, b)$  incurred by converting  $W$ -deleterious alleles (blue curves, above 0), and the expected  
219 fitness cost  $H_S(s, b)$  incurred by converting  $S$ -deleterious alleles (red curves, below 0), both weighted by  
220 the distribution of selective effects (DFE). These are plotted as functions of  $s$ , for 3 different values of  
221  $b$ . Weighting  $H$  by the DFE gives a better sense of the relative contributions of mutations with different  
222 selection coefficients to the total cost and gain. Also, with this weighting, averaging  $H_W$  and  $H_S$  over the  
223 DFE simply amounts to computing the area under the two curves, which thus directly correspond to  $G_W(b)$   
224 and  $G_S(b)$ , respectively. The parameter values used for Figure 4 correspond to the simulation trajectory  
225 displayed in Figure 1, for the co-dominant and recessive cases.

226 As  $b$  increases,  $W$  alleles segregate at a lower frequency (Figure 4A&B) and  $S$  alleles at a higher frequency  
227 (Figure 4C&D). Correlatively, the expected gain contributed by converting  $W$  alleles (Figure 4E&F, blue  
228 curves) decreases, and the cost contributed by converting  $S$  alleles (red curves) increases with population-  
229 level  $b$ . The intermediate value of  $b$  used in Figure 4 is precisely the one for which the areas under the two  
230 curves in panel C are equal (shaded areas in blue and red) – it is thus the predicted evolutionary equilibrium  
231 (of note, the areas under the two curves may not look equal to each other on the figure, in particular in the  
232 recessive case, but this is only because the two curves extend much further to the right than is shown).

233 Importantly, the way the two compartments, Weak and Strong, react to changes in population-level  
234  $b$  is very different. On one side,  $W$ -deleterious polymorphisms are only moderately affected, and this,  
235 mostly in the range of small selective effects. In contrast  $S$ -deleterious polymorphisms are strongly affected.  
236 More specifically, increasing  $b$  leads to a surge in the segregation of  $S$ -deleterious mutations of intermediate  
237 strength, for which gBGC and selection are of the same order of magnitude. This surge translates into a peak  
238 in the expected cost (red curve, bottom panel), whose area increases with  $b$ . Translating these observations  
239 in terms of the net selection acting on gBGC, the fitness advantage is mostly contributed by converting  $W$   
240 strongly-deleterious mutations, and is essentially a constant. The fitness cost, on the other hand, is mostly

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

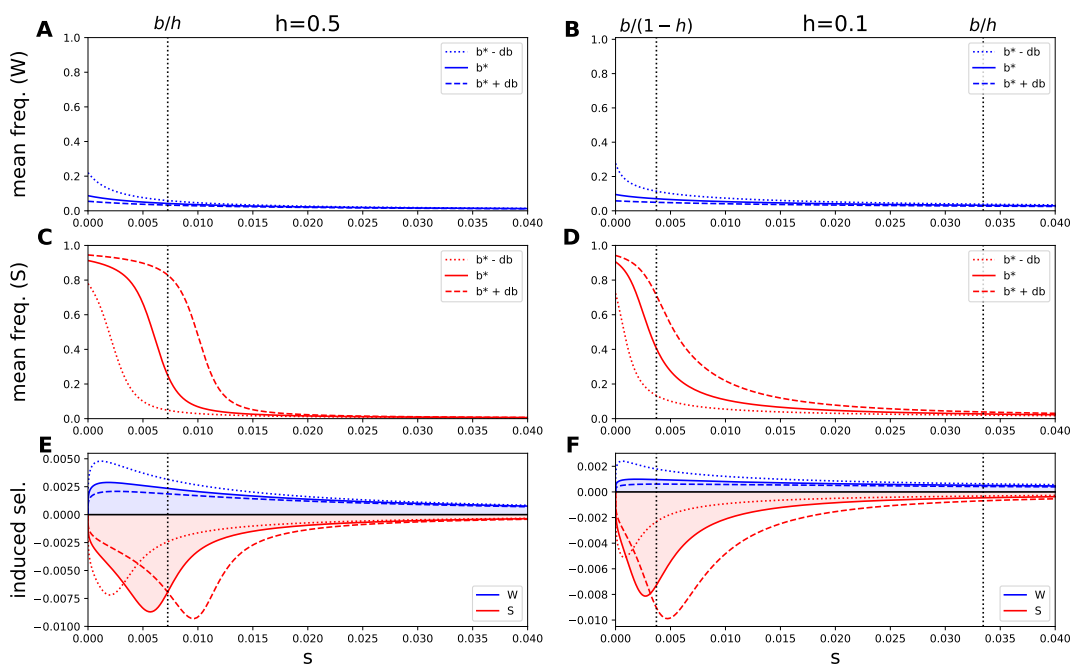


Figure 4: Mean segregation frequency of  $W$  alleles (A&B),  $S$  alleles (C&D), and induced selection (E&F), as a function of  $s$ , under the co-dominant (A,C&E) and recessive (B,D&F) settings, for different values of  $b$ : plain lines correspond to the equilibrium value of  $b$ , dashed lines to a slightly increased  $b$ , and dotted lines to a slightly decreased  $b$ . Blue lines correspond to  $W$  alleles, while red lines correspond to  $S$  alleles.

241 contributed by  $S$  mutations of selective effects of the order of  $b$ . This fitness cost varies strongly with  $b$  and  
 242 is the main factor responsible for modulating the selection induced on gBGC modifiers, as a function of the  
 243 population  $b$ .

244 Of note, the exact patterns differ between the co-dominant and the recessive cases. In the co-dominant  
 245 case, the peak in the conversion cost is around  $s \simeq b/h$ , the value for which gBGC and selection exactly  
 246 compensate each other. In the recessive case, the region that contributes to the increased cost when gBGC  
 247 increases is between  $b/(1-h) \leq s \leq b/h$ , or equivalently  $hs \leq b \leq (1-h)s$ . This is the range for which the  
 248 strength of gBGC is stronger than selection against the heterozygote but weaker than selection against the  
 249 homozygote for the deleterious mutant. As a result, these GC-deleterious polymorphisms tend to segregate  
 250 at intermediate frequencies, as if they were over-dominant (i.e. advantageous when in one copy, deleterious  
 251 when in two copies), resulting in a higher fraction of heterozygotes in the population, and thus a substantial  
 252 cost against gBGC (Glémin, 2010). This can explain why the strength of selection around the equilibrium  
 253 value of  $b$  is higher in the partly recessive case.

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

254 **gBGC is partially buffered against changes in population size**

255 A somewhat paradoxical consequence of gBGC is the extreme sensitivity of equilibrium base composition  
256 to even mild variation in its population-scaled intensity  $B = 4Nb$  (Eyre-Walker, 1999). Quantitatively, the  
257 neutral equilibrium GC/AT composition ratio scaling exponentially with  $B$ , which can quickly lead to very  
258 large GC content even for moderate increase in  $N$ . For instance, based on the current estimate of  $b$  in  
259 humans, increasing effective population size by a factor 10 would imply a long-term neutral equilibrium GC  
260 content greater than 99% in the 10% most highly recombining fraction of the genome. How to explain, then,  
261 that gBGC does not more often lead to diverging base composition across species?

262 Implicit in the argument just exposed is that the strength of gBGC is fixed, while population size varies, or  
263 at least, that there is no internal mechanism for tuning the raw intensity of gBGC ( $b$ ) depending on effective  
264 population size ( $N$ ), so as to somehow guarantee that  $B = 4Nb$  never becomes too large. Yet, if gBGC  
265 is under stabilizing selection, this raises the possibility for such an internal mechanism to spontaneously  
266 emerge. This fundamentally depends on how the evolutionary optimum  $b^*$  scales with population size.

267 To examine this point, the optimal value  $b^*$  predicted by the model was computed (using the semi-  
268 analytical approximation) over a broad range of values of  $N$  between  $10^2$  and  $10^6$ . For this experiment, a  
269 mutation rate of  $u = 10^{-8}$  was assumed (for  $S \rightarrow W$  mutations), and a bias of  $\lambda = 2$ . Both the co-dominant  
270 case ( $h = 0.5$ ) and the partially recessive case ( $h = 0.1$ ) were considered.

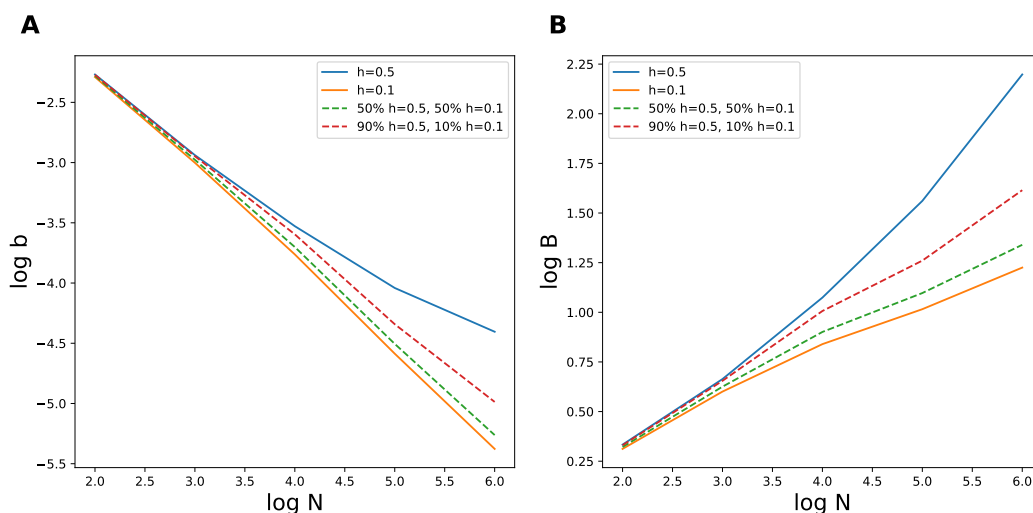


Figure 5: Scaling of  $b^*$  (A) and  $B^* = 4Nb^*$  (B) as a function of  $N$ , under the co-dominant ( $h = 0.5$ ) and recessive ( $h = 0.1$ ) settings (plain curves), or assuming a mixture of co-dominant and recessive mutations (dashed curves).

271 In all cases (Figure 5), whether co-dominant or recessive,  $b^*$  decreases with  $N$ . The trend is moderate  
272 in the co-dominant case but more pronounced in the recessive case. In both cases, the decrease is less  
273 than linear, such that  $B = 4Nb$  still increases as a function of  $N$ . This increase is quite substantial in  
274 the co-dominant case, with  $B$  reaching values above 10 for population sizes of  $N = 10^4$  and above 100 for

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

275  $N > 3.10^5$ . In the recessive case, on the other hand,  $B$  is much less responsive to changes in population  
 276 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  
 277 magnitude for  $N$ .

278 Interestingly, a mixture of 50% co-dominant and 50% partially recessive ( $h = 0.1$ ) essentially behaves like  
 279 the pure partially recessive case (all mutations with  $h = 0.1$ ). Even a small proportion of 10% of partially  
 280 recessive positions, mixed with 90% of co-dominant positions, shows substantially more stable levels of gBGC  
 281 as a function of  $N$  (Figure 5, dashed lines). Recessive mutations thus appear to represent an efficient buffer  
 282 against changes in population-scaled gBGC induced by changes in population size.

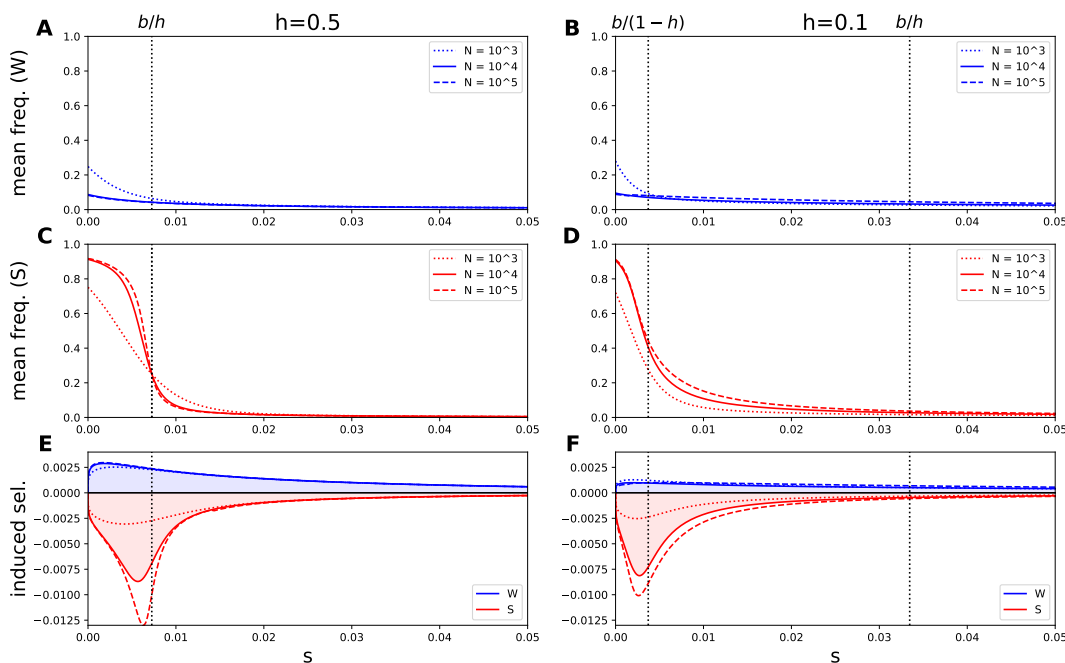


Figure 6: Mean segregation frequency of  $W$  alleles (A&B),  $S$  alleles (C&D), and induced selection (E&F), as a function of  $s$ , under the co-dominant (A,C&E) and recessive (B,D&F) settings, for different values of  $N$ : plain lines correspond to the equilibrium value of  $b$ , dashed lines to a slightly increased  $N$ , and dotted lines to a slightly decreased  $N$ . Blue lines correspond to  $W$  alleles, while red lines correspond to  $S$  alleles.

283 The fundamental reason why  $b^*$  decreases with  $N$  can be understood by examining the structure of the  
 284 induced selective response (Figure 6). As mentioned above, the mutation-segregation balance essentially  
 285 takes the form of a tradeoff between, on one side, a net error-correcting effect on strongly deleterious muta-  
 286 tions (more often deleterious towards  $W$  than towards  $S$ ) and, on the other side, a conversion load mostly  
 287 contributed by  $S$ -deleterious mutations with selection coefficients of the order of  $b$ . The first component,  
 288 being in the strong selection regime, is essentially insensitive to  $N$  (Figure 6E&F, blue curves). The sec-  
 289 ond component, on the other hand, precisely because of the compensation between gBGC and selection,  
 290 is effectively in a regime dominated by drift, and thus, in many respects, has an evolutionary dynamics

291 resembling nearly-neutral evolution. As such, its mean heterozygosity is strongly influenced by changes in  
292 effective population size, and more precisely, will tend to increase with  $N$ . Since biased gene conversion  
293 is in direct proportion to the amount of heterozygosity, the conversion cost itself will also increase with  $N$   
294 (Figure 6E&F, red curves). Altogether, GC-deleterious mutations with selective effects of the order of  $b$  are  
295 efficiently mobilized (i.e. contribute more to standing variation) upon an increase in  $N$  and thus represent  
296 a key force buffering  $B^*$  against changes in population size.

297 Of note, and as already explored above (Figure 4), in the co-dominant case (Figure 6, A,C&E), the range  
298 of GC-deleterious mutations that are mobilized consists of a relatively narrow peak around  $b/h$ . In contrast,  
299 in the recessive case, a good fraction of the range comprised between  $b/(1-h)$  and  $b/h$  (the two dotted  
300 vertical lines on Figure 6B,D&F), corresponding to the co-dominant regime, is mobilized, thus contributing  
301 a much more responsive buffer against changes in  $N$  – which can easily dominate the overall response even if  
302 recessive mutations represent a minority of the total standing variation, as observed above (Figure 6, dashed  
303 lines).

#### 304 **gBGC and the genetic load**

305 gBGC is often depicted as a force that interferes with selection, and that causes a significant deleterious  
306 burden (Galtier and Duret, 2007; Berglund et al., 2009; Galtier et al., 2009; Necşulea et al., 2011). On  
307 the other hand, our results reflect those of previous studies showing that BGC confers a significant fitness  
308 advantage by correcting the most common class of mutations (here  $S \mapsto W$ ) (Bengtsson and Uyenoyama,  
309 1990). But as pointed out by Glémin (2010), the levels of gBGC that evolve naturally are not necessarily  
310 the ones that minimize the average genetic load of a population. The average genetic load of a population  
311 can be decomposed into the load of  $W$  deleterious alleles:

$$L_W = \langle \mathbb{E}_{\phi_{s,b}^W} [2hsx(1-x) + sx^2] \rangle, \quad (8)$$

312 where  $x$  is a random variable drawn from the equilibrium frequency distribution of  $W$  alleles  $\phi_{s,b}^W(x)$ , and  
313 that of  $S$  deleterious alleles:

$$L_S = \langle \mathbb{E}_{\phi_{s,b}^S} [2hsx(1-x) + sx^2] \rangle, \quad (9)$$

314 where  $x$  is a random variable drawn from the equilibrium frequency distribution of  $S$  alleles  $\phi_{s,b}^S(x)$ . In both  
315 cases, and as above, the inner expectation is on  $x$  and the outer expectation (angle brackets) is on  $s$  (drawn  
316 from the gamma DFE).

317 Using a population size of  $N = 10,000$ , a mutation rate of  $u = 10^{-8}$  and a mean selection coefficient  
318  $hs = 0.01$ , we computed the average genetic load as a function of  $b$  for  $h = 0.5$  and  $h = 0.1$  (Figure 7A&B).  
319 The load is minimized for a very small value of  $b$  compared to the one that naturally evolves. The level  
320 of gBGC that minimizes the average genetic load corresponds to the level that equalizes the frequencies of  
321  $W$  and  $S$  alleles, leading to an average GC content of 0.5 (Figure 7C&D). It is worth noting that when  
322 the average frequencies of  $W$  and  $S$  alleles are equal, it does not mean that they are distributed evenly  
323 in heterozygotes. In fact,  $W$  deleterious alleles are more often heterozygous (Figure 7 E&F), because they  
324 are numerous due to a high  $S \mapsto W$  mutation rate, but at low frequency because of gBGC. Conversely,  $S$   
325 deleterious alleles are less numerous due to a high  $S \mapsto W$  mutation rate, but more often at high frequency

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

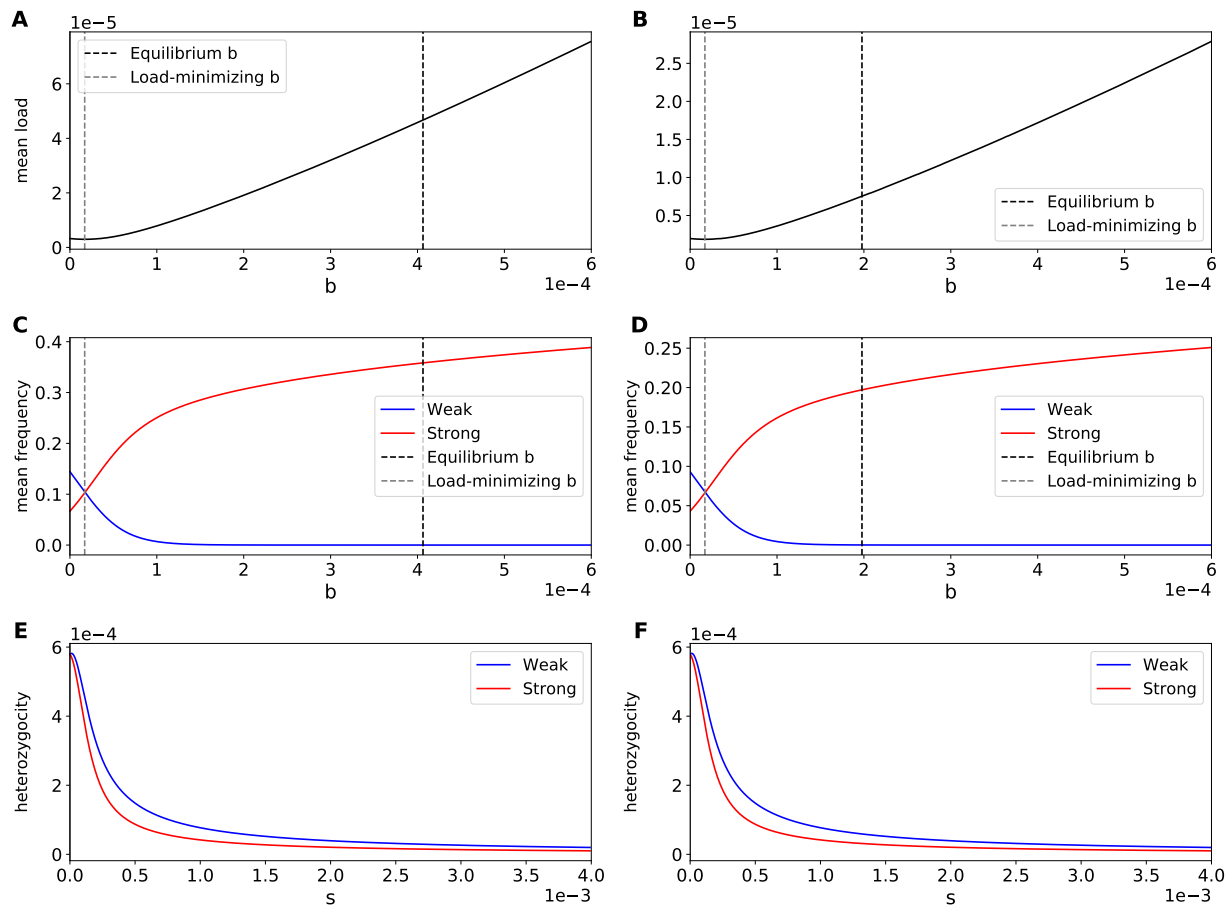


Figure 7: A&B: average deleterious load in a population as a function of  $b$ . The black lines shows the value equilibrium value of  $b$  and the grey line shows the value of  $b$  that minimizes the average load. C&D: frequency of W and S alleles as a function of  $b$ . E&F: heterozygosity for W and 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:  $h = 0.1$

326 because of gBGC and thus more often homozygous. Therefore, when the mean load in the population is  
 327 minimal, there still is an individual advantage to convert W deleterious alleles more often for heterozygotes.

328 **Empirical calibration**

329 The modeling work presented thus far suggests that biased gene conversion in favour of GC can in principle  
 330 evolve as a consequence of the mutation bias towards AT and that its intensity can also be modulated in  
 331 an adaptive manner as a function of key parameters, in particular effective population size. An important  
 332 question that remains is whether the model provides quantitatively reasonable predictions when confronted  
 333 to current empirical knowledge about the strength of gBGC in various species.

334 Humans and the mouse represent good cases to consider. To account for the substantial heterogeneity in  
 335 recombination rates across the genome of these two species, the theoretical calculations were adapted so as  
 336 to allow for a mixture of gBGC strength, whose mean is under the control of the modifier. Quantitatively,

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

		$h\bar{s} = 0.01$				$h\bar{s} = 0.05$			
$h$	$N$	$B_h^*$	stdev	$p(B < 0)$	$B_h^*$	stdev	$p(B < 0)$		
	$10^4$	1.5	0.9	0.050	3.3	1.4	0.009		
0.5	$10^5$	4.8	1.7	0.003	11.8	0.9	< 0.001		
	$10^6$	18.6	0.9	< 0.001	58.3	0.6	< 0.001		
	$10^4$	0.8	0.4	0.029	1.1	0.4	0.006		
0.1	$10^5$	1.2	0.4	0.003	1.5	0.4	< 0.001		
	$10^6$	2.8	0.7	< 0.001	3.6	0.8	< 0.001		
50% 0.1 : 50% 0.5	$10^4$	0.9	0.5	0.028	1.3	0.5	0.004		
	$10^5$	1.5	0.5	0.002	1.9	0.5	< 0.001		
	$10^6$	3.8	0.9	< 0.001	5.2	1.1	< 0.001		
10% 0.1 : 90% 0.5	$10^4$	1.2	0.7	0.033	1.8	0.6	0.002		
	$10^5$	2.3	0.7	< 0.001	3.0	0.7	< 0.001		
	$10^6$	7.8	2.3	< 0.001	10.5	1.0	< 0.001		

Table 1: Numerical estimates of scaled intensity of gBGC  $B^* = 4Nb^*$  (mean over the genome), equilibrium standard deviation and probability of a negative gBGC, for different parameter values for  $N$ ,  $h$ ,  $h\bar{s}$ .

337 based on the estimate that about 90% of the recombination is concentrated in about 10% of the genome  
338 (Smagulova et al., 2011; Pratto et al., 2014), we assume that 10% of the genome experiences a 100 times  
339 stronger gBGC than the remaining 90%. For the other parameters, assuming that  $\sim 10\%$  of the genome is  
340 under selection (Rands et al., 2014), for a genome of total size 3 Gb, this gives  $L = 3.10^8$  positions. In  
341 humans, the mutation rate is  $u = 3.10^{-8}$  per base pair and per generation. In the mouse, the mutation rate  
342 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  
343 of  $\lambda = 2$ , the value used here. Current estimates of the DFE suggest a shape parameter  $a$  between 0.15 and  
344 0.25 (Eyre-Walker et al., 2006; Castellano et al., 2019). Here, we used three values for  $a$ : 0.1, 0.2 and 0.3.  
345 The mean selection coefficient under this DFE is difficult to estimate. In humans, recent estimates are of the  
346 order of  $h\bar{s} = 0.01$  to 0.05, both of which were tried in what follows. Finally, the co-dominant and partially  
347 recessive cases are considered, as well as the 50:50 and 90:10 mixtures of these two dominance regimes, with  
348 population sizes varying from  $N = 10^4$  to  $N = 10^6$ , so as to cover most of the range of what can be expected  
349 more generally in mammals.

350 The estimates of  $B^* = 4Nb^*$  (mean over the genome) under these parameter values are reported in Table  
351 1. Under co-dominant selection, the predicted values for  $B^*$  range from 1.5 to 124, showing quite some  
352 sensitivity to effective population size, mean selection strength across the genome, and shape of the DFE  
353 (Table 1 and Table S2&3). In contrast, assuming partially recessive mutations returns a much narrower  
354 range of estimates, from 0.8 to 6. Fitting the model assuming a mix of co-dominant and recessive mutations  
355 (last rows of Table 1) suggests that a moderate fraction of recessive mutations is sufficient to make  $B$  less  
356 responsive to changes in  $N$ .



357 Empirical estimates of gBGC in Humans are of the order of  $B = 0.3$  for the genome-wide mean, and  
358 around  $B = 5.2$  to  $6.5$  in the top 20% regions of high recombination (Duret and Arndt, 2008). The theoretical  
359 predictions (Table 1) are globally higher than these empirical estimates, although they get reasonably close  
360 to them (predicted  $B$  of the order of 1) for the lower values of  $\bar{s}$  or assuming the presence of recessive  
361 mutations. This, together with the rather extreme results obtained for the largest population size under the  
362 co-dominant case, suggest that recessive mutations may play a role in buffering gBGC.

363 Finally, the predicted evolutionary variance is small, although not negligible for low population size, for  
364 which  $b^*$  is predicted to be negative around 1 to 5% of the time. This suggests that induced selection on  
365 gBGC may not always be sufficiently powerful to guarantee a bias towards Strong over the whole range of  
366 molecular evolutionary regimes susceptible to be observed across mammals – although, even then, it may  
367 still represent a sufficiently strong selective force preventing gBGC to become unreasonably large.

### 368 **3 Discussion**

369 In this study, we developed a model to characterize the evolution of gBGC by means of natural selection. We  
370 first showed that, in the presence of a mutation bias towards AT, gBGC was under relatively weak stabilizing  
371 selection around a positive value of the transmission bias, in agreement with a previous study (Bengtsson and  
372 Uyenoyama, 1990). The equilibrium value of the transmission bias ( $b^*$ ) corresponds to the one that equalizes  
373 the fitness gain of converting strongly deleterious AT mutations in heterozygotes and the fitness cost of  
374 transmitting slightly deleterious GC mutations to offsprings. This balance depends both of the strength of  
375 the mutation bias towards AT, but also on the mean fitness and dominance effects of deleterious mutations.  
376 When even few deleterious mutations are recessive, the cost of transmitting slightly deleterious GC alleles  
377 becomes quickly higher, and  $b^*$  decreases. Importantly,  $b^*$  is negatively correlated to effective population  
378 size. In fact, the fitness gain of correcting strongly deleterious AT mutations is essentially independent of  
379 effective population size, while the cost of transmitting slightly deleterious GC alleles increases quickly with  
380 it. This could contribute to the absence, or the weak positive correlation between the population-scaled  
381 gBGC coefficient ( $B = 4N_e b$ ) and effective population size ( $N_e$ ) reported in several clades of eukaryotes  
382 (Lartillot, 2013; Clément et al., 2017; Galtier et al., 2018; Galtier, 2021; Boman et al., 2021).

### 383 **Decoupling the short- and long-term effect of gBGC**

384 gBGC is often described as an evolutionary force that antagonizes natural selection. It has even earned the  
385 nickname of "Achilles' heel of genomes" (Galtier and Duret, 2007). Galtier et al. (2018) proposed that the  
386 negative relationship between gBGC and effective population size observed in angiosperms and animals could  
387 also arise from a drift-barrier effect, where a low effective population size imposes a limit to the efficacy of  
388 selection against gBGC. Here we show that despite a deleterious effect at the population level, gBGC is still  
389 (weakly) positively selected. Therefore, the levels of gBGC observed in animals may not be counter-selected  
390 at all. In this view, the pervasive existence of gBGC in eukaryotes is not explained by a limited efficiency of  
391 negative selection due to drift, but by the short-term advantage of biasing gene conversion towards GC that  
392 limits the long-term reproductive capacity of the population as a whole.

### 393 **The strength of selection acting on gBGC**

394 The strength of stabilizing selection on gBGC according to our model is weak, and as a consequence, the  
395 equilibrium variance can be substantial. Of note, our estimate of the equilibrium variance depends on  
396 several assumptions. First, we ignored linkage and only considered the direct conversion gain/cost in fitness  
397 in one generation, while in fact, a gBGC modifier will be statistically linked to the deleterious AT alleles  
398 it corrects over several generations. By considering only the direct gain/cost at one generation, we might  
399 be underestimating the strength of selection acting on a modifier, and thus overestimating the evolutionary  
400 variance. However, the agreement of our analytical estimate with the simulations (which do incorporate the  
401 effect of linkage) suggests that the impact of linkage is not major.

402 Second, when computing the selection induced on a modifier, we implicitly assume that the population has  
403 time to reach mutation-selection-drift-gBGC equilibrium between each modification of  $b$  (low mutation rate  
404 at modifier loci). When the number and effect sizes of loci that can influence the strength of gBGC are large  
405 enough, such that the population does not have time to reach mutation-selection-drift-gBGC equilibrium  
406 between two consecutive modifications of  $b$ , the short-term benefit/cost of converting with a bias  $b$  is not  
407 coupled to its long-term benefit/cost. In this case, one should observe increased oscillations around  $b^*$ ,  
408 and thus increased evolutionary variance. This point is confirmed by running the simulator under a higher  
409 mutation rate at the modifier locus (Figure S1&2).

410 Empirically, not much is known about the genetic architecture of gBGC, so the effective mutation rate  
411 at the gBGC modifiers is difficult to estimate. Essentially, however, the results obtained here, which show  
412 a reasonable match with the simulated variance under linkage and assuming a low mutation rate at the  
413 modifier, give the best-case scenario among all possible genetic architectures for gBGC, i.e. the scenario for  
414 which stabilizing selection on  $b$  is tightest. Even so, it remains weak, at least too weak to always guarantee  
415 a positive gBGC under empirically reasonable conditions (Table 1). On the other hand, selection against  
416 excessively high levels of gBGC might still be efficient, depending on the exact distribution of fitness and  
417 dominance effects.

### 418 **The penetrance of a somatic repair bias in meiosis**

419 In our derivation, we have focussed on the consequences of biased repair in the germ line, disregarding all  
420 considerations about somatic constraints. In reality, however, it is very likely that the mechanisms that  
421 bias DNA repair towards GC in meiosis are the same as those that operate in somatic cells. Most single  
422 nucleotide DNA damages involve wrongly incorporated As, Ts, or even Us. Repair enzymes that minimize  
423 the somatic mutation rate should therefore be GC-biased. In this sense, in mammals, the base excision  
424 repair pathway has DNA glycosylases for excising adenines and thymines, but none for guanine or cytosine  
425 (Krokan and Bjørås, 2013).

426 Accounting for these somatic constraints leads to a different perspective on the evolution of gBGC, which  
427 could then be seen as an indirect consequence of the shared repair machinery between meiosis and somatic  
428 repair. In this context, modifiers can still act on the strength of gBGC, although now by modulating the  
429 penetrance of the structurally GC-biased repair system in meiosis. The simulation model could easily be

430 adapted to incorporate such somatic constraints, essentially by assuming that the modifiers of gBGC would  
431 act multiplicatively on  $b$ , itself a priori assumed positive. The semi-analytical derivation is less dependent  
432 on such details, as it merely quantifies the selection induced at equilibrium on any modifier of gBGC.

433 These considerations lead to a re-interpretation of the results presented here. What they fundamentally  
434 suggest is that, depending on the exact distribution of fitness and dominance effects, there might be enough  
435 selection for limiting the penetrance of the somatic repair bias in meiosis, if this ever leads to overly strong  
436 gBGC (deleterious at the individual level). In this view, the expected strength of gBGC should lie between  
437 the somatic repair bias (strongly GC biased) and  $b^*$ . This could explain why meiotic gene conversion appears  
438 to be universally GC-biased, despite the weak selection preventing it from being AT-biased in low  $N_e$  species  
439 (Table 1). Of note, even if selection to limit the penetrance of the somatic repair bias is maximally effective,  
440 the expected value of  $b$  still induces a substantial load at the population level.

#### 441 **gBGC and effective population size**

442 In mammals, there is a (weak) correlation between effective population size  $N_e$  and the population-scaled  
443 gBGC coefficient ( $B = 4N_e b$ ) (Lartillot, 2013; Galtier, 2021). This correlation is also observed among human  
444 populations (Glémin et al., 2015; Subramanian, 2019), and effective population size seems to explain the  
445 difference in  $B$  between two passerine species (Barton and Zeng, 2021). However, in *Leptidea* butterflies  
446 there is no relationship between  $B$  and genetic diversity, suggesting that the transmission bias  $b$  is lower  
447 in species/populations of higher effective population size (Boman et al., 2021). Finally, no correlation has  
448 been observed between  $B$  and  $N_e$  in 29 animal species (Galtier et al., 2018), or in 11 species of angiosperms  
449 (Clément et al., 2017).

450 The most probable hypothesis so far is that in animals and plants, there is a negative correlation between  
451 the repair bias  $b_0$  and effective population size (Galtier et al., 2018). Several arguments have been proposed to  
452 explain this negative relationship. As previously said, Galtier et al. (2018) proposed a drift-barrier hypothesis:  
453 assuming that gBGC is deleterious, it can be more efficiently counter-selected in species with higher  $N_e$ . Our  
454 modeling work provides another interpretation, by pointing out that the evolutionary optimum  $b^*$  is itself  
455 negatively correlated with  $N_e$ .

456 On the other hand, it has been shown that depending on the repair mechanisms, the intensity of the bias  
457 could be negatively correlated with heterozygosity (Lesecque et al., 2013; Li et al., 2019). As heterozygosity  
458 is supposed to be proportional to  $N_e$ , this can also explain why we observe no correlation between  $B$  and  $N_e$   
459 (Clément et al., 2017; Galtier et al., 2018; Boman et al., 2021), while we still expect one under selection only.  
460 However, the switch to such heterozygosity-dependent mechanisms could also be an adaptive response to the  
461 increasing cost of gBGC, and the two hypotheses are not mutually exclusive. Nevertheless, these hypothesis  
462 remain verbal, and a proper modelling of the molecular mechanisms of gBGC and their selective advantage  
463 is needed to put them to the test.

## 464 Empirical relevance

465 We highlighted that gBGC being deleterious at the population level is not an indicator that it is negatively  
466 selected. It is therefore unclear whether the levels of gBGC currently found in eukaryotes are actually  
467 negatively selected. Here, we computed the expected  $b^*$  under empirically realistic parameters, and recover  
468 a rather high  $b^*$ . Of note, in the context of a heterogeneous recombination landscape, such as considered  
469 above, most of the selection induced on the modifier is contributed by those positions that are under the  
470 strongest gBGC, which correspond to the highly recombining regions of the genome. Our model therefore  
471 predicts even higher equilibrium values for  $b^*$  under more homogeneous recombination landscapes (Table  
472 S1).

473 It is important to note, however, that this estimation is sensitive to parameters that are very difficult to  
474 estimate reliably. Notably, the size of the genome that is under selection (Rands et al., 2014), the DFE and  
475 more specifically the size of the compartment of strongly deleterious mutations. Moreover,  $b^*$  is strongly  
476 sensitive to the distribution of dominance effects, about which little is known (Billiard et al., 2021). Finally,  
477 it relies on the assumption that half the genome under selection has a GC allele as optimal and the other  
478 half an AT allele. This assumption is intuitive and is made in almost all models of gBGC (Bengtsson, 1990;  
479 Glémin, 2010; Bolívar et al., 2016; Corcoran et al., 2017). However, when using empirical fitness landscapes  
480 in protein coding genes instead of an arbitrary distribution of selective effects, it appears that AT encoded  
481 amino-acids are more often optimal than GC-ones, which seems to be partly due to the structure of the  
482 genetic code (Joseph, 2024). Under this scenario, a slight mutational bias is actually beneficial, and thus  $b^*$   
483 should be lower.

484 Overall, while the present model significantly improves our understanding of the selective pressures  
485 exerted on gBGC, it is by no mean an attempt to accurately predict the strength of gBGC *in natura*.

## 486 4 Methods

### 487 4.1 Model

488 The model assumes a population of fixed size  $N$  diploid individuals, randomly mating and with non-  
489 overlapping generations. The genome is composed of a single chromosome. Since neutral loci don't have  
490 any impact on the evolution of gBGC, they are not explicitly modeled. As a result, the chromosome is  
491 assumed to consist of  $L$  bi-allelic positions, with two alternative alleles,  $W$  (weak) or  $S$  (strong), that are  
492 all under selection with locus-specific selective strengths. The model also invokes a modifier locus placed  
493 somewhere along the chromosome (in the experiments conducted here, at one third of the total length of the  
494 chromosome).

495 For a given selected position  $i$ ,  $1 \leq i \leq L$ , either the  $W$  allele or the  $S$  allele is considered deleterious with  
496 probability  $1/2$ , in which case the selection strength  $s_i$  acting on the deleterious allele is randomly drawn  
497 from a gamma distribution of mean  $\bar{s}$  and shape parameter  $a$ . All selected loci share the same dominance  
498 coefficient  $h$ . In the following, the co-dominant case  $h = 1/2$  and the recessive case, where the deleterious  
499 allele is recessive with  $0 < h < 1/2$ , are both considered. The selective effects are assumed additive over

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

500 loci. Thus, assuming locus  $i$  is such that  $W$  is the deleterious allele, then the log-fitness contribution is 0  
 501 for genotype  $SS$ ,  $hs_i$  for genotype  $SW$  and  $s_i$  for genotype  $WW$  (and conversely for loci for which  $S$  is the  
 502 deleterious allele). Letting  $Q_{ij}^1, Q_{ij}^2 \in \{0, 1\}^2$  stand for the genotype of diploid individual  $j$  at position  $i$ ,  
 503 with the convention that 1 stands for the deleterious allele (which can be either  $S$  or  $W$  depending on the  
 504 locus), the total Malthusian (log) fitness of individual  $j$  is then given by:

$$\ln W_j = - \sum_{i=1}^L (Q_{ij}^1(1 - Q_{ij}^2) + Q_{ij}^2(1 - Q_{ij}^1)) hs_i + Q_{ij}^1 Q_{ij}^2 s_i. \quad (10)$$

505 The selected positions undergo recurrent mutations between  $W$  and  $S$ . Allele  $W$  mutates towards  $S$  at rate  
 506  $u$ , and allele  $S$  mutates towards  $W$  at rate  $\lambda u$  per generation.

507 The modifier locus encodes an additive quantitative trait controlling the bias of gene conversion. For  
 508 individual  $j$ , with genotype  $(z_j^1, z_j^2) \in R^2$  at the modifier locus, the bias is then equal to:

$$\beta_j = \frac{1}{2} (z_j^1 + z_j^2).$$

509 How this bias exactly impacts gene conversion during meiosis is described below. The modifier locus mutates  
 510 are rate  $w$ , in which case the quantitative contribution of the mutant allele is equal to that of its parent,  
 511 plus a normally distributed increment, of mean 0 and standard deviation  $\Delta z$ :

$$z' \sim N(z, \Delta z^2).$$

512 A simplified version of meiosis is implemented as follows. Consider individual  $j$ . First, each selected  
 513 position that happens to be heterozygous in this individual undergoes gene conversion with probability  $\alpha$ ,  
 514 in which case conversion is towards the  $S$  allele with probability  $(1 + \beta_j)/2$  and towards the  $W$  allele with  
 515 probability  $(1 - \beta_j)/2$ , with  $\beta_j$  such as defined above (equation.). Second, a cross-over point is chosen  
 516 uniformly at random over the chromosome, and two recombinant chromosomes are produced by swapping  
 517 the segments on both sides of the cross-over point. Thus, both the rate of cross-over and the rate of gene  
 518 conversion are considered fixed and invariant across individuals, while the bias of the gene conversion events  
 519 is allowed to vary between individuals, based on the genotype at the modifying locus. Of note, a positive  
 520 (resp. negative) value for  $\beta_j$  results in biased gene conversion towards  $S$  (resp. towards  $W$ ). Quantitatively,  
 521 at a given selected position at which individual  $j$  is heterozygous, the net proportions of gametes produced  
 522 by this individual bearing the  $S$  allele is:

$$q_S = (1 - \alpha) \frac{1}{2} + \alpha \frac{1 + \beta}{2} = \frac{1 + \alpha\beta}{2} = \frac{1 + b}{2},$$

523 with  $b = \alpha\beta$ . Similarly, the proportion of gametes with the  $W$  allele is  $q_W = \frac{1-b}{2}$ .

524 The overall life cycle runs as follows. First, all individuals of the current generation undergo mutations  
 525 both at the modifying and at the selected loci, with mutation rates such as given above. Next, each individual  
 526 of the next generation is produced by first randomly choosing two parents in the current generation, each  
 527 with a probability proportional to its fitness  $W$  (such as given by equation 10 above). Each of the two chosen  
 528 individuals then undergoes a meiosis, producing a pair of gametes, one of which is randomly picked out and  
 529 paired with the gamete produced by applying the same procedure to the other individual. Of note, only one  
 530 gamete per meiosis is kept for the next generation, the other one being discarded.

531 Altogether, the parameters of the model are:

- 532  $N$ : population size
- 533  $L$ : number of loci under selection
- 534  $\bar{s}$ : mean selection strength at the selected loci
- 535  $a$ : shape of the distribution of selection strengths across loci
- 536  $h$ : dominance coefficient
- 537  $u$ : basal mutation rate at the selected loci
- 538  $\lambda$ : mutation bias ( $S \rightarrow W$  relative to  $W \rightarrow S$ )
- 539  $w$ : mutation rate at the modifier locus
- 540  $\Delta z$ : mean effect size of the mutations at the modifier locus
- 541  $\alpha$ : gene conversion rate (per generation and per selected locus)

## 542 4.2 Theory / Analytical approximation

543 Here, a semi-analytical approximation is derived for determining the equilibrium value of the net strength  
544 of biased gene conversion  $b$  as well as its evolutionary variance. This derivation assumes a low mutation  
545 rate at the modifier locus (low  $w$ ), such that the population is, at any time, approximately monomorphic for  
546 the strength of gBGC, and all selected loci are at mutation-selection-drift-conversion equilibrium under this  
547 value of gBGC. The derivation also assumes that linkage both among selected loci and between the modifier  
548 and the selected loci is negligible. The first condition implies that background selection is weak, and that the  
549 mutation-selection-drift-conversion equilibrium can be determined independently at each locus. The second  
550 is motivated by the fact that, in practice, most selected loci are sufficiently far from the modifier, such that  
551 most of the induced selection is contributed by loci that not tightly linked with the modifier.

552 Consider a population monomorphic at the modifier locus for an allele of strength  $\beta$ , at equilibrium under  
553 a gBGC of strength  $b = \alpha\beta$ . In this population, a mutant at the modifier locus, of size  $2\delta\beta$  appears in an  
554 individual. This individual thus has a gBGC strength of  $b' = b + \delta b$  in its germline, with  $\delta b = \alpha \delta\beta$ . We want  
555 to determine the net selective advantage or disadvantage incurred by this individual, owing to its departure  
556 from the population-level gBGC. This selection will be indirectly contributed by the effect of biased gene  
557 conversion on the selected loci across the genome. Therefore, in the following, this will be called the selection  
558 *induced* on the gBGC modifier, or induced selection for short.

559 Under efficient linkage dissipation, induced selection is the sum of the contributions of all selected loci.  
560 Consider in a first step a single locus at which  $W$  is the deleterious allele, with selection  $s$ , dominance  $h$  and  
561 segregating in the population at frequency  $x$ . Given  $x$ , the probability for the individual to be heterozygous  
562 at this position is:

$$P(x) = 2x(1-x),$$

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

563 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.  
 564 In a random mating population, this will result in an average fitness gain in the offspring of:

$$\begin{aligned} \ln f_W(x, s, h) &= \frac{1+b'}{2}((1-x) \times 0 + x \times (-hs)) + \frac{1-b'}{2}((1-x) \times (-hs) + x \times (-s)) \\ &= b' \frac{s}{2} [h + x(1-2h)] \\ &= b' C(s, h, x), \end{aligned}$$

565 where:

$$C(s, h, x) = \frac{s}{2} [h + x(1-2h)].$$

566 Of note, if  $b' > 0$ , this is indeed a gain, since on average, S alleles, which have a higher fitness at that position,  
 567 are over-transmitted. Next, to assess the fate of the gBGC mutant, one should discount the equivalent gain,  
 568 but under a gBGC equal to  $b$  in the population, such that the average selective advantage contributed by  
 569 the selected position under consideration to the individual bearing the mutant allele for the modifier (now  
 570 accounting for the probability for this individual to be a heterozygote at the selected locus) is:

$$\delta \ln f_W(x, s, h) = P(x) C(s, h, x) \delta b.$$

571 Equation 11 gives the cost conditional on the frequency  $x$  of the W allele at the focal selected position and  
 572 conditional on the selection coefficient  $s$ . This needs to be averaged over  $x$  at mutation-selection-conversion-  
 573 drift equilibrium (here noted  $\phi_{s,b}^W$ ) and then summed over the distribution of selective effects across the  $L/2$   
 574 loci being deleterious towards the W allele:

$$\begin{aligned} G_W(b) &= \frac{\delta \ln f_W}{\delta b} \\ &= \frac{L}{2} \langle H_W(s, b) \rangle, \end{aligned}$$

575 where the angle brackets stand for the expectation over  $s$  under the DFE, and,

$$H_W(s, b) = E_{\phi_{s,b}^W} [P \times C]$$

576 is the expectation over  $x$  under  $\phi_{b,s}^W$  of  $P(x) C(s, h, x)$ . In other words, it is the net gain induced by conversion  
 577 events at loci that are W-deleterious, with selection coefficient  $s$  and dominance coefficient  $h$ . In turn, the  
 578 distribution  $\phi_{b,s}^W$  is given by (Wright, Glemin):

$$\phi_{b,s}^W(x) = \frac{1}{Z_{b,s}^W} x^{4Nv-1} (1-x)^{4Nu-1} e^{-4Nx(b+s(h(1-x)+x(1-h)))},$$

579 where  $Z_{b,s}^W$  is the normalization constant:

$$Z_{b,s}^W = \int x^{4Nv-1} (1-x)^{4Nu-1} e^{-4Nx(b+s(h(1-x)+x(1-h)))} dx.$$

580 A similar derivation is done for a locus at which W is the deleterious allele, which, by symmetry, gives:

$$\begin{aligned} G_S(b) &= \frac{\delta \ln f_S}{\delta b} \\ &= -\frac{L}{2} \langle H_S(s, b) \rangle, \end{aligned}$$

581 where

$$H_S(s, b) = E_{\phi_{s,b}^S} [P \times C]$$

582 and

$$\phi_{b,s}^S(x) = \frac{1}{Z_{b,s}^S} x^{4Nu-1} (1-x)^{4Nv-1} e^{-4Nx(-b+s(h(1-x)+x(1-h)))},$$

583 with normalization constant:

$$Z_{b,s}^S = \int x^{4Nu-1} (1-x)^{4Nv-1} e^{-4Nx(-b+s(h(1-x)+x(1-h)))} dx.$$

584 Of note,  $P(x)$  and  $C(s, h, x)$  are positive for all  $x$ , and thus, increasing biased gene conversion towards the  
585 strong alleles results in a net gain over  $W$ -deleterious loci, but a net a loss over  $S$ -deleterious loci. Whether  
586 the mutant for gBGC is favoured by this induced selection will depend on the balance between these two  
587 components. In other words, the total selection induced on the modifier is:

$$\begin{aligned} \frac{\delta \ln f}{\delta b} &= G(b) \\ &= G_W(b) - G_s(b). \end{aligned}$$

### 588 4.3 Implementation

589 The model was implemented in C++, using openMP for parallelizing the computations. For all results  
590 presented here, it was run under population sizes of size  $N = 1000$ , with  $L = 10000$  selected loci, for a  
591 total of 210 000 generations, discarding the first 10 000 generations (burn-in) and subsampling 1 every 100  
592 generations, upon which averages and standard deviations for quantities of interest were computed on the  
593 remaining 2000 points.

594 Numerical integration and solving was done in Python, using the scipy library for numerical integration  
595 over the allele frequency distributions. For integrating over the gamma distribution of selective effects, the  
596 gamma distribution was discretized into  $n = 300$  points, corresponding to the mid-points of the successive  
597  $1/n$  quantiles, and then the integral over the distribution was approximated as the equal-weighted average  
598 of the integrand over these  $n$  values for  $hs$ .

### 599 Acknowledgments

600 We are grateful to Sylvain Glémin, Laurent Duret and Nicolas Galtier for their comments on previous versions  
601 of this manuscript. We also would like to thank Sylvain Glémin for insightful discussions on the subject  
602 and on the model. **Funding:** Agence Nationale de la Recherche, Grant ANR-19-CE12-0019 / HotRec.  
603 **Competing interests:** The author declare no conflicts of interest. **Data and materials availability:**  
604 All the scripts necessary to reproduce this study are available at <https://gitlab.in2p3.fr/bayesiancook/gbgc>

### 605 References

606 Barton, H. J. and Zeng, K. (2021). The effective population size modulates the strength of GC biased gene  
607 conversion in two passerines.  
608 Bengtsson, B. O. (1986). Biased conversion as the primary function of recombination. *Genetics Research*,  
609 47(1):77–80.



- 610 Bengtsson, B. O. (1990). The effect of biased conversion on the mutation load. *Genetics Research*, 55(3):183–  
611 187.
- 612 Bengtsson, B. O. and Uyenoyama, M. K. (1990). Evolution of the segregation ratio: Modification of gene  
613 conversion and meiotic drive. *Theoretical Population Biology*, 38(2):192–218.
- 614 Berglund, J., Pollard, K. S., and Webster, M. T. (2009). Hotspots of Biased Nucleotide Substitutions in  
615 Human Genes. *PLOS Biology*, 7(1):e1000026.
- 616 Billiard, S., Castric, V., and Llaurens, V. (2021). The integrative biology of genetic dominance. *Biological  
617 Reviews*, 96(6):2925–2942.
- 618 Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. (2019). GC-biased gene conversion  
619 conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol*, 20(1):5.
- 620 Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. (2016). Recombination Rate Variation Modulates  
621 Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an  
622 Avian System. *Mol Biol Evol*, 33(1):216–227.
- 623 Boman, J., Mugal, C. F., and Backström, N. (2021). The Effects of GC-Biased Gene Conversion on Patterns  
624 of Genetic Diversity among and across Butterfly Genomes. *Genome Biology and Evolution*, 13(5).
- 625 Brown, T. C. and Jiricny, J. (1987). A specific mismatch repair event protects mammalian cells from loss of  
626 5-methylcytosine. *Cell*, 50(6):945–950.
- 627 Castellano, D., Macià, M. C., Tataru, P., Bataillon, T., and Munch, K. (2019). Comparison of the Full  
628 Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes. *Genetics*, 213(3):953–  
629 966.
- 630 Clément, Y., Sarah, G., Holtz, Y., Homa, F., Pointet, S., Contreras, S., Nabholz, B., Sabot, F., Sauné,  
631 L., Ardisson, M., Bacilieri, R., Besnard, G., Berger, A., Cardi, C., Bellis, F. D., Fouet, O., Jourda,  
632 C., Khadari, B., Lanaud, C., Leroy, T., Pot, D., Sauvage, C., Scarcelli, N., Tregear, J., Vigouroux, Y.,  
633 Yahiaoui, N., Ruiz, M., Santoni, S., Labouisse, J.-P., Pham, J.-L., David, J., and Glémin, S. (2017).  
634 Evolutionary forces affecting synonymous variations in plant genomes. *PLOS Genetics*, 13(5):e1006799.
- 635 Corcoran, P., Gossmann, T. I., Barton, H. J., The Great Tit HapMap Consortium, Slate, J., and Zeng, K.  
636 (2017). Determinants of the Efficacy of Natural Selection on Coding and Noncoding Variability in Two  
637 Passerine Species. *Genome Biology and Evolution*, 9(11):2987–3007.
- 638 Duret, L. and Arndt, P. F. (2008). The Impact of Recombination on Nucleotide Substitutions in the Human  
639 Genome. *PLOS Genetics*, 4(5):e1000071.
- 640 Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic  
641 Landscapes. *Annu. Rev. Genom. Hum. Genet.*, 10(1):285–311.
- 642 Eyre-Walker, A. (1999). Evidence of Selection on Silent Site Base Composition in Mammals: Potential  
643 Implications for the Evolution of Isochores and Junk DNA. *Genetics*, 152(2):675–683.
- 644 Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev  
645 Genet*, 8(8):610–618.

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

- 646 Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The Distribution of Fitness Effects of New Deleterious  
647 Amino Acid Mutations in Humans. *Genetics*, 173(2):891–900.
- 648 Galtier, N. (2021). Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer*  
649 *Community Journal*, 1.
- 650 Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of  
651 molecular evolution. *Trends in Genetics*, 23(6):273–277.
- 652 Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation  
653 of deleterious amino acid changes in primates. *Trends in Genetics*, 25(1):1–5.
- 654 Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content Evolution in Mammalian  
655 Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, 159(2):907–911.
- 656 Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. (2018).  
657 Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size,  
658 and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5):1092–1103.
- 659 Glémin, S. (2010). Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and  
660 Inbreeding Depression. *Genetics*, 185(3):939–959.
- 661 Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of  
662 GC-biased gene conversion in the human genome. *Genome Res.*, 25(8):1215–1228.
- 663 Joseph, J. (2024). Increased Positive Selection in Highly Recombining Genes Does not Necessarily Reflect  
664 an Evolutionary Advantage of Recombination. *Molecular Biology and Evolution*, 41(6):msae107.
- 665 Kostka, D., Hubisz, M. J., Siepel, A., and Pollard, K. S. (2012). The Role of GC-Biased Gene Conver-  
666 sion in Shaping the Fastest Evolving Regions of the Human Genome. *Molecular Biology and Evolution*,  
667 29(3):1047–1057.
- 668 Krokan, H. E. and Bjørås, M. (2013). Base Excision Repair. *Cold Spring Harb Perspect Biol*, 5(4):a012583.
- 669 Lachance, J. and Tishkoff, S. A. (2014). Biased Gene Conversion Skews Allele Frequencies in Human Pop-  
670 ulations, Increasing the Disease Burden of Recessive Alleles. *The American Journal of Human Genetics*,  
671 95(4):408–420.
- 672 Lartillot, N. (2013). Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the  
673 Evolutionary Dynamics of Recombination Landscapes. *Molecular Biology and Evolution*, 30(3):489–502.
- 674 Lesecque, Y., Mouchiroud, D., and Duret, L. (2013). GC-Biased Gene Conversion in Yeast Is Specifically  
675 Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance. *Molecular Biology and*  
676 *Evolution*, 30(6):1409–1419.
- 677 Li, R., Bitoun, E., Altomose, N., Davies, R. W., Davies, B., and Myers, S. R. (2019). A high-resolution map  
678 of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat*  
679 *Commun*, 10(1):3900.
- 680 Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., Patterson, C., Gregory, C.,  
681 Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W. R., Muscarella, M. E., Luo, H., Lennon, J. T.,

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

- 682 Brun, Y. V., and Lynch, M. (2018). Evolutionary determinants of genome-wide nucleotide composition.  
683 *Nat Ecol Evol*, 2(2):237–240.
- 684 Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of  
685 meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485.
- 686 Meunier, J. and Duret, L. (2004). Recombination Drives the Evolution of GC-Content in the Human Genome.  
687 *Molecular Biology and Evolution*, 21(6):984–990.
- 688 Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National*  
689 *Academy of Sciences*, 80(20):6278–6281.
- 690 Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., and Duret, L. (2011).  
691 Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mu-*  
692 *tation*, 32(2):198–206.
- 693 Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and*  
694 *Systematics*, 23:263–286.
- 695 Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. (2012). Evidence for widespread  
696 GC-biased gene conversion in eukaryotes. *Genome biology and evolution*, 4(7):675–682.
- 697 Pouyet, F., Aeschbacher, S., Thiéry, A., and Excoffier, L. (2018). Background selection and biased gene  
698 conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7:e36317.
- 699 Pouyet, F., Mouchiroud, D., Duret, L., and Sémon, M. (2017). Recombination, meiotic expression and  
700 human codon usage. *eLife*, 6:e27344.
- 701 Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., and Camerini-Otero, R. D. (2014). Recom-  
702 bination initiation maps of individual human genomes. *Science*, 346(6211):1256442–1256442.
- 703 Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained:  
704 Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLOS Genetics*,  
705 10(7):e1004525.
- 706 Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. (2010).  
707 Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Trans-*  
708 *actions of the Royal Society B: Biological Sciences*, 365(1552):2571–2580.
- 709 Roman, H. (1985). Gene conversion and crossing-over. *Environmental Mutagenesis*, 7(6):923–932.
- 710 Smagulova, F., Gregoret, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V.  
711 (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Na-*  
712 *ture*, 472(7343):375–378.
- 713 Smeds, L., Mugal, C. F., Qvarnström, A., and Ellegren, H. (2016). High-Resolution Mapping of Crossover  
714 and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS*  
715 *Genetics*, 12(5):e1006044.
- 716 Subramanian, S. (2019). Population size influences the type of nucleotide variations in humans. *BMC Genet*,  
717 20(1):1–12.

THE EVOLUTION OF GC-BIASED GENE CONVERSION BY MEANS OF NATURAL SELECTION

- 718 Webster, M. T., Axelsson, E., and Ellegren, H. (2006). Strong Regional Biases in Nucleotide Substitution  
719 in the Chicken Genome. *Molecular Biology and Evolution*, 23(6):1203–1216.
- 720 Winkler, H. (1930). Die Konversion der Gene : eine vererbungstheoretische Untersuchung. *G Fischer*.