Fine-scale quantification of GC-biased gene conversion intensity in mammals.

Nicolas Galtier

ISEM - Univ Montpellier - CNRS - IRD

Abstract

GC-biased gene conversion (gBGC) is a molecular evolutionary force that favours GC over AT alleles irrespective of their fitness effect. Quantifying the variation in time and across genomes of its intensity is key to properly interpret patterns of molecular evolution. In particular, the existing literature is unclear regarding the relationship between gBGC strength and species effective population size, N_e. Here we analysed the nucleotide substitution pattern in coding sequences of closely related species of mammals, thus accessing a high resolution map of the intensity of gBGC. Our maximum likelihood approach shows that gBGC is pervasive, highly variable among species and genes, and of strength positively correlated with N_e in mammals. We estimate that gBGC explains up to 60% of the total amount of synonymous AT→GC substitutions. We show that the fine-scale analysis of gBGC-induced nucleotide substitutions has the potential to inform on various aspects of molecular evolution, such as the distribution of fitness effects of mutations and the dynamics of recombination hotspots.

1 Introduction

GC-biased gene conversion (gBGC) is a recombination-associated transmission bias by which G and C alleles are favoured over A and T alleles. This evolutionary force was discovered in the 2000's from the analyses of early population genomic data sets [22, 30, 76, 80], and experimentally confirmed later on [49, 65, 83]. gBGC manifests itself as a GC-bias that affects both non-functional and functional sequences and is correlated with the local recombination rate [28]. gBGC has a strong impact on patterns of variation genome wide in mammals [10, 20, 64, 69] and many other taxa [11, 25, 33, 44, 48, 55, 58, 61, 78]. gBGC can mimic the effect of natural selection and confound its detection by generating patterns of clustered AT→GC substitutions, distorted site frequency spectra and altered non-synonymous/synonymous ratios [4, 15, 19, 28, 42, 66, 73]. Importantly, because it favours G and C alleles irrespective of their fitness effect, gBGC tends to counteract natural selection and increase the deleterious mutation load [3, 32, 41, 59].

The abundant body of literature reviewed above demonstrates a significant effect of gBGC in a large number of genomes. Only a few studies, however, have attempted to quantify its strength - a harder task. gBGC results from a DNA repair bias involving paired chromosomes at meiosis, and operating in the immediate neighborhood of DNA double strand breaks. The genome average transmission bias, b, is therefore expected to be proportional to the recombination rate, gene conversion tract length, and repair bias. The effect of gBGC on genome evolution is also expected to be dependent on the intensity of drift: being a directional force, gBGC is only effective if stronger than the stochastic component of allele frequency evolution. The intensity of drift is inversely related to the effective population size N_e, so that the strength of gBGC is usually measured by the $B = 4N_eb$ parameter. Glémin et al. [35] used genome-wide resequencing data to estimate B at the megabase scale throughout the human genome.

Fitting various population genetic models to polarised GC vs. AT site frequency spectra, Glémin et al. [35] estimated the genome average B to be in the weak selection range, around 0.4, with B reaching a value above 5 in 1%-2% of the genome. This variance among genomic regions in gBGC strength is interpreted as reflecting the existence of recombination hotpots in humans [8, 20, 76].

Similar analyses have been performed in a number of non-human taxa. In the fruit fly *Drosophila melanogaster*, no evidence for gBGC has been reported, albeit a weak effect on the X chromosome [31, 68]. In contrast, Wallberg et al. [78] estimated the genome average B to be above 5 in the honey bee *Apis mellifera*, again with substantial variation between low-recombining and high-recombining regions. Note that N_e is expected to be much smaller in *Homo sapiens* and the eusocial *A. mellifera* than in *D. melanogaster* [72]. Galtier et al. [33] analysed site frequency spectrum at synonymous positions in the coding sequences of 30 species of animals. They estimated that the average B at third codon positions varies between 0 and 2 among species, without any significant relationship with N_e-related life history traits. These comparisons among distantly-related animals revealed substantial variation in the intensity of gBGC among species, but, somewhat paradoxically, no detectable effect of N_e.

Another attempt to quantify the strength of gBGC is to get information from between-species divergence data, instead of within-species polymorphism data. Capra et al. [8] simultaneously modelled the effects of purifying selection and gBGC during the human/chimpanzee divergence and estimated that in apes 0.33% of the genome is undergoing gBGC at rate $B = 3$. This is lower than the estimates provided by Glémin et al. [35, see above], presumably because Capra et al. [8] assumed a constant gBGC rate at any location of the genome, whereas recombination hotspots are known to be highly dynamic in apes [1, 46]. Using a method that combines polymorphism and divergence data, De Maio et al. [18] estimated the average B to be of the order of 0.3-0.7 in apes, consistent with Glémin et al. [35]. Lartillot [43] analysed coding sequence divergence in 33 species of placental mammals and estimated the among-gene and among-species variation of B. He found that the average B varied among species from ~ 0.1 (in apes) to 3-5 (in bats and lagomorphs), with an among-gene standard deviation of B as high as twice the mean. Lartillot [43] detected a significant, negative correlation between B and species body mass. Body mass being strongly and negatively correlated with population density in mammals [16], this result suggests that N_e might be a determinant of the strength of gBGC in mammals, in agreement with theoretical expectations. Elaborating on the approach of De Maio et al. [18], Borges et al. [5] also reported a positive relationship between the population scaled gBGC coefficient and N_e across species/populations of apes.

So on one hand comparative analyses of site-frequency spectra among animals did not reveal any effect of N_e on the strength of gBGC, while on the other hand the analysis of the substitution pattern in mammals is consistent with a N_e effect. Also a bit surprisingly, the estimated range of variation of B across mammals [0.1-5, 43] is wide enough to contain all the estimates of B reported in any species of animals so far. Lartillot [43] analysed a subset of currently annotated mammalian orthologs (1329 exons in the largest data set), and importantly, relatively ancient divergences, at the family or order level, thus capturing the average effect of gBGC across dozens of million years. Here we analyse a large set of genes from closely related species in four families of mammals, thus accessing a high-resolution map of the effect of gBGC on coding sequences, both in time and across the genome. We focus on two key features of gBGC-driven molecular evolution, namely clustered AT→GC substitutions, and an excess of AT→GC over GC→AT substitutions compared to the mutation process. Estimating B in 40 lineages of mammals, we show that gBGC explains a substantial fraction of synonymous and non-synonymous AT→GC substitutions, that N_e is a strong predictor of the intensity of gBGC in mammals, and that large-N_e and small-N_e taxa differ substantially in how gBGC is distributed among and within genes.

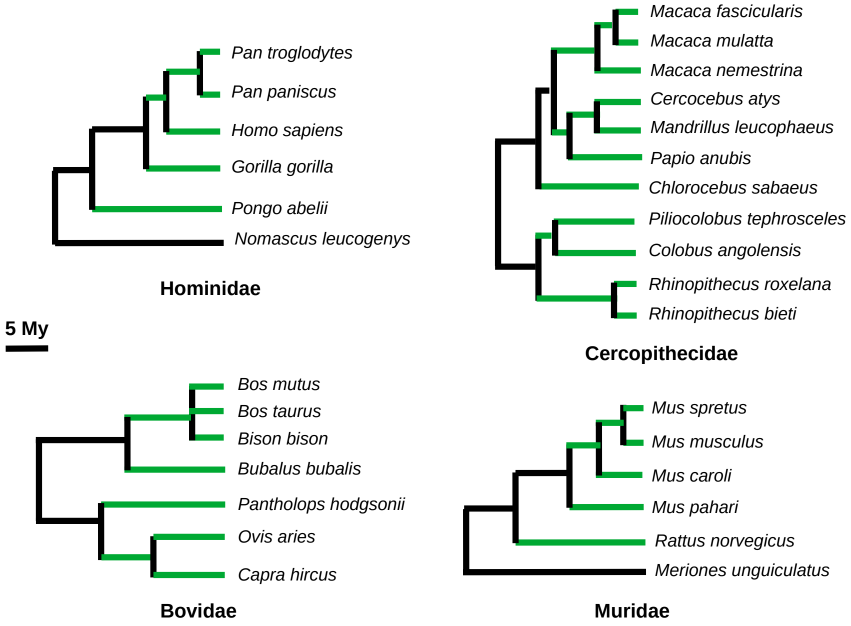

Figure 1: The four families, 32 species and 40 lineages of mammals (green branches) analysed here. Branch lengths are proportional to the estimated divergence time.

2 Results

2.1 Overview

We analysed patterns of AT→GC (i.e., Weak to Strong, or WS), GC→AT (SW) and GC-conservative (SSWW) coding sequence nucleotide substitutions in 40 recently diverged lineages (branches) from four families of mammals (Fig.1), namely Hominidae (humans and apes), Cercopithecidae (old world monkeys), Bovidae (cattle, sheep and allies) and Muridae (mice, rats, gerbils). A total of 1,104,917 third codon position synonymous substitutions and 514,552 first or second codon position non-synonymous substitutions were called. The median number of substitutions across branches was 24,960, and the minimum was 3927. The overall ratio of non-synonymous to synonymous substitutions, d_N/d_S, was 0.233; the family-specific d_N/d_S ratio was 0.275, 0.252, 0.228 and 0.213 in Hominidae, Cercopithecidae, Bovidae and Muridae, respectively. The d_N/d_S ratio is a marker of N_e in mammals, with small populations experiencing a higher substitution load, hence a higher d_N/d_S [60, 62, 70]. These results therefore indicate that the four families of our data set rank in the Hominidae < Cercopithecidae < Bovidae < Muridae order as far as N_e is concerned, consistent with previous analyses [43, 71].

2.2 Substitution clustering

Focusing on synonymous substitutions, we calculated Moran's I [53], a statistics that measures spatial aurocorrelation and was adjusted to target the 400 bp scale. This index therefore measures the tendency for substitutions (of a specific sort) having appeared in a given branch to be located less than 400 bp apart. Fig.2 shows the distribution among branches of the average centered Moran's I, separately for WS and SW synonymous substitutions. The centered Moran's I for SW substitutions was very close to zero in all branches from all four families, indicating very little, if any, clustering of substitutions. WS substitutions behaved differently: the centered Moran's I was close to zero in Hominidae, perceptibly positive in Cercopithecidae, and reached much higher values in Bovidae and Muridae, demonstrating the existence of clusters of synonymous WS substitutions in these two families. This pattern - clustering of WS but not SW substitutions - is a signature of gBGC [*e.g.* 19]; its intensity appears to increase with N_e across the four families analysed here.

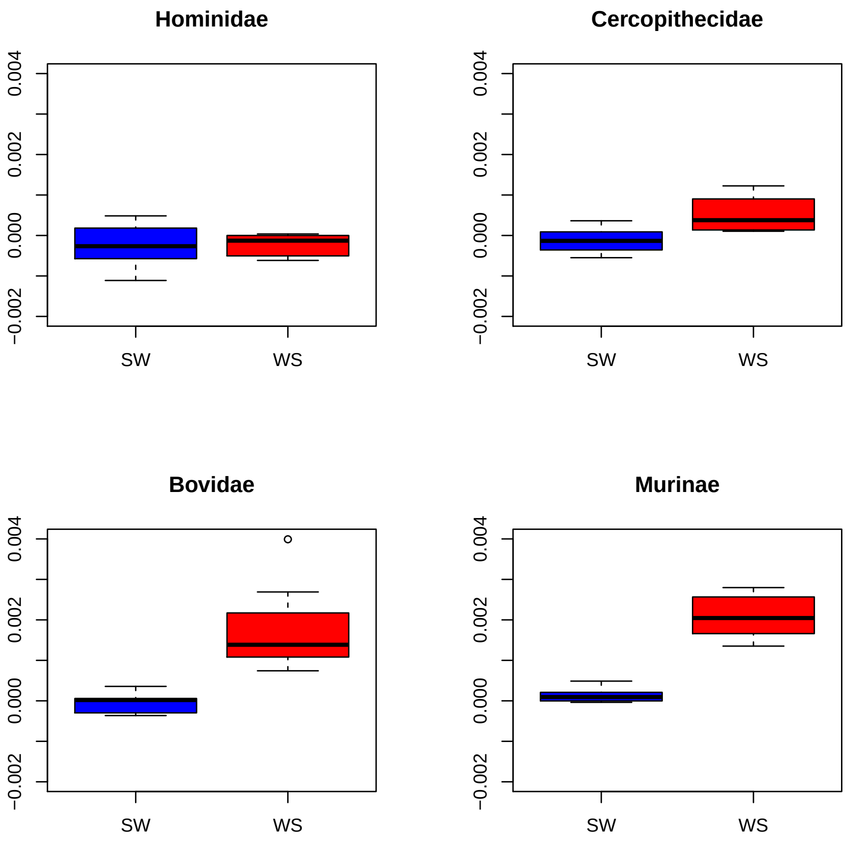

Figure 2: Distribution of centered Moran's I for WS and SW synonymous substitutions in four families of mammals. Only branches in which at least 100 genes had at least 3 inferred substitutions were included.

Simulations were performed in order to assess the amount of clustering needed to explain the observed values of Moran's I. Our simulation procedure considers two levels of clustering, one at the 500 bp scale and one at the 40 bp scale, while accounting for the intron-exon structure and the among-genes variance in mutation rate and GC-content (see Methods). In Muridae and Bovidae, we were able to replicate the observed values of Moran's I when 15-40% of the simulated substitutions appeared in clusters. This percentage was 0-10% in Cercopithecidae, and non-existent in Hominidae (Supplementary Fig. S1).

2.3 Estimating B

For each branch we estimated the population gBGC coefficient $B = 4N_eb$ and its variation based on synonymous WS, SW and SSWW synonymous substitution counts. Various models were fitted to the data via the maximum likelihood (ML) method, assuming that the mutation process is known [75]. Model M1 assumes a constant intensity of gBGC, B, among and within genes. Model M2 considers two categories of genes, each with its own gBGC intensity, assumed to be shared by all sites within a gene. M2 led to a rejection of M1 by a likelihood ratio test ($p-val < 0.05$) in 36 branches out of 40. The M3z model assumes three categories of genes, which we below denote "cold" ($B = 0$), "mild", and "hot".

M3z rejected M1 in 39 branches out of 40, and M2 in 27 branches. There was, therefore, strong evidence for a variable B across genes in this data set.

Then we fitted models that assume some variation of B both among and within genes. Model M3h considers three categories of genes that differ in terms of the prevalence, q, of gBGC hotspots. gBGC is assumed to operate at intensity B_h within hotspots, and zero outside hotspots. Model M3sh is a simplified version of M3h obtained when q approaches zero. Applying these two models led to a dramatic increase in log-likelihood for most branches (Supplementary Table S1), which is indicative of the existence of substantial within-gene variation in gBGC intensity. Model M3h rejected M3sh by a likelihood ratio test only in one branch out of 40 (*Bison bison* terminal branch, Bovidae), consistent with the idea that gBGC hotspots occupy a small fraction of coding sequence length.

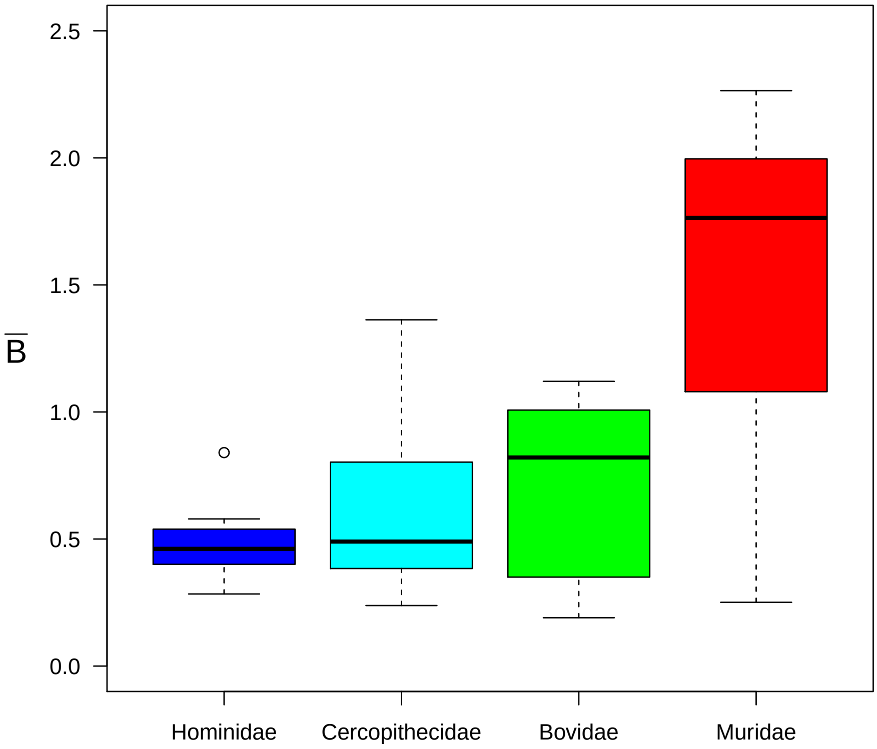

Figure 3: Distribution of the average estimated B (M3sh model). One outlying data point is missing from the figure: the estimated average B was 3.86 in the (*Capra hircus*, *Ovis aries*) ancestral branch (Bovidae).

The across-genes average gBGC intensity, $\bar{B}$, varied among models, with models allowing for more variation in B usually yielding a higher $\bar{B}$ (Supplementary Table S1). Below we report estimates of $\bar{B}$ obtained under the M3sh model. These were very similar to estimates obtained by averaging $\bar{B}$ across the M1, M2, M3z, M3sh and M3h models, weighting by the AIC of each model [63, Supplementary Fig. S2].

Fig.3 shows the distribution of $\bar{B}$ among branches in the four analysed families. The median $\bar{B}$ was just below 0.5 in primates, 0.82 in Bovidae and 1.76 in Muridae. We calculated the across genes relative standard deviation (RSD) of B, which is the ratio of the standard deviation by the average B. The RSD would be expected to be constant across branches if the across-genes distribution of the intensity of gBGC only differed among branches by a coefficient of proportionality. We found that the RSD was generally rather high (median RSD across branches: 1.8), and substantially smaller in Muridae (median: 1.2) than in the other three families (median Bovidae: 1.8; median Hominidae: 1.7; median Cercopithecidae: 2.1). This suggests that the intensity of gBGC is more evenly distributed among genes in Muridae than in the other taxa. Of note, this result superficially appears to contradict the analysis illustrated by Fig.2, which shows that the clustering of WS substitutions is maximal in Muridae. Importantly, the Moran's I analysis (Fig.2) addresses the within-gene clustering of substitutions, whereas in the RSD analysis we consider the among-gene variation in gBGC intensity.

We estimated in each branch the number of WS substitutions that would be expected in the absence of gBGC. This was achieved by forcing $B = 0$ for all categories of genes under the M3sh model (see Methods). We found that gBGC results in a substantial excess of WS substitutions, which varies from typically 30% in primates to typically 60% in Muridae (Supplementary Fig. S3). No effect of gBGC on SW substitutions is expected under the M3sh model (see Methods).

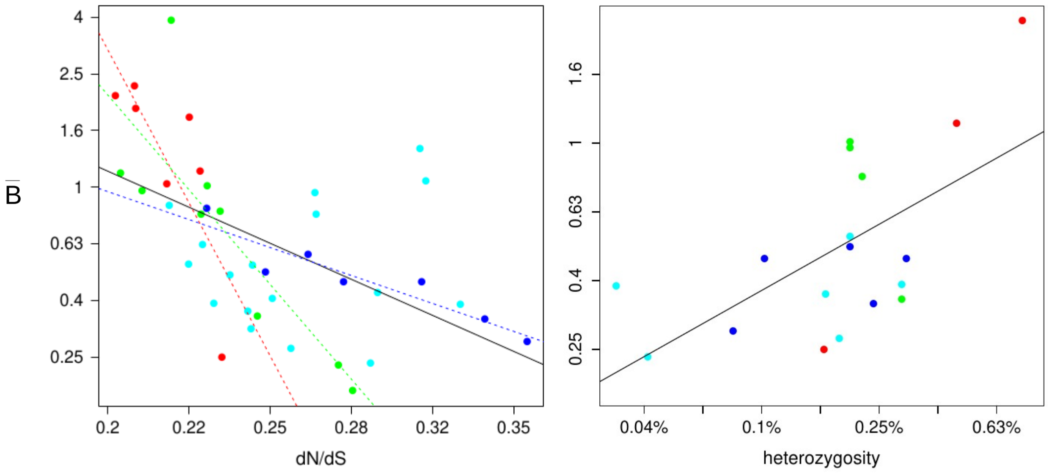

Figure 4: Relationship between the average estimated B and dN/dS (left panel, $n = 40$ branches) or heterozygosity (right panel, $n = 18$ species) across 40 mammalian lineages in log-transformed scales. B was estimated under the M3sh model. Blue: Hominidae; cyan: Cercopithecidae; green: Bovidae; red: Muridae; black line: regression line for the whole data set; colored dotted lines: family-specific regression lines

2.4 Correlates of B

We correlated the log-transformed estimated $\bar{B}$ with log-transformed branch-specific d_N/d_S ratio and found a significantly negative relationship ($n = 40$; $r^2 = 0.24$; p-val=0.0013). The correlation coefficient of the $\bar{B}$ vs. d_N/d_S relationship was also significantly negative when calculated within Hominidae ($n = 7$; $r^2 = 0.83$; p-val=0.0043), within Bovidae ($n = 9$; $r^2 = 0.79$; p-val=0.0013) and within Muridae ($n = 7$; $r^2 = 0.57$; p-val=0.049). No significant relationship was detected within Cercopithecidae (fig.4, left). Very similar results were obtained when we correlated the estimated $\bar{B}$ with d_N/d_S calculated based on SSWW substitutions only, i.e., a statistics essentially independent of gBGC: the squared correlation coefficients were 0.82 (p-val=0.0047), 0.68 (p-val=0.0059) and 0.73 (p-val=0.0139) within Hominidae, Bovidae and Muridae, respectively, and 0.23 (p-val=0.0017) for the whole data set (all variables log-transformed). A literature search yielded estimates of heterozygosity (*i.e.*, within-species genetic diversity), π, in 18 species of our data set. The estimated $\bar{B}$ was positively correlated with π ($r^2 = 0.73$; p-val=2.1×10^{-5}; fig.4, right). The sample size was here too small to investigate the within-family relationships. B was also found to be negatively correlated with species longevity ($r^2 = 0.36$, p-val=0.0026) and log-transformed body mass ($r^2 = 0.22$, p-val=0.017).

2.5 Substitution clustering conditional on B

Fig.2 revealed virtually no clustering of WS substitutions in Hominidae, even though the analysis of substitution counts demonstrated a significant impact of gBGC on coding sequences in this family (Fig.3). To test whether the spatial distribution of WS substitutions really differs between mammalian families, we analysed substitution clustering conditional on B. For each branch, we first fitted to WS, SW and SSWW substitution counts a gBGC model, M5f, assuming five categories of genes undergoing distinct gBGC intensities, from $B = 0$ in the coldest category to $B = 10$ in the hottest one. We assigned each gene to one of these gBGC intensity categories, and calculated the average Moran's I for WS substitutions separately for the five categories (Fig.5). This was also done using the hotspot version of this five-category model, M5shf (Supplementary Fig. S4). The size of dots in Fig.5 and Supplementary Fig. S4 reflects the proportions of the five classes of genes in each family, genes from distinct branches being here merged. We found that the average Moran's I increased with gBGC intensity, as expected, but varied strongly among families in every gBGC category, with Muridae consistently showing the highest average Moran's I, and Hominidae the lowest, at all gBGC intensities. This result indicates that the level of clustering of WS substitutions differs across families to an extant that cannot be explained just by differences in average B.

3 Discussion

Analysing the substitution pattern in coding sequences across four mammalian families, we checked two predictions of the gBGC model, namely a clustering of WS substitutions and an excess of WS over SW substitutions compared to the mutation pattern. Both approaches revealed a conspicuous effect of gBGC in mammalian coding sequences.

3.1 Ubiquitous gBGC in mammals

Dreszer et al. [19] investigated the substitution pattern in the human genome and showed that clusters or nearby substitutions tend to be enriched in the WS sort. The effect, although significant, was not particularly strong: the proportion of WS substitutions in clusters reached 0.55, whereas it was 0.44 on average (their figure 1A). Analyzing exon evolution in apes, Berglund et al. [3] and Galtier et al. [32] identified a few dozens of GC-biased exons, out of >10,000 analyzed exons. Here we applied a distinct but related approach to mammalian coding sequences, and reveal only a weak, if any, tendency for WS synonymous substitutions to be clustered in Hominidae, consistent with previous research. The trend, however, was obvious in Cercopithecidae, and strong in Bovidae and Muridae (Fig.2). Our simulations suggest that >15%, and maybe up to 40%, of WS synonymous substitutions appear as clusters in these two families. These WS substitution clusters likely reflect a localized effect of gBGC at recombination hotspots. Here we show that such clusters, although anecdotical in humans and apes, are a major component of the substitution pattern in other families of mammals.

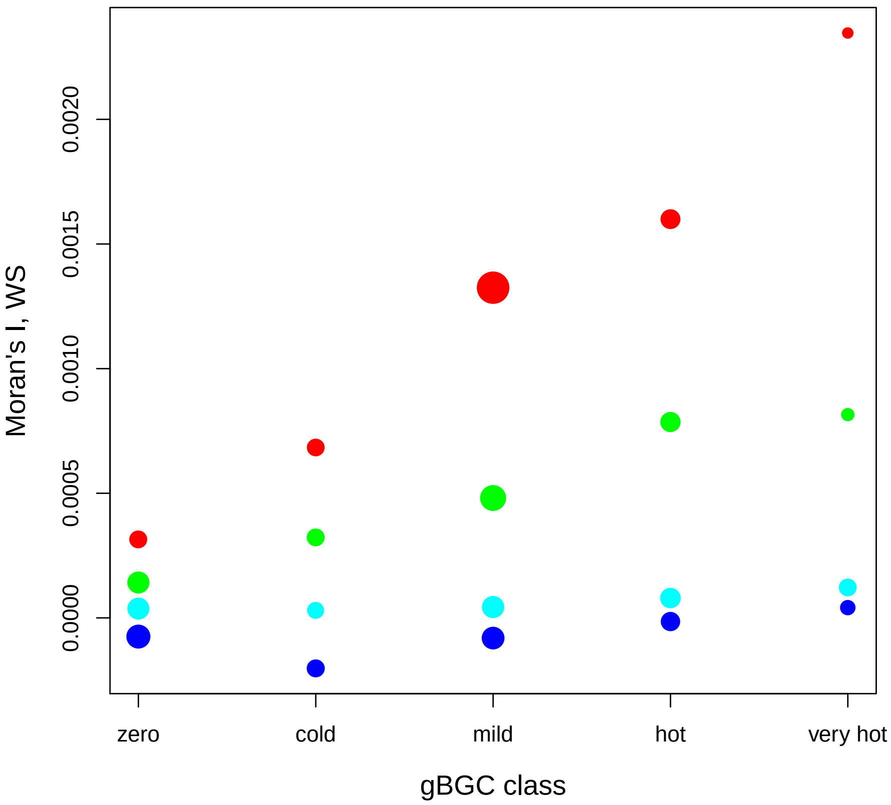

Figure 5: Average centered Moran's I as a function of average estimated gBGC strength B. Each dot is for a class of genes in a family (all species merged). Genes were assigned to classes under the M5f model. Dot size reflects the relative number of genes in the considered class. Blue: Hominidae; cyan: Cercopithecidae; green: Bovidae; red: Muridae.

Our estimate of the B parameter, which measures the average intensity of gBGC across genes, varied between 0.2 and 3.9 among the 40 analysed lineages. In primates, the median estimated $\bar{B}$ was ~ 0.5, i.e., in the range of previously published values: 0.1 in hominoids [43], 0.38 in humans [35], 0.35-0.7 in apes [18]. Our estimates of $\bar{B}$ in Bovidae ($\sim 0.5 - 1$) and Muridae ($\sim 1 - 2$) are also quite similar to those obtained by Lartillot [43] in the *Bos taurus* (Bovidae), *Mus musculus* and *Rattus rattus* (Muridae) lineages. In Bovidae, we found a positive relationship between the estimated $\bar{B}$ and branch age (in million years), defined as the average between the date of the top and bottom nodes of a branch ($n = 9$ branches; $r^2 = 0.75$; p-val=0.003). This is consistent with the hypothesis of a high ancestral N_e in this taxon, as also suggested by fossil data and d_N/d_S-based reconstructions [26, 27]. Our study could not confirm the report by Romiguier et al. [69] and Lartillot [43] of a particularly strong gBGC in bats, tenrecs and lagomorphs due to the unavailability of fully-sequenced, closely related species in sufficient numbers in these taxa.

The estimated genome average B was in the nearly neutral zone in the four families analysed here. Even so, gBGC was found to be pervasive and strongly impact the substitution process in coding sequences. Fitting a three-category, hotspot model across genes, we estimate that 30 to 60% of the WS synonymous substitutions can be attributed to gBGC in mammals. It should be noted that this estimate, as well as the estimates of B we report in this work, is dependent on the assumption of a known and constant mutation process. Here we used the WS, SW, SS and WW mutations rates obtained from Smith et al. [75], who analysed >130,000 *de novo* mutations inferred from mother/father/child trios in humans - a very large data set. Milholland et al. [52] compared the germline mutation pattern of *H. sapiens* and *M. musculus* and did not detect any conspicuous difference between the two species in the proportions of SW, WS and SSWW mutations, and neither did Wang et al. [79] when comparing *Macaca mulatta* (Cercopithecidae) to *H. sapiens*. So the existing literature does not seem to question our assumption of constant relative mutation rates in mammals - but note that no such data is available in Bovidae, to our knowledge. Also note that the clustering analysis (Fig.2) does not make any assumption regarding the mutation process.

It should be noted that our estimate of B in this analysis is based on substitution counts inferred via a parsimony-based approach. This way of counting substitutions is not devoid of potential problems. Maximum parsimony substitution inference is known to be biased towards common-to-rare changes [21]. However, the relatively recent divergence times we are considering presumably keeps this effect to a minimum. Indeed the longest branch across all four trees (Fig. 1), the *Rattus norvegicus* terminal branch, has a length below 0.08 substitutions per site, which is the minimal length for which this problem was detectable in [21]. Using closely related species, on the other hand, runs into another potential bias: when closely related species are analysed, there is a risk that within-species polymorphism contributes a non-negligible fraction of the observed sequence variation, biasing the estimation of quantities such as the d_N/d_S ratio [56]. This bias likely affects the estimation of the relative SW and WS substitution rates as well, since the expected SW/WS rate ratio differs between polymorphism and divergence when gBGC is at work. More work would be needed to confirm and quantify the effect of this bias on our analysis.

3.2 A significant effect of the effective population size

Although a significant effect of gBGC was detected in all four analysed families, its intensity varied conspicuously among families, Muridae being the most strongly impacted, followed by Bovidae, Cercopithecidae, and Hominidae. It is noticeable that gBGC ranks family in the same order as N_e, as measured by the family-average d_N/d_S ratio, this order being consistently recovered in nearly all the analyses we performed. This is in line with the expectation that the intensity of gBGC should be higher in large than in small populations. An effect of N_e was also detected by correlating $\bar{B}$ with the d_N/d_S ratio across branches, and with heterozygosity across species (Fig.4). These analyses confirmed the significance of the effect, both among and within families, thus corroborating the relationship uncovered by Lartillot [43] at a deeper phylogenetic scale.

Interestingly, the slope of the $log(\bar{B})$ vs. $log(d_N/d_S)$ relationship differed conspicuously among families in Fig.4. This result might tell something about the strength of selection on amino-acid changing mutations in mammalian coding sequences. Welch et al. [81] showed that, assuming a Gamma distribution of deleterious effects of non-synonymous mutations, the d_N/d_S ratio is expected to be proportional to ${N_e}^{-\beta}$, where β is the shape parameter of the Gamma distribution. So under this assumption, and since B is proportional to N_e, the slope of the $log(\bar{B})$ vs. $log(d_N/d_S)$ relationship should equal $-1/\beta$. This rationale yields estimates of β equal to 0.51 in Hominidae, 0.15 in Bovidae, and 0.09 in Muridae. These figures differ considerably from estimates obtained by site frequency spectrum analyses, *i.e.*, $\beta \sim 0.15$ in Hominidae and ~ 0.2 in Muridae [9, 29, 38]. More work is needed to understand the origin and meaning of this discrepancy. At any rate, our results suggest that gBGC analysis could constitute a new source of information on the variation in N_e among species, and might enrich the ongoing discussion on this issue [e.g 7, 29].

The among lineages correlation between the estimated $\bar{B}$ and the d_N/d_S ratio we report here in mammals, which confirms Lartillot [43]'s results, contrasts with the absence of such a correlation at the Metazoa scale. Large-N_e fruit flies and marine molluscs, for instance, are less strongly impacted by gBGC than small-N_e bees and amniotes [33, 68, 78]. The simplest explanation for this is that b, the transmission bias, probably differs much between distantly related taxa, due to differences in recombination rate, repair bias and/or conversion tract length. For instance, the recombination rate is known to be particularly high in honey bees [82].

Two recent studies experimentally assessed the intensity of gBGC in mice via crosses followed by sperm [34] or progeny [47] whole genome sequencing. Both estimated that b is lower in mice than in humans - maybe five times lower, although this figure requires confirmation. Gautier [34] invoked purifying selection against gBGC to explain this result. Indeed, because of its deleterious effects [3, 32, 59], gBGC as a process could be counter-selected, and purifying selection being more effective in large than in small population, this verbal model would predict a lower b in mice than in human. Adapted to our results, this hypothesis of a negative correlation between N_e and b would imply that the range of variation in B should be narrower than the range of variation in N_e among mammalian lineages. We indeed observed a narrower variation in the magnitude of the estimated $\bar{B}$ (standard deviation of $log(\bar{B})$: 0.63) than of heterozygosity (standard deviation of $log(\pi)$: 0.74; same 18 species used in these two calculations, see Fig.4, right panel). The difference, however, is not particularly pronounced, and does not suggest the existence of a strong negative relationship between b and N_e in mammals. That said, not only N_e influences the variation of π: the mutation rate also matters. Among-species differences in per generation mutation rate, if any, should be taken into account for a better assessment of the b vs. N_e relationship.

3.3 Recombination hotspots dynamics

The among-gene variation in gBGC intensity, measured by the relative standard deviation of B, was found to be substantial in all branches, while lower in Muridae than in the other three families - a pattern also reported by Lartillot [43]. gBGC seems to be more evenly distributed across the genome in this taxon, consistent with previous reports that GC3 in murid rodents has been increasing and was homogenised since the common ancestor of this family [10, 54, 67, 69]. Muridae appears to be a peculiar group of mammals with this respect [69], and one should keep this in mind when interpreting patterns of gBGC-related evolution in this taxon. Of note, the existing literature does not suggest that the recombination map is less heterogeneous in mouse or rat than in primates [6, 40, 50].

The within-gene heterogeneity in B, in contrast, was more pronounced in Muridae than in Bovidae and, particularly, primates (Fig.2), and this was true even when we controlled for gene-specific B (Fig.5): for any particular intensity of gBGC, WS substitutions tend to be more clustered in large-N_e than in small-N_e species. This intriguing result might be interpreted in relation with the dynamics of recombination hotspots. To result in a cluster of WS substitutions, a recombination hotspot must be active during a sufficiently long period of time for several WS alleles to reach a high population frequency. We suggest that in primates gBGC does not generate a pattern of highly clustered WS substitutions because heterozygosity is low and recombination hotspots are short-lived in this taxon. Indeed, recombination hotspots are known to be particularly ephemeral in Hominidae, due to a transmission distortion associated with the Red Queen-like evolution of the major hotspot determining gene PRDM9 [1, 14, 57]. For instance, Lesecque et al. [46] showed that denisovians and modern humans did not share the same recombination hotpots, while the level of divergence between these two genomes is of the order of one synonymous substitution per gene on average [51]. Our results might suggest that the situation differs in other taxa of mammals, maybe in a way related to N_e. At any rate, heterozygosity is higher in large-N_e species, which increases the probability that a given local episode of gBGC results in more than one WS substitutions, irrespective of recombination hotspot lifespan. A deeper understanding of this result would require to account for the many factors influencing the turnover time of PRDM9 alleles and recombination hotspots [45], the length of gene conversion tracts, as well as the population mutation rate and fixation probability of WS mutations.

4 Concluding remarks

Quantifying gBGC in closely related species of mammals, we report a pervasive effect on the nucleotide substitution process, a positive relationship with N_e, and a complex pattern of variation within and among genes. This work also demonstrates that the analysis of gBGC has the potential to illuminate various aspects of molecular evolution, including the distribution of fitness effect of mutations and the dynamics of recombination hotspots. The apparent lack of a N_e effect on gBGC intensity at the Metazoa scale is an unresolved question that requires further quantification of the strength of gBGC in non-vertebrate taxa.

5 Material and Methods

5.1 Sequence data

Mammalian coding sequence alignments were downloaded from the Orthomam v10 database [74]. The four families of mammals represented by at least six species in OrthoMam v10 were selected, namely Hominidae (six species, 11,859 genes), Cercopithecidae (eleven species, 10,834 genes), Bovidae (seven species, 9527 genes), and Muridae (six species, 11,758 genes). In Bovidae, the *Bos indicus* sequences were not considered since this taxon is a subspecies of *Bos taurus*. In all four families the phylogenetic histories of the sampled species are well documented, with the exception of the unresolved relationship between cattle, yak and bison [23, 24, 37, 77, Fig1]. Nodes were dated based on the TimeTree website (http://www.timetree.org/) using the median date estimates.

5.2 Substitution mapping

For each of the four data sets, nucleotide substitutions were mapped to the resolved branches of the trees using a stringent parsimony approach. For any given branch, an X→Y substitution was recorded if and only if all species descending from the considered branch carried state Y, and all other species carried state X. All positions not matching this exact pattern, including positions with missing data or gaps, were disregarded. Branches connected to the root of the tree were excluded, as well as branches whose number of descending species was higher than half the total number of sampled species in the family. A total of 40 distinct branches were considered - seven in Hominidae and Muridae, nine in Bovidae, 17 in Cercopithecidae (Fig.1). For each branch and each coding sequence, the number and positions of non-synonymous and synonymous substitutions were recorded, distinguishing the AT→GC (WS), GC→AT (SW) and GC-conservative (SSWW) sorts. Only synonymous substitutions occurring at third codon positions, and non-synonymous substitutions occurring at first or second codon positions, were counted. For any given branch, substitutions that mapped to consecutive sites were ignored, and genes in which the per base pair substitution rate was higher than ten times the across-genes median rate were discarded (implying that the number of analysed genes could slightly differ among lineages within a family). The last two steps aimed at diminishing the effect of misaligned regions.

5.3 Clustering analysis (synonymous substitutions)

For each branch and each gene of length above 400 bp, we calculated Moran's I index (Moran 1950) separately for WS and SW synonymous substitutions. We used a weight matrix defined as follows: the weight equalled one for any two substitutions distant of 400 bp or less, and zero for any two substitutions more distant than 400 bp. Window widths of 200 bp and 100 bp gave qualitatively similar results. For each branch and each sort of substitutions, Moran's I was averaged across genes, excluding genes with less than three substitutions of the considered sort. Moran's I has a negative expectation of $-1/(l-1)$ under the null hypothesis of no spatial autocorrelation, where l is the number of third codon positions of the considered gene. Here we used the centered version of the statistics, $I + 1/(l-1)$, the expectation of which is zero in the absence of substitution clustering.

5.4 Clustering: simulations

We downloaded from the Ensembl database coding sequence annotations in one representative species per family, namely *Homo sapiens* (Hominidae), *Macaca mulatta* (Cercopithecidae), *Bos taurus* (Bovidae) and *Mus musculus* (Muridae), disregarding coding sequences shorter than 400 bp. Then we simulated substitution data in a hypothetical branch by iteratively sampling the location of third-codon-position substitutions across coding sequences using the following method:
(initiation:) randomly sample the location of the first substitution among the third codon positions of all genes;
(iteration:)
- with probability $1 - p_{clust}$, randomly sample the location of the $(n+1)^{th}$ substitution among the third codon positions of all genes;
- with probability p_{clust}, randomly sample the location of the $(n+1)^{th}$ substitution in the neighborhood of the n^{th} substitution (clustered substitutions).

More precisely, conditional on the n^{th} and $(n+1)^{th}$ substitutions being clustered,
- with probability p_{CO} the $(n+1)^{th}$ substitution was randomly sampled in a window of width l_{CO} centered on the location of the n^{th} substitution, and
- with probability $p_{NCO} = 1 - p_{CO}$ the $(n+1)^{th}$ substitution was randomly sampled in a window of width l_{NCO} centered on the location of the n^{th} substitution.

This was intended to represent the fact that gene conversion tracts associated to crossing-over and non-crossing-over events are of different lengths [13]. If the sampled location of the $(n+1)^{th}$ substitution reached beyond the boundaries of the exon carrying the n^{th} substitution, then the $(n+1)^{th}$ substitution was ignored. Our procedure also accounted for the existence of variation in mutation rate among exons: we assumed that one half of the exons had a mutation rate γ times as high as the other half. We separately simulated WS and SW substitution data, accounting for the distribution of GC-content at third codon positions - hence, the availability of W and S sites - in the four groups. A gene with GC3=90% was 10 times more likely to host a SW substitution than a gene with GC3=10% in our simulations.

Two parameters of the simulation procedure were varied among conditions, namely the per third codon position density of substitutions (taking values in $\{0.0003, 0.001, 0.003, 0.01, 0.03\}$) and the probability p_{clust} for two successive substitutions to be clustered (taking values in $\{0, 0.1, 0.2, 0.3, 0.4\}$). The other parameters were fixed to constant values estimated from the literature. Parameters l_{CO} and l_{NCO} were set to 500 and 40 bp, respectively [13, 39, 47, 83]. Parameter γ was set to 3, ensuring an among-exon mutation rate relative standard deviation of 0.5, in agreement with figure 4 in Smith et al. [75]. Finally, parameter p_{CO} was set to 0.86, according to the following rationale: the CO/NCO odds ratio for the first substitution to occur in a gene conversion tracts is $(l_{CO}n_{CO})/(l_{NCO}n_{NCO})$, where n_{CO} and n_{NCO} are the number of crossing-over and non crossing-over events, respectively; the CO/NCO odds ratio for the second substitution to occur in the same gene conversion tract is l_{CO}/l_{NCO}; so the CO/NCO odds ratio for the occurrence of a pair of clustered substitutions is the product, p, of the two terms above, and $p_{CO} = p/(1+p)$. Using $n_{CO}/n_{NCO} = 0.1$ [2, 12], $l_{CO} = 500$ and $l_{NCO} = 40$ we obtain $p_{CO} = 0.86$.

5.5 Maximum likelihood estimation of gBGC strength

For each branch and each gene, we counted the numbers of inferred WS, SW and SSWW synonymous substitutions at third codon positions. Then we fitted mutation/gBGC/drift models to these observations in the maximum likelihood framework.

Consider a coding sequence of length l evolving in a panmictic diploid population of constant size N_e under neutrality during a period T of time. The expected number of substitutions, n^*, depends on the mutation rate μ and fixation probability f:

$$n^* = 2N_e l \mu f T \tag{1}$$

Assuming a homogeneous gBGC intensity of b, the fixation probability of WS, SW and SSWW mutations can be written as:

$$f_1 = \frac{2b}{1 - e^{-4N_e b}} \tag{2}$$

$$f_2 = \frac{2b}{e^{4N_e b} - 1} \tag{3}$$

$$f_3 = 1/2N_e \tag{4}$$

Here and below, subscript 1, 2 and 3 respectively refer to the WS, SW and SSWW sorts of change. Substituting in equation 1 and only considering third codon positions, we get the expected number of synonymous substitutions of the three sorts:

$$n_1^* = l_W \mu_{WS} \frac{B}{1 - e^{-B}} T \tag{5}$$

$$n_2^* = l_S \mu_{SW} \frac{B}{e^{-B} - 1} T \tag{6}$$

$$n_3^* = (l_W \mu_{WW} + l_S \mu_{SS}) T \tag{7}$$

where $B = 4N_e b$, l_W and l_S are the number of AT- and GC-ending codons, respectively, in the considered coding sequence, and μ_{WS}, μ_{SW}, μ_{SS} and μ_{WW} are the corresponding mutation rates.

Assuming that the number of WS substitutions is Poisson distributed, the probability of observing n WS substitutions given B and T is given by the following function:

$$\phi_1(n|B,T) = \frac{n_1^{*n}}{n!} e^{-n_1^*} \tag{8}$$

and similarly for SW and SSWW substitutions:

$$\phi_2(n|B,T) = \frac{n_2^{*n}}{n!} e^{-n_2^*} \tag{9}$$

$$\phi_3(n|B,T) = \frac{{n_3^*}^n}{n!} e^{-n_3^*} \tag{10}$$

We modelled the variation of B among genes using discrete distributions. Assume there are k categories of genes, with each category including a fraction p_k of the genes and characterised by a population-scaled gBGC intensity B_k, assumed to be constant within genes. For a gene at which n_1, n_2, and n_3 substitutions of the three sorts are observed, the likelihood can be written as:

$$L = \sum_k p_k \phi_1(n_1|B_k,T)\phi_2(n_2|B_k,T)\phi_3(n_3|B_k,T) \tag{11}$$

We considered various models that differ in how B varies across genes. Under model M1, a constant B across genes is assumed. Model M2 defines two categories of genes, each with its own gBGC intensity. Model M3z (for zero) has three categories of genes, among which one has a gBGC intensity of zero, the other two being free parameters. Model M5f (for fixed) has five categories of genes with fixed gBGC intensities equal to $B_1 = 0$, $B_2 = 0.333$, $B_3 = 1$, $B_4 = 3.333$ and $B_5 = 10$, respectively. In all four models the proportions of genes in the various categories were free to vary, T was assumed to be shared among genes and the relative mutation rates μ_{WS}, μ_{SW}, μ_{SS} and μ_{WW} were set to empirical estimates obtained from Smith et al. (2018), i.e., $\mu_{WS} = 5.21$, $\mu_{SW} = 10.90$, $\mu_{SS} = 2.07$ and $\mu_{WW} = 1$.

The models above assume a homogeneous rate of gBGC among positions within a gene. To account for the existence of hotspots of gBGC, we modelled the within-gene variation of B by assuming that only a fraction q of the positions undergo gBGC at rate B_h, the other positions evolving neutrally. Under this assumption, the expected number of WS and SW substitutions are given by:

$$n_1^* = l_W \mu_{WS}[(1-q) + \frac{qB_h}{1-e^{-B_h}}]T \tag{12}$$

$$n_2^* = l_S \mu_{SW}[(1-q) + \frac{qB_h}{e^{B_h}-1}]T \tag{13}$$

while n_3^* is given by equation 7.

Equations 12 and 13 simplify if q is assumed to be much smaller than 1 and B_h much higher than 1; under these assumptions, we have:

$$n_1^* = l_W \mu_{WS}(1+qB_h)T \tag{14}$$

$$n_2^* = l_S \mu_{SW} T \tag{15}$$

Parameters q and B_h only appear as a product in equations 14 and 15, saving one degree of freedom. The simplified equation 15 exhibits the main difference between hotspot and gene-homogeneous models, which concerns SW substitutions. In gene-homogeneous models, the expected number of SW substitutions is decreased in genes experiencing strong gBGC, whereas hotspot models predict nearly no influence of gBGC on the SW substitution rate if q is sufficiently small.

The among-gene variation in gBGC strength was here modeled via categories of genes that differed with respect to the prevalence of hotspots, q, while sharing the same intensity of gBGC within hotspots, B_h. Specifically, we considered a three-category model in which the "coldest" category had no hotspot, i.e., $q_1 = 0$. The fraction of hotspots in the other two categories, q_2 and q_3, and the relative prevalence of the three categories, p_1, p_2 and p_3, as well as T and B_h, were free to vary. This model was called M3h (for hotspot); its predictions are given by equations 12, 13 and 7. A simplified hotspot model, M3sh (for simplified hotspot), was also implemented by instead using equations 14, 15 and 7. M3sh is a special case of M3h assuming that the fraction of sites affected by gBGC within a gene is small. We also considered a simplified five-category hotspot model, M5shf, with fixed values for the $q_k B_h$ product equal to 0, 0.333, 1, 3.33 and 10, respectively.

The overall likelihood was obtained by multiplying the likelihoods of distinct genes. Parameters were estimated in the maximum likelihood (ML) framework. Likelihood maximization was achieved via home-made C++ programs using the Bio++ library [36].

The across gene categories average estimated intensity of gBGC was computed as

$$\bar{B} = \sum_k \hat{p_k} \hat{B_k} \tag{16}$$

under gene-homogeneous models and

$$\bar{B} = \sum_k \hat{p_k} \hat{q_k} \hat{B_h} \tag{17}$$

under hotspot models, where the k index is for gene categories and the hat denotes ML estimate. The across genes standard deviation of the intensity of gBGC was calculated similarly.

Akaike's Information Criterion (AIC) was calculated for all models, the number of estimated parameters being 2, 4, 5, 5, 5, 5 and 6, respectively, for M1, M2, M3z, M5f, M3sh, M5shf and M3h. AIC weights [63] were used to calculate an across-model estimate of the B parameter. The parametrisation of the various models is recapitulated in Table 1.

For each gene, the expected number of WS substitutions in the absence of gBGC was estimated as:

$$m_1^* = l_W \mu_{WS} \hat{T} \tag{18}$$

and the excess WS substitutions due to gBGC were estimated as $(n_1^* - m_1^*)/m_1^*$. No depletion of SW substitutions is expected under M3sh (compare equations 15 and 18). Finally, each gene was assigned to one of the gBGC categories by selecting the category k maximising the following posterior probability:

$$\phi_1(n_1|B_k, T)\phi_2(n_2|B_k, T)\phi_3(n_3|B_k, T) \tag{19}$$

All these calculations were achieved separately for the 40 branches of the data set.

model	cat.^{a}	nb fixed^b	fixed^c	nb optimised^d	optimised^e
all^f			$\mu_{WS}, \mu_{SW}, \mu_{SS}, \mu_{WW}$		T
M1	1	4		2	B
M2	2	4		4	B_1,B_2,p_1
M3z	3	5	B_1	5	B_2,B_3,p_1,p_2
M5f	5	9	B_1-B_5	5	p_1-p_4
M3h	3	5	q_1	6	q_2,q_3,B_h,p_1,p_2
M3sh	3	5	$q_1 B_h$	5	$q_2 B_h$,$q_3 B_h$,p_1,p_2
M5shf	5	9	$q_1 B_h$-$q_5 B_h$	5	p_1-p_4

Table 1: Parametrisation of the models used in this analysis. a : number of categories of genes; b: number of fixed parameters; c: list of fixed parameters; d: number of optimized parameters; e: list of optimized parameters; f: parameters shared by all models

5.6 Additional variables

For every branch, the numbers of non-synonymous substitutions of the WS, SW and SSWW sorts at first and second codon positions were computed, summed across genes, and used to calculate branch-specific d_N/d_S ratios. A literature survey was conducted in search for genome-wide estimates of within-species diversity, or heterozygosity, π. Such estimates were collected in 18 of the analysed species, as reported in Supplementary Table S2. Data on species longevity and body mass were obtained from the AnAge data base [17] and are also reported in Supplementary Table S2.

6 Supplementary Material, Data accessibility

All the data sets, programs and scripts used in this study are available from:
`https://osf.io/fx54q/?view_only=1109ca2f66e74ad99f0d76ac93d40fc5`

7 Acknowledgments

The author is grateful to Laurent Duret, Nicolas Lartillot, Sylvain Glémin, Benoît Nabholz and Jonathan Romiguier for very useful comments, and to Adam Eyre-Walker for sharing data on *de novo* mutation in humans. This work was supported by Agence Nationale de la Recherche project ANR-19-CE12-0019 (HotRec).

References

[1] A Auton, A Fledel-Alon, S Pfeifer, O Venn, L Ségurel, T Street, E M Leffler, R Bowden, I Aneas, J Broxholme, P Humburg, Z Iqbal, G Lunter, J Maller, R D Hernandez, C Melton, A Venkat, M A Nobrega, R Bontrop, S Myers, P Donnelly, M Przeworski, and G McVean. A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336(6078):193–198, 2012.

[2] F Baudat and B de Massy. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res*, 15(5):565–577, 2007.

[3] J Berglund, K S Pollard, and M T Webster. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol*, 7(1):e26, 2009.

[4] P Bolívar, C F Mugal, M Rossi, A Nater, M Wang, L Dutoit, and H Ellegren. Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Mol Biol Evol*, 35(10):2475–2486, 2018.

[5] R Borges, G J Szöllősi, and C Kosiol. Quantifying gc-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*, 212(4):1321–1336, 2019.

[6] H Brunschwig, L Levi, E Ben-David, R W Williams, B Yakir, and S Shifman. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*, 191(3):757–764, 2012.

[7] V Buffalo. Why do species get a thin slice of π? Revisiting Lewontin's paradox of variation. *bioRxiv*, doi: https://doi.org/10.1101/2021.02.03.429633, 2021.

[8] J A Capra, M J Hubisz, D Kostka, K S Pollard, and A Siepel. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet*, 9(8):e1003684, 2013.

[9] D Castellano, M C Macià, P Tataru, T Bataillon, and K Munch. Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. *Genetics*, 213(3):953–966, 2019.

[10] Y Clément and P F Arndt. Substitution patterns are under different influences in primates and rodents. *Genome Biol Evol*, 3:236–245, 2011.

[11] Y Clément, G Sarah, Y Holtz, F Homa, S Pointet, S Contreras, B Nabholz, F Sabot, L Sauné, M Ardisson, R Bacilieri, G Besnard, A Berger, C Cardi, F De Bellis, O Fouet, C Jourda, B Khadari, C Lanaud, T Leroy, D Pot, C Sauvage, N Scarcelli, J Tregear, Y Vigouroux, N Yahiaoui, M Ruiz, S Santoni, J P Labouisse, J L Pham, J David, and S Glémin. Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet*, 13(5):e1006799, 2017.

[12] F Cole, L Kauppi, J Lange, I Roig, R Wang, S Keeney, and M Jasin. Homeostatic control of recombination is implemented progressively in mouse meiosis. *Nat Cell Biol*, 14(4):424–430, 2012.

[13] F Cole, F Baudat, C Grey, S Keeney, B de Massy, and M Jasin. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat Genet*, 46(10): 1072–1080, 2014.

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

[15] P Corcoran, T I Gossmann, HJ Barton, Great-Tit-HapMap-Consortium, J Slate, and K Zeng. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. *Genome Biol Evol*, 9(11):2987–3007, 2017.

[16] J Damuth. A macroevolutionary explanation for energy equivalence in the scaling of body size and population density. *Am Nat*, 169(5):621–631, 2007.

[17] J P de Magalhães and J A Costa. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol*, 22(8):1770–1774, 2009.

[18] N De Maio, C Schlötterer, and C Kosiol. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol Biol Evol*, 30(10):2249–2262, 2013.

[19] T R Dreszer, G D Wall, D Haussler, and K S Pollard. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res*, 17(10):1420–1430, 2007.

[20] L Duret and P F Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, 4(5):e1000071, 2008.

[21] A Eyre-Walker. Problems with parsimony in sequences of biased base composition. *J Mol Evol*, 47 (6):686–690, 1998.

[22] A Eyre-Walker. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, 152(2):675–683, 1999.

[23] P H Fabre, A Rodrigues, and E J Douzery. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol*, 53(3):808–825, 2009.

[24] P H Fabre, L Hautier, D Dimitrov, and E J Douzery. A glimpse on the pattern of rodent diversification: a phylogenetic approach. *BMC Evol Biol*, 12:88, 2012.

[25] E Figuet, M Ballenghien, J Romiguier, and N Galtier. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol Evol*, 7(1):240–250, 2014.

[26] E Figuet, J Romiguier, J Y Dutheil, and N Galtier. Mitochondrial DNA as a tool for reconstructing past life-history traits in mammals. *J Evol Biol*, 27(5):899–910, 2014.

[27] E Figuet, M Ballenghien, N Lartillot, and N Galtier. Reconstruction of body mass evolution in the Cetartiodactyla and mammals using phylogenomic data. *bioRxiv*, doi: https://doi.org/10.1101/139147, peer-reviewed and recommended by Peer Community In Evolutionary Biology, https://evolbiol.peercommunityin.org/articles/rec?id=67, 2017.

[28] N Galtier and L Duret. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends Genet*, 23(6):273–277, 2007.

[29] N Galtier and M Rousselle. How much does n_e vary among species? *Genetics*, 216(2):559–572, 2020.

[30] N Galtier, G Piganeau, D Mouchiroud, and L Duret. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159(2):907–911, 2001.

[31] N Galtier, E Bazin, and N Bierne. GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics*, 172(1):221–228, 2006.

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

[33] N Galtier, C Roux, M Rousselle, J Romiguier, E Figuet, S Glémin, N Bierne, and L Duret. Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Mol Biol Evol*, 35(5):1092–1103, 2018.

[34] Maud Gautier. *La recombinaison comme moteur de l'évolution des génomes : caractérisation de la conversion génique biaisée chez la souris*. PhD thesis, Ecole Doctorale E2M2, Université Claude Bernard Lyon 1, 9 2019.

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

[36] L Guéguen, S Gaillard, B Boussau, M Gouy, M Groussin, N C Rochette, T Bigot, D Fournier, F Pouyet, V Cahais, A Bernard, C Scornavacca, B Nabholz, A Haudry, L Dachary, N Galtier, K Belkhir, and J Y Dutheil. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol*, 30(8):1745–1750, 2013.

[37] A Hassanin, F Delsuc, A Ropiquet, C Hammer, B Jansen van Vuuren, C Matthee, M Ruiz-Garcia, F Catzeflis, V Areskoug, T T Nguyen, and A Couloux. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C R Biol*, 335(1):32–50, 2012.

[38] C D Huber, B Y Kim, C D Marsden, and K E Lohmueller. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci U S A*, 114(17):4465–4470, 2017.

[39] A J Jeffreys and C A May. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*, 36(2):151–156, 2004.

[40] M I Jensen-Seaman, T S Furey, B A Payseur, Y Lu, K M Roskin, C F Chen, M A Thomas, D Haussler, and H J Jacob. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res*, 14(4):528–538, 2004.

[41] J Lachance and S A Tishkoff. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am J Hum Genet*, 95(4):408–420, 2014.

[42] N Lartillot. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol Biol Evol*, 30(2):356–368, 2012.

[43] N Lartillot. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol Biol Evol*, 30(3):489–502, 2013.

[44] F Lassalle, S Périan, T Bataillon, X Nesme, L Duret, and V Daubin. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet*, 11(2):e1004941, 2015.

[45] T Latrille, L Duret, and N Lartillot. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. *Philos Trans R Soc Lond B Biol Sci*, 372(1736):20160463, 2017.

[46] Y Lesecque, S Glémin, N Lartillot, D Mouchiroud, and L Duret. The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS Genet*, 10(11): e1004790, 2014.

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

[48] H Long, W Sung, S Kucukyildirim, E Williams, S F Miller, W Guo, C Patterson, C Gregory, C Strauss, C Stone, C Berne, D Kysela, W R Shoemaker, M E Muscarella, H Luo, J T Lennon, Y V Brun, and M Lynch. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol*, 2(2):237–240, 2018.

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

[50] G A McVean, S R Myers, S Hunt, P Deloukas, D R Bentley, and P Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, 2004.

[51] M Meyer, M Kircher, M T Gansauge, H Li, F Racimo, S Mallick, J G Schraiber, F Jay, K Prüfer, C de Filippo, P H Sudmant, C Alkan, Q Fu, R Do, N Rohland, A Tandon, M Siebauer, R E Green, K Bryc, A W Briggs, U Stenzel, J Dabney, J Shendure, J Kitzman, M F Hammer, M V Shunkov, A P Derevianko, N Patterson, A M Andrés, E E Eichler, M Slatkin, D Reich, J Kelso, and S Pääbo. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338 (6104):222–226, 2012.

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

[53] P A Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1):17–23, 1950.

[54] D Mouchiroud, C Gautier, and G Bernardi. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J Mol Evol*, 27(4):311–320, 1988.

[55] C F Mugal, P F Arndt, and H Ellegren. Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Mol Biol Evol*, 30(7):1700–1712, 2013.

[56] C F Mugal, V E Kutschera, F Botero-Castro, J B W Wolf, and I Kaj. Polymorphism data assist estimation of the nonsynonymous over synonymous fixation rate ratio ω for closely related species. *Mol Biol Evol*, 37(1):260–279, 2020.

[57] S Myers, R Bowden, A Tumian, R E Bontrop, C Freeman, T S MacFie, G McVean, and P Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

[58] B Nabholz, A Künstner, R Wang, E D Jarvis, and H Ellegren. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol Biol Evol*, 28(8):2197–2210, 2011.

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

[60] S I Nikolaev, J I Montoya-Burgos, K Popadin, L Parand, Margulies EH, and S E Antonarakis. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A*, 104(51):20443–20448, 2007.

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

[62] K Popadin, L V Polishchuk, L Mamirova, D Knorre, and K Gunbin. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A*, 104(33):13390–13395, 2007.

[63] D Posada and T R Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*, 53(5): 793–808, 2004.

[64] R Pracana, A D Hargreaves, J F Mulley, and P W H Holland. Runaway GC evolution in gerbil genomes. *Mol Biol Evol*, 37(8):2197–2210, 2020.

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

[66] A Ratnakumar, S Mousset, S Glémin, J Berglund, N Galtier, L Duret, and M T Webster. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci*, 365(1552):2571–2580, 2010.

[67] M Robinson, C Gautier, and D Mouchiroud. Evolution of isochores in rodents. *Mol Biol Evol*, 14 (8):823–828, 1997.

[68] M C Robinson, E A Stone, and N D Singh. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. *Mol Biol Evol*, 31(2):425–433, 2014.

[69] J Romiguier, V Ranwez, E J Douzery, and N Galtier. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*, 20 (8):1001–1009, 2010.

[70] J Romiguier, E Figuet, N Galtier, E J Douzery, B Boussau, J Y Dutheil, and V Ranwez. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One*, 7(3):e33852, 2012.

[71] J Romiguier, V Ranwez, E J Douzery, and N Galtier. Genomic evidence for large, long-lived ancestors to placental mammals. *Mol Biol Evol*, 30(1):5–13, 2013.

[72] J Romiguier, J Lourenco, P Gayral, N Faivre, L A Weinert, S Ravel, M Ballenghien, V Cahais, A Bernard, E Loire, L Keller, and N Galtier. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *J Evol Biol*, 27(3):593–603, 2014.

[73] M Rousselle, A Laverré, E Figuet, B Nabholz, and N Galtier. Influence of recombination and GC-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. *Mol Biol Evol*, 36(3):458–471, 2019.

[74] C Scornavacca, K Belkhir, J Lopez, R Dernat, F Delsuc, E J P Douzery, and V Ranwez. Orthomam v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol*, 36(4):861–862, 2019.

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

[76] C C Spencer, P Deloukas, S Hunt, J Mullikin, S Myers, B Silverman, P Donnelly, D Bentley, and G McVean. The influence of recombination on human genetic diversity. *PLoS Genet*, 2(9):e148, 2006.

[77] D Vanderpool, B Q Minh, R Lanfear, D Hughes, S Murali, R A Harris, M Raveendran, D M Muzny, M S Hibbins, R J Williamson, R A Gibbs, K C Worley, J Rogers, and M W Hahn. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biol*, 18(12):e3000954, 2020.

[78] A Wallberg, S Glémin, and M T Webster. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genet*, 11(4):e1005189, 2015.

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

[80] M T Webster and N G Smith. Fixation biases affecting human SNPs. *Trends Genet*, 20(3):122–126, 2004.

[81] J J Welch, A Eyre-Walker, and D Waxman. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J Mol Evol*, 67(4):418–426, 2008.

[82] L Wilfert, J Gadau, and P Schmid-Hempel. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity (Edinb)*, 98(4):189–197, 2007.

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