RESEARCH ARTICLE

Unusual mammalian usage of TGA stop codons reveals that sequence conservation need not imply purifying selection

Alexander Thomas Ho*, Laurence Daniel Hurst

Milner Centre for Evolution, University of Bath, Bath, United Kingdom

* a.t.ho@bath.ac.uk

OPEN ACCESS

Citation: Ho AT, Hurst LD (2022) Unusual mammalian usage of TGA stop codons reveals that sequence conservation need not imply purifying selection. PLoS Biol 20(5): e3001588. https://doi.org/10.1371/journal.pbio.3001588

Academic Editor: Claudia Bank, University of Bern, SWITZERLAND

Received: February 17, 2022

Accepted: April 20, 2022

Published: May 12, 2022

Data Availability Statement: The raw data used are all publicly available from the repositories outlined in the methods section. All data manipulation was performed using bespoke Python 3.6 scripts, while statistical analyses and data visualisations were performed using R 3.3.3. These scripts are hosted at https://github.com/ath32/gBGC where they are organised into folders corresponding to each part of the methods section. The underlying data for each figure may be found in the corresponding source data as outlined in the figure legends.

Abstract

The assumption that conservation of sequence implies the action of purifying selection is central to diverse methodologies to infer functional importance. GC-biased gene conversion (gBGC), a meiotic mismatch repair bias strongly favouring GC over AT, can in principle mimic the action of selection, this being thought to be especially important in mammals. As mutation is GC→AT biased, to demonstrate that gBGC does indeed cause false signals requires evidence that an AT-rich residue is selectively optimal compared to its more GC-rich allele, while showing also that the GC-rich alternative is conserved. We propose that mammalian stop codon evolution provides a robust test case. Although in most taxa TAA is the optimal stop codon, TGA is both abundant and conserved in mammalian genomes. We show that this mammalian exceptionalism is well explained by gBGC mimicking purifying selection and that TAA is the selectively optimal codon. Supportive of gBGC, we observe (i) TGA usage trends are consistent at the focal stop codon and elsewhere (in UTR sequences); (ii) that higher TGA usage and higher TAA→TGA substitution rates are predicted by a high recombination rate; and (iii) across species the difference in TAA <-> TGA substitution rates between GC-rich and GC-poor genes is largest in genomes that possess higher between-gene GC variation. TAA optimality is supported both by enrichment in highly expressed genes and trends associated with effective population size. High TGA usage and high TAA→TGA rates in mammals are thus consistent with gBGC's predicted ability to "drive" deleterious mutations and supports the hypothesis that sequence conservation need not be indicative of purifying selection. A general trend for GC-rich trinucleotides to reside at frequencies far above their mutational equilibrium in high recombining domains supports the generality of these results.

Introduction

If at a given site in DNA a mutation appears in a population and is eliminated by selection owing to its deleterious effects, the site in question will tend to be more conserved between species than comparable neutrally evolving sequence. This simple logic underpins the notion that the functionality of sequence can be inferred from its degree of conservation—for

Funding: This work was supported by the European Research Council (grant EvoGenMed ERC-2014-ADG 669207 to L.D.H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: CDS, coding sequence; CRE, *cis*-regulatory element; gBGC, GC-biased gene conversion; HEG, highly expressed gene; HRG, highly recombining gene; LEG, lowly expressed gene; LRG, lowly recombining gene; ncRNA, noncoding RNA; (O-E)/E, (Observed-Expected)/Expected; PGLS, phylogenetically generalized least squares; pTGA, predicted TGA usage; TR, translational readthrough.

discussion, see Ponting [1]. It is explicit in, for example, molecular evolutionary tests for purifying selection (e.g., Ka/Ks test [2–5]) attempts to identify sites prone to disease-causing mutations [6,7], and estimates of the proportion of DNA within a genome that is "functional" [8].

These methods assume, however, that no force other than selection can deterministically act to alter the frequency of extant alleles. Over the past 2 decades, GC-biased gene conversion (gBGC) has been established as a potentially important influence on allele frequencies [9], mimicking selection [10–12]. The process of gBGC results from a repair bias favouring G/C alleles over A/T alleles during GC:AT mismatch repair in a (commonly assumed to be meiotic) heteroduplex [13,14]. In humans, at non-crossover gene conversion events 67.6% of GC:AT mismatches favour the GC allele [15]. It is probably as a consequence of this bias, coupled with the regionalisation of recombination domains over extended time periods, that mammals, alongside birds and possibly other amniotes [16], have genomes with large (>300 Mb) blocks of relatively homogeneous higher or lower GC content (isochores) [10,11,17]. Importantly, assuming consistency of local recombination rates over evolutionary time and a correlation between crossover rates and non-crossover rates [18], gBGC also can explain the relatively strong correlation between GC content of these blocks and local recombination rates in mammals [19–22] (but see also [23,24]). Consistent with such models, SNP analysis reveals the predicted fixation bias for AT→GC mutations in GC-rich domains, even after allowing for nonequilibrium GC content [25,26].

While the human conversion bias is strong, defining the expected impact of gBGC on the human genome is not trivial. For example, in any given generation, the net effect of bias is a function of the length of the relevant conversion tracts, the commonality of AT:GC mismatches within the tracts and the rate of initiation of such tracts. Williams and colleagues [18] estimate a mean rate in human non-crossover events (where there is the strong GC:AT bias) of 5.9×10^{-6} per bp per generation. More generally, Glemin and colleagues [27] estimate that the net effect on substitutions is on average in the nearly neutral area. However, as recombination occurs primarily within recombination hotspots approximately 2% of the human genome is subject to strong gBGC in any generation [27]. Over the longer term, as the location of recombination hotspots evolves rapidly, they predict that a large fraction of the genome is affected by short episodes of strong gBGC [27]. Galtier [28] estimates that approximately 60% of all synonymous AT→GC substitutions are influenced by gBGC.

Strong gene conversion is, however, not phylogenetically universal. In the best-resolved instance, yeast, where meiotic tetrads can be directly studied, the bias is extremely weak at best. The highest estimates suggests that the GC allele is the donor allele in 50.62% of cases [11,29]. Further analysis report a lesser bias [30], with a further large study reporting weak bias in the opposite direction [31]. Meta-analysis of over 100,000 GC:AT mismatch resolutions in *Saccharomyces cerevisiae* determined a net segregation of 50.03%, only just in favour of the GC alleles and not significantly different from 50:50 segregation [31]. To date, strong conversion has been observed in only a few taxa [31], mammals [11], and birds [32,33], being the 2 well-described exceptions, though weaker and nonregionalised gBGC is suspected in many taxa [21].

In terms of the population genetical influence, the action of gBGC is directly comparable to meiotic drive (alias segregation distortion) [34]. In this sense, gBGC may be said to "drive" alleles. In turn, such drive can mimic positive selection [35]. Importantly, it has previously been noted that gBGC can (and in birds and mammals regularly does) create false signals of positive selection by promoting the spread from rare to common of AT→GC mutations [12,36–40]. However, as is implicit in all such models [41], gBGC could also mimic the action of purifying selection. A GC allele at fixation mutating to a selectively advantageous AT allele would be forced by gBGC to eliminate the AT allele, causing conservation of the deleterious GC allele.

Mimicry of positive selection owing to gBGC in mammals is thought to be common and, to date, analyses have focused on the substitutional process, rather than the conservation process [12,36–40]. We are aware of no clear example of gBGC causing false signals of purifying selection. A core difficulty is finding a circumstance where gBGC makes predictions different from those of mutation bias and selectionist models. Differentiating between the effects of gBGC and mutation bias tends to be relatively straightforward, as mutation is near-universally GC→AT biased [42–46], while gBGC is biased in the opposite direction. More problematic is the possibility that the GC state is also the selectively optimal state. If so, then both gBGC and selection make the same predictions of conservation of GC and covariation with the recombination rate. Given Bengtsson's argument that gBGC may be biased in this direction to counter a deleterious GC→AT biased mutational process [47], it may well be unusual to have the selectively optimal state being promoted by mutation bias but not by gBGC. Indeed, in *Drosophila*, for example, "optimal" codons tend to end in G or C [48]. Codon optimality may also not be adequate to define the direction of selection; however, as such selection may also be contingent on the overall GC-richness of the sequence (owing to RNA structure effects [41]). Thus, the core difficulty in establishing gBGC as a cause of false signals of purifying selection and of conservation of deleterious alleles is to identify a case where we can have confidence (and independently verify) that the AT state is selectively optimal compared to its GC-richer allele.

Here, we suggest that mammalian stop codon usage may provide an exceptional test case. Across all domains of life, the 3 stop codons, TAA, TGA, and TAG, are not used equally [49], with TAA being commonly, if not universally, selectively favoured [49]. This is probably owing, in large part, to selective avoidance of translational readthrough (TR). During TR, the stop codon is missed by its cognate release factor [50] due to the misbinding of a near-cognate tRNA [51,52], leading to the erroneous translation of the 3′ UTR and the generation of potentially deleterious protein products [53]. Each stop codon has a distinct intrinsic error rate such that TGA>TAG>TAA in bacteria [54–59] and eukaryotes [55,60] (including humans [61]). TR rate reduction in any given gene might thus be achieved by selection for TAA.

Evocation of such selection presumes that TR is usually deleterious [62,63]. This is likely as the formation of C-terminal extensions cause energetic wastage [64] as well as problems with protein stability [65–67], aggregation [68,69], and localisation [70,71]. Alternatively, in the absence of another 3′ in-frame stop codon, both the readthrough transcript and nascent protein are likely to be degraded when the translational machinery reaches the polyA+ tail [72,73]. In addition to reducing TR costs, TAA also has several other benefits: There may be selection for fast release of the ribosome to prevent ribosomal traffic jams [74], and it is robust to 2 mistranscription events (TAA→TGA, TAA→TAG) while the 2 other stop codons are resilient to just one (TGA→TAA, TAG→TAA).

It is then noteworthy that stop codon usage in mammals is different from that seen elsewhere [49,75]: TGA is more often conserved than TAA [76] and, unusually, the substitution rate of TAA→TGA is higher than the reverse [49]. Despite the fact that in humans, TAA is disproportionately employed in highly expressed genes (HEGs) [77]; this signal of conservation has been interpreted as evidence that purifying selection is operating to preserve TGA in mammals [76]. Gene conversion would, however, oppose fixation of TGA→TAA mutations (while also favouring TAA→TGA) and hence mimic purifying selection on TGA, even if selection were operating in the opposite direction. Biased gene conversion, thought to be especially influential in humans [15], could thus resolve the exceptionalism of TGA conservation in mammals.

Here, we evaluate this suggestion. Duret and Galtier [11] provide a series of tests for differentiating gBGC from selection, noting that the trend to the higher GC state should be correlated with recombination and common to all sites regardless of functional status. We consider several analyses that examined these predictions finding all to be robustly supported. However, to be confident that TAA underusage at the focal stop codon is indeed maladaptive, we also need evidence that TAA is the optimal stop codon. We consider several tests, all of which support this. Finally, we show that complex mutational biases cannot fully explain the TAA/TGA usage trends and confirm a general pattern for GC-rich trinucleotides to reside at frequencies far above their mutational equilibria in GC-rich (high recombining) domains. The latter results are consistent with broadscale patterns of conservation of GC-rich residues owing to gBGC. The same analysis resolves the trinucleotide usage in domains not likely to be subject to gBGC is as expected from a model of complex mutation bias. Indeed, these models predict higher TGA usage than TAG usage in these domains. However, different trinucleotides of same nucleotide content (such as TGA and TAG) have repeatable differences in the extent to which they are subject to fixation bias in GC-rich isochores. The cause of these previously unknown complex fixation biases is unresolved.

Results

Bias towards TGA usage is also evident in the 5′ and 3′ UTR

The gBGC hypothesis predicts that, because the AT→GC bias in the mismatch repair process is nonspecific to terminating stop codons, stop codon usage at the focal stop need not be greatly different to usage of the same trinucleotides seen elsewhere in the genome. To address this, we analyse "stop" codon usage at the focal termination site and in human 5′ and 3′ UTR sequences irrespective of reading frame. This controls for effects of transcription-coupled mutational bias. A model supposing that TGA stop codons are optimal in mammals predicts the patterns of stop codon usage as a function of GC content should not be seen in 5′ and 3′ UTR sequence.

We first establish how intronic GC, as a proxy for isochore GC, covaries with stop codon usage at the focal termination codon. Consistent with the observations of Seoighe and colleagues [76] and Belinky and colleagues [49], we find TGA to be the most common stop in the primate lineage (Fig 1). Not only is TGA the most common stop, but also its usage significantly and positively covaries with intronic GC content in humans when both metrics are calculated in 10% percentile bins (n = approximately 1,000 genes) (Spearman's rank; $p < 2.2 \times 10^{-16}$, rho = 0.99, $n = 10$). TAG usage is also correlated with intronic GC content (Spearman's rank; $p = 0.0014$, rho = 0.89, $n = 10$). TAA frequency is negatively correlated with intronic GC content (Spearman's rank; $p < 2.2 \times 10^{-16}$, rho = –0.99, $n = 10$). As predicted by a gBGC model, we see the same trends in noncoding sequences. TAA frequency is negatively correlated with intronic GC content in both 5′ and 3′ UTR sequence (Spearman's rank; both $p < 2.2 \times 10^{-16}$, both rho = –0.99, $n = 10$). TGA is positively correlated with intronic GC content in both 5′ and 3′ UTR sequence (Spearman's rank; both $p < 2.2 \times 10^{-16}$, both rho = 1, $n = 10$). TAG is uncorrelated with intronic GC content in both 5′ (Spearman's rank; $p = 0.10$, rho = 0.55, $n = 10$) and 3′ UTR sequence (Spearman's rank; $p = 0.61$, rho = 0.19, $n = 10$). Analysis on a gene-by-gene basis (instead of using binned data) using linear regression models supports these conclusions, and the same trends in stop codon usage can be seen in intronic sequence against GC3 (GC3 being used in this circumstance as intronic stop usage predicted by intronic GC would be nonindependent; S1 Table). This is strong evidence that the trends in canonical stop usage are approximately the same as the trends in stop usage outside of the canonical termination context.

High TGA usage is strongly predicted by high recombination rate

Biased gene conversion can explain a strong correlation between the local recombination rate and substitution-derived GC* in primates [22,78], GC* here being the predicted fixation bias

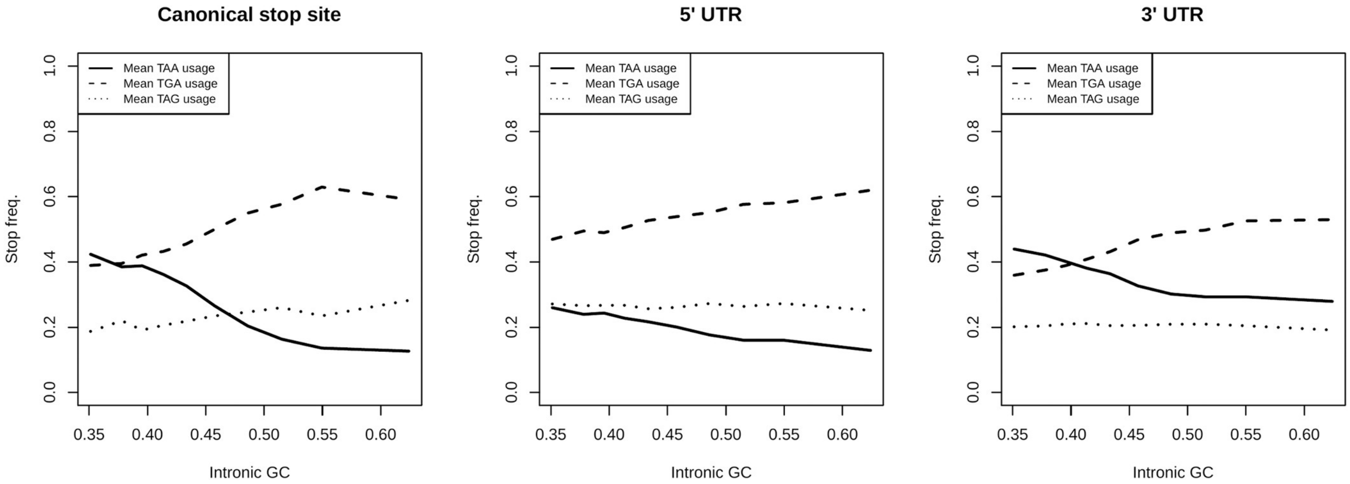

Fig 1. Stop codon frequencies (relative to the usage of all stops) at the canonical stop site, in the 5′ UTR, and in the 3′ UTR at 10 equal-sized bins of various intronic GC contents in the genome. TAA frequency is negatively correlated with intronic GC content in all 3 sequences (Spearman's rank; all $p < 2.2 \times 10^{-16}$, all rho = –0.99, $n = 10$). TGA is positively correlated with intronic GC content in all 3 sequences (Spearman's rank; all $p < 2.2 \times 10^{-16}$, rho = 0.99 for CDS, rho = 1 for both UTRs, $n = 10$). TAG usage is positively correlated with intronic GC content at the canonical stop site (Spearman's rank; $p = 0.0014$, rho = 0.89, $n = 10$) but is uncorrelated with intronic GC content in both 5′ (Spearman's rank; $p = 0.10$, rho = 0.55, $n = 10$) and 3′ UTR sequences (Spearman's rank; $p = 0.61$, rho = 0.19, $n = 10$). Underlying data can be found in S1 Data.

https://doi.org/10.1371/journal.pbio.3001588.g001

determined equilibrium value rather than a nonequilibrium observed value. Similarly, such a model could predict high TGA usage in domains of high recombination. If TAA is optimal, the selection would not predict this as Hill–Robertson interference predicts more efficient selection with higher recombination rates.

To consider the effect of recombination on stop codon usage, we consider both local instantaneous measures of recombination (from the HapMap 2 project, see Methods) and broader scale analysis. The disadvantage of the former analysis is that local recombination rates are not stationary over evolution time so current estimates need not reflect the past history that influences stop codon usage. One problem with the latter is low samples size. Indeed, genome segments with consistently high recombination rates that could make for an ideal test are the pseudoautosomal regions (PAR1 and PAR2). However, there are few pseudoautosomal genes. As predicted by the gBGC model, these regions have high GC content relative to the chromosome average, reportedly 48% in PAR1 compared to 39% in the rest of the X chromosome [79]. In support of the gBGC model explaining high TGA usage, we also find that TGA is used much more often in PAR1 genes (71.4%, using 1 candidate transcript per gene annotated in this region) compared to the genome-wide average (52.4%). While these are not significantly different ($P > 0.05$), statistical comparison of TGA usage between these 2 values is, however, underpowered due to there being a low number of annotated genes which we may extract ($n = 14$).

A better "gross" scale analysis is to consider chromosome size as smaller chromosomes are associated with higher recombination rate per bp [21]. As predicted by the gBGC model in the human genome, we find autosomal size (bp length) to be negatively associated with GC content (Spearman's rank; $p = 0.0078$, rho = –0.56, $n = 22$) and TGA usage (Spearman's rank; $p = 0.0094$, rho = –0.55, $n = 22$) (S1 Fig).

To test whether local recombination rate is predictive of stop codon usage in humans, we employ logistic regression modelling considering all genes, using local recombination rate as the independent variable. Here, we consider the recombination rate which for humans is valid as gBGC-associated non-crossover and crossover events are highly correlated [18]. We find that high recombination rate is significantly predictive of higher TGA usage (coefficient = 0.017, $p = 0.023$) and lower TAA usage (coefficient = –0.046, $p = 1 \times 10^{-6}$), these being the directions predicted by the gBGC hypothesis. Indeed, we find the same trends in noncoding sequences when using linear models to predict trinucleotide frequencies as TAA, TGA, and TAG may appear more than once (unlike at the canonical stop). High recombination rate significantly predicts higher TGA trinucleotide frequency in the 5′ UTR (coefficient = 0.0032, $p = 0.012$), in the 3′ UTR (coefficient = 0.0053, $p < 2.2 \times 10^{-16}$), and in intronic sequence (coefficient = 0.0054, $p < 2.2 \times 10^{-16}$). It also significantly predicts lower TAA trinucleotide frequency in the 3′ UTR (coefficient = –0.0050, $p = 3.5 \times 10^{-14}$) and in intronic sequence (coefficient = –0.0043, $p < 2.2 \times 10^{-16}$), but not in the 5′ UTR where the regression coefficient is negative but not significant (coefficient = –0.0011, $p = 0.28$). These results are all consistent with gBGC promoting TGA over TAA in domains of high recombination both at the focal stop codon and elsewhere.

Net flux to TGA stop codons is highest in GC-rich and highly recombining genes

(i) Increased TAA→TGA substitution in GC-rich regions is common to mammalian and avian lineages, but not lineages that possess weak gBGC. The above considers observed patterns of usage. We can also consider evidence from recent substitution events. Here, we consider flux, meaning the substitution rate from state A to state B (e.g., TAA→TGA) per occurrence of state A in the ancestral sequence. To calculate flux rates, we consider species trios, assign an ancestral state to the internal node by maximum likelihood, and calculate rates of change from this ancestral state to a derived state per incidence of the ancestral state. This is comparable to a prior method [49], excepting for our use of likelihood instead of parsimony.

The gBGC hypotheses predicts that TAA→TGA flux in the mammalian lineage should be highest in GC-rich isochores. More generally, it predicts that in species with gBGC strong and regionalised enough to cause high variation between genes in GC content, that the TAA→TGA flux should be especially accentuated in GC-rich domains. By contrast, species less influenced by gBGC should not show similar accentuation of TAA→TGA flux. We thus test whether the intragenomic difference in TAA<→TGA flux between the highest and lowest by GC is greater when the difference between the mean GC of the 2 partitions (high GC and low GC) is itself greater or when the intragenomic variance in GC is higher.

From the TAA→TGA and TGA→TAA flux rates, we may then adapt the standard formulae (e.g., as used by Long and colleagues [44], see also Li [80] and Bulmer [81]) to calculate TGA content from these flux rates alone, predicted TGA usage (pTGA) (see Methods). This provides a single metric of the relative substitution rate between the 2 stop codons. This we do for the top (GC-rich) and bottom (GC-poor) 50% of genes by GC content, assayed by calculating the intronic GC content of each orthologue from 1 candidate species from the trio, to determine whether the TAA→TGA rate increases with GC pressure.

We calculate the difference in pTGA between GC-rich and GC-poor genes for 4 mammalian set of species trios (within primates, mice, dogs, and cows) and 4 nonmammalian species trios (birds, nematodes, fruitflies, and plants) (see https://github.com/ath32/gBGC for species lists). To assay the extent of pTGA deviation, we calculate (Observed-Expected)/Expected ((O-E)/E) where O is pTGA of the GC-rich set and E is that for the GC-poor set of genes. To assign significance, we compare observed pTGA deviation scores to null simulations that calculate pTGA for 2 null groups of genes according to the net genomic TAA→TGA and TGA→TAA rates (see Methods).

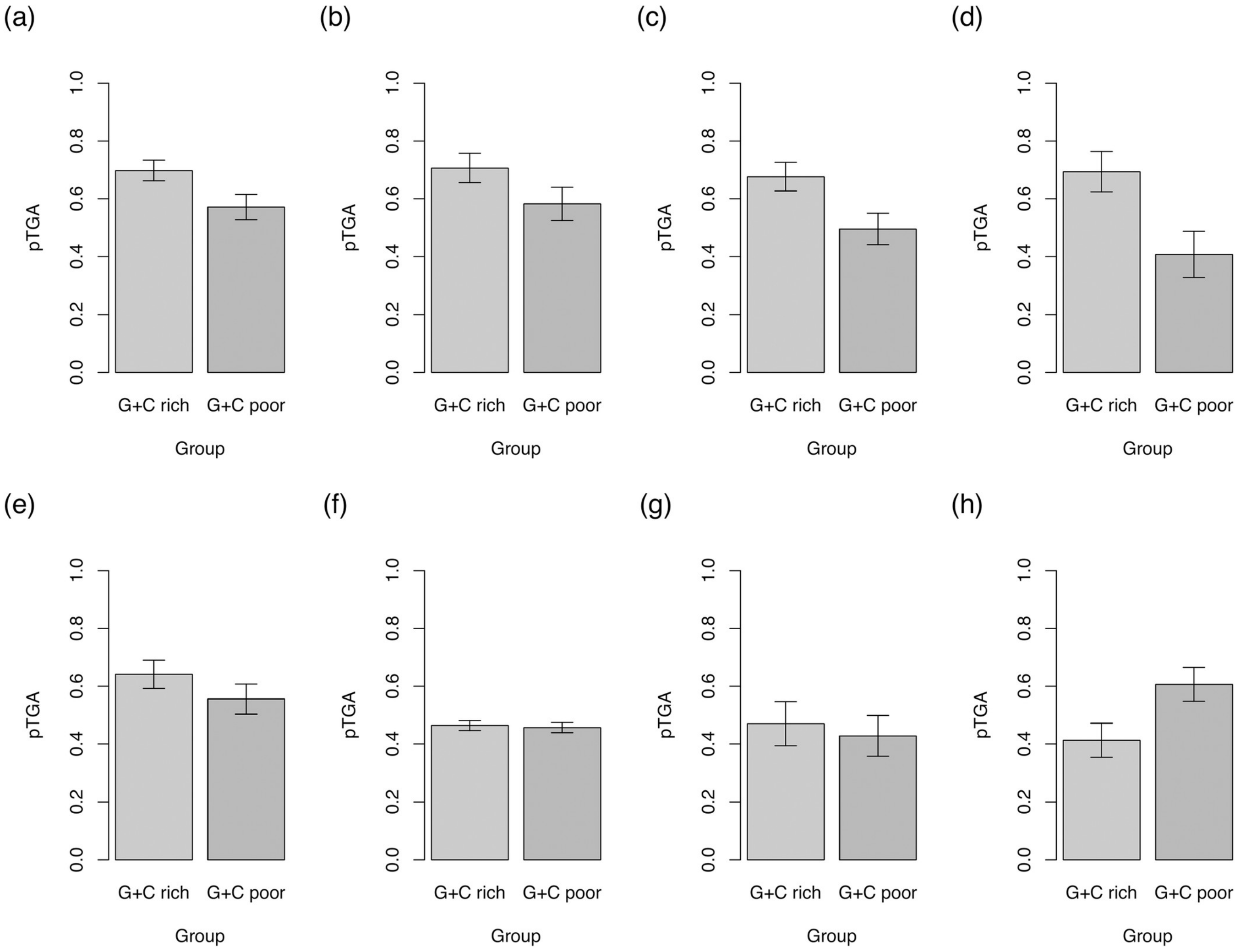

Fig 2. pTGA derived from TAA→TGA and TGA→TAA flux for the top 50% of genes by GC content and bottom 50% of genes by GC content in 4 mammalian (a–d) and 4 nonmammalian (e–h) lineages. pTGA is calculated as 1/(1+(TGA→TAA/TAA→TGA)) and hence represents the balance between the 2 dominant stop codon flux events. Error bars show standard deviation calculated from 10,000 bootstraps generated by resampling genes in each bin with replacement. Underlying data can be found in S2 Data. Trios analysed are primates (a), dogs (b), cows (c), mice (d), birds (e), nematodes (f), fruit flies (g), and plants (h). Species lists are available at https://github.com/ath32/gBGC.

https://doi.org/10.1371/journal.pbio.3001588.g002

Consistent with the hypothesis that gBGC drives high TGA usage in GC-rich isochores, pTGA is higher in GC-rich genes than GC-poor genes across the 4 mammalian lineages. The difference between the gene groups is greater than expected by chance in all 4 cases (primates: $p = 0.014$, dog: $p = 0.040$, cow: $p < 0.0001$, mouse: $p < 0.0001$). Of the nonmammalian lineages, pTGA in GC-rich genes exceeds pTGA in GC-poor genes in birds ($p = 0.174$), flies ($p = 0.427$), and nematodes ($p = 0.231$) but none of the observed differences are significantly different to null. Possibly related to the selfing biology of *Arabidopsis* [82], pTGA is lower in GC-rich genes than GC-poor genes (nevertheless, $p = 1$ using the same test as the other lineages, Fig 2H).

The prediction of the gBGC model is that the between-species variation in intragenomic flux difference should be predicted by the extent of GC variation within the genome. For this analysis, we calculate GC variation as the difference in mean intronic GC content between the 2 sets of genes analysed in Fig 2 and call this GC. We also estimate the variance in GC3

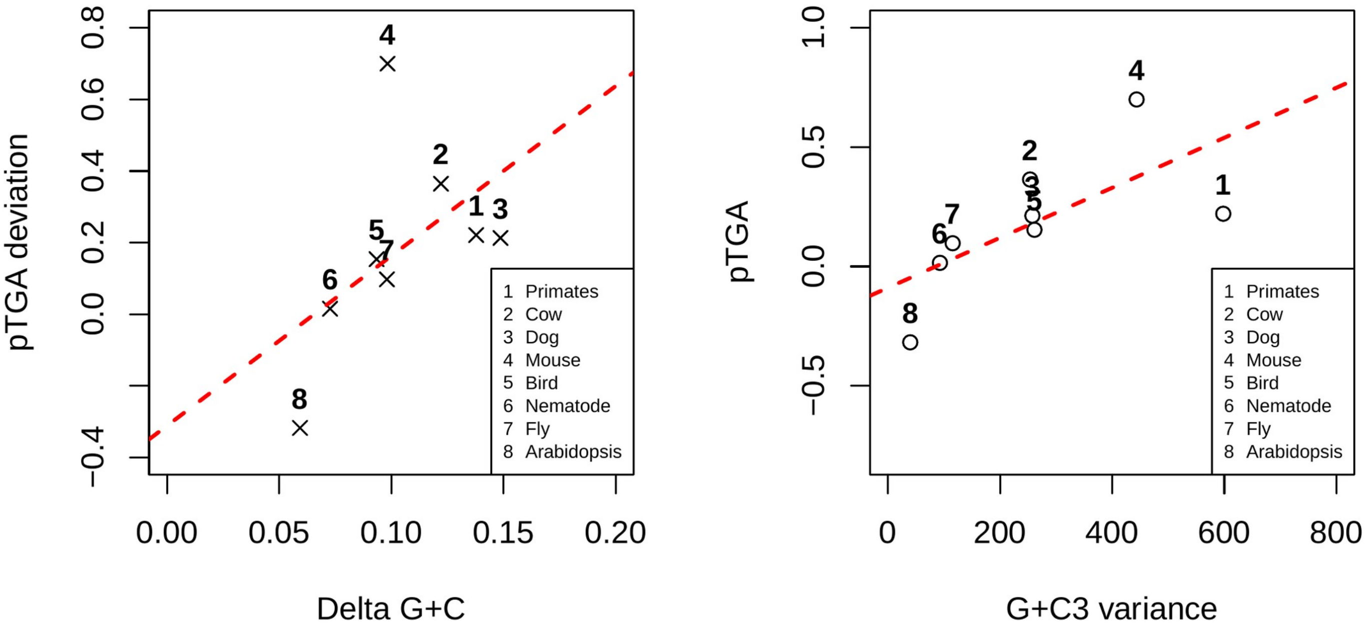

Fig 3. pTGA deviation between the top 50% and bottom 50% of genes by GC content as a function of (a) the difference in GC content between the 2 gene bins, "delta GC," and (b) coding sequence GC3 content variance across a sample of 4 mammalian and 4 nonmammalian lineages. pTGA is calculated as 1/(1 +(TGA→TAA/TAA→TGA)) and hence represents the balance between the 2 dominant stop codon flux events. pTGA deviation is calculated as (O-E)/E where O is the pTGA score of GC-rich genes and E is the pTGA score of GC-poor genes. pTGA deviation is positively correlated with both delta GC (Spearman's rank; $p = 0.046$, rho = 0.74, $n = 8$) and GC3 variance (Spearman's rank; $p = 0.028$, rho = 0.79, $n = 8$). Underlying data can be found in S3 Data. (O-E)/E, (Observed-Expected)/Expected; pTGA, predicted TGA usage.

https://doi.org/10.1371/journal.pbio.3001588.g003

between all genes. Consistent with the gBGC hypothesis for explaining TGA usage trends, analysing our 8 lineages we find pTGA deviation is significantly correlated with both GC (Spearman's rank; $p = 0.046$, rho = 0.74, $n = 8$) and genomic variance in coding sequence GC3 (Spearman's rank; $p = 0.028$, rho = 0.79, $n = 8$) (Fig 3).

This suggests that species with pronounced TAA→TGA flux in their GC-rich domains (mammals) also tend to have more variation between their GC-richest and GC-poorest genes. Broadly, these results accord with what is known about gBGC across these species. The (O-E)/E values are higher in mammals (primates = 0.221, cows = 0.365, dogs = 0.213, mice = 0.700) and birds (birds = 0.154) than in invertebrates (nematodes = 0.015, fly = 0.097) and plants (*Arabidopsis* = –0.318). Birds are expected to resemble mammals as they too have pronounced gBGC [11,83]. However, small chromosomes and associated high recombination rates probably mean that most genes in birds are subject to considerable gBGC, it being notable that the predicted pTGA is high for both gene groups (Fig 2E). Non-isochore-containing genomes of invertebrates may possess AT→GC-biased gene conversion, albeit with much weaker [31,41,84] or less regionalised effects. *Arabidopsis* being an almost obligate inbreeder is expected to be most affected by mutation bias and least affected by gBGC [82], although it has yet to reach equilibrium [82].

(ii) TAA→TGA flux is higher in highly recombining genes than lowly recombining genes. Just as gBGC predicts TAA→TGA flux to positively covary with GC content, as gBGC is coupled tightly to recombination, it also predicts a positive relationship with recombination rate. To assess this, using data from the HapMap2 project, we first define highly recombining genes (HRGs) as the top 50% of genes by recombination rate and lowly recombining genes (LRGs) as the bottom 50%. Adapting our stop codon flux methodology, we then calculated the flux rates for TAA→TGA and TGA→TAA for HRGs and LRGs and used these rates to calculate pTGA for both groups (Fig 4). Significance was once again determined by comparing the observed pTGA deviation to those observed in null simulations that assume uniform genomic TAA→TGA and TGA→TAA rates. Consistent with the hypothesis that gBGC drives high TGA usage in highly recombining regions, pTGA is higher in HRGs than LRGs, (p = 0.049). The pTGA deviation score between HRGs and LRGs is 0.172, slightly less than observed between GC-rich and GC-poor genes in the same genome (0.221).

No evidence to support TGA optimality in eukaryotes

The evidence from nontermination sites (i.e., UTRs, introns) supports the hypothesis that whatever causes unusual TGA usage trends in most mammals, it cannot be explained by selection on the focal termination codon alone. Also, as predicted by the gBGC model, the TAA→TGA flux is stronger in domains of high GC/high recombination. Nonetheless, to have a case that gBGC acts against the direction of selection, we need also to be able to confident that selection does not prefer TGA. Outside of the focal termination codon, this is hard to assay but at the focal stop codon, we can gather further evidence.

First, selection on any genic feature is classically assumed to predict that usage of that feature will be most common in HEGs [62,85,86] as selection is strongest in HEGs. Overusage of "optimal" codons in HEGs is a case in point [87,88]. In the current context, the opportunity for deleterious readthrough (or other stop codon error) should scale linearly with the amount of protein product, so protein levels are a good metric for assaying strength of selection on the stop codon. Hence, if TGA usage were to be explained by selection, TGA usage is predicted to positively correlate with protein level. Prior data appeared to contradict this, suggesting that human HEGs (the opposite being lowly expressed gene (LEGs)) preferentially use TAA stop codons [77]. However, possible covariation between expression level and GC content [89–91] could disturb the ability to make correct inference. We ask whether TAA or TGA are overemployed in HEGs after controlling for GC content.

Second, the efficiency of both selection [92,93] and gBGC [28,32] are expected to vary with the effective population size (N_e), both being more effective when N_e is high. The gBGC effect is however complicated by the fact that selection may also modify the effect of gBGC, reducing its impact if deleterious [28], such selection in turn also being dependent on N_e. Most evidence suggests that gBGC is more influential when N_e is high (but see also [94]). However, we know the direction of gBGC, and it must act against TAA. Thus, across eukaryotes our expectation is that if TAA is optimal (and gBGC is relatively less important), its usage will increase with N_e. However, if gBGC is unexpectedly important outside of mammals or if TGA is optimal then TGA will increase with N_e. We previously observed this not to be the case, with TAA increasing with N_e [95]. However, the possibility remains that for LEGs TGA might be optimal and causing the focal termination codon trends (despite similar behaviour in 3′ UTR). We test this extension.

(i) High expression level strongly predicts high TAA usage controlling for GC. To test the predictive power of expression level on stop codon usage, we consider a series of logistic regression models. Each gene was assigned a 1 (present) or 0 (absent) in 3 different columns, TAA, TGA, and TAG, depending on its stop codon identity. These scores were included as the dependent variable in several logistic regression models, with protein abundance (for which we employ the natural log to promote a normal distribution) an independent predictor. We control for GC content by fitting multivariate models that include GC3 content (Table 1). Collinearity between GC content and protein abundance need not be a concern as the computed variance inflation factors are very low (less than 1.1 for all models).

Consistent with prior observations of stop codon covariance with GC content [77,96], we find TAA usage to be negatively (indicated by the sign of the coefficient), and TGA to be

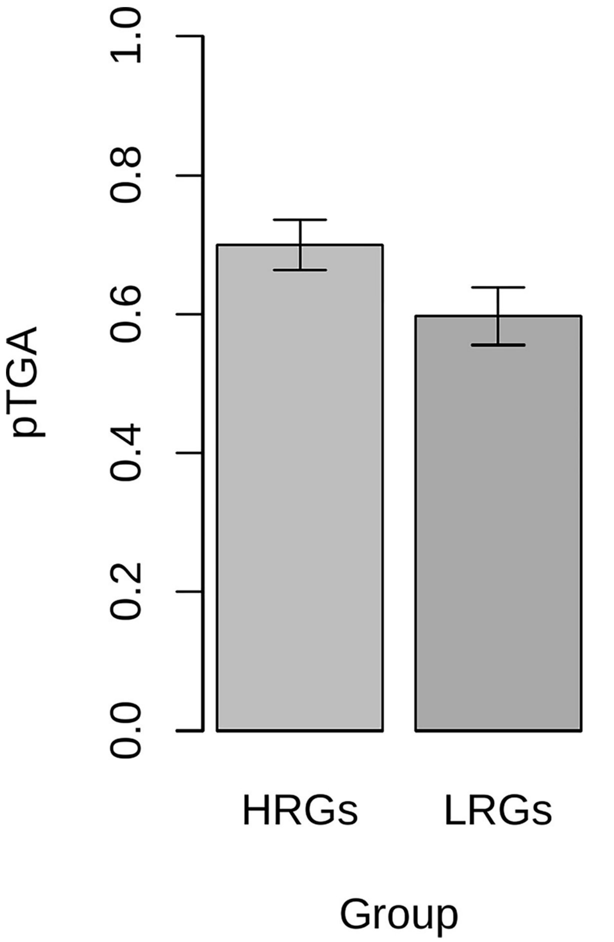

Fig 4. pTGA derived from TAA→TGA and TGA→TAA flux for the top 50% of genes by recombination rate (HRGs) and bottom 50% of genes by recombination rate (LRGs) in the human genome. pTGA is calculated as 1/(1+(TGA→TAA/TAA→TGA)) and hence represents the balance between the 2 dominant stop codon flux events. Error bars show standard deviation calculated from 10,000 bootstraps generated by resampling genes in each bin with replacement. Underlying data can be found in S4 Data. HRG, highly recombining gene; LRG, lowly recombining gene; pTGA, predicted TGA usage.

https://doi.org/10.1371/journal.pbio.3001588.g004

positively, correlated with GC3 in all 4 species trios tested. By the same coefficient analysis, we find that high TAA stop codon usage is predicted by high expression level in all 4 mammalian lineages [77], contra to the possibility that TGA has become the favoured stop codon in mammals. Both protein abundance and GC3 are consistently significant predictors of stop codon

Table 1. Results from multivariate logistic regression analysis that assess the extent to which gene expression and gene coding sequence GC content can predict stop codon usage in mammalian genes.

Stop	Parameter	Primates			Dog			Cow			Mouse		
		Coef.	Std. error	*p*-value	Coef.	Std. Error	*p*-value	Coef.	Std. Error	*p*-value	Coef.	Std. error	*p*-value
TAA	Log (PxAbundance)	0.023	0.007	6E-4	0.071	0.014	4E-7	0.071	0.008	2E-16	0.013	0.006	0.039
	GC3	−0.038	0.001	2E-16	−0.033	0.002	2E-16	−0.040	0.002	2E-16	−0.034	0.002	2E-16
TGA	Log (PxAbundance)	−0.015	0.006	0.009	−0.039	0.012	0.001	−0.050	0.007	3E-13	−0.006	0.005	0.273
	GC3	0.019	0.001	2E-16	0.018	0.002	2E-16	0.022	0.001	2E-16	0.019	0.001	2E-16

https://doi.org/10.1371/journal.pbio.3001588.t001

usage in our 3 mammalian lineages. In 8/8 models, the coefficients of protein abundance are consistent with TAA preference over TGA in HEGs. Assuming that gene expression levels in orthologous genes are stable, stop codon usage reliably informs us of the stop that is preferred by selection.

(ii) Across taxa, lowly expressed genes also prefer TAA over TGA. While the above analyses provide support for the hypothesis that TAA, and not TGA, is preferred in HEGs, there is however, a further possibility, namely, that while TAA may well be preferred by HEGs, TGA may be optimal in LEGs. If this were to be the case, TGA might increase in genome-wide usage if most genes are not "highly" expressed. This we test by phylogenetically generalized least squares (PGLS) regression analysis that compares TGA enrichment (at the primary stop codon compared to downstream, to remove any GC covariance) in LEGs to effective population size (N_e) for several eukaryotic species controlling for phylogenetic topology (see PGLS in Methods).

We find N_e to be a significant negative predictor of TGA enrichment in LEGs (PGLS; estimate = −0.060, $p = 0.012$). By contrast, TAA enrichment in LEGs is positively, if not significantly, associated with N_e (PGLS; estimate = 0.073, $p = 0.078$). When we consider HEGs, N_e positively and significantly correlates with TAA enrichment (PGLS; estimate = 0.059, $p = 0.0014$) but is negatively, if not significantly, associated with TGA enrichment (PGLS; estimate = −0.044, $p = 0.17$). These results are not consistent with a selective preference for TGA stop codons at any expression level. These same results also indicate that gBGC is not an important force in most of the species examined as gBGC should also be more influential when N_e is high and force increased usage of TGA [32].

TAA→TGA flux cannot be explained by mutation bias in humans

The above evidence indicates that whatever causes TGA conservation, it is neither specific to the termination site nor explained by selection for termination efficiency at the termination site. In principle, the trends we have seen could be explained by mutation bias. However, mutation bias tends to be GC→AT biased so should favour TAA not TGA [42–46]. Nonetheless, the possibility remains either that some more complex *k*-mer bias might exist or that mutation bias varies by isochore. Indeed, nucleotide pools can vary through the cell cycle potentially altering local mutation bias [97]. Moreover, CpG to TpG rates are high in humans [98–101], and thus creation of new stop codons away from the focal stop (e.g., within 3′ UTR) via CpGA to TpGA could be common. We could imagine, for example, that focal stop codons commonly mutate to a sense codon this being rescued by a 3′ UTR preexisting stop. If so, stop codon usage could be determined by mutational processes away from the focal termination codon. The same model does not however predict TAA→TGA flux at orthologous termination sites. That CpG deamination rate may also correlate negatively with GC content [101] also renders this an unlikely explanation.

We consider the relative rates of human germline de novo mutations derived from family trio data [102]. From the mutation rate of each class of mutational event, we calculate rates per occurrence of the ancestral nucleotide and generate a mutational matrix. From this, we calculate the neutral equilibrium frequencies of all nucleotides (denoted N^*), dinucleotides, or codons (see Methods). From N^* predictions, we may predict the equilibrium GC frequency (GC^*). Under the assumption that nucleotide contents are stationary, deviation of the observed nucleotide content from predicted equilibrium provides an indication of the direction of any fixation bias [44]. However, equilibrium status is disputed [103] and the predicted equilibrium can vary with complexity of the mutational model (mononucleotide, dinucleotide, etc.).

Consistent with previous analyses [42–46], from a dataset of 108,778 observed de novo mutations, we find an overall GC→AT skewed mutational profile that hence fails to predict observed stop usage (S2 Table). Might, however, variation in mutation bias between isochores explain increasing usage of TGA and decreasing usage of TAA as domains become more GC-rich? To assay whether the above mutational profile covaries with intronic GC in a similar way to stop codon flux, we first repeat the above analysis for mutations found in different isochore GC contents (see also Smith and colleagues [46]). For each mononucleotide change, the local GC content (10 kb window) was calculated. Mutations were then ordered by GC and split into 10% percentile bins of equal size (approximately 10,000 mutations each). From each of these bins and their associated mutational spectra and nucleotide contents, we recalculate GC* and TGA* (Fig 5, orange points). We find our GC* and TGA* predictions for each bin to be consistent between isochores of different GC content, indicating that mutation bias is not driving the trends we see in TGA usage nor TAA→TGA stop codon flux—and see also Smith and colleagues [46]. If anything, mutation bias is increasingly GC→AT biased at high GC as the local GC content around de novo mutations is negatively correlated with their predicted GC* (Spearman's rank; $p = 0.024$, rho = –0.72, $n = 10$) and TGA* (Spearman's rank; $p = 0.035$, rho = –0.68, $n = 10$).

The above approach makes no allowances for more complex dinucleotide effects nor the possibility that some stop codons might be generated by mutations within CDS or within 3′ UTR sequences when the focal stop mutates. Given that there is hypermutability at CpG residues, leading to TpG residues [98–101] that are likely to affect the mutation-drift equilibrium frequency of TGA, we expand our analysis to consider the 16 × 16 dinucleotide mutational matrix. We also apply a model in which we generate null sequences from the equilibrium mutational matrix in a Markov process, hence allowing for within UTR mutational events. We consider the relative frequencies of the 3 stop codons in such sequence and how they vary by

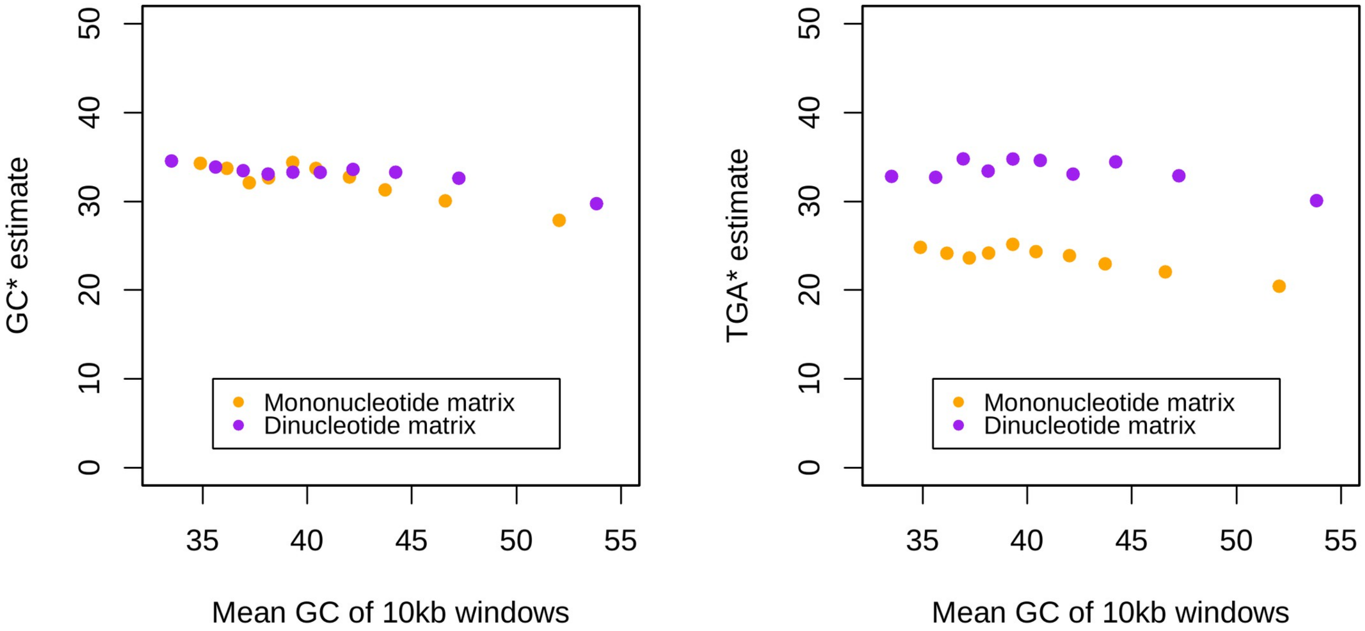

Fig 5. Predicted GC equilibrium (GC*) and relative TGA equilibrium (TGA*) frequencies across isochore GC contents derived from mononucleotide (orange) and dinucleotide (purple) mutational matrices. Standard deviations for the datapoints are minuscule and hence error bars are not shown (approximately 0.5% for mononucleotide estimates of TGA* and GC*, approximately 0.5% for dinucleotide estimates of TGA*, and approximately 0.1 for dinucleotide estimates of GC*). Underlying data can be found in S5 Data.

https://doi.org/10.1371/journal.pbio.3001588.g005

local GC. Consistent with the mononucleotide results, we find dinucleotide-derived GC* and TGA* to be lower than observed in the genome (40.9% and 52.4%, respectively) and, importantly, flat across GC contents (Fig 5, purple points). While TGA* derived from the dinucleotide matrix exceeds TGA* derived from the mononucleotide matrix, this is probably as a consequence of permitting CpG hypermutation generating potentially premature stop codons. We conclude that the absence of evidence for increasing GC* with GC content strongly argues against mutation bias as an explanation for higher TAA→TGA flux and higher TGA usage in GC-rich isochores.

Mutation bias predicts trinucleotide usage in GC poor domains and TAG rarity

Above we have generated a mutational expectation for all trinucleotides but focused on TGA. This allows us to ask a series of further questions. For example, for all trinucleotides might a mutational null match what we see in GC-poor domains, as expected if these are less subject to gBGC? In addition, can mutation explain any trends in stop codon usage in GC-poor domains, for example, the observation that TAG is underused compared with TGA?

We find that observed trinucleotide frequencies from GC-poor sequences (the bottom 20% of genes by GC content) are accurately predicted by a GC-poor mutational matrix (derived from the bottom 20% of de novo mutations by surrounding 10 kb GC content) for all sequence that is not CDS ($r^2 > 0.9$; Fig 6). This strongly supports the hypothesis that mutation bias alone may explain trinucleotide trends in GC-poor domains outside of the coding context. In addition, while one can always consider more complex *k*-mer–dependent mutational models, our extension from dinucleotides rates appears to be robust. Importantly, in such GC-poor isochores TAG equilibrium is lower than TGA equilibrium (S2 Fig). This indicates mutation bias operates differently on the 2, going some way to explain why TAG and TGA behave differently.

gBGC predicts deviations from mutational expectations for all trinucleotides

The previous analysis suggests that in low GC domains, *k*-mer trends are well predicted by mutation bias alone (Fig 6). By contrast, in GC-rich domains, there exists a substitutional bias to TGA that is incompatible with mutation bias alone (Fig 5). Is the TAA→TGA fixation bias in high GC domains illustrative of a broader pattern? Were gBGC mimicking purifying selection we expect that GC-rich trinucleotides should be most deviant from their mutational null in GC-rich domains. We hence extend the above analysis to consider the extent to which all trinucleotides deviate from mutational equilibrium as a function of their isochore of residence. In this instance, however, we cannot be confident that the GC-rich residue is selectively deleterious (as with TGA). Moreover, even when optimal codons are known to be GC ending selection at exon ends can commonly be in the opposite direction to enable accurate splicing [104], adding complexity.

Using mutational profiles from the relevant isochore, we calculate trinucleotide frequencies that represent our mutational null and compare these to observed trinucleotide frequencies in the genome. To test the hypothesis that a fixation "boost" in GC-rich isochores acts differently on GC-rich trinucleotides, we calculate a fixation boost metric. Specifically, we first calculate a (O-E)/E score for the top 20% of sequences by GC content, where expected is the mutational equilibrium frequency derived from the top 20% of de novo mutations assaying their surrounding 10 kb GC content. This metric we term deviation 1, or *D1* for short. We then repeat this for the bottom 20% of sequences by GC content using their equivalent set of de novo

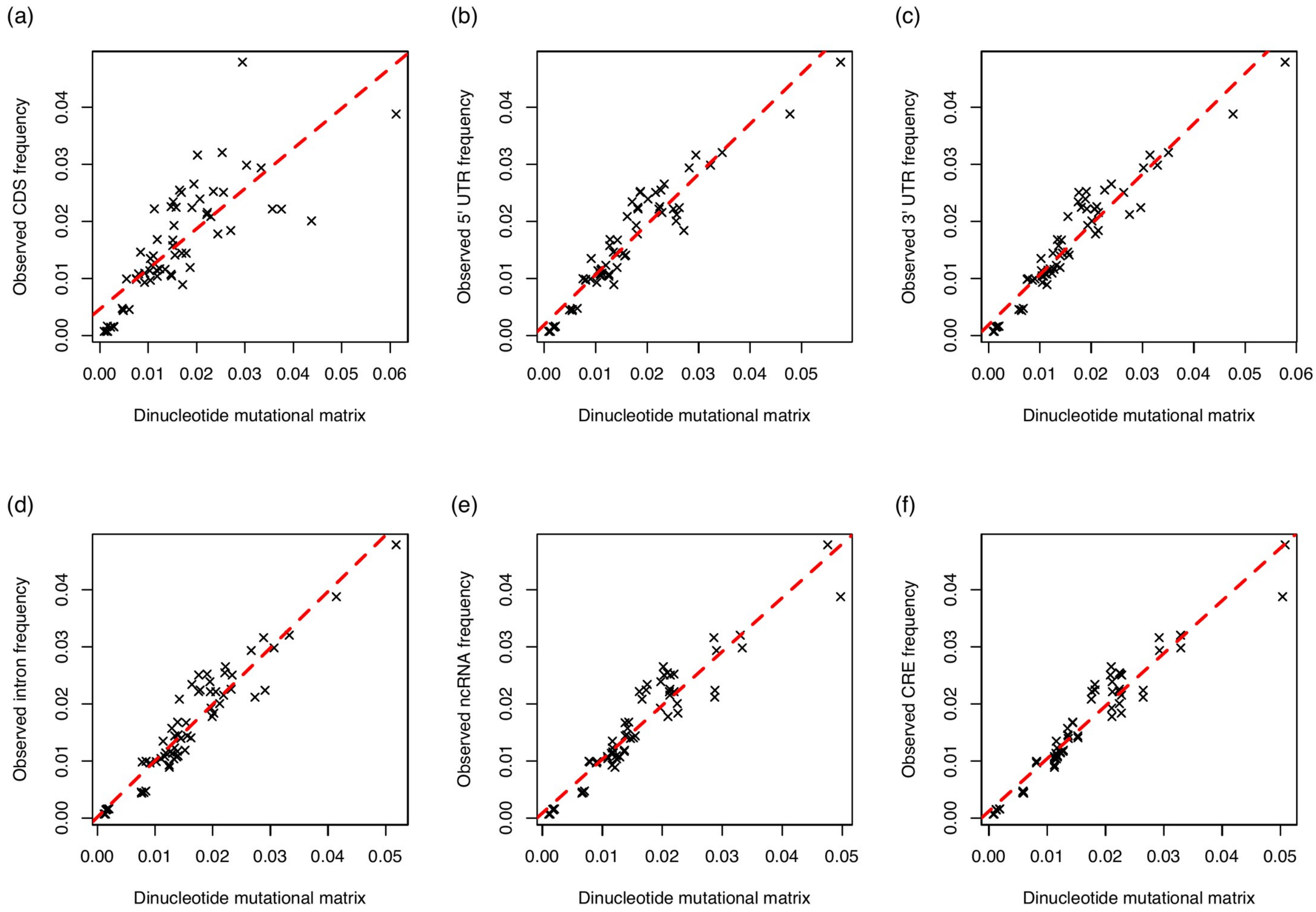

Fig 6. Observed (a) CDS, (b) 5′ UTR, (c) 3′ UTR, (d) intronic, (e) ncRNA, (f) CRE trinucleotide frequencies as a function of the expected frequencies of the same trinucleotides derived from a dinucleotide mutational matrix. Expected frequencies were calculated simulated DNA sequences derived from dinucleotide equilibrium frequencies. Dinucleotide frequencies were calculated from a sample of de novo mutations taking place in the bottom 20% of sequences by GC content to avoid potential GC-coupled fixation biases. Expected frequencies accurately predict what is seen in real CDS sequence (linear regression; $p = 7.7 \times 10^{-15}$, adjusted $r^2 = 0.62$), 5′ UTR sequence (linear regression; $p < 2.2 \times 10^{-16}$, adjusted $r^2 = 0.90$), 3′ UTR sequence (linear regression; $p < 2.2 \times 10^{-16}$, adjusted $r^2 = 0.91$), intronic sequence (linear regression; $p < 2.2 \times 10^{-16}$, adjusted $r^2 = 0.90$), ncRNA sequence (linear regression; $p < 2.2 \times 10^{-16}$, adjusted $r^2 = 0.90$), and CRE sequence (linear regression; $p < 2.2 \times 10^{-16}$, adjusted $r^2 = 0.93$). Underlying data can be found in S6 Data. CDS, coding sequence; CRE, *cis*-regulatory element; ncRNA, noncoding RNA.

https://doi.org/10.1371/journal.pbio.3001588.g006

mutations, receiving *D2*. Given the above results (Fig 6), we expect the bottom 20% to be closest to mutation equilibrium, hence having a low *D2* score. By contrast, if there is a GC-correlated fixation bias, *D1* should be high for the GC-rich trinucleotides. We thus consider, for each trinucleotide, the difference between *D1* and *D2* values, this reflecting the shift in fixation process associated with domains of high GC. Using this metric, trinucleotides may be ranked by the "boost" they receive from GC-coupled fixation bias within their GC class. Thus, we classify all trinucleotides into 1 of 4 classes by GC% (0%, 33%, 66%, 100%). Within the 0 class are trinucleotides with no G or C (e.g., AAA, ATA, TTA) and within the 100% class by contrast are those with no A or T (e.g., GGC, GGG), for example.

We find that the more GC-rich the class of trinucleotides, the more they exceed their mutational equilibrium in high GC isochores (0% < 33% < 66% < 100%) (for statistics, see Fig 7). This strongly supports the notion that the trinucleotide content of isochores derives from a fixation bias, rather than mutation bias, favouring GC residues, as gBGC would predict. More

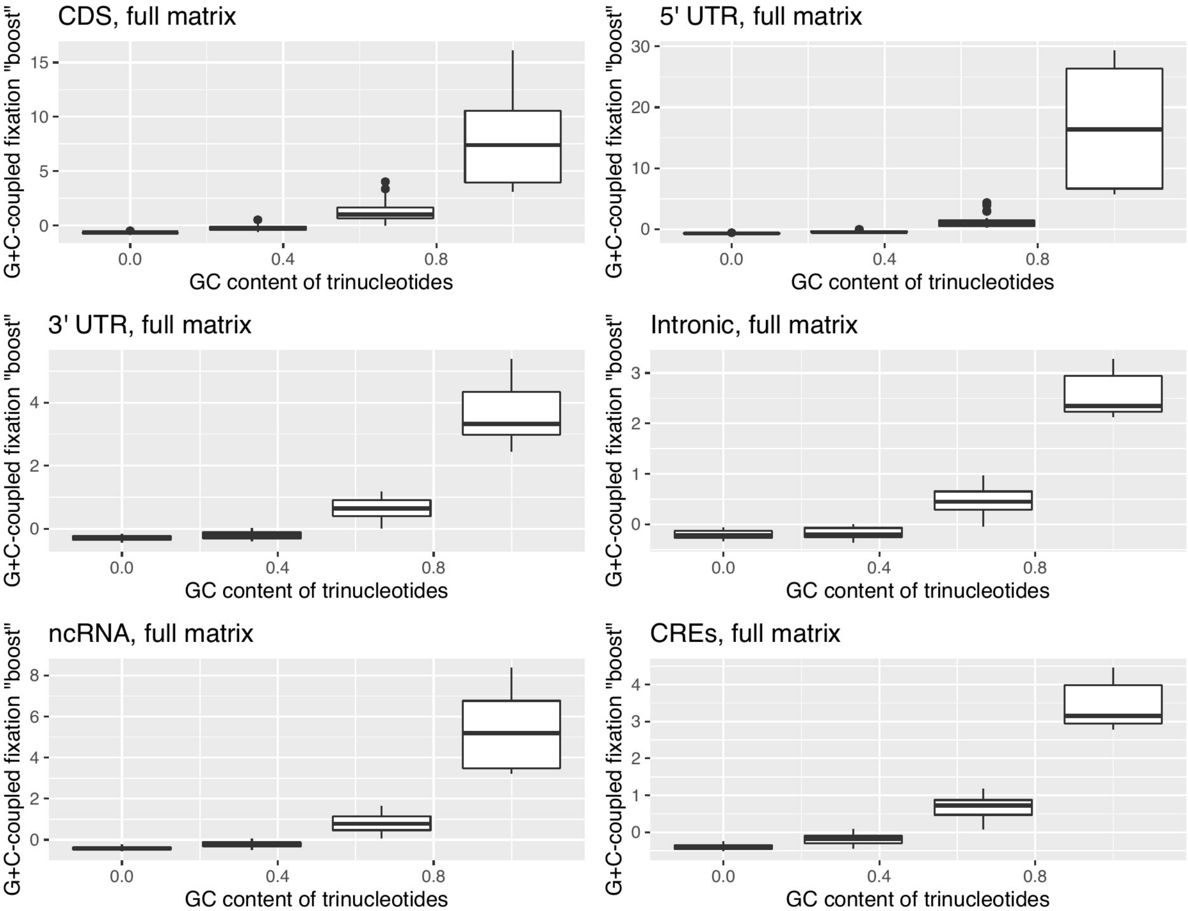

Fig 7. Deviation scores, (O-E)/E, describing the difference in GC-coupled fixation "boost" for the 4 GC classes of trinucleotides. Deviation between fixed and mutational equilibrium frequencies for each trinucleotide in the top 20% of sequences by GC content, D1, was calculated as (O-E)/E, where expected is the mutational equilibrium frequency. This was repeated for the bottom 20% of sequences by GC content to receive D2. As we predict GC-rich sequences to be subjected to stronger biased gene conversion, we predict D1 > D2. To compare D1 and D2, we once again calculate (O-E)/E, which we dub the GC-coupled fixation "boost". In all sequences, GC content is positively correlated with this "boost" metric (Spearman's rank; all $p < 2.2 \times 10^{-16}$; rho = 0.92 in CDS, rho = 0.94 in 5′ UTR, rho = 0.90 in 3′ UTR, rho = 0.87 in introns, rho = 0.92 in ncRNA, rho = 0.93 in CREs, $n = 64$ in all tests). Underlying data can be found in S7 Data. CDS, coding sequence; CRE, *cis*-regulatory element; ncRNA, noncoding RNA; (O-E)/E, (Observed-Expected)/Expected.

https://doi.org/10.1371/journal.pbio.3001588.g007

generally then, we have strong reason to suspect the gBGC-mediated fixation bias causes false signals of purifying selection at GC-rich residues in GC-rich isochores that extend far beyond the specific context of TAA→TGA flux.

We assess this possibility a second way by considering flux between all 2-fold synonymous codon pairs, all ending G:A or C:T, in genes of increasing recombination rate. Considering all 2-fold synonymous codon pairs en masse, we find that the flux to the GC-rich codons are most strongly favoured at high recombination rates, consistent with possible gBGC action (S3 Fig). Before Bonferroni correction, this is true for 10 of the 12 two-fold synonymous codon pairs individually (Binomial test with null probability = 0.5; $p = 0.039$). This too is supportive of a gBGC-mediated fixation bias that is much more general than the stop codon example. Unlike with TAA and TGA flux, however, we cannot in these examples be sure which (if either) is the selectively optimal state. The 2 exceptions are Leucine (TTA<->TTG) and Glutamine (CAA<->CAG) where the ratio of flux increasing GC and decreasing GC is invariant to recombination rate (S4 Fig). That both CAA<->CAG and TAA<->TAG are unrepresentative of the more general trend is noteworthy.

Trinucleotides have stereotypical fixation biases

We have observed that high TGA usage and high TAA→TGA fixation bias is especially common in GC-rich isochores, but TAG usage does not behave in the same way. Is this difference between 2 GC-matched trinucleotides particular to TAG and TGA? The CAA→CAG result would suggest not. We can address this by considering within GC-class variation in the fixation "boost" scores calculated above.

Not only do we find substantial variation between trinucleotides of the same class (S5 Fig), but also we find the ranking within each GC class to be remarkably consistent between sequence types (5′ UTR, 3′ UTR, noncoding RNA (ncRNA), *cis*-regulatory elements (CREs), introns) (Fig 8). We exclude coding sequence from this analysis to negate the impacts of coding selection. Within the most populated GC classes (33% and 66%), ranks are significantly

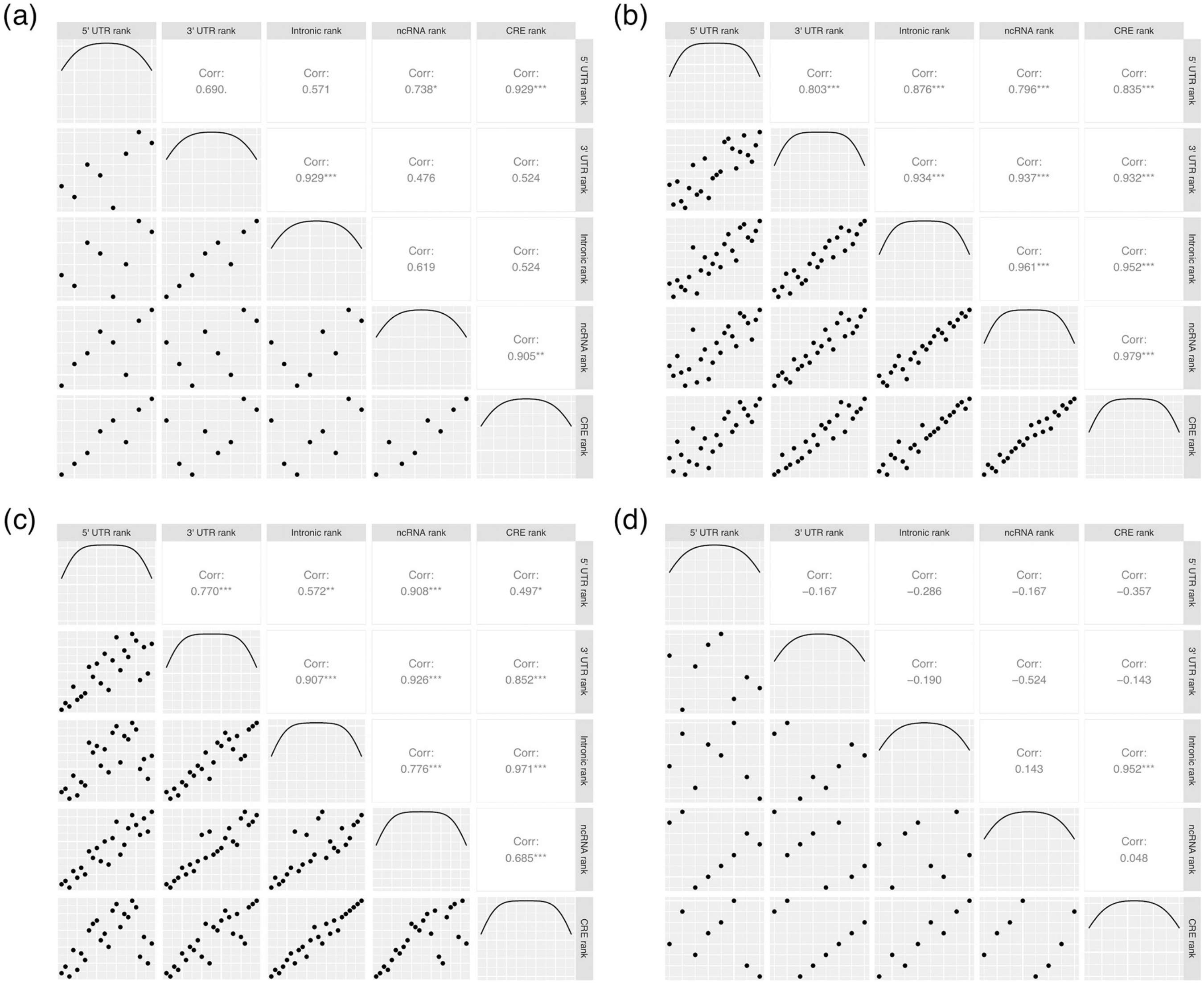

Fig 8. Correlation analysis of trinucleotide ranks (by their gBGC "boost" metric) within the 4 GC classes (a) 0%, (b) 33%, (c) 66%, and (d) 100%. Within the 33% and 66% GC classes, ranks are significantly correlated in all comparisons ($p < 0.01$). This is not true of the 0% and 100% GC classes, correlation analyses within which are underpowered ($n = 8$ trinucleotides in each class compared to 24 in the 33% and 66% classes). Correlation statistics were calculated using Pearson's method. Underlying data can be found in S8 Data. CRE, *cis*-regulatory element; gBGC, GC-biased gene conversion; ncRNA, noncoding RNA.

https://doi.org/10.1371/journal.pbio.3001588.g008

correlated in all comparisons (Pearson's method; all $p < 0.01$). This supports the hypothesis of a consistent isochore dependent fixation bias that acts differently on different trinucleotides of the same GC content. We note that the nonexpressed CRE versus intron comparison gives exceptionally high repeatability indicating that transcription coupled repair/mutation probably does not explain these trends.

Within the TAG/TGA case study, we find TAG to be less "boosted" than other A, G, T-containing trinucleotides, second only to GTA trinucleotides (S6A Fig). By contrast, TGA is the most promoted by fixation bias, with the one exception of AGT in the 5′ UTR (S6A Fig). Fixation bias correlated with GC-content, hence appears to contribute to the differences in frequency between TGA and TAG trinucleotides outside of, and possibly also within, the stop codon context. That TCA also receives a consistent higher GC-coupled fixation boost than TAC (S6B Fig) favours that the fixation bias is dependent on nucleotide context rather than stop codon functionality. We also recall that while high recombination rate favours flux to the GC-rich state at 2-fold degenerate sites, Glutamine (CAA→CAG) is one exception to this rule (S4 Fig). If CAA→CAG is suffering a similar fate to TAA→TAG, this too would be supportive of a general nucleotide context-dependent trend in fixation bias affecting TAG rather than selection for termination efficiency.

Discussion

The assumption that sequence conservation implies purifying selection and hence optimality of the preserved sequence underpins many enterprises, from medical diagnostics to evolutionary analyses of the proportion of sequence that is functional. While there has been prior consideration that tests for positive selection might be impacted by gBGC mimicking selection's signatures [12,35–40], there has been less attention paid to the problem that it might also explain sequence conservation, despite this being a logical necessity [41]. We identified the case of stop codon usage in mammals as a test case because prior evidence suggested a contradiction: TAA looks to be optimal (as elsewhere) but TGA was nonetheless conserved. We reasoned that gBGC might explain this and resolve the exceptionalism of mammalian stop codon usage. Our data strongly support this. We see TGA usage is higher in GC-rich and highly recombinogenic domains, with the same trends also being seen in noncoding sequence. Increased TAA→TGA flux is also seen in GC-rich regions and regions of high recombination. Multiple lines of evidence suggest that at the focal termination codon TGA is not optimal and hence that gBGC can act against the direction of selection. The results satisfy all criteria proposed by Duret and Galtier [11] for differentiating gBGC from selection. Across species a greater flux of TAA→TGA in the GC-richer genes is associated with a greater intragenomic variance in GC content, consistent with the above trends being predicted, broadly speaking, by the extent to which a species is isochoric.

Is the TAA/TGA enigma a special case or indicative of a more general trend? We observe that deviation of all trinucleotides from mutational equilibrium in GC-rich domains is strongly predicted by their GC content. The TAA→TGA trend in high GC domains can be considered a special example. More generally then, we have strong reason to suspect the gBGC mediated fixation bias will cause false signals of purifying selection at GC-rich residues in GC-rich isochores that extend beyond the specific context of TAA→TGA flux. This example is however unusual in that we have confidence that the substitutional process at the focal termination codon context forces conservation of a nonoptimal codon, a trend that can be partly overcome by stronger selection for optimality in HEGs.

There is, however, another possibility to explain deviation from mutational equilibrium in domains of high GC, this being that some form of selection favours GC-rich sequence. As

Hill–Robertson interference [105] is reduced in domains of high recombination selection should be more effective in such domains, causing a fixation bias. One can imagine many possible modes of such selection, for example on DNA structure [106–108] or on nucleosome positioning [109–112]. Unlike gBGC that predicts GC enrichment, any selection model must, after the fact, explain why GC-rich trinucleotides are favoured. Such models are unconvincing for several reasons.

First, in the current context TGA is not selectively favourable at the focal termination codon but nonetheless conserved. This suggests we must evoke a force other than selection to explain TGA conservation (assuming selection on stop codon functionality to be the strongest mode of selection at the focal stop). Why we should not similarly evoke the same force outside of the termination context seems like special pleading.

Second, the strength of selection (and associated load) in species with low N_e (mammals and birds) is problematic. Consider the hypothesis that TA dinucleotides could lead to accidental incidences of, for example, "TATA" boxes in eukaryotes and "Pribnow" boxes in bacteria (i.e., the TATAA motif). More generally, TA features in many key regulatory motifs that would be inappropriate in most DNA regions in both eukaryotic and prokaryotic genomes [113,114]. To date, this is probably the best (if not only) model for selection against TA in all taxa. This could, in principle, explain why TAG is underused compared with TGA. Indeed, within the trinucleotides with only A and T, ATA and TAT, the 2 that are core to TATA box, are consistently the 2 with the lowest "boost" (S6 Fig). In bacteria and archaea, the strength of selection against such spurious binding is estimated to be around $N_e s = -0.09$ and thus within the range of nearly neutral mutations for these species [115]. If then *Escherichia coli*'s N_e is of the order of 10^8 [116], then s must be approximately $-0.09/10^8 = -9 \times 10^{-10}$. For a mutation to be under selection in humans s approximately 1/2 N_e must hold. In a species with N_e approximately 10,000 (e.g., humans), then this value of s (i.e., 1/20,000) is much greater than 9×10^{-10} estimated for selection against spurious binding. Thus, unless the selective cost of spurious binding is very much greater in humans than in bacteria, it is hard to see how selection can be efficient enough to remove point mutations that introduce spurious binding sites.

We do not presume that mutation bias and selection have no role. Indeed, in GC-poor domains mutation bias appears to provide a robust fit to the observed trends and explains the differential usage of TAG and TGA. Further, HEGs overemploy TAA. However, for a full explanation of TGA conservation, especially in GC-rich domains, we need to evoke some other force, of which biased gene conversion is a good possibility, not least because it predicts high GC trinucleotides should be given a fixation boost in GC-rich domains, as observed.

We do not wish to claim that TAA is optimal for all genes. There could be many reasons that, for some genes, TGA is optimal. One possibility could be that TGA might be the least leaky in some contexts but as the experimental evidence contradicts this possibility [61], we do not consider this reasonable. Alternatively, TGA may be TR-prone and "leaky," but that leakiness is selectively favoured in some instances. High rates of TR may beneficially increase proteome diversity [117]. Indeed, a few examples of functional readthrough have been described [118,119], though the commonality of this in mammals is unknown. Alternatively, readthrough may be part of a gene regulatory mechanism [76,120]. Indeed, the discovery of TGA conservation prompted speculation that TGA might be commonly optimal in humans as it enables novel gene expression control. Specifically, it was suggested that ribosomes that read through the primary stop codon stall and form a queue from the next in-frame stop (or ribosome pausing factor), filling the space between the 2 stops and eventually infringing upon the 3′ end of the coding sequence itself. At this point, translation of this mRNA molecule is blocked [120]. The fact that readthrough occurs at a low (but not very low) rate thus allows the mRNA molecule to be translated a relatively tightly regulated number of times prior to degradation.

Generally, however, it is unclear how any adaptive TR model might explain mammalian exceptionalism in stop codon usage. Given that TGA optimality cannot explain why TGA is also favoured in noncanonical stop contexts, the above arguments are, by Occam's razor, not needed to explain general trends. Moreover, were there selection for TR, one might expect this to be common to all eukaryotes and therefore predict higher TGA usage in species with high N_e (not just mammals), but this is not seen [95]. Instead TAA usage correlates positively with N_e [95], as expected if it is the optimal stop codon (although there are mechanisms that are rare in high N_e species but common in mammals, a high density of exonic splice enhancers to define intron–exon junctions being a case in point [121]).

Why do trinucleotides of the same nucleotide content have different fixation boosts?

Our evocation of gBGC to explain the general trends in GC-rich domains is not a complete explanation. Importantly, we see repeatable trends whereby GC-matched trinucleotides show consistent differences in levels of fixation bias "boost" in GC-rich isochores. For example, TAG is among the least "boosted" trinucleotides in the 33% GC class, compared to TGA which more highly exceeds its mutational equilibrium at high GC isochores. Similarly, TAG usage appears largely uncorrelated with local GC content. Any model (selection, mutation, or gene conversion) evoking a relationship between simple GC pressure and differences in nucleotide content cannot obviously account for a difference in boost between nucleotide-matched trinucleotides (e.g., TAG and TGA).

Given the ability of our complex mutation bias model to predict trinucleotide usage in low GC domains (Fig 6), we assume that our mutation bias estimation in GC-rich domains is also largely accurate. If so, complex mutation bias is unable to explain the repeatable boost scores (Fig 8). In principle, there could be several remaining classes of explanation. First, selection might act differently on underlying di or trimers. For example, regarding TAG and TGA, selection on TA or AG residues may be different to that on TG or GA ones. We can find no convincing evidence for this that can explain the universality of TAG avoidance (see S1 Text). One also needs to evoke selection that is strong enough throughout the human genome, which appears unlikely for reasons given above.

A further possibility is an interaction between complex mutation bias and gBGC making certain trinucleotides more liable to conversion owing to their relative commonality in populations. With a difference in mutational equilibria, the incidence of TAA/TAG meiotic heteroduplex mismatches (or sense/antisense ones to be more precise) is highly likely to be lower than that of TAA/TGA mismatches. Thus, gBGC may more commonly act on TAA/TGA. Overall, however, we see no correlation between our gBGC boost score and mutational equilibrium in any GC class of trinucleotides (Spearman's rank; $p > 0.05$ for 0%, 33%, 66%, and 100% GC trinucleotides). Pairwise comparison of all possible trinucleotide combinations also indicates that the trinucleotide with the higher mutational equilibrium does not necessarily receive the higher boost (Binomial test with null probability = 0.5, $p = 0.17$). This may reflect the fact that common trinucleotides are also more commonly substrates to be converted.

Finally, like mutation, gBGC may be contingent on the local sequence context such that, for example, TAG and CAG are relatively unaffected by gBGC, while TGA is affected. This could explain similar trends in bacteria and eukaryotes if, as is claimed, gBGC also operates in bacteria [122]. Complex specificity might be expected as many protein–nucleic acid interactions are contingent on local sequence context. For example, APOBEC3/A/B induced mutations account for many C→T and C→G mutations but occur predominantly in the context of TC [A/T] [123,124]. More specifically, several DNA repair processes are known to be affected by local sequence context [125] including, at least in bacteria, mismatch repair [126], the process underpinning gBGC. Here, sequence contexts that enhance localised DNA flexibility are associated with mismatch repair activation [126] (see also [127,128]). Similar evidence for a role of local DNA flexibility has been found in yeast [128,129]. The biological response elicited by CTG and CGG repeats in human trinucleotide repeat disorders may be mediated by their increased flexibility indicative of a relationship between local flexibility and trinucleotide content [130]. Evidence in humans for more effective repair of flexible DNA owing to local sequence context [131] suggests that an association between DNA mismatch repair and DNA flexibility may have relevance to understanding fixation biases in GC-rich domains. If flexibility is the core factor, then we might expect that a trinucleotide and its antisense should have similar boost scores as both feature in the same 3 base pairs of DNA (one on the Crick strand, the other on Watson). In our data, however, we find that the difference in gBGC "boost" between sense and antisense trinucleotides is no smaller than randomised trinucleotide comparisons ($p > 0.05$ regardless of the sequence analysed). This suggests that DNA flexibility alone cannot explain gBGC boost. Despite this, direct analysis of the sequence context associated with gBGC would be valuable. Preliminary data is consistent with k-mer dependency, especially CG contingency [15].

Methods

General methods

All data manipulation was performed using bespoke Python 3.6 scripts. Statistical analyses and data visualisations were performed using R 3.3.3. All scripts required for replication of the described analyses can be found at https://github.com/ath32/gBGC. While stop codons function at the mRNA level, we here analyse chromosomal DNA sequences and therefore refer to the 3 stops as TAA, TGA, and TAG.

Inferring stop codon switches from eukaryotic triplets

Lists of one-to-one orthologous genes were downloaded for a diverse variety of species triplets from the main Ensembl repository (release 101), Ensembl plants (release 46), or Ensembl metazoan (release 46): [1] primates; *Homo sapiens*, *Otolemur garnettii*, *Callithrix jacchus*, [2] cows; *Bison bison bison*, *Bos grunniens*, *Bos taurus*, [3] dogs; *Canis lupus familiaris*, *Ursus americanus*, *Vulpes vulpes*, [4] mice/rodents; *Mus musculus*, *Mus spretus*, *Rattus norvegicus*, [5] birds; *Gallus gallus*, *Anas platyrhynchos platyrhynchos*, *Meleagris_gallopavo*, [6] flies; *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Drosophila simulans*, [7] nematodes; *Caenorhabditis briggsae*, *Caenorhabditis remanei*, *Caenorhabditis elegans*, [8] plants; *Arabidopsis halleri*, *Arabidopsis lyrata*, *Arabidopsis thaliana*. Orthologous genes were extracted from the respective genomes using whole-genome sequence and gene annotation data downloaded from the same sources. Genes were filtered to retain genes with CDS length divisible by 3, no premature stop codons, and stop codons TAA, TGA, or TAG. Genes from each species triplet that met our quality controls were aligned using MAFFT with the -linsi algorithm [132].

Rather than using parsimony as done previously [49,133], stop codon switches were reconstructed using a maximum likelihood approach. For each species triplet, ancestral nucleotide states for the internal node between the 2 ingroups were inferred by maximum likelihood using IQTree v2.1.2 with the -asr flag [134,135]. This analysis does have one limitation in that we do not control for the possibility of parallel substitutions; however, we assume this effect to be small. To calculate stop codon flux rates, we compute the inferred ancestral stop codon state at the internal node and calculate transition from this ancestral state to a derived state (per incidence of the ancestral state).

Under the assumption that intronic GC reflects isochore GC content, intronic nucleotide sequences were extracted from 1 candidate genome within a species trio (e.g., the human genome was used as a representative of the primate triplet) using the appropriate GFF and WGS files downloaded from Ensembl (release 101). From the resulting spectra of intronic GC contents, genes were binned into GC-rich and GC-poor sets such that there were an equal number of flux events in each. This allowed effective segregation of genes by isochoric GC content for comparison of their stop codon flux rates. For primates, the same was done to compare HRGs and LRGs—see "Pseudo-autosomal regions, chromosome size, and local recombination rates" below for the source of recombination data.

Predicting equilibrium TGA content using flux data

The pTGA for a given lineage was calculated by adapting the formulae outlined by Long and colleagues [44]. In their study, given a spectrum of de novo mutations, they propose the equilibrium GC content, P_n, can be calculated from the GC→AT mutation rate divided by the reciprocal rate, m, such that:

We adapt this equation to the stop codon exemplar. As TAA and TGA stop codon usage covary in opposite directions with genomic GC content, we consider their usage to be dependent on one another. Due to the unusual biology of TAG, not least that it remains lowly used irrespective of genomic GC content, we exclude fluxes involving TAG from this calculation. Our proposed equation for calculating equilibrium TGA content, pTGA, from the ratio of TGA→TAA divided by TAA→TGA, s, is:

Null simulations to assign significance to observed pTGA deviation between 2 groups of genes

The difference in pTGA observed between 2 gene groups ("A and B," GC-rich and GC-poor genes, or HRGs and LRGs) may be assigned significance by comparisons to simulated null gene groups. First, by analysing all genes en masse we can calculate a genomic rate of TAA→TGA per TAA and for TGA→TAA per TGA. For each group of genes, we may then calculate null pTGA scores that control for these rates.

For each gene in the group, we determine the ancestral stop codon (of which we are only interested in TAA or TGA) and record the number of each. If the ancestral stop codon is TAA, we generate a random number between 0 and 1 and if equal to or below the genomic TAA→TGA rate we record a null TAA→TGA flux event. If the ancestral stop codon is TGA we generate a random number between 0 and 1 and if equal to or below the genomic TGA→TAA rate, we record a null TGA→TAA flux event. By this method, we thus receive null counts of TAA→TGA and TGA→TAA which may be divided by the ancestral counts of TAA and TGA to receive null flux rates. From these rates, we may calculate null pTGA, and thus by repeating this process 1,000 times we create a null distribution of pTGA for the gene group. Repeating this method for both gene groups, we have a distribution for gene group A and gene group B.

Next, we randomly sample with replacement 1 pTGA score from each of the 2 distributions, receiving a random pair. For each random pair, we calculate the deviation between that sampled from group A and group B and repeat this process 10,000 times to create a null distribution of differences. We then compare the observed difference between the real gene groups to this distribution, asking how many simulants have as high a difference as the observed one (n). The significance of the observed difference beyond null may be represented as $p = n / m$ where m is the number of random pairs considered.

Calculating mutational equilibria

The equilibrium content of all 4 nucleotides (indicated N*) may be estimated using the full mutational spectrum [136,137]. A full spectrum of 108,778 de novo mutations (from 1,548 Icelandic human family trios) was downloaded from the supplementary material of Jonsson, Sulem [102]. Knowing the rate of flux between every nucleotide (normalised to the occurrence of each nucleotide), we calculate the mutational equilibrium states of all nucleotides and GC content exactly as outlined in [137]. The same theory can be applied to the 3 stop codons to predict their equilibrium frequencies as follows, where TAA′ indicates the frequency of TAA after some period of time:

TAA′ = TAA (1 – TAA→TGA – TAA→TAG) + TGA (TGA→TAA) + TAG (TAG→TAA)

TGA′ = TGA (1 – TGA→TAA – TGA→TAG) + TAA (TAA→TGA) + TAG (TAG→TGA)

TAG′ = TAG (1 – TAG→TAA – TAG→TGA) + TAA (TAA→TAG) + TGA (TGA→TAG)

For equilibrium calculation, these simultaneous equations are solved such that TAA′ = TAA, etc. We are solving for gain = loss for each stop codon:

TAA (TAA→TGA + TAA→TAG) = TGA (TGA→TAA) + TAG (TAG→TAA)

TGA (TGA→TAA + TGA→TAG) = TAA (TAA→TGA) + TAG (TAG→TGA)

TAG (TAG→TAA + TAG→TGA) = TAA (TAA→TAG) + TGA (TGA→TAG)

Note that in these equations, we ignore the possibility of mutations from stop codons to sense codons. These we assume to be very rare and, should they occur, highly deleterious via the creation of C-terminal extensions. To constrain the results such that all equilibrium frequencies sum to 1, we replace 1 arbitrarily chosen stop codon frequency with 1—the sum of the other 2. While this would be achieved most accurately using precise mutational flux data between TAA, TGA, and TAG, this is not captured within the Jonsson [102] dataset. Instead, we estimate flux between the 3 stops using null frequencies proposed by Belinky and colleagues [49]. In their paper, they suggest the substitution control for TAA>TGA and TAA>TAG is A>G, for TGA>TAA and TAG>TAA is G>A, and for TGA>TAG and TAG>TGA is 2 × A>G × G>A.

The full spectrum of 108,778 de novo mutations may also be analysed using a 16 × 16 dinucleotide mutation matrix by tracing each mutation back to the reference genome and inferring dinucleotide changes. From the resultant matrix, we estimate the equilibrium frequencies of each dinucleotide by adapting the simultaneous equations above to consider flux into and away from each dinucleotide. An estimated GC* may then be calculated from the 16 dinucleotide equilibria, whereas TGA* (and other trinucleotide equilibrium frequencies) may instead be estimated by incorporating the 16 equilibria into Markov models, simulating null sequences, and calculating trinucleotide frequencies from these (see "Markov models for simulating null sequences").

Gene expression metrics

To assess the role of gene expression in mammalian stop codon evolution, we consider experimentally derived protein abundance data downloaded for *H. sapiens*, *B. taurus*, *C. familiaris*, and *M. musculus* from PaxDb [138]. As selection acts on protein activity, not mRNA levels, we consider this a robust measure. For species where multiple datasets are available, we employ the whole organism integrated set for maximum coverage of the proteome (see https://github.com/ath32/gBGC for accessions list).

Pseudo-autosomal regions, chromosome size, and local recombination rates

To assay the impact of recombination, we employed (a) chromosome size as a proxy of long-term recombination rate per bp; (b) pseudoautosomal localization, this being known to be highly recombinogenic; and (c) estimated recent recombination rates.

For the latter, we employed recombination rates generated by the HapMap2 project [139] using coordinates lifted to the hg19/GRCh37 human genome build by Adam Auton (available: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/). For this analysis, we hence use the GRCh37 human genome build and annotations, downloaded from NCBI and available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/ (last accessed 24 September 2020). For logistic regression modelling, each gene was assigned an estimated recombination rate equal to the average recombination rate of all its internal SNPs from the genetic map.

To assess the possible correlation between equilibrium GC content and recombination rate (see S7 Fig), we instead employ recombination rate bands directly assayed from 15,257 parent offspring pairs at 10 kb resolution. This we consider to be the better data to use for this analysis as de novo mutations may be reasonably assigned the recombination rate of the 10 kb band it falls within. The data were downloaded from https://www.decode.com/addendum/ (last accessed 14 September 2020) [140].

Coordinates of the 2 regions (PAR1 and PAR2) were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/grc/human, last accessed 14 September 2020). Chromosome sizes employed are base pair lengths derived from human genome build hg38.

Assessing the predictive abilities of gene expression and recombination rate

To determine whether expression and recombination rate can correctly predict the observed trends in stop codon usage, we employ logistic regression. Stop codon usage and GC3 content was captured alongside gene expression data or recombination data (depending on the feature to be examined). Models were fit and examined using the glm function in R with the "family = binomial" parameter. This produces a coefficient for each independent feature and associates a *p*-value for its predictive significance. We control for GC content by including GC3 content in a multivariate model when assessing expression level metrics. For the analysis of stop codon usage in null sequences, we instead use linear regression, also using glm in R, as more than 1 "stop codon" may be present in each sequence.

PGLS analysis of TGA enrichment and effective population size (N_e)

A phylogenetically controlled test of correlation between N_e and TGA enrichment in LEGs (lowest 25% of genes by protein abundance—see "Gene expression metrics" above) were facilitated by PGLS using the "caper" R package (https://CRAN.R-project.org/package=caper). N_e estimates are from species with well-resolved estimates of mutation rate and well-described polymorphism data, and are the same as used in Ho and Hurst [95]. Pagel's lambda (λ) was predicted by maximum likelihood. Species used in this analysis were the same as published in our previous analysis [95], with the input phylogenetic trees generated using TimeTree (http://www.timetree.org/) and available in our GitHub repository along with the data required to repeat this analysis. TGA enrichment scores in LEGs were calculated such that:

where mean TGA usage downstream is calculated from downstream codon positions +1 to +6. "Usage" refers to the relative frequency of TGA compared with the other stop codons TAA and TAA at position *n*, such that:

Markov models for simulating null sequences

Null trinucleotide frequencies were generated from a null model that controls for underlying mono- or dinucleotide mutation rates. To achieve this, we first calculate mutational equilibrium frequencies for all mono- or dinucleotides—see "Calculating mutational equilibria" above and Rice and colleagues [137]. We next simulate 10,000 sequences (of average coding sequence length) using Markov models in a similar way to that outlined by Ho and Hurst [141]. The first nucleotide/dinucleotide of each simulant is selected at random according to equilibrium nucleotide/dinucleotide frequencies. The following nucleotide is selected from a second set of frequencies: given the prior nucleotide in the simulation, what is the probability that the next nucleotide should be A, C, G, or T. As all trinucleotides occur in these simulated sequences at a rate dictated by a derived mutational matrix, trinucleotide frequencies in the real sequences that are deviant from the simulations indicates enrichment or under-enrichment beyond chance.

Supporting information

S1 Fig. The relationships of autosome length with GC content and TGA usage in the human genome. Autosomal size (bp length) is negatively associated with G+C content (Spearman's rank; $p = 0.0078$, rho = –0.56, $n = 22$) and TGA usage (Spearman's rank; $p = 0.0094$, rho = –0.55, $n = 22$). Underlying data can be found in S9 Data.
(PDF)

S2 Fig. Trinucleotide frequencies in 6 sets of different genomic sequences (between 0%–36.31% GC content) compared to dinucleotide matrix-derived equilibrium predictions. The GC range used is the bottom 20% of genes to avoid the possible confounding effects of biased gene conversion. CDS refers to coding sequence, CREs to *cis*-regulatory elements. "xDinuc matrix" refers to equilibrium estimates of trinucleotide frequencies derived from a dinucleotide mutational matrix. Underlying data can be found in S6 Data. CDS, coding sequence; CRE, *cis*-regulatory element; ncRNA, noncoding RNA.
(PDF)

S3 Fig. The rate of flux increasing GC content at 2-fold degenerate sites divided by the rate of flux decreasing GC content at the same sites across 10 gene bins of increasing recombination rate. Flux to the G+C-rich codons is most strongly favoured at high recombination rates (Spearman's rank; $p < 2.2 \times 10^{-16}$, rho = 0.99), consistent with the possible action of GC-biased gene conversion. Underlying data can be found in S10 Data.
(PDF)

S4 Fig. The rate of flux increasing GC content at 2-fold degenerate sites divided by the rate of flux decreasing GC content at the same sites across 10 gene bins of increasing recombination rate for each appropriate amino acid. Flux increasing GC content are significantly favoured in regions with higher recombination rate in 10 of the 12 amino acids before Bonferroni correction (Spearman's rank tests; $p < 0.05$), the 2 exceptions to this being Leucine and Glutamine. Underlying data can be found in S10 Data.
(PDF)

S5 Fig. Deviation scores, (O-E)/E, describing the difference in gBGC "boost" for each trinucleotide individually. The normalised differences, (O-E)/E, between estimated trinucleotide mutational equilibrium frequencies (calculated from DNMs) and fixed trinucleotide frequencies (from 10 kb sequences surrounding those mutations) were calculated for GC-rich (top 20%, 45.5%–100%, "group 5") and GC-poor (bottom 20%, 0%–36.3%, "group 1")

sequences surrounding 108,778 DNMs. As we predict, GC-rich sequences to be subjected to stronger biased gene conversion, we predict a larger differential between fixed and equilibrium frequency, *D*, for GC-rich trinucleotides in GC-rich sequences. The extent to which a trinucleotide is "boosted" by biased gene conversion can hence be accessed by measuring the difference, (O-E)/E, in *D* between the GC-richest and GC-poorest sequences. Trinucleotides are ordered from low to high according to the extent they are "boosted" by biased gene conversion. Underlying data can be found in S7 data. DNM, de novo mutation; (O-E)/E, (Observed-Expected)/Expected.
(PDF)

S6 Fig. Trinucleotides containing (a) A, G, and T and (b) A, C, and T within the 33% GC-content class of trinucleotides ranked by gBGC "boost" scores. TGA receives a consistent higher GC-coupled fixation boost than TAG which performs the second worst (after GTA). TCA similarly receives a consistently higher GC-coupled fixation boost than TAC. Sequences analysed include CRE, 5′ UTR (5), intronic (intron), ncRNA, (ncrna), and 3′ UTR (3). CDS sequences are excluded from this analysis as they are much more prone to selection and other potential fixation biases. Underlying data can be found in S8 data. CDS, coding sequence; CRE, *cis*-regulatory element; gBGC, GC-biased gene conversion; ncRNA, noncoding RNA.
(PDF)

S7 Fig. Predicted G+C equilibrium (G+C*) and TGA equilibrium (TGA*) frequencies from de novo mutations of various recombination rates. Mutations were assigned a recombination rate based upon their local 10 kbp environment. Mutations in nonrecombining regions were discarded. The remaining mutations were split into bins of equal size (approximately 5,000 mutations) for the calculation of GC* and TGA*. Recombination rate is not correlated with GC* (Spearman's rank; $p = 0.58$, rho = −0.2) nor TGA* (Spearman's rank; $p = 0.63$, rho = −0.18) when estimated from de novo mutations. Underlying data can be found in S11 Data.
(PDF)

S1 Table. Results of linear regression models predicting stop codon (TAA, TGA, TAG) trinucleotide usage as a function of intronic G+C content in 5′ and 3′ UTR sequences and as a function of coding sequence GC3 content in intronic sequences.
(PDF)

S2 Table. The 4 × 4 mutational matrix for 108,778 observed de novo mutations in 1,548 human trios. Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome. Under the assumption that the observed number of mutations is a Poisson variable, 95% confidence intervals (CI) were calculated using the Poisson.test function in R.
(PDF)

S1 Text. Possible selective explanations for TAG avoidance compared with TGA.
(PDF)

S1 Data. Underlying data for Fig 1.
(XLSX)

S2 Data. Underlying data for Fig 2.
(XLSX)

S3 Data. Underlying data for Fig 3.
(XLSX)

S4 Data. Underling data for Fig 4.
(XLSX)

S5 Data. Underlying data for Fig 5.
(XLSX)

S6 Data. Underlying data for Figs 6 and S2.
(XLSX)

S7 Data. Underlying data for Figs 7 and S5.
(XLSX)

S8 Data. Underling data for Figs 8 and S6.
(XLSX)

S9 Data. Underling data for S1 Fig.
(XLSX)

S10 Data. Underlying data for S3 Fig.
(XLSX)

S11 Data. Underlying data for S7 Fig.
(XLSX)

Author Contributions

Conceptualization: Alexander Thomas Ho, Laurence Daniel Hurst.

Data curation: Alexander Thomas Ho, Laurence Daniel Hurst.

Formal analysis: Alexander Thomas Ho.

Funding acquisition: Laurence Daniel Hurst.

Investigation: Alexander Thomas Ho.

Methodology: Alexander Thomas Ho, Laurence Daniel Hurst.

Project administration: Alexander Thomas Ho, Laurence Daniel Hurst.

Software: Alexander Thomas Ho.

Supervision: Laurence Daniel Hurst.

Validation: Alexander Thomas Ho, Laurence Daniel Hurst.

Visualization: Alexander Thomas Ho.

Writing – original draft: Alexander Thomas Ho.

Writing – review & editing: Alexander Thomas Ho, Laurence Daniel Hurst.

References

1. Ponting CP. Biological function in the twilight zone of sequence conservation. BMC Biol. 2017; 15 (1):1–9. https://doi.org/10.1186/s12915-016-0343-5 PMID: 28100223
2. Nielsen R, Yang ZH. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998; 148(3):929–36. https://doi.org/10.1093/genetics/148.3.929 PMID: 9539414
3. Yang ZH, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000; 15(12):496–503. https://doi.org/10.1016/s0169-5347(00)01994-7 PMID: 11114436

4. Pond SLK, Frost SDW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005; 22(5):1208–22. https://doi.org/10.1093/molbev/msi105 PMID: 15703242

5. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 2002; 18 (9):486–7. https://doi.org/10.1016/s0168-9525(02)02722-1 PMID: 12175810

6. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods. 2010; 7(4):250–1. https://doi.org/10.1038/nmeth0410-250 PMID: 20354513

7. Sun H, Yu GJ. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. Sci Rep. 2019; 9(1):1667. https://doi.org/10.1038/s41598-018-38189-9 PMID: 30733553

8. Ponting CP. The functional repertoires of metazoan genomes. Nat Rev Genet. 2008; 9(9):689–98. https://doi.org/10.1038/nrg2413 PMID: 18663365

9. Lachance J, Tishkoff SA. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. Am J Hum Genet. 2014; 95(4):408–20. https://doi.org/10.1016/j.ajhg.2014.09.008 PMID: 25279983

10. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. Genetics. 2001; 159(2):907–11. https://doi.org/10.1093/genetics/159.2.907 PMID: 11693127

11. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009; 10(1):285–311. https://doi.org/10.1146/annurev-genom-082908-150001 PMID: 19630562

12. Galtier N, Duret L, Glemin S, Ranwez V. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 2009; 25(1):1–5. https://doi.org/10.1016/j.tig.2008.10.011 PMID: 19027980

13. Brown TC, Jiricny J. Different base base mispairs are corrected with different efficiencies and specificities in monkey kidney-cells. Cell. 1988; 54(5):705–11. https://doi.org/10.1016/s0092-8674(88)80015-1 PMID: 2842064

14. Brown TC, Jiricny J. Repair of base base mismatches in simian and human-cells. Genome. 1989; 31 (2):578–83. https://doi.org/10.1139/g89-107 PMID: 2561110

15. Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, et al. The rate of meiotic gene conversion varies by sex and age. Nat Genet. 2016; 48(11):1377–84. https://doi.org/10.1038/ng.3669 PMID: 27643539

16. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, et al. The mosaic genome of warm-blooded vertebrates. Science. 1985; 228(4702):953–8. https://doi.org/10.1126/science.4001930 PMID: 4001930

17. Eyre-Walker A, Hurst LD. The evolution of isochores. Nat Rev Genet. 2001; 2(7):549–55. https://doi.org/10.1038/35080577 PMID: 11433361

18. Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, et al. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife. 2015; 4(1):e04637. https://doi.org/10.7554/eLife.04637 PMID: 25806687

19. Fullerton SM, Bernardo Carvalho A, Clark AG. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 2001; 18(6):1139–42. https://doi.org/10.1093/oxfordjournals.molbev.a003886 PMID: 11371603

20. Eyre-Walker A. Recombination and mammalian genome evolution. Proc R Soc Lond B Biol Sci. 1993; 252(1335):237–43. https://doi.org/10.1098/rspb.1993.0071 PMID: 8394585

21. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol Evol. 2012; 4(7):675–82. https://doi.org/10.1093/gbe/evs052 PMID: 22628461

22. Duret L, Arndt PF. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 2008; 4(5):e1000071. https://doi.org/10.1371/journal.pgen.1000071 PMID: 18464896

23. Marsolier-Kergoat MC, Yeramian E. GC Content and Recombination: Reassessing the Causal Effects for the Saccharomyces cerevisiae Genome. Genetics. 2009; 183(1):31–8. https://doi.org/10.1534/genetics.109.105049 PMID: 19546316

24. Kiktev DA, Sheng ZW, Lobachev KS, Petes TD. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2018; 115(30):E7109–E18. https://doi.org/10.1073/pnas.1807334115 PMID: 29987035

25. Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. The evolution of isochores: evidence from SNP frequency distributions. Genetics. 2002; 162(4):1805–10. https://doi.org/10.1093/genetics/162.4.1805 PMID: 12524350

26. Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. Vanishing GC-rich isochores in mammalian genomes. Genetics. 2002; 162(4):1837–47. https://doi.org/10.1093/genetics/162.4.1837 PMID: 12524353

27. Glemin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. Quantification of GC-biased gene conversion in the human genome. Genome Res. 2015; 25(8):1215–28. https://doi.org/10.1101/gr.185488.114 PMID: 25995268

28. Galtier N. Fine-scale quantification of GC-biased gene conversion intensity in mammals. bioRxiv. 2021. https://doi.org/10.1101/2021.05.05.442789

29. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature. 2008; 454(7203):479–85. https://doi.org/10.1038/nature07135 PMID: 18615017

30. Liu HX, Maclean CJ, Zhang JZ. Evolution of the yeast recombination landscape. Mol Biol Evol. 2019; 36(2):412–22. https://doi.org/10.1093/molbev/msy233 PMID: 30535029

31. Liu HX, Huang J, Sun XG, Li J, Hu YW, Yu LY, et al. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. Nat Ecol Evol. 2018; 2(1):164–73. https://doi.org/10.1038/s41559-017-0372-7 PMID: 29158556

32. Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 2014; 15(12):549. https://doi.org/10.1186/s13059-014-0549-1 PMID: 25496599

33. Rousselle M, Laverre A, Figuet E, Nabholz B, Galtier N. Influence of recombination and GC-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. Mol Biol Evol. 2019; 36(3):458–71. https://doi.org/10.1093/molbev/msy243 PMID: 30590692

34. Gutz H, Leslie JF. Gene conversion: a hitherto overlooked parameter in population genetics. Genetics. 1976; 83(4):861–6. https://doi.org/10.1093/genetics/83.4.861 PMID: 971809

35. Nagylaki T. Evolution of a large population under gene conversion. Proc Natl Acad Sci U S A. 1983; 80 (19):5941–5. https://doi.org/10.1073/pnas.80.19.5941 PMID: 6577463

36. Berglund J, Pollard KS, Webster MT. Hotspots of Biased Nucleotide Substitutions in Human Genes. PLoS Biol. 2009; 7(1):45–62. https://doi.org/10.1371/journal.pbio.1000026 PMID: 19175294

37. Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, et al. Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc Lond B Biol Sci. 2010; 365(1552):2571–80. https://doi.org/10.1098/rstb.2010.0007 PMID: 20643747

38. Dreszer TR, Wall GD, Haussler D, Pollard KS. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. Genome Res. 2007; 17(10):1420–30. https://doi.org/10.1101/gr.6395807 PMID: 17785536

39. Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. Genome Biol Evol. 2017; 9 (11):2987–3007. https://doi.org/10.1093/gbe/evx213 PMID: 29045655

40. Bolivar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, et al. 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. 2018; 35(10):2475–86. https://doi.org/10.1093/molbev/msy149 PMID: 30085180

41. Harrison RJ, Charlesworth B. Biased gene conversion affects patterns of codon usage and amino acid usage in the Saccharomyces sensu stricto group of yeasts. Mol Biol Evol. 2011; 28(1):117–29. https://doi.org/10.1093/molbev/msq191 PMID: 20656793

42. Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet. 2010; 6(9):e1001115. https://doi.org/10.1371/journal.pgen.1001115 PMID: 20838599

43. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 2008; 105(27):9272–7. https://doi.org/10.1073/pnas.0803466105 PMID: 18583475

44. Long HA, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo WF, et al. Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol. 2018; 2(2):237–40. https://doi.org/10.1038/s41559-017-0425-y PMID: 29292397

45. Smith NGC, Eyre-Walker A. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol Biol Evol. 2001; 18(6):982–6. https://doi.org/10.1093/oxfordjournals.molbev.a003899 PMID: 11371586

46. Smith TCA, Arndt PF, Eyre-Walker A. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLoS Genet. 2018; 14(3):e1007254. https://doi.org/10.1371/journal.pgen.1007254 PMID: 29590096

47. Bengtsson BO. Biased conversion as the primary function of recombination. Genet Res. 1985; 47 (1):77–80.

48. Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of Drosophila. BMC Evol Biol. 2007; 7(1):226. https://doi.org/10.1186/1471-2148-7-226 PMID: 18005411

49. Belinky F, Babenko VN, Rogozin IB, Koonin EV. Purifying and positive selection in the evolution of stop codons. Sci Rep. 2018; 8(1):9260. https://doi.org/10.1038/s41598-018-27570-3 PMID: 29915293

50. Rodnina MV, Korniy N, Klimova M, Karki P, Peng BZ, Senyushkina T, et al. Translational recoding: canonical translation mechanisms reinterpreted. Nucleic Acids Res. 2020; 48(3):1056–67. https://doi.org/10.1093/nar/gkz783 PMID: 31511883

51. Roy B, Leszyk JD, Mangus DA, Jacobson A. Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. Proc Natl Acad Sci U S A. 2015; 112(10):3038–43. https://doi.org/10.1073/pnas.1424127112 PMID: 25733896

52. Beznoskova P, Gunisova S, Valasek LS. Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. RNA. 2016; 22(3):456–66. https://doi.org/10.1261/rna.054452.115 PMID: 26759455

53. Rodnina MV. The ribosome in action: Tuning of translational efficiency and protein folding. Protein Sci. 2016; 25(8):1390–406. https://doi.org/10.1002/pro.2950 PMID: 27198711

54. Strigini P, Brickman E. Analysis of specific misreading in Escherichia coli. J Mol Biol. 1973; 75(4):659–72. https://doi.org/10.1016/0022-2836(73)90299-4 PMID: 4581523

55. Parker J. Errors and alternatives in reading the universal genetic code. Microbiol Rev. 1989; 53 (3):273–98. https://doi.org/10.1128/mr.53.3.273-298.1989 PMID: 2677635

56. Meng SY, Hui JO, Haniu M, Tsai LB. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in Escherichia coli. Biochem Biophys Res Commun. 1995; 211(1):40–8. https://doi.org/10.1006/bbrc.1995.1775 PMID: 7779107

57. Sanchez JC, Padron G, Santana H, Herrera L. Elimination of an HuIFN alpha 2b readthrough species, produced in Escherichia coli, by replacing its natural translational stop signal. J Biotechnol. 1998; 63 (3):179–86. https://doi.org/10.1016/s0168-1656(98)00073-x PMID: 9803532

58. Roth JR. UGA nonsense mutations in Salmonella-typhimurium. J Bacteriol. 1970; 102(2):467–75. https://doi.org/10.1128/jb.102.2.467-475.1970 PMID: 4315894

59. Ryden SM, Isaksson LA. A temperature-sensitive mutant of Escherichia-coli that shows enhanced misreading of UAG/A and increased efficiency for some transfer-RNA nonsense suppressors. Mol Gen Genet. 1984; 193(1):38–45. https://doi.org/10.1007/BF00327411 PMID: 6419024

60. Geller AI, Rich A. A UGA termination suppression tRNATrp active in rabbit reticulocytes. Nature. 1980; 283(5742):41–6. https://doi.org/10.1038/283041a0 PMID: 7350525

61. Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. Nucleic Acids Res. 2018; 46 (4):1927–44. https://doi.org/10.1093/nar/gkx1315 PMID: 29325104

62. Li C, Zhang J. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. PLoS Genet. 2019; 15(5):e1008141. https://doi.org/10.1371/journal.pgen.1008141 PMID: 31120886

63. Arribere JA, Cenik ES, Jain N, Hess GT, Lee CH, Bassik MC, et al. Translation readthrough mitigation. Nature. 2016; 534(7609):719–23. https://doi.org/10.1038/nature18308 PMID: 27281202

64. Wagner A. Robustness, evolvability, and neutrality. FEBS Lett. 2005; 579(8):1772–8. https://doi.org/10.1016/j.febslet.2005.01.063 PMID: 15763550

65. Clegg JB, Weatherall DJ, Milner PF. Haemoglobin constant spring—a chain termination mutant? Nature. 1971; 234(5328):337–40. https://doi.org/10.1038/234337a0 PMID: 4944483

66. Namy O, Duchateau-Nguyen G, Rousset JP. Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. Mol Microbiol. 2002; 43(3):641–52. https://doi.org/10.1046/j.1365-2958.2002.02770.x PMID: 11929521

67. Pang SY, Wang WH, Rich B, David R, Chang YT, Carbunaru G, et al. A novel nonstop mutation in the stop codon and a novel missense mutation in the type II 3 beta-hydroxysteroid dehydrogenase (3 beta-HSD) gene causing, respectively, nonclassic and classic 3 beta-HSD deficiency congenital adrenal hyperplasia. J Clin Endocrinol Metab. 2002; 87(6):2556–63. https://doi.org/10.1210/jcem.87.6.8559 PMID: 12050213

68. Vidal R, Frangione B, Rostagno A, Mead S, Revesz T, Plant G, et al. A stop-codon mutation in the BRI gene associated with familial British dementia. Nature. 1999; 399(6738):776–81. https://doi.org/10.1038/21637 PMID: 10391242

69. Vidal R, Revesz T, Rostagno A, Kim E, Holton JL, Bek T, et al. A decamer duplication in the 3 ' region of the BRI gene originates an amyloid peptide that is associated with dementia in a Danish kindred. Proc Natl Acad Sci U S A. 2000; 97(9):4920–5. https://doi.org/10.1073/pnas.080076097 PMID: 10781099

70. Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L, et al. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. New Engl J Med. 2005; 352(3):254–66. https://doi.org/10.1056/NEJMoa041974 PMID: 15659725

71. Hollingsworth TJ, Gross AK. The severe autosomal dominant retinitis pigmentosa rhodopsin mutant Ter349Glu mislocalizes and induces rapid rod cell death. J Biol Chem. 2013; 288(40):29047–55. https://doi.org/10.1074/jbc.M113.495184 PMID: 23940033

72. Dimitrova LN, Kuroha K, Tatematsu T, Inada T. Nascent Peptide-dependent Translation Arrest Leads to Not4p-mediated Protein Degradation by the Proteasome. J Biol Chem. 2009; 284(16):10343–52. https://doi.org/10.1074/jbc.M808840200 PMID: 19204001

73. Klauer AA, van Hoof A. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. Wiley Interdiscip Rev RNA. 2012; 3(5):649–60. https://doi.org/10.1002/wrna.1124 PMID: 22740367

74. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell. 2010; 141(2):344–54. https://doi.org/10.1016/j.cell.2010.03.031 PMID: 20403328

75. McCaughan KK, Brown CM, Dalphin ME, Berry MJ, Tate WP. Translational termination efficiency in mammals is influenced by the base following the stop codon. Proc Natl Acad Sci U S A. 1995; 92 (12):5431–5. https://doi.org/10.1073/pnas.92.12.5431 PMID: 7777525

76. Seoighe C, Kiniry SJ, Peters A, Baranov PV, Yang H. Selection shapes synonymous stop codon use in mammals. J Mol Evol. 2020; 88(7):549–61. https://doi.org/10.1007/s00239-020-09957-x PMID: 32617614

77. Trotta E. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. BMC Genomics. 2016; 17(17):366. https://doi.org/10.1186/s12864-016-2692-4 PMID: 27188984

78. Meunier J, Duret L. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 2004; 21(6):984–90. https://doi.org/10.1093/molbev/msh070 PMID: 14963104

79. Blaschke RJ, Rappold G. The pseudoautosomal regions, SHOX and disease. Curr Opin Genet Dev. 2006; 16(3):233–9. https://doi.org/10.1016/j.gde.2006.04.004 PMID: 16650979

80. Li W-H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 1987; 24(4):337–45. https://doi.org/10.1007/BF02134132 PMID: 3110426

81. Bulmer M. The Selection-Mutation-Drift Theory of Synonymous Codon Usage. Genetics. 1991; 129 (3):897–907. https://doi.org/10.1093/genetics/129.3.897 PMID: 1752426

82. Marais G, Charlesworth B, Wright SI. Recombination and base composition: the case of the highly self-fertilizing plant Arabidopsis thaliana. Genome Biol. 2004; 5(7):R45. https://doi.org/10.1186/gb-2004-5-7-r45 PMID: 15239830

83. Smeds L, Mugal CF, Qvarnstrom A, Ellegren H. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. PLoS Genet. 2016; 12(5):e1006044. https://doi.org/10.1371/journal.pgen.1006044 PMID: 27219623

84. Robinson MC, Stone EA, Singh ND. Population Genomic Analysis Reveals No Evidence for GC-Biased Gene Conversion in Drosophila melanogaster. Mol Biol Evol. 2014; 31(2):425–33. https://doi.org/10.1093/molbev/mst220 PMID: 24214536

85. Xu C, Park JK, Zhang JZ. Evidence that alternative transcriptional initiation is largely nonadaptive. PLoS Biol. 2019; 17(3):e3000197. https://doi.org/10.1371/journal.pbio.3000197 PMID: 30883542

86. Xu C, Zhang JZ. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. Cell Syst. 2018; 6(6):734–42. https://doi.org/10.1016/j.cels.2018.05.007 PMID: 29886108

87. de Oliveira JL, Morales AC, Hurst LD, Urrutia AO, Thompson CRL, Wolf JB. Inferring adaptive codon preference to understand sources of selection shaping codon usage bias. Mol Biol Evol. 2021; 38 (8):3247–66. https://doi.org/10.1093/molbev/msab099 PMID: 33871580

88. Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 2002; 12 (6):640–9. https://doi.org/10.1016/s0959-437x(02)00353-2 PMID: 12433576

89. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. 2006; 4(6):933–42. https://doi.org/10.1371/journal.pbio.0040180 PMID: 16700628

90. Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, et al. Codon usage and splicing jointly influence mRNA localization. Cell Syst. 2020; 10(4):351–62. https://doi.org/10.1016/j.cels.2020.03.001 PMID: 32275854

91. Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. A unification of mosaic structures in the human genome. Hum Mol Genet. 2003; 12(19):2411–5. https://doi.org/10.1093/hmg/ddg251 PMID: 12915446

92. Ohta T. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst. 1992; 23(1):263–86.

93. Lynch M. The Origins of Genome Architecture. Sunderland, MA: Sinauer Assocs., Inc.; 2007.

94. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glemin S, et al. Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol. 2018; 35(5):1092–103. https://doi.org/10.1093/molbev/msy015 PMID: 29390090

95. Ho AT, Hurst LD. Effective population size predicts local rates but not local mitigation of read-through errors in eukaryotic genes. Mol Biol Evol. 2020; 38(1):244–62.

96. Korkmaz G, Holm M, Wiens T, Sanyal S. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem. 2014; 289(44):30334–42. https://doi.org/10.1074/jbc.M114.606632 PMID: 25217634

97. Wolfe KH, Sharp PM, Li WH. Mutation rates differ among regions of the mammalian genome. Nature. 1989; 337(6204):283–5. https://doi.org/10.1038/337283a0 PMID: 2911369

98. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer. 2014; 14(12):786–800. https://doi.org/10.1038/nrc3816 PMID: 25568919

99. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980; 287 (5782):560–1. https://doi.org/10.1038/287560a0 PMID: 6999365

100. Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci U S A. 1990; 87(12):4692–6. https://doi.org/10.1073/pnas.87.12.4692 PMID: 2352943

101. Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 2005; 22(3):650–8. https://doi.org/10.1093/molbev/msi043 PMID: 15537806

102. Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017; 549(7673):519–22. https://doi.org/10.1038/nature24018 PMID: 28959963

103. Sun JH, Ai SM, Luo HJ, Gao B. Estimation of the equilibrium GC content of human genome. 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology. Hangzhou, China: IEEE; 2019. p. 12–7.

104. Warnecke T, Hurst LD. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. Mol Biol Evol. 2007; 24 (12):2755–62. https://doi.org/10.1093/molbev/msm210 PMID: 17905999

105. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966; 8(3):269–94. PMID: 5980116

106. Basham B, Schroth GP, Ho PS. An A-DNA triplet code—Thermodynamic rules for predicting A-DNA and B-DNA. Proc Natl Acad Sci U S A. 1995; 92(14):6464–8. https://doi.org/10.1073/pnas.92.14.6464 PMID: 7604014

107. Babbitt GA, Schulze KV. Codons support the maintenance of intrinsic DNA polymer flexibility over evolutionary timescales. Genome Biol Evol. 2012; 4(9):954–65. https://doi.org/10.1093/gbe/evs073 PMID: 22936074

108. Vinogradov AE. Bendable genes of warm-blooded vertebrates. Mol Biol Evol. 2001; 18(12):2195–200. https://doi.org/10.1093/oxfordjournals.molbev.a003766 PMID: 11719569

109. Warnecke T, Batada NN, Hurst LD. The impact of the nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet. 2008; 4(11):e1000250. https://doi.org/10.1371/journal.pgen.1000250 PMID: 18989456

110. Prendergast JGD, Semple CAM. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res. 2011; 21(11):1777–87. https://doi.org/10.1101/gr.122275.111 PMID: 21903742

111. Langley SA, Karpen GH, Langley CH. Nucleosomes shape DNA polymorphism and divergence. PLoS Genet. 2014; 10(7):e1004457. https://doi.org/10.1371/journal.pgen.1004457 PMID: 24991813

112. Babbitt GA, Cotter CR. Functional conservation of nucleosome formation selectively biases presumably neutral molecular variation in yeast genomes. Genome Biol Evol. 2011; 3(1):15–22. https://doi.org/10.1093/gbe/evq081 PMID: 21135411

113. Karlin S, Mrazek J. Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci U S A. 1997; 94(19):10227–32. https://doi.org/10.1073/pnas.94.19.10227 PMID: 9294192

114. Mrazek J, Karls AC. In silico simulations of occurrence of transcription factor binding sites in bacterial genomes. BMC Evol Biol. 2019; 19(1):67. https://doi.org/10.1186/s12862-019-1381-8 PMID: 30823869

115. Hahn MW, Stajich JE, Wray GA. The effects of selection against spurious transcription factor binding sites. Mol Biol Evol. 2003; 20(6):901–6. https://doi.org/10.1093/molbev/msg096 PMID: 12716998

116. Berg OG. Selection intensity for codon bias and the effective population size of Escherichia coli. Genetics. 1996; 142(4):1379–82. https://doi.org/10.1093/genetics/142.4.1379 PMID: 8846914

117. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife. 2013; 2(1):e01179. https://doi.org/10.7554/eLife.01179 PMID: 24302569

118. Schueren F, Thoms S. Functional translational readthrough: a systems biology perspective. PLoS Genet. 2016; 12(8):e1006196. https://doi.org/10.1371/journal.pgen.1006196 PMID: 27490485

119. Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, et al. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res. 2011; 21(12):2096–113. https://doi.org/10.1101/gr.119974.110 PMID: 21994247

120. Yordanova MM, Loughran G, Zhdanov AV, Mariotti M, Kiniry SJ, Oconnor PBF, et al. AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation. Nature. 2018; 553 (7688):356–60. https://doi.org/10.1038/nature25174 PMID: 29310120

121. Wu XM, Hurst LD. Why Selection Might Be Stronger When Populations Are Small: Intron Size and Density Predict within and between-Species Usage of Exonic Splice Associated cis-Motifs. Mol Biol Evol. 2015; 32(7):1847–61. https://doi.org/10.1093/molbev/msv069 PMID: 25771198

122. Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. PLoS Genet. 2015; 11(2):e1004941. https://doi.org/10.1371/journal.pgen.1004941 PMID: 25659072

123. Seplyarskiy VB, Andrianova MA, Bazykin GA. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. Genome Res. 2017; 27(2):175–84. https://doi.org/10.1101/gr.210336.116 PMID: 27940951

124. Chen J, MacCarthy T. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Comp Biol. 2017; 13(3):e1005471. https://doi.org/10.1371/journal.pcbi.1005471 PMID: 28362825

125. Cai Y, Patel DJ, Broyde S, Geacintov NE. Base Sequence Context Effects on Nucleotide Excision Repair. J Nucleic Acids. 2010; 2010:174252. https://doi.org/10.4061/2010/174252 PMID: 20871811

126. Mazurek A, Johnson CN, Germann MW, Fishel R. Sequence context effect for hMSH2-hMSH6 mismatch-dependent activation. Proc Natl Acad Sci U S A. 2009; 106(11):4177–82. https://doi.org/10.1073/pnas.0808572106 PMID: 19237577

127. Wang H, Yang Y, Schofield MJ, Du CW, Fridman Y, Lee SD, et al. DNA bending and unbending by MutS govern mismatch recognition and specificity. Proc Natl Acad Sci U S A. 2003; 100(25):14822–7. https://doi.org/10.1073/pnas.2433654100 PMID: 14634210

128. Isaacs RJ, Rayens WS, Spielmann HP. Structural differences in the NOE-derived structure of G-T mismatched DNA relative to normal DNA are correlated with differences in C-13 relaxation-based internal dynamics. J Mol Biol. 2002; 319(1):191–207. https://doi.org/10.1016/S0022-2836(02)00265-6 PMID: 12051946

129. Li Y, Lombardo Z, Joshi M, Hingorani MM, Mukerji I. Mismatch recognition by Saccharomyces cerevisiae Msh2-Msh6: Role of structure and dynamics. Int J Mol Sci. 2019; 20(17):4271. https://doi.org/10.3390/ijms20174271 PMID: 31480444

130. Bacolla A, Gellibolian R, Shimizu M, Amirhaeri S, Kang S, Ohshima K, et al. Flexible DNA: Genetically unstable CTG center dot CAG and CGG center dot CCG from human hereditary neuromuscular disease genes. J Biol Chem. 1997; 272(27):16783–92. https://doi.org/10.1074/jbc.272.27.16783 PMID: 9201983

131. Ruzicka M, Soucek R, Kulhanek P, Radova L, Fajkusova L, Reblova K. Bending of DNA duplexes with mutation motifs. DNA Res. 2019; 26(4):341–52. https://doi.org/10.1093/dnares/dsz013 PMID: 31230075

132. Katoh K, Kuma K-i, Miyata T, Toh H. Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome Inform. 2005; 16(1):22–33. PMID: 16362903

133. Rogozin IB, Belinky F, Pavlenko V, Shabalina SA, Kristensen DM, Koonin EV. Evolutionary switches between two serine codon sets are driven by selection. Proc Natl Acad Sci U S A. 2016; 113 (46):13109–13. https://doi.org/10.1073/pnas.1615832113 PMID: 27799560

134. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

135. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020; 37(5):1530–4. https://doi.org/10.1093/molbev/msaa015 PMID: 32011700

136. Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. Atypical at skew in firmicute genomes results from selection and not from mutation. PLoS Genet. 2011; 7(9):e1002283. https://doi.org/10.1371/journal.pgen.1002283 PMID: 21935355

137. Rice AM, Morales AC, Ho AT, Mordstein C, Muhlhausen S, Watson S, et al. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol. 2020; 38(1):67–83.

138. Wang MC, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015; 15 (18):3163–8. https://doi.org/10.1002/pmic.201400441 PMID: 25656970

139. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449(7164):851–61. https://doi.org/10.1038/nature06258 PMID: 17943122

140. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010; 467 (7319):1099–103. https://doi.org/10.1038/nature09525 PMID: 20981099

141. Ho AT, Hurst LD. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. PLoS Genet. 2019; 15(9):e1008386. https://doi.org/10.1371/journal.pgen.1008386 PMID: 31527909