



PHD

**Errors, stop codons, and the evolution of genomes
(Alternative Format Thesis)**

Ho, Alex

Award date:
2023

Awarding institution:
University of Bath

[Link to publication](#)

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy

If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Errors, stop codons, and the evolution of genomes

Alexander Thomas Ho

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

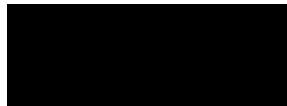
March 2022

COPYRIGHT:

Attention is drawn to the fact that copyright of this thesis rests with the author. A copy of this thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that they must not copy it or use material from it except as permitted by law or with the consent of the author.

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree. I am the author of this thesis, and the work described therein was carried out by myself personally, with the exceptions of the two appendices containing work from collaborators. For these manuscripts, contribution details have been included.

SIGNED:



Acknowledgements

Thank you, Laurence, for your support and mentorship throughout my undergraduate studies and PhD. Thank you for believing in me and thank you for providing the perfect environment for me to develop as a scientist. I have learned so much from our thought-provoking conversations.

To my colleagues in the lab – Alan, Atahualpa, Dana, Liam, Lucy, Rosina and Will – thank you for making the last ~3.5 years thoroughly enjoyable. I'll miss the lunches, the laughs, and the drunken dancing at conferences. To Christine, Laura, Bethan, Greg, and Samir – thank you for those SARS-CoV-2 collaborations that kept me busy (but smiling) during the early months of the pandemic.

To my Mum and Dad, as always, thank you for your support. Undertaking a PhD was a step into the unknown and I couldn't have done it without you. Thanks also to my girlfriend Danielle, for supporting me through the final stretch.

Finally, thank you to the European Research Council for providing the funding that supported all my work.

Table of Contents

Summary	4
Abbreviations	5
Chapter 1: Introduction	6
Chapter 2: In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons	24
Published manuscript	26
Supplementary information	59
Chapter 3: Effective population size predicts local rates but not local mitigation of read-through errors	89
Published manuscript	91
Supplementary information	111
Chapter 4: Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage	121
Published manuscript	123
Supplementary information	135
Chapter 5: Sequence conservation need not imply purifying selection: evidence from mammalian stop codon usage	140
Unpublished manuscript	142
Supplementary information	197
Chapter 6: (Discussion part 1) Stop codon usage as a window into genome evolution: mutation, selection, biased gene conversion and the TAG paradox	211
Unpublished manuscript	213
Chapter 7: Discussion part 2	245
Appendix 1: Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design	270
Appendix 2: Causes and consequences of purifying selection on SARS-CoV-2	289

Summary

The canonical view of protein evolution is one where selection acts upon the final product of gene expression. As gene expression is extremely prone to error, and these errors are deleterious, it is however becoming increasingly apparent that selection acts far earlier to oppose erroneous protein synthesis. Genomes may evolve to become more error-proof by preventing their occurrence or by mitigating their impacts. In this thesis I seek to resolve several questions regarding error control using translational read-through (TR) as an exemplar. TR occurs when the stop codon of an mRNA transcript is missed by the termination machinery during translation, leading to the continuation of translation into the 3' UTR and potential generation of C-terminally extended proteins. TR prevention and mitigation can be achieved by selection for stop codons. TR rate can be reduced by using the least error-prone stop codon, TAA, to terminate translation while TR may theoretically be mitigated by 3' in-frame additional stop codons (ASCs) which could act as a fail-safe mechanism. In Chapter 2, I ask how often we see evidence for error mitigation strategies in response to TR. I present evidence for ASC enrichment in some, not all, unicellular eukaryotes but no such evidence in multicellular species or bacteria. I note that the strength of selection for TR mitigation should depend on the TR rate, thus in Chapter 3 I investigate how ASCs and TAA stop codons co-evolve, asking whether there is a preferred evolutionary route for error handling. I observe that TAA enrichment significantly correlates with effective population size (N_e), while ASC enrichment does not. As nearly neutral theory predicts that selection is most efficient in species with high N_e , from these results I infer that error prevention might be optimal. If TAA is positively selected to prevent TR, how then might we explain variation in the usage of the non-optimal stops, TGA and TAG? In Chapter 4, I re-examine the long-standing hypothesis that bacterial stop codon usage adapts to the cellular abundance of RF1 and RF2 release factors, finding evidence to the contrary. I note also that TGA is enigmatically highly abundant and highly conserved in mammals despite its high intrinsic TR rate. This, however, I find in Chapter 5 to be better explained by the action of GC-biased gene conversion than selection for TGA stop codons or mutation bias. All the above I frame within the wider literature in a review article presented as “discussion part 1” in Chapter 6. During the pandemic I also contributed to two papers concerning the evolution of SARS-CoV-2, which are presented as appendices.

Abbreviations

A	Adenine
ASC	Additional stop codon
C	Cytosine
EMBL	European Molecular Biology Laboratory
G	Guanine
gBGC	GC-biased gene conversion
mRNA	Messenger RNA
N_e	Effective population size
NCBI	National Centre for Biotechnology Information
ncRNA	Non-coding RNA
NMD	Nonsense-mediated decay
NSD	Non-stop decay
ORF	Open reading frame
OSC	Out-of-frame stop codon
PTC	Premature termination codon
RNA	Ribonucleic acid
RF	Release factor
T	Thymine
TAA	Ochre stop codon
TAG	Amber stop codon
TGA	Opal / umber stop codon
TR	Translational read-through
tRNA	Transfer RNA
U	Uracil
UTR	Untranslated region

Chapter 1

Introduction

Gene expression is inherently prone to error

The canonical view of protein evolution is one where selection acts upon the final product of gene expression. How well does a protein do its job within the cell? How well does an enzyme's active site fit to its substrate? Under this framework, one expects selection to merely discriminate between what might be described as "good" protein and "bad" protein. It is, however, becoming increasingly apparent that the multi-step process of gene expression is extremely prone to errors (Drummond and Wilke 2009). A newer view considers that selection acts much earlier in a protein's lifecycle to oppose erroneous protein synthesis (Warnecke and Hurst 2011).

At all stages from transcription to translation to post-translational modifications, molecular errors are ubiquitous (Drummond and Wilke 2009). These molecular errors do not refer to mutations or any other heritable error that might influence the fitness of the next generation, but to non-heritable "phenotypic" errors that impact only the fitness of the cell where gene expression is taking place. During transcription, such errors include nucleotide misincorporations by RNA polymerase that may lead to reduced functionality and expression of the resultant protein (Parker 1989; Ogle and Ramakrishnan 2005; Drummond and Wilke 2009; Wong, et al. 2018). During splicing, phenotypic errors include intron retention that leads to the inclusion of uneconomical sequence in mature mRNA (Wilhelm, et al. 2008; Drummond and Wilke 2009). During translation, phenotypic errors include accidental ribosomal slippage, where the ribosome skips one or more nucleotides and resumes translation in an incorrect reading frame (Drummond and Wilke 2009; Seligmann 2019), amino acid misincorporations (sometimes called mistranslations) (Parker 1989; Ogle and Ramakrishnan 2005), and translational read-through (TR), where translation fails to terminate at the stop codon of an mRNA transcript and thus continues into the 3' UTR (Doronina and Brown 2006; Namy and Rousset 2010; Rodnina 2018).

The rate of any given phenotypic error varies according to the molecular machinery required at that stage of gene expression. Estimated from the number of mistakes present in end-product RNAs, overall transcription fidelity has an error rate of approximately 1 in 1,000 to 1 in 100,000 across prokaryotes and eukaryotes (James, et al. 2017), with misincorporations introduced by mistakes RNA polymerase but

often corrected by proof-reading machinery. In humans, from quantitative PCR against exon-exon boundaries, it is estimated that erroneous intron retention occurs at a rate of $\sim 5.7 \times 10^{-6}$ to 2.3×10^{-2} per correctly spliced intron (Fox-Walsh and Hertel 2009). Across the tree of life, erroneous ribosomal slippage occurs approximately once per 10,000 to 100,000 codons (Parker 1989; Kurland 1992), thought to be promoted by slippery sequences that can take the form of A|AAB|BBC where AAA and BBB are identical nucleotides and C is any other nucleotide (Brierley, et al. 1992; Licznar, et al. 2003; Napthine, et al. 2003). Phenotypic rates such as these are typically orders of magnitude higher than the spontaneous mutation rate. Direct sequencing in humans predicts a mutation rate of $\sim 1.1-1.4 \times 10^{-8}$ per bp per generation depending on the genomic region (Lynch 2010; Kong, et al. 2012; Milholland, et al. 2017; Rodriguez-Galindo, et al. 2020), while bacterial mutation rates typically range from $\sim 1 \times 10^{-7}$ to $\sim 1 \times 10^{-9}$ (Westra, et al. 2017; Chevallereau, et al. 2019), but have been estimated as low as 2.3×10^{-11} per bp per generation (Drake 1991). In certain clinical and natural environments bacterial mutations can occur at ~ 100 times greater frequency (LeClerc, et al. 1996; Matic, et al. 1997; Oliver, et al. 2000) but even these rates are dwarfed by those of phenotypic errors.

Are phenotypic errors truly deleterious?

Whether selection should oppose phenotypic errors or not depends upon their fitness effects. Theoretically, phenotypic errors need not be deleterious. Just as mutations introduce the genetic novelty for selection to act upon, phenotypic errors could also place new coding sequences and protein domains in the view of selection (Masel 2006; Whitehead, et al. 2008). Nucleotide misincorporations might mirror *de novo* mutations by creating a mRNA transcript containing a single base change. If retained introns do not contain a premature termination codon (PTC) (Ge and Porse 2014), they may activate or repress translation initiation if located in the 5' UTR by introducing upstream open reading frames (ORFs) (Tahmasebi, et al. 2016), introduce cis-elements that affect mRNA stability or translational efficiency if located in the 3' UTR (Sun, et al. 2010), or contribute amino acids to the protein product if located in coding sequence without disrupting the reading frame (Jacob and Smith 2017). Perhaps the largest impact on coding information can result from ribosomal slippage, where reading frame disruption may have profound effects on amino acid composition, not

least when the frameshift occurs at the 5' end (Ketteler 2012). Indeed, programmed frameshifting is commonly utilised in viruses (Maia, et al. 1996; Brierley and Dos Ramos 2006; Dulude, et al. 2006; Ketteler 2012), and less commonly in bacteria (Craigie and Caskey 1986; Gupta, et al. 2013) and eukaryotes (Clark, et al. 2007; Baranov, et al. 2011), to maximise the coding potential of their genomes. One could argue that phenotypic errors provide even more selective utility than *de novo* mutations given much more “novelty” may be tested at one time.

Whilst the results of phenotypic error can sometimes be functional, we can be confident that most are deleterious. One clear indicator of this is that quality control pathways to ensure successful gene expression are abundant across the tree of life. Both bacteria and eukaryotes, for example, have evolved pathways to degrade transcripts containing PTCs (Belasco 2010). Eukaryotes selectively degrade PTC-containing mRNAs via nonsense mediated decay (NMD) after recognising irregular exon junctions downstream of the stop codon (Maquat and Carmichael 2001; Belasco 2010; Lykke-Andersen and Jensen 2015). Bacterial genes do not contain introns and thus typical NMD is not possible, but their PTC-containing transcripts are nevertheless rapidly degraded (Nilsson, et al. 1987), possibly via internal cleavage by RNase E (Arnold, et al. 1998; Baker and Mackie 2003). To protect against translational read-through, both groups also contain non-stop decay (NSD) pathways that detect transcripts lacking a stop codon and degrade them using 3' endonucleases (van Hoof, et al. 2002; Richards, et al. 2006). At the protein level, bacteria and eukaryotes possess a wide repertoire of chaperone proteins that function to ensure nascent proteins fold correctly, are protected from heat shock, and do not aggregate (see Saibil 2013 for review). What would be the need for such diverse quality control if phenotypic errors were not deleterious?

More evidence of evolution to minimise phenotypic errors can be found in molecular evolution studies, many of which come from the lab of Jianzhi Zhang. The hypothesis that phenotypic errors are deleterious predicts that errors should be minimised in highly expressed genes, the fitness costs of erroneous gene expression being a product of the error rate multiplied by the expression level. Zhang hence predicts that the sequence evolution of highly expressed genes is constrained to minimise errors, observing for example that transcriptional start site diversity (Xu, et al. 2019),

polyadenylation diversity (Xu and Zhang 2018), and the presence of translational read-through motifs (Li and Zhang 2019) all covary negatively with expression. Non-optimal transcription initiations, polyadenylations, and translation terminations are thus most-often deleterious and purged by purifying selection. Indeed, Zhang and colleagues find several more similar examples (Xu and Zhang 2014; Liu and Zhang 2018a, b; Jiang and Zhang 2019; Xu and Zhang 2020, 2021).

Error handling at the sequence level is not only controlled by purifying selection, but by positive selection for error proofing motifs. Mistranslations may be prevented by selection for optimal codon usage, which improves translational accuracy and lowers the rate of amino acid misincorporations (Akashi 1994; Powell and Moriyama 1997; Stoletzki and Eyre-Walker 2007; Punde, et al. 2019). To mitigate the impacts of ribosomal slippage, out-of-frame stop codons (OSCs) are under selection downstream of “slippery” codons to terminate translation should the ribosome find itself translating sequence in the +1-reading frame (Seligmann and Pollock 2004; Abrahams and Hurst 2018; Seligmann 2019). It is probably for this error proofing purpose that adenine is enriched at the +4-nucleotide site at the start of bacterial genes, ATGA motifs containing the ATG start codon with a TGA stop codon hidden in the +1 reading frame for the immediate cessation of erroneous protein synthesis should it be necessary (Abrahams and Hurst 2017). In-frame additional stop codons (ASCs) may also be under selection in eukaryotic (Liang, et al. 2005; Adachi and Cavalcanti 2009; Fleming and Cavalcanti 2019), but possibly not bacterial (Major, et al. 2002; Korkmaz, et al. 2014), 3' UTR sequences to act as a fail-safe mechanism that protects against translational read-through by providing a second opportunity for translation to terminate.

Error proofing selection during molecular evolution works in tandem with cellular quality control to minimise the consequences of gene expression error. For example, in genomic locations that are invisible to NMD, such as the coding sequence of intronless genes in mammals (Zhang, et al. 1998; Maquat and Li 2001; Brocke, et al. 2002) and the final exon of any coding sequence (Le Hir, et al. 2001; Lindeboom, et al. 2016), codons that are only one point mutation away from a PTC are avoided to prevent the need for NMD in the first place (Cusack, et al. 2011). For translational read-through errors, while NSD pathways are present in bacteria and eukaryotes (van

Hoof, et al. 2002; Richards, et al. 2006), termination motifs that minimise TR rate are consistently enriched in highly expressed genes to prevent the need for degradation (Korkmaz, et al. 2014; Trotta 2016; Wei and Xia 2017; Cridge, et al. 2018).

Error prevention and mitigation may be achieved locally or globally

While supporting the argument that phenotypic errors are deleterious, the above evidence also demonstrates that error control can be achieved at two different scales. Mutations that reduce cellular error rates, by improving proofreading or degradation pathways for example, can be considered “global” solutions because they affect every qualifying gene or transcript in the genome. Mutations that have less far-reaching effects, such as those producing OSCs to protect against frameshifts or PTCs to mitigate the consequences of intron retention, are by contrast “local” solutions because they operate only at a single locus. As originally theorised and modelled by Rajon and Masel (2011), local solutions hence require repeated bouts of selection and fixation to produce genome-wide effects while global solutions need to evolve only once.

Rajon and Masel (2011) argue that selection against error forms a positive feedback loop, such that only global or only local solutions evolve in any given population. The differences in fitness effects between global and local solutions hence makes several predictions with what to expect in species with different effective population size (N_e) when evoking nearly neutral theory (Ohta 1992). They hypothesize that local solutions evolve readily in species with large N_e as selection is efficient and deleterious sequences resulting from error are purged. By contrast, low N_e species that are more sensitive to drift accumulate deleterious sequences as they cannot easily fix local solutions, hence selection strongly favours global error solutions for a genome-wide remedy. Intermediate population sizes, they find in their simulations, are bistable and either global or local solutions might result (Rajon and Masel 2011).

There is, however, a limitation to the Rajon and Masel (2011) models as they fail to consider that both global and local solutions may take the form of error prevention or error mitigation. Error prevention refers to any adaptation that reduced the rate at which phenotypic errors occur, while error mitigation refers to adaptations that reduce the impact of such errors. Within the exemplar of translational read-through, Rajon

and Masel (2011) assume global solutions to reduce error rate (without naming a candidate pathway to facilitate this) and local solutions to provide mitigation (via selection for ASCs). Their models therefore fail to account for global error mitigation and local error rate reduction despite both being known TR error control mechanisms. Global error mitigation against TR, for example, occurs via NSD pathways that degrade transcripts lacking a stop (van Hoof, et al. 2002; Richards, et al. 2006) and also via mRNA surveillance systems that repress the expression of nonstop mRNAs and destabilise their translated products (Ito-Harashima, et al. 2007). TR error rate reduction can be facilitated locally by changing the stop codon used by a transcript, TAA being the most reliable stop codon followed by TAG and TGA in both eukaryotes and bacteria (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen, et al. 1993; Meng, et al. 1995; Sanchez, et al. 1998; Tate, et al. 1999; Wei, et al. 2016; Cridge, et al. 2018). Error handling should thus be described not simply by global versus local solutions, but by a 2x2 grid of global/local and rate/mitigation solutions. As the strength of selection acting upon each of the four axes is dependent upon the others, how they might co-evolve is an open question.

Questions

In this thesis, using the case example of TR, I aim to answer several questions of how genomes evolve to control phenotypic errors.

How often do we see evidence of local error mitigation? In Chapter 2, I ask how often we see evidence of such strategies against TR errors. In doing so I examine the hypothesis that ASCs are under selection to act as a “fail-safe” mechanism in 3’ UTR sequences to provide a second opportunity for translation to terminate should this not occur at the canonical stop site. Indeed, this hypothesis has been the topic of some debate, with evidence of ASC enrichment beyond nucleotide expectations being observed in yeast (Liang, et al. 2005) and ciliates (Adachi and Cavalcanti 2009; Fleming and Cavalcanti 2019) but not in bacteria (Major, et al. 2002; Korkmaz, et al. 2014). Past studies have, however, focused on relatively few species and hence there is a gap for a systematic multi-species analysis that I here hope to fill.

How do local error mitigation and local error prevention strategies co-evolve? Noting that the strength of any selection acting upon ASCs should depend on the stop codon used for termination, the three stops having distinct TR rates, in Chapter 3 I ask how TAA stop codon usage and ASCs co-evolve. To achieve this I follow Rajon and Masel (2011) in considering both error control strategies against the predictions of nearly neutral theory with regards to N_e . The hypothesis that local solutions are the preferred route for species with large N_e predicts both TAA and ASC enrichment to correlate positively with N_e . If one, and not the other, were to correlate with N_e this could be inferred as a selective preference for either prevention or mitigation.

How might we explain between-species variation in non-optimal stop codon usage? As all the available experimental data for TR rates point towards TAA optimality, the simplest null expectation is that TAA should be the most abundant stop codon in all genomes, especially in bacteria due to their large N_e . In Chapter 4, I consider a long-standing hypothesis that the between-species stop codon usage trends of bacteria can be explained by variation in the relative abundance of class I release factors (RFs) (Sharp and Bulmer 1988; Wei, et al. 2016), RF1 binding TAA and TAG and RF2 binding TAA and TGA (Rodnina 2018). To assess this hypothesis it is useful to note that eukaryotes and archaea contain only one RF, eRF1 (Inagaki and Doolittle 2000; Jackson, et al. 2012) and aRF1 (Kobayashi, et al. 2012) respectively, that recognises all three stop codons. If the stop codon usage trends of bacteria match those observed in eukaryotes and archaea, this would suggest that RF biology is not needed to explain such trends and hence that other hypotheses are needed to explain imperfect stop codon usage.

How can we explain the high abundance and conservation of TGA in mammals? Related to the previous question is how to explain the unusual high abundance and conservation of non-optimal TGA stop codons observed in mammalian genomes (Belinky, et al. 2018; Seoighe, et al. 2020). In Chapter 5, I consider three hypotheses that could explain this phenomenon: mutation bias, selection for TGA, and GC-biased gene conversion (gBGC). To differentiate between these, I analyse both stop codon usage and follow the Belinky, et al. (2018) methodology to analyse stop codon flux (the rate at which one stop codon changes to another, per incidence of the ancestral stop codon). Mutation bias can be easily assessed by inferring a mononucleotide or

dinucleotide mutational matrix from *de novo* mutations. Selection for TGA may be assayed by regression analysis that predicts stop codon usage as a function of gene expression, under the assumption that highly expressed genes are under the greatest selection pressure for error control. The gBGC hypothesis may be investigated by assessing the relationship between stop codon usage and recombination rate, gBGC being tightly linked to the mismatch repair process (Brown and Jiricny 1988, 1989) during homologous recombination (Mugal, et al. 2015).

I summarise all the above in a commissioned review article focused on stop codon usage, Chapter 6. This I present as “Discussion part 1” as this manuscript places the results of the previous chapters within the wider literature. Following “Discussion part 2” where I discuss other unanswered questions and the methodological limitations of my work, I present two published papers concerning the molecular evolution of SARS-CoV-2 as appendices. These are slightly removed from the story of my thesis but nonetheless represent work undertaken during my PhD.

References

Abrahams L, Hurst LD. 2017. Adenine enrichment at the fourth CDS residue in bacterial genes is consistent with error proofing for +1 frameshifts. *Mol. Biol. Evol.* 34(12): 3064-3080.

Abrahams L, Hurst LD. 2018. Refining the ambush hypothesis: Evidence that GC- and AT-rich bacteria employ different frameshift defence strategies. *Genome Biol. Evol.* 10(4): 1153-1173.

Adachi M, Cavalcanti AR. 2009. Tandem stop codons in ciliates that reassign stop codons. *J. Mol. Evol.* 68(4): 424-431.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics.* 136(3): 927-935.

Arnold TE, Yu J, Belasco JG. 1998. mRNA stabilization by the ompA 5' untranslated region: Two protective elements hinder distinct pathways for mRNA degradation. *RNA.* 4(3): 319-330.

- Baker KE, Mackie GA. 2003. Ectopic RNase E sites promote bypass of 5'-end-dependent mRNA decay in *Escherichia coli*. *Mol. Microbiol.* 47(1): 75-88.
- Baranov PV, Wills NM, Barriscale KA, Firth AE, Jud MC, Letsou A, Manning G, Atkins JF. 2011. Programmed ribosomal frameshifting in the expression of the regulator of intestinal stem cell proliferation, adenomatous polyposis coli (APC). *RNA Biol.* 8(4): 637-647.
- Belasco JG. 2010. All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nat. Rev. Mol. Cell Biol.* 11(7): 467-478.
- Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci. Rep.* 8(1): 9260.
- Brierley I, Dos Ramos FJ. 2006. Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res.* 119(1): 29-42.
- Brierley I, Jenner AJ, Inglis SC. 1992. Mutational analysis of the slippery-sequence component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.* 227(2): 463-479.
- Brocke KS, Neu-Yilik G, Gehring NH, Hentze MW, Kulozik AE. 2002. The human intronless melanocortin 4-receptor gene is NMD insensitive. *Hum. Mol. Genet.* 11(3): 331-335.
- Brown TC, Jiricny J. 1988. Different base base mispairs are corrected with different efficiencies and specificities in monkey kidney-cells. *Cell.* 54(5): 705-711.
- Brown TC, Jiricny J. 1989. Repair of base base mismatches in simian and human-cells. *Genome.* 31(2): 578-583.
- Chevallereau A, Meaden S, van Houte S, Westra ER, Rollie C. 2019. The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity. *Philos. Trans. R. Soc. B.* 374(1772).
- Clark MB, Jaenicke M, Gottesbuehren U, Kleffmann T, Legge M, Poole ES, Tate WP. 2007. Mammalian gene PEG10 expresses two reading frames by high efficiency-1 frameshifting in embryonic-associated tissues. *J. Biol. Chem.* 282(52): 37359-37369.

Craig WJ, Caskey CT. 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*. 322(6076): 273-275.

Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res*. 46(4): 1927-1944.

Cusack BP, Arndt PF, Duret L, Crollius HR. 2011. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet*. 7(10): e1002276.

Doronina VA, Brown JD. 2006. When nonsense makes sense and vice versa: Non-canonical decoding events at stop codons in eukaryotes. *Mol. Biol*. 40(4): 731-741.

Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA*. 88(16): 7160-7164.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet*. 10(10): 715-724.

Dulude D, Berchiche YA, Gendron K, Brakier-Gingras L, Heveker N. 2006. Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. *Virology*. 345(1): 127-136.

Fleming I, Cavalcanti ARO. 2019. Selection for tandem stop codons in ciliate species with reassigned stop codons. *PLoS One*. 14(11): e0225804.

Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl. Acad. Sci. U.S.A.* 106(6): 1766-1771.

Ge Y, Porse BT. 2014. The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. *Bioessays*. 36(3): 236-243.

Geller AI, Rich A. 1980. A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature*. 283(5742): 41-46.

- Gupta P, Kannan K, Mankin AS, Vazquez-Laslop N. 2013. Regulation of gene expression by macrolide-induced ribosomal frameshifting. *Mol. Cell.* 52(5): 629-642.
- Inagaki Y, Doolittle WF. 2000. Evolution of the eukaryotic translation termination system: Origins of release factors. *Mol. Biol. Evol.* 17(6): 882-889.
- Ito-Harashima S, Kuroha K, Tatematsu T, Inada T. 2007. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* 21(5): 519-524.
- Jackson RJ, Hellen CUT, Pestova TV. 2012. Termination and post-termination events in eukaryotic translation. In: Marintchev A, editor. *Advances in Protein Chemistry and Structural Biology, Vol 86: Fidelity and Quality Control in Gene Expression.* p. 45-93.
- Jacob AG, Smith CWJ. 2017. Intron retention as a component of regulated gene expression programs. *Hum. Genet.* 136(9): 1043-1057.
- James K, Gamba P, Cockell SJ, Zenkin N. 2017. Misincorporation by RNA polymerase is a major source of transcription pausing in vivo. *Nucleic Acids Res.* 45(3): 1105-1113.
- Jiang D, Zhang J. 2019. The preponderance of nonsynonymous A-to-I RNA editing in coleoids is nonadaptive. *Nat. Commun.* 10(1): 5411.
- Jorgensen F, Adamski FM, Tate WP, Kurland CG. 1993. Release factor-dependent false stops are infrequent in *Escherichia coli*. *J. Mol. Biol.* 230(1): 41-50.
- Ketteler R. 2012. On programmed ribosomal frameshifting: the alternative proteomes. *Front. Genet.* 3(1): 242.
- Kobayashi K, Saito K, Ishitani R, Ito K, Nureki O. 2012. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1 alpha complex. *Nucleic Acids Res.* 40(18): 9319-9328.
- Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature.* 488(7412): 471-475.

- Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* 289(44): 30334-30342.
- Kurland CG. 1992. Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.* 26(1): 29-50.
- Le Hir H, Gatfield D, Izaurralde E, Moore MJ. 2001. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.* 20(17): 4987-4997.
- LeClerc JE, Li BG, Payne WL, Cebula TA. 1996. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science.* 274(5290): 1208-1211.
- Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15(5): e1008141.
- Liang H, Cavalcanti AR, Landweber LF. 2005. Conservation of tandem stop codons in yeasts. *Genome Biol.* 6(4): R31.
- Liczner P, Mejlhede N, Prere MF, Wills N, Gesteland RF, Atkins JF, Fayet O. 2003. Programmed translational-1 frameshifting on hexanucleotide motifs and the wobble properties of tRNAs. *EMBO J.* 22(18): 4770-4778.
- Lindeboom RGH, Supek F, Lehner B. 2016. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* 48(10): 1112-1118.
- Liu Z, Zhang J. 2018a. Human C-to-U coding RNA editing is largely nonadaptive. *Mol. Biol. Evol.* 35(4): 963-969.
- Liu Z, Zhang J. 2018b. Most m(6)A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Mol. Biol. Evol.* 35(3): 666-675.
- Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* 16(11): 665-677.
- Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences.* 107(3): 961-968.

- Maia IG, Seron K, Haenni AL, Bernardi F. 1996. Gene expression from viral RNA genomes. *Plant Mol. Biol.* 32(1-2): 367-391.
- Major LL, Edgar TD, Yee Yip P, Isaksson LA, Tate WP. 2002. Tandem termination signals: myth or reality? *FEBS Lett.* 514(1): 84-89.
- Maquat LE, Carmichael GG. 2001. Quality control of mRNA function. *Cell.* 104(2): 173-176.
- Maquat LE, Li XJ. 2001. Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. *RNA.* 7(3): 445-456.
- Masel J. 2006. Cryptic genetic variation is enriched for potential adaptations. *Genetics.* 172(3): 1985-1991.
- Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, Denamur E, Elion J. 1997. Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science.* 277(5333): 1833-1834.
- Meng SY, Hui JO, Haniu M, Tsai LB. 1995. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 211(1): 40-48.
- Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. 2017. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* 8(1): 15183.
- Mugal CF, Weber CC, Ellegren H. 2015. GC-biased gene conversion links the recombination landscape and demography to genomic base composition GC-biased gene conversion drives genomic base composition across a wide range of species. *Bioessays.* 37(12): 1317-1326.
- Namy O, Rousset JP. 2010. Specification of standard amino acids by stop codons. In: Atkins JF, Gesteland RF, editors. *Recoding: Expansion of Decoding Rules Enriches Gene Expression.* p. 79-100.

- Napthine S, Vidakovic M, Ginary R, Namy O, Brierley I. 2003. Prokaryotic-style frameshifting in a plant translation system: conservation of an unusual single-tRNA slippage event. *EMBO J.* 22(15): 3941-3950.
- Nilsson G, Belasco JG, Cohen SN, Vongabain A. 1987. Effect of premature termination of translation on mRNA stability depends on the site of ribosome release. *Proc. Natl. Acad. Sci. U.S.A.* 84(14): 4890-4894.
- Ogle JM, Ramakrishnan V. 2005. Structural insights into translational fidelity. *Annu. Rev. Biochem.* 74(1): 129-177.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. System.* 23(1): 263-286.
- Oliver A, Canton R, Campo P, Baquero F, Blazquez J. 2000. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science.* 288(5469): 1251-1253.
- Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev.* 53(3): 273-298.
- Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.* 94(15): 7784-7790.
- Punde N, Kookan J, Leary D, Legler PM, Angov E. 2019. Codon harmonization reduces amino acid misincorporation in bacterially expressed *P. falciparum* proteins and improves their immunogenicity. *AMB Express.* 9(1): 167.
- Rajon E, Masel J. 2011. Evolution of molecular error rates and the consequences for evolvability. *Proc. Natl. Acad. Sci. U.S.A.* 108(3): 1082-1087.
- Richards J, Mehta P, Karzai AW. 2006. RNase R degrades non-stop mRNAs selectively in an SmpB-tmRNA-dependent manner. *Mol. Microbiol.* 62(6): 1700-1712.
- Rodnina MV. 2018. Translation in prokaryotes. *Cold Spring Harb. Perspect. Biol.* 10(9): a032664.

- Rodriguez-Galindo M, Casillas S, Weghorn D, Barbadilla A. 2020. Germline de novo mutation rates on exons versus introns in humans. *Nat. Commun.* 11(1): 3304.
- Saibil H. 2013. Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.* 14(10): 630-642.
- Sanchez JC, Padron G, Santana H, Herrera L. 1998. Elimination of an HuIFN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J. Biotechnol.* 63(3): 179-186.
- Seligmann H. 2019. Localized context-dependent effects of the “ambush” hypothesis: More off-frame stop codons downstream of shifty codons. *DNA Cell Biol.* 38(8): 786-795.
- Seligmann H, Pollock DD. 2004. The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23(10): 701-705.
- Seoighe C, Kiniry SJ, Peters A, Baranov PV, Yang H. 2020. Selection shapes synonymous stop codon use in mammals. *J. Mol. Evol.* 88(7): 549-561.
- Sharp PM, Bulmer M. 1988. Selective differences among translation termination codons. *Gene.* 63(1): 141-145.
- Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. *Mol. Biol. Evol.* 24(2): 374-381.
- Strigini P, Brickman E. 1973. Analysis of specific misreading in *Escherichia coli*. *J. Mol. Biol.* 75(4): 659-672.
- Sun S, Zhang Z, Sinha R, Karni R, Krainer AR. 2010. SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat. Struct. Mol. Biol.* 17(3): 306-312.
- Tahmasebi S, Jafarnejad SM, Tam IS, Gonatopoulos-Pournatzis T, Matta-Camacho E, Tsukumo Y, Yanagiya A, Li W, Atlasi Y, Caron M, et al. 2016. Control of embryonic stem cell self-renewal and differentiation via coordinated alternative splicing and translation of YY2. *Proc. Natl. Acad. Sci. U.S.A.* 113(44): 12360-12367.

Tate WP, Mansell JB, Mannering SA, Irvine JH, Major LL, Wilson DN. 1999. UGA: a dual signal for 'stop' and for recoding in protein synthesis. *Biochemistry (Mosc)*. 64(12): 1342-1353.

Trotta E. 2016. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics*. 17(17): 366.

van Hoof A, Frischmeyer PA, Dietz HC, Parker R. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science*. 295(5563): 2262-2264.

Warnecke T, Hurst LD. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat. Rev. Genet*. 12(12): 875-881.

Wei Y, Xia X. 2017. The role of +4U as an extended translation termination signal in bacteria. *Genetics*. 205(2): 539-549.

Wei YL, Wang J, Xia XH. 2016. Coevolution between stop codon usage and release factors in bacterial species. *Mol. Biol. Evol*. 33(9): 2357-2367.

Westra ER, Sunderhauf D, Landsberger M, Buckling A. 2017. Mechanisms and consequences of diversity-generating immune strategies. *Nat. Rev. Immunol*. 17(11): 719-728.

Whitehead DJ, Wilke CO, Vernazobres D, Bornberg-Bauer E. 2008. The look-ahead effect of phenotypic mutations. *Biol. Direct*. 3(1): 18.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 453(7199): 1239-1243.

Wong HE, Huang C, Jr., Zhang Z. 2018. Amino acid misincorporation in recombinant proteins. *Biotechnol. Adv*. 36(1): 168-181.

Xu C, Park JK, Zhang JZ. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biol*. 17(3): e3000197.

Xu C, Zhang J. 2020. Mammalian alternative translation initiation is mostly nonadaptive. *Mol. Biol. Evol.* 37(7): 2015-2028.

Xu C, Zhang J. 2021. Mammalian circular RNAs result largely from splicing errors. *Cell Rep.* 36(4): 109439.

Xu C, Zhang JZ. 2018. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. *Cell Sys.* 6(6): 734-742.

Xu G, Zhang J. 2014. Human coding RNA editing is generally nonadaptive. *Proc. Natl. Acad. Sci. U.S.A.* 111(10): 3769-3774.

Zhang J, Sun XL, Qian YM, LaDuca JP, Maquat LE. 1998. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol. Cell. Biol.* 18(9): 5272-5283.

Chapter 2

In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons

Alexander T. Ho and Laurence D. Hurst

PLoS Genetics, 15(9): e1008386.

This chapter contains work published on 17th September 2019 at PLoS Genetics, the original and sole place of publication. It thus contains analysis of publicly available data using bespoke scripts that are freely available at the locations cited within the paper. The paper is open access and I have permission as the author to include the article in full in this thesis (<https://journals.plos.org/plosgenetics/s/licenses-and-copyright>). The latest version of the published article can be found by following the address: <https://doi.org/10.1371/journal.pgen.1008386>.

Pre-amble

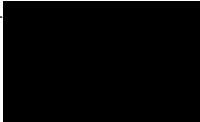
How often do we see evidence of local error mitigation?

In this first published chapter, I ask how often we see evidence of local error mitigation strategies against translational read-through (TR) errors. To achieve this, I examine the hypothesis that in-frame additional stop codons (ASCs) are under selection to act as a “fail-safe” mechanism in 3' UTR sequences to provide a second opportunity for translation to terminate should this not occur at the canonical stop site. While TR rates may be reduced by any given genome by increasing its usage of TAA stop codons, the least error prone stop, TR to some degree is inevitable. The fail-safe hypothesis states that ASCs enables the translation of TR-affected transcripts to terminate only a short distance downstream of the canonical stop codon, minimising the energy wastage from translating the 3' UTR and limiting the size of the resultant C-terminal extensions.

The fail-safe hypothesis is particularly interesting because it receives mixed support from the literature. In eukaryotes, there is strong evidence for ASC enrichment in some yeast and ciliate species but very little in other eukaryotic taxa. In bacteria, bioinformatic studies have failed to identify any genomes enriched beyond null for ASCs. At first look, ASCs appear enriched immediately downstream from the canonical stop codon in bacterial 3' UTRs, but this is confounded by a general preference for thymine as the immediately proximal nucleotide (the so-called +4 base).

To date most studies have focused on very few species and there is hence a gap in the literature for a systematic multi-species analysis. I hope to fill this gap with this chapter, presenting bioinformatic analyses that test a wealth of predictions made by the fail-safe hypothesis in ~650 phylogenetically distinct bacterial genomes and ~70 unicellular eukaryotes. Corroborating prior studies, I find no evidence to support ASC selection in eubacteria and observe ASC enrichment in yeast and ciliates. In addition, I observe evidence of ASC enrichment in several more unicellular eukaryotes for the first time.

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:			
In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons			
Publication status (tick one)			
Draft manuscript	<input type="checkbox"/>	Submitted	<input type="checkbox"/>
In review	<input type="checkbox"/>	Accepted	<input type="checkbox"/>
Published	<input checked="" type="checkbox"/>		
Publication details (reference)	Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. PLoS Genet. 15(9): e1008386.		
Copyright status (tick the appropriate statement)			
I hold the copyright for this material	<input type="checkbox"/>	Copyright is retained by the publisher, but I have been given permission to replicate the material here	<input checked="" type="checkbox"/>
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	<p>The candidate contributed to / considerably contributed to / predominantly executed the...</p> <p>Formulation of ideas: 100%</p> <p>Design of methodology: 100%</p> <p>Bioinformatic analyses: 100%</p> <p>Experimental work: N/a</p> <p>Presentation of data in journal format: 100%</p>		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed		Date	03/03/2022

RESEARCH ARTICLE

In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons

Alexander T. Ho *, Laurence D. Hurst 

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

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



Abstract

Errors throughout gene expression are likely deleterious, hence genomes are under selection to ameliorate their consequences. Additional stop codons (ASCs) are in-frame non-sense 'codons' downstream of the primary stop which may be read by translational machinery should the primary stop have been accidentally read through. Prior evidence in several eukaryotes suggests that ASCs are selected to prevent potentially-deleterious consequences of read-through. We extend this evidence showing that enrichment of ASCs is common but not universal for single cell eukaryotes. By contrast, there is limited evidence as to whether the same is true in other taxa. Here, we provide the first systematic test of the hypothesis that ASCs act as a fail-safe mechanism in eubacteria, a group with high read-through rates. Contra to the predictions of the hypothesis we find: there is paucity, not enrichment, of ASCs downstream; substitutions that degrade stops are more frequent in-frame than out-of-frame in 3' sequence; highly expressed genes are no more likely to have ASCs than lowly expressed genes; usage of the leakiest primary stop (TGA) in highly expressed genes does not predict ASC enrichment even compared to usage of non-leaky stops (TAA) in lowly expressed genes, beyond downstream codon + 1. Any effect at the codon immediately proximal to the primary stop can be accounted for by a preference for a T/U residue immediately following the stop, although if anything, TT- and TC- starting codons are preferred. We conclude that there is no compelling evidence for ASC selection in eubacteria. This presents an unusual case in which the same error could be solved by the same mechanism in eukaryotes and prokaryotes but is not. We discuss two possible explanations: that, owing to the absence of nonsense mediated decay, bacteria may solve read-through via gene truncation and in eukaryotes certain prion states cause raised read-through rates.

OPEN ACCESS

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

Editor: Xavier Didelot, University of Warwick, UNITED KINGDOM

Received: June 11, 2019

Accepted: August 27, 2019

Published: September 17, 2019

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pgen.1008386>

Copyright: © 2019 Ho, Hurst. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Scripts can be found at <https://github.com/ath32/ASCs>. All other relevant data are within the manuscript and its Supporting Information files.

Author summary

In all organisms, gene expression is error-prone. One such error, translational read-through, occurs where the primary stop codon of an expressed gene is missed by the translational machinery. Failure to terminate is likely to be costly, hence genomes are

Funding: This work was supported by the European Research Council (Grant EvoGenMed ERC-2014-ADG 669207 to L.D.H). For more information regarding ERC activities, please visit <https://erc.europa.eu/>. 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.

under selection to prevent this from happening. One proposed error-proofing strategy involves in-frame proximal additional stop codons (ASCs) which may act as a 'fail-safe' mechanism by providing another opportunity for translation to terminate. There is evidence for ASC enrichment in several eukaryotes. We extend this evidence showing it to be common but not universal in single celled eukaryotes. However, the situation in bacteria is poorly understood, despite bacteria having high read-through rates. Here, we test the fail-safe hypothesis within a broad range of bacteria. To our surprise, we find that not only are ASCs not enriched, but they may even be selected against. This provides evidence for an unusual circumstance where eukaryotes and prokaryotes could solve the same problem the same way but don't. What are we to make of this? We suggest that if read-through is the problem, ASCs are not necessarily the expected solution. Owing to the absence of nonsense-mediated decay, a process that makes gene truncation in eukaryotes less viable, we propose bacteria may rescue a leaky stop by mutation that creates a new stop upstream. Alternatively, raised read-through rates in some particular conditions in eukaryotes might explain the difference.

Introduction

Errors throughout transcription, translation, and post-translational modification can, and do, happen all the time [1–5]. Whilst an invaluable source of novelty that drives evolution [6], the majority of these errors are likely deleterious [1–3, 6–8]. Genomes may therefore be under selection to mitigate their consequences. This has been supported by bioinformatic studies of stop codon usage in gene locations other than that of the canonical stop. For example, it has been suggested that adenine enrichment at the fourth coding sequence residue in bacterial genes may promote translation termination following a frameshift event at the initiating ATG that allows an out-of-frame stop codon to be read [9, 10]. In 5' leading regions, in-frame stop codons are enriched and postulated to rapidly terminate premature translations [11] (i.e. those that occur before the ribosome reaches the recognised start codon of the mRNA). Selection on the primary stop codon is also thought to be error related [12–14]. Experimental evidence from bacterial studies suggest the three stops differ in their read-through rates [14–21]. Notably the least leaky of the three, TAA, is the preferred codon, especially in the most highly expressed genes [13]. In this study, we consider the hypothesis that additional stop codons (ASCs) occur after the primary stop codon as a fail-safe mechanism to minimise the costs of stop codon readthrough [22]. This question is important both as a means to address the importance of error-proofing to genome evolution but potentially also for optimal transgene design.

Although ribosomes normally terminate translation at stop codons there is a chance that an ectopic amino acid is inserted, allowing translation to continue in the same frame for the generation of extended polypeptides [23, 24]. The primary cause appears to be aberrant recognition by near-cognate tRNAs [25, 26] or other tRNA species [27]. While read-through rates vary depending both on the stop codon and local sequence context [28], read-through rates are typically orders of magnitude higher than the mutation rate [29–33], rendering read-through a potentially significant fitness-modifying trait. While there may be beneficial consequences, such as increased proteome diversity [34], the best evidence suggests that it is largely non-adaptive [8]. Selection for the least leaky stop in highly expressed genes [13] provides strong support for the notion that selection acts to reduce read-through rates as it is most commonly a deleterious error. Possible costs include energetic wastage owing to unnecessary

translation [35] and creation of potentially toxic or sticky novel peptides. Resource wastage can be acute if the ribosome needs to be recovered, as can happen, for example, if it moves into a polyA tail as both RNA and protein can be targeted for destruction [36–38]. In theory, the presence of ASCs downstream may alleviate some of these costs by reducing the amount of additional amino acids added to erroneous polypeptide chains [39] and preventing polyA associated destruction. We herein refer to such a system as the 'fail-safe' hypothesis.

The fail-safe hypothesis has been most thoroughly examined in eukaryotes, notably in yeast [39], and two ciliate species which have reassigned their genetic code such that TGA is the only stop codon [40]. In yeasts, a statistical excess of UAA at the third codon downstream of TAA-terminating genes points towards a maintenance of ASCs by selection in a manner dependent on expression level [39]. This was corroborated in ciliates, where ASCs appear downstream of the primary stop more often than expected by chance given the base composition of 3' regions [40]. Given that the excess is larger in ciliates than in yeast, it was proposed that ASCs are under variable selection intensity dependent on readthrough rate, which in turn may vary between species [40]. This, however, remains *post hoc* speculation.

In bacteria tests of the fail-safe hypothesis are lacking. One study found tandem ASCs (those which immediately follow the primary stop) are over-represented, being seen in 7% of *E. coli* genes [41]. However, the experimentally estimated termination efficiency of tandem stops were below the expected rate and it was postulated that *prima facie* over-representation in the genome could be attributed instead to the preference for a tetranucleotide containing +4U, thought to improve the termination efficiency of the primary stop [41–45]. +4U in this context refers to the base immediately after the primary stop. A +4U base biases the first codon after the primary stop towards a second stop codon as all stops start T/U.

More recently, one study has widened the investigation to ASCs in the following 5 in-frame codon positions. Such ASCs are reported in 8% of *E. coli* genes [13], however, although this figure concurs with the findings of Major and colleagues [41], the authors do not comment on whether this is higher, lower or the same as expected given more codon positions are being considered. More generally, it is unknown whether ASC frequency downstream is higher than expected under a GC-controlled null in any eubacteria. Preliminary data weakly argue against the fail-safe hypothesis as there is no preference for UAA, UGA or UAG as an ASC downstream of the primary stop [13]. While, however, one might imagine selection that favours ASCs might also be strong enough to bias usage towards the strongest stop (UAA), this is a second order effect compared with selection for any ASC in leaky genes.

Differential leakiness of stop codons in eubacteria provides a foundation for testing the fail-safe hypothesis. While UAA [29], UGA [30, 31], and UAG [29, 32, 33] are all subject to readthrough, they do so to differing degrees. The mechanistic basis for this variation is thought to relate to the specificity and abundance of release factors. The stop codons are recognised by a class I release factor [46–49], with their dissociation mediated by class II release factors following peptide release [50]. In bacterial lineages decoded according to translation table 11 (TT11), the class I release factors responsible are RF1 and RF2. UAG is recognised by RF1, UGA is recognised by RF2, and UAA is recognised by both RF1 and RF2 [48, 51, 52]. It is thought that the ability of UAA to bind both RF1 and RF2 contributes to it being the least 'leaky' stop. No matter what the mechanism, the selection of ASCs is likely to be highest in UGA-terminating genes and weakest for UAA, all else being equal.

In addition to termination efficiency, there are at least two other predictors of stop codon usage, GC pressure and expression level, when comparing across genes and genomes. While between genomes genomic GC is a strong predictor of UAA and UGA alone, UAG and UGA, with identical nucleotide contents, show dissimilar trends, UGA usage being positively correlated with genomic GC while UAG usage is uncorrelated [13, 14, 53]. This is conjectured to

relate to co-evolution between RF1:RF2 ratios and GC content [14]. Within genomes it is considered that highly expressed genes should be under selection to employ UAA this being the least leaky. Indeed, while across bacteria UAA usage is well predicted by GC pressure, it is found to be enriched in highly expressed genes (HEGs) even in GC rich genomes [13, 14]. The resistance to GC pressure in HEGs is consistent with the notion that the net effect of read-through is a combined function of the per translation leakage rate and the number of translation events any given transcript is subject to.

Here we provide the first systematic test of the fail-safe hypothesis applied to eubacteria. We interrogate the 3' UTRs of a large sample of phylogenetically relatively independent bacterial species for enrichment of ASCs. In acknowledgment of prior studies, we control for GC pressure [13, 14, 53]. We ask whether we can detect ASCs at rates higher than expected given underlying nucleotide content, and whether 3' UTR codon switches seen in closely related species are biased towards ASC deposition compared to null (determined by out of frame rates). Further, we ask whether highly expressed genes have more ASCs and whether expression level and primary stop usage predicts ASC usage. The most extreme difference should be between highly expressed TGA ending genes, which should have strong ASC selection, and lowly expressed TAA ending genes in which fail-safe selection should be the weakest. We also ask if the presence of an ASC predicts the downstream presence of further ASCs and whether mollicutes employing only two stops under-employ the codon that isn't a stop.

The tests are, however, complicated by the fact that stop codon efficiency is also dictated by local genomic context [28]. Indeed, it has been observed that nucleotide substitution rate increases with downstream distance from the stop codon with no obvious plateau within the next six downstream 'codons' [12], bringing attention to this region as a potential influencer of termination efficiency. Such regions may be directly involved in the formation of termination complexes that include the ribosome [45]. As noted, one downstream element thought to affect termination is the nucleotide at position +4 [41, 42, 54, 55]. In eukaryotes, +4C is associated with an increase to ca. 3% readthrough in certain genomic contexts [55], whereas +4U is highly preserved in all three domains of life and thought to reduce readthrough rate via improved cross-linking with RF2 [42]. This is problematic as it tends to increase the frequency of 3' in-frame stops at the first downstream codon compared to the simplest null model. At a greater scale, at least a hexanucleotide sequence may affect termination efficiency [44, 55, 56]. Whilst this evidence was found in eukaryotes, it cannot be discounted that the local genomic context affecting readthrough rates in bacteria could extend beyond the fourth site nucleotide. Thus, we attempt to control for downstream motif preferences, in addition to GC content, in our assessment of whether ASCs are selected for error-control. We find that, in contrast to eukaryotes, the great majority of our evidence argues against the notion that 3' ASCs are selectively favoured. We speculate as to why this might be.

Results

Nucleotide controlled simulations suggest genome-wide avoidance, not enrichment, of ASCs

A prediction of the fail-safe 3' stop hypothesis is that stop codons should be enriched immediately after the primary stop. Thus, we assessed genomes for ASC enrichment through comparison against a null model where downstream 3' codons are chosen according to dinucleotide content only. This was achieved by the simulation of 10,000 dinucleotide-controlled 3' UTRs per genome, the calculated mean ASC frequencies being the 'expected' value and the Z-score being the deviation from this mean normalised to the standard deviation of the simulations. A positive Z-score is an instance where ASCs are overused compared to null.

The null neutral expectation was that there is no difference between the ASC frequencies of the real genomes and simulated sequences hence 50:50 split of positive and negative Z-scores. We instead find there to be significant variation from this ratio when considering the UTR as a whole but, unexpectedly, with an excess of instances of under-usage of stops (from codon position +1 to +6; 13/644 $Z > 0$, $p < 2.2 \times 10^{-16}$, two-tailed binomial test). The same under usage is seen at all sites when considered individually ($p < 2.2 \times 10^{-16}$ for all positions, two-tailed binomial tests; 89/644 $Z > 0$ at position +1, 56/644 $Z > 0$ at position +2, 36/644 $Z > 0$ at position +3, 35/644 $Z > 0$ at position +4, 48/644 $Z > 0$ at position +5, 40/644 $Z > 0$ at position +6). All significant findings survive multi-test correction ($p < 0.05/6$).

These results accord with what we see if we consider the proportion of genomes showing significant deviation compared to null ($|Z| > 1.96$). In this instance, the null expectation of the binomial test is no longer 50:50, rather that 95% of genomes will not be significantly deviated and 5% will. There is a significant variation from this ratio when considering UTR *en mass* ($p < 2.2 \times 10^{-16}$ for the whole UTR (553/644 genomes) and at each position ($p < 2.2 \times 10^{-16}$ at position +1 (177/644 genomes), $p < 2.2 \times 10^{-16}$ at position +2 (129/644), $p < 2.2 \times 10^{-16}$ at position +3 (136/644), $p < 2.2 \times 10^{-16}$ at position +4 (113/644), $p < 2.2 \times 10^{-16}$ at position +5 (92/644), $p = 3.1 \times 10^{-12}$ at position +6 (77/644), two-tailed binomial tests), all surviving multi-test correction ($p < 0.05/6$). Closer examination again indicates that significant enrichment (one-tailed test, therefore we now use $Z > 1.64$) occurs less than expected by chance ($p < 1.6 \times 10^{-13}$ for the whole UTR (1/644), $p = 2.8 \times 10^{-10}$ at position +1 (4/644), $p = 1.6 \times 10^{-13}$ at position +2 (1/644), $p = 4.5 \times 10^{-15}$ at position +3 (0/644), $p = 4.5 \times 10^{-15}$ at position +4 (0/644), $p = 1.6 \times 10^{-13}$ at position +5 (1/644), $p = 4.5 \times 10^{-15}$ at position +6 (0/644), one-tailed binomial tests). Indeed, when we consider under-enrichment ($Z < -1.64$), we find more significant results than expected by chance ($p < 2.2 \times 10^{-16}$ for whole UTR (570/644), $p < 2.2 \times 10^{-16}$ at position +1 (230/644), $p < 2.2 \times 10^{-16}$ at position +2 (206/644), $p < 2.2 \times 10^{-16}$ at position +3 (204/644), $p < 2.2 \times 10^{-16}$ at position +4 (176/644), $p < 2.2 \times 10^{-16}$ at position +5 (135/644), $p = 2.9 \times 10^{-12}$ at position +6 (119/644), one-tailed binomial tests). These results provide no *prima facie* support for the fail-safe hypothesis and, if anything, argue for ASC avoidance.

Is there anything peculiar about the genomes for which we find under usage of ASCs? As all three stop codon variants are AT-rich by nature, they are more likely to appear in AT-rich genomes by chance. The fail-safe hypothesis therefore predicts selection to retain ASCs most strongly in GC-rich genomes, where a dearth of ASCs is expected in the absence of selection. Our results are contra to this prediction, as we find a significant negative correlation between Z-score and GC3 content ($p < 2.2 \times 10^{-16}$, $\rho = -0.64$, Spearman's rank correlation) (Fig 1). This trend is consistent at all positions +1 to +6 (S1 Fig) with the magnitude of the gradient decreasing with 3' distance (S2 Fig). This result is repeated when considering raw ASC frequency instead of Z-score (S3 Fig). Indeed, it appears that it is where ASCs are predicted to be most needed that they most under-employed.

3' codon switches from stop to non-stop are more common in-frame than out-of-frame, suggesting avoidance of ASCs

Above, we not only find no evidence for ASC enrichment but for ASC avoidance. Could this be because genomes specifically remove ASCs at a higher rate than chance? Alternatively, perhaps switches from non-stop to stop occur at a lower rate than expected. We investigate both of these possibilities through analysing codon switches from stop to non-stop, and vice versa, in orthologous gene triplets. We employ 29 sets of triplet species (a paired ingroup and an outgroup) and consider the results *en mass*. For null expectations we employ the comparable rate (stop->non-stop, non-stop->stop) in the +1 reading frame of the 3' domain.

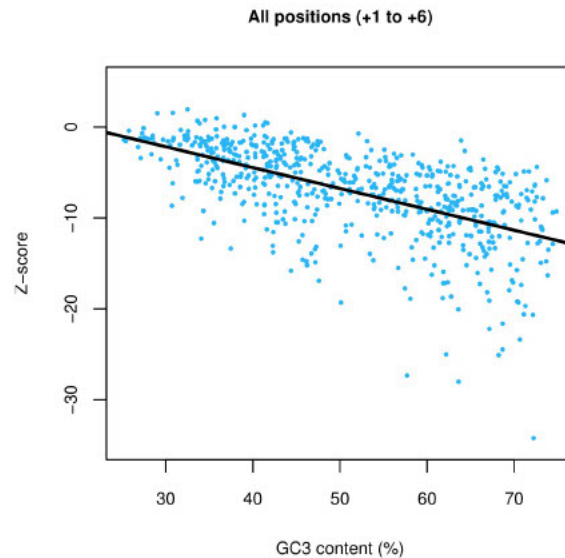


Fig 1. Z-scores, measuring deviation in ASC frequency from a null model (10,000 simulations), plotted against the genomic GC3 content of filtered TT11 bacterial genomes. A significant negative relationship is observed between Z-score and genomic GC3 across positions +1 to +6 ($p < 2.2 \times 10^{-16}$, $\rho = -0.64$, Spearman's rank correlation).
<https://doi.org/10.1371/journal.pgen.1008386.g001>

Considering all codons (Table 1) regardless of position in our dataset of the orthologous genes, we find the frequency of in-frame codon switches from non-stop to stop in 3' UTR codons to be no different to the same switch in out-of-frame codons of the same sequences ($p = 0.31$, $\chi^2 = 1.0$, Chi² test). Consistent with selection to avoid ASCs, switches from stop to non-stop occur significantly more often in-frame than out-of-frame ($p = 0.0024$, $\chi^2 = 9.2$, Chi² test). Hence not only are in-frame stops not deposited in 3' UTRs more than chance, they are if anything avoided. Both of these results corroborate the findings of our initial binomial tests and argue strongly against the fail-safe hypothesis in bacteria.

Considering each position individually tells a similar story. For the vast majority of codon switches at each position, there is no difference in switch rate between in-frame and out-of-frame codons (S1 Table). Exceptions to this are found at position +4, where switches from stop to non-stop are significantly more common in-frame than out-of-frame, and at position +5,

Table 1. Codon switch (from stop to non-stop and non-stop to stop) counts and frequencies compared between the in-frame and out-of-frame 3' UTR codons of 29 triplets of closely related bacterial genomes.

Switch	In-frame			+1 Frame-shift			Chi ² p-val
	Ancestral codons	Switch count	Switch frequency	Ancestral codons	Switch count	Switch frequency	
Stop > NS	2,693	655	0.243	3,146	679	0.216	0.0024
NS > Stop	51,525	664	0.0129	52,372	704	0.0134	0.31

<https://doi.org/10.1371/journal.pgen.1008386.t001>

where switches from non-stop to stop are significantly less common in-frame than out-of-frame. Both results are consistent with rejection of the fail-safe hypothesis, however do not survive even generous Bonferroni correction ($p > 0.05/6$). At position +1, switches from non-stop to stop are significantly more common in-frame than out-of-frame, though this is likely explained by selection for +4T.

No enrichment of ASCs in highly expressed genes

The above tests provide no support for the fail-safe hypothesis but consider genes equally, regardless of the primary stop codon and expression level. Selection for termination efficiency is thought to be highest in HEGs [13, 57] under the assumption that the net effect of readthrough is a function of the number of translation events the transcripts of any given gene are subject to. If the fail-safe hypothesis of ASCs is true, we therefore expect ASC frequencies to be significantly higher in HEGs than LEGs. This, however, does not seem to be the case. Unlike what is seen in yeast [39], there were no significant differences between the ASC frequencies of HEGs and LEGs at any position even before multi-test correction ($p = 0.95$ for whole UTR, $p > 0.05$ for all positions, Wilcoxon signed-rank tests; S4 Fig), suggesting that either expression level has no influence over the negative effects of readthrough or ASCs do not significantly affect the ability of a transcript to avoid these consequences. This test is however limited by small genome sample size. Through manually adding enrichment of stop codons to our data we find that a ~35% increase in HEGs compared to LEGs is required to retrieve a signal. Hence, we can be confident that ASC frequencies in our HEGs dataset do not exceed those seen in LEGs by this margin. We cannot investigate codon switches in highly and lowly expressed gene groups, as the PaxDb database does not contain compatible data to match the ATGC data we used for this analysis.

No enrichment of ASCs in TGA-terminating HEGs compared to TAA-terminating LEGs

The HEG/LEG analysis, whilst also negative, does not allow for covariance between expression level and usage of different stop codons. Notably the least leaky stop (TAA) is also the preferred one in the highly expressed genes [13, 14], which has the potential to dampen any differences between HEGs and LEGs. Under the fail-safe hypothesis, we expect TGA-terminating HEGs (high readthrough, high expression) to have the strongest selection for ASCs and TAA-terminating LEGs (low readthrough, low expression) to have the weakest. However, we find no significant difference between these groups when considering the whole UTR ($p = 0.36$, Wilcoxon signed-rank test). Aside from position +1, there is no significant difference between TGA-terminating HEGs and TAA-terminating LEGs at a single position scale ($p = 0.060$ for position +2, $p = 1$ for position +3, $p = 0.83$ for position +4, $p = 0.60$ for position +5, $p = 0.62$ for position +6, Wilcoxon signed-rank tests). Even at position +1 the enrichment of ASC in the TAG/HEG class is a barely significant trend ($p = 0.041$, Wilcoxon signed-rank test) that does not survive Bonferroni correction ($p > 0.05/6$) (Fig 2). We thus find no evidence to support the notion of ASC selection, apart from a possible very weak effect at position +1.

ASC enrichment at position +1 is peculiar to TGA-terminating genes

If the above exceptional result at position +1 is owing to selection we might also expect the enrichment to be seen in other TAG and TAA highly expressed genes. If, alternatively, it is a motif preference associated with TGA, we might expect it to be seen in lowly expressed TGA terminating genes but not necessarily elsewhere.

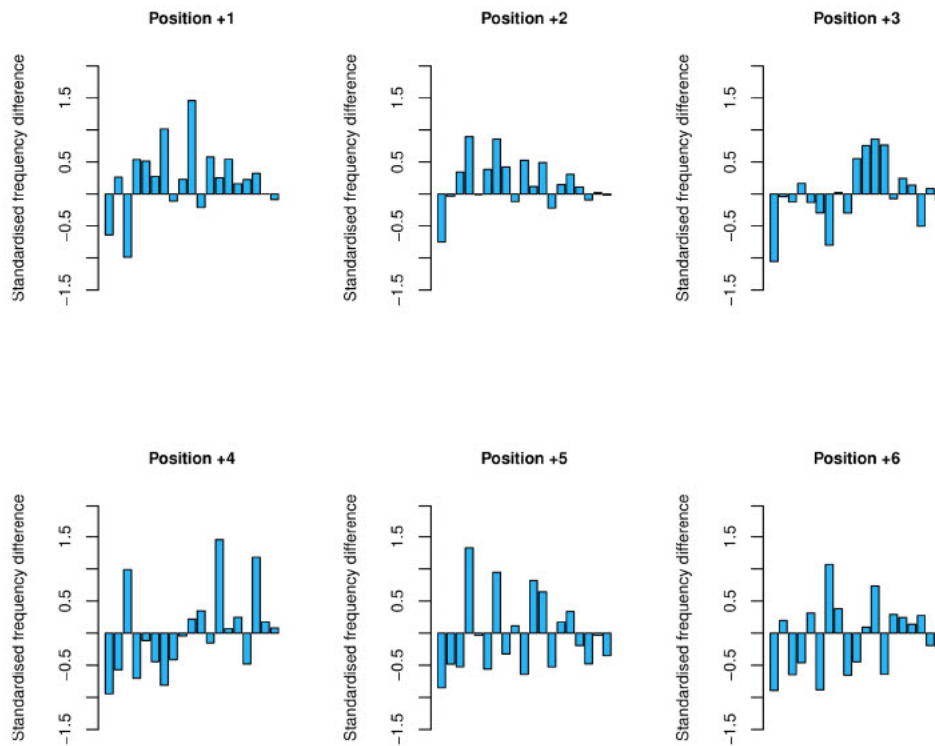


Fig 2. ASC frequencies of TGA-terminating HEGs compared to TAA-terminating LEGs. Each bar represents one genome, with bar heights representative of the standardised difference between the two groups. There is no significant difference between TGA-terminating HEGs and TAA-terminating LEGs at positions +2 to +6 ($p = 0.060$ for position +2, $p = 1$ for position +3, $p = 0.83$ for position +4, $p = 0.60$ for position +5, $p = 0.62$ for position +6, Wilcoxon signed-rank tests). There is prima facie significant difference between TGA-terminating HEGs and TAA-terminating LEGs at position +1 ($p = 0.041$, Wilcoxon signed-rank test), but this does not survive Bonferroni correction.

<https://doi.org/10.1371/journal.pgen.1008386.g002>

To examine these possibilities we consider all combinations of expression level and primary stops in the assessment of ASC frequency (Fig 3). Considering the whole UTR (+1 to +6) we find evidence for heterogeneity when considering all genes regardless of expression level ($p = 0.01$, $\chi = 8.79$, Kruskal-Wallis). However, if we remove position +1 from this analysis, significant heterogeneity cannot be recovered ($p = 0.57$, $\chi = 1.12$, Kruskal-Wallis). Indeed, we find that ASC enrichment is particular to position +1 and a peculiarity of TGA terminating genes weakly seen at all expression levels. We established this by first testing for heterogeneity between ASC usage dependent on the primary stop at position +1. When considering all genes ($p = 1.9 \times 10^{-15}$, $\chi = 67.81$, Kruskal-Wallis) and LEGs ($p = 0.032$, $\chi = 6.91$, Kruskal-Wallis) we see evidence for such heterogeneity. For HEGs ASC usage is highest for TGA terminating

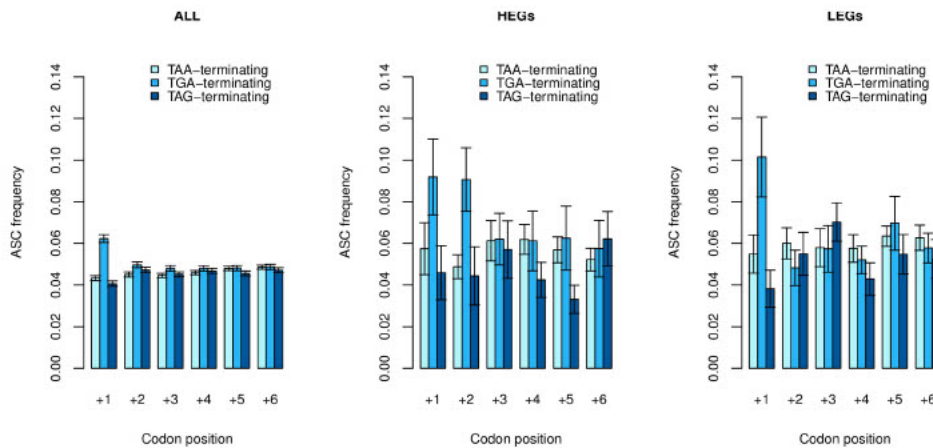


Fig 3. ASC frequencies calculated in TAA, TGA and TAG-terminating genes for all genes, highly expressed genes (HEGs) and lowly expressed genes (LEGs). Error bars represent bootstrapped standard error. We find significant differences between primary stop groups at position +1 when considering all genes ($p = 1.89 \times 10^{-25}$, $\chi = 67.81$, Kruskal-Wallis) and LEGs ($p = 0.032$, $\chi = 6.91$, Kruskal-Wallis), but not HEGs ($p = 0.14$, $\chi = 3.97$, Kruskal-Wallis). We instead observe significant enrichment at position +2 in HEGs ($p = 0.029$, $\chi = 7.09$, Kruskal-Wallis). Signals at position +1 in LEGs and at position +2 in HEGs do not survive Bonferroni correction. For all other positions, there was no significant difference in any expression group ($p > 0.05$, Kruskal-Wallis).

<https://doi.org/10.1371/journal.pgen.1008386.g003>

genes but not significantly so ($p = 0.14$, $\chi = 3.97$, Kruskal-Wallis). Similarly the significance at position +1 in LEGs does not survive Bonferroni correction ($p < 0.05/6$). With some evidence for heterogeneity, we proceed to post-hoc Wilcoxon signed-rank tests for the two significant cases these indicating in each case, enrichment is highest in TGA-terminating genes (position +1 all genes: TGA > TAA, $p < 2.2 \times 10^{-16}$; TGA > TAG, $p < 2.2 \times 10^{-16}$; position +1 LEGs: TGA > TAA, $p = 3.0 \times 10^{-3}$; TGA > TAG, $p = 1.3 \times 10^{-3}$, Wilcoxon signed-rank tests). That we do not find significant deviation between primary stops at position +1 in HEGs is surprising, however likely comes as a direct consequence of small sample size.

Confirming the lack of signal outside of position 1, for all such positions, there is no significant difference in any expression group ($p > 0.05$, Kruskal-Wallis), with one exception, this being significant enrichment at position +2 in HEGs ($p = 0.029$, $\chi = 7.09$, Kruskal-Wallis). Here too the effect is most pronounced for TGA terminating genes (position +2 HEGs: TGA > TAA, $p = 6.9 \times 10^{-3}$; TGA > TAG, $p = 0.027$, Wilcoxon signed-rank tests), but neither the original test nor the substest survive Bonferroni correction.

+4T nucleotide preference rather than ASC selection best explains ASC enrichment at codon 1

Above we have shown that TGA-terminating genes are commonly immediately followed by ASCs. There are two hypotheses for this: (i) a general enrichment of thymine at the fourth coding residue that enables more effective termination [13, 41, 42], most especially true for TGA due to its unique recognition by RF2 alone, and (ii) an enrichment of ASCs in response to TGA leakiness. Several lines of evidence argue in favour of the former.

First, we sought to establish whether there was general +4T enrichment. To this end we calculated the frequency of T-starting codons at position +1 and compared it to the average T-starting codon frequency from positions +1 to +6. T-starting codons at position +1 were found to be enriched compared to other downstream positions ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test). However, this is not necessarily attributable to the presence of position +1 ASCs. In repeating the same methodology, we find the frequency of all non-stop T-starting codons to be significantly enriched at position +1 compared to the UTR average in genes that don't have a position +1 ASC ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test). This effect is most heavily influenced by TGA-terminating genes, in which T-starting non-stop codons are more enriched at position +1 compared to the UTR average ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test) than seen in TAA-terminating genes ($p = 4.7 \times 10^{-14}$, Wilcoxon signed-rank test) and TAG-terminating ($p = 0.9951$, Wilcoxon signed-rank test) genes.

Second, we find ASC frequencies at position +1 in HEGs and LEGs are not significantly different ($p = 0.66$, Wilcoxon signed-rank test). In absolute terms the enrichment in LEGs is if anything higher. This is contra to the fail-safe prediction that ASCs should be most greatly enriched in HEGs.

Third, if the effect is owing to translation termination signals favouring +4T, then +4T enrichment might be expected to be most profound in TGA terminating genes and weakest in TAG terminating genes as RF2 crosslinking [43, 44] would be irrelevant for RF1-recruiting TAG. As TAA can use RF2 or RF1 it should be intermediate. To investigate this, we analysed the relative usage of thymine against adenine, cytosine, and guanine at the fourth site as a function of primary stop usage (Fig 4). Considering all genes this not only confirmed T enrichment compared to the next most frequent nucleotide, unique to TGA-terminating genes (T > A: $p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test) but, consistent with the RF2 crosslinking hypothesis, the +4T usage was in the order TGA > TAA > TAG. +4T frequency is significantly higher in TGA-terminating genes than TAG-terminating genes ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test) and TAA-terminating genes than TAG-terminating genes ($p = 7.5 \times 10^{-3}$, Wilcoxon signed-rank test).

The strength of +4T enrichment in TGA and weakness in TAG-terminating genes is underscored when we consider HEGs and LEGs separately. Thymine frequency at the fourth site significantly exceeded the next highest nucleotide regardless of the primary stop in HEGs, in the predicted order (T > A, $p = 2.9 \times 10^{-4}$ in TGA-terminating genes; $p = 0.013$ in TAA-terminating genes; $p = 0.045$ in TAG-terminating genes, Wilcoxon signed-rank tests). The signal in TAG-terminating genes in this instance does not withstand multi-test correction ($p > 0.05/3$). In LEGs, too, raw +4T frequency is found in the expected order TGA > TAA > TAG, with enriched frequencies of thymine evident only in TGA-terminating genes (T > A: $p = 1.2 \times 10^{-4}$, Wilcoxon signed-rank test).

Analysis of T starting codons suggest a [TAA|TGA]T[T/C] motif

The above results suggest that any weak stop excess at codon position +1 is not owing to selection for stops *per se*. Is the enrichment for T-starting codons the same for all such codons, stops included, or might some classes be especially preferred, suggesting some further motif structures? To investigate this, we calculated an enrichment score for each T-starting codon (Fig 5). We notice an enrichment of TC and TT-starting codons at position +1, particularly in HEGs and TGA-terminating genes. Indeed, we propose that there may be a fifth nucleotide site preference for thymine or cytosine in +4T-containing genes as part of a wider motif beneficial for translation termination. Consistent with this, the enrichment of stop codons at position +1 is unremarkable compared to other T-starting codons. This is, too, consistent with our

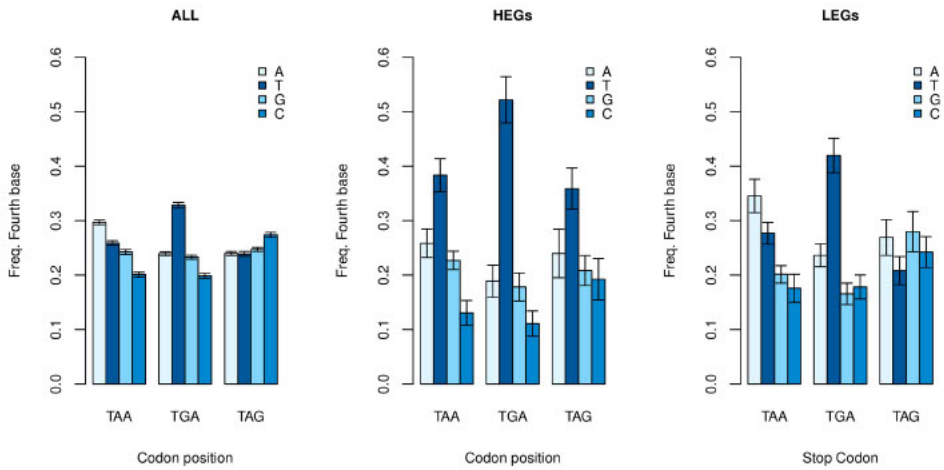


Fig 4. Assessment of fourth base nucleotide frequencies as a function of primary stop usage. Standard errors represent bootstrapped standard error. In all genes, not only is +4T enriched, compared to the next highest base, in TGA-terminating genes ($T > A$; $p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test), but consistent with the RF2 crosslinking hypothesis, the +4T was in the order TGA>TAA>TAG.

<https://doi.org/10.1371/journal.pgen.1008386.g004>

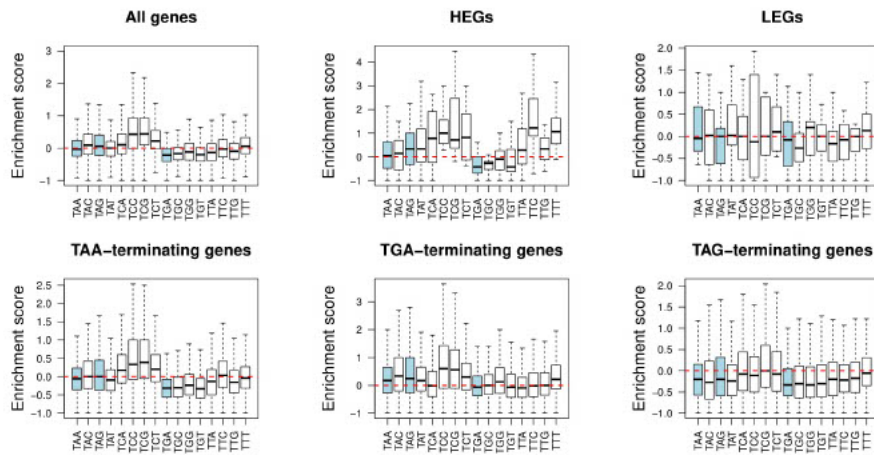


Fig 5. Enrichment of T-starting codons at position +1. Enrichment scores calculated for each T-starting codon at position +1 such that: Enrichment Score = $[F1 / \text{mean}(F3 + F4 + F5 + F6)] - 1$, where F1 = frequency at position +1 etc. Stop codons, highlighted in blue, show no remarkable enrichment compared to other T-starting codons.

<https://doi.org/10.1371/journal.pgen.1008386.g005>

+4T-controlled simulation experiment (S5 Fig), which finds that increased ASC frequencies at position +1 are the direct consequence of +4T enrichment. Further analysis suggests that TT is preferred in HEGs regardless of the primary stop. This partially reflects an AT bias in our genome set and more generally the preference for TT in AT rich genomes and TC in GC rich ones (S6 Fig).

No evidence that ASCs are enriched with downstream T, apart from TGA

As stops appear to prefer a +4T to enable stop codon recognition, we can ask whether this is also true of ASCs. We thus test the null that ASCs are as likely to have a downstream T as primary stops. For all genes, ASCs have significantly less chance to be immediately followed by a T than do primary stops ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test). The same is seen in HEGs ($p = 2.4 \times 10^{-6}$, Wilcoxon signed-rank test), though not LEGs ($p = 0.078$, Wilcoxon signed-rank test). The fail-safe hypothesis however does not necessarily predict selection termination functionality at ASCs to match that of primary stops. A more generous null is to ask whether ASCs have more T at the +4 site than do non-ASC codons in the 3' region. We actually find for all genes that ASCs have lower chance of having this ($p < 1.5 \times 10^{-12}$, Wilcoxon signed-rank test). The same is seen in both HEGs ($p = 7.0 \times 10^{-3}$, Wilcoxon signed-rank test) and LEGs ($p = 0.027$, Wilcoxon signed-rank test).

A more specific approach assesses each stop codon variant individually. As +4T enrichment appears to be peculiar to TGA-terminating genes, we expect TGAT to be more common as an ASC than TAAT, and even more so compared to TAGT. All three stop variants are significantly less likely to be followed by T when in-frame downstream than when located at the primary stop site (TAA $p < 2.2 \times 10^{-16}$; TGA $p < 2.2 \times 10^{-16}$; TAG $p < 4.6 \times 10^{-5}$, Wilcoxon signed-rank tests). Though whilst TAA ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test) and TAG ($p < 2.5 \times 10^{-12}$, Wilcoxon signed-rank test) are less likely to possess a 3' neighbouring T than non-ASC codons, TGA is significantly more likely to ($p < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test). Hence there is exceptionalism of TGA which falls in line with the expectations of the fail-safe hypothesis. Indeed, ASC +4T frequencies are found in the expected pattern TGA > TAA > TAG (TGA > TAA $p < 2.2 \times 10^{-16}$; TGA > TAG $p < 2.2 \times 10^{-16}$; TAA > TAG 6.0×10^{-5} , Wilcoxon signed-rank tests). We do, however, find a contradictory result in that TGAT is no more common in HEGs than LEGs ($p = 0.56$, Wilcoxon signed-rank test), though this is affected by low genome sample sizes.

One might suggest that the enrichment of T following TGA in 3' positions compared to other non-stop codons could be attributed to dinucleotide preference. We control for this by comparing 3' TGA to non-stop codons with third nucleotide A, finding again that TGAT to be significantly more common ($p = 2.9 \times 10^{-5}$, Wilcoxon signed-rank test).

No evidence that ASC presence predicts reduced downstream ASC frequency

The above analyses provide little support for the fail-safe hypothesis as any weak site +1 trends appear better explained by +4T motif presence. The observation of ASC enrichment at codon +2 in TGA terminating HEGs (sensitive to Bonferroni correction) and the enrichment of 3' TGAT are the only results that doesn't obviously fit with this otherwise profound rejection of the hypothesis. Given this, and the difficulties allowing for complex GC pressure and motif issues, we consider alternative tests.

In theory, if ASCs function in the termination of translation, it is unlikely that an ASC will be followed by another. The combined action of the primary stop and the ASC should terminate translation such that net readthrough rates are negligible and there is no selection for a

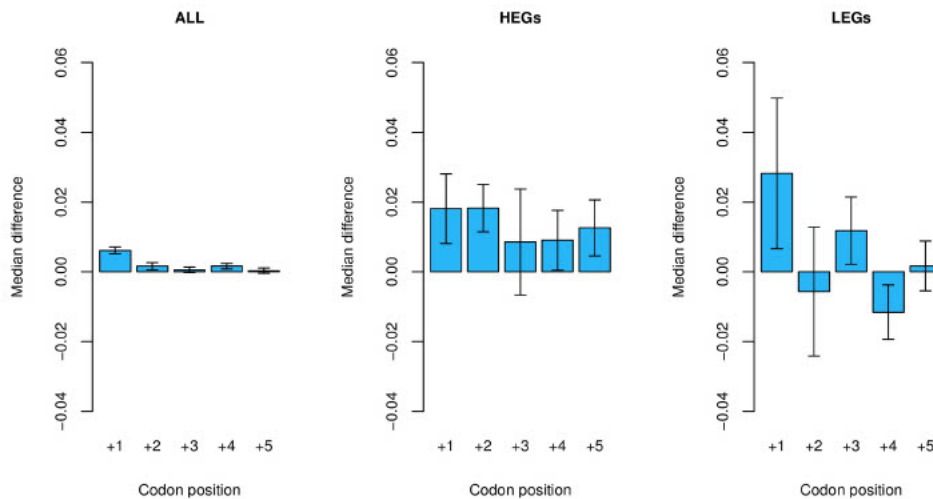


Fig 6. ASC-containing genes (at position +N, where N is a downstream codon position from +1 to +5) compared against ASC-absent genes for the presence of another ASC at the next codon position. Bars represent the median difference between ASC-absent and ASC-containing genes at each focal position. Error bars represent bootstrapped standard error. In the 'all genes' group, where an ASC is present at position +1 there is a significantly reduced chance of having another ASC at position +2 (Wilcoxon signed-rank test: $p = 3.6 \times 10^{-3}$).

<https://doi.org/10.1371/journal.pgen.1008386.g006>

third stop. We thus test the null hypothesis that ASC-containing genes, where the stop codon lies before (and including) codon +N, have an equal chance of possessing a further ASC downstream. We compare downstream ASC frequencies of ASC-containing and ASC-absent genes and see no evidence that possession of a stop predicts low rates of downstream stops ($p = 0.83$ where the focal codon is position +1, $p = 0.76$ for position +2, $p = 0.77$ for position +3, $p = 0.78$ for position +4 and $p = 0.92$ for position +5, one-tailed Wilcoxon signed-rank tests). This provides no support for the fail-safe hypothesis.

We can ask a more detailed question, namely whether having a stop at position N predicts the absence of a stop at the next codon position (+N+1, rather than generically downstream). In this case 'N' refers to each position from +1 to +5 (position +6 could not be tested in this instance as this would require analysis of ASCs at position +7, which is not considered). Where we consider all genes, ASC-absent genes demonstrate no significant excess of ASCs at position +N+1 over ASC-containing genes at all positions ($p > 0.05$), except where the focal codon position was position +1 ($p = 3.6 \times 10^{-3}$, Wilcoxon signed-rank test; Fig 6). Were this owing to selection, we expect to find a stronger signal in HEGs than in LEGs. However, there is no significant difference between HEGs and LEGs at any position ($p > 0.05$ for all positions +1 to +5, Wilcoxon signed-rank tests). A significant signal can only be found in HEGs at position +2 ($p = 0.027$, Wilcoxon signed-rank test), although this result does not survive multi-test correction ($p > 0.05/5$). We do however notice that the magnitude of the effect is actually greater in LEGs, which is contra to the fail-safe hypothesis.

We conclude that these tests provide no robust evidence that the presence of a stop codon predicts the presence/absence of further stops and if any such effects exist they are specific to

the domain in the immediate vicinity of the primary stop, suggesting that hidden motifs might be a viable alternative explanation.

TGA is under-used in TT4 mollicute UTRs compared to GC matched genomes that use TGA

The mollicute bacteria provide for a "natural" experiment as some genomes employ TT4 in which only TAA and TAG are used for chain termination. Hence, as TGA functions as a stop codon in TT11 genomes, it is expected under the fail-safe hypothesis that TGA frequency 3' of the primary stop in TT4 genomes should be consistently lower than that in TT11 genomes.

We tested this hypothesis by fitting a LOESS model (span = 2/3) for positions +1 to +6 usage of TGA against genomic GC3 in TT11 genomes of the full genome set. These models allowed the prediction of TGA frequencies of TT4 mollicute genomes at each position given their genomic GC3 content. TGA frequency was significantly reduced in TT4 genomes compared to predicted by the LOESS model at positions +3 and +5 ($p = 1.5 \times 10^{-1}$ for position +1; $p = 9.8 \times 10^{-2}$ for position +2; $p = 9.7 \times 10^{-4}$ for position +3; $p = 7.9 \times 10^{-1}$ for position +4; $p = 3.4 \times 10^{-4}$ for position +5; $p = 7.7 \times 10^{-2}$ for position +6, Wilcoxon signed-rank tests). For comparison, TT11 mollicute genomes do not significantly under-use TGA at any position ($p > 0.05$, Wilcoxon signed-rank tests). In TT4 genomes, lack of underrepresentation at position +1 possibly accords with the utility of +4T and similar motifs adjacent to the primary stop. The poverty of TGA at positions +3 and +5 survives multi-test correction and is consistent with the possibility that TGA maintains a function in TT11 genomes beyond its role in TT4 genomes. Why TGA is not underused at positions +2, +4 and +6 is unexplained. We do, however, find that when considering the whole UTR (positions +1 to +6) TGA is used significantly less often in TT4 genomes than predicted ($p = 3.8 \times 10^{-6}$, Wilcoxon signed-rank test). We acknowledge the limitations of LOESS modelling, which include those relating to the arbitrary nature of kernel/span function, and therefore validate this result with a different test design (S7 Fig). Given the above we also asked whether TAA, TGA, and TAG codon switches occur at different rates in TT4 genomes. We find no significant differences (S2 Table) but strongly caution that the results are limited by drastically reduced gene sample size.

The above results are consistent with the hypothesis that TGA is underused in 3' domains when it isn't employed as a stop codon, compared with its usage in genomes of similar GC content when it can function as a stop. However, if TGA is underrepresented in TT4 decoded genomes due to its selection for error-proofing in TT11 decoded species, we expect the magnitude of this under-enrichment to consistently surpass all other codons. We thus investigated all 64 codons using the same LOESS methodology and ranked them by their one-tailed Wilcoxon signed-rank test p-value (S3 Table). We find TGA to be just the 25th most under-enriched codon at position +1, 20th at position +2, 4th at position +3, 49th at position +4, 2nd at position +5, and 16th at position +6. Instead, we find codons CCG (1st at positions +1, +4, +6), GTG (2nd at position +1, 3rd at positions +4 and +6), and TAT (1st at position +2, 2nd at position +3, 4th at position +1) among the more commonly underrepresented codons at specific positions. Assertions that there is something special about TGA, specifically relating to translational termination, therefore remains speculative.

Between-genome primary stop codon usage is reflected in downstream positions

The disconnect between TAG and TGA usage as a primary stop has been attributed to co-evolution between RFI:RF2 ratios and GC content [14]. If true, this renders stop usage tightly

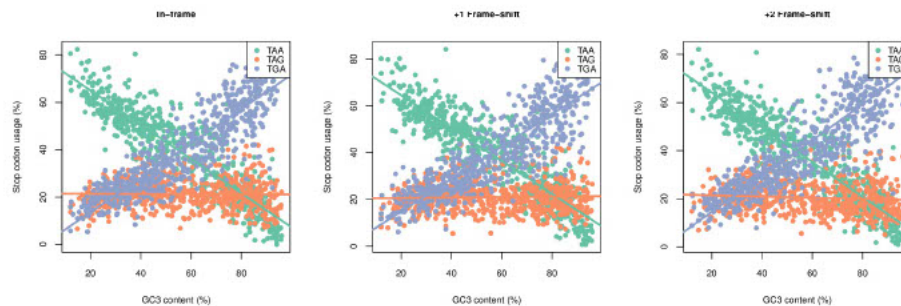


Fig 7. Relative usage of each stop codon in 3' UTRs plotted against GC3 content for TT11 bacterial species. Surprisingly, we find that trends in the decoupled TGA and TAG to be consistent across all three reading frames. Spearman's rank correlation information can be found in [S5 Table](#).

<https://doi.org/10.1371/journal.pgen.1008386.g007>

coupled to the mechanistic basis of translational termination. Are then these trends in TAA, TGA and TAG usage also seen downstream?

First, we analysed the relative usage of TAA, TGA and TAG at the primary site so as to repeat the findings of Korkmaz and colleagues (2014) with our genome set ([S8 Fig](#)). As expected we find that TAA-usage is negatively correlated with genomic GC3 ($\rho = -0.92$, $p < 2.2 \times 10^{-16}$, Spearman's rank correlation), TGA-usage is positively correlated with genomic GC3 ($\rho = 0.88$, $p < 2.2 \times 10^{-16}$, Spearman's rank correlation), and TAG-usage shows no significant correlation and remains at low levels regardless of genomic GC3 ($\rho = -0.017$, $p = 0.663$, Spearman's rank correlation). We then returned our focus to downstream. Surprisingly, we find that trends in TGA and TAG usage remains clearly decoupled despite their equal GC content. Indeed, trends in stop codon usage are remarkably similar between positions +0 to +6 ([S4 Table](#)).

That stop codon usage at the primary stop consistent in 3' positions implies either a) that the release factor hypothesis [22] regarding the decoupled usage of TGA and TAG usage is wrong or b) ASCs are, despite all the other negative data, under selection as fail-safe codons. We can investigate this by considering all three reading frames: should the relative codon usage of ASCs remain consistent in +1 and +2 frame-shift environments we can be relatively confident that usage is not controlled by selection relating to translational readthrough or termination. This is exactly what we find ([Fig 7](#)), and this is consistent with the bulk of the evidence described in our study. Thus, we suggest that the RFI:RF2 ratio is not the correct explanation for the differential stop usage as a function of GC and we are instead missing some important information regarding TGA and TAG usage.

Lack of ASC enrichment in bacterial genomes reveals a discrepancy with single-celled eukaryotes

The above bacterial evidence against ASC enrichment is in contrast to that seen in yeast and ciliates [39, 40]. Do prokaryotes and eukaryotes truly differ in their propensity to use ASCs to control translational read-through rates? Alternatively, might there be a reporting bias in which only significant effects surface in the published literature, thereby giving a skewed view of the commonality of fail-safe stops? Additionally, there are several ways to evaluate the fail-safe hypothesis and it could be that our methods would fail to report effects in the eukaryotic

species within which ASC enrichment has been observed. For example, while we employ a dinucleotide control, Adachi and Cavalcanti in the prior ciliate analysis [40] employ a method that considers the rate of occurrence of the first 3' stop as a function of downstream position given an underlying rate at which stops are observed in 3' UTR.

To ask whether our method would recover enrichment where previously claimed, we consider ASC enrichment in *T. thermophila*, *P. tetraurelia* and *S. cerevisiae* via the calculation of Z-scores, i.e. using the same method described earlier (Fig 8). Significant enrichment ($Z > 1.64$, for one tailed test of enrichment) is detectable using whole 3' UTR frequencies in the two ciliates but not in yeast. The latter negative result is not surprising as, unlike in ciliates, yeast enrichment is only detectable at position +3 and predominantly only when the primary stop is TAA [39]. Indeed, we find that position +3 is unusual in being enriched ($Z > 0$) in ASCs in all genes and in TAA-terminating yeast genes (Fig 8), although in neither is the effect significant ($Z = 0.93$ for all genes, $Z = 0.70$ for TAA-terminating genes).

These results suggest that our method can capture some but not all of the prior claims. Nevertheless, we extend the Z-score analysis to a set of 68 single-celled eukaryotes to investigate whether the spread of Z-scores matches that of the bacteria. We propose that single-celled eukaryotes are the fairest comparators to eubacteria as they are likely to both have large effective population sizes and, being single celled, would suffer the immediate consequences of any fitness costs of read-through. Multi-cellular organisms, by contrast, might be able to buffer fitness loss in one cell, for example by apoptosis and cell replacement. A genome is considered 'enriched' if it contains significant ASC enrichment at one or more positions ($Z > 2.33$, Bonferroni corrected one tailed). Interestingly, we find 20/68 of our eukaryotic genomes to be enriched, compared to 0/644 of our bacteria, these proportions being significantly different ($p < 0.0001$, $\chi^2 = 184.3$, Chi^2 test).

An alternative metric is to consider the number of genomes showing enrichment, defined by chi-squared, above dinucleotide controlled null frequencies at each position. For this we

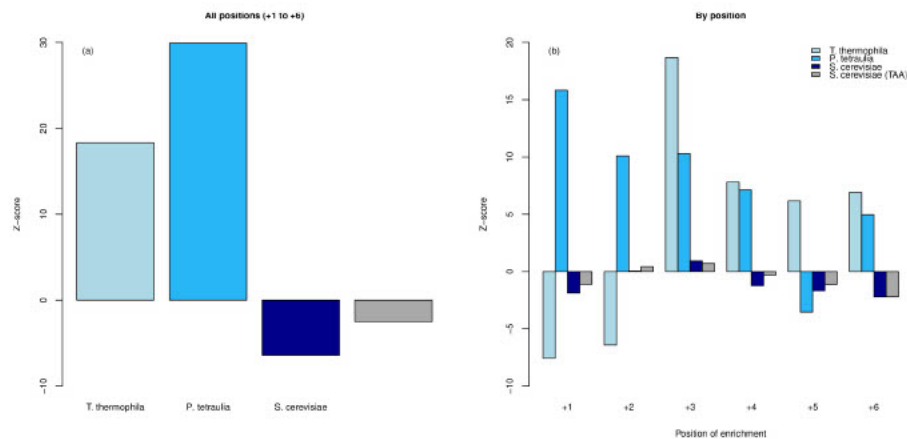


Fig 8. Z-score analysis of previously analysed eukaryotic genomes. Z-scores representing deviation from dinucleotide controlled null simulations over the whole 3' UTR sequence (a) and then each position individually (b) for three eukaryotic genomes.

<https://doi.org/10.1371/journal.pgen.1008386.g008>

employ a Chi^2 p-value $< 0.05/n$, where n is the number of positions tested, and apply this to our TT11 bacterial set of genomes and our set of 68 single-celled eukaryotes. Having defined positional ASC enrichment as $p < 0.01$ ($0.05/5$) as we analyse five positions (+2 to +6), the probability of a genome not possessing significant ASC enrichment at one or more positions is 0.99^5 (approximately 0.951). There is hence a $1 - 0.99^5$ (approximately 0.049) probability that a genome will contain significant enrichment at one or more positions. Hence, our null is that 4.9% of our genomes are expected to show ASC enrichment by chance alone. In our eukaryotic set, we find over-representation of genomes containing significant ASC enrichment compared to this null prediction (21/68, $p = 6.12 \times 10^{-12}$, one-tailed-binomial test with $p = 0.049$, expected = 3). Such a result supports evidence for ASC enrichment in eukaryotic systems [39, 40], however we note that whilst ASC enrichment is commonplace, it is not universal nor consistent in its position. By contrast in bacteria, using this same method, we find that significantly fewer bacterial genomes show enrichment than expected by null (21/644, $p = 0.028$, one-tailed binomial test with $p = 0.049$, expected = 32, Fig 9), consistent with a broad claim that eubacteria seem to avoid ASCs. Moreover, the observed proportions of 21/644 in bacteria and 21/68 in eukaryotes are significantly different ($p < 0.0001$, $\chi^2 = 79.6$, Chi^2 test), corroborating the results of our Z-score analysis.

We also repeat the Chi^2 comparison using an alternative null model as proposed by Adachi and Cavalcanti [40]. This too confirms the same results (Fig 10), namely avoidance of ASCs in bacteria, enrichment in single-celled eukaryotes. Indeed, this mode of analysis reports enrichment at one or more positions in 32 of 68 eukaryote genomes and only 7 of 644 bacterial genomes, these proportions being different ($\chi^2 = 242.3$, $p < 0.0001$, Chi^2 test).

The conclusions that there is indeed a discrepancy between bacterial and eukaryotic propensity to select for ASCs is hence both real and largely resilient to methodological nuance. With respect to the eukaryotes, we corroborate significant ASC enrichment (using at least one methodology) in the previously analysed yeast [39] (*S. cerevisiae*, plus *C. albicans*) and ciliates [40] (*P. tetraurelia*, *T. thermophila*). We note that the two ciliate species analysed in the prior study [40] possess a re-assigned translation table (TGA is the only stop). We not only recover ASC enrichment in these re-assigned ciliates (plus *S. lemnae*), but a translation table 11 (TGA, TAA and TAG are all stops) ciliate as well (*S. coeruleus*). Of our methodologies, the two dinucleotide-controlled analyses (Z-score: 20 enriched genomes, S7 Table; Chi^2 analysis: 21 enriched genomes, S8 Table) appear to be the most stringent in detecting eukaryotic ASC enrichment. Identification of enrichment using the Adachi and Cavalcanti null model [40] is more generous (32 enriched genomes, S9 Table). We do, however, note that ASC enrichment at one or more positions is recovered by all three methods in 15 eukaryotic genomes, indicating reasonable overlap between the tests.

Discussion

Our results suggest that, unlike in yeast, ciliates and some other protists, the error-proofing role of ASCs in bacteria is minimal at best. We began by testing the most obvious prediction of the fail-safe hypothesis, that stop codons should be enriched downstream of the primary stop codon. Having found no evidence for this at a genome-wide level, we considered the conservation of ASCs and found evidence that stops are less preserved than expected, this too being consistent with apparent avoidance. Additionally, we compared highly expressed and lowly expressed genes, seeing no differences. Comparing TGA-terminating HEGs and TAA-terminating LEGs we found TGA-terminating HEGs do not contain significantly higher ASC frequencies, except at position +1. The effect seen at +1 is not the result of selection for stops, but rather a knock-on consequence of selection for T-starting codons at the first codon

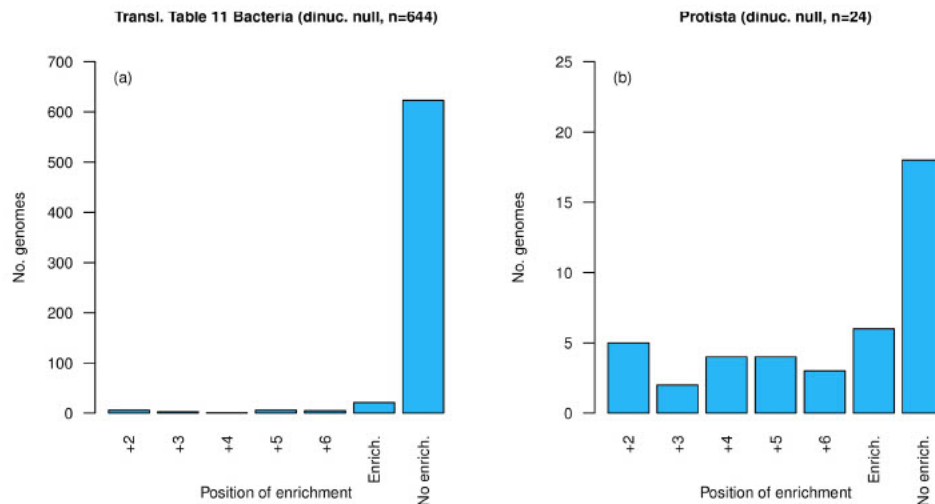


Fig 9. Number of genomes showing enrichment over dinucleotide-controlled null at each position, excluding position +1, in two genome sets (a) translation table 11 bacteria (n = 644) and (b) single-celled eukaryotes (n = 68). Genomes showing enrichment are underrepresented in the bacterial set (21/644, $p = 0.028$, one-tailed binomial test, expected = 32) and overrepresented in the eukaryotic set (21/68, $p = 6.12 \times 10^{-12}$, one-tailed-binomial test, expected = 3). 'Enrich.' is the total number of genomes with enrichment at one or more positions. 'No enrich.' is the total number of genomes with no enrichment at any position.

<https://doi.org/10.1371/journal.pgen.1008386.g009>

downstream, the trend being seen for non-stop T-starting codons too. Indeed, in the context of other T-starting codons stop codon usage is not simply unremarkable, the trend seems to be enrichment for non-stops, TT and TC being preferred residues. While it is suggestive that the leakiest codon (TGA) is the one associated with ASC enrichment at site +1, this trend is better explained by reference to the notion that RF2 cross-links with the adjacent +4T and TGA uses only RF2. Perhaps an informative test would compare species with defective/absent RF2 to those without, however we find no such genomes in our genome set. These results suggest bacteria and eukaryotes are different in the usage of fail-safe stops. Using several alternative methodologies to compare ASC enrichment in bacteria to protists, we validate that ASC enrichment is found in single celled eukaryotes more often than in bacteria. Our findings therefore highlight a discrepancy in the way that bacterial and eukaryotic genomes evolve in response to translational read-through. With respect to bacterial transgenes, our results thus do not support any major adjustments to their design or experimental protocols, beyond using TAA or TAAT[T/C] for termination.

A few results were consistent with the fail-safe hypothesis but not overwhelmingly so. While having a stop codon at any given position doesn't predict a dearth of downstream stops, if there is a stop at position +1 there is less likely to be one at position +2. However, the magnitude of this effect is greater in LEGs than HEGs questioning the overall relevance of this to the fail-safe hypothesis. Given too that the effects are seen exclusively in proximity to the primary stop, selection on unrecognised motifs is a viable and probably better alternative explanation. That TT4 mollicutes contain fewer TGAs in their 3' domain than expected is also enigmatic.

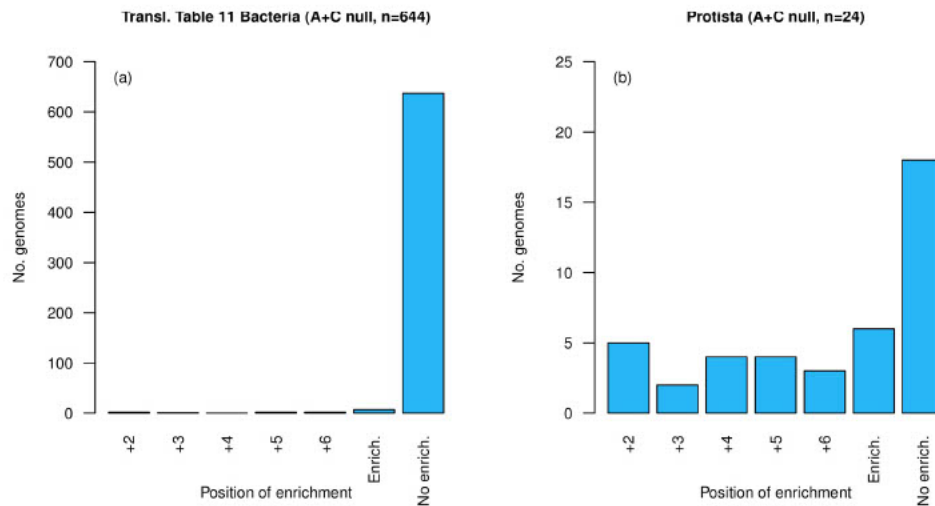


Fig 10. Number of genomes showing enrichment over A+C null (see methods) at each position, excluding position +1, in two genome sets (a) translation table 11 bacteria (n = 644) and (b) single-celled eukaryotes (n = 68). Genomes showing enrichment are underrepresented in the bacterial set (7/644, $p = 9.5 \times 10^{-8}$, one-tailed binomial test with $p = 0.049$, expected = 32) and overrepresented in the protists set (32/68, $p = < 2.2 \times 10^{-16}$, one-tailed binomial test with $p = 0.049$, expected = 3). 'Enrich' is the total number of genomes with enrichment at one or more positions. 'No enrich.' is the total number of genomes with no enrichment at any position.

<https://doi.org/10.1371/journal.pgen.1008386.g010>

That TT4 mollicutes contain less 3' UTR TGA than TT11 genomes (after control for GC content) is consistent with selection impacting TGA levels in 3' domains of TT11-decoded species. However, in TT11 genomes we see no evidence for ASCs beyond null levels and indeed, *prima facie* they seem to be avoided more often than enriched compared to GC controlled nulls. Furthermore, that some sense codons are even more consistently under-used at 3' UTR sites, for reasons that are unknown, suggests that there is a gap in our knowledge of the biology of these 3' ends.

A third possibly consistent result is that ASC usage of the three stops as a function of GC content matches that of the primary stop. The patterns for the primary stop were speculated to reflect co-evolution between GC content and RF1:RF2 ratios [14] but this remains to be verified. That we see the same broad trends at all downstream positions, despite all the other evidence against these ASCs being functional stops, we suggest more profoundly questions the RF1:RF2 ratio model than it supports ASC functionality. In accord with the under usage of TGA and other codons in TT4 genomes, perhaps more complicated dinucleotide or trinucleotide preferences should be considered.

This leaves one outstanding observation, namely that 3' TGA tend to be followed by T more than expected, even given the rate of T-starting codons with an A immediately prior. How can we explain this? We suggest a hypothesis that might explain the curious observations against a backdrop of a large body of negative evidence. First, we wish to discount the possibility that the lack of evidence for selection on ASCs relates to read-through not being a strong enough selective force. Experimental estimates in *E. coli* and *S. typhimurium* suggest that the

read-through rates are really very high. A read-through event at a TAA-terminating site can occur at frequencies between $>1 \times 10^{-5}$ – 9×10^{-4} [29], and at a TAG between 1.1×10^{-4} – 7×10^{-3} [28, 29, 32, 33]. If ASCs do meaningfully function in chain termination, one would have expected to find a signal in TGA-terminating genes, where readthrough may occur at rates of 1 in 1000 translation events up to 1 in 100 [15, 30, 31]. Thus, numbers suggest a potentially high rate of readthrough. Second, it is most likely because of this that stop codons are themselves subjected to selection for efficient termination. This is probably why TGA-terminating genes are rarely highly expressed—where such selection is expected to be strongest and TAA is over-represented in the set of highly expressed genes even in GC rich genomes [13]. Consistent with this, Belinky and colleagues (2018) found that stop codon switches occur significantly more frequently than the equivalent substitutions in non-coding DNA. Given this we assume that selection against read-through is a significant force.

We can then question whether, if read-through is the problem, ASCs are the expected solution in bacteria. Evidence from stop codon usage, especially in highly expressed genes suggests that there is selection for TAA enrichment as the stop. We could presume that in many cases this means simply a non-TAA stop mutates to be TAA and is selectively favoured, especially if the gene is highly expressed. However, there are other possibilities. For example, imagine that we have a highly expressed gene using TGA and so possesses high read-through rates. Imagine too that upstream are sense codons which could mutate in one step to TAA or indeed any stop. This would introduce a premature stop (assuming the context is otherwise fine) with, importantly, a guaranteed fail-safe stop downstream i.e. the original primary stop. There would be a benefit from lower net read-through rates (we presume nearly all genes will terminate at or before the second, original, stop) and a benefit from reduced translation costs when the new earlier stop functions. Moreover, the sequence now immediately 3' of the new stop will, if read-through happens, be sense codons of a recently functional protein, so there should be no toxicity of this additional sequence. All of these benefits suggest this is a viable path for evolution, the major cost owing to reduced gene length affecting protein function. However, tolerating such a cost appears to be possible, with stop codon shifts in 5' directions now thought to have an under-appreciated influence on gene shortening [58]. If the net benefits of reduction in read-through is greater than this cost then the system will have evolved towards reduced net readthrough.

Could this also explain why we detect no enrichment of stop codons in the 3' domain as, until the first ASC, selection would have recently been on this to perform as coding sequence? Might this explain the apparent general rarity of downstream T following an ASC? A stop lacking the +4T would be especially leaky and so especially favour rescue by creation of an earlier stop. The one exception could be TGAT. If this, like TGA, remains relatively leaky (unlike TAAT) then selection could still favour 3' stop creation. Might this also go some way to explaining the mollicutes result? If TGA wasn't a stop there is no reason it would by necessity feature in the 3' domain as the abandoned stop and so might appear at low frequency in the mollicutes.

An alternative trajectory to rescue a leaky TGA would be for TGA to mutate but to a sense codon. This could be favoured if the run on then meets a less leaky stop codon shortly downstream. The shortening process we suggest would be more common than the lengthening for several reasons. First, especially in highly expressed genes, addition of amino acids is likely to be costly, whereas loss would come with an energetic saving. Second, in the shortening process there are multiple potential sites that could mutate to a new upstream stop, while in the latter the mutation is required at the stop codon. Third in the gene shortening mode, at the time of mutation, at least one downstream site will be an ASC (the old primary stop), thus the system comes with guaranteed ASC protection. By contrast, gene extension could replace a leaky stop

with, at best a less leaky stop, but no guaranteed fail-safe ASC. Fourth, there is no guarantee on extension that the extension isn't toxic, while for read-through after shortening this would not be an issue. Thus, we suggest there may be a process to shorten highly expressed genes to enable evolution of protection from read-through that might be particular to prokaryotes. The difficulty with this model seems to be that the rate at which this would need to occur might have to be rather high. Whether this predicts any pattern is unclear as genes cannot continue to shorten indefinitely.

Might a propensity to gene shortening as a mechanism to cope with read-through also explain why ASC enrichment isn't seen in bacteria but is in eukaryotes? In eukaryotes the mutation creating this new upstream stop could be trapped by eukaryote-specific nonsense mediated decay (NMD) making gene shortening a non-viable solution. Perhaps for eukaryotes ASCs are the only viable solution (although how NMD knows a 3' stop isn't the true stop and the real primary stop not a premature stop is unknown). The model is consistent with HEGs generally being shorter (S6 Table) but this is not a discriminating prediction as a simple translational cost argument would predict the same. Arguing against such a model however is the finding that stops in the vicinity of the true stop might not trigger NMD, the stops having to be 3' of the last intron, at least in some species [59–61].

An alternative possibility to explain the eukaryote-prokaryote divide concerns the possibility that in some eukaryotes read-through rates can be greatly increased. Notably, the yeast prion [PSI⁺] state has been linked to extensive read-through via the misfolding of release factor Sup35p [6, 62, 63]. It is tempting to speculate that this provides a possible mechanism for increased selection of ASCs in yeast not seen in bacteria. Though the [PSI⁺] system in yeast is possibly best studied, it now appears that prion-like systems are present throughout the tree of life [64, 65], including bacteria [64]. Not all prion-like states affect translation termination, however. The identification of species susceptible to prion-induced increased translational read-through rate could provide a means to test the fail-safe hypothesis in the future. Such a model predicts co-incidence between genomes with ASC selection and prion-like systems affecting translation. Indeed, we are unaware of any bacterial prion system disrupting translational termination which would be consistent with the absence of ASC selection. The closest resemblance that we are aware of is with a system in *Clostridium botulinum* affecting a domain of transcription (not translation) termination factor Rho (Cb-Rho) [66].

Above we have presumed that read-through rates are the same in all genomes, with the possible exception of prion mediated read-through. In this context we note a further striking peculiarity, that ASC rates (Z-score deviation from dinucleotide controlled null) are especially low in GC-rich organisms. GC-rich organisms are typically thought to be those with stronger selection as the underlying mutational bias is towards AT [67, 68]. Assuming this reflects higher effective population sizes in GC-rich organisms, the lower Z-scores in GC-rich organisms is enigmatic—if anything one might expect selection to favour more ASCs if selection is strong. It is also enigmatic as in GC-rich genomes the span to the next random stop in the 3' domain is likely to be longer as stops are AT-rich, hence GC-rich genomes should also be under selection to conserve ASCs. However, this assumes all else is equal. If AT-rich bacteria are subject to higher read-through rates, the GC-trend might make some sense. Such a model would fit in the broader context of the possibility of stronger selection against error creation when populations are large and selection efficient [69]. Comparably, GC-rich organisms have a broader spectrum of tRNAs thought to reduce ribosomal frameshifting rates [70]. Might this also reduce read-through rates? An alternative possibility is that in GC-rich genomes, random ASCs are less likely to function as stops. If for example AT-richness in the vicinity of a stop is needed to enable stop functioning, then a random ASC in a GC-rich genome is, for example, unlikely to have a +4T and might thus be ineffective. Indeed, experimentally tandem stops

appear not to have the expected level of read-through suggesting particular context requirements [41, 42]. We suggest that experimental determination of read-through rates in organisms with different tRNA profiles would be informative.

Methods

General methods

All analyses were performed using bespoke Python 3.6 scripts. Statistical analyses and data visualisations were performed using R 3.3.3. Scripts can be found at <https://github.com/ath32/ASCs>. Whilst it is acknowledged that stop codons function at the mRNA level, in this analysis we have analysed chromosomal DNA sequences and henceforth refer to the three stops as TAA, TGA and TAG and to +4U enrichment as +4T. Please note that in all other contexts +1, +2 etc refer to the position of downstream codons, not nucleotides, with +1 being the codon immediately after the primary stop.

Genome downloads and filtering

Whole-genome sequences for 3,727 bacterial genomes were downloaded from the European Molecular Biology Laboratory (EMBL) database (<http://www.ebi.ac.uk/genomes/bacteria.html>, last accessed 1st August 2018). For the majority of the analyses, genomes were filtered to include only one genome per genus, so as to prevent over-sampling from the very well surveyed groups and hence to reduce any bias attributable to phylogenetic nonindependence. So as to exclude plasmids, incomplete genomes or very small genomes we retained only those genomes larger than 500,000 base pairs. This generated a sample of 650 genomes, 644 that employ translation table (TT) 11 and 6 using TT4, in which TGA no longer functions as a stop. The exception to this filtering was the specific analysis of mollusc and TT4 genomes, which were filtered directly from the raw sample of 3,727 genomes (106 and 94 genomes respectively). Of these genomes, only those with > 100 genes were considered for analysis.

For every gene in each genome, a sequence inclusive of the primary stop followed by 27 nucleotides of the 3' UTR was extracted by applying coding sequence coordinates to the total genomic sequence attainable in the EMBL files. Only genes with 3' intergenic space of >30 base pairs were considered for analysis, thus ensuring a sample of genes with sufficient 3' UTR length. Resultant sequences were filtered to retain only those 3' sequences made up exclusively of A, T, G and C, those from genes with one stop after the initiating codon, and those from a gene body with a nucleotide length that is a multiple of three. Genomic GC values were calculated from the whole genome sequence. GC3 values are unweighted means of per gene GC3 value.

Our single-celled eukaryotic set were downloaded and filtered much in the same way. 70 eukaryote genomes of unique genus were downloaded from the full Ensembl Protist set (<https://protists.ensembl.org/species.html>, last accessed 8th August 2019). Similar to the ciliates analysis by Adachi and Cavalcanti [40], we extracted a sequence inclusive of the primary stop followed by 97 nucleotides of the 3' UTR from each gene. As with the bacterial genomes, we do this by applying annotated coding sequence coordinates to the total genomic sequence. Only genes with 3' intergenic space of >100 base pairs were considered for analysis to ensure a sample of genes with sufficient 3' UTR length. Extracted 3' UTR sequences were subjected to the same filters as with the bacterial ones. We increased our sample with the addition of two yeast species via bespoke downloads—*S. cerevisiae* (yeastgenome.org) and *C. albicans* (candidagenome.org). For *S. cerevisiae*, annotated 3' UTR coordinates were applied to the whole genome sequence to extract the appropriate sequence. For *C. albicans*, 3' UTR sequences were located downstream from the first in-frame stop codon of downloadable ORFs (that contain

intergenic sequence). We exclude genomes with < 500 qualifying 3' UTR sequences, leaving a final sample of 68 genomes.

Protein abundance data

Experimental protein abundance data were downloaded for all genomes available from PaxDb [71]. Corresponding whole genome sequence files were downloaded from the European Molecular Biology Laboratory (EMBL) database. PaxDb external IDs and EMBL locus tags were extracted and matched to generate a sample of genomes and genes for which both PaxDb and EMBL sequence data were available ($n = 24$). In these genomes, qualifying genes that feature in the top and bottom quartiles of PaxDb data were defined as highly expressed genes (HEGs) and lowly expressed genes (LEGs) respectively. Only genomes with >100 qualifying HEGs and >100 qualifying LEGs were considered ($n = 22$). In reporting our results, we refer to the analysis of three gene groups: HEGs and LEGs which contain the qualifying genes of the 22 genomes for which there was available gene expression data, and 'all genes' where the qualifying genes of all filtered genomes are considered regardless of expression level.

Simulations

ASC frequencies for codon positions +1 to +6 were compared to expected frequencies generated from a null model where sequence is dictated solely by 3' UTR dinucleotide content. To achieve this, we simulated 10,000 UTR sequences for each genome using Markov models to preserve reading frame context at the dinucleotide level. ASCs are likely to occur by chance in every genome at a given rate that is dependent on its dinucleotide content. Hence the observation of ASC frequencies that exceed our null represents enrichment beyond chance. Nucleotide frequencies used in the Markov decision process were determined by generating a string containing the 3' UTRs of all qualifying genes from a given genome. The raw frequencies of each nucleotide within this string were calculated for the selection of the first base of each simulation. Overlapping dinucleotide frequencies were calculated for the selection of following simulated nucleotides according to the previously selected nucleotide. Simulations were complete once 21 nucleotides in length (equivalent to a primary stop followed by 6 downstream codons).

For each genome, ASC frequencies were calculated and compared to the mean ASC frequencies from the 10,000 simulated sequences at each of the 6 downstream codon positions. Comparison to null was established through the calculation of Z-scores under the assumption of a normal distribution to assess the magnitude of deviation from null in standard deviation units. Z-scores were used to complete various binomial tests using the `binom_test` function from the SciPy stats R package [72].

Translation table 4 mollicute analyses

The mollicute group contains both TT11 and TT4 genomes, allowing a side-by-side comparison in closely related species. TGA is not used as a stop codon in TT4 genomes. Hence, if observed TGA frequency is lower in TT4 genomes than in TT11 genomes, this implies selection upon TGA as an ASC in TT11 genomes. We design two tests to investigate whether TGA is underused in TT4 genomes.

(i) Frequency of TGA at codon positions +1 to +6 was plotted against genomic GC3 content in TT11 genomes from the full genome set ($n = 644$). A LOESS model was fit to allow the prediction of TGA frequency of TT11 and TT4 mollicute genomes according to their GC3 content at each position. TGA frequencies at each position for mollicute genomes were calculated

and compared to their predicted values. The fail-safe hypothesis predicts under enrichment of TGA in the TT4 genomes, but not TT11 ones.

(ii) Frequencies of TGA at positions +1 to +6 were calculated for TT4 mollicute genomes and compared to those of GC3 content-matched TT11 genomes from the full genome set. TT11 genomes were selected for comparison if their genomic GC3 content lies within 3.5% of the focal TT4 genome. Mean TGA frequencies for each position were calculated for selected TT11 genomes and compared with the corresponding TT4 genome frequency.

Third stop frequency as a function of presence/absence of a prior ASC

Genes with an ASC were compared to those without. The null expectation is that those containing an ASC before (and including) position +N have an equal chance of possessing another ASC downstream as genes without one. Two groups of genes were thus extracted for each position—those with an ASC up to position N and those without. ASC frequencies of each group were calculated for downstream positions up to position +6 and compared. Given the nature of this experiment, no data is available for position +6 (as there is no further downstream position to use to calculate ASCs within our chosen intergenic range). To consider more localised nucleotide preferences, we also repeat this methodology considering just the following base (+N+1) instead of all downstream positions.

Nucleotide enrichment at fourth and fifth nucleotide sites

For the analysis of the fourth nucleotide site of the primary stop codon, raw nucleotide frequencies (A, T, G, C) were calculated. Fourth site T enrichment relative to null was investigated through the comparison of T-starting codon frequency at position +1 to the mean frequency of T-starting codons throughout the 3' UTR (+1 to +6) using a Wilcoxon signed-rank test.

The analysis of the fifth nucleotide site of +4T-containing genes was completed in a similar manner. Raw nucleotide frequencies at nucleotide position +5 of genes were calculated, plotted for visual comparison and used in the completion of statistical analysis. Fifth site T and fifth site C enrichment relative to null was investigated through the comparison of TT/TC-starting codon frequency at position +1 to the mean frequency of TT/TC-starting codons throughout the 3' UTR (+1 to +6) of the given genome using a Wilcoxon signed-rank test.

3' stop codon switch analysis

Analysis of stop codon switches (from non-stop to stop, or vice versa) was completed by adapting a methodology described in previous studies [12, 73, 74]. Orthologous gene information for closely related species were downloaded from the Alignable Tightly Genome Clusters (ATGC) database [75]. Corresponding whole genome sequence data was downloaded from NCBI [76]. Where possible, the same triplets (containing two closely related ingroup species and one outgroup to allow the reconstruction of mutations by a parsimony approach) were downloaded as used in previous studies [12, 73, 74]. In total, 29 ATGC triplet clusters were considered in the analysis (8 of the 37 clusters used in prior studies were ineligible).

All gene sequences from each ATGC-COG (Cluster of Orthologous genes) were aligned using the *-linsi* parameter of MAFFT [77]. Aligned genes without gaps downstream of the primary stop, from all genomes, were considered together in the codon switch analysis. Ancestral codons were inferred where the outgroup codon matched at least one of the ingroup codons. A switch was recorded where one of the ingroup codons differed from both the other ingroup codon and the outgroup codon (and thus the inferred ancestral codon). Frequencies of switches from non-stop to stop and stop to non-stop amongst 'in-frame' 3' UTR codons were

calculated. These were compared to null frequencies, calculated through the analysis of the same sequences but with +1 frameshift.

Bacteria and single-celled eukaryote comparisons

To provide fair comparison between prokaryotes and eukaryotes we adopt the same set of methodologies for both genome sets. Due to the nature of ASC enrichment in eukaryotes not being universally specific to a particular codon position, we count the number of genomes in each set that possess ASC enrichment at one or more site (between +2 to +6). ASCs at a particular position were considered to be enriched if they produced a positive Chi^2 value and a p-value below 0.05/5 (after Bonferroni correction) when compared to the mean from a dinucleotide controlled null (see 'Simulations' section of these methods). As we set our p-value threshold at 0.01, the probability of a genome possessing significant ASC enrichment at one or more positions by chance is 0.99^5 (approximately 0.951). Therefore, there is a $1-0.99^5$ (approximately 0.049) probability that a genome will contain significant enrichment at one or more positions by chance. We determined whether the number of genomes containing enrichment in each set was higher, lower, or as expected by using binomial tests under the null expectation that 4.9% of genomes possess enrichment purely by chance.

We additionally repeat the analysis using the null model proposed by Adachi and Cavalcanti [40]. In their analysis of ASC enrichment in ciliates, they consider the probability of finding the first in-frame stop codon as a function of 3' distance from the primary stop. The probability of finding the first stop at position +1 is equal to the probability of finding a stop at any position, p . The probability p is calculated for each genome by concatenating the first 100 non-coding nucleotides downstream of each gene, scanning this sequence for in-frame stops, and dividing the total number of stops by the total number of codon positions considered. At position +2, the probability of finding the first stop is the probability of not finding a stop at any position upstream, in this case position +1, multiplied by the probability of finding a stop at any position. This concept is recursively applied with each position downstream such that first ASC probability = $p[1-p]^{n-1}$, where n is the focal codon position. For each position +1 to +6 we calculate ASC probability and multiply this by the total number of UTR sequences analysed to determine the expected number. We then apply a Chi^2 test. To determine whether the number of genomes showing significant enrichment at one or more sites is higher, lower, or as expected, we apply a binomial test as described above.

Supporting information

S1 Fig. Z-scores measuring deviation from a null model (10,000 simulations) plotted against the genomic GC3 content of filtered TT11 bacterial genomes. Significant negative relationships were observed between Z-score and genomic GC3 content at each position (Spearman's rank: $p < 2.2 \times 10^{-16}$ for all positions; $\rho = -0.61$ at position +1, $\rho = -0.65$ at position +2, $\rho = -0.51$ at position +3, $\rho = -0.45$ at position +4, $\rho = -0.41$ at position +5, $\rho = -0.46$ at position +6). (TIF)

S2 Fig. Gradient analysis of Z-score plotted against genomic GC3. Raw gradients (of Z-score plotted against genomic GC3) plotted at each of the six positions downstream (A). Absolute gradients plotted against codon position (B). Our expectation under the fail-safe hypothesis is that at codon position +1, stops will be largely resistant to GC pressure while at position +6 this resilience will be diminished. We thus predict that looking across genomes, the plot of ASC usage against GC content should be flatter at site +1 than at site +6. Interestingly, there is

a significant correlation between absolute gradient and distance from the primary stop (Spearman's rank: $p = 2.8 \times 10^{-3}$, $\rho = -1$). We therefore infer that either the presence of ASCs is more resilient to GC pressure when located further downstream relative to the primary stop, or ASCs are actively selected against at positions closest to the primary stop. Both of these inferences go against the fail-safe hypothesis.

(TIF)

S3 Fig. Raw ASC frequencies at positions +1 to +6 plotted against genomic GC3 content. There is a significant negative correlation between the variables at all positions tested (Spearman's rank: $p < 2.2 \times 10^{-16}$ for all positions; $\rho = -0.92$ for position +1, $\rho = -0.95$ for position +2, $\rho = -0.95$ for position +3, $\rho = -0.94$ for position +4, $\rho = -0.93$ for position +5, $\rho = -0.94$ for position +6). The gradient of the linear model fitted for position +1 is significantly different than that of position +6 ($p = 2.035732 \times 10^{-10}$), with position +6 having the more negative gradient.

(TIF)

S4 Fig. Gradient frequencies of HEGs compared to LEGs. Each bar represents one genome, with genomes ordered by genomic GC3 content from left to right. Bar heights represents the raw difference between in ASC frequency between the two groups tested. A positive difference represents enrichment in the HEGs group, a negative difference represents enrichment in the LEGs group. There is a no significant difference between HEGs and LEGs at any position (Wilcoxon signed-rank test: $p > 0.05/6$).

(TIF)

S5 Fig. Z-scores, measuring deviation of observed ASC frequencies in +4T-containing genes from +4T-containing null simulations, plotted against genomic GC3 content. Only position +1 and position +2 are considered as these are the only sites where a signal for ASC enrichment has been noted. We find Z-scores to be negatively correlated with genomic GC3 when considering all genes at position +1 (Spearman's rank: $\rho = -0.3054869$, $p < 2.2 \times 10^{-16}$) and position +2 (Spearman's rank: $\rho = -0.1880088$, $p = 1.62 \times 10^{-06}$). There is no relationship between Z-score and genomic GC3 in HEGs or LEGs at either position (Spearman's rank, $p > 0.05$).

(TIF)

S6 Fig. Assessment of fifth site nucleotide preferences. Fifth site nucleotide frequencies in +4T-containing genes of different primary stop and expression level (A). Frequencies of TC (B) and TT-starting (C) codons at position +1 compared to the average frequency of the respective codons between positions +1 to +6. Positive scores represent enrichment whilst negative scores represent under-representation.

(TIF)

S7 Fig. ASC frequencies of TT4 mollicute genomes calculated and compared to those of GC-matched TT11 genomes. Each bar represents the frequency difference between a mollicute genome and the average of its GC-matched TT11 genomes. TGA was underrepresented at positions +3 and +5 only (Wilcoxon signed-rank tests: $p = 0.11$ for position +1; $p = 0.15$ for position +2; $p = 1.5 \times 10^{-3}$ for position +3; $p = 0.70$ for position +4; $p = 6.8 \times 10^{-4}$ for position +5; $p = 0.11$ for position +6).

(TIF)

S8 Fig. Relative usage of each stop codon as the primary stop (position +0) and as ASCs at positions +1 to +6. Contra to our expectations, we find codon usage at positions +1 to +6 to

be consistent with that of the primary stop.

(TIF)

S1 Table. Positional codon switch (from stop to non-stop and non-stop to stop) counts and frequencies compared between the in-frame and out-of-frame 3' UTR codons of 29 triplets of closely related bacterial genomes.

(TIF)

S2 Table. Codon switch (from stop to non-stop and non-stop to stop) counts and frequencies compared between the in-frame and out-of-frame 3' UTR codons of a triplet of TT4 mollicute genomes.

(TIF)

S3 Table. 3' UTR frequencies of all codons in TT4 mollicutes compared to prediction by LOESS model fitted to TT11 codon frequencies. Observed frequencies were compared to predicted frequencies using one-tailed Wilcoxon-signed rank tests, the p-values from which are found in the table below. A significant p-value represents significant under-enrichment in the TT4 genomes. Stop codons are highlighted in blue.

(TIF)

S4 Table. Association between relative stop codon usage and GC3 content assessed by Spearman's Rank tests at each downstream position.

(TIF)

S5 Table. Association between relative stop codon usage and GC3 content assessed by Spearman's Rank tests in each reading frame.

(TIF)

S6 Table. Association between gene length and gene expression level assessed at genome-wide level. Gene expression (represented by experimental protein abundance data) and gene nucleotide lengths were used in Spearman's rank tests.

(TIF)

S7 Table. Z-score analysis of single-celled eukaryotic genomes. Z-scores represent deviation in ASC frequency from dinucleotide-controlled simulations.

(TIF)

S8 Table. Chi² analysis of ASC frequency single-celled eukaryotic genomes against dinucleotide-controlled simulations.

(TIF)

S9 Table. Chi² analysis of ASC frequency single-celled eukaryotic genomes against the Adachi and Cavalcanti null.

(TIF)

S1 Text. Supporting text for [S1 Fig](#), [S2 Fig](#) and [S3 Fig](#).

(DOCX)

S2 Text. Supporting text for [S1 Table](#) and [S2 Table](#).

(DOCX)

S3 Text. Supporting text for [S4 Fig](#).

(DOCX)

S4 Text. Supporting text for [S5 Fig](#).

(DOCX)

S5 Text. Supporting text for [S6 Fig.](#)
(DOCX)

S6 Text. Supporting text for [S7 Fig.](#)
(DOCX)

S7 Text. Supporting text for [S3 Table.](#)
(DOCX)

S8 Text. Supporting text for [S8 Fig.](#), [S4 Table](#) and [S5 Table.](#)
(DOCX)

S9 Text. Supporting text for [S6 Table.](#)
(DOCX)

S10 Text. Supporting text for [S7 Table](#), [S8 Table](#) and [S9 Table.](#)
(DOCX)

Author Contributions

Conceptualization: Alexander T. Ho, Laurence D. Hurst.

Data curation: Alexander T. Ho.

Formal analysis: Alexander T. Ho, Laurence D. Hurst.

Funding acquisition: Laurence D. Hurst.

Investigation: Alexander T. Ho.

Methodology: Alexander T. Ho, Laurence D. Hurst.

Project administration: Alexander T. Ho, Laurence D. Hurst.

Supervision: Laurence D. Hurst.

Visualization: Alexander T. Ho.

Writing – original draft: Alexander T. Ho.

Writing – review & editing: Alexander T. Ho, Laurence D. Hurst.

References

1. Warnecke T, Hurst LD. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat Rev Genet.* 2011; 12(12):875–81. <https://doi.org/10.1038/nrg3092> WOS:000297252500013. PMID: 22094950
2. Fu Q, Liu CJ, Zhang X, Zhai ZS, Wang YZ, Hu MX, et al. Glucocorticoid receptor regulates expression of microRNA-22 and downstream signaling pathway in apoptosis of pancreatic acinar cells. *World Journal of Gastroenterology.* 2018; 24(45):5120–30. <https://doi.org/10.3748/wjg.v24.i45.5120> WOS:000452759500007. PMID: 30568389
3. Liu Z, Zhang JZ. Human C-to-U coding RNA editing is largely nonadaptive. *Mol Biol Evol.* 2018; 35(4):963–9. <https://doi.org/10.1093/molbev/msy011> WOS:000431889000014. PMID: 29385526
4. Liu Z, Zhang JZ. Most m(6)A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Mol Biol Evol.* 2018; 35(3):666–75. <https://doi.org/10.1093/molbev/msx320> WOS:000427260700013. PMID: 29228327
5. Yang JR, Maclean CJ, Park C, Zhao HB, Zhang JZ. Intra and interspecific variations of gene expression levels in yeast are largely neutral: (Nei Lecture, SMBE 2016, Gold Coast). *Mol Biol Evol.* 2017; 34(9):2125–39. <https://doi.org/10.1093/molbev/msx171> WOS:000408307400001. PMID: 28575451

6. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 2009; 10(10):715–24. <https://doi.org/10.1038/nrg2662> ISI:000269965100014. PMID: 19763154
7. 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> WOS:000462993700035. PMID: 30883542
8. Chuan L, Zhang J. Stop-codon read-through arises largely from molecular errors and is generally non-adaptive. *PLoS Genet.* 2019; 15(5):e1008141. <https://doi.org/10.1371/journal.pgen.1008141> PMID: 31120886
9. Abrahams L, Hurst LD. Adenine enrichment at the fourth CDS residue in bacterial genes is consistent with error proofing for +1 frameshifts. *Mol Biol Evol.* 2017; 34(12):3064–80. <https://doi.org/10.1093/molbev/msx223> WOS:000416178900003. PMID: 28961919
10. Seligmann H, Pollock DD. The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 2004; 23(10):701–5. <https://doi.org/10.1089/dna.2004.23.701> WOS:000225034200012. PMID: 15585128
11. Seligmann H. Cost minimization of ribosomal frameshifts. *J Theor Biol.* 2007; 249(1):162–7. <https://doi.org/10.1016/j.jtbi.2007.07.007> WOS:000250847700014. PMID: 17706680
12. 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> WOS:000435448300003. PMID: 29915293
13. 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> WOS:000344549700016. PMID: 25217634
14. Wei YL, Wang J, Xia XH. Coevolution between stop codon usage and release factors in bacterial species. *Mol Biol Evol.* 2016; 33(9):2357–67. <https://doi.org/10.1093/molbev/msw107> WOS:000381702500016. PMID: 27297468
15. 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](https://doi.org/10.1016/0022-2836(73)90299-4) WOS:A1973P411600007. PMID: 4581523
16. Geller AI, Rich A. UGA termination suppression transfer RNA^{trp} active in rabbit reticulocytes. *Nature.* 1980; 283(5742):41–6. <https://doi.org/10.1038/283041a0> WOS:A1980JA27900029. PMID: 7350525
17. Parker J. Errors and alternatives in reading the universal genetic code. *Microbiol Rev.* 1989; 53(3):273–98. WOS:A1989AN47600001. PMID: 2677635
18. Jorgensen F, Adamski FM, Tate WP, Kurland CG. Release factor-dependent false stops are infrequent in *Escherichia coli*. *J Mol Biol.* 1993; 230(1):41–50. <https://doi.org/10.1006/jmbi.1993.1124> WOS:A1993KR93300007. PMID: 8450549
19. Meng SY, Hui JO, Hanlu 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> WOS:A1995RB8500007. PMID: 7779107
20. 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. WOS:000076580900002. PMID: 9803532
21. Tate WP, Mansell JB, Mannering SA, Irvine JH, Major LL, Wilson DN. UGA: a dual signal for "stop" and for recoding in protein synthesis. *Biochemistry-Moscow.* 1999; 64(12):1342–53. WOS:000084900100002. PMID: 10648957
22. Nichols JL. Nucleotide sequence from polypeptide chain termination region of coat protein cistron in bacteriophage-R17 RNA. *Nature.* 1970; 225(5228):147–51. <https://doi.org/10.1038/225147a0> WOS:A1970F007100021. PMID: 5409960
23. Doronina VA, Brown JD. When nonsense makes sense and vice versa: Non-canonical decoding events at stop codons in eukaryotes. *Mol Biol.* 2006; 40(4):731–41. WOS:000239572100018.
24. Namy O, Rousset JP. Specification of standard amino acids by stop codons. In: Atkins JF, Gesteland RF, editors. *Recoding: Expansion of Decoding Rules Enriches Gene Expression.* Nucleic Acids and Molecular Biology. 242010. p. 79–100.
25. 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 USA.* 2015; 112(10):3038–43. <https://doi.org/10.1073/pnas.1424127112> WOS:000350646500044. PMID: 25733896
26. 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/ma.054452.115> WOS:000371365400013. PMID: 26759455

27. 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> WOS:000297918600011. PMID: 21994247
28. Bossi L, Roth JR. The influence of codon context on genetic-code translation. *Nature.* 1980; 286(5769):123–7. <https://doi.org/10.1038/286123a0> WOS:A1980JY73500026. PMID: 7402305
29. 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> WOS:A1984RY11900006. PMID: 6419024
30. Sambrook JF, Fan DP, Brenner S. A strong suppressor specific for UGA. *Nature.* 1967; 214(5087):452–3. <https://doi.org/10.1038/214452a0> WOS:A19679255200007. PMID: 5340340
31. Roth JR. UGA nonsense mutations in *Salmonella typhimurium*. *J Bacteriol.* 1970; 102(2):467–75. WOS:A1970G085800022. PMID: 4315894
32. Bossi L. Context effects—translation of UAG codon by suppressor transfer-RNA is affected by the sequence following UAG in the message. *J Mol Biol.* 1983; 164(1):73–87. [https://doi.org/10.1016/0022-2836\(83\)90088-8](https://doi.org/10.1016/0022-2836(83)90088-8) WOS:A1983QD31100005. PMID: 6188841
33. Miller JH, Albertini AM. Effects of surrounding sequence on the suppression of nonsense codons. *J Mol Biol.* 1983; 164(1):59–71. [https://doi.org/10.1016/0022-2836\(83\)90087-6](https://doi.org/10.1016/0022-2836(83)90087-6) WOS:A1983QD31100004. PMID: 6188840
34. 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. <https://doi.org/10.7554/eLife.01179> WOS:000328643800002. PMID: 24302569
35. Wagner A. Energy constraints on the evolution of gene expression. *Mol Biol Evol.* 2005; 22(6):1365–74. ISI:000229279100001. <https://doi.org/10.1093/molbev/msi126> PMID: 15758206
36. Klauer AA, van Hoof A. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *WIREs RNA.* 2012; 3(5):649–60. <https://doi.org/10.1002/wrna.1124> WOS:000307731000005. PMID: 22740367
37. Ito-Harashima S, Kuroha K, Tatematsu T, Inada T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* 2007; 21(5):519–24. <https://doi.org/10.1101/gad.1490207> WOS:000244760600003. PMID: 17344413
38. 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> WOS:000265104600008. PMID: 19204001
39. Liang H, Cavalcanti ARO, Landweber LF. Conservation of tandem stop codons in yeasts. *Genome Biol.* 2005; 6(4):R31. <https://doi.org/10.1186/gb-2005-6-4-r31> WOS:000228436000008. PMID: 15833118
40. Adachi M, Cavalcanti ARO. Tandem stop codons in ciliates that reassign stop codons. *J Mol Evol.* 2009; 68(4):424–31. <https://doi.org/10.1007/s00239-009-9220-y> WOS:000265145500011. PMID: 19294453
41. Major LL, Edgar TD, Yip PY, Isaksson LA, Tate WP. Tandem termination signals: myth or reality? *FEBS Lett.* 2002; 514(1):84–9. [https://doi.org/10.1016/s0014-5793\(02\)02301-3](https://doi.org/10.1016/s0014-5793(02)02301-3) WOS:000174640300017. PMID: 11904187
42. Wei YL, Xia XH. The role of +4U as an extended translation termination signal in bacteria. *Genetics.* 2017; 205(2):539–49. <https://doi.org/10.1534/genetics.116.193961> WOS:000394144900007. PMID: 27903612
43. Brown CM, Tate WP. Direct recognition of messenger-RNA stop signals by *Escherichia coli* polypeptide-chain release factor-2. *J Biol Chem.* 1994; 269(52):33164–70. WOS:A1994QA63800066. PMID: 7806547
44. Poole ES, Major LL, Mannering SA, Tate WP. Translational termination in *Escherichia coli*: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. *Nucleic Acids Res.* 1998; 26(4):954–60. <https://doi.org/10.1093/nar/26.4.954> WOS:000072238300012. PMID: 9461453
45. Tate WP, Cridge AG, Brown CM. 'Stop' in protein synthesis is modulated with exquisite subtlety by an extended RNA translation signal. *Biochem Soc Trans.* 2018; 46:1615–25. <https://doi.org/10.1042/BST20180190> WOS:000453394200020. PMID: 30420414
46. Capecci MR. Polypeptide chain termination in vitro—isolation of a release factor. *Proc Natl Acad Sci USA.* 1967; 58(3):1144–51. <https://doi.org/10.1073/pnas.58.3.1144> WOS:A19679929200055. PMID: 5233840

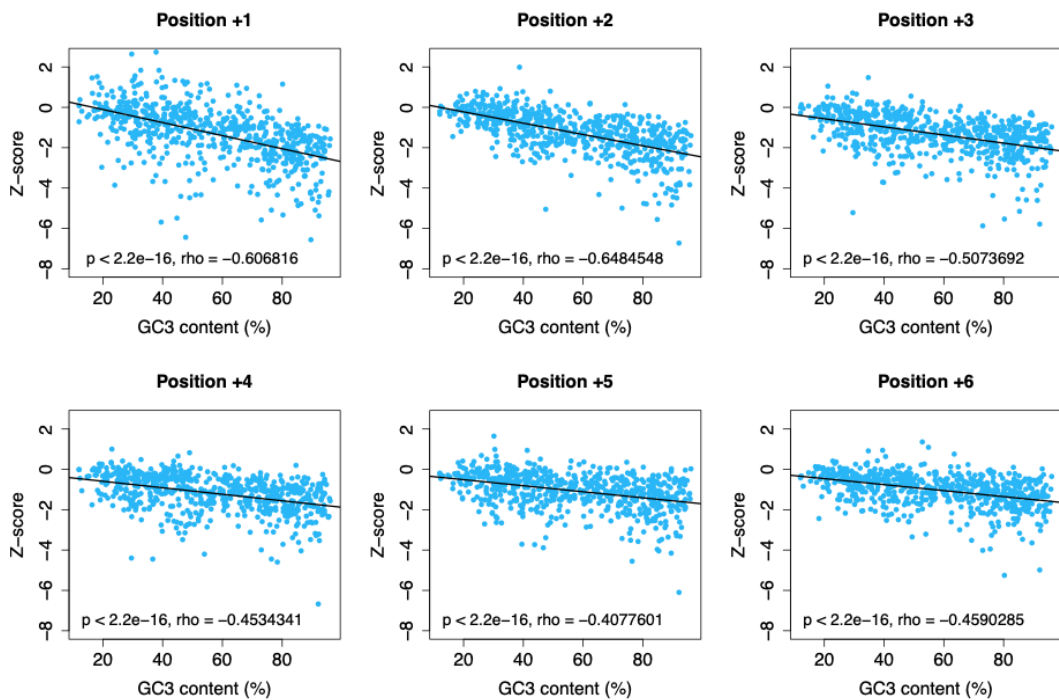
47. Caskey CT, Tompkins R, Scolnick E, Caryk T, Nirenberg M. Sequential translation of trinucleotide codons for initiation and termination of protein synthesis. *Science*. 1968; 162(3849):135–8. <https://doi.org/10.1126/science.162.3849.135> WOS:A1968B850100026. PMID: 4877370
48. Scolnick E, Tompkins R, Caskey T, Nirenberg M. Release factors differing in specificity for terminator codons. *Proc Natl Acad Sci USA*. 1968; 61(2):768–74. <https://doi.org/10.1073/pnas.61.2.768> WOS:A1968C065000064. PMID: 4879404
49. Petry S, Brodersen DE, Murphy FV, Dunham CM, Selmer M, Tarry MJ, et al. Crystal structures of the ribosome in complex with release factors RF1 and RF2 bound to a cognate stop codon. *Cell*. 2005; 123(7):1255–66. <https://doi.org/10.1016/j.cell.2005.09.039> WOS:000234584500014. PMID: 16377566
50. Freistoffer DV, Pavlov MY, MacDougall J, Buckingham RH, Ehrenberg M. Release factor RF3 in E-coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. *EMBO J*. 1997; 16(13):4126–33. <https://doi.org/10.1093/emboj/16.13.4126> WOS:A1997XJ99000037. PMID: 9233821
51. Milman G, Goldstein J, Scolnick E, Caskey T. Peptide chain termination, 3. Stimulation of in vitro termination. *Proc Natl Acad Sci USA*. 1969; 63(1):183–90. <https://doi.org/10.1073/pnas.63.1.183> WOS:A1969D694400031. PMID: 4897024
52. Scolnick EM, Caskey CT. Peptide chain termination, 5. Role of release factors in mRNA terminator codon recognition. *Proc Natl Acad Sci USA*. 1969; 64(4):1235–41. <https://doi.org/10.1073/pnas.64.4.1235> WOS:A1969F268900015. PMID: 4916922
53. Povlotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. Stop codons in bacteria are not selectively equivalent. *Biol Direct*. 2012; 7(1):30–30. <https://doi.org/10.1186/1745-6150-7-30> WOS:000315725200001. PMID: 22974057
54. Poole ES, Brown CM, Tate WP. The identity of the base following the stop codon determines the efficiency of in-vivo translational termination in Escherichia-coli. *EMBO J*. 1995; 14(1):151–8. WOS:A1995QB06100017. PMID: 7828587
55. 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> WOS:000426293300033. PMID: 29325104
56. Namy O, Hatin I, Rousset JP. Impact of the six nucleotides downstream of the stop codon on translation termination. *Embo Reports*. 2001; 2(9):787–93. <https://doi.org/10.1093/embo-reports/kve176> WOS:000171287400013. PMID: 11520858
57. Sharp PM, Bulmer M. Selective differences among translation termination codons. *Gene*. 1988; 63(1):141–5. [https://doi.org/10.1016/0378-1119\(88\)90553-7](https://doi.org/10.1016/0378-1119(88)90553-7) WOS:A1988M52000014. PMID: 3133285
58. Vakhrusheva AA, Kazanov MD, Mironov AA, Bazykin GA. Evolution of prokaryotic genes by shift of stop codons. *J Mol Evol*. 2011; 72(2):138–46. <https://doi.org/10.1007/s00239-010-9408-1> WOS:000288808400002. PMID: 21082168
59. Lindeboom RGH, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet*. 2016; 48(10):1112–8. <https://doi.org/10.1038/ng.3664> WOS:000384391600006. PMID: 27618451
60. Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, Hagemeier C, et al. Binary specification of nonsense codons by splicing and cytoplasmic translation. *EMBO J*. 1998; 17(12):3484–94. <https://doi.org/10.1093/emboj/17.12.3484> WOS:000074363800026. PMID: 9628884
61. Zhang J, Sun XL, Qian YM, LaDuca JP, Maquet LE. At least one intron is required for the nonsense-mediated decay of triphosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol*. 1998; 18(9):5272–83. <https://doi.org/10.1128/mcb.18.9.5272> WOS:000075484300035. PMID: 9710612
62. True HL, Lindquist SL. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*. 2000; 407(6803):477–83. <https://doi.org/10.1038/35035005> WOS:000089727400040. PMID: 11028992
63. Wickner RB, Masison DC, Edsles HK. PSI and URE3 as yeast prions. *Yeast*. 1995; 11(16):1671–85. <https://doi.org/10.1002/yea.320111609> WOS:A1995TQ24100007. PMID: 8720070
64. Harrison PM. Evolutionary behaviour of bacterial prion-like proteins. *Plos One*. 2019; 14(3). <https://doi.org/10.1371/journal.pone.0213030> WOS:000460371800018. PMID: 30835736
65. Angarica VE, Ventura S, Sancho J. Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains. *BMC Genomics*. 2013; 14. <https://doi.org/10.1186/1471-2164-14-316> WOS:000318944100001. PMID: 23663289
66. Yuan AH, Hochschild A. A bacterial global regulator forms a prion. *Science*. 2017; 355(6321):198–201. <https://doi.org/10.1126/science.aai7776> WOS:000391743700047. PMID: 28082594

67. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 2010; 6(9):e1001115. e1001115 <https://doi.org/10.1371/journal.pgen.1001115> ISI:000282369200053. PMID: 20838599
68. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 2010; 6(9):e1001107. e1001107 <https://doi.org/10.1371/journal.pgen.1001107> ISI:000282369200045. PMID: 20838593
69. 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> WOS:000360585900015. PMID: 25771198
70. Warnecke T, Huang Y, Przytycka TM, Hurst LD. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. *Genome Biol Evol.* 2010; 2(1):636–45. <https://doi.org/10.1093/gbe/evq049> WOS:000291467300007. PMID: 20688751
71. 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> WOS:000362503900007. PMID: 25656970
72. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python 2001. Available from: <http://www.scipy.org/>.
73. Belinky F, Rogozin IB, Koonin EV. Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Sci Rep.* 2017; 7. <https://doi.org/10.1038/s41598-017-12619-6> WOS:000412032000031. PMID: 28963504
74. 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 USA.* 2016; 113(46):13109–13. <https://doi.org/10.1073/pnas.1615832113> WOS:000388970100068. PMID: 27799560
75. Kristensen DM, Wolf YI, Koonin EV. ATGC database and ATGC-COGs: an updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation. *Nucleic Acids Res.* 2017; 45(D1):D210–D8. <https://doi.org/10.1093/nar/gkw934> WOS:000396575500032. PMID: 28053163
76. Tatusova T, DiCuccio M, Badretdin A, Chelvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 2016; 44(14):6614–24. <https://doi.org/10.1093/nar/gkw569> WOS:000382999900013. PMID: 27342282
77. Katoh K, Kuma K-I, Miyata T, Toh H. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome informatics International Conference on Genome Informatics.* 2005; 16(1):22–33. MEDLINE:16362903. PMID: 16362903

Supplementary information for: In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons

Alexander T. Ho and Laurence D. Hurst
PLoS Genetics, 15(9): e1008386.

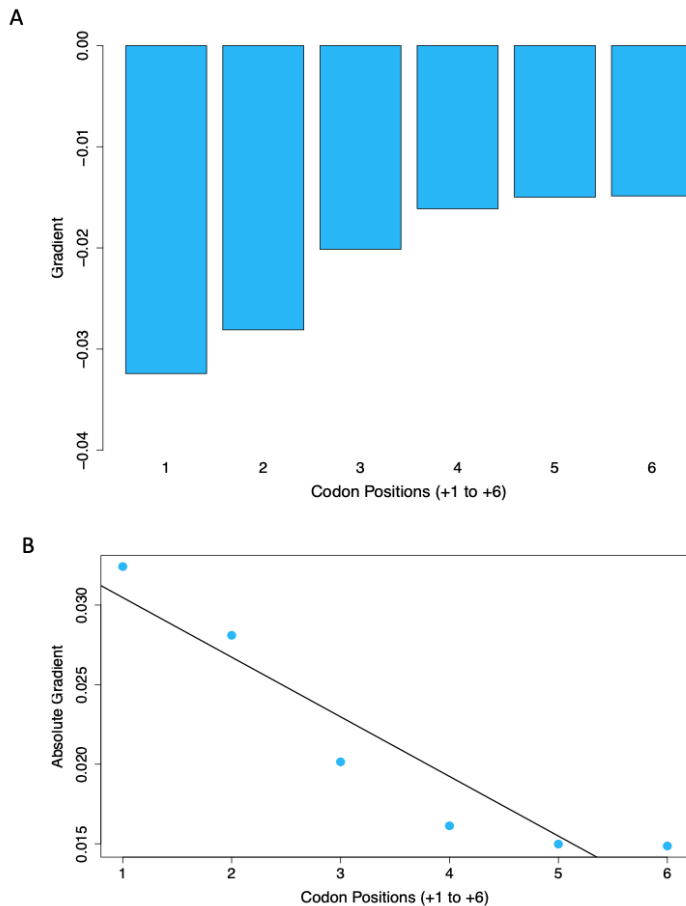
Note: Some of the supplementary figures and tables are extremely small when viewed in this document. To view these best, please refer to the online source: <https://doi.org/10.1371/journal.pgen.1008386>.



S1 Fig. Z-scores measuring deviation from a null model (10,000 simulations) plotted against the genomic GC3 content of filtered TT11 bacterial genomes.

Significant negative relationships were observed between Z-score and genomic GC3 content at each position (Spearman's rank: $p < 2.2 \times 10^{-16}$ for all positions; $\rho = -0.61$ at position +1, $\rho = -0.65$ at position +2, $\rho = -0.51$ at position +3, $\rho = -0.45$ at position +4, $\rho = -0.41$ at position +5, $\rho = -0.46$ at position +6).

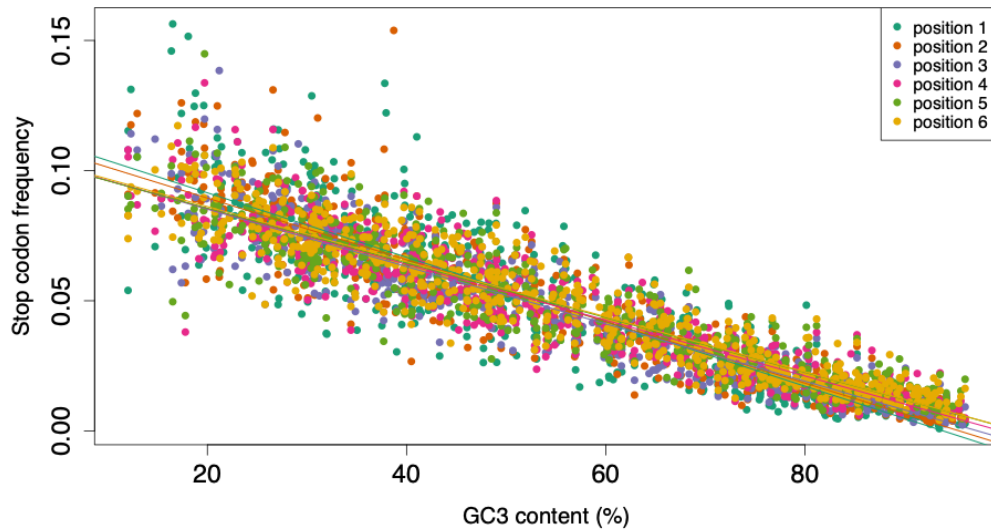
<https://doi.org/10.1371/journal.pgen.1008386.s001>



S2 Fig. Gradient analysis of Z-score plotted against genomic GC3.

Raw gradients (of Z-score plotted against genomic GC3) plotted at each of the six positions downstream (A). Absolute gradients plotted against codon position (B). Our expectation under the fail-safe hypothesis is that at codon position +1, stops will be largely resistant to GC pressure while at position +6 this resilience will be diminished. We thus predict that looking across genomes, the plot of ASC usage against GC content should be flatter at site +1 than at site +6. Interestingly, there is a significant correlation between absolute gradient and distance from the primary stop (Spearman's rank: $p = 2.8 \times 10^{-3}$ $\rho = -1$). We therefore infer that either the presence of ASCs is more resilient to GC pressure when located further downstream relative to the primary stop, or ASCs are actively selected against at positions closest to the primary stop. Both of these inferences go against the fail-safe hypothesis.

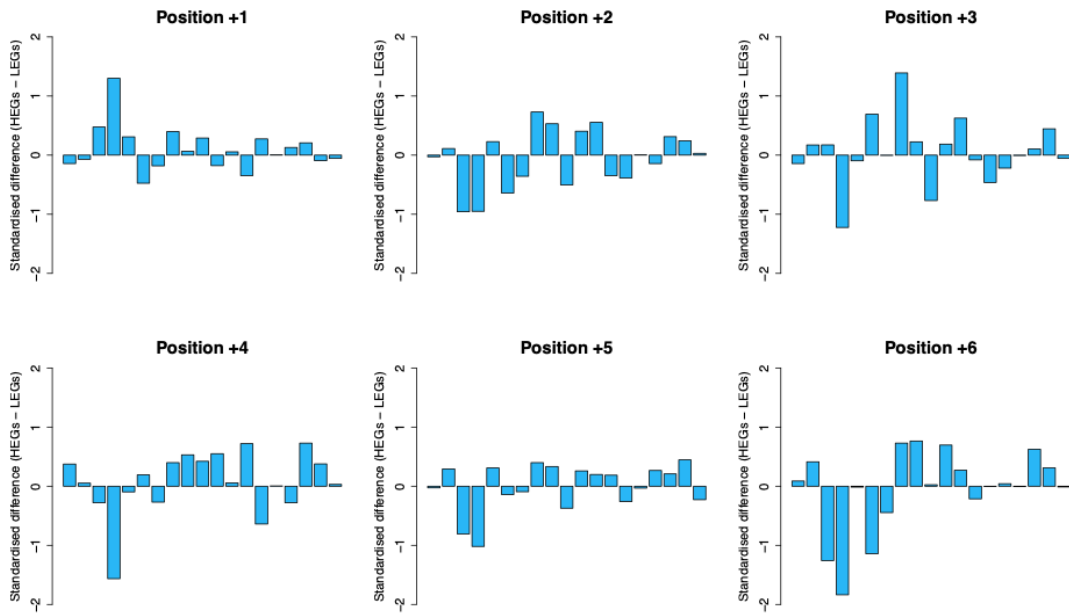
<https://doi.org/10.1371/journal.pgen.1008386.s002>



S3 Fig. Raw ASC frequencies at positions +1 to +6 plotted against genomic GC3 content.

There is a significant negative correlation between the variables at all positions tested (Spearman's rank: $p < 2.2 \times 10^{-16}$ for all positions; $\rho = -0.92$ for position +1, $\rho = -0.95$ for position +2, $\rho = -0.95$ for position +3, $\rho = -0.94$ for position +4, $\rho = -0.93$ for position +5, $\rho = -0.94$ for position +6). The gradient of the linear model fitted for position +1 is significantly different than that of position +6 ($p = 2.035732 \times 10^{-10}$), with position +6 having the more negative gradient.

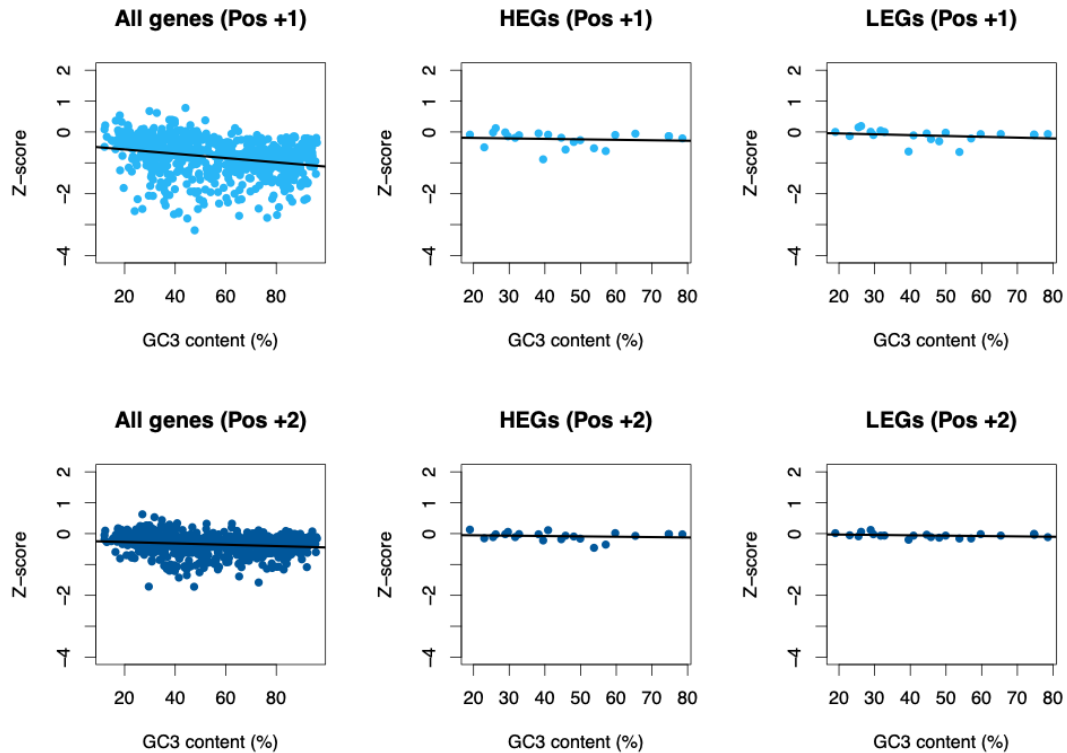
<https://doi.org/10.1371/journal.pgen.1008386.s003>



S4 Fig. Gradient frequencies of HEGs compared to LEGs.

Each bar represents one genome, with genomes ordered by genomic GC3 content from left to right. Bar heights represents the raw difference between in ASC frequency between the two groups tested. A positive difference represents enrichment in the HEGs group, a negative difference represents enrichment in the LEGs group. There is a no significant difference between HEGs and LEGs at any position (Wilcoxon signed-rank test: $p > 0.05/6$).

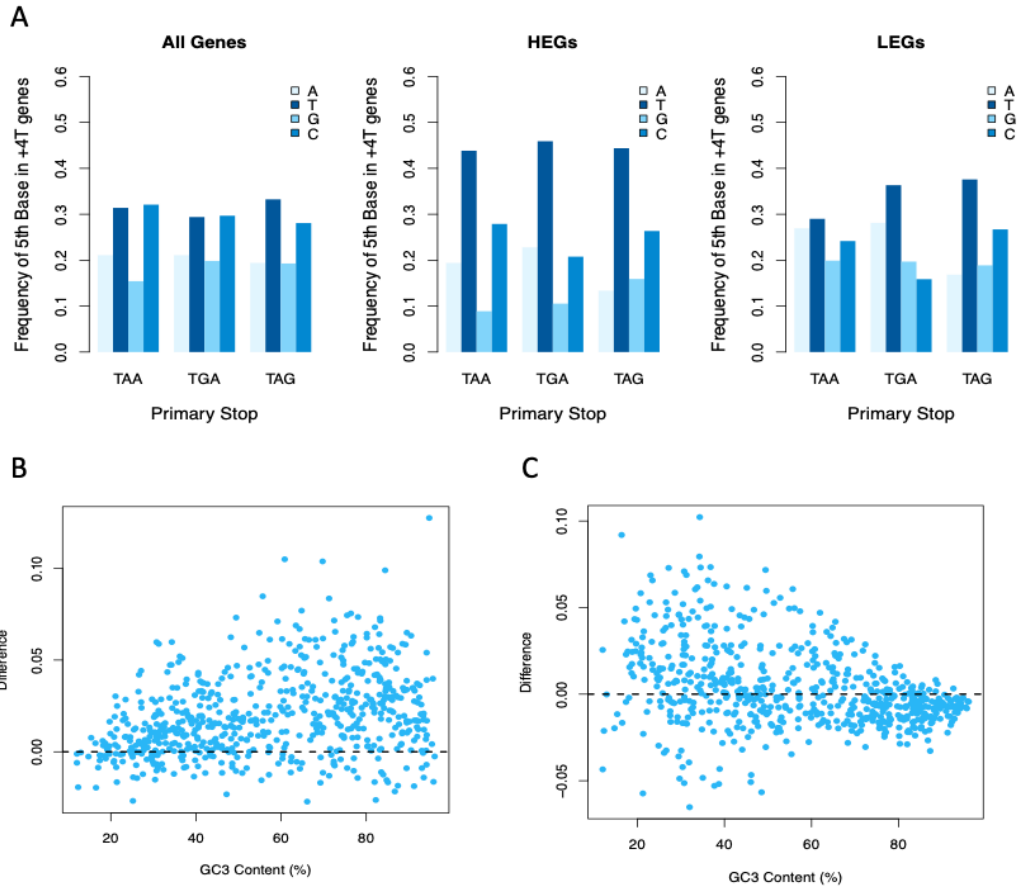
<https://doi.org/10.1371/journal.pgen.1008386.s004>



S5 Fig. Z-scores, measuring deviation of observed ASC frequencies in +4T-containing genes from +4T-containing null simulations, plotted against genomic GC3 content.

Only position +1 and position +2 are considered as these are the only sites where a signal for ASC enrichment has been noted. We find Z-scores to be negatively correlated with genomic GC3 when considering all genes at position +1 (Spearman's rank: $\rho = -0.3054869$, $p < 2.2 \times 10^{-16}$) and position +2 (Spearman's rank: $\rho = -0.1880088$, $p = 1.62 \times 10^{-06}$). There is no relationship between Z-score and genomic GC3 in HEGs or LEGs at either position (Spearman's rank, $p > 0.05$).

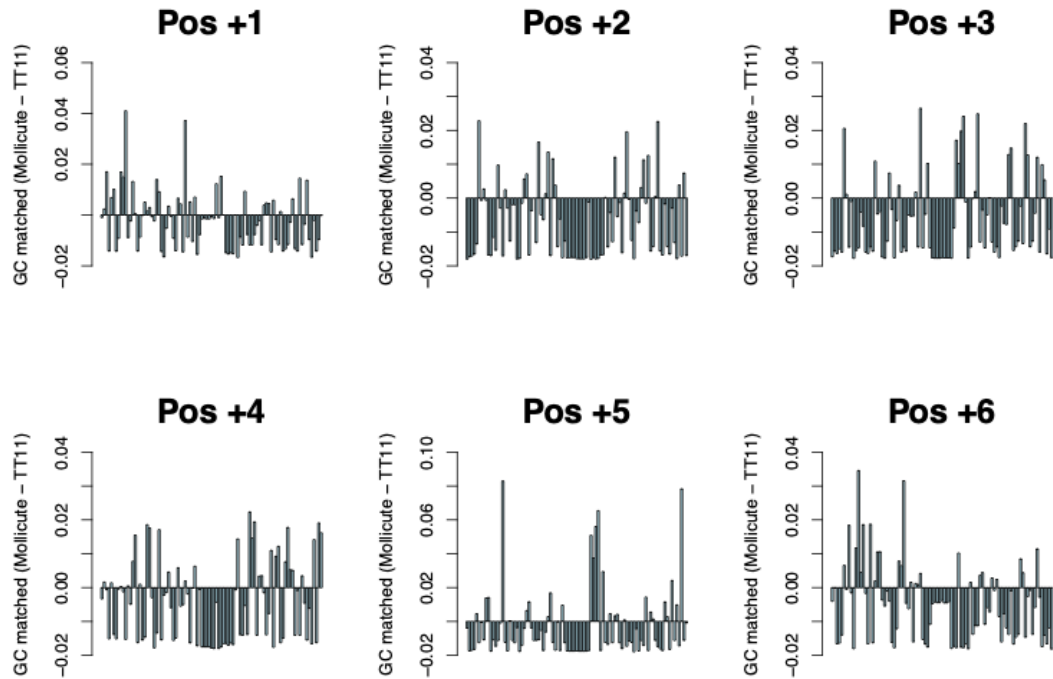
<https://doi.org/10.1371/journal.pgen.1008386.s005>



S6 Fig. Assessment of fifth site nucleotide preferences.

Fifth site nucleotide frequencies in +4T-containing genes of different primary stop and expression level (A). Frequencies of TC (B) and TT-starting (C) codons at position +1 compared to the average frequency of the respective codons between positions +1 to +6. Positive scores represent enrichment whilst negative scores represent under-representation.

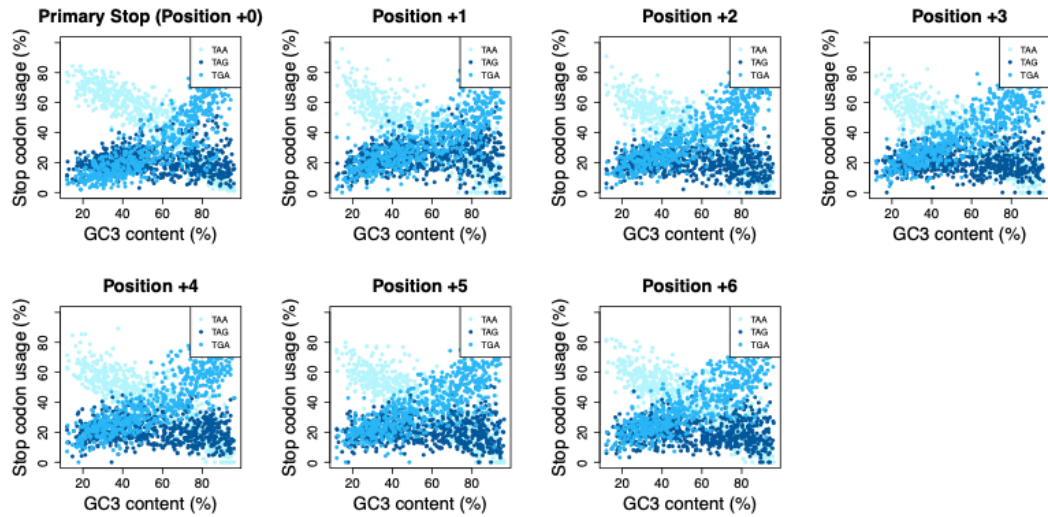
<https://doi.org/10.1371/journal.pgen.1008386.s006>



S7 Fig. ASC frequencies of TT4 mollicute genomes calculated and compared to those of GC-matched TT11 genomes.

Each bar represents the frequency difference between a mollicute genome and the average of its GC-matched TT11 genomes. TGA was underrepresented at positions +3 and +5 only (Wilcoxon signed-rank tests: $p = 0.11$ for position +1; $p = 0.15$ for position +2; $p = 1.5 \times 10^{-3}$ for position +3; $p = 0.70$ for position +4; $p = 6.8 \times 10^{-4}$ for position +5; $p = 0.11$ for position +6).

<https://doi.org/10.1371/journal.pgen.1008386.s007>



S8 Fig. Relative usage of each stop codon as the primary stop (position +0) and as ASCs at positions +1 to +6.

Contra to our expectations, we find codon usage at positions +1 to +6 to be consistent with that of the primary stop.

<https://doi.org/10.1371/journal.pgen.1008386.s008>

S1 Table. Positional codon switch (from stop to non-stop and non-stop to stop) counts and frequencies compared between the in-frame and out-of-frame 3' UTR codons of 29 triplets of closely related bacterial genomes.

<https://doi.org/10.1371/journal.pgen.1008386.s009>

Position	Switch	In-frame			+1 Frame-shift			Chisq p-val
		Ancestral codons	Switch count	Switch frequency	Ancestral codons	Switch count	Switch frequency	
+1	Stop > NS	679	139	0.205	591	132	0.223	0.313
	NS > Stop	11,254	152	0.0135	10,911	118	0.0108	0.00573
+2	Stop > NS	504	65	0.129	634	94	0.148	0.267
	NS > Stop	8,049	47	0.00584	7,979	45	0.00564	0.812
+3	Stop > NS	494	118	0.239	600	120	0.200	0.0534
	NS > Stop	6,276	49	0.00781	6,235	54	0.00866	0.468
+4	Stop > NS	526	139	0.264	583	123	0.211	0.00783
	NS > Stop	5,332	43	0.00806	5,304	41	0.00773	0.781
+5	Stop > NS	485	111	0.229	586	121	0.206	0.267
	NS > Stop	4,533	31	0.00684	4,476	48	0.0107	0.0120
+6	Stop > NS	507	120	0.234	-	-	-	-
	NS > Stop	3,847	33	0.00858	-	-	-	-

S2 Table. Codon switch (from stop to non-stop and non-stop to stop) counts and frequencies compared between the in-frame and out-of-frame 3' UTR codons of a triplet of TT4 mollicute genomes.

<https://doi.org/10.1371/journal.pgen.1008386.s010>

Switch	In-frame			+1 Frame-shift			Chi-squared p-val
	Ancestral codons	Switch count	Switch frequency	Ancestral codons	Switch count	Switch frequency	
TGA > other	2	1	0.5	3	1	0.333	0.683
other > TGA	314	3	0.00955	312	0	0	N/A
TAA > other	27	6	0.222	38	11	0.289	0.672
other > TAA	300	9	0.03	294	7	0.0238	0.487
TAG > other	8	1	0.125	4	2	0.5	0.134
other > TAG	307	2	0.00651	315	3	0.00952	0.589

S3 Table. 3' UTR frequencies of all codons in TT4 mollicutes compared to prediction by LOESS model fitted to TT11 codon frequencies.

Observed frequencies were compared to predicted frequencies using one-tailed Wilcoxon-signed rank tests, the p-values from which are found in the table below. A significant p-value represents significant under-enrichment in the TT4 genomes. Stop codons are highlighted in blue.

<https://doi.org/10.1371/journal.pgen.1008386.s011>

	+0	+1	+2	+3	+4	+5	+6
tga	3.81E-06	1.52E-01	9.82E-02	9.65E-04	7.91E-01	3.36E-04	7.70E-02
taa	9.78E-01	7.39E-01	3.00E-02	9.02E-01	2.34E-01	8.68E-01	3.51E-01
tag	1.00E+00	2.90E-01	9.35E-01	4.66E-01	5.34E-01	9.73E-01	8.94E-01
ttt	1.00E+00	6.80E-01	4.48E-03	7.66E-01	9.16E-01	8.37E-02	1.23E-01
ttc	1.00E+00	9.46E-01	9.98E-01	3.00E-02	4.94E-02	6.65E-01	6.01E-01
tta	1.00E+00	9.98E-01	1.00E+00	7.66E-01	6.65E-01	2.75E-01	4.49E-01
ttg	1.00E+00	1.23E-01	8.38E-01	2.69E-02	6.33E-01	1.96E-01	3.83E-01
tct	1.00E+00	6.65E-01	4.49E-01	3.85E-03	3.05E-01	2.21E-01	9.23E-01
tcc	1.00E+00	9.73E-01	1.04E-02	1.52E-01	4.48E-03	7.70E-02	3.99E-01
tca	1.00E+00	6.80E-01	5.94E-02	2.16E-02	7.70E-02	4.66E-01	3.83E-01
tcg	1.00E+00	3.99E-01	2.80E-03	5.34E-05	1.34E-02	7.90E-04	1.34E-02
tat	1.00E+00	1.91E-05	3.81E-06	2.67E-05	2.00E-03	2.75E-01	2.34E-01
tac	1.00E+00	2.48E-01	9.09E-01	7.08E-02	1.68E-03	3.35E-01	9.51E-01
tgt	1.00E+00	1.52E-01	2.80E-03	6.94E-03	2.67E-05	5.34E-01	2.09E-01
tgc	1.00E+00	3.68E-02	4.66E-01	3.83E-01	7.97E-03	8.37E-02	7.52E-01
tgg	1.00E+00	7.25E-01	2.48E-01	4.16E-01	9.96E-01	1.34E-02	1.64E-04
ctt	1.00E+00	2.61E-01	7.91E-01	3.83E-01	5.67E-01	1.34E-02	9.85E-01
ctc	1.00E+00	9.67E-01	5.67E-01	1.32E-01	8.77E-01	3.20E-01	7.10E-01
cta	1.00E+00	7.39E-01	6.80E-01	7.70E-02	9.96E-01	9.76E-01	6.33E-01
ctg	1.00E+00	8.04E-01	1.14E-01	7.08E-02	6.49E-01	7.10E-01	1.34E-02
cct	1.00E+00	8.04E-01	3.51E-01	6.95E-01	3.00E-02	3.51E-01	7.97E-03
ccc	1.00E+00	7.63E-06	6.49E-02	1.18E-02	9.87E-01	2.69E-02	8.37E-02
cca	1.00E+00	7.25E-01	9.67E-01	4.94E-02	9.94E-01	2.90E-01	8.68E-01
ccg	1.00E+00	3.81E-06	1.34E-02	1.34E-02	7.63E-06	9.65E-04	3.81E-06
cat	1.00E+00	2.75E-01	4.48E-03	4.07E-02	8.37E-02	6.65E-01	4.66E-01
cac	1.00E+00	4.16E-01	1.85E-01	4.94E-02	2.69E-02	9.02E-01	1.32E-01
caa	1.00E+00	3.20E-01	9.81E-01	9.97E-01	9.87E-01	9.91E-01	5.94E-02
cag	1.00E+00	7.08E-02	9.63E-01	2.61E-01	4.16E-01	8.38E-01	1.52E-01
cgt	1.00E+00	2.69E-02	1.85E-01	3.51E-01	6.49E-02	1.17E-03	9.65E-04

S4 Table. Association between relative stop codon usage and GC3 content assessed by Spearman's Rank tests at each downstream position.

<https://doi.org/10.1371/journal.pgen.1008386.s012>

Position	TAA		TGA		TAG	
	Rho	P-value	Rho	P-value	Rho	P-value
+0 (primary)	-0.92	< 2.2e-16	0.88	< 2.2e-16	-0.017	0.68
+1	-0.86	< 2.2e-16	0.81	< 2.2e-16	0.25	1.4e-10
+2	-0.88	< 2.2e-16	0.87	< 2.2e-16	-0.098	0.013
+3	-0.87	< 2.2e-16	0.86	< 2.2e-16	-0.044	0.26
+4	-0.87	< 2.2e-16	0.87	< 2.2e-16	-0.080	0.043
+5	-0.87	< 2.2e-16	0.84	< 2.2e-16	-0.065	0.099
+6	-0.85	< 2.2e-16	0.85	< 2.2e-16	-0.094	0.017

S5 Table. Association between relative stop codon usage and GC3 content assessed by Spearman's Rank tests in each reading frame.

<https://doi.org/10.1371/journal.pgen.1008386.s013>

Reading Frame	TAA		TGA		TAG	
	Rho	P-value	Rho	P-value	Rho	P-value
+0 (in-frame)	-0.91	< 2.2e-16	0.90	< 2.2e-16	-0.038	0.33
+1 Frame-shift	-0.91	< 2.2e-16	0.89	< 2.2e-16	0.0032	0.93
+2 Frame-shift	-0.91	< 2.2e-16	0.90	< 2.2e-16	-0.12	0.0021

S6 Table. Association between gene length and gene expression level assessed at genome-wide level.

Gene expression (represented by experimental protein abundance data) and gene nucleotide lengths were used in Spearman's rank tests.

<https://doi.org/10.1371/journal.pgen.1008386.s014>

Accession	Rho	P-value	GC	GC3
AE000511	0.07	0.069	38.9	40.9
AE002098	-0.19	3.4 x 10 ⁻⁰⁴	51.5	59.7
AE004092	-0.20	1.6 x 10 ⁻⁰⁷	38.5	31.7
AE005176	-0.22	9.2 x 10 ⁻¹¹	35.3	25.5
AE006468	-0.12	8.3 x 10 ⁻⁰⁷	52.2	57.0
AE014299	-0.19	6.2 x 10 ⁻⁰⁹	46.0	45.8
AE015928	-0.35	5.3 x 10 ⁻¹⁵	42.8	44.6
AE016823	-0.19	3.1 x 10 ⁻¹²	35.0	29.6
AE017225	-0.15	8.0 x 10 ⁻⁰⁶	35.4	26.2
AE017285	0.12	2.8 x 10 ⁻⁰³	63.1	74.8
AE017354	-0.28	1.2 x 10 ⁻⁰⁸	38.3	32.8
AL111168	-0.18	0.021	30.5	19.0
AL123456	-0.03	0.217	65.6	78.5
AL590842	-0.07	0.034	47.6	48.1
AP009552	-0.07	1.5 x 10 ⁻⁰⁵	42.3	39.5
AP012205	-0.32	1.4 x 10 ⁻²³	47.7	49.9
BA000017	-0.10	1.4 x 10 ⁻⁰³	32.9	23.0
BX897699	-0.03	0.362	38.2	28.9
CP001114	-0.02	0.580	63.4	74.7
CP001219	0.16	4.2 x 10 ⁻⁰⁴	58.8	65.4
CP002077	-0.04	0.828	40.0	38.2
U00096	0.02	0.281	50.8	53.8

S7 Table. Z-score analysis of single-celled eukaryotic genomes.

Z-scores represent deviation in ASC frequency from dinucleotide-controlled simulations.

<https://doi.org/10.1371/journal.pgen.1008386.s015>

Accession	P1_Zscore	P2_Zscore	P3_Zscore	P4_Zscore	P5_Zscore	P6_Zscore
Stentor	17.214486 2	16.300777 7	31.195794 8	33.158078	11.438646 1	22.711428 6
Thecamonas	0.2485532	-0.5486816	-0.3570446	1.666992 4	-1.0994509	-1.9425618
Monosiga	-5.778071	-3.0600379	-0.1587398	-2.8215427	-2.3994836	-3.4617629
Leptomonas	-6.9867224	-5.8914975	-4.9573856	-5.2714793	-6.4365065	-4.3273514
Spironucleus	0.2556468 4	-2.0934388	0.1984575 8	-2.4697767	-5.0593736	-1.0689865
Thalassiosira	-7.2106504	-2.7773431	-1.5170654	-3.2354253	-2.0603369	-6.1148792
Fragilariopsis	12.165440 9	5.6689043 5	-0.7344367	-0.2325782	-6.5846695	-5.5514482
Tetrahymena	-9.5355621	-9.5224858	16.694526 1	14.168474	7.7138865 4	6.4075122
Eimeria	-0.64554	-1.3912806	-1.3734585	-1.0877742	-1.3231437	-1.5382534
Hondaea	-2.1474402	2.7466343 1	-0.8782693	-0.606092	-1.4642002	-2.4275714
Thraustotheca	4.0855874 1	-0.104448	4.5030245 5	2.5201426 9	2.8987120 1	1.6965946 2
Hyaloperonospora	-7.9905895	-6.9140936	-4.9420549	-3.4496433	-2.8538775	-3.5711925
Giardia	-1.3493263	0.9734487 1	-0.104576	3.8635684 1	-0.0858364	-0.9747457
Leishmania	-2.7919178	-4.4033335	-4.8610368	-3.3408571	-4.0266067	-2.6465693
Emiliana	-19.036334	-3.1636671	-12.59843	-10.463621	-10.484026	-11.255094
Angomonas	0.9809295	0.9223442 5	-0.6203595	-1.7965483	-2.0369631	-0.51251
Phytophthora	-14.649737	-8.7220324	2.1604806 6	-2.9762246	-7.6872962	-3.4788864
Tieghemostelium	3.0072314 7	7.7130178 9	2.8162430 3	2.1479780 1	-0.9994664	-1.3822488
Candida	0.3013322	2.7977670 2	4.4653262 7	0.5053384 7	0.3033465 7	0.0721173 1
Cyclospora	-3.5379054	-4.646197	-4.0069516	-3.7283106	-0.9387594	-2.6259193
Chlamydomonas	1.9616005 5	-6.7150778	-3.0238859	2.0577938 4	-2.6385147	-1.9300849
Sphaeroforma	-6.4164981	-3.3318298	-3.0061861	-1.9853232	-1.4030009	-1.6683429
Cryptosporidium	-1.7576225	0.1985980 8	0.1362086 1	1.1690775 7	1.0070053 2	-0.8927566
Dictyostelium	-2.5746227	8.8283988 5	10.559259 4	7.6467870 7	5.7821215 2	2.2293688 4
Albugo	-2.6789789	-1.6038712	-1.9638053	-2.093771	-3.2618513	-3.765025
Carpediemonas	-0.0533575	-0.0122037	-0.0147494	0.1074201 2	0.1080163 9	0.0666392 8
Nothophytophthora	-21.162682 a	-22.413307	-13.073031	-17.439237	-5.9900116	-13.804912

S8 Table. Chi2 analysis of ASC frequency single-celled eukaryotic genomes against dinucleotide-controlled simulations.

<https://doi.org/10.1371/journal.pgen.1008386.s016>

Genus	+1_chi	+1_p	+2_chi	+2_p	+3_chi	+3_p	+4_chi	+4_p	+5_chi	+5_p	+6_chi	+6_p
Thelliera	0.53488236	0.46456141	6.726688	0.00948808	4.89309955	0.02869424	5.90521189	0.01509615	12.0110903	0.00052885	6.726688	0.00948808
Phytophthora	89.0436018	3.86E-21	26.360742	2.83E-07	0.57181939	0.44953884	0.36072475	0.548104	8.12758095	0.00439969	0.13612882	0.71216022
Pseudoactinium	43.102845	5.19E-11	4.8692567	0.02733928	18.1016127	5.72E-05	6.70406647	0.00953933	20.740041	5.29E-06	12.4750041	0.00042943
Leptomonas	41.1649291	1.40E-10	26.7508252	2.31E-07	28.8491365	7.82E-08	30.4748562	3.38E-08	51.2887158	7.93E-13	16.224645	5.63E-05
Hyalopromonas	37.1185819	1.11E-09	32.4627975	3.21E-08	13.2801328	0.00202683	7.48540842	0.0062201	4.10122295	0.04285221	1.62526534	0.20235917
Toxoplasma	11.562398	0.00067851	10.175854	0.0014282	7.0962027	0.00772549	2.9971693	0.1070555	20.743171	5.17E-06	13.876184	8.48E-05
Reticulicoccus	52.7151146	3.86E-13	30.5603013	3.24E-08	57.6964438	3.06E-14	60.7600425	6.63E-15	75.4886025	3.77E-18	64.6879638	8.77E-16
Perkinsus	36.8142532	1.30E-09	0.75976592	0.38340179	1.84304886	0.17455946	0.46887996	0.49350473	0.5976488	0.43947652	0.99472277	0.31859882
Ichthyophthirius	18.9694677	1.33E-05	25.1639731	5.27E-07	0.07805853	0.77999274	5.347137	0.0175709	11.855383	0.00058788	19.9412044	7.99E-06
Chlamydomonas	9.45049096	0.00211093	6.6105106	0.01013763	3.7393421	0.05314625	0.02575933	0.87250274	1.14218483	0.00029114	0.15063086	0.69793321
Aureococcus	2.645208	0.10385668	1.0977975	0.29474996	6.43025954	0.01121941	36.0944176	1.88E-09	13.6393383	0.0022125	5.6190468	0.01776631
Strigomonas	3.18136224	0.07448261	3.18136224	0.07448261	2.46026713	0.1671367	4.89947647	0.02686884	4.89947647	0.02686884	3.9942962	0.04565834
Monocilia	37.5916249	8.72E-10	17.1511509	0.19032807	0.07402179	0.78548861	2.6079376	0.10636681	3.2323903	0.0211924	10.250468	0.00136661
Albugo	0.52757584	0.46762782	6.81173653	0.00905607	6.35158842	0.00172762	1.71442083	0.19041279	10.4832999	0.05678552	0.02437691	0.02437691
Phytomonas	0.87142369	0.35095147	18.9074657	1.37E-05	6.17855198	0.01293881	6.17855198	0.01293881	8.71028007	0.00318419	8.23821755	0.00409579
Oxytricha	0.06014946	0.80927907	4.33559072	0.03733688	1.3542457	0.24451225	0.04928936	0.4811351	0.06014046	0.80927407	21.2349306	4.06E-06
Entamoeba	33.2007346	8.31E-09	15.9400615	6.54E-05	1.44148439	0.22989929	3.1024696	0.07817362	1.4412382	0.2850845	14.4413282	0.0001446
Nannochloropsis	8.91847643	0.002823	0.31733957	0.57321087	0.18379652	0.66813067	0.69315133	0.40509456	0.00061866	0.3801564	3.64607364	0.05620156
Actina	31.63238	1.86E-08	9.0886256	0.00257189	15.3194067	9.08E-05	5.6843297	0.01711261	12.313569	0.0004967	24.047475	9.40E-07
Pseudoocochilium	0.00486505	0.94439273	5.32452683	0.0210732	6.37746517	0.01155785	0.23787741	0.62574338	3.18630541	0.07425766	0.5870687	0.443555
Ectocarpus	0.01969957	0.8883794	1.42663023	0.2323157	2.27551231	0.13143169	4.90673755	0.02675157	2.51855662	0.11251353	7.21572880	0.00722656
Styloncheia	0.63096602	0.42170019	17.6011937	2.12E-05	7.0185779	0.00808232	7.57631416	0.00591402	12.6882875	0.00051287	14.3120719	0.00015887
Candida	0.25834858	0.61125991	11.4125258	0.00072951	41.0208138	3.51E-10	1.94702828	0.02045246	0.88628061	0.0001831	0.9892039	0.0001831
Plasmodium	0.24772417	0.16188214	1.01216138	0.1343856	1.92456229	0.16534426	0.56533954	0.45211683	0.07011695	0.79116658	0.59774452	0.43943989
Angomonas	5.0537887	0.02457215	6.60190701	0.01018925	0.93046894	0.33489436	5.42948957	0.1979947	11.394221	0.00051122	1.45361323	0.22749096
Leishmania	3.8083388	0.00997951	12.508434	0.00040512	12.508434	0.00040512	8.56747046	0.0034222	12.8434818	0.00032105	7.8884964	0.00030395
Perkinsella	6.4718721	0.0109595	16.6859869	4.41E-05	1.71973233	0.18977213	2.62289887	0.10533179	3.25595691	0.07116439	0.9388147	0.33258286
Acanthamoeba	21.6185241	3.33E-06	12.6495372	0.00037566	25.9044775	5.59E-07	19.4773002	1.02E-05	51.3191315	8.8E-13	21.938882	2.82E-06
Giardia	0.9385797	0.33264372	5.83740966	0.0154889	1.60400993	0.20313671	31.4648933	2.03E-08	0.47485243	0.48912414	0.00557776	0.94046581
Cytospora	0.36195268	0.54742376	6.2221713	0.0261603	8.05670728	0.00453355	0.83174667	0.36176823	14.825465	0.00014374	11.0226803	0.00090004
Trypanosoma	0.58932587	0.43898815	5.64367557	0.01753844	7.29667321	0.00690822	0.12552301	0.72311986	3.63116186	0.05670722	12.2768766	0.00045861
Pythium	30.0032247	4.31E-08	11.2718669	0.00078921	8.44636204	0.00365775	0.04254699	0.8364657	1.7797018	0.18218504	0.11671937	0.7326202
Fragillipollis	88.2525559	5.76E-21	20.7368661	6.36E-06	1.1283049	0.28813622	6.05561338	0.01386216	0.92406363	0.33641033	3.6651522	0.0558177
Stramonocleus	0.07406639	0.7855056	11.8857447	0.00056565	0.71193982	0.39880059	8.71198001	0.00316124	21.2042466	4.13E-06	9.7124551	0.00187246
Notophytophthora	201.974094	7.75E-46	177.968877	1.34E-40	89.4348831	2.41E-30	96.49860378	8.93E-33	271.348495	1.90E-07	74.4617936	6.16E-18
Thyriostomella	6.7292362	0.00948087	32.5348276	1.17E-08	6.90644156	0.00085885	6.5559074	0.01045492	0.699801	0.04285057	0.0377873	0.84529242
Trichomonas	174.874253	6.54E-40	32.8651299	9.88E-09	16.1629663	5.81E-05	17.0939691	3.56E-05	5.43197559	0.0197713	12.6996315	0.0003673
Babesia	12.1632094	0.00048733	21.034594	4.51E-06	0.4746204	0.49174421	0.4305622	0.00180001	21.034594	1.51E-06	199.05063	3.37E-45
Aphanizomenon	4.55640957	0.0372976	0.01885649	0.89077871	0.25010117	0.61700244	0.57013871	0.45020579	0.09443215	0.7861664	6.7007035	0.40087885
Guillardia	24.6245748	6.97E-07	0.35545807	0.55103959	2.96117936	0.00828571	9.87908994	0.00167168	0.18554009	4.07433113	0.0433395	0.435395
Tetrahymena	28.4886276	9.52E-08	17.1373815	3.48E-05	123.853355	9.07E-29	44.717875	2.28E-11	4.21543795	0.04005727	5.7352081	0.01662809
Saprolegnia	19.036867	1.28E-05	71.4911897	3.69E-06	5.23324077	0.02719098	1.19973197	0.27327526	2.0214394	0.15502998	0.18748299	0.6650827
Salsipoea	2.63707292	0.10439573	4.7597402	0.02916336	3.53166682	0.06020723	1.78403067	0.18165516	0.00202201	0.89176164	1.33E-05	1.33E-05
Naegleria	0.0903997	0.76366663	3.31262369	0.06874965	11.7622024	0.00060446	3.10055061	0.07826582	14.0631328	0.0001822	0.92394333	0.33644171
Emetia	2.3531271	0.12496863	8.9120369	0.00283288	8.34310529	0.00387003	2.885024	0.08403807	5.78345124	0.01617775	7.9430607	0.00024111
Thraustotheca	0.00262331	0.95915162	0.61573362	0.43263713	17.1880967	3.39E-05	22.6297788	1.96E-06	5.92642878	0.0014544	5.3627598	0.01662818
Plasmodium	49.7897238	1.71E-12	25.9134912	3.57E-07	58.3955373	7.14E-14	65.6382545	5.42E-16	46.3141119	1.01E-11	65.6382545	5.42E-16
Sphaeroformia	41.4735983	1.20E-10	7.13682608	0.00755166	13.0580007	0.00020398	7.1682908	0.00755166	4.8418909	0.02777642	7.41911635	0.00045309
Capsaspora	6.84921288	0.00886093	15.1015092	0.00010188	21.3511506	1.68E-06	11.6864996	0.00062955	0.00129003	11.2343258	0.00080098	0.00080098
Thamnomonas	2.02620338	0.15460641	1.09125057	0.29619429	0.00025798	0.98718501	4.75712892	0.02917718	0.00843193	0.92683664	0.38709701	0.5338293
Planoprototellum	130.555327	3.13E-30	2.7438459	0.09752603	8.18413883	0.00024591	9.37968711	0.00196808	0.56228584	0.78515956	0.37565779	0.37565779
Steatorhiza	191.98706	1.17E-43	84.1654882	4.33E-20	285.500945	5.84E-43	163.468685	1.98E-37	89.1617589	3.84E-21	123.939121	8.89E-29
Blastocystis	8.78831211	0.00030167	2.94418093	0.08618738	2.022261	0.15500818	0.06094227	0.80501279	2.022261	0.15500818	4.65073551	0.03103992
Gregarina	21.1186377	4.32E-06	11.6512764	0.00064158	17.2408679	3.29E-05	4.21848042	0.03999114	0.2523262	0.81543738	1.1943495	0.27422706
Emilia	53.4787811	2.61E-13	3.11357448	0.07764239	12.0579931	0.00051571	3.9933748	0.04567952	3.9933748	0.04567952	18.0481311	2.15E-05
Hammondia	53.7439294	2.28E-13	24.9321844	0.0816278	19.8362112	8.44E-06	25.544028	4.35E-07	23.1686481	1.48E-06	34.720048	3.81E-09
Cryptosporidium	5.81513155	0.01582777	1.87818212	0.17053814	0.02105999	0.88461809	1.73517796	0.18775097	2.85683167	0.00909805	1.05478544	0.3044079
Thalassiozoa	3.9186607	0.0472767	0.0360395	0.8544945	1.3549874	0.2444785	0.03360505	0.8544945	0.5360723	0.4640651	0.9635389	0.32673717
Bigelovia	248.363402	4.37E-56	34.911169	3.45E-09	5.43651385	0.01971999	8.66430055	0.00273036	0.07329808	0.04914779	0.37953718	0.5378175
Hondaria	1.94793257	0.16280881	23.3216718	1.37E-06	0.66612364	0.41617379	2.98194477	0.12952448	0.48006645	0.48883922	2.99082839	0.14810956
Caenidictia	16.5452446	4.75E-05	67.578769	2.03E-16	11.54564	1.64E-26	193.294567	6.07E-44	8.7883266	6.28E-34	28.300689	2.31E-07
Saccharomyces	27.0760583	1.96E-07	0.59068705	0.44215356	20.1598022	7.12E-05	7.12481524	0.00760242	9.68059932	0.00183368	8.17509564	0.00424574
Peronospora	23.4913251	1.25E-06	34.2309517	4.89E-09	8.85545092	0.00292215	8.85545092	0.00292215	4.07033523	0.04364261	6.34775847	0.01173927
Besnoitia	48.803477	2.83E-12	22.919705	1.60E-06	2.2189225	0.1383809	8.650126	0.0032469	25.7726497	0.84E-07	9.3127728	0.00227336
Symbiodinium	84.283753	4.29E-20	0.56192616									

S9 Table. Chi2 analysis of ASC frequency single-celled eukaryotic genomes against the Adachi and Cavalcanti null.

<https://doi.org/10.1371/journal.pgen.1008386.s017>

Genus	+1_chi	+1_p	+2_chi	+2_p	+3_chi	+3_p	+4_chi	+4_p	+5_chi	+5_p	+6_chi	+6_p
Alveolates	9.83101861	0.00171933	1.13494401	0.28672323	0.04749952	0.82747286	0.34405332	0.55749886	0.33314478	0.56381317	1.35437074	0.24451547
Phytophthora	34.3962239	4.50E-09	1.45153216	0.22838711	26.0971565	3.71E-07	10.744279	0.00105169	0.16692375	0.74367512	3.55892207	0.05922819
Rhizaria	10.8304161	0.00099846	38.8733681	4.52E-10	0.95643445	0.3380685	0.00063907	0.97983171	0.02491588	0.52059355	0.47058669	0.00050682
Leptomonas	0.07718845	0.78114433	2.78809624	0.09437631	1.63859182	0.20051879	0.97959684	0.32229836	1.90791304	0.16719527	12.0940689	0.00050682
Hydrogomonaspora	12.4272669	0.0041388	10.6337578	0.00112042	1.8540355	0.17331477	0.05164833	0.83201948	0.63201814	0.8567376	0.78086003	0.3788402
Toxoplasma	0.52341301	0.46938941	0.38926712	0.53268489	0.01995965	0.88764882	1.95002275	0.1625844	6.12627932	0.01331869	3.54851508	0.05959888
Reticulomyxa	0.05313503	0.81739212	0.02033609	0.88654705	3.45981107	0.0628775	8.81109903	0.00299404	0.02733585	0.0031106	5.37505924	0.02042665
Perkinsus	95.9479348	1.8E-27	0.0016113	0.96748065	3.91453576	0.0478023	0.01083407	0.9170974	0.22373809	0.63621175	0.12142905	0.72748071
Ichthyophthirius	3.4100029	0.06480233	10.1286011	0.00145986	0.65747837	0.41745159	2.20832318	0.1372759	5.77176736	0.01628655	8.55183858	0.00345111
Chlamydomonas	26.2803593	2.95E-07	4.61240669	0.03174166	0.62731962	0.283402	0.11378836	0.73587169	0.13551126	0.71278485	2.1329545	0.1441588
Autrocooccus	3.8068518	0.05106668	0.97159462	0.32428281	7.47180993	0.00526277	36.9327225	1.22E-09	13.698491	0.00021462	3.8878891	0.04635952
Strimonas	0.44584033	0.50431096	0.21937272	0.63951829	0.01863784	0.80149083	0.48189904	0.60882127	0.83047113	0.82214616	0.42739843	0.51312292
Monosiga	19.7298571	8.92E-06	0.17371542	0.67683068	2.52452552	0.11208849	0.02382184	0.8773301	0.03295279	0.85592324	2.57449404	0.10859882
Albugo	4.2785205	0.03893727	0.12478996	0.72388882	1.09574443	0.29520393	1.4156095	0.2412741	1.99596273	0.15773882	0.01958805	0.88875018
Phytomonas	1.6453307	0.18959537	6.80597111	0.00908322	0.95640291	0.45317746	1.3081295	0.1649623	0.7531288	0.38391031	1.71641397	0.1901553
Oxytricha	0.13948974	0.7078788	2.47635969	0.11564946	0.06308117	0.80169067	0.24376503	0.63572526	0.42959640	0.51206659	0.93832029	0.58777224
Entamoeba	42.6538892	6.53E-11	26.4918916	2.65E-07	1.09590083	0.79516746	1.6023185	0.21161378	1.01915742	0.31273895	5.73197076	0.01665916
Nematochloasus	0.03712842	0.61513811	4.2944964	0.01460977	1.95189913	0.14238614	0.54573444	0.46062029	4.44338989	0.03038663	1.01617707	0.3137992
Achlya	14.1568171	0.00016819	0.98715973	0.32043755	37.3833372	9.88E-10	20.9033753	4.83E-06	19.5468097	8.82E-06	30.004148	4.31E-08
Pseudocnemelembus	1.45098207	0.22837044	2.17002203	0.14072422	4.23523897	0.03959305	0.84848416	0.35829851	8.042786	0.00456827	1.73139334	0.18928275
Ectocarpus	10.6752264	0.00150256	0.95344751	0.33831317	0.49487718	0.48367401	0.00716	0.95781329	0.13928262	0.7087805	1.61512288	0.16638425
Stylocheilichia	3.35127357	0.06715304	2.53768005	0.11116153	0.77707974	0.37803692	1.41013402	0.23503415	3.18179163	0.07446304	0.25615101	0.61266317
Candida	2.82673378	0.09270673	12.5391759	0.00039851	49.7355214	1.76E-12	0.00536888	0.94163259	0.80173941	0.34960576	0.55433838	0.00053838
Plasmodium	0.5303402	0.46464662	2.04708617	0.15250822	1.62280603	0.20273838	0.20102733	0.65410151	0.34600424	0.5563888	0.06543352	0.79751712
Angomonas	39.2569786	3.72E-10	12.7915181	0.00034819	0.01094644	0.91667319	1.6527131	0.19859007	5.41340089	0.01998275	0.76853666	0.38066176
Lentiniopsis	10.9341093	0.0009441	1.93446466	0.16339481	1.27886945	0.25792331	3.91106871	0.04786908	1.03949507	0.93793893	2.68988958	0.10099949
Perkinsella	0.03712842	0.84718847	4.2944964	0.02965047	1.19853156	0.27861597	1.12056202	0.38979738	0.59485403	0.44054821	12.8138723	0.00044024
Acanthamoeba	136.2819516	1.73E-31	0.0555427	0.81302019	2.93486119	0.08666611	1.81547121	0.1778533	0.0007855	0.55232028	0.45735081	0.00034528
Giarinia	0.1126696	0.72188264	2.75128753	0.00536743	2.33297633	1.16659548	6.89150705	0.00866062	11.0262345	0.93706055	0.74252671	0.38885588
Cystosporia	18.6514886	1.57E-05	0.0070745	0.93295021	0.46850881	0.48367401	2.38281781	0.12287609	3.11735753	0.07753945	0.75230004	0.8355395
Trypanosoma	26.6260511	2.47E-07	0.84106053	0.35909964	34.0912786	5.26E-09	4.79300772	0.02857549	0.07480709	0.78447095	2.68731703	0.10111991
Pythium	6.00172316	0.01429191	0.8376528	0.84794373	0.02643452	0.87084366	10.3057665	0.00132615	19.5156283	8.09E-06	4.64308154	0.03178974
Fragilariopsis	164.002748	1.51E-37	6.62290763	0.01006703	1.01495252	0.33773296	0.42122016	0.51672804	461.00014	0.00653874	17.0102388	3.70E-05
Spinococcus	1.16176229	0.28110936	11.021431	0.00090054	0.02946815	0.86338873	4.4816555	0.03426052	1.7482469	4.80E-05	6.11038457	0.01348999
Nothophytophthora	67.6905714	1.91E-16	50.3894322	1.26E-12	2.5552481	0.10992856	0.70293923	0.00801822	3.99563118	0.04861836	5.21589919	0.02238127
Highmestomium	12.6312928	0.0003792	9.01149425	0.00268287	0.97768976	0.32770983	0.09173021	0.76198997	1.1220592	0.0423286	2.80064662	0.0942283
Trichomonas	70.9098907	3.74E-17	0.00084305	0.9763644	1.9494761	0.1659295	0.14187881	0.70424208	1.11033952	0.0371941	0.93132329	0.8348697
Baobea	17.3561868	3.10E-05	2.11026629	4.35E-06	0.71545515	0.39763876	3.97768081	0.0461107	1.5927609	0.6553058	2.09420134	1.84E-07
Aphanomyces	27.0063712	2.03E-07	0.99213893	0.31912627	0.10821364	0.74218644	0.12865752	0.71982804	0.48998965	0.47995311	0.06238176	0.80277032
Plasmodium	0.5303402	0.46464662	2.04708617	0.15250822	1.62280603	0.20273838	0.20102733	0.65410151	0.34600424	0.5563888	0.06543352	0.79751712
Tetrahymena	48.7675045	2.88E-12	2.9157679	2.85E-06	1.1621939	4.26E-27	29.185663	6.58E-08	1.58011113	0.0806511	4.30632044	0.03797101
Saprolegnia	0.00689099	0.93181757	0.17849511	0.67266938	1.64944415	0.19903188	5.13822303	0.01805477	11.9740072	34.254184	0.00015971	0.00015971
Salpingoeca	93.9766138	5.29E-27	2.3080773	0.12204665	1.96774448	0.16064748	2.97920413	0.08545709	1.30275109	0.25371132	13.245189	0.00078628
Naegleria	0.55824549	0.45496803	6.38983758	0.01147755	11.1330224	0.00082263	0.44504218	0.46635056	1.23897414	1.2554091	1.23897414	0.2666885
Elmeria	1.24003522	0.26546427	1.26031611	0.26159126	1.06995971	0.3099455	0.04934067	0.8241456	0.00207829	0.96364411	0.3779331	0.56104741
Thraustotheca	3.18795603	0.07418271	0.07318468	0.00139053	22.4349902	2.17E-06	24.398881	8.51E-07	4.7222152	0.0297725	6.2743119	0.112125
Sphaeroformia	8.16547767	0.00426952	22.6928586	1.90E-06	0.80175291	0.0560979	0.08245629	0.77399501	0.27108273	0.6888802	0.67066039	0.0006039
Capsospora	14.2574808	0.00015943	0.04989534	0.82324548	2.0640305	0.15080858	0.07481427	0.78452028	1.49587737	0.22130679	0.00035813	0.98490353
Thecamonas	11.7852556	0.00059702	7.70850243	0.00549613	1.20187729	0.27294769	11.6873989	0.00062925	0.85205816	0.06836876	0.79372684	0.0006039
Planorproteum	254.204207	2.21E-57	17.6698888	2.63E-05	26.5027099	2.63E-07	19.6181477	9.46E-06	2.12931298	0.14450551	0.04830794	0.33015037
Stentor	90.3126066	2.03E-21	6.90291802	0.00818208	99.994516	1.56E-23	6.37835027	0.01555205	1.37600156	0.24062963	0.04832272	0.82400999
Blastocystis	0.01207707	0.91024562	0.05924715	0.80789883	0.79651892	0.3721362	2.60468477	0.10654834	0.79279345	0.06190025	0.80513735	0.0006039
Gregarina	13.5466954	0.0002327	6.53823609	0.01058806	9.5224164	0.00202972	9.14810778	0.00248975	3.8668188	0.04924979	6.45385501	0.01107719
Emiliana	16.5951053	4.62E-05	28.4463388	9.61E-08	0.85813935	0.35420075	0.70238937	0.0158189	0.2287183	0.6234582	6.6532999	0.00089724
Monodina	6.129539	0.01329415	0.1351514	0.71316066	0.8	0.37109337	0.01985736	0.68793619	0.31880729	0.00591499	2.13093339	0.1445281
Cryptosporidium	6.57447636	0.01034514	0.36720643	0.54453096	0.00060939	0.92607296	0.15168343	0.69693178	0.80048441	0.36858872	2.41631872	0.12007881
Thalassiosira	0.0102986	0.9187808	2.5049459	0.13316411	8.21401865	0.0041568	1.7442349	0.2583947	0.0842267	0.7661835	0.10113854	0.73989645
Bignoniella	447.66509	2.33E-99	9.24581179	0.00236033	1.25394937	0.2617428	0.60818468	0.43543428	3.97354846	0.04622027	23.0707692	1.56E-06
Hyndea	6.08908178	0.01360198	64.9249941	7.78E-16	9.43626665	0.0021736	7.83996044	0.00511037	0.28229814	0.59518888	0.41584861	0.51901581
Cavendishia	13.0775903	0.00029805	2.73991658	0.07266938	5.01460559	0.05213433	22.184294	2.48E-06	5.94346486	0.014856	20.734882	5.27E-06
Saccharomyces	15.6402332	7.66E-05	0.1624803	0.68647464	94.9837138	3.34E-09	2.7348291	0.11341144	1.15538423	0.2823827	0.1604095	0.68878764
Peronospora	3.18413284	0.07435644	8.04102564	0.00457296	5.88E-05	0.99382928	0.01576264	0.90008866	0.13276618	0.17558014	0.03059293	0.86055393
Besnoitia	0.7795008	0.00531963	0.00995172	0.22069522	13.6284478	0.00022718	4.9785922	0.02566894	0.21188653	0.7221246	0.97434003	0.3240006
Symbiodinium	37.1123507	1.13E-09	4.09780041	0.0429365	3.8897854	0.0480923	1					

S1 Text. Supporting text for S1 Fig, S2 Fig and S3 Fig.

<https://doi.org/10.1371/journal.pgen.1008386.s018>

A prediction of the fail-safe hypothesis is that the closer a site is to the primary stop codon, the stronger the selection should be to preserve an ASC (all else being equal), closer stops having the effect of reducing costs of error prone readthrough. One way to test this hypothesis is to consider the extent to which ASC frequency is resilient to underlying GC pressure. Our expectation under the fail-safe hypothesis is that at codon position +1, stops will be largely resistant to GC pressure while at position +6 this resilience will be relatively diminished. We thus predict that looking across genomes, the plot of ASC usage against GC content should be flatter at site +1 than at site +6.

To evaluate this, Z-scores were calculated using the mean ASC frequencies and standard deviations, as calculated from the aforementioned simulations, for each genome and plotted against GC3 content with the slope determined by linear regression of Z-score predicted by GC3 (S1 Fig). While for reasons unknown, significant negative relationships were observed between Z-score and genomic GC3 content at each position (Spearman's rank: $p < 2.2 \times 10^{-16}$ for all positions; $\rho = -0.61$ at position +1, $\rho = -0.65$ at position +2, $\rho = -0.51$ at position +3, $\rho = -0.45$ at position +4, $\rho = -0.41$ at position +5, $\rho = -0.46$ at position +6), unexpectedly the negative slope is steeper at position +1 than position +6 ($p = 5.2 \times 10^{-14}$) contra to fail-safe hypothesis expectations. Indeed, there is a significant negative correlation between absolute gradient and distance from the primary stop (Spearman's rank: $p = 2.7 \times 10^{-3}$, $\rho = -1$; S2 Fig), confirming that slope becomes flatter with 3' distance. Using absolute stop codon frequency, rather than Z-score, confirms the same result (S3 Fig). These results additionally indicate that in GC-rich genomes, fail-safe ASCs are if anything avoided (note the Z scores control for dinucleotide content) and provide no support for the hypothesis that ASCs in closest proximity to the primary stop are more strongly preserved. This test comes with the caveat that the costs of stops further downstream may be relatively weak and thus failure to meet expectations is not necessarily strong evidence against the hypothesis. This being said, it is notable the trend is against our expectations.

S2 Text. Supporting text for S1 Table and S2 Table.

<https://doi.org/10.1371/journal.pgen.1008386.s019>

Do we find no evidence for ASC enrichment because genomes specifically remove ASCs at a higher rate than chance? Perhaps switches from non-stop to stop occur at a lower rate than chance, and hence ASCs are a difficult evolutionary solution to stop codon readthrough? Further to the whole-UTR analysis in the main manuscript, we investigate both of these possibilities through analysing codon switches from stop to non-stop, and vice versa, in all downstream codon positions separately (S1 Table). Additionally, we look at a triplet of mollicute genomes to analyse TAA, TGA and TAG separately (S2 Table).

S3 Text. Supporting text for S4 Fig.

<https://doi.org/10.1371/journal.pgen.1008386.s020>

Selection for termination efficiency is thought to be highest in HEGs due to the notion that the net effect of readthrough is a function of the number of translation events any given transcript is subject to. If the fail-safe hypothesis of ASCs is true, we therefore expect ASC frequencies to be significantly higher in HEGs than LEGs. This, however, does not seem to be the case. Standardised differences in ASC frequency for each genome $[(\text{HEGs} - \text{LEGs})/\text{mean}(\text{HEGs} + \text{LEGs})]$ were calculated and are presented in S4 Fig. There were no significant differences between the ASC frequencies of HEGs and LEGs at any position (Wilcoxon signed-rank tests, $p = 0.40$ for position +1, $p = 0.68$ for position +2, $p = 0.62$ for position +3, $p = 0.57$ for position +4, $p = 0.83$ for position +5, $p = 0.77$ for position +6), suggesting that either expression level has no influence over the negative effects of readthrough or ASCs do not significantly affect the ability of a transcript to avoid these consequences.

S4 Text. Supporting text for S5 Fig.

<https://doi.org/10.1371/journal.pgen.1008386.s021>

Are stop codons enriched when we allow for +4T enrichment? To compensate for the detected +4T enrichment, we returned to our initial null simulation experiment. This time we considered only +4T containing genes and adjusted our Markov models such that only +4T-containing genes were produced via simulation. Sequences were generated such that thymine was the first base 100% of the time, with following nucleotides selected according to dinucleotide frequencies. Therefore, these +4T-containing simulated sequences produced a null model appropriate for comparison with the +4T-containing genes from real genomes. In acknowledgement of a possible weak signals identified at position +1 (see main paper) and position +2 in HEGs (see Fig 3 of the main paper), we consider only these two positions. The null neutral expectation was that there is no difference between the ASC frequencies of the real genomes and simulated sequences. To assess this, we calculated Z-scores and completed binomial tests. Given our simulated sequences are built upon dinucleotide content alone we expect a random distribution of ASCs, thus our null expectation is a 50:50 split of positive and negative Z-scores. We find there to be significant variation from this ratio at both positions in all genes (Binomial tests: 34/644 $Z > 0$, $p < 2.2 \times 10^{-16}$ for position +1; 67/644 $Z > 0$ at position +2). These results are repeated in HEGs (Binomial tests: 1/22 $Z > 0$, $p = 1.1 \times 10^{-5}$ for position 1; 4/22 $Z > 0$, $p = 4.3 \times 10^{-3}$ for position 2) and LEGs (Binomial tests: 5/21 $Z > 0$, $p = 0.026$ for position 1; 3/21 $Z > 0$, $p = 1.4 \times 10^{-3}$ for position 2). The result at position +1 in LEGs however does not survive multi-test correction ($p > 0.05/2$). As with the original simulations, in contrast to the prediction of enrichment per the fail-safe hypothesis, we note that deviation is due to under usage of ASCs (note the rarity of instances of $Z > 0$).

As before, we next looked at the proportion of genomes showing significant deviation from null ($|Z| > 1.96$). In this instance, the null expectation of the binomial test is no longer 50:50, rather that 95% of genomes will not be significantly deviated and 5% will. In all genes, there no significant deviation from this ratio at position +1 (Binomial test: 36/644, $p = 0.4693$) but significant deviation at position +2 (Binomial test: 0/644, $p = 7.0 \times 10^{-15}$). Closer examination indicates that significant difference at position +2

is due to under enrichment (Binomial test, alternative = 'lower': 0/644 $Z > 1.64$, $p = 4.5 \times 10^{-5}$ at position +1 and position +2). In HEGs and LEGs, there are no significant deviations from null ($|Z| > 1.96$).

The direction and magnitude of deviation from null was once again considered using the calculation of Z-scores. As discussed in the analysis of our previous simulations, the fail-safe hypothesis predicts resistance to GC pressure at position +1 and thus a flat slope when Z-score is plotted against genomic GC3. We repeat this analysis with our +4T-controlled null model (S5 Fig). Consistent with our original simulation-based analysis, we find Z-scores to be negatively correlated with genomic GC3 when considering all genes at position +1 (Spearman's rank: $\rho = -0.31$, $p < 2.2 \times 10^{-16}$) and position +2 (Spearman's rank: $\rho = -0.19$, $p = 1.6 \times 10^{-6}$). It therefore appears that ASCs are in fact avoided rather than enriched, consistent with our initial simulation experiment. No such negative relationships are found in HEGs and LEGs at either position (Spearman's rank: $p > 0.05$), however low Z-score magnitudes ($|Z| < 1.96$) indicate that ASCs frequencies defy the fail-safe prediction of enrichment at the very least.

S5 Text. Supporting text for S6 Fig.

<https://doi.org/10.1371/journal.pgen.1008386.s022>

Building on the observation that there may be a preference for fifth site thymine or cytosine in +4T-containing genes, we look at fifth site nucleotide frequencies. +4T-containing genes were extracted and fifth site nucleotide frequencies were calculated and compared (S6 Fig). Consistent with the enrichment of TC and TT-starting codons, there is preference for either thymine or cytosine when considering all genes. Fifth site T and C are both found in significantly higher frequency than the next most common nucleotide in TAA-terminating genes (Wilcoxon signed-rank tests: $T > A$, $p < 2.2 \times 10^{-16}$; $C > A$, $p < 2.2 \times 10^{-16}$), TGA-terminating genes (Wilcoxon signed-rank tests: $T > A$, $p < 2.2 \times 10^{-16}$; $C > A$, $p = 4.2 \times 10^{-08}$), and TAG-terminating genes (Wilcoxon signed-rank tests: $T > A$, $p < 2.2 \times 10^{-16}$; $C > A$, $p = 3.7 \times 10^{-15}$). Interestingly, in HEGs a fifth site T is preferred over C in all three groups (Wilcoxon signed-rank test: $p = 3.7 \times 10^{-3}$ in TAA-terminating genes; $p = 6.2 \times 10^{-3}$ in TGA-terminating genes; $p = 0.034$ in TAG-terminating genes), suggesting fifth site T is most optimal. In LEGs, there is no significant difference between any of the nucleotides at the fifth site of +4T-containing TAA-terminating genes (Kruskal-Wallis: $\chi = 7.503$, $p = 0.057$). Adjusting for Bonferroni correction ($p > 0.05/3$), thymine is not present in significantly higher frequency than the next highest base in TGA-terminating LEGs ($T > G - W = 130.5$, $p = 0.026$) or in TAG-terminating LEGs ($T > A - W = 114$, $p = 0.11$).

The above test doesn't control for GC pressure and may thus reflect an excess of AT rich genomes in our sample. To address this, we compare frequencies of TC or TT-starting codon frequency at position +1 to the average frequency of the respective codons between positions +1 to +6. In agreement with our frequency plots, we find TC-starting codons to be significantly enriched at position +1 (Wilcoxon signed-rank test: $p < 2.2 \times 10^{-16}$), and thus fifth site cytosine to be enriched in +4T-containing genes. However, we unexpectedly find no enrichment of TT-starting codons compared to null (Wilcoxon signed-rank test: $p = 0.26$). As this result is not consistent with our expectation, we cannot rule out the possibility that enrichment arises merely due to GC nucleotide pressure. Consistent with this we find that TT-codon usage at position

+1 decreases with genomic GC3, whereas TC-codon enrichment increases (S6 Fig). This proposes a hypothetical model where fifth site thymine is favoured at low GC, and cytosine at high GC in +4T-containing genes.

S6 Text. Supporting text for S7 Fig.

<https://doi.org/10.1371/journal.pgen.1008386.s023>

We acknowledge the limitations of LOESS modelling, which include those relating to the arbitrary nature of kernel/span function, and therefore validate the LOESS result with a different test design. Mollicute ASC frequencies were compared to GC-matched TT11 genomes (S7 Fig). In TT4 genomes, only TAA and TAG are used for chain termination. Hence, as TGA functions as a stop codon in TT11 genomes, it is expected under the fail-safe hypothesis that TGA frequency 3' of the primary stop in TT4 genomes should be consistently lower than that in TT11 genomes. For each mollicute genome analysed, TT11 genomes of GC3 content within 3.5% were selected. The ASC frequencies of these genomes were calculated, and averages were calculated for each position. We find TGA to be underrepresented at positions +3 and +5 only (Wilcoxon signed-rank tests: $p = 0.11$ for position +1; $p = 0.15$ for position +2; $p = 1.5 \times 10^{-3}$ for position +3; $p = 0.70$ for position +4; $p = 6.8 \times 10^{-4}$ for position +5; $p = 0.11$ for position +6). This result is consistent with our LOESS analysis.

S7 Text. Supporting text for S3 Table.

<https://doi.org/10.1371/journal.pgen.1008386.s024>

Prior mollicutes analysis (see main paper and S4 Fig) agrees with the hypothesis that TGA is underused in 3' domains when it isn't employed as a stop codon, compared with its usage in genomes of similar GC content when it can function as a stop. Whilst it remains possible that other codons may also be under-used in these genomes, for other reasons, our hypothesis does predict TGA to be among the most strongly under-enriched. We thus investigated all 64 codons using the aforementioned LOESS methodology and ranked them by their one-tailed Wilcoxon signed-rank test p-value (S3 Table). We find TGA to be the 25th most under-enriched codon at position +1, 20th at position +2, 4th at position +3, 49th at position +4, 2nd at position +5, and 16th at position +6. Instead, we find codons CCG (1st at positions +1, +4, +6), GTG (2nd at position 1, 3rd at positions +4 and +6), and TAT (1st at position +2, 2nd at position +3, 4th at position +1) among the more commonly underrepresented codons at specific positions. It therefore appears premature to presume that there is something special about TGA selection in bacterial 3' UTRs relating to translational termination.

S8 Text. Supporting text for S8 Fig, S4 Table and S5 Table.

<https://doi.org/10.1371/journal.pgen.1008386.s025>

Given that the vast bulk of our evidence argues against ASCs functioning as fail-safe stop codons, we predict that there is no reason to maintain relative codon usage downstream of the primary stop. In contradiction, we find striking similarities between position +0 (primary stop) and positions +1 to +6 (S8 Fig). Surprisingly, we find that trends in TGA and TAG usage remains clearly decoupled despite their equal GC content (S4 Table).

This result implies that stop codon usage at the primary stop is the same as usage at 3' positions. However, this would not necessarily be the case if we were to find that relative codon usage is also maintained across the two other reading frames. We find this to be true (S5 Table), and hence question the validity of current hypotheses that release factor abundance explains the decoupling of TGA and TAG usage.

S9 Text. Supporting text for S6 Table.

<https://doi.org/10.1371/journal.pgen.1008386.s026>

Our proposed hypothesis of gene shortening via the conversion of candidate stop codons (those that are upstream of the primary stop and are one-point mutation away from being a stop codon) to stop codons predicts that HEGs are longer in nucleotide length than LEGs. We consider only preliminary tests to indicate whether this may be true at a genomic level. Specifically, we assess the correlation between nucleotide length and protein abundance (S6 Table). Indeed, 18/22 demonstrate a negative trend of which 14/22 are significant.

S10 Text. Supporting text for S7 Table, S8 Table and S9 Table.

<https://doi.org/10.1371/journal.pgen.1008386.s027>

For the identification of ASC enrichment in eukaryotic genomes we apply three methodologies – Z-score deviation from dinucleotide-controlled null (S7 Table), Chi2 with dinucleotide-controlled null (S8 Table), and Chi2 with the Adachi and Cavalcanti null (S9 Table). Please find the detailed results of these analyses for each genome over the next few pages.

Chapter 3

Effective population size predicts local rates but not local mitigation of read-through errors

Alexander T. Ho and Laurence D. Hurst
Molecular Biology & Evolution, msa210.

This chapter contains work published on 14th August 2020 at MBE, the original and sole place of publication. It thus contains analysis of publicly available data using bespoke scripts that are freely available at the locations cited within the paper. The paper is open access and I have permission as the author to include the article in full (https://academic.oup.com/journals/pages/authors/production_and_publication/online_licensing). The latest version of the published article can be found by following the address: <https://doi.org/10.1093/molbev/msaa210>.

Pre-amble

How do local error mitigation and local error prevention strategies co-evolve?



In the previous chapter, I note that the strength of selection acting upon additional stop codons (ASCs) should depend on the error rate of the stop codon used for termination. If an error-prone stop codon is used to terminate translation then ASCs might be under strong selection, while reliable termination at the canonical site might render ASCs unnecessary. Indeed, the causal arrow could theoretically face the opposite direction. If error mitigation by ASCs were to efficiently ameliorate the consequences of translational read-through (TR) then genomes might not need to adapt their stop codon usage to reduce TR error rates. The co-dependence of error prevention and error mitigation strategies upon each other is thus an intriguing problem.

Chapter 2 provided one line of evidence supporting error prevention being the preferred solution. While the highly expressed genes of bacteria and humans prefer termination with TAA, the most reliable stop codon, I found no evidence that such bacterial genes are enriched for ASCs. In this chapter, I investigate how well both error control strategies accord with the predictions of nearly neutral theory. Under nearly neutral theorem, species with large effective population size (N_e) have more efficient selection to purge deleterious mutations and thus fix optimal sequence motifs. The tendency for any given species' genome to be enriched for TAA or fail-safe ASCs should thus depend upon its N_e . Without co-dependency between the two error solutions, one might predict both TAA and ASC enrichment to correlate positively with N_e . If one, and not the other, were to correlate with N_e this could be inferred as a selective preference for either prevention or mitigation.

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:	
Effective population size predicts local rates but not local mitigation of read-through errors	
Publication status (tick one)	
Draft manuscript <input type="checkbox"/> Submitted <input type="checkbox"/> In review <input type="checkbox"/> Accepted <input type="checkbox"/> Published <input checked="" type="checkbox"/>	
Publication details (reference)	Ho AT, Hurst LD. 2020. Effective population size predicts local rates but not local mitigation of read-through errors in eukaryotic genes. Mol. Biol. Evol. 38(1): 244–262.
Copyright status (tick the appropriate statement)	
I hold the copyright for this material <input type="checkbox"/> Copyright is retained by the publisher, but I have been given permission to replicate the material here <input checked="" type="checkbox"/>	
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the... Formulation of ideas: 100% Design of methodology: 100% Bioinformatic analyses: 100% Experimental work: N/a Presentation of data in journal format: 100%
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.
Signed	Date 03/03/2022

Effective Population Size Predicts Local Rates but Not Local Mitigation of Read-through Errors

Alexander T. Ho ^{*}, and Laurence D. Hurst 

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

Associate editor: Claus Wilke

*Corresponding author: E-mail: a.t.ho@bath.ac.uk

Abstract

In correctly predicting that selection efficiency is positively correlated with the effective population size (N_e), the nearly neutral theory provides a coherent understanding of between-species variation in numerous genomic parameters, including heritable error (germline mutation) rates. Does the same theory also explain variation in phenotypic error rates and in abundance of error mitigation mechanisms? Translational read-through provides a model to investigate both issues as it is common, mostly nonadaptive, and has good proxy for rate (TAA being the least leaky stop codon) and potential error mitigation via “fail-safe” 3' additional stop codons (ASCs). Prior theory of translational read-through has suggested that when population sizes are high, weak selection for local mitigation can be effective thus predicting a positive correlation between ASC enrichment and N_e . Contra to prediction, we find that ASC enrichment is not correlated with N_e . ASC enrichment, although highly phylogenetically patchy, is, however, more common both in unicellular species and in genes expressed in unicellular modes in multicellular species. By contrast, N_e does positively correlate with TAA enrichment. These results imply that local phenotypic error rates, not local mitigation rates, are consistent with a drift barrier/nearly neutral model.

Key words: nearly neutral theory, molecular evolution, translational read-through, additional stop codons, phenotypic error, error mitigation.

Introduction

Genomes vary in multiple parameters that typically covary. Some genomes, like ours for example, are “bloated” in the sense that they have large introns, many introns (per bp of coding sequence), large intergene distances, and a high load of transposable elements (TEs) (Lynch and Conery 2003). As a consequence, we have an especially low density of exon to immature transcript size (Warnecke et al. 2008) and a low gene density (Lynch and Conery 2003). Other eukaryotic genomes, such as that of yeast (*Saccharomyces cerevisiae*), are by contrast lithe: Introns are rare and small, intergene distance is low, TE load is relatively light, and gene density is thus high (Lynch and Conery 2003).

How are we to understand not just the variation between genomes in such parameters but also the tendency for multiple measures of genomic “economy” to positively correlate one with the other? Lynch (2007) has forcefully argued that the nearly neutral theory (Ohta 1992) can explain both problems. This proposes that when the effective population size (N_e) is large, selection is relatively efficient at removing deleterious mutations. By contrast when N_e is low, mutations, such as TE insertions, that would be “seen” as deleterious in species with large N_e are instead only weakly deleterious or effectively neutral and hence able to be fixed owing to drift. Consistent with this, the level of constraint on protein

sequence evolution is higher when N_e (or proxies to it) is high (Keightley and Eyre-Walker 2000). More particularly, species with low N_e are expected to accumulate uneconomical sequence, leading to genome bloating. Consistent with this model, intron density and size covary negatively with the effective population size (Lynch and Conery 2003; Wu and Hurst 2015).

The genome bloating concerns what generically might be regarded as aspects of genomic anatomy. Can the same body of theory also explain genomic behavior? Lynch et al. (2016) have also argued that, although the mutation rate is under selection to be as low as possible (Kimura 1967), low N_e forces a drift barrier preventing especially low mutation rates in species with small N_e . As a consequence, genomes of species with low N_e have both bloated genomes and high mutation rates (both per bp per generation and especially per genome) (Sung et al. 2012; Lynch et al. 2016).

The mutations considered in these models are one class of error, these being heritable errors. One can also ask about selection on nonheritable (somatic) mutations and nonheritable nonmutational “phenotypic” errors (Burger et al. 2006; Willensdorfer et al. 2007), such as accidental mistranslation, frameshifts, stop-codon read-through, missplicing, misfolding, and so on. Here, we consider between-species variation in phenotypic errors. Errors like these are ubiquitous, occur at

high rates, and are typically deleterious (Warnecke and Hurst 2011; Yang et al. 2017; Fu et al. 2018; Liu and Zhang 2018a, 2018b; Li and Zhang 2019).

To resolve these errors, genomes can employ “global” or “local” solutions (Rajon and Masel 2011). Global solutions involve strategies achieved by altering the molecular machinery required for gene expression and hence moderate errors at multiple sites. By contrast, local solutions are employed to ameliorate error at a specific site, or gene. In turn, each class of solution (global/local) can either affect the rate of error or mitigate errors once they have occurred. The pathway to detect and recycle misfolded proteins (Chen et al. 2011; Jackson and Hewitt 2016) may be considered as a global error mitigation device, preventing the buildup of potentially toxic misfolded proteins, whereas employment of chaperones, to direct the correct folding of proteins, may be considered part of a system of global error rate modification, preventing misfolding in the first place. Other examples of global mitigation solutions include improving the machinery required for proofreading during transcription (Zenkin et al. 2006; Gamba and Zenkin 2018) and nonsense-mediated decay (NMD) to trap misspliced transcripts (Kawashima et al. 2009; Tabrez et al. 2017). Further, the genetic code appears to be structured in a manner that reduces the impact of mistranslation (Freeland and Hurst 1998) and, in addition, misacylated tRNAs may tend to mismatch with codons that code for the misloaded amino acids (Seligmann 2011), both of which may be considered as global mitigation strategies.

Here we focus on the problem of selection on local error control devices, in particular to understand how selection on local error rate and local error mitigation vary with N_e . Although it has been suggested that selection for local mutation rate modification (i.e., heritable errors) is too weak (Hodgkinson and Eyre-Walker 2011; Chen and Zhang 2013), even in species with large N_e (Chen and Zhang 2013), selection on local phenotypic error handling may be different as the underlying rates are higher. As regards local rate modifiers, examples include usage of strong splice sites or exonic splice enhancers (ESEs) to increase splicing accuracy of a specific exon or the use of optimal codons to modify rates of amino acid misincorporation at one codon within a gene (Stoletzki and Eyre-Walker 2007). Numerous local error mitigation mechanisms have also been suggested. Codons mutationally adjacent to stop codons (those one mutation away) are avoided at the 3'-end of human genes (Cusack et al. 2011) where NMD cannot recognize premature stop codons (Zhang et al. 1998). Although this will not affect the rate of mistranscription, it will ensure that mistranscription costs are reduced as especially harmful premature stops codons are less likely to be a consequence of mistranscription. Out-of-frame stop codons are thought to promote translation termination following erroneous frameshift events (Seligmann and Pollock 2004) and have been found to be enriched downstream of frameshift-prone codons (Seligmann 2019). Similarly, in eukaryotes, the presence of in-frame stop codons in introns implies selection to degrade by NMD erroneously spliced mRNA in which introns are retained (Jaillon et al. 2008;

Sayani et al. 2008; Brogna and Wen 2009; Ramani et al. 2009; Behringer and Hall 2016). Intronic stop codons occur earlier in the intron than expected by chance, consistent with selection to minimize waste and trigger NMD as soon as possible (Behringer and Hall 2016), although the presence of an in-frame intronic stop codon is no guarantee of NMD-mediated removal on intron retention (Sayani et al. 2008).

The distinction between local error rate and error mitigation control is not always unambiguous. As noted above, selection on codon usage is suggested to alter the rate of amino acid misincorporation (Stoletzki and Eyre-Walker 2007) or mitigate mistranscription events (Cusack et al. 2011). Comparably, in addition to global rate modification, chaperones could in principle mitigate the misfolding effects of mistranscription and mistranslation. Similarly, owing to great enrichment of A at coding site +4, the trinucleotide TGA is greatly enriched at positions 2–4 in bacterial genes (which start NTC) (Abrahams and Hurst 2017). As TGA is a stop codon, this might possibly be to enable rapid frameshift correction, that is, stopping a misaligned ribosome and putting it back into frame by enabling a one base shuttle backward, thereby reducing the net rate of out-of-frame initiation. Alternatively, if the +1-stop codon enables rapid release of the ribosome, it acts to minimize the costs, rather than reducing the rate. Logically it is possible that both occur.

Here we consider what may be a good exemplar for considering relative selection strength on local error rates and local error mitigation, namely the rate and mitigation of translational read-through. Read-through happens when the primary stop codon of an expressed gene is not recognized by its release factor (Roy et al. 2015; Beznoskova et al. 2016) leading to translation of the 3'-untranslated region (UTR) (Doronina and Brown 2006; Namy and Rousset 2010)—see Rodnina et al. (2020) for a recent review. There are some hypothetical advantages of read-through, such as increased proteome diversity (Dunn et al. 2013) and access to additional C-terminal protein domains at low abundance (well described in Pancrustacea [Jungreis et al. 2011], mammals [Eswarappa et al. 2014], yeast [Namy et al. 2003], for example). Read-through may also enable selection to purge deleterious 3'-UTR sequence (Giacomelli et al. 2007; Rajon and Masel 2011; Kosinski and Masel 2020). However, the best evidence suggests that it is typically nonadaptive and arises due to molecular error (Li and Zhang 2019).

The costs of C-terminal extension via read-through have multiple mechanisms. In the absence of a fail-safe stop codon, we might expect degradation of both RNA and nascent protein when the translating ribosome reaches the polyA tail (Dimitrova et al. 2009; Klauer and van Hoof 2012). Should protein be produced following termination at a 3' fail-safe stop codon there may yet be problems with protein localization (Falini et al. 2005; Hollingsworth and Gross 2013), protein aggregation (Vidal et al. 1999, 2000), and protein stability (Clegg et al. 1971; Namy et al. 2002; Pang et al. 2002; Inada and Aiba 2005; Shibata et al. 2015) causing reduced titer (Arribere et al. 2016). Aside from these, even in a best-case

scenario there is likely to be energetic wastage from unnecessary 3'-UTR translation (Wagner 2005).

One reason read-through is a useful exemplar for broad-scale pan taxon analysis is that, unlike the case of splicing error, where different species employ different SR proteins and ESEs and have different intron lengths and densities, the molecular biology of termination is similar across eukaryotes (and to some degree within prokaryotes and archaea) (Capechi 1967; Grentzmann et al. 1994; Mikuni et al. 1994; Stansfield et al. 1995; Zavialov et al. 2001; Salas-Marco and Bedwell 2004; Alkalaeva et al. 2006; Cao et al. 2007; Dever and Green 2012; Kobayashi et al. 2012).

Furthermore, from genomic analysis alone we can make inferences concerning error rates. This is because both prokaryotes and eukaryotes preferentially use the least leaky stop codon (Strigni and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen et al. 1993; Meng et al. 1995; Sanchez et al. 1998; Tate et al. 1999; Wei et al. 2016; Cridge et al. 2018), TAA, to terminate translation, the preference being strongest where the costs of erroneous read-through would be highest, namely in highly expressed genes (HEGs) (Korkmaz et al. 2014; Trotta 2016). We do not exclude the possibility of other modes of selection acting in favor of TAA. There may, for example, be selection for fast release of the ribosome to prevent ribosomal traffic jams (Tuller et al. 2010). Conserved TAA repeats at specific sites in tRNAs overlapping mRNAs in mtDNA might imply utility beyond its function as a stop codon (Faure and Barthélémy 2019). Furthermore, TAA is robust to two mistranscription events (TAA->TGA, TAG) whereas the two other stop codons are resilient to only one (TGA->TAA, TAG->TAA). We can, however, discern that at least some TAA selection relates to translational read-through by examination of 3' flanking sequence known to alter read-through rates (Bossi and Roth 1980; Wei and Xia 2017; Cridge et al. 2018). Enrichment of these flanking motifs across genes, aligned with evidence for TAA preference in HEGs, provides solid evidence that read-through is a significant, although not necessarily unique, selection pressure.

A third reason that read-through is a good exemplar is because there is prior evidence for an easy to define error mitigation mechanism. Notably, 3' in-frame additional stop codons (ASCs) may ameliorate translational error costs by providing a second opportunity (a fail-safe mechanism) to terminate translation (Nichols 1970; Major et al. 2002; Liang et al. 2005; Adachi and Cavalcanti 2009; Fleming and Cavalcanti 2019). ASCs have sometimes been referred to as "tandem stops"; however, we prefer the "ASC" terminology to avoid possible confusion relating to their proximity to the primary stop. The term "tandem stop codon," for example, sometimes only refers to the immediately proximal in-frame codon position. Similarly, we note that ASCs are distinct from out-of-frame stop codons, these being stops that lie out-of-frame in coding sequence, possibly to ameliorate frameshift errors (Seligmann and Pollock 2004; Abrahams and Hurst 2017). One might expect that selection for ASCs might be stronger the closer they are to the focal stop codon. For this reason, and following prior evidence of enrichment specifically at sites very close to the focal stop (Liang et al. 2005;

Adachi and Cavalcanti 2009; Ho and Hurst 2019), we here consider ASC enrichment in the following six in-frame "codon" positions.

Theoretical expectations regarding N_e and the selection on error rate control and error mitigation are not as simple as stronger selection, and hence greater commonality of both, when N_e is high (Rajon and Masel 2011; Meer et al. 2020). The situation is especially complex as the global/local and rate/mitigation distinctions provide four mutually dependent axes for selection. Selection on mitigation and rate have the potential to be negatively associated: If rates are low, mitigation is unnecessary; if mitigation is effective, selection on rate reduction diminishes (Rajon and Masel 2011). Similarly, if global error rates are low or global mitigation mechanisms effective, selection for local effects will be weaker and vice versa. These dynamics are even more complicated as correlations can be accentuated by subsequent evolution. If, for example, error read-through rates are low, then 3' downstream regions are effectively shielded from selection so enabling accumulation of mutations that render read-through more deleterious should it happen, intensifying selection to reduce rates (Rajon and Masel 2011; Meer et al. 2020). This sort of positive feedback loop produces, it is argued, two attractors: mostly deleterious consequences of read-through (no ASCs) coupled with low read-through rates, and mostly benign read-through (owing to ASCs and other devices) coupled with high read-through rates (Rajon and Masel 2011; Meer et al. 2020).

The question then is how the occupancy of these two solutions, assuming these to be the only two stable solutions, might be affected by changes in N_e . Unlike global solutions, local solutions must evolve multiple times in order to affect error handling for multiple genes. Each event must have a low selection coefficient associated with it (as opposed to global modifiers). Considering the case of read-through errors in particular, it was thus argued that a strategy of high error rate with common mitigation is expected under high N_e (Rajon and Masel 2011; Meer et al. 2020). Conversely as N_e declines, the solution could shift to globally regulated low error rates (low read-through rates) and absence of mitigation (reduced selection for ASCs).

In support of their model, especially as regards read-through, Rajon and Masel (2011) argue that yeast has a large population size, high read-through error rates with effective local mitigation of read-through, citing previously observed ASC enrichment (Liang et al. 2005). However, there was no comparator from taxa with smaller or larger N_e . Our recent demonstration (Ho and Hurst 2019) that in bacteria (that we presume to mostly have even higher N_e) there is no evidence for ASC enrichment would appear to contradict the prediction of enrichment for local mitigation when N_e is high (see also Korkmaz et al. [2014]). However, as bacterial and eukaryotic termination mechanisms are not identical the comparison may not be fair.

Meer et al. (2020) argue that the high mistranscription rate in *Escherichia coli* compared with species such as *S. cerevisiae* supports the view of higher error rates when N_e is also high. Meer et al. (2020) also note, however, the possibility of local selection to reduce error rates when N_e is high as 1) selection is efficient and 2) intrinsic error rates are high. The recognition

of the potential relevance of local selection on error rate questions in turn, the assumptions of the original model. The model of Rajon and Masei (2011) assumes that error rate is a globally regulated process associated with trade-offs in translational velocity and growth rate, whereas mitigation is locally regulated (e.g., by selection for ASCs). Local modulation of rate via change in stop codon usage (leaky vs. nonleaky) is not considered. This complicates matters as rate (stop codon choice) and mitigation (ASC selection) are both local variables. As such, both are subject to low selective coefficients and so more likely to be recognized by selection when N_e is high. In this regard, Rajon and Masei (2011) make no prediction about TAA usage as a function of N_e . However, just as Meer et al. (2020) predict, and observe, lower mistranscription rates in HEGs than in lowly expressed genes (LEGs) when both N_e and global mistranscription rates are high (e.g., in *E. coli*), so too one can ask whether a greater HEG/LEG TAA disparity is seen when N_e is high.

Rather than attempting to extend theory to consider the balance between local rate and local mitigation (in a 2×2 framework), we shall instead attempt to provide a robust empirical base for theory to address. We shall consider rates of usage of TAA as a local rate reducing modifier and of ASCs as local mitigators of errors. Aside from N_e , however, we also ask about alternative possibly relevant parameters. For example, often when considering error mitigation, a distinction between unicellular and multicellular organisms may be relevant. We presume that any given gene expression error is more threatening to organismal survival in unicellular species compared with multicellular ones. In multicellular species, there are at least two mechanisms through which gene expression fitness effects could be ameliorated. Firstly, low fitness cells generated by molecular error may be removed by apoptosis and subsequently replaced through new cell proliferation and differentiation (Bergmann and Steller 2010; Brock et al. 2019). Secondly, in multicells the reduced productivity of low fitness cells could be ameliorated by the functional redundancy of its neighbors. These avenues are not equally open to all cells within a multicellular species. Indeed, for this sort of reason selection against erroneous protein translation is thought to be more stringent in neurons (Drummond and Wilke 2008). These same avenues for compensation are probably less open to unicell species also. Aside from cellularity, it may be important to consider genome anatomical features such as gene length, intergenic distances, and GC content. As stop codons are GC-poor, GC-rich genomes might be under stronger selection to preserve TAA or ASCs whereas AT-rich genomes have a higher probability of an in-frame ASC by chance. The costs of producing potentially deleterious read-through transcripts might also vary in terms of the proportion of the sequence added or in terms of the absolute length added.

Results

Evidence for Selection against Translational Read-through in Eukaryotes

We first sought to strengthen prior evidence (Li and Zhang 2019) that translational read-through is indeed opposed by

natural selection in eukaryotes. Given that nucleotides in close downstream proximity to the stop are implicated in stop codon recognition (Bossi and Roth 1980; Cridge et al. 2018; Tate et al. 2018) and hence are under selection to modify translational read-through, we ask 1) whether there is evidence for selective constraint in the vicinity of the stop codon, 2) whether overrepresented motifs reflect selection for read-through suppression specifically, and 3) whether TAA is overrepresented in HEGs. We note that analysis of preferences S' of the focal stop codon is complicated by selection on amino acid content that need have little or nothing to do with stop codon recognition. This notion is supported in yeast, S' codon usage being uncorrelated with known effects on translation termination (Williams et al. 2004). Indeed, amino acid choice has recently been shown to majorly impact protein expression and decay (Weber et al. 2020). For these reasons, we focus attention on $3'$ effects but present S' effects for context.

Constraint on Substitution Rate Surrounding Stop Codons Is Most Acute Near the Stop Codon

In bacteria, substitution rate gradually increases with the $3'$ distance from the stop codon (Belinky et al. 2018). We here apply the same species triplet method to consider substitution rates surrounding the stop codon in several eukaryotic species (see Materials and Methods). Consistent with the bacterial observations, we find substitution rate to be constrained in close proximity to the primary stop codon in TAA-, TGA-, and TAG-terminating genes across all of our eukaryotic groups (fig. 1). Although the shape of the downstream substitution rate curve is variable between groups, substitution rate is always lowest in close proximity to the stop codon, this being most evident in *Caenorhabditis*, *Drosophila*, and *Arabidopsis* (fig. 1).

Preference for Motifs That Decrease Read-through Rates in the Immediate Vicinity of the Stop Codon Is Commonplace in Eukaryotes

Constraint on substitution rate alone, however, need not be evidence for selection against translational read-through. Stop codon recognition is not the sole function of the $3'$ -UTR sequence, these sequences also containing regulatory motifs, binding sites for translational regulators, and so on (Kuersten and Goodwin 2003; Mayr 2019). To ascertain whether substitution constraint was attributable to selection for read-through modifying motifs, we assessed the sequence surrounding stop codons in HEGs for significant nucleotide enrichments and depletions (compared with global levels; fig. 2) and ask whether they relate to known read-through modulators (Cridge et al. 2018). Looking for enrichment in HEGs (compared with all genes) allows us to focus our analysis on identifying motifs that may decrease read-through, under the assumption that these genes are where the costs of aberrant stop recognition are the most extreme.

We find that certain site-specific nucleotide enrichments ($P < 0.05$) are consistent (>3 genomes; supplementary table T1 for TAA-terminating genes, supplementary table T2 for

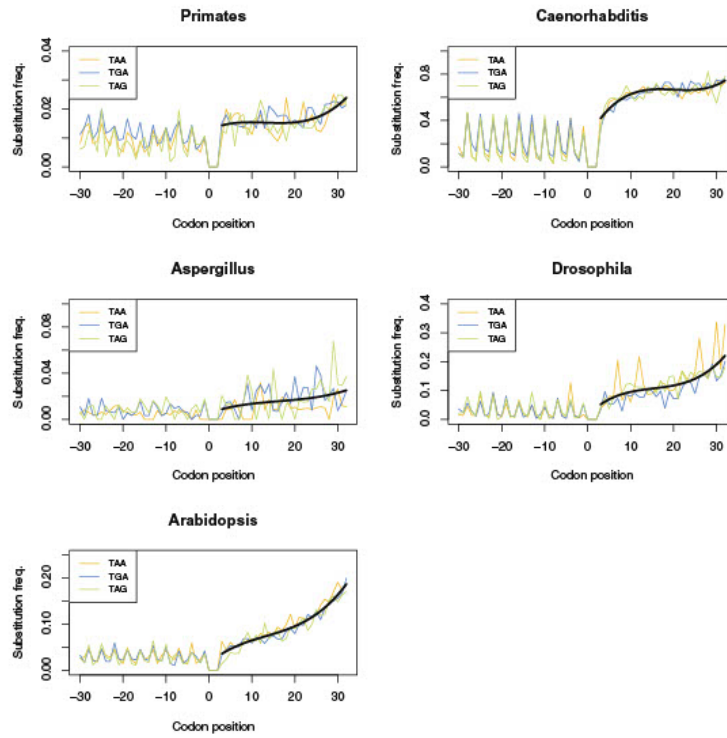


FIG. 1. Substitution frequencies of TAA-, TGA-, and TAG-terminating genes at nucleotide positions surrounding the primary stop codon in five eukaryotic groups. Though the profile of change in substitution rate downstream to the stop codon is different between groups, constraint on substitution rate is relieved with increased 3' distance in TAA-, TGA-, and TAG-terminating genes across all of groups. The black line represents a fitted polynomial line of the average substitution rate across all stop variants.

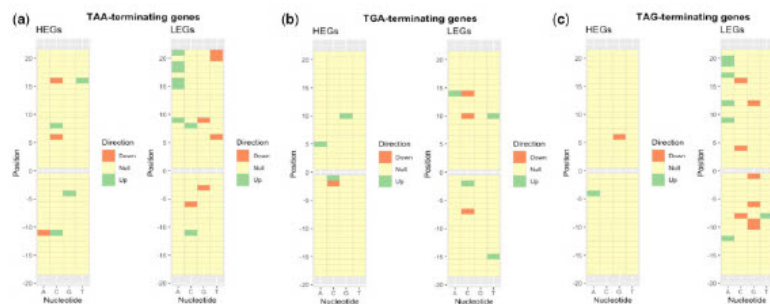


FIG. 2. Heat map showing significant nucleotide enrichments and depletions at positions surrounding (a) TAA, (b) TGA, and (c) TAG stop codons in highly expressed human genes. Significant enrichments and depletions in HEGs were determined by chi-square tests ($P < 0.05$) relative to a null expectation from all genes (regardless of expression level).

248

Downloaded from https://academic.oup.com/mbe/article/38/1/244/5892771 by University of Bath user on 03 January 2022

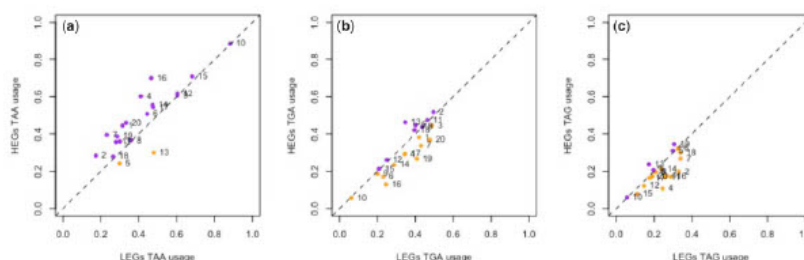


Fig. 3. Difference in the usage of (a) TAA, (b) TGA, and (c) TAG stop codons between HEGs and LEGs in 20 eukaryotic species. HEGs are the top quartile of genes expressed according to experimentally derived protein abundance data. LEGs are defined as the bottom quartile of expressed genes. The dotted line in each plot represents equal codon usage between HEGs and LEGs, hence points above the line represent overusage in LEGs (colored purple) and points under the line represent overusage in HEGs (colored orange). In our sample, 18/20 genomes contain higher TAA frequency in HEGs compared with LEGs. Numbered data points correspond to the following species: 1, *Gallus gallus*; 2, *Bos taurus*; 3, *Homo sapiens*; 4, *Xenopus tropicalis*; 5, *Aspergillus niger*; 6, *Drosophila melanogaster*; 7, *Chlamydomonas reinhardtii*; 8, *Arabidopsis thaliana*; 9, *Schizosaccharomyces pombe*; 10, *Dictyostelium discoideum*; 11, *Equus caballus*; 12, *Apis mellifera*; 13, *Rattus norvegicus*; 14, *Saccharomyces cerevisiae*; 15, *Plasmodium falciparum*; 16, *Anopheles gambiae*; 17, *Caenorhabditis elegans*; 18, *Oryza sativa*; 19, *Trypanosoma brucei*; 20, *Danio rerio*.

TGA-terminating genes, supplementary table T3 for TAG-terminating genes, Supplementary Material online) among eukaryotes. The observed trends are also consistent with selection to mitigate read-through. Notably, many of these common nucleotide preferences (eg, +4G or +5C following TAA) have previously been experimentally determined to decrease read-through rate (Cridge et al. 2018). We hence conclude that translational read-through is indeed a significant error in gene expression that triggers local error rate selection on 3' sequence in response.

TAA Stop Codons Are More Strongly Preferred in Highly Expressed Eukaryotic Genes

Having established that read-through is a significant selection pressure, we next assess whether TAA enrichment is a common evolutionary response. The assumption that TAA is the least leaky stop predicts that TAA stops should be the most common across all genomes, all else being equal. However, all else is not equal, the most common stop for any given genome being well predicted by GC content which is highly variable between species (supplementary fig. S1, Supplementary Material online). As expected if there is some form of GC pressure, the relative usage of TAA is negatively correlated with GC content ($P = 3.2 \times 10^{-6}$, $\rho = -0.803$; Spearman's rank), with TGA ($P = 0.00083$, $\rho = 0.636$; Spearman's rank) and TAG ($P = 0.00012$, $\rho = 0.705$; Spearman's rank) positively correlated. Similar to the trends observed in bacteria previously (Korkmaz et al. 2014), TAG is universally disfavored despite its identical nucleotide composition to TGA.

Given the above, rather than simply considering raw TAA usage between genomes, a fairer way to address whether there might be selection favoring TAA is to ask whether TAA is preferred in HEGs compared with LEGs within the same genome, expression level being a key modifier in the evolutionary dynamics of local error traps (Xiong et al. 2017).

Consistent with TAA selection, across a data set of 20 species (15 multicellular and 5 unicellular) for which we have proteomic data we find 18/20 possess higher TAA usage in HEGs (fig. 3). This significantly exceeds the simplest null expectation of a 50:50 split of TAA preference between HEGs and LEGs ($P = 0.0002$, one-tailed binomial test with null $P = 0.5$). Moreover, in HEGs, the observed TAA stop frequencies across our species are significantly higher than those of TGA ($P = 0.0047$; Wilcoxon signed-rank test) and TAG ($P = 1.33 \times 10^{-8}$; Wilcoxon signed-rank test) in the same species. This contrasts with what is seen in LEGs, where we recover no significant difference between TAA and TGA frequency across our data set ($P = 0.29$; Wilcoxon signed-rank test). In LEGs, TAA frequencies are, however, higher than TAG ($P = 0.00029$; Wilcoxon signed-rank test) possibly reflecting the fact that TAG is the leakiest stop and least favored.

ASCs Are Enriched in HEGs Predominantly in Genomes Where ASCs Are Globally Enriched

As with TAA stop codons we can also ask whether ASCs in the first six in-frame codon positions are preferred in HEGs (fig. 4), well-described ASC enrichment being previously witnessed in such proximity to the focal stop (Nichols 1970; Major et al. 2002; Liang et al. 2005; Adachi and Cavalcanti 2009; Fleming and Cavalcanti 2019; Ho and Hurst 2019). Using the same data set, we find that only 7/20 genomes possess an excess of ASCs in HEGs compared with LEGs when considering genes that end in any stop. This is no different than expected under the 50:50 null ($P = 0.26$, two-tailed binomial test with null $P = 0.5$). This might, however, be complicated by the fact that TAA-ending genes are also less leaky and highly expressed. However, we do not observe any deviation from this null across any of the primary stop groups either (7/20 genomes when considering TAA-terminating genes, 7/20 considering TGA-terminating genes,

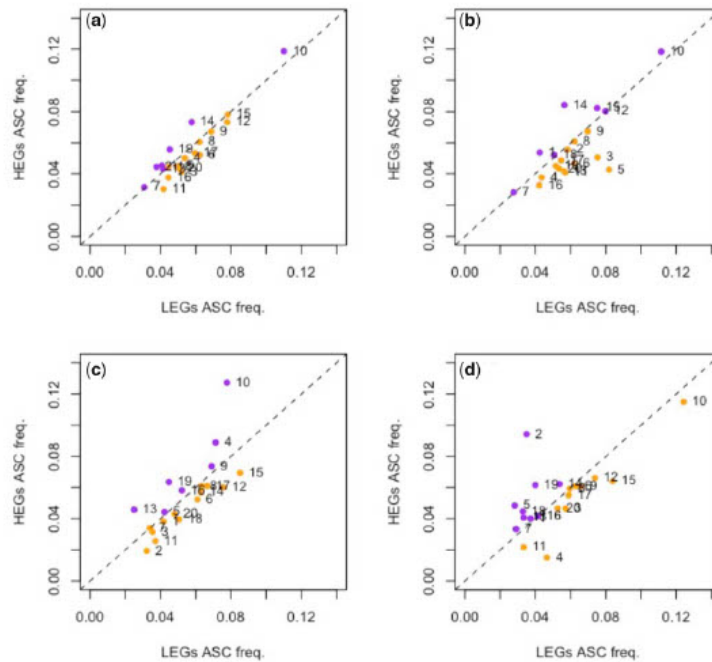


FIG. 4. Difference in ASC frequency between genes of high and low expression across (a) all genes, (b) TAA-terminating genes, (c) TGA-terminating genes, and (d) TAG-terminating genes in 20 eukaryotic species. HEGs are the top quartile of genes expressed according to experimentally derived protein abundance data. LEGs are defined as the bottom quartile of expressed genes. The dotted line in each plot represents equal ASC frequency in HEGs and LEGs, hence points above the line represent overusage in HEGs (colored purple) and points under the line represent overusage in LEGs (colored orange). In our sample, 7/20 genomes contain higher ASC frequency in HEGs compared with LEGs when considering all genes. Numbered data points correspond to the following species: 1, *Gallus gallus*; 2, *Bos taurus*; 3, *Homo sapiens*; 4, *Xenopus tropicalis*; 5, *Aspergillus niger*; 6, *Drosophila melanogaster*; 7, *Chlamydomonas reinhardtii*; 8, *Arabidopsis thaliana*; 9, *Schizosaccharomyces pombe*; 10, *Dictyostelium discoideum*; 11, *Equus caballus*; 12, *Apis mellifera*; 13, *Rattus norvegicus*; 14, *Saccharomyces cerevisiae*; 15, *Plasmodium falciparum*; 16, *Anopheles gambiae*; 17, *Caenorhabditis elegans*; 18, *Oryza sativa*; 19, *Trypanosoma brucei*; 20, *Danio rerio*.

9/20 considering TAG-terminating genes, all results $P > 0.05$, two-tailed binomial tests with null $P = 0.5$).

Although prima facie the above suggests that no selection is acting upon ASCs, we instead suggest that this owes to the phylogenetic patchiness of ASC enrichment. Across vertebrates, plants, fungi, and invertebrates we find some genomes that have significant ASC enrichment, and others that do not (supplementary fig. S2, Supplementary Material online), suggesting that strong ASC selection is not common to all species, but particular to a few. Specifically, we see no evidence for such enrichment in vertebrates as a group, but more species show enrichment in plants, fungi, and invertebrates than expected by chance (supplementary fig. S2, Supplementary Material online). If only some genomes, for whatever reason, have selection for ASCs these should be enriched for genomes showing an excess in HEGs compared

with LEGs. Indeed, of the seven genomes that possess an excess of ASCs in HEGs compared with LEGs (considering all genes), four contain significant ASC enrichment. By contrast, only one genome contains significant ASC enrichment out of the 13 which have ASC excess in LEGs compared with HEGs. These proportions are significantly different ($P = 0.031$, Fisher's exact test), suggesting that if ASCs are under selection it is in HEGs that they are most common. However, the same result also suggests that in many species ASCs are not under strong selection.

N_e Predicts TAA Usage, but Not ASC Enrichment

As we reported previously (Ho and Hurst 2019), some but not all unicellular eukaryotes show evidence of statistically significant ASC enrichment. As described above, this patchiness of ASC enrichment is observed in multicellular eukaryotes too.

We note that in all multicellular groups we have analyzed, ASC enrichment is rarer than seen in unicellular species (supplementary fig. S3, Supplementary Material online). Similarly, pooling all multicellular HEGs together and all unicell HEGs together we find TAA usage in unicells (62.7%) to significantly exceed that of multicells (43.7%) ($P < 2.2 \times 10^{-16}$, $\chi^2 = 597.1$, chi-square test 1 df). Such variation within groups, and between unicells and multicells, could potentially be explained by N_e or cellularity but our analysis has so far failed to control for phylogeny.

To test for correlation between TAA enrichment and N_e , we gather a sample of species for which we have an N_e estimate. As this sample contained a few species pairs that are especially phylogenetically close (and thus especially influential in the face of parameter estimation error), we pruned the species sample (and phylogenetic tree) to remove closely related species pairs with low species divergence times, leaving one of the two (e.g., human–chimp was resolved to human alone) (supplementary fig. S4, Supplementary Material online). For each genome in our reduced species list ($n = 15$), we calculate a TAA enrichment score taking into account background nucleotide usage (see Materials and Methods) and compare this parameter with N_e in phylogenetically controlled regression analyses (using phylogenetic generalized least squares [PGLS] tests). We find robust evidence to support a positive relationship between N_e and TAA enrichment (adjusted $r^2 = 0.55$, $P = 0.00098$, $\lambda = 0.0$; PGLS).

To test the comparable prediction for ASC enrichment, we calculated an ASC enrichment score for each genome. Interestingly, although a phylogenetically uncontrolled analysis reports significance in the direction expected (supplementary fig. S5, Supplementary Material online), a significant relationship between ASC enrichment and N_e was not recovered (adjusted $r^2 = -0.07$, $P = 0.85$, $\lambda = 1.0$; PGLS). This is because ASC enrichment shows a high rate of phylogenetic autocorrelation ($P = 0.03$ for $\lambda = 0.0$, $P = 1.0$ for $\lambda = 1.0$), the high λ value suggesting that the trait is evolving as expected given the tree topology alone.

These results suggest that, although TAA enrichment and ASC enrichment are both adaptations to translational read-through, TAA usage is consistent with expectations from the nearly neutral theory but ASC enrichment is not. Instead, its distribution appears to be patchy.

No Significant Correlation between N_e and TAA HEG/LEG Disparity

Meer et al. (2020) note that when N_e is high, there is a greater disparity in mistranscriptional error rates between HEGs and LEGs than there is when N_e is low. Here, we ask whether there is similarly a greater HEG/LEG TAA disparity when N_e is high. To test this, we employ protein abundance data to identify HEGs and LEGs for the species in our tree (due to lack of available data, we are reduced to $n = 11$). The variable to be measured for association with N_e was TAA frequency in HEGs divided by TAA frequency in LEGs (that we call TAA disparity). We find no significant relationship between TAA disparity and N_e in phylogenetic-controlled analysis ($P > 0.05$) (adjusted $r^2 = 0.19$, $P = 0.10$, $\lambda = 0.0$; PGLS). We note too

that the effect size measured by adjusted r^2 at 0.19 is substantially lower than that observed for the TAA enrichment– N_e effect (adjusted $r^2 = 0.55$). One possible caveat, however, is that the sample size here is a little lower than in the N_e –TAA enrichment correlation ($n = 11$ and $n = 15$) and λ is low indicating that phylogeny alone cannot explain all of the data. However, using the same species as in the TAA analysis (i.e., with $n = 11$), N_e remains strongly and significantly correlated with TAA enrichment (see above, and Materials and Methods) (adjusted $r^2 = 0.39$, $P = 0.024$, $\lambda = 0.0$; PGLS). This suggests that we have enough statistical power to detect a correlation between HEG/LEG TAA disparity and N_e at least if there was one of the same magnitude as seen with TAA enrichment.

Cellularity Predicts ASC Enrichment but Not TAA Usage

The results described so far suggest that high N_e genomes favor the most effective stop codon, especially in HEGs. However, N_e appears to have no ability to predict between-species variation in ASCs, at least after phylogenetic control. What might then explain such variation? We ask whether cellularity may be a predictor as multicellularity may protect against gene expression error by either cell redundancy or cell replacement.

First, we ask whether cellularity (considered as a binary trait in PGLS analysis using the same species tree) predicts TAA and ASC enrichment. We find that it does not for TAA enrichment (adjusted $r^2 = 0.022$, $P = 0.27$, $\lambda = 0.0$; PGLS) but does for ASC enrichment (table 1). This test, although suggestive of a role for cellularity in prediction of ASCs, could be criticized as it overlooks the possible interaction between the TAA and ASCs, namely a gene with TAA may not require ASCs, although in yeast ASCs are used most commonly when associated with TAA (Liang et al. 2005). To consider this issue we divide genes into TAA ending and non-TAA ending (table 1). We find that the connection between cellularity and ASC usage is unaffected. However, there emerges the possibility of ASC enrichment in non-TAA-ending genes being predicted by N_e , although this is sensitive to Bonferroni correction (at $P < 0.05/3$).

As cellularity and N_e are likely to covary, the further (and possibly fairer) comparison is to consider the ASC and TAA enrichment jointly by both cellularity and N_e (table 2). This we do using a multiple regression model within PGLS. The resulting model, using the ASC enrichment scores calculated from all genes, has a significant fit to the data (adjusted $r^2 = 0.45$, $P = 0.011$, $\lambda = 0.0$; multiple regression PGLS) with cellularity remaining a significant predictor ($P = 0.015$), unlike N_e ($P = 0.77$). The presence of a significant relationship for cellularity, but not N_e , with ASC enrichment is also evident both when we restrict our gene sets to non-TAA- and TAA-ending genes. Notably, the earlier observed weak correlation between ASC enrichment and N_e when TAA genes are excluded is removed upon control for cellularity. The same multiple regression method for TAA enrichment finds, as before, that N_e is a significant predictor ($P = 0.0018$), unlike cellularity ($P = 0.37$).

Table 1. ASC Enrichment Scores Assessed for a Relationship with N_e or Cellularity by Linear Regression Using PGLS.

Dependent Variable	Gene Set	N_e	Cellularity
ASC enrichment	All genes	$P = 0.85, r^2 = -0.07$	$P = 0.0021, r^2 = 0.49$
	Non-TAA-ending genes	$P = 0.041, r^2 = 0.23$	$P = 0.0024, r^2 = 0.48$
	TAA-ending genes	$P = 0.98, r^2 = -0.08$	$P = 0.0026, r^2 = 0.48$

Note.—ASC enrichment score was calculated for three different sets of genes for each eukaryotic genome in our data set: all genes, non-TAA-ending genes, TAA-ending genes. The resultant scores were then assessed for a relationship with either N_e or cellularity in a phylogenetically controlled manner. P -values and r^2 -values are given for each scenario. r^2 -values given are the adjusted r^2 -values, hence why some are negative.

Table 2. ASC Enrichment Scores Assessed for a Relationship with N_e or Cellularity by Including Both Parameters in Multiple Regression PGLS.

Dependent Variable	Gene Set	N_e	Cellularity	Adjusted r^2
ASC enrichment	All genes	$P = 0.77$	$P = 0.015$	$r^2 = 0.45$
	Non-TAA-ending genes	$P = 0.25$	$P = 0.032$	$r^2 = 0.50$
	TAA-ending genes	$P = 0.12$	$P = 0.0013$	$r^2 = 0.54$

Note.—ASC enrichment score was calculated for three different sets of genes for each eukaryotic genome in our data set: all genes, non-TAA-ending genes, TAA-ending genes. The resultant scores were then assessed for a relationship with either N_e or cellularity in a phylogenetically controlled multiple regression. P -values are given for each coefficient and the adjusted r^2 -value is reported for each overall model.

These results suggest that N_e , but not cellularity, predicts the usage of the least leaky stop codon, consistent with classical nearly neutral theory. By contrast, enrichment of ASCs is predicted by cellularity and not by N_e , the latter being contrary to the predictions of Rajon and Masel (2011).

GC Content May Play a Minor Role in TAA Enrichment

Aside from N_e and cellularity, it is possible that genome architecture plays a role in both TAA and ASC selection. Might such factors help explain the patchiness of ASC enrichment across species of the same taxonomic group or cellular state? For example, as stop codons are AT-rich, GC-rich genomes contain fewer TAA stops and 3' ASCs by chance and hence might be under higher selection pressure to preserve existing ones. Additionally, shorter average gene size might modulate the intensity of selection, possibly because the costs associated with the misprocessing of long genes are higher owing to greater wastage. Larger average 3' intergenic distance may also ensure that TAA primary stops and ASCs are under stronger selection in order to minimize the amount of misprocessing following stop codon read-through.

Considering all three variables in a multiple regression, we find GC content to be the lone significant coefficient when predicting TAA enrichment ($P = 0.028$) despite the overall model having a near significant fit to the data (adjusted $r^2 = 0.30, P = 0.075, \lambda = 0.0$; PGLS). This relationship is positive, consistent with the view that TAA stops are increasingly preserved in GC-rich genomes. Using the same methodology, ASC enrichment is not predicted by GC content ($P = 0.99$), median gene body length ($P = 0.53$) or median 3' intergenic distance ($P = 0.52$), the overall model being a nonsignificant fit to the data (adjusted $r^2 = -0.18, P = 0.83, \lambda = 1.0$; PGLS).

The above results suggest the most relevant model, at least for TAA usage, may be one in which GC, cellularity, and N_e are employed as predictors. In such a model, GC content does not remain a significant predictor of TAA usage (overall model: adjusted $r^2 = 0.60, P = 0.0042, \lambda = 0.0$; $N_e: P = 0.013$;

cellularity: $P = 0.97$; GC: $P = 0.13$; PGLS). The same model still finds that cellularity ($P = 0.028$), but neither N_e ($P = 0.96$) nor GC ($P = 0.69$), is a predictor of ASC enrichment (overall model: adjusted $r^2 = 0.41, P = 0.031, \lambda = 0.0$; PGLS).

We conclude that GC content plays no more than a minor role in TAA selection at the primary site and that genome architecture is otherwise unimportant in the identification of genome-wide TAA and ASC enrichment.

Marginal Evidence That Genes Associated with Expression in Unicell Mode Contain More ASCs Than Genes Associated with Multicellular Expression in the Same Organism

The above analyses suggest a role for N_e alone in determining the usage of error-preventing TAA, whereas impact mitigating ASCs were predicted by cellularity. The latter, being a novel result, merits further consideration. Indeed, it would be helpful to have a further means to test the cellularity model controlling for N_e . We suggest that this could be achieved by comparing genes expressed exclusively in the unicell mode with those expressed exclusively in the multicell mode in the same species. We consider two such comparisons: between pollen-specific genes and genes expressed more often in the whole plant body (for brevity, pollen-reduced genes) in *Arabidopsis thaliana* and between the unicellular free-living amoeboid phase and the multicellular phase in the cellular slime mold *Dictyostelium discoideum*. Neither comparison is perfect but to some extent as a pair they control for each other's weaknesses. In *A. thaliana* we are, for example, comparing common multicellular expression with rare single cell expression, whereas in *D. discoideum* the unicell mode of expression is the common mode of gene expression. However, in *Arabidopsis* we also have a difference between haploid and diploid expression which is uncontrolled.

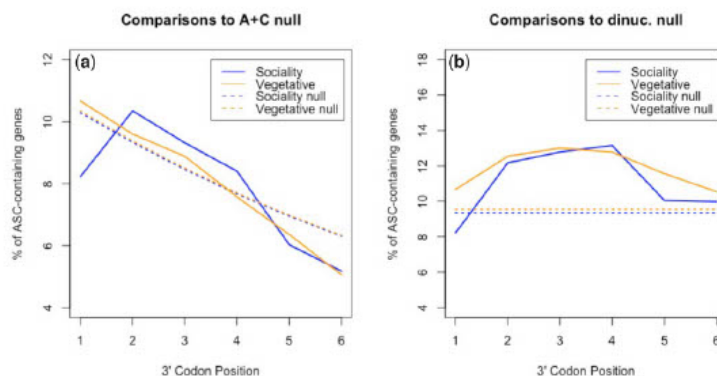


FIG. 5. Assessment of ASC enrichment against (a) the Adachi & Cavalcanti null and (b) dinucleotide-controlled simulations in the 3'-UTRs of sociality- and vegetative-growth associated genes in *Dictyostelium discoideum*. ASC frequencies in vegetative stage expressed genes are enriched ($P < 0.001$; chi-square tests with 1 df) across all positions compared with the dinucleotide null. Against the same null ASCs in sociality genes are enriched at positions +2 to +4 ($P < 0.001$; chi-square tests with 1 df), suggesting that these are the most optimum locations for ASC enrichment within *D. discoideum* within our chosen UTR range. Against the A+C null, vegetative gene ASC frequency is nondeviant ($P = 0.32$; chi-square test with 1 df) whereas sociality gene ASC frequency is significantly lower than expected ($P = 0.0089$, chi-square test with 1 df).

Dictyostelium discoideum Unicell-Expressed Genes Have an Excess of 1 ASCs

The cellularity hypothesis predicts ASCs to be enriched in vegetative (single cell) expressed genes compared with sociality (multicellular) genes. Considering all six 3' codon positions together this is observed ($\chi^2 = 4.76$, $P = 0.029$; chi-square test with 1 df). Examined on a site-by-site basis, ASCs are significantly enriched in vegetative genes compared with sociality genes at position +1 ($P = 0.0035$, chi-square test with 1 df), but no other position within our chosen UTR range (positions +2 to +6: $P > 0.05$). Although there is no significant difference in ASC frequency between vegetative and social genes across positions +2 to +6, ASC frequencies in vegetative stage expressed genes are nonetheless strongly enriched ($P < 0.001$; chi-square tests with 1 df; fig. 5) across all positions compared with dinucleotide-controlled null. ASCs in sociality genes are also enriched beyond dinucleotide expectations at positions +2 to +4 ($P < 0.001$; chi-square tests with 1 df), suggesting that these are the most optimum locations for ASC enrichment within the species within our chosen UTR range. The position +1 difference between vegetative and sociality genes can also be observed when comparing genes against the Adachi and Cavalcanti (2009) null (see Materials and Methods), vegetative gene ASC frequency being nondeviant ($P = 0.32$; chi-square test with 1 df) and sociality gene ASC frequency being significantly lower than expected ($P = 0.0089$, chi-square test with 1 df). This difference is not only consistent with our cellularity prediction but also prima facie consistent with the possible prediction of the fail-safe hypothesis that ASCs should be most strongly selected immediately after the primary stop codon to minimize the error made following read-through.

Might this effect alternatively be owing to a more general thymine nucleotide preference (+4T) following the primary stop that affects position +1 ASC frequency, as seen in bacteria (Major et al. 2002; Wei and Xia 2017)? Contra to this possibility, we find T-starting codons (excluding TGA, TAA, and TAG) to be significantly enriched in sociality genes rather than vegetative genes at this site ($P < 0.0001$, chi-square test with 1 df). This is the opposite to what would be expected if +4T enrichment were to explain the ASC difference observed between vegetative and social genes.

Arabidopsis thaliana Unicell-Expressed Genes Have an Excess of 1 ASCs

Per the cellularity hypothesis, we predict pollen-specific genes to be more likely to contain ASCs than pollen-reduced genes, in spite of them being less expressed. However, UTR-wide ASC frequency (all positions +1 to +6) is not significantly deviant between the two gene sets ($\chi^2 = 1.33$, $P = 0.25$; chi-square test with 1 df). Considering each position in isolation, we find that ASCs in pollen-specific genes are, however, significantly enriched compared with pollen-reduced genes at position +1 ($P = 0.015$, chi-square test with 1 df), consistent with prior evidence for ASC selection at this site in *A. thaliana* (Kochetov et al. 2011). There is no significant difference at any other position within our chosen 3'-UTR range (pos +2: $P = 0.39$, pos +3: $P = 0.90$, pos +4: $P = 0.56$, pos +5: $P = 0.87$, pos +6: $P = 0.93$). Again, we acknowledge the possibility that the position +1 ASC difference occurs due to nucleotide preference in proximity to the primary stop. We reject this possibility, finding no significant difference in T-starting codon (excluding TAA, TGA, and TAG) frequency

253

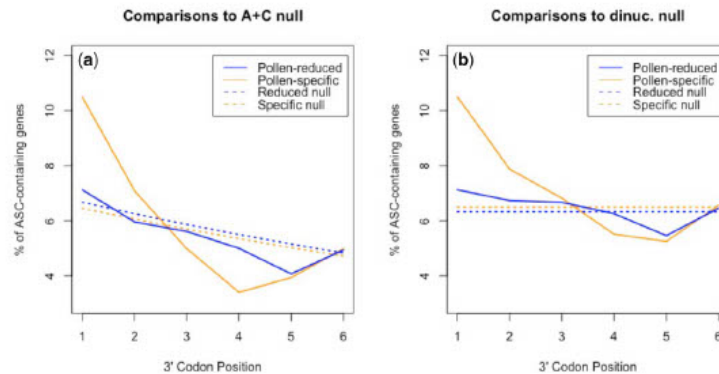


FIG. 6. Assessment of ASC enrichment against (a) the Adachi & Cavalcanti null and (b) dinucleotide-controlled simulations in the 3'-UTRs of pollen-specific and pollen-reduced genes in *Arabidopsis thaliana*. Against dinucleotide-controlled simulations, the ASC frequencies of both sets of genes are significantly enriched at position +1 (pollen-specific genes: $P = 0.0022$, pollen-reduced genes: $P = 0.031$, chi-square tests with 1 df). Consistent with the cellularity hypothesis, significance is an order of magnitude weaker in the case of pollen-reduced genes. When compared with the A+C null, there is evidence of ASC enrichment at position +1 ($P = 0.0019$, chi-square test with 1 df) in pollen-specific but not pollen-reduced genes ($P = 0.23$, chi-square test with 1 df).

at position +1 between pollen-specific and pollen-reduced genes ($\chi^2 = 1.7$, $P = 0.20$; chi-square test with 1 df).

Does this result truly reflect a difference between unicell and multicell expressed genes or might the signal observed merely represent a difference between plant tissues, irrespective of cellularity? We can test this by comparing multicellular tissues. Consistent with there being no difference between tissues of the same state, UTR-wide ASC frequencies between leaf-specific and non-leaf-specific genes are nondeviant ($\chi^2 = 0.24$, $P = 0.63$; chi-square test with 1 df). Taking each position in isolation, there are no differences anywhere between positions +1 to +6 ($P > 0.05$). Similarly, comparing silique-specific genes to non-silique-specific genes also finds no evidence to support deviant ASC frequencies (UTR-wide and all positional results $P > 0.05$; chi-square tests with 1 df).

In our analysis of ASC enrichment in multicellular species (supplementary fig. S1, Supplementary Material online), we detected significant ASC enrichment at position +1 in *A. thaliana*. Is this still the case in pollen-reduced genes that are rarely expressed in the unicell mode or might the trend be predominantly owing to the pollen-expressed genes? To assess this, pollen-specific and pollen-reduced genes were compared with both dinucleotide-controlled and A+C null at position +1. Against dinucleotide-controlled simulations, the ASC frequencies of both sets of genes are significantly enriched at this position (pollen-specific genes: $P = 0.0022$, pollen-reduced genes: $P = 0.031$, chi-square tests with 1 df; fig. 6). However, when compared with the A+C null, there is evidence of ASC enrichment at position +1 ($P = 0.0019$, chi-square test with 1 df) in pollen-specific but not pollen-reduced genes ($P = 0.23$, chi-square test with 1 df). The pollen case study hence concurs with our evidence that

unicellularity may play some role in determining selection for error mitigation.

Weak Evidence for TAA Enrichment in Single-Cell Expressed Genes

Above, we have considered ASC usage as a function of cellularity within the same species. If the prior results comparing between species hold, we do not expect to see much, if any, evidence for TAA enrichment in the single-cell phase. In the slime mold, TAA is found at a slightly higher frequency in vegetative genes (89.5%) than multicell genes (87.7%) ($\chi^2 = 4.4$, $P = 0.036$; chi-square test with 1 df). In *Arabidopsis*, TAA usage is higher in absolute terms in pollen-specific genes (40.1% of genes, compared with 36.0% in pollen-reduced genes) but not significantly so ($\chi^2 = 2.5$, $P = 0.12$; chi-square test with 1 df). We suggest that these present weak support at best for a coupling between cellularity and TAA usage.

Discussion

When considering the evolution of the rate of heritable errors (i.e., mutations), the drift-barrier model for the evolution of the mutation rate (Sung et al. 2012; Lynch et al. 2016) proposes that heritable error rates will be higher when N_e is lower. By contrast, Rajon and Masel (2011) suggest that in species of large effective population size (N_e) there is more effective selection favoring local error mitigation (i.e., more ASCs) hence relaxing selection on global rate modification. We find that the prediction of Rajon and Masel (2011) for ASC enrichment to be higher when N_e is high to not be supported. We also find that their further result of greater HEG/LEG

disparity when N_e is high (Meer et al. 2020) is not replicated as regards selection for TAA.

By contrast, we do find that TAA preference, and hence local error rate, is predicted by N_e although whether absolute global rates also vary with N_e we cannot address. This conclusion assumes that TAA, the preferred stop codon in HEGs in nearly all of our species, because it is associated with lower read-through rates. We showed both enrichment of motifs associated with reduced read-through in HEGs and a general preference for TAA in HEGs, both indicative of selection on TAA to enable low read-through rates. It would be valuable to empirically test this prediction across multiple eukaryotic species, but it is promising that available experimental data are consistent (Cridge et al. 2018). With this caveat, we suggest therefore that our results provide *prima facie* support for the drift-barrier model being applied to understand the fate of mutations affecting the rate of local phenotypic errors.

Why Might the Prior Model Be Wrong?

Rajon and Masel (2011) predict higher ASC usage when N_e is high. This we did not observe. Why might the model of Rajon and Masel (2011) be wrong? We consider several possibilities. Firstly, this may reflect the fact that selection on ASCs is but one mode of locally selected mitigation on read-through. In yeast, potential C-terminal extensions may become pre-adapted for read-through events (evidenced by higher intrinsic structural disorder) (Kosinski and Masel 2020). In mammalian cells, increased hydrophobicity in 3'-UTR encoded sequence has been linked to more efficient translation arrest when termination fails (Hashimoto et al. 2019). However, unless the relative importance of ASCs as a mode of local mitigation itself varies with N_e , there is no reason to suppose that the prediction of Rajon and Masel (2011), that they supported by reference to selection on ASCs specifically, is incorrect.

Secondly, might our analysis be too conservative? That we find N_e to predict ASC enrichment in phylogenetically uncontrolled tests is a provocative result given the use of phylogenetic control in similar studies has been contentious. For example, the associations between genome complexity and N_e described by Lynch and Conery (2003) were observed without control for phylogeny and were subsequently found to not be robust to phylogenetic control (Whitney and Garland 2010). However, more recently, a relationship between N_e and intron size/number has been recovered using PGLS, albeit with more data points and more recent N_e estimates (Wu and Hurst 2015). On balance it seems that the most stringent tests are those that are phylogenetically controlled and hence, to err on the side of caution, we prefer the argument that there is no link between N_e and ASCs. However, given the findings of Wu and Hurst (2015), we acknowledge the possibility that the lack of observed relationship between these two variables may not be resilient to improved sample size and improved N_e estimation. We note too that N_e estimation makes an assumption that the populations are at equilibrium which need not be true but should just factor as a noise variable in the analysis. Nevertheless, the association between N_e and ASC

enrichment must be, at the very least, weaker than that observed between N_e and TAA enrichment given that we find a significant relationship between these traits using the same test with the same data.

Thirdly, the inability of this model to correctly predict the data may stem from the fact that it is importantly incomplete. Rajon and Masel (2011) assume that the only local selection is through the mitigation route (e.g., ASC selection), rate being modified by global trade-offs between translational fidelity and replication rate. As local selection is only available to species with high N_e , they infer that mitigation (by assumption the only mode of local selection) is favored when N_e is high. However, they do not consider the case of local rate modifiers (stop codon usage), which also should respond most efficiently to selection when N_e is high. That ASC selection is not predicted by N_e but local rate modifiers are suggests that their model is incomplete (in an important manner). If so, this suggests caution in assuming veracity of downstream inferences and suggests that it is important to include local rate and mitigation (and global mitigation if this too is relevant). That we could not substantiate their extension (Meer et al. 2020) which assumed higher absolute error rates when N_e is high, causing greater TAA HEG-LEG disparity, is similarly compatible with a problem with model specification.

We suggest that extended models could quite easily explain our observations. Given that local error rates are lower when N_e is high (higher TAA usage), we might expect the lack of clear correlation with patterns of ASC enrichment (and deviation from prior predictions regarding ASC usage): When local error rates are low, selection for ASCs is low because mistakes are rare, when error rates are high this is because selection is too weak to reduce local error rate and hence selection for ASCs must also be weak. We probably need other variables, such as cellularity, to explain between-species variation in ASC enrichment.

Remaining Conundrums

Why Might Selection Act More on Local Error Rates Than on Local Error Mitigation

Above we suggest a synthesis in which N_e modulates the efficiency of local selection such that error rates are lower when N_e is high which in turn dislocates any mitigation selection (ASCs). It leaves, however, several unanswered questions. Firstly, why might selection act more on local error rates than on local error mitigation? The logic of the Rajon and Masel (2011) model is that selection for local effects is associated with small selective coefficients and so most relevant in species with high N_e . This renders the apparent preference for selection on rate over mitigation enigmatic. If an error, such as a read-through, has a mean cost c per event (energy lost through translation of the 3'-UTR, deleterious protein products, etc.) and a rate r (proportional to the number of translation events over time), then the net cost is $c \times r$. A mutation reducing the error rate by delta r (dr) should be associated with positive selection of strength equal to $c \cdot dr$. Similarly, however, reducing the cost by dc has selection of strength

$d_{c,r}$. Given the symmetrical nature of these two, it is not at first sight obvious why selection should be focused more on the rate of error than on the cost per error. This might be because mutations affecting rate are more common than those affecting cost. This seems unlikely as rate-affecting mutations must act at or near the stop codon whereas cost can be reduced by any mutation in 3'-UTR that enables an earlier ASC.

Alternatively, the nature of the mutational effects may be different such that $c_{d,r} > d_{c,r}$, that is, reducing error rate is more visible to selection than reducing error cost. We suggest that one possible reason, at least in our system, is that TAA is so much more efficient than TGA that a TGA->TAA mutation may have an order of magnitude effect on the error rate, as evidenced in bacteria (Sambrook et al. 1967; Roth 1970; Strigini and Brickman 1973; Ryden and Isaksson 1984), but an ASC might save only a relatively small proportion of energy. For example, in AT-rich genomes there might be a stop a certain distance from the focal stop just by chance. An earlier stop codon is likely to reduce costs but not by orders of magnitude (indeed Rajon and Masel [2011] assume that any ASC renders read-through effectively neutral). In addition, ASCs may not be a perfect solution to reducing costs as they may be less effective if not within a correct context (Major et al. 2002). If so, $c_{d,r} \gg d_{c,r}$ may hold and we expect selection on local rate more than local mitigation of costs. That AT-rich genomes likely have a fail-safe stop codon by accident may also explain why we find TAA enrichment to be predicted by GC content. In GC-rich species, the costs of read-through are higher as the distance to the nearest downstream accidental stop codon is longer. Hence, the selection for TAA is stronger than when an incidental ASC is found.

If the above logic is correct, then the results derived here need not be generalizable to other error-prone systems. Although Rajon and Masel (2011) emphasize that the translational read-through system may be a generalizable exemplar (as we too indeed assumed) of error control, if the above logic is correct it would suggest that the preference for local error rate selection is largely contingent on a peculiarity of the system (high error rate variance between stops). For this reason, and contra Rajon and Masel (2011), we caution against generalizing.

If further caution against generalizing is needed, consider the case of selection on splicing. Our model for stronger TAA selection when N_e is high, and no correspondence with error mitigation, might appear to be at odds with evidence for the increased use of another local error rate modifier, ESEs, to reduce the rate of error-prone splicing in low N_e species, these having large and frequent introns (Wu and Hurst 2015). In this case, however, it is proposed that not simply are error rates higher with low N_e , but they are also subject to a ratchet-like accumulation of insertions, each degrading splicing levels that bit more. As a consequence, the accumulation of many splice degrading insertions can enable selection for one exonic mutation enabling increased splice rates (hence increased ESE density, especially in proximity to large introns). In the case of stop codons, there is only one stop codon per gene so there is less possibility of an accumulation of stop

codon degrading mutations. We note that the possibility that weakened local selection might itself increase error rates is not permitted in the models of Rajon and Masel (2011) or Meer et al. (2020).

Why Might Cellularity Matter for ASCs but Not for TAA?

A second enigma concerns the cellularity result. Although N_e does not predict ASC enrichment, that is, local error mitigation, in phylogenetically controlled tests across species, even controlling for N_e , single-celled status predicts ASC enrichment. Comparison of unicell- and multicell-expressed genes within the same species provides some further, albeit marginal, support for this possibility. In *A. thaliana* and *D. discoideum*, ASC frequency immediately proximal to the primary stop is significantly higher in unicell-associated genes compared with multicell-associated genes.

We considered looking at cellularity as a variable as a priori we thought that costs of read-through errors would be different in cellular and multicellular species. The cell replacement argument, indeed, can be evoked to explain the stronger purifying selection on brain-expressed genes (Drummond and Wilke 2008), as neurons cannot be replaced following the accumulation of improperly folded protein. Why then does cellularity matter for error mitigation (ASC usage) but not for error rate (TAA usage)? Were we to have found that for TAA both N_e and cellularity matter, the logic would have been easier to discern (although there is a weak hint of this in the within-species analyses). The result is further compounded by the observation that ASC enrichment in yeasts is most pronounced in TAA-terminating genes (Liang et al. 2005), suggesting that the two processes act synergistically to safeguard HEGs.

We have no good answer to this enigma. It is possible that cellularity does not matter per se, it just happens to covary with some other variable. Indeed, across species we see a strong trend, but the within-species trend is much less robust. One possibility relates to a third parameter we have little or no access to. For example, it is known that one consequence of some prior states is greatly increased rates of translational read-through (Wickner et al. 1995; Harrison 2019). If the distribution of this problem is phylogenetically patchy, but the effect is also more acute in the unicell mode or more commonly seen in unicellular organisms, then this could go some way to explain the phenomenon.

Some sort of enigmatic phylogenetic patchiness seems to be required to explain our between-species ASC enrichment data. Although we found evidence for variable ASC enrichment among different eukaryote groups, there is also considerable unexplained intragroup variability. Indeed, in all phylogenetic groups (vertebrates, invertebrates, plants, fungi, and unicell eukaryotes) we see that some, but not the majority, of the species present genome-wide ASC enrichment. That the species with ASC enrichment in HEGs compared with LECs tend to be those with absolute ASC enrichment underscores this enigmatic patchiness. It is possible that some genomes simply do not value ASC error mitigation and instead rely upon their efficient nonstop decay or other

degradation mechanisms (Kosinski and Masel 2020). It is hence important to note that all genomes likely employ a wide range of error mitigation mechanisms (both local and global) and these may not coevolve identically in all species. Nevertheless, we considered three further possible predictors (GC content, gene size, and 3' intergenic distance) but none was strongly predictive at genome-wide level by PGLS tests. This suggests that even allowing for cellularity and N_e there remains some very patchy predictor of ASC enrichment that we have been unable to discern.

We can support the notion of patchiness by reference to what may be happening in prokaryotes. In some bacteria, stalled ribosomes on mRNAs that do not contain a stop codon (or have had their stop codon read-through) may be rescued by alternative release factors such as ArfA (Keiler and Feaga 2014). One might predict, then, that ArfA-containing genomes have less propensity to select for fail-safe ASCs as the impact of read-through is reduced. We indeed find *prima facie* evidence to suggest that bacterial species with an annotated ArfA gene possess significantly lower ASC frequencies (supplementary fig. S6, Supplementary Material online). Could it be that ASCs selection is dependent on an error mitigation mechanism being missing from some eukaryotes? Understanding such possibilities and access to pan-taxon, high resolution measures of absolute read-through rates would be invaluable.

Materials and Methods

General Methods

All data manipulation was performed using bespoke Python 3.6 scripts. Statistical analyses and data visualizations were performed using R 3.3.3. All scripts required for replication of the described analyses can be found at <https://github.com/ath32/eASCs>. We acknowledge that stop codons function at the mRNA level; however, here we analyze chromosomal DNA sequences and therefore refer to the three stops as TAA, TGA, and TAG.

Extraction and Filtering of 3'-UTR Sequences

Whole-genome sequence and gene annotation data were downloaded from Ensembl release 97 (<https://www.ensembl.org/info/about/species.html>, last accessed September 12, 2019) and EnsemblGenomes release 45 (<http://ensemblgenomes.org>, last accessed September 12, 2019). The main Ensembl set contains primarily vertebrate genomes ($n = 216$), Ensembl Metazoa contains invertebrate genomes ($n = 77$), Ensembl Plants contains plant genomes ($n = 62$), Ensembl Fungi contains fungal genomes ($n = 1,014$), and Ensembl Protists contains unicelled eukaryote genomes ($n = 236$). For all sets, genomes were filtered to retain just one genome per genus to reduce biases due to phylogenetic non-independence that may occur due to oversampling. Species sets for each group were then manually curated to move incorrectly placed species. *Caenorhabditis elegans* and *S. cerevisiae* were removed from the vertebrates set as they are not vertebrates. Unicellular (algae) species in the plants set were removed and added to the unicellular set if not

already present. Nondimorphic yeast species were removed from the fungal set and added to the unicellular set if not already present. *Candida albicans* was also added to the unicell set via bespoke download (available from www.candida-genome.org, last accessed September 12, 2019). This left a final sample of 104 vertebrates, 41 invertebrates, 22 plants, 21 fungi, and 71 unicellular eukaryotes. A full species list for each taxonomic group can be found in Source Data, Supplementary Material online.

Similar to prior analyses (Adachi and Cavalcanti 2009; Ho and Hurst 2019), for every gene in each genome a sequence inclusive of the primary stop followed by 97 nucleotides of the 3'-UTR was extracted by reference to the annotated coding sequence coordinates. Only genes with 3' intergenic space of >100 bp were considered. Resultant sequences were filtered to retain only those 3' sequences made up exclusively of A, T, G, and C, those from genes with one stop after the initiating codon, and those from a gene body with a nucleotide length that is a multiple of 3.

Inferring Substitution Rate

Lists of one-to-one orthologous genes were downloaded for a diverse variety of species triplets from the appropriate Ensembl Biomart repository: 1) primates; *Homo sapiens*, *Macaca mulatta*, *Pan troglodytes*, 2) nematodes; *Caenorhabditis briggsae*, *Caenorhabditis remanei*, and *Caenorhabditis elegans*, 3) *Aspergillus*; *Aspergillus flavus*, *Aspergillus niger*, *Aspergillus oryzae*, 4) *Drosophila*; *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Drosophila simulans*, and 5) *Arabidopsis*; *Arabidopsis halleri*, *Arabidopsis lyrata*, *Arabidopsis thaliana*. Orthologous genes were extracted from the respective genomes and filtered to retain genes with coding sequence of length 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 (Katoh et al. 2005). Alignments with gaps < 10 codons upstream or < 10 "codons" downstream were discarded from further analysis.

Mutations in coding sequence or in the immediate 3'-UTR were reconstructed using a parsimony approach as previously described (Rogozin et al. 2016; Belinky et al. 2018). As each species triplet contains two ingroups and one clear outgroup, ancestral nucleotides can be inferred for each position where the outgroup nucleotide matches that of at least one ingroup. For analysis of substitution rate, we infer mutations at all nucleotide positions from ten codons upstream to ten codons downstream of the stop for TAA-, TGA-, and TAG-terminating genes (where all three orthologs agree on the stop codon). Dividing these mutational counts by the number of valid TAA-, TGA-, and TAG-terminating genes allows the calculation of mutational frequency per site.

Comparing Stop Codon Frequencies between HEGs and LEGs

Experimentally derived protein abundance data were downloaded for all available eukaryotic genomes from PaxDb (Wang et al. 2015). Corresponding whole-genome sequence files were downloaded from the appropriate European

Molecular Biology Laboratory (EMBL) database. A list of the species included can be found in Source Data, [Supplementary Material](#) online. PaxDb external IDs and EMBL locus tags were extracted and matched to generate a sample of genomes and genes for which both PaxDb and EMBL sequence data were available for >400 genes. This filtering produced a sample of 20 eukaryotic genomes, 15 of which belong to multicellular species and 5 belong to unicells. In these genomes, genes that met our filtering criteria that feature in the top and bottom quartiles of expression were defined as HEGs and LEGs, respectively. The frequencies of each primary stop codon (TAA, TGA, and TAG) at each expression level were then calculated and compared. We calculate a standardized frequency difference (SFD) for each codon such that:

$$\text{SFD} = \frac{\text{HEG frequency} - \text{LEG frequency}}{\text{LEG frequency}}$$

Determining Nucleotide Enrichment in HEGs

HEGs were identified and extracted as explained in the previous section. A, C, G, and T counts were counted at each site within our chosen range surrounding the primary stop codon. These counts were compared with a genome-wide null, these being the frequency of each nucleotide at the same positions in all genes regardless of expression level. Comparable “null” counts were calculated as the genome-wide frequency multiplied by the number of genes in the highly expressed set, allowing a comparison of the real observed HEG counts to the null counts using chi-square tests. Significant nucleotide enrichments or depletions were called if the chi-square tests produced a P -value < 0.05 (before Bonferroni correction).

Recognition of ASC Enrichment in Multicellular Eukaryotes

As found in previous studies (Liang et al. 2005; Adachi and Cavalcanti 2009; Ho and Hurst 2019), ASC enrichment in eukaryotes is unlikely to be universally specific to one particular 3' codon position. Hence, we repeat the methodology previously published in our assessment of ASC enrichment in unicellular eukaryotic genomes (Ho and Hurst 2019) in counting the number of genomes in each taxonomic grouping (vertebrates, invertebrates, etc.) that possess ASC enrichment (as determined by chi-square tests) at one or more sites. ASCs at a particular position were considered to be enriched if they were found in raw excess to null expectation and their comparison to null produced a chi-square P -value below 0.05/6 (Bonferroni-corrected, ~ 0.0083). Our P -value threshold dictates that the probability of a genome possessing no significant ASC enrichment at one or more positions by chance is $(1 - 0.0083)^6$ (~ 0.951). Therefore, there is a $1 - 0.951$ (~ 0.049) probability that a genome will contain significant enrichment at one or more positions by chance. We use this probability in a series of binomial tests to consider whether the number of genomes in our data set possessing ASC enrichment was higher, lower, or as expected due to chance. This methodology was repeated for two distinct null models: i) dinucleotide-controlled simulations (Ho and

Hurst 2019) and ii) a degrading frequency null adapted from that first proposed by Adachi and Cavalcanti (2009) and used in our previous ASC analysis (Ho and Hurst 2019):

- i. The dinucleotide-controlled null involves the simulation of 10,000 bespoke null 3'-UTR sequences for a particular genome. Control for genome-specific dinucleotide preferences is facilitated by the capture of nucleotide and dinucleotide frequencies in a Markov-like decision process that directs nucleotide selection in the creation of each simulated sequence. ASC frequencies are calculated in the simulants for comparison with the real genome.
- ii. The adapted Adachi and Cavalcanti (2009) (A+C) null considers only the first in-frame ASC of each UTR sequence. The null ASC frequency expectation at a given position is considered as the probability of not finding a stop at any position upstream multiplied by the probability of finding a stop at any position: First ASC probability = $p[1 - p]^{(n-1)}$, where n is the focal codon position and p is the ASC frequency at any in-frame UTR position.

Calculating an ASC Enrichment Score

To assess the relationship of ASC enrichment with any variable first requires the calculation of an enrichment score. To do this, we first calculate a positional enrichment score (PES) from positions +1 to +6 individually such that:

$$\text{PES} = \frac{\text{Observed} - \text{Expected}}{\text{Expected}},$$

where “observed” is the raw ASC count in the genome at a particular position and “expected” is the expected frequency for that position under the Adachi and Cavalcanti (2009) null hypothesis. The overall enrichment score for each genome used for the correlation analysis was the mean positional enrichment score across all positions. Scores were calculated for 24 genomes for which an existing N_e estimate was available in the literature (Gossmann et al. 2012; Lynch et al. 2016) or a bespoke N_e estimate was possible.

Calculating a TAA Stop Codon Enrichment Score

Similar to how an ASC enrichment score is required for correlation analyses, we must calculate a variable to quantify the extent to which TAA usage is increased in a given genome. For this purpose, we calculate a TAA enrichment score such that:

$$\text{TAA enrichment score} = \frac{\text{TAA usage at primary site} - \text{mean TAA usage downstream}}{\text{mean TAA usage downstream}},$$

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

$$\text{TAA usage} = \frac{\text{TAA freq.}}{\text{TAA freq.} + \text{TGA freq.} + \text{TAG freq.}}$$

Estimation of N_e

New N_e estimations were calculated using previously published species nucleotide diversity (π) and mutation rate (μ) such that:

$$\text{Effective population size } (N_e) = \frac{\pi}{4\mu}$$

All nucleotide diversity, mutation rate, and estimated effective population size values used in this study can be found in Source Data, [Supplementary Material](#) online.

Derivation of GC, Gene Length, and 3' Intergenic Distance

With N_e estimated as above and cellularity considered as a binary trait (0 for multicells and 1 for unicells), we could examine the relationship between enrichment score and these two variables. In addition, GC content was calculated from all of the extracted UTR sequences of a given genome. Median gene body lengths and 3' intergenic distances were calculated for each genome given Ensembl annotations (Source Data, [Supplementary Material](#) online).

PGLS Analysis

Phylogenetically controlled tests were facilitated by PGLS using the caper package in R (<https://CRAN.R-project.org/package=caper>), with lambda (λ) predicted by maximum likelihood. Pagel's lambda statistic (between 0 and 1) reveals the extent to which the phylogeny correctly predicts the covariance observed between species, such that $\lambda = 0$ suggests each data point is phylogenetically independent and $\lambda = 1$ suggests traits are evolving as predicted by tree topology alone. Note that adjusted r^2 -values reported by PGLS may be negative if the fitted model performs worse than null. The phylogenetic trees required for this analysis were generated using TimeTree ([Kumar et al. 2017](#)) and are available at <https://github.com/ath32/eASCs> in nexus format.

Intraspecies Comparisons of Unicell- and Multicell-Expressed Genes

The comparison of genes associated with unicellular development to those associated with multicellular development within the same organism controls for N_e in assessing the role of cellularity in error mitigation selection. Our cellularity hypothesis predicts that unicellular-expressed genes contain more ASCs than multicellular-expressed genes. We test this prediction in two phylogenetically distinct organisms: *D. discoideum* and *A. thaliana*.

In *D. discoideum*, we compare genes associated with vegetative (unicell) growth to social (multicell) growth using data from [de Oliveira et al. \(2019\)](#). In their study, sociality genes were defined as those expressed >90% of the time during the sociality growth phase (> 1 h following nutrient starvation). We consider any genes not included in their social genes list

(available in the source data of their paper) to be associated with vegetative growth.

In *A. thaliana*, we compare genes enriched in pollen (unicell) with those depleted in pollen (multicell). To facilitate this, we acquired a list of pollen-specific and pollen-reduced genes from [supplementary table 1 of Pina et al. \(2005\)](#). Pollen-specific genes are those called present in pollen but absent in seedlings, leaves, siliques, and roots. Pollen-reduced genes are those expressed less often in pollen compared with other tissues.

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to give special thanks to Joanna Masel for her suggestions on how best to frame our results within the literature. We also extend thanks to Atahualpa Castillo-Morales and colleagues for sharing their *Dictyostelium discoideum* expression data and Alan Rice for proof reading. This work was supported by the European Research Council (Grant EvoGenMed ERC-2014-ADG 669207 to L.D.H.).

References

- Abrahams L, Hurst LD. 2017. Adenine enrichment at the fourth CDS residue in bacterial genes is consistent with error proofing for +1 frameshifts. *Mol Biol Evol.* 34(12):3064–3080.
- Adachi M, Cavalcanti AR. 2009. Tandem stop codons in ciliates that reassign stop codons. *J Mol Evol.* 68(4):424–431.
- Alkalaeva EZ, Pisarev AV, Frolova LY, Kisselev LL, Pestova TV. 2006. In vitro reconstitution of eukaryotic translation reveals cooperativity between release factors eRF1 and eRF3. *Cell* 125(6):1125–1136.
- Arribere JA, Cenik ES, Jain N, Hess GT, Lee CH, Bassik MC, Fire AZ. 2016. Translation readthrough mitigation. *Nature* 534(7609):719–723.
- Behringer MG, Hall DW. 2016. Selection on position of nonsense codons in introns. *Genetics* 204(3):1239–1248.
- Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci Rep.* 8(1):9260.
- Bergmann A, Steller H. 2010. Apoptosis, stem cells, and tissue regeneration. *Sci Signal.* 3(145):re8.
- Beznoskova P, Gunisova S, Valasek LS. 2016. Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. *RNA* 22:456–466.
- Bossi L, Roth JR. 1980. The influence of codon context on genetic-code translation. *Nature* 286(5769):123–127.
- Brock CK, Wallin ST, Ruiz OE, Samms KM, Mandal A, Sumner EA, Eisenhoffer GT. 2019. Stem cell proliferation is induced by apoptotic bodies from dying cells during epithelial tissue maintenance. *Nat Commun.* 10(1):1044.
- Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol.* 16(2):107–113.
- Burger R, Willensdorfer M, Nowak MA. 2006. Why are phenotypic mutation rates much higher than genotypic mutation rates? *Genetics* 172(1):197–206.
- Capecchi MR. 1967. Polypeptide chain termination in vitro - isolation of a release factor. *Proc Natl Acad Sci U S A.* 58(3):1144–1151.
- Chen B, Retzlaff M, Roos T, Frydman J. 2011. Cellular strategies of protein quality control. *Cold Spring Harb Perspect Biol.* 3(8):a004374.
- Chen XZ, Zhang JZ. 2013. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol Biol Evol.* 30(7):1559–1562.
- Clegg JB, Weatherall DJ, Milner PF. 1971. Haemoglobin constant spring - a chain termination mutant? *Nature* 234(5328):337–340.

- Grigge AC, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 46(4):1927–1944.
- Cusack BP, Arndt PF, Duret L, Crollius HR. 2011. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet.* 7(10):e1002276.
- de Oliveira JL, Morales AC, Stewart B, Gruenheit N, Engelmoer J, Brown SB, de Brito RA, Hurst LD, Urrutia AO, Thompson CRL, et al. 2019. Conditional expression explains molecular evolution of social genes in a microbe. *Nat Commun.* 10:Article number 3284.
- Dever TE, Green R. 2012. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb Perspect Biol.* 4(7):a013706.
- Dimirova LN, Kuroha K, Tatematsu T, Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J Biol Chem.* 284(16):10343–10352.
- Doronina VA, Brown JD. 2006. When nonsense makes sense and vice versa: non-canonical decoding events at stop codons in eukaryotes. *Mol Biol.* 40:731–741.
- Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.
- Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *BioRx* 2:e01179.
- Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, Willard B, Graham LM, DiCorleto PE, Fox PL. 2014. Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell* 157(7):1605–1618.
- Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L, La Starza R, DiVenio D, Colombo E, Santucci A, et al. 2005. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med.* 352(3):254–266.
- Faure E, Barthélémy R. 2019. Possible functionality of start and stop codons present at specific and conserved positions in animal mitochondrial genes specifying tRNA. *Int J Zool Stud.* 4:23–29.
- Fleming I, Cavalcanti ARO. 2019. Selection for tandem stop codons in dilute species with reassigned stop codons. *PLoS One* 14(11):e0225804.
- Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. *J Mol Biol.* 47(3):238–248.
- Fu Q, Liu Q, Zhang X, Zhai ZS, Wang YZ, Hu MX, Xu XL, Zhang HW, Qin T. 2018. Glucocorticoid receptor regulates expression of microRNA-22 and downstream signaling pathway in apoptosis of pancreatic acinar cells. *World J Gastroenterol.* 24(45):5120–5130.
- Gamba P, Zenkin N. 2018. Transcription fidelity and its roles in the cell. *Curr Opin Microbiol.* 42:13–18.
- Gao HX, Zhou ZH, Rawat U, Huang C, Bouakaz L, Wang CH, Cheng ZH, Liu YY, Zavialov A, Gursky R, et al. 2007. RF3 induces ribosomal conformational changes responsible for dissociation of class I release factors. *Cell* 129(5):929–941.
- Geller AI, Rich A. 1980. A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature* 283(5742):41–46.
- Giacomelli MG, Hancock AS, Masel J. 2007. The conversion of 3' UTRs into coding regions. *Mol Biol Evol.* 24(2):457–464.
- Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol.* 4(5):658–667.
- Gratzmann G, Brechemier-Baey D, Heurgue V, Mora L, Buckingham RH. 1994. Localization and characterization of the gene encoding release factor RF3 in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 91(13):5848–5852.
- Harrison PM. 2019. Evolutionary behaviour of bacterial prion-like proteins. *PLoS One* 14(3):e0213030.
- Hashimoto S, Nobuta R, Izawa T, Inada T. 2019. Translation arrest as a protein quality control system for aberrant translation of the 3'-UTR in mammalian cells. *FEBS Lett.* 593(8):777–787.
- Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. *PLoS Genet.* 15(9):e1008386.
- Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet.* 12(11):756–766.
- Hollingsworth TJ, Gross AK. 2013. The severe autosomal dominant retinitis pigmentosa rhodopsin mutant Ter349Glu mislocalizes and induces rapid rod cell death. *J Biol Chem.* 288(40):29047–29055.
- Inada T, Aiba H. 2005. Translation of aberrant mRNAs lacking a termination codon or with a shortened 3'-UTR is repressed after initiation in yeast. *EMBO J.* 24(8):1584–1595.
- Jackson MP, Hewitt EW. 2016. Cellular proteostasis: degradation of misfolded proteins by lysosomes. *Essays Biochem.* 60(2):173–180.
- Jailon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Saulemont B, Nowadki M, Serrano V, Porcel BM, Séguenot B, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* 451(7176):359–362.
- Jorgensen F, Adamski FM, Tate WP, Kurland CG. 1993. Release factor-dependent false stops are infrequent in *Escherichia coli*. *J Mol Biol.* 230(1):41–50.
- Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res.* 21(12):2096–2113.
- Katoh K, Kuma K-I, Miyata T, Toh H. 2005. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform.* 16(1):22–33.
- Kawashima T, Pellegrini M, Chanfreau GF. 2009. Nonsense-mediated mRNA decay mutes the splicing defects of spliceosome component mutations. *RNA* 15(12):2236–2247.
- Keightley PD, Eyre-Walker A. 2000. Deleterious mutations and the evolution of sex. *Science* 290(5490):331–333.
- Keiler KC, Feaga HA. 2014. Resolving nonstop translation complexes is a matter of life or death. *J Bacteriol.* 196(12):2123–2130.
- Kimura M. 1967. On the evolutionary adjustment of mutation rates. *Genet Res.* 9(1):23–34.
- Klauer AA, van Hoof A. 2012. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *Wiley Interdiscip Rev RNA* 3(5):649–660.
- Kobayashi K, Saito K, Ishitani R, Ito K, Nureki O. 2012. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1 alpha complex. *Nucleic Acids Res.* 40(18):9319–9328.
- Kochetov AV, Volkova OA, Poliakov A, Dubchak I, Rogozin IB. 2011. Tandem termination signal in plant mRNAs. *Gene* 481(1):1–6.
- Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem.* 289(44):30334–30342.
- Kosinski L, Masel J. 2020. Readthrough errors purge deleterious cryptic sequences, facilitating the birth of coding sequences. *Mol Biol Evol.* 37(6):1761–1774.
- Kuersten S, Goodwin EB. 2003. The power of the 3' UTR: translational control and development. *Nat Rev Genet.* 4(8):626–637.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol.* 34(7):1812–1819.
- Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15(5):e1008141.
- Liang H, Cavalcanti AR, Landweber LF. 2005. Conservation of tandem stop codons in yeasts. *Genome Biol.* 6(4):R31.
- Liu Z, Zhang JZ. 2018a. Human C-to-U coding RNA editing is largely nonadaptive. *Mol Biol Evol.* 35(4):963–969.
- Liu Z, Zhang JZ. 2018b. Most m(6A) RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Mol Biol Evol.* 35(3):666–675.
- Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
- Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* 17(11):704–714.

- Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302(5649):1401–1404.
- Major LL, Edgar TD, Yee Yip P, Isaksson LA, Tate WP. 2002. Tandem termination signals: myth or reality? *FEBS Lett*. 514(1):84–89.
- Mayr C. 2019. What are 3' UTRs doing? *Cold Spring Harb Perspect Biol*. 11(10):a034728.
- Meer KM, Nelson PG, Xiong K, Masel J. 2020. High transcriptional error rates vary as a function of gene expression level. *Genome Biol Evol*. 12(1):3754–3761.
- Meng SY, Hui JO, Haniu M, Tsai LB. 1995. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem Biophys Res Commun*. 211(1):40–48.
- Mikuni O, Ito K, Moffat J, Matsumura K, McCaughan K, Nobukuni T, Tate W, Nakamura Y. 1994. Identification of the PRFC gene, which encodes peptide-chain-release factor-3 of *Escherichia coli*. *Proc Natl Acad Sci U S A*. 91(13):5798–5802.
- Namy O, Duchateau-Nguyen G, Hatin I, Hermann-Le Denmat S, Termier M, Rousset JP. 2010. Identification of stop codon read-through genes in *Saccharomyces cerevisiae*. *Nucleic Acids Res*. 38(9):2289–2296.
- Namy O, Duchateau-Nguyen G, Rousset JP. 2002. Translational read-through of the PDE2 stop codon modulates cAMP levels in *Saccharomyces cerevisiae*. *Mol Microbiol*. 43(3):641–652.
- Namy O, Rousset JP. 2010. Specification of standard amino acids by stop codons. In: Atkins JF, Gesteland RF, editors. *Recoding: expansion of decoding rules enriches gene expression*. New York (NY): Springer. p. 79–100.
- Nichols JL. 1970. Nucleotide sequence from polypeptide chain termination region of coat protein cistron in bacteriophage-R17 RNA. *Nature* 225(5228):147–151.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23(1):263–286.
- Pang SY, Wang WH, Rich B, David R, Chang YT, Carbanaru G, Myers SE, Howie AF, Smillie KJ, Mason JJ. 2002. 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*. 87(6):2556–2563.
- Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev*. 53(3):273–298.
- Pina C, Pinto F, Feijo JA, Becker JD. 2005. Gene family analysis of the *Arabidopsis* pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. *Plant Physiol*. 138(2):744–756.
- Rajon E, Masel J. 2011. Evolution of molecular error rates and the consequences for evolvability. *Proc Natl Acad Sci U S A*. 108(3):1082–1087.
- Ramani AK, Nelson AC, Kapranov P, Bell I, Gingeras TR, Fraser AG. 2009. High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective *Caenorhabditis elegans*. *Genome Biol*. 10(9):R101.
- Rodnina MV, Komiya N, Klimova M, Karki P, Peng BZ, Senyushkina T, Belardinelli R, Maracci C, Wohlgenuth I, Samatova E, et al. 2020. Translational recoding: canonical translation mechanisms reinterpreted. *Nucleic Acids Res*. 48(3):1056–1067.
- Rogozin IB, Belyuk F, Pavlenko V, Shabalina SA, Kristensen DM, Koonin EV. 2016. Evolutionary switches between two serine codon sets are driven by selection. *Proc Natl Acad Sci U S A*. 113(46):13109–13113.
- Roth JR. 1970. UGA nonsense mutations in *Salmonella typhimurium*. *J Bacteriol*. 102(2):467–475.
- Roy B, Leszyk JD, Mangus DA, Jacobson A. 2015. Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *Proc Natl Acad Sci U S A*. 112(10):3038–3043.
- Ryden SM, Isaksson LA. 1984. 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*. 193(1):38–45.
- Salas-Marco J, Bedwell DM. 2004. GTP hydrolysis by eRF3 facilitates stop codon decoding during eukaryotic translation termination. *Mol Cell Biol*. 24(17):7769–7778.
- Sambrook JF, Fan DP, Brenner S. 1967. A strong suppressor specific for UGA. *Nature* 214(5087):452–453.
- Sanchez JC, Padron G, Santana H, Herrera L. 1998. Elimination of an HuiFN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J Biotechnol*. 63(3):179–186.
- Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF. 2008. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Mol Cell*. 31(3):360–370.
- Seligmann H. 2011. Error compensation of tRNA misacylation by codon-anticodon mismatch prevents translational amino acid misinsertion. *Comput Biol Chem*. 35(2):81–95.
- Seligmann H. 2019. Localized context-dependent effects of the “ambush” hypothesis: more off-frame stop codons downstream of shifty codons. *DNA Cell Biol*. 38(8):786–795.
- Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol*. 23(10):701–705.
- Shibata N, Ohoka N, Sugaki Y, Onodera C, Inoue M, Sakuraba Y, Takakura D, Hashii N, Kawasaki N, Gondo Y, et al. 2015. Degradation of stop codon read-through mutant proteins via the ubiquitin-proteasome system causes hereditary disorders. *J Biol Chem*. 290(47):28428–28437.
- Stansfield I, Jones KM, Kushnirov VV, Dagkesamanskaya AR, Poznyakovskii AI, Paushkin SV, Nierras CR, Cox BS, Ter-Avanesyan MD, Tuite MF. 1995. The products of the SUP45 (ERF1) and SUP35 genes interact to mediate translation termination in *Saccharomyces cerevisiae*. *EMBO J*. 14(17):4365–4373.
- Stoletzki N, Eyre-Walker A. 2006. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol*. 24(2):374–381.
- Strigini P, Bridoman E. 1973. Analysis of specific misreading in *Escherichia coli*. *J Mol Biol*. 75(4):659–672.
- Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A*. 109(45):18488–18492.
- Tabrez SS, Shama RD, Jain V, Siddiqui AA, Mukhopadhyay A. 2017. Differential alternative splicing coupled to nonsense-mediated decay of mRNA ensures dietary restriction-induced longevity. *Nat Commun*. 8(1):306.
- Tate WP, Cridge AG, Brown CM. 2018. ‘Stop’ in protein synthesis is modulated with exquisite subtlety by an extended RNA translation signal. *Biochem Soc Trans*. 46(6):1615–1625.
- Tate WP, Marsell JB, Mannering SA, Irvine JH, Major LL, Wilson DN. 1999. UGA: a dual signal for ‘stop’ and for recoding in protein synthesis. *Biochemistry (Mosc)*. 64(12):1342–1353.
- Trotta E. 2016. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics*. 17(1):366.
- Tuller T, Carmi A, Vestsgaard K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*. 141(2):344–354.
- Vidal R, Frangione B, Rostagno A, Mead S, Revez T, Plant G, Ghiso J. 1999. A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature*. 399(6738):776–781.
- Vidal R, Revez T, Rostagno A, Kim E, Holton JL, Bek T, Bojsen-Moller M, Braendgaard H, Plant G, Ghiso J, et al. 2000. 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*. 97(9):4920–4925.
- Wagner A. 2005. Energy constraints on the evolution of gene expression. *Mol Biol Evol*. 22(6):1365–1374.
- Wang MC, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues and cell lines. *Proteomics*. 15(18):3163–3168.
- Warnecke T, Hurst LD. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat Rev Genet*. 12(12):875–881.

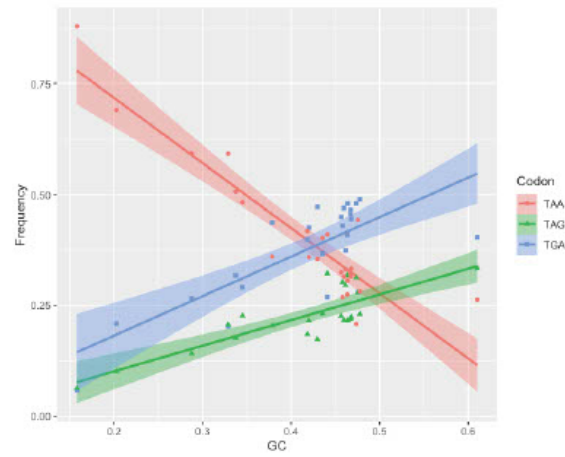
- Wamidek T, Pamley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9(2):r29.
- Weber M, Burgos R, Yus E, Yang J-S, Lluch-Senar M, Serrano L. 2020. Impact of C-terminal amino acid composition on protein expression in bacteria. *Mol Syst Biol.* 16(5):e9208.
- Wei Y, Wang J, Xia X. 2016. Coevolution between stop codon usage and release factors in bacterial species. *Mol Biol Evol.* 33(9):2357–2367.
- Wei Y, Xia X. 2017. The role of +4U as an extended translation termination signal in bacteria. *Genetics* 205(2):539–549.
- Whitney KD, Garland T. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet.* 6(8):e1001080.
- Wickner RB, Masison DC, Edskes HK. 1995. PSI and URE3 as yeast prions. *Yeast* 11(16):1671–1685.
- Willensdorfer M, Burger R, Nowak MA. 2007. Phenotypic mutation rates and the abundance of abnormal proteins in yeast. *PLoS Comp Biol.* 3:2058–2071.
- Williams I, Richardson J, Starkey A, Stansfield I. 2004. Genome-wide prediction of stop codon readthrough during translation in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 32(22):6605–6616.
- Wu XM, Hurst LD. 2015. 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.* 32(7):1847–1861.
- Xiong K, McEntee JP, Porfiro DJ, Masel J. 2017. Drift barriers to quality control when genes are expressed at different levels. *Genetics* 205(1):397–407.
- Yang JR, Maclean CJ, Park C, Zhao HB, Zhang JZ. 2017. Intra and inter-specific variations of gene expression levels in yeast are largely neutral (Nei Lecture, SMCBE 2016, Gold Coast). *Mol Biol Evol.* 34(9):2125–2139.
- Zavialov AV, Buckingham RH, Ehrenberg M. 2001. A posttermination ribosomal complex is the guanine nucleotide exchange factor for peptide release factor RF3. *Cell* 107(1):115–124.
- Zenkin N, Yuzenkova Y, Severinov K. 2006. Transcript-assisted transcriptional proofreading. *Science* 313(5786):518–520.
- Zhang J, Sun XL, Qian YM, LaDuca JP, Maquat LE. 1998. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol.* 18(9):5272–5283.

Supplementary information for: Effective population size predicts local rates but not local mitigation of read-through errors

Alexander T. Ho and Laurence D. Hurst
Molecular Biology & Evolution, msa210.

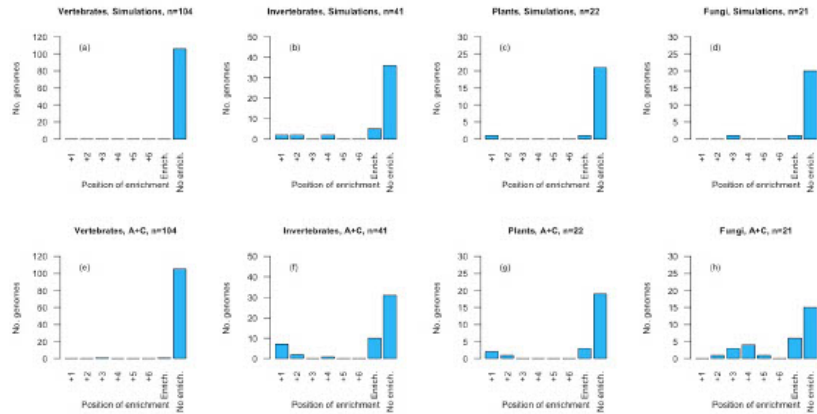
Note: Some of the supplementary figures and tables are extremely small when viewed in this document. To view these best, please refer to the online source: <https://doi.org/10.1093/molbev/msaa210>.

Supplementary Information – Figures followed by Tables



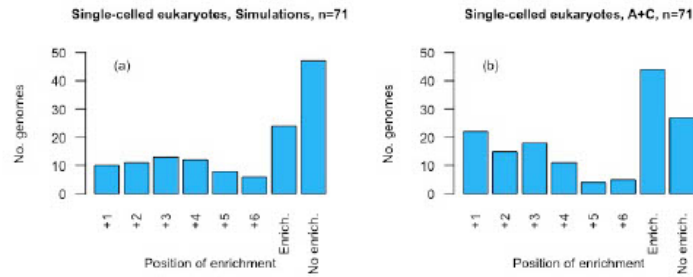
Supplementary fig. S1. Eukaryotic stop codon usage as a function of GC content calculated from the same genes of each genome. The frequencies of TAA, TGA and TAG stop codons were calculated for 25 eukaryotic genomes and assessed for a relationship with GC content. TAA frequency is negatively correlated GC content ($p = 3.2 \times 10^{-6}$, $\rho = -0.803$; Spearman's rank), with TGA ($p = 0.00083$, $\rho = 0.636$; Spearman's rank) and TAG ($p = 0.00012$, $\rho = 0.705$; Spearman's rank) positively correlated. Despite its relationship with GC, TAG is always the least frequent stop codon in all cases.

Supplementary Information – Figures followed by Tables



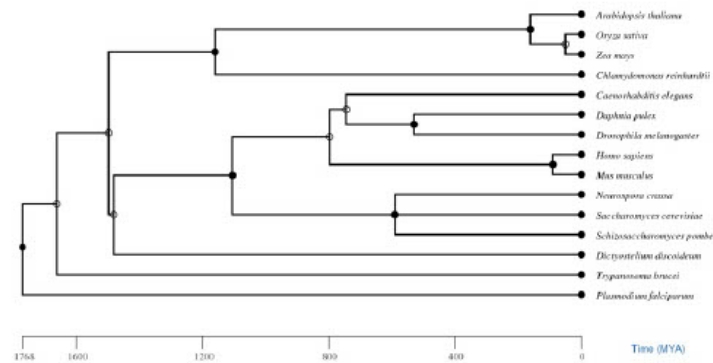
Supplementary fig. S2. Assessment of ASC enriched genomes in four multicellular eukaryote groups. ASC frequencies in each genome were compared to two null models: dinucleotide-controlled simulations (a-d) and the degraded null proposed by Adachi and Cavalcanti (e-h). In vertebrates enriched genomes are significantly under-represented against both dinucleotide-controlled simulations (0/104, $p = 0.0053$, one-tailed binomial test, expected 5.1) and the A+C null (1/104, $p = 0.034$, one-tailed binomial test, expected 5.1). In the invertebrate set enriched genomes are significantly over-represented against both dinucleotide-controlled simulants (5/41, $p = 0.049$, one-tailed binomial test, expected 2.0) and the A+C null (10/41, $p = 2.2 \times 10^{-5}$, one-tailed binomial test, expected 2.0). In plants and fungi there is no significant difference in the number of enriched genomes compared to dinucleotide null (plants: 1/22, $p = 1$, two-tailed binomial test, expected 1.1; fungi: 1/21, $p = 1$, two-tailed binomial test, expected 1.0), but significant over-representation compared to the A+C null (plants: 3/22, $p = 0.0091$, one-tailed binomial test, expected 1.1; fungi: 6/21, $p = 4.0 \times 10^{-4}$, one-tailed binomial test, expected 1.0).

Supplementary Information – Figures followed by Tables



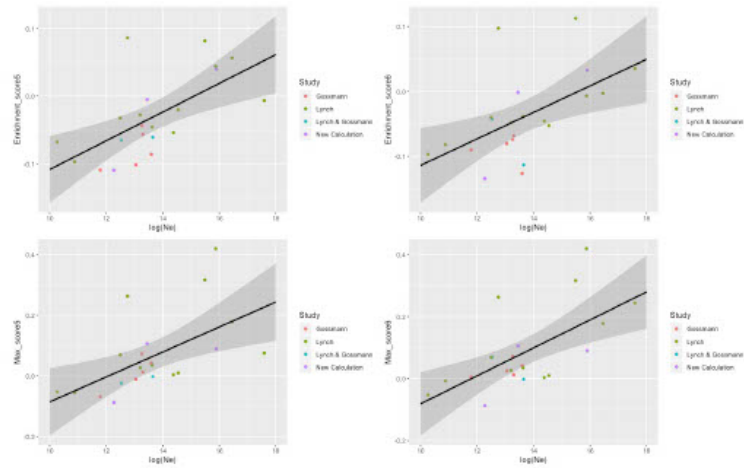
Supplementary fig. S3. Assessment of ASC enriched genomes in unicellular eukaryotes, considering position +1. ASC frequencies in each unicell genome were compared to two null models: dinucleotide-controlled simulations (a) and the degrading null proposed by Adachi and Cavalcanti (b). How do these levels of enrichment compare with those seen in multicellular eukaryotes? Despite the invertebrate set (Supplementary Figure 1) containing the largest proportions of enriched genomes (5/41 simulations and 10/41 A+C), this proportion is significantly lower ($p = 0.022$, $\chi^2 = 5.2$, Chi² test on dinucleotide null proportions; $p = 2.7 \times 10^{-4}$, $\chi^2 = 13.2$, Chi² test for A+C null proportions) than those seen in unicellular species (24/71 simulations and 44/71 A+C). This indicates that, while ASC enrichment is present and common among some multicellular eukaryotic groups, it is rarer than in unicellular organisms.

Supplementary Information – Figures followed by Tables



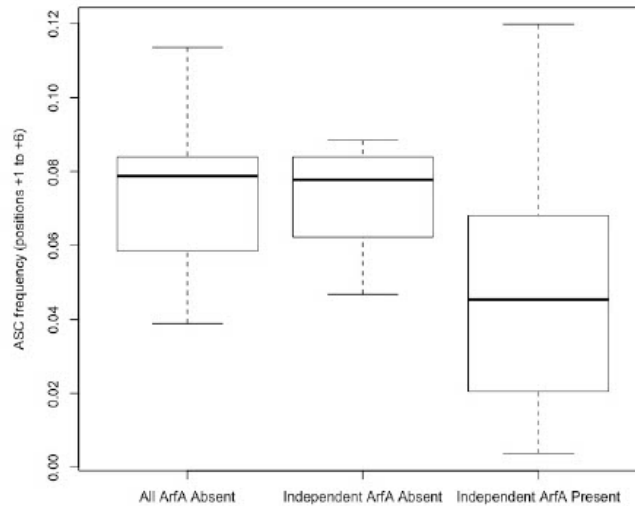
Supplementary fig. S4. Pruned phylogenetic tree describing the eukaryotic species used in PGLS analysis. 15 species were used in our phylogenetically-controlled tests for correlation, pruned from 24 species to remove species with low divergence time. The tree was derived using TimeTree, which requires a species list to be uploaded.

Supplementary Information – Figures followed by Tables



Supplementary fig. S5. Correlation analysis between N_e and four measures of ASC enrichment in 24 eukaryotes. To investigate the possibility of a relationship between N_e and ASC enrichment, we calculate ASC enrichment scores for each genome using two methods. First, we consider a score which takes the average ASC enrichment at each position (from +1 to +6 downstream; see methods of the main paper). Second, given genes possessing an ASC are unlikely to select for a third stop, we consider just the maximum score (at any position from +1 to +6). There is, however, an argument that position +1 should be ignored when considering ASC enrichment due to the possible selection of extended termination motifs immediately proximal to the primary stop. For each method, we hence calculate one score that includes position +1 and one score that excludes it. We find that all four measures of genome ASC enrichment are positively correlated with $\log(N_e)$ before Bonferroni correction (enrichment score including +1: $p = 0.0080$, enrichment score excluding +1: $p = 0.0090$, max score including +1: $p = 0.025$, max score excluding +1: $p = 0.010$).

Supplementary Information – Figures followed by Tables



Supplementary fig. S6. Additional stop codon (ASC) frequency comparison between bacterial genomes with and without an annotated ArfA gene. ArfA is associated with ribosome rescue in mRNAs that do not contain a stop codon in bacteria, hence we predict genomes without an ArfA annotated gene to have greater selection for fail-safe ASCs. To test this prediction, we calculate ASC frequencies for all ArfA-absent genomes (n=212) available for download from EMBL, for all ArfA-absent genomes that are relatively phylogenetically independent (one genome per genus, n=6) and for similarly independent ArfA-present genomes (n = 639). Considering all ArfA-absent genomes, ASC frequencies are significantly lower than observed in the ArfA-present group ($p = 2.9 \times 10^{-15}$, Wilcoxon signed-rank test). This is corroborated when using just independent ArfA-absent species ($p = 0.0060$, Wilcoxon signed-rank test).

Supplementary Information – Figures followed by Tables

Supplementary table T2. Consensus sequences for TGA-terminating highly expressed genes in 19 eukaryotes. Nucleotides A, T, G and C are called if there exists significant enrichment ($p < 0.05$) for these bases compared to null expectations (generated from genes of all expression level in the genome) according to χ^2 tests. A minus sign indicates significant under-enrichment compared to null. An 'N' is called if there exists no significant deviation one way or another for all bases at this position.

Species	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21						
<i>Espea_cakulius</i>	N	N	N	N	N	N	-T/C	N	N	N	N	-T/G	N	N	N	T	A	N	T	G	A	N	N	N	N	N	N	N	-G/C	N	N	N	C	C	A	N	N	N	N	N	N				
<i>Apia_mullera</i>	N	N	N	C	N	N	N	N	N	N	N	-G	N	N	N	N	-A	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N				
<i>Typoceroscus_broad</i>	A/T	N	N	N	N	N	-T	N	N	N	N	N	N	-T/C	A/T	-T	N	A	N	T	G	A	G/C	G	N	N	N	N	N	N	N	-T/G	N	G	N	N	N	A/T	-C	N	N				
<i>Chlamydomonas_rehderii</i>	N	-G	-A	N	N	N	N	N	N	N	N	-A	C	N	N	N	N	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	-A	N	N			
<i>Homo_sapiens</i>	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	-C	C	T	G	A	N	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N		
<i>Bos_taurus</i>	N	N	N	-T	-C	-A	-A/C	C	-A/G	G	G	N	A	N	C	A/C	N	T	G	A	N	N	-A/T	-A/G	N	-T	N	N	-A/C	N	-A	-T/G	N	-A/C	-G	N	-T	N	N	N	N				
<i>Dilysidatum_dendroideum</i>	N	N	N	N	N	N	N	N	N	N	N	-C	N	N	N	N	G	N	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
<i>Gallus_gallus</i>	A/C	N	N	N	A	N	N	N	N	-T	N	G	N	N	N	N	N	A	N	T	G	A	A/C	N	N	N	T/C	N	-T	-G	N	T/G	N	N	N	N	N	N	N	N	N	N	N	N	
<i>Saccharomyces_cerevisiae</i>	-C	N	-G	-C	N	N	G	C	N	N	A	N	N	N	N	N	N	N	N	T	G	A	T/C	-A	N	N	N	-A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
<i>Oryza_sativa</i>	N	C	N	N	N	N	N	N	-T/C	N	N	C	N	N	N	A	N	N	T	G	A	N	G	N	N	G	N	N	N	N	N	N	N	N	N	N	N	N	N	N	-C	N	N		
<i>Xenopus_tropicalis</i>	C	N	C	N	N	N	N	G	N	-A/C	N	G	G	G	N	G	N	T	G	A	A	T	N	T	N	N	T	N	N	N	N	N	N	N	N	N	N	N	N	N	N				
<i>Pleurodon_sabiparvus</i>	N	N	N	N	N	N	-T	N	N	N	N	N	N	N	-G	N	N	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
<i>Arabidopsis_thaliana</i>	N	N	A	N	N	N	N	N	N	N	N	N	T	N	N	N	N	N	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	T	N	N	N
<i>Asophthalma_gambelae</i>	N	N	N	N	-G	N	N	N	N	N	N	N	N	N	N	A/G	A/T	-A/C	T	G	A	N	-T	N	N	N	C	N	-T	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
<i>Rattus_norvegicus</i>	N	N	-A	N	N	N	G	N	N	N	N	N	N	N	N	N	-T/C	-G	-A/T/G	T	G	A	N	N	N	N	N	-A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	C	C	
<i>Schistosoma_mansoni</i>	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
<i>Aspergillus_niger</i>	N	N	N	N	-C	N	N	N	N	N	N	N	N	N	N	N	N	N	N	T	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	A	N
<i>Danio_rerio</i>	N	N	G	G	N	A/C	N	N	N	-T	N	N	N	N	G	G	A	N	T	G	A	-C	N	N	N	-A/C	N	-A	N	C	-G	T	N	-A/C	N	N	N	N	N	N	N	N	N		
<i>Caenorhabditis_elegans</i>	N	N	N	N	N	G	N	A	N	N	N	C	N	-T	N	N	N	-T/C	T	G	A	-C	-G	N	C	C	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	-A/T	

Chapter 4

Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage

Alexander T. Ho and Laurence D. Hurst
Molecular Biology & Evolution, msab326.

This chapter contains work published on 9th November 2021 at MBE, the original and sole place of publication. It thus contains analysis of publicly available data using bespoke scripts that are freely available at the locations cited within the paper. The paper is open access and I have permission as the author to include the article in full (https://academic.oup.com/journals/pages/authors/production_and_publication/online_licensing). The latest version of the published article can be found by following the address: <https://doi.org/10.1093/molbev/msab326>.

Pre-amble

How might we explain between-species variation in non-optimal stop codon usage?

As all the available experimental data for TR rates point towards TAA optimality, the simplest null expectation is that TAA should be the most abundant stop codon in all genomes. Why, then, is this not the case? Chapter 3 provides a parsimonious answer for eukaryotic species: those with low effective population size (N_e) have less efficient selection, are more susceptible to drift, and thus cannot increase their TAA usage. It is less clear why TAA usage does not always dominate in bacterial genomes as all species in this group have large N_e . In this chapter, I re-analyse the long-standing hypothesis that the between-species stop codon usage trends of bacteria can be explained by variation in the relative abundance of class I release factors (RFs), RF1 and RF2.



During translation termination, TAG is recognised by RF1, TGA is recognised by RF2, and TAA may be recognised by either RF1 or RF2. The RF hypothesis may thus explain both a preference for TAA, but also between-genome differences in TGA and TAG usage. Indeed, this has been supported by observed correlations between the RF1:RF2 ratio and the TAG:TGA usage across several bacterial species. In Chapter 2, however, between-species stop codon usage trends in bacteria observed at the canonical stop codon site were also seen both in-frame and out-of-frame in the 3' UTR. This clearly cannot be explained by the RF hypothesis as stop codon trinucleotides in the 3' UTR almost never participate in translation termination.

Here, then, I test several further predictions of the RF hypothesis, questioning whether it is more parsimonious that release factor abundance is adapted to stop codon usage than vice versa. Lack of support for the RF hypothesis would be potentially influential in resolving the underlying causes of stop codon usage by directing attention towards other mutational, selective, or other hypotheses that may be needed to explain the abundance imperfect stop codons in many genomes.

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:				
Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage				
Publication status (tick one)				
Draft manuscript <input type="checkbox"/> Submitted <input type="checkbox"/> In review <input type="checkbox"/> Accepted <input type="checkbox"/> Published <input checked="" type="checkbox"/>				
Publication details (reference)	Ho AT, Hurst LD. 2021. Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage. Mol. Biol. Evol. 39(1): msab326.			
Copyright status (tick the appropriate statement)				
I hold the copyright for this material <input type="checkbox"/> Copyright is retained by the publisher, but I have been given permission to replicate the material here <input checked="" type="checkbox"/>				
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the... Formulation of ideas: 100% Design of methodology: 100% Bioinformatic analyses: 100% Experimental work: N/a Presentation of data in journal format: 100%			
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.			
Signed	<table border="1" style="width: 100%;"> <tr> <td style="width: 70%;"></td> <td style="width: 10%;">Date</td> <td style="width: 20%;">03/03/2022</td> </tr> </table>		Date	03/03/2022
	Date	03/03/2022		

Variation in Release Factor Abundance Is Not Needed to Explain Trends in Bacterial Stop Codon Usage

Alexander T. Ho * and Laurence D. Hurst 

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

*Corresponding author: E-mail: a.tho@bath.ac.uk

Associate Editor: Xuhua Xia

Abstract

In bacteria stop codons are recognized by one of two class I release factors (RF1) recognizing TAG, RF2 recognizing TGA, and TAA being recognized by both. Variation across bacteria in the relative abundance of RF1 and RF2 is thus hypothesized to select for different TGA/TAG usage. This has been supported by correlations between TAG:TGA ratios and RF1:RF2 ratios across multiple bacterial species, potentially also explaining why TAG usage is approximately constant despite extensive variation in GC content. It is, however, possible that stop codon trends are determined by other forces and that RF ratios adapt to stop codon usage, rather than vice versa. Here, we determine which direction of the causal arrow is the more parsimonious. Our results support the notion that RF1/RF2 ratios become adapted to stop codon usage as the same trends, notably the anomalous TAG behavior, are seen in contexts where RF1:RF2 ratios cannot be, or are unlikely to be, causative, that is, at 3' untranslated sites never used for translation termination, in intragenomic analyses, and a cross archaeal species (that possess only one RF1). We conclude that specifics of RF biology are unlikely to fully explain TGA/TAG relative usage. We discuss why the causal relationships for the evolution of synonymous stop codon usage might be different from those affecting synonymous sense codon usage, noting that transitions between TGA and TAG require two-point mutations one of which is likely to be deleterious.

Key words: release factor, stop codons, translation termination, molecular evolution.

Introduction

Most amino acids are encoded by more than one sense codon (Plotkin and Kudla 2011). The three stop codons (TAA, TGA, and TAG) are similarly synonymous in function. In outline, therefore, the well-explored evolution of synonymous sense codons and the less well-explored evolution of stop codon usage might be seen as two sides of the same coin. Being involved in translation termination rather than polypeptide chain extension (Rodnina 2018), however, stop codons do not share molecular machinery with their coding counterparts. Rather than becoming bound to a cognate tRNA (with an accompanied amino acid), the stop codons are instead recognized first by a class I release factor (RF) (Jackson et al. 2012; Rodnina 2018). The resulting complex then becomes bound by a class II RF that mediates peptide release and ribosomal dissociation signaling the end of translation (Jackson et al. 2012; Rodnina 2018).

The precise details of the nature of class I release factors differ between taxa with potential relevance for trends in stop codon usage. In bacteria, TAG is decoded by class I release factor RF1, TGA by class I release factor RF2, and TAA by both RF1 and RF2 (Rodnina 2018). It is most likely that RF1 and RF2 arose via a duplication present in the common ancestor of all bacteria (Burroughs and Aravind 2019). The other domains of life possess one universal RF (Frolova et al. 1994; Inagaki and Doolittle 2000; Jackson et al. 2012; Kobayashi et al. 2012).

While the archaeal RF (aRF1) and the eukaryotic one (eRF1) have a very similar catalytic mechanism to the bacterial RFs, the archaeo-eukaryotic lineage RFs appear to be evolutionarily unrelated to bacterial RF1 and RF2 (Inagaki and Doolittle 2000; Vestergaard et al. 2001; Burroughs and Aravind 2019).

Across bacterial species, the usage of the three stop codons varies considerably (Povolotskaya et al. 2012; Korkmaz et al. 2014; Belinky et al. 2018; Ho and Hurst 2019). This variation is not parsimoniously explained by simple covariation with GC content. While TAA usage is negatively correlated with GC and TGA usage is strongly positively correlated, TAG usage, despite having an identical nucleotide content to TGA, is mostly low and unresponsive to GC pressure (Povolotskaya et al. 2012; Korkmaz et al. 2014; Ho and Hurst 2019). In all bacterial species, on average about 20% of all stops is TAG no matter what their GC content (assayed as mean GC3 or GC of the whole genome). As the read-through rate of TAG is consistently lower than that of TGA across species (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen et al. 1993; Meng et al. 1995; Sanchez et al. 1998; Tate et al. 1999; Wei et al. 2016; Cridge et al. 2018), the avoidance of TAG, aligned with evidence for the selective importance of read-through avoidance (Bossi and Roth 1980; Wei and Xia 2017; Cridge et al. 2018; Li and Zhang 2019) renders TAG avoidance especially enigmatic. It is suggested that variation in RF1/RF2 cellular abundance may explain this and other

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Mol. Biol. Evol. doi:10.1093/molbev/msab326 Advance Access publication November 09, 2021

1

features of stop codon usage (Sharp and Bulmer 1988; Korkmaz et al. 2014; Wei et al. 2016).

Sharp and Bulmer (1988) presented one of the earliest RF-based hypotheses to explain bacterial stop codon usage noticing 1) that TAA is the most common stop codon in bacteria and 2) that TAA stop codons can be decoded by both RF1 and RF2 class I release factors while TGA and TAG can only bind one. We now also know that TAA is the preferred stop codon in highly expressed bacterial genes (Korkmaz et al. 2014) consistent with this model. It has additionally been postulated by Wei et al. (2016), by extension of the codon-anticodon adaptation hypothesis (Ikemura 1981, 1992; Akashi and Eyre-Walker 1998; Xia 1998; van Wieringh et al. 2011; Prabhakaran et al. 2014; Chithambaram et al. 2014a, 2014b), that bacterial species adjust their stop codon usage to their relative expression of RF1 and RF2, a model we dub the "release factor (RF) hypothesis". Not only would this explain the apparent preference for TAA, but also potentially the difference in the behavior of TGA and TAG usage against GC content (Povolotskaya et al. 2012; Korkmaz et al. 2014; Wei et al. 2016; Ho and Hurst 2019).

While the RF hypothesis provides an attractive model for explaining TAA optimality in bacteria, both prokaryotic and eukaryotic evidence suggests it is not necessary. While the highly expressed genes of bacteria preferentially use TAA stops (Korkmaz et al. 2014; Wei et al. 2016), so do those of humans (Trotta 2016) where all three stops are decoded by a universal RF. Recently, we also showed that this effect probably is not unique to humans but common across eukaryotes as, controlling for local GC content, TAA usage positively correlates with effective population size (N_e) (Ho and Hurst 2021). The strongest current explanation for universal TAA optimality is selection for reduced translational read-through (TR), the failure to terminate translation. Indeed, experimentally derived TR rates in both prokaryotes and eukaryotes demonstrate TAA to be the least "leaky" (followed by TAG, with TGA being the most prone to TR) (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen et al. 1993; Meng et al. 1995; Sanchez et al. 1998; Tate et al. 1999; Wei et al. 2016; Cridge et al. 2018). We can be confident that TAA is under selection for this purpose due to the similar enrichment of TR-modulating 3' flanking motifs that reduce TR rates (Bossi and Roth 1980; Wei and Xia 2017; Cridge et al. 2018). At the very least, TR provides a significant selection pressure for TAA regardless of the cellular RF environment.

Nevertheless, the RF hypothesis does receive support from correlations between RF1:RF2 ratios (assessed at mRNA level by real-time qPCR in *Escherichia coli*, *Mycobacterium smegmatis*, and *Bacillus subtilis* and protein level by Western blot in *E. coli*) and TAG:TGA relative stop codon usage observed in a few different bacteria (Korkmaz et al. 2014). It has been noted, however, that analysis of mRNA levels is less informative than protein abundance data as RF2 is translationally regulated (Craigen et al. 1985; Dorly et al. 1990; Wei et al. 2016). Subsequent assessment of the RF1:RF2 and TAG:TGA

correlation using protein abundance data by Wei et al. (2016) nonetheless corroborated the Korkmaz et al. (2014) result in a wider range of species spanning Proteobacteria, Firmicutes, Cyanobacteria, Actinobacteria, and Spirochetes ($n = 14$). Crucially, Wei et al. (2016) also identify that RF2 is exceptionally low when A + T content at codon third sites (AT3) is high across species. While the RF1:RF2 ratio provides a rationale to explain why TAG and TGA behave differently, that RF2 abundance covaries with AT3 could explain why this difference may vary with GC pressure. With regards to between-species TAG usage trends across GC contents, the authors speculate that at low GC contents mutation bias favors TAA (the most AT-rich stop codon) over TGA and TAG, at mid-range GC contents TAA is favored by selection while TGA is preferred over TAG as RF2 levels exceed RF1, and at high GC contents RF2 is exceptionally high which favors TGA and keeps TAG at low frequency (Wei et al. 2016). The RF hypothesis can hence theoretically explain both the preference for TAA stops (even if it is not the only hypothesis) and the unusual biology of TAG.

Key to understanding the necessity of the RF hypothesis is to solve the TGA/TAG problem. While the above evidence is consistent with the notion that stop codon usage adapts to RF1/RF2 relative levels the causal arrow could predominantly be in the opposite direction: RF1/RF2 usage may adapt to stop codon usage rather than vice versa.

One approach to support the notion that RF1:RF2 ratios adapt to stop usage trends would be to resolve the causes of the anomalous behavior of TAG with respect to GC content. If, for example, we could demonstrate some process that explains invariant TAG usage across genomes that differ widely in GC content then that would lend considerable strength to the notion that another force acts on TAG usage and RF1:RF2 ratios instead respond to equilibrium TAG/TGA ratios. This method is currently problematic. Previously it has been assumed that GC content at putatively neutral sites in a genome must indicate the mutational bias in that genome (Knight et al. 2001). If true, then the neutral AT%/neutral GC% ratio should predict $f(\text{TAA})/f(\text{TGA})$ or $f(\text{TAA})/f(\text{TAG})$ with a line of slope 1. Deviation from this line could then be employed to infer a fixation bias. While this assumes no complex k -mer dependent mutation biases (e.g. CpG hypermutability), by far the greater difficulty is that analyses of rare SNPs (Hershberg and Petrov 2010; Hildebrand et al. 2010), mutation accumulation lines (Long et al. 2018), and parent-offspring trios (Smith et al. 2018) report that mutation appears to be universally GC->AT biased. As a consequence, nucleotide content at putatively neutral sites exceeds mutational equilibrium in GC-rich genomes and in GC-rich domains within genomes (Smith et al. 2018). For the same reason, we consider the force causing high GC content to be "GC pressure", but leave unresolved exactly what the force is, beyond knowing that it is not mutation bias. No matter what the cause, we cannot infer the role of selection by inferring differences between nucleotide usages at a focal site (the termination codon in our case) and some putatively "neutral" site.

An alternative approach that side-steps these problems, and the one taken here, is to ask whether the same

anomalous trend in TAG usage as a function of GC content is seen when the RF1:RF2 release factor hypothesis does not apply. Here, we focus on the discrepancy between TGA and TAG usage as a function of GC content. We consider several tests. First, if RF1 and RF2 abundance were to explain between-species stop codon usage trends in bacteria then one would not expect to see the same trends in stop codon trinucleotides outside of the canonical termination context. To consider this, we extend our prior analysis (Ho and Hurst 2019) and examine trends in 3' UTR including after the first downstream stop, allowing for the possibility that the first 3' UTR stop codon may be a fail-safe codon (Major et al. 2002; Liang et al. 2005; Adachi and Cavalcanti 2009). Second, assuming that termination control should be approximately the same for all genes within any given genome, any covariance between stop codon usage and GC content observed between bacteria genomes should not be repeated in intra-genomic analysis. Third, we take advantage of the fact that the other domains of life possess one universal RF (Frolova et al. 1994; Inagaki and Doolittle 2000; Jackson et al. 2012; Kobayashi et al. 2012) and ask whether the trends seen across bacteria are seen across these groups, most notably Archaea these being close relatives of bacteria. If they are, this would lend weight to the hypothesis that stop codon usage trends do not require an RF1/RF2 mediated rationale.

Given that TGA and TAG have identical nucleotide contents, an anomalous trend in TAG usage has two possible diagnostics. We could consider evidence of a significant positive relationship between TGA usage and GC content in combination with no significant correlation between TAG usage and GC content as one diagnostic. The finding of no significant correlation for TAG versus GC is, however, potentially sensitive to sampling and sample size. We thus also consider as a more general alternative a difference in slope between TAG and TGA usage as a function of GC, with the slope being steeper for TGA.

We test the above predictions using *in silico* analysis of bacterial, eukaryotic, and archaeal whole-genome sequences. In agreement with our recent studies (Ho and Hurst 2019, 2021), we find stop codon usage to be highly consistent at the canonical stop site and at genomic loci not involved in translation termination. We too find that stop codon usage trends are consistent between bacteria and archaea. Intra-genomically, stop codon usage can be predicted by local (genic) 3' UTR GC content within the genomes of most bacteria and in humans with TAG anomalous in both. These results are largely consistent with the hypothesis that there are (unspecified) forces external to RF abundance that dictate genome-wide stop codon usage in bacteria and elsewhere. Evoking Occam's razor, we propose that in bacteria, it is more likely that RF expression adapts to stop codon usage rather than vice versa. We suggest that the evolution of synonymous stop codons and the evolution of synonymous sense codons may well operate according to different principles, the possible reasons for which we discuss. We also provide additional data indicating that TAA optimality appears to be universal regardless of RF diversity.

Results

Between-Species Stop Codon Usage Trends in Bacteria are Consistent Outside of the Canonical Termination Site

Our first test of the RF hypothesis concerns stopping codon usage inside and outside of the canonical termination context. As TAA, TGA, and TAG trinucleotides do not function in translation termination in untranslated sequence, the RF hypothesis poses that there is no reason why their usage (and cross-species trends) should reflect what is seen at the canonical stop codon site. Here, then, we consider the relative usage of trinucleotides TAA, TGA, and TAG at the canonical stop site and the 3' UTRs across bacteria. Note that for this analysis, we use 3' UTR GC content as our proxy for GC pressure to mitigate the confounding impacts of expression level and codon usage bias. We start by considering whether stop codon usage trends at the focal termination site are the same as those in 3' UTR. Previously we found that stop codon usage is consistent between the canonical stop site and the 3' UTR across bacteria (Ho and Hurst 2019). However, as 3' additional stop codons (ASCs) have been proposed as a potential fail-safe mechanism to prevent phenotypic error (Major et al. 2002; Liang et al. 2005; Adachi and Cavalcanti 2009) the first in-frame "stop codon" downstream of the canonical stop might be subjected to similar termination selection pressures (possibly mediated by RF1:RF2 ratios). To control for this, we expand our analysis to consider only in-frame "codons" downstream of the first occurring in frame ASC.

We find TGA is positively correlated with GC content at all three sites (fig. 1; Spearman's rank: all $P < 2.2 \times 10^{-16}$, $\rho = 0.89$ at the canonical stop, $\rho = 0.93$ in the 3' UTR, $\rho = 0.86$ downstream of ASCs), TAA is negatively correlated with GC content at all three sites (Spearman's rank: all $P < 2.2 \times 10^{-16}$, $\rho = -0.93$ at the canonical stop, $\rho = -0.95$ in the 3' UTR, $\rho = -0.87$ downstream of ASCs), and TAG is unresponsive to GC pressure at all three sites (Spearman's rank: $P = 0.79$, $\rho = -0.010$ at the canonical stop, $P = 0.42$, $\rho = -0.032$ in the 3' UTR, $P = 0.059$, $\rho = -0.074$ downstream of ASCs). Notably, TAG usage is decoupled from TGA usage in 3' UTR sequences in the same way as at the canonical stop site. TGA and TAG usage slopes against 3' UTR GC content are significantly different from each other when considering the canonical stop (TGA: 0.013, TAG: 0.00028), 3' UTR trinucleotides (TGA: 0.011, TAG: -1.1×10^{-5}), and 3' UTR in-frame codons downstream of the first occurring ASC (TGA: 0.012, TAG: -0.00056). These results, and prior similar results for out-of-frame "stop" codons in 3' UTR (Ho and Hurst 2019), suggest that RF1:RF2 dynamics are not required to explain differential trends seen for TGA and TAG.

The Relationships between Stop Codon Usage and GC Content Observed Between-Species Are Also Observed in Within-Genome Analysis

The RF hypothesis predicts variation in stop codon usage between species possessing different RF1:RF2 profiles. It conversely predicts that no such variation should exist between

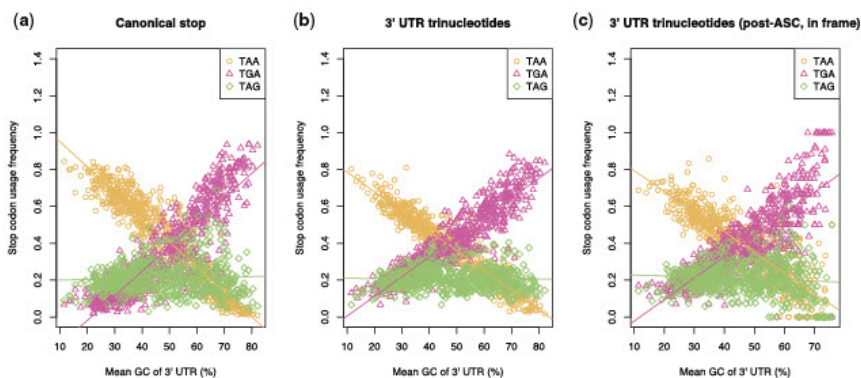


Fig. 1. Stop codon usage at (a) the canonical stop codon site, (b) any 3' UTR site irrespective of frame, and (c) in-frame positions downstream of an ASCs as a function of 3' UTR GC content across 644 phylogenetically independent bacteria. TGA is positively correlated with GC content at all three sites (Spearman's rank: all $P < 2.2 \times 10^{-16}$, $\rho = 0.89$ at the canonical stop, $\rho = 0.93$ in the 3' UTR, $\rho = 0.86$ downstream of ASCs), TAA is negatively correlated with GC content at all three sites (Spearman's rank: all $P < 2.2 \times 10^{-16}$, $\rho = -0.93$ at the canonical stop, $\rho = -0.95$ in the 3' UTR, $\rho = -0.87$ downstream of ASCs), and TAG is unresponsive to GC pressure at all three sites (Spearman's rank: $P = 0.79$, $\rho = -0.10$ at the canonical stop, $P = 0.42$, $\rho = -0.032$ in the 3' UTR, $P = 0.059$, $\rho = -0.074$ downstream of ASCs).

genes subject to equal RF1 and RF2 availability. As genes within the same genome should be exposed to approximately the same RF environment within any given cell the hypothesis that stop codon usage is dictated by forces other than RF1:RF2 ratios predicts that GC content and stop codon usage should be intra-genomically correlated. A key question then is whether intra-genomically TGA and TAG respond differently to local GC pressure.

To test for a relationship between stop codon usage and GC content of the same gene, logistic regression was used to predict stop codon usage (as a binary variable where 1 = present and 0 = absent for TAA, TGA, and TAG stop codons of a particular gene) using GC content calculated from the 3' UTR (Source data, available <https://github.com/ath32/RF>). Given the highly noisy nature of the GC content of 3' UTR (compared to mean UTR GC across all genes in a genome) and the fact that each gene presents only a single stop codon, this analysis is likely to be noisy and underpowered.

We find that TAA usage is well predicted by, and negatively associated with, 3' UTR GC content across our sample of 644 bacteria. In 640/644 species, there is a negative coefficient when predicting TAA usage with 3' UTR GC, significantly more than expected by null chance (Binomial test with null $P = 0.5$, $P < 2.2 \times 10^{-16}$). In 624/644 species, the predictive nature of 3' UTR GC on TAA usage is both negative and significant with $P < 0.05$, this also far exceeding chance (Binomial test with null $P = 0.05$, $P < 2.2 \times 10^{-16}$). As predicted by GC pressure alone, we find the opposite trends for both TGA and TAG usage. 3' UTR GC content is a positive predictor of TGA usage in 622/644 species and of TAG usage in 545/644 species, both ratios being more than expected by

chance (Binomial tests with null $P = 0.5$, both $P < 2.2 \times 10^{-16}$). This positive predictive power is significant in 540/644 species with respect to TGA usage and 413/644 species with respect to TAG usage, more than chance in both cases (Binomial tests with null $P = 0.05$, both $P < 2.2 \times 10^{-16}$).

That we see evidence for covariance with GC content for both TGA and TAG could be considered consistent with predictions of the RF hypothesis—when RF variation is removed TGA usage and TAG usage respond similarly with GC content. We, however, find evidence of an intragenomic disconnect between TGA and TAG. The estimated coefficient in models predicting TGA usage using 3' UTR GC content is higher in absolute terms than equivalent models that predict TAG usage in 373/644 cases, this being more than expected by chance (Binomial test with null ratio = 0.5, $P = 6.7 \times 10^{-5}$). This implies TGA usage is more strongly coupled with local GC pressure than TAG usage, hence the factors underlying the between-species TGA/TAG difference could also be present within genomes where RF environment is approximately the same for all genes.

There, however, exists the possibility that for most bacterial genomes that GC content is approximately the same for all genes and hence the signals described above are mostly noise. To mitigate this, we analyze a published subset of bacteria defined as having unusually high intragenomic GC content variation (Daubin and Perriere 2003). We find the same trends are robustly supported (Supplementary table 1, Supplementary Material online). TAA is significantly predicted by genic GC3 in 10/18 genomes, GC3 being a negative predictor in all 10 cases (Binomial test with null ratio = 0.5, $P = 0.0020$). TGA is significantly predicted by genic GC3 in 9/

18 genomes, GC3 being a positive predictor in 8/9 cases (Binomial test with null ratio = 0.5, $P = 0.039$). TAG is significantly predicted by genic GC3 in 10/18 genomes, with GC3 a positive predictor in 5/10 cases but a negative predictor in the other 5. For the great majority of genomes within this sample (14/18, 77.8%), the TGA versus GC slope is also more positive than the TAG versus GC slope, consistent with the between-species analysis.

There too exists the possibility that RF1 and RF2 abundance varies throughout the cell cycle. To control for this, one might consider intragenomic analysis of a genome that possesses both substantial GC content variation and only one release factor for translation termination. We hence consider the human genome which possesses both traits due to its isochore structure (Eyre-Walker and Hurst 2001; Galtier et al. 2001; Duret and Galtier 2009) and one universal class I release factor (eRF1). Alongside 3' UTR GC content, in the human genome, we may too consider coding sequence GC3 content given the relative lack of strong synonymous codon usage bias (Vogel et al. 2010; Plotkin and Kudla 2011). We find that not only can intragenomic stop codon usage be predicted by GC pressure in most bacteria; there too exists a relationship between stop codon usage and GC content in the human genome.

When genes are grouped by their 3' UTR GC content into 10 equal bins (fig. 2a), we find TGA usage is positively correlated with 3' UTR GC content (Spearman's rank: $P = 0.0035$, $\rho = 0.85$) and TAA usage is negatively correlated with 3' UTR GC content (Spearman's rank: $P = 0.0068$, $\rho = -0.82$). TAG usage does not significantly correlate with 3' UTR GC content (Spearman's rank: $P = 0.080$, $\rho = 0.59$), in agreement with the disconnect observed between TGA and TAG usage between bacterial species. Results are slightly different when we employ GC3 as the measure of local GC content (fig. 2b). Here, we find TGA usage is positively correlated with GC3 content (Spearman's rank: $P < 2.2 \times 10^{-16}$,

$\rho = 0.98$), TAA usage is negatively correlated with GC3 content (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = -0.99$), and, unlike the 3' UTR GC measure, TAG usage is positively correlated with GC3 content (Spearman's rank: $P = 0.0020$, $\rho = 0.88$). However, TAG usage is nonetheless decoupled from TGA usage as indicated by a significantly shallower slope (Z-test on slopes; 0.0016 for TAG, 0.0049 for TGA, $P = 4.8 \times 10^{-9}$) and lower frequency in absolute terms at all GC3 contents. We conclude that the anomalous behavior of TAG can be seen in the absence of RF1:RF2 ratio variation.

Despite Their Shared Release Factor Recognition, TGA, and TAG Usage are Decoupled across Archaea

Just as the RF hypothesis makes predictions as to what to expect in between-species and within-species analysis of bacteria, it too makes predictions regarding the closely related archaea. As archaea use only one class I RF (aRF1) to decode all three stop codons, stop codon usage variation between archaeal species cannot be attributed to RF abundance. If between-species trends were to match the trends seen in bacteria, this would suggest forces external to RF abundance are more significant influencers of stop codon usage. We hence here analyze the stop codon usage of 106 archaeal genomes as a function of 3' UTR GC content (fig. 3).

We find that, on initial examination, the stop codon usage trends observed in archaea do somewhat resemble what we see in bacteria (fig. 1). TAA and TGA trends are consistent with bacteria observations in archaea, TGA usage being positively correlated with 3' UTR GC content (Spearman's rank $P < 2.2 \times 10^{-16}$, $\rho = 0.76$) and TAA usage being negatively correlated with 3' UTR GC content (Spearman's rank $P < 2.2 \times 10^{-16}$, $\rho = -0.90$). TAG usage appears to behave slightly differently between the two domains, however. In bacteria, TAG is unresponsive to GC pressure (fig. 1), while TAG is positively correlated with 3' UTR GC content across archaea (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = 0.57$).

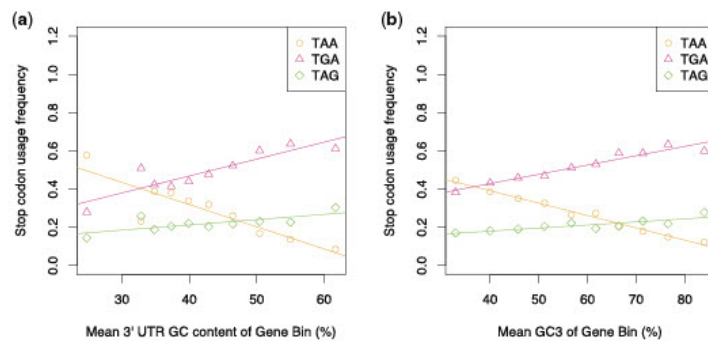


FIG. 2. Stop codon usage as a function of (a) 3' UTR GC content and (b) GC3 content across the human genome. TGA usage is positively correlated with both 3' UTR GC content (Spearman's rank: $P = 0.0035$, $\rho = 0.85$) and GC3 content (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = 0.98$). TAA usage is negatively correlated with both 3' UTR GC content (Spearman's rank: $P = 0.0068$, $\rho = -0.82$) and GC3 content (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = -0.99$). TAG usage is not significantly correlated with 3' UTR GC content (Spearman's rank: $P = 0.080$, $\rho = 0.59$) but is positively correlated with GC3 content (Spearman's rank: $P = 0.0020$, $\rho = 0.88$).

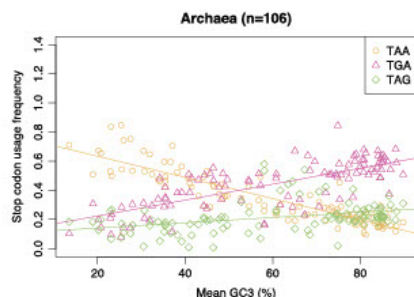


FIG. 3. Stop codon usage as a function of 3' UTR GC content across 106 archaea. TGA usage is positively correlated with 3' UTR GC content (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = 0.76$). TAA usage is negatively correlated with 3' UTR GC content (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = -0.90$). TAG is positively correlated with 3' UTR GC content (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = 0.57$).

Despite this, there is evidence that TGA and TAG usage are differently correlated in archaea, as observed in bacteria. In archaea, the slope of TAG usage against 3' UTR GC (0.0032) is significantly shallower than the comparable TGA usage trend (0.0079) (Z-test on slopes; $P = 2.1 \times 10^{-7}$). The RF hypothesis gives no parsimonious explanation for this given that TGA and TAG stop codons are decoded by the same class I release factor in archaea. All these results are repeated when mean GC3 is used as the proxy for GC pressure in the analysis of archaea (supplementary fig 1, Supplementary Material online).

The above analysis, however, is in some regards unfair. Close examination of the bacterial data suggests that at the highest GC levels TAG usage is especially low (fig. 1). The archaeal genomes are, however, not represented in this more extreme end of GC contents. Might the finding of a positive correlation between TAG usage and GC content in archaea but not in bacteria reflect this sampling issue? To assess this possibility, for each archaeal species we select the nearest bacterial species by mean 3' UTR GC content to be used for comparison between domains. For the two equally sized GC-matched sets of genomes, we then repeat the correlation analysis and test for differences between linear models fitted to TAG usage against 3' UTR GC content (fig. 4).

In this GC-matched bacterial data set, we find TAG usage to be positively correlated with 3' UTR GC content (Spearman's rank: $P = 0.0037$, $\rho = 0.28$), consistent with the archaeal result. Thus, the prima facie modest differences between archaea and bacteria in TAG usage trends appear mostly to be explained by sampling differences. This underscores the notion that TAG's weak response to GC pressure is independent of RF1:RF2. Note, however, that the slopes of TAG usage against 3' UTR GC content are still slightly different between the two groups (Z-test: archaea slope = 0.0032, GC-matched bacteria slope = 0.0013, $P = 0.01$).

We conclude that the stop codon usage trends against GC pressure in bacteria and archaea are approximately the same

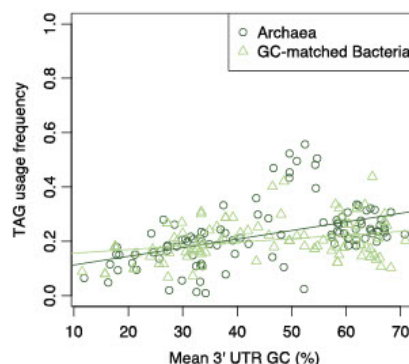


FIG. 4. Stop codon usage as a function of 3' UTR GC content across 106 archaea and GC-matched bacteria. TAG is positively correlated with 3' UTR GC content in both archaea (Spearman's rank: $P < 2.2 \times 10^{-16}$, $\rho = 0.57$) and GC-matched bacteria (Spearman's rank: $P = 0.0037$, $\rho = 0.28$). The slopes of linear regressions fitted to the two sets of data are significantly different (Z-test: archaea slope = 0.0032, bacteria slope = 0.0013, $P = 0.0097$).

with TAG and TGA being discordant in both. This is not predicted by the RF hypothesis which proposes that the disconnect between TGA and TAG usage in bacteria can be parsimoniously explained by RF1:RF2 ratios without the need to evoke other forces.

We have also analyzed the stop codon usage trends in a sample of eukaryotes, finding TGA and TAG usage to behave near-identically with 3' UTR GC content in simple linear regression analysis (supplementary fig 2, Supplementary Material online) and in phylogenetically controlled analysis using PGLS (supplementary table 2, Supplementary Material online). The relevance of this in explaining the bacterial trends is, however, unclear. We see no reason why eukaryotic trends should diminish the archaeal result which supports that the hypothesis that forces external to RF environment is required to explain the TAG/TGA usage disconnect.

TAA Is the Most Enriched Stop Codon across Bacterial, Eukaryote, and Archaeal Species

Above we have presented several lines of evidence that one need not evoke the RF hypothesis to explain stop codon usage trends in bacteria, most notably the disconnect between TGA and TAG usage (which is observed at the canonical site, in the 3' UTR, across archaea, in intra-genome bacterial analysis, and in the human genome). We next move away from the TGA/TAG enigma to consider a different prediction of the RF hypothesis: that TAA is optimal in bacteria because it is decoded by both RF1 and RF2 (while TGA and TAG are decoded uniquely by one RF only). As eukaryotes and archaea possess only one class I release factor in their termination machinery, the RF hypothesis predicts TAA to be no more enriched than TGA and TAG. To test this,

we calculate stop codon enrichment compared to null expectation for all bacterial, eukaryotic, and archaeal species. The null frequencies in question are produced from 10,000 dinucleotide-controlled simulations (see methods), with enrichment or under-enrichment of stop codon frequency beyond null chance determined by (O-E)/E scores.

Consistent with the RF hypothesis, TAA enrichment exceeds that of TGA and TAG in 74.5% of bacterial species (480/644), this being significantly more than a simple null of 33.3% (one third, as one of the three stops must be the most enriched) (Binomial test; $P < 2.2 \times 10^{-16}$). We also find TAA to be the most enriched stop in 81.8% (18/22) of our eukaryote genomes (more than expected by chance, $P = 4.1 \times 10^{-6}$) and 54.7% (58/106) of our archaeal genomes ($P = 4.9 \times 10^{-6}$), hence forces outside of RF abundance are needed to explain TAA enrichment in such species. This complements evidence that TAA is also preferred in highly expressed genes in numerous taxa (Korkmaz et al. 2014; Trotta 2016; Ho and Hurst 2021).

Discussion

The RF hypothesis poses that bacterial species match their stop codon usage to their relative expression of RF1 and RF2 (Sharp and Bulmer 1988; Korkmaz et al. 2014; Wei et al. 2016). However, while RF1:RF2 ratios are indeed predictive of stop codon usage trends across bacteria, the causal arrow could be running in the opposite direction. Our data are more supportive of this direction of the causal arrow: between-species trends in stop codon usage are consistent in 3' UTR trinucleotides that do not function in translation termination, between-species trends in stop codon usage are consistent in within-genome analysis (controlling for RF abundance) in bacteria and in humans, between-species trends in stop codon usage are consistent in archaea which possess only one RF. Furthermore, that TAA is the most enriched stop codon (against dinucleotide-controlled null) across all three domains of life suggests that this phenomenon also cannot be explained by the RF hypothesis. We do not wish to suggest that RF abundance need not play a role in shaping stop codon usage in bacteria. However, it is apparent that other significant forces are needed to explain stop codon usage trends, not least because the same trends are seen in contexts where RF1:RF2 ratios cannot be, or are unlikely to be, causative.

A priori, the notion that RF usage adapts to stop codon usage is possibly more parsimonious from a viewpoint of evolutionary accessibility. If we imagine that RF1:RF2 ratios have, for some reason, shifted, it is not easy to see how TGA ↔ TAG exchanges might evolve in response to such shifts. To move between these two stop codons, we need a minimum of two mutational events. As the failure to terminate is deleterious (Li and Zhang 2019), we may also assume the intermediate to most often be the third stop codon, TAA, it being one mutation from either TGA or TAG. That two mutations are needed renders such adaptation to the RF pool difficult. That one needs to route via the optimal stop codon renders the process requiring even more special explanation. Imagine, for instance, that selection could favor TGA → TAA

because TGA's release factor is limiting. If so, then TAA → TAG is likely also to be deleterious as the sum of RF1 and RF2 must be greater than the sum of either (and TAA is more generally optimal). By contrast, if one evokes the notion that other forces act on stop codon usage (as our data strongly supports) then adapting the RF1:RF2 pool to better fit such usage is a trivial question of adjusting expression levels of one of either gene. If TGA does not have enough of its release factor then shifting RF1:RF2 levels is a much more evolutionarily accessible solution than two mutational events, one of which is deleterious. On a priori grounds, then adaptation of RF1:RF2 to stop usage seems the more likely mechanism.

We could imagine several experimental tests of this. First, there are strains available with all TAGs replaced with TAA stop codon (Lajoie et al. 2013). Given this, if the RF1:RF2 ratio evolves in response to stop codon usage, as we suggest is more likely, then over several generations we would expect RF1 levels to decrease. Indeed, an RF1 deletion may be favored, much as RF2 is absent from molluscs that no longer employ TGA as a termination codon (Grosjean et al. 2014). Conversely, we could alter RF1:RF2 ratios by modifying promoters or post-transcriptional regulation. Imagine for example that we could increase RF1 levels leaving RF2 levels static. In principle, TAG should now be favored over TGA under the RF1:RF2 ratio hypothesis. However, we predict that there would be little increase in TAG as one required mutation, TAA → TAG would be just as deleterious as before. The ratio could adjust nonetheless because of an increased substitution rate from TGA to TAA. However, selection always favors TGA → TAA as TAA is optimal. The change in this rate is thus also expected to be marginal. One could argue that deletion of either RF1 or RF2 must cause selection to increase TGA or TAG ratios as one of the two may well be no longer functional as a stop codon. However, this is not an informative experiment as the effect is probably so catastrophic that no organism could tolerate it (or be competitive) unless they already had abolished usage of one of the two G containing stop codons. Thus, the weak selection context involving small changes to RF1:RF2 ratios is by far the more biologically plausible condition.

We note that the hypothesis that RF1:RF2 is adapted to stop codon usage, rather than vice versa as proposed by the RF hypothesis, is not necessarily in disagreement with the observations of Korkmaz et al. (2014) or Wei et al. (2016). The observed correlations between RF1:RF2 and TAG:TGA usage can be explained by either direction of causality or, indeed, by no causality at all and the existence of other covariates. How to explain the disconnect between TAG and TGA usage at high genomic GC content is more difficult. Recalling the model proposed by Wei et al. (2016), the RF hypothesis is not needed to support mutation bias favoring TAA over TGA/TAG at low GC nor selection favoring TAA over TGA/TAG at mid-GC ranges. At high GC contents, however, unknown forces external to RF1:RF2 would likely be needed to explain the large differences observed between TGA and TAG usage.

What then could explain this? The evidence we report above gives several clues. The disconnect between TGA and

TAG is 1) common to bacteria and archaea (and possibly some eukaryotes as evidenced by intra-human trends) and 2) unrelated to translation termination given that the between-species stop codon usage trends against GC content are consistent even when we consider untranslated sequences. There are a few candidate hypotheses to explain the enigma of TAG usage. Under-usage of TAG might come as a result of complex mutation biases that disfavor the trinucleotide (or the internal dinucleotides) in all sequences. Alternatively, there could exist purifying selection or other complex fixation biases (GC-biased gene conversion, for example) against such trinucleotides and dinucleotides. The calculation of TGA and TAG mutational equilibrium frequencies across a wide range of bacteria would go some way to resolving possible discriminating mutation biases. Comparisons of these mutational equilibria to TGA and TAG fixed frequencies would similarly give an indication of whether discriminating fixation biases are at play. While we do not have evidence regarding the viability of complex (k-mer contingent) selection or complex biased gene conversion, evidence does support simple (mononucleotide-level) action of both. Possible evidence for selection favoring GC comes from the finding that in nitrogen-fixing bacteria, GC content tends to be higher than in related non-nitrogen fixing bacteria (McEwan et al. 1998), this possibly relating to GC being more nitrogen costly than A and T. More generally, unrelated organisms from the same ecology have similar GC contents (Foerster et al. 2005). Evidence for biased gene conversion comes from observations of higher GC in recombining genomes (Lassalle et al. 2015) and an association between the presence of the nonhomologous end-joining DNA double-strand break repair pathway and GC content (Weissman et al. 2019).

Aside from explaining TGA and TAG usage differences at high GC content, there is a second possible problem with our interpretation: how to explain the adaptation of sense codon usage bias to tRNA pools. There are clear parallels between the relationships of stop codons with release factors and sense codons with tRNAs, so why should stop codons and release factors coevolve differently with the causal arrow being predominantly in the direction of RF1:RF2 adapting to stop usage rather than vice versa? There are several differences that we think mean that selection on stop codon usage and on sense codon usage might be different.

First, as we discussed above, the problem with stop codon usage is how differential selection for TGA and TAG owing to RF1:RF2 ratios could manifest as selection on mutations altering stop codon usage (two mutations via an optimal intermediary). By contrast, synonymous codons tend to be only one mutational event away from each other. Thus, in the case where codon and tRNAs are out of supply-demand equilibrium (Qian et al. 2012), it is easier to see how a point mutation (unpreferred->optimal codon) can be a viable evolutionarily accessible route.

Second, each gene has only one stop codon but multiple copies of codons for many amino acids. It is hence easier to see why a tRNA might be translationally rate limiting. Consequently, one can envisage cases when selection favors

a tRNA duplication as it provides more of a translational resource for a fast-growing organism (Higgs and Ran 2008). When this happens, supply of that tRNA is likely to exceed demand and so selection could be on codons, especially in highly expressed genes, to use that over-dosed tRNA. By contrast, it is less clear whether, with one-stop codon per gene, selection would prefer more release factor in absolute terms. Indeed it is notable that there appears to be coadaptation between tRNA gene copy numbers and amino acid compositions in all three domains of life (McFarlane and Whitehall 2009; Du et al. 2017), while to our knowledge, we do not see similar duplications of *prfA* (encoding RF1) or *prfB* (encoding RF2) across bacteria.

Related to this is the problem of the rate at which RF1 and RF2 can be re-used after being employed in a termination function. In theory, there need be no lag period, while tRNAs must first be amino-acylated. In sum, then, it is not clear that RF1:RF2 levels need to be limiting in the same way tRNA pools might be limiting, meaning selection for RF1 or RF2 duplication that places the system out of supply-demand equilibrium is unlikely to be commonplace. This being said, in vitro experiments suggest different RF concentrations might improve termination, albeit with trade-offs, in certain conditions (Abdalaal et al. 2020) and can affect TR rates (though not translational efficiency per se) (Le Goff et al. 1997).

A final possibility is that sense codon usage and stop codon usage may not be so different and it is codon/stop usage bias that comes first in both cases. However, models presuming that codons and tRNA abundances coadapt, with selection for more tRNA when translation is limiting, provide a good account of data (Shields 1990; Higgs and Ran 2008; Ran and Higgs 2010). This is similarly a wealth of evidence for correlations between tRNA abundance and synonymous codon usage (Varenne et al. 1984; Sorensen et al. 1989; Gingold and Pilpel 2011). We caution, however, that there is no experimental evidence (that we know of) that tRNA abundance change precedes codon usage adaptation. One could imagine a shift in equilibrium introduced by a shift in codon usage, resulting from a myriad of factors, resulting in selection for altered tRNA abundances.

We note also that theoretical parallels cannot be drawn between stop codons and start codons. While ATG is the near-exclusive start codon in eukaryotes and preferred over other NTG start codons in bacteria (Belinky et al. 2017), there does not exist a simple RF1:RF2-like analog and hence there is no RF-like hypothesis to test with regards start codon usage. Indeed, in some species, GTG start codons appear to be poorly transcribed but more readily translated than ATG (Panicker et al. 2015). This suggests a complexity beyond that seen in the selection of stop codon usage.

Materials and Methods

General Methods

All data manipulation was performed using bespoke Python 3.6 scripts. Statistical analyses and data visualizations were performed using R 3.3.3. All scripts required for replication

of the described analyses can be found at <https://github.com/ath32/RF>.

Genome Downloads and Filtering of CDS and 3' UTR Sequences

A total of 3,727 bacterial genomes were downloaded from the EMBL database (<http://www.ebi.ac.uk/genomes/bacteria.html>, accessed August 1, 2018), 380 archaeal genomes were downloaded from advanced search of the NCBI assemblies (<https://www.ncbi.nlm.nih.gov/assembly/advanced/>, accessed July 2, 2021), and 21 eukaryotic genomes were downloaded from Ensembl (<https://www.ensembl.org/index.html?redirect=no>, accessed July 2, 2021) or Ensembl Protists (<https://protists.ensembl.org/index.html>, accessed July 2, 2021). For bacterial and archaeal genomes, we filter to retain just one genome per genus to reduce possible bias due to phylogenetic nonindependence. We too filter to retain only genomes over 500,000 base pairs in length to exclude any extremely small (e.g., plasmid) or incomplete genomes. Only genomes decoded by translation table 11 (possessing all three stop codons) were considered for analysis. This leaves a final sample of 644 bacterial, 106 archaeal, and 21 eukaryotic species.

For extraction of coding and 3' UTR sequences, we obtain the relevant data from the appropriate accompanying GFF files from the same repositories. Coding sequences for all genomes were filtered to retain only those starting with ATG and terminating in TAA, TGA, or TAG. 3' UTR sequences were obtained in several different ways. Due to a lack of appropriate annotation, for bacteria and archaea we filter genes to retain only those with >30 nts of 3' intergenic space and assume the 30 nucleotides downstream of the stop codon to be 3' UTR sequences. For most eukaryotes, we extract coding sequences and exonic sequences and define exonic sequence downstream of the stop codon to be 3' UTR sequence. This method cannot be used for single-celled eukaryotes downloaded from the Ensembl Protist repository, hence for these, we filter genes to retain only those with >100 nts of 3' intergenic space and assume the 100 nts downstream of the stop codon to be 3' UTR.

Establishing Between-Genome Trends in Stop Codon Usage

Stop codon usage frequencies were calculated at the canonical termination site of all coding sequences for each genome in the bacteria, eukaryote, and archaea data sets. This was repeated at 3' UTR null sites for comparisons to genomic regions where stop codons do not function in translation termination. Linear models were fitted to stop codon usage frequencies and mean 3' UTR GC content to determine between-species trends. Trends were tested for correlation using Spearman's rank tests.

Intraspecies Logistic Regression Analysis of Bacteria Species

Coding sequences for our bacterial data set were downloaded from Ensembl bacteria (release 51) as described above. For each species, we calculate the 3' UTR GC content, coding sequence GC3 content, and identify the stop codon used for

each gene. We capture this information in CSV files with the stop codon information captured as presence (scored 1) or absence (scored 0) of TAA, TGA, and TAG. The extent to which stop codon usage, as binary variables, may be predicted by GC content in each genome was investigated using logistic regression facilitated by the "glm" function and "family = binomial" parameter in R.

Intra-Genome Analysis of *H. sapiens*

3' UTR GC content was calculated for every gene in the *H. sapiens* genome. Genes were subsequently split evenly into 10 bins by their 3' UTR GC content. In each bin, stop codon usage frequencies at the canonical stop site were calculated. A linear model was then fit to these stop codon usage frequencies and the mean intronic GC contents of the bins to assess the intra-genomic trend. This was repeated using coding sequence GC3 content instead of 3' UTR GC content. Tests for correlation were facilitated by Spearman's rank tests.

PGLS Analysis of Eukaryotes

Phylogenetically controlled regression analyses were completed to test for correlation between 3' UTR GC content (or GC3 content) and stop codon usage using the PGLS function of the caper package in R (<https://CRAN.R-project.org/package=caper>). Pagel's lambda was predicted by maximum likelihood. The phylogenetic trees required for this analysis were generated using TimeTree (Kumar et al. 2017) and are available at <https://github.com/ath32/RF> in nexus format. Note that this analysis was only used for eukaryotic species where we can be confident in the phylogenetic relationships between species.

Dinucleotide-Controlled Simulations

Stop codon usage frequencies were compared to expected frequencies generated from 10,000 dinucleotide-controlled simulations. Simulations used Markov models to preserve reading-frame context at dinucleotide resolution as outlined by Ho and Hurst (2019). The first nucleotide of each simulated sequence was selected according to nucleotide frequencies in the coding sequences. Subsequent nucleotides were selected according to dinucleotide frequencies observed in coding sequences until one codon, or three nucleotides, in length. Deviations of the real frequencies compared to null expected frequencies were calculated as:

$$\text{Deviation} = \frac{\text{Observed} - \text{Expected}}{\text{Expected}}.$$

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the European Research Council (grant EvoGenMed ERC-2014-ADG 669207 to L.D.H.).

Data Availability

The data underlying this article are available in the article and in its online [supplementary material](#).

References

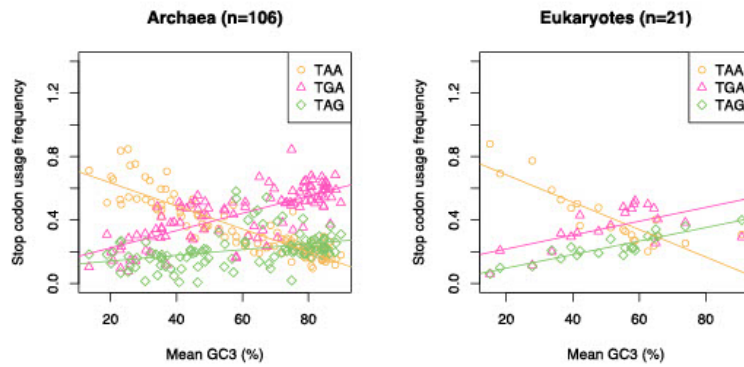
- Abdalaal H, Pundir S, Ge X, Sanyal S, Nasvall J. 2020. Collateral toxicity limits the evolution of bacterial release factor 2 toward total omnipotence. *Mol Biol Evol.* 37(10):2918–2930.
- Adachi M, Cavalcanti AR. 2009. Tandem stop codons in ciliates that reassign stop codons. *J Mol Evol.* 68(4):424–431.
- Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. *Curr Opin Genet Dev.* 8(6):688–693.
- Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci Rep.* 8(1):9260.
- Belinky F, Rogozin IB, Koonin EV. 2017. Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Sci Rep.* 7(1):12422.
- Bossi L, Roth JR. 1980. The influence of codon context on genetic-code translation. *Nature* 286(5769):123–127.
- Burroughs AM, Aravind L. 2019. The origin and evolution of release factors: implications for translation termination, ribosome rescue and quality control pathways. *Int J Mol Sci.* 20(8):1981.
- Chithambaram S, Prabhakaran R, Xia X. 2014a. Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. *Mol Biol Evol.* 31(6):1606–1617.
- Chithambaram S, Prabhakaran R, Xia X. 2014b. The effect of mutation and selection on codon adaptation in *Escherichia coli* bacteriophage. *Genetics* 197(1):301–315.
- Craigen WJ, Cook RG, Tate WP, Caskey CT. 1985. Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. *Proc Natl Acad Sci USA.* 82(11):3616–3620.
- Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 46(4):1927–1944.
- Daubin V, Perriere G. 2003. G+C structuring along the genome: a common feature in prokaryotes. *Mol Biol Evol.* 20(4):471–483.
- Dorly BC, Edgar CD, Adamski FM, Tate WP. 1990. Frameshift autoregulation in the gene for *Escherichia coli* release factor 2: partly functional mutants result in frameshift enhancement. *Nucleic Acids Res.* 18(22):6517–6522.
- Du MZ, Wei W, Qin L, Liu S, Zhang AY, Zhang Y, Zhou H, Guo FB. 2017. Co-adaptation of tRNA gene copy number and amino acid usage influences translation rates in three life domains. *DNA Res.* 24(6):623–633.
- Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10(1):285–311.
- Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet.* 2(7):549–555.
- Foerster KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6(12):1208–1213.
- Frolova L, Le Goff X, Rasmussen HH, Cheperegin S, Drugeon G, Kress M, Amman I, Haenni AL, Celis JE, Philippe M. 1994. A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. *Nature* 372(6507):701–703.
- Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159(2):907–911.
- Geller AI, Rich A. 1980. A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature* 283(5742):41–46.
- Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol.* 7(1):481.
- Grosjean H, Breton M, Sirand-Pugnet P, Tardy F, Thiaucourt F, Citi C, Barre A, Yoshizawa S, Fourmy D, de Crécy-Lagard V, et al. 2014. Predicting the minimal translation apparatus: lessons from the reductive evolution of molluscs. *PLoS Genet.* 10(5):e1004363.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9):e1001115.
- Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol.* 25(11):2279–2291.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9):e1001107.
- Ho AT, Hurst LD. 2021. Effective population size predicts local rates but not local mitigation of read-through errors in eukaryotic genes. *Mol Biol Evol.* 38(1):244–262.
- Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. *PLoS Genet.* 15(9):e1008386.
- Ikemura T. 1992. Correlation between codon usage and tRNA content in microorganisms. In: Hatfield DL, Lee B, Pirtle PM, editors. *Transfer RNA in protein synthesis*. Boca Raton (FL): CRC Press. p. 87–111.
- Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol.* 151(3):389–409.
- Inagaki Y, Doolittle WF. 2000. Evolution of the eukaryotic translation termination system: origins of release factors. *Mol Biol Evol.* 17(6):882–889.
- Jackson RJ, Hellen CUT, Pestova TV. 2012. Termination and post-termination events in eukaryotic translation. In: Marinichev A, editor. *Advances in protein chemistry and structural biology*. Vol. 86. Fidelity and quality control in gene expression. Cambridge (MA): Academic Press. p. 45–93.
- Jorgensen F, Adamski FM, Tate WP, Kurland CG. 1993. Release factor-dependent false stops are infrequent in *Escherichia coli*. *J Mol Biol.* 230(1):41–50.
- Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* 2(4):RESEARCH0010.
- Kobayashi K, Saito K, Ishitani R, Ito K, Nureki O. 2012. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1 alpha complex. *Nucleic Acids Res.* 40(18):9319–9328.
- Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem.* 289(4):30334–30342.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol.* 34(7):1812–1819.
- Lajoie MJ, Rovner AJ, Goodman DB, Aemi H-R, Haimovich AD, Kuznetsov G, Mercer JA, Wang HH, Carr PA, Mosberg JA, et al. 2013. Genomically recoded organisms expand biological functions. *Science* 342(6156):357–360.
- Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet.* 11(2):e1004941.
- Le Goff X, Philippe M, Jean-Jean O. 1997. Overexpression of human release factor 1 alone was an antisuppressor effect in human cells. *Mol Cell Biol.* 17(6):3164–3172.
- Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15(5):e1008141.
- Liang H, Cavalcanti AR, Landweber LF. 2005. Conservation of tandem stop codons in yeasts. *Genome Biol.* 6(4):R31.
- Long H, Sung W, Kucukylidirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.* 2(2):237–240.
- Major LL, Edgar TD, Yee Yip P, Isaksson LA, Tate WP. 2002. Tandem termination signals: myth or reality? *FEBS Lett.* 514(1):84–89.
- McEwan CEA, Gatherer D, McEwan NR. 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* 128(2):173–178.

- McFarlane RJ, Whitehall SK. 2009. tRNA genes in eukaryotic genome organization and reorganization. *Cell Cycle* 8(19):3102–3106.
- Meng SY, Hui JO, Haniu M, Tsai LB. 1995. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem Biophys Res Commun*. 211(1):40–48.
- Panicker S, Browning GF, Markham PF. 2015. The effect of an alternate start codon on heterologous expression of a PhoA fusion protein in *Mycoplasma gallisepticum*. *PLoS One* 10(5):e0127911.
- Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev*. 53(3):273–298.
- Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*. 12(1):32–42.
- Povolotskaya IS, Kondrashov FA, Lelida A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biol Direct*. 7(1):30.
- Prabhakaran R, Chithambaram S, Xia X. 2014. Aeromonas phages encode tRNAs for their overused codons. *Int J Comput Biol Drug Des*. 7(2–3):168–182.
- Qian WF, Yang JR, Pearson NM, Maclean C, Zhang JZ. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet*. 8(3):e1002603.
- Ran W, Higgs PG. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol Biol Evol*. 27(9):2129–2140.
- Rodnina MV. 2018. Translation in prokaryotes. *Cold Spring Harb Perspect Biol*. 10(9):a032664.
- Sanchez JC, Padron G, Santana H, Herrera L. 1998. Elimination of an HufN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J Biotechnol*. 63(3):179–186.
- Sharp PM, Bulmer M. 1988. Selective differences among translation termination codons. *Gene* 63(1):141–145.
- Shields DC. 1990. Switches in species-specific codon preferences: the influence of mutation biases. *J Mol Evol*. 31(2):71–80.
- Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet*. 14(3):e1007254.
- Sorensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol*. 207(2):365–377.
- Strigini P, Brickman E. 1973. Analysis of specific misreading in *Escherichia coli*. *J Mol Biol*. 75(4):659–672.
- Tate WP, Marsell JB, Mannering SA, Irvine JH, Major LL, Wilson DN. 1999. UGA: a dual signal for 'stop' and for recoding in protein synthesis. *Biochemistry (Mosc)*. 64(12):1342–1353.
- Trotta E. 2016. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics* 17(17):366.
- van Wieringh A, Ragonnet-Cronin M, Prandkeviciene E, Pavon-Eternod M, Kleiman L, Xia X. 2011. HIV-1 modulates the tRNA pool to improve translation efficiency. *Mol Biol Evol*. 28(6):1827–1834.
- Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol*. 180(3):549–576.
- Vestergaard B, Van LB, Andersen GR, Nyborg J, Buckingham RH, Kjeldgaard M. 2001. Bacterial polypeptide release factor RF2 is structurally distinct from eukaryotic eRF1. *Mol Cell*. 8(6):1375–1382.
- Vogel C, Abreu RD, Ko DJ, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte BM, Penalva LO. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol*. 6(1):400.
- Wei Y, Xia X. 2017. The role of +4U as an extended translation termination signal in bacteria. *Genetics* 205(2):539–549.
- Wei YL, Wang J, Xia XH. 2016. Coevolution between stop codon usage and release factors in bacterial species. *Mol Biol Evol*. 33(9):2357–2367.
- Weissman JL, Fagan WF, Johnson PLF. 2019. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS Genet*. 15(11):e1008493.
- Xia XH. 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics*. 149(1):37–44.

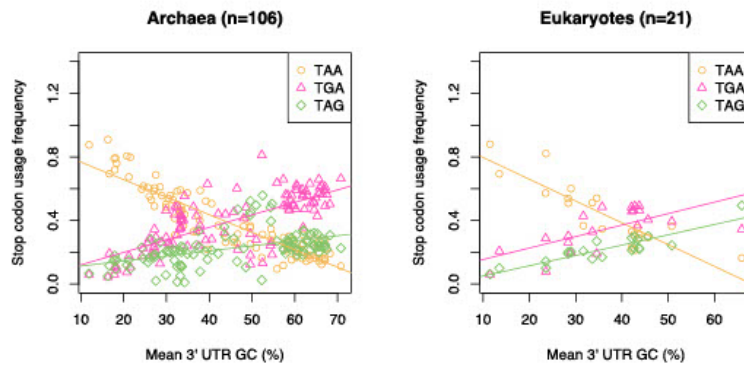
Supplementary information for: Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage

Alexander T. Ho and Laurence D. Hurst
Molecular Biology & Evolution, msab326.

Note: Some of the supplementary figures and tables are extremely small when viewed in this document. To view these best, please refer to the online source: <https://doi.org/10.1093/molbev/msab326>.



Supplementary figure 1. Stop codon usage as a function of GC3 content across 106 archaea and 21 eukaryotes. TAA usage is negatively correlated with GC3 content across archaea (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = -0.89$) and eukaryotes (Spearman's rank: $p = 0.00059$, $\rho = -0.86$). TGA usage is positively correlated with GC3 content across archaea (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = 0.76$) and eukaryotes (Spearman's rank: $p = 0.012$, $\rho = 0.56$). TAG is positively correlated with GC3 content across archaea (Spearman's rank: $p = 1.1 \times 10^{-7}$, $\rho = 0.49$) and eukaryotes (Spearman's rank: $p = 9.5 \times 10^{-7}$, $\rho = 0.88$).



Supplementary figure 2. Stop codon usage as a function of 3' UTR GC content across 106 archaea and 21 eukaryotes. TAA usage is negatively correlated with 3' UTR GC content in archaea (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = -0.90$) and eukaryotes (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = -0.85$). TGA usage is positively correlated with 3' UTR GC content (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = 0.76$) and eukaryotes (Spearman's rank: $p = 0.0013$, $\rho = 0.66$). TAG is positively correlated with 3' UTR GC content (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = 0.57$) and eukaryotes (Spearman's rank: $p < 2.2 \times 10^{-16}$, $\rho = 0.85$).

Supplementary table 1. Results from logistic regression models predicting intra-genome stop codon usage as a function of GC3 content in 18 bacterial species with unusually high GC3 variance.

Species	TAA		TGA		TAG		TGA slope > TAG slope?	1st Quar. GC3	Median GC3	3rd Quar. GC3
	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value				
<i>Bacillus halodurans</i>	-0.02841	4.74E-07	0.011942	0.0724	0.031814	8.74E-06	-	38.79	42.39	45.61
<i>Bacillus subtilis</i>	-0.005203	0.141961	0.004036	0.31	0.003759	0.441	+	37.84	43.6	48.36
<i>Brucella melitensis</i>	-0.001085	0.8379	0.006046	0.213	-0.009717	0.1398	+	62.9	67.22	70.53
<i>Chlamydia trachomatis</i>	-0.029705	0.000139	0.040822	1.13E-05	0.003748	0.654823	+	30.312	34.032	37.326
<i>Deinococcus radiodurans</i>	-0.0007681	0.8812	0.005232	0.259	-0.010696	0.1175	+	81.33	85.25	87.76
<i>Escherichia coli</i>	-0.0007345	0.80073	0.00871	0.00465	-0.023183	4.70E-06	+	48.5	54.93	59.81
<i>Listeria monocytogenes</i>	-0.047371	2.16E-09	0.038682	3.03E-05	0.03836	0.00036	+	25.35	28.42	31.84
<i>Mycobacterium tuberculosis</i>	-0.02757	0.000165	-0.015345	0.00553	0.037028	3.56E-09	-	76	79.41	82.54
<i>Mycoplasma genitalium</i>	-0.02574	0.112	-3.42E-15	1	0.02574	0.112	-	18.775	22.528	27.414
<i>Mycoplasma pneumoniae</i>	0.007537	0.479	-9.63E-16	1	-0.007537	0.479	+	34.91	41.12	45.81
<i>Pasteurella multocida</i>	0.0003631	0.965	0.003999	0.725	-0.003915	0.706	+	31.25	34.78	37.96
<i>Pseudomonas aeruginosa</i>	-0.052776	<2e-16	0.049092	<2e-16	-0.024599	<2e-16	+	78.74	86.44	89.8
<i>Ralstonia solanacearum</i>	-0.02692	1.73E-07	0.029694	1.71E-11	-0.018197	0.00109	+	82.77	87.6	90.6
<i>Salmonella enterica</i>	0.003003	0.355	0.002948	0.396	-0.014898	0.00401	+	53	59.34	63.56
<i>Sinorhizobium meliloti</i>	-0.020731	1.49E-06	0.019668	1.52E-08	-0.009917	0.0189	+	72.5	77.95	81.45
<i>Staphylococcus aureus</i>	-0.055096	1.89E-09	0.0497	3.57E-05	0.04103	0.000152	+	19.831	22.549	25.424
<i>Thermotoga maritima</i>	-0.035571	5.52E-05	0.033554	2.24E-05	-0.01379	0.2672	+	48.78	52.54	56.21
<i>Vibrio cholerae</i>	-0.010944	0.0283	0.003428	0.563	0.014622	0.0262	-	44.23	48.67	52.3

Supplementary table 2. Results from Phylogenetic Generalized Least Squares (PGLS) analysis testing for correlation between stop codon usage and two proxies of GC pressure. Both genomic 3' UTR GC content and genomic GC3 content are significant predictors of TAA, TGA and TAG usage when controlling for phylogeny. The direction of the correlation of stop codon usage with 3' UTR GC content is indicated by the sign of the estimate. Lambda was computed by maximum likelihood and explains the extent to which the traits are evolving as expected given tree topology alone (0 = each point is phylogenetically independent, 1 = traits are evolving as expected given the phylogenetic relationships).

Stop codon usage	Predictor variable	Estimate	Adjusted r²	P-value	Lambda (ML)
TAA	Mean GC3	-0.0091	0.80	1.6 x 10 ⁻⁸	0.83
	Mean 3' UTR GC	-0.014	0.84	5.1 x 10 ⁻⁶	0.986
TGA	Mean GC3	0.0041	0.37	0.012	0.98
	Mean 3' UTR GC	0.0076	0.56	0.0024	1.000
TAG	Mean GC3	0.0042	0.80	1.1 x 10 ⁻⁵	0.80
	Mean 3' UTR GC	0.0058	0.66	2.7 x 10 ⁻⁴	0.000

Chapter 5

Sequence conservation need not imply purifying selection: evidence from mammalian stop codon usage

Alexander T. Ho and Laurence D. Hurst

PLoS Biology (under review)

This chapter contains a draft manuscript currently under review at PLoS Biology as an invited re-submission.

Pre-amble

How can we explain the high abundance and conservation of TGA in mammals?

Related to the problem of explaining non-optimal stop codon usage in bacteria is the question of how to explain the unusual high abundance and conservation of TGA stop codons observed in mammalian genomes. Three lines of evidence immediately refute selective hypotheses related to minimising translational read-through: highly expressed human genes prefer TAA stop codons, TAA enrichment positively correlates with effective population size (N_e) in eukaryotes (chapter 3), and TAA is the most enriched stop codon in eukaryotes, bacteria, and archaea (chapter 4). Nearly neutral theory might partially explain high TGA abundance in these species, mammals having low N_e and hence inefficient selection to increase TAA usage. It cannot, however, explain why TGA might be more conserved than TAA and TAG stop codons. In this chapter I assess three broad hypotheses that could explain this phenomenon in mammalian taxa: mutation bias, a unique mammalian selective preference for TGA, and GC-biased gene conversion (gBGC).

To differentiate between the three hypotheses, I analyse both stop codon usage and flux (the rate at which one stop codon changes to another, per incidence of the ancestral stop codon). Mutation bias can be assessed by inferring a mutational matrix from *de novo* mutations and determining the extent of AT->GC or GC->AT bias. Selection for TGA may be assayed by regression analysis that predicts stop codon usage as a function of gene expression, assuming highly expressed genes to be under the greatest selection pressure for error control. I investigate the gBGC hypothesis by assessing the relationship between stop codon usage and recombination rate, gBGC being tightly linked to the mismatch repair process during homologous recombination.

If gBGC were to provide a parsimonious framework for mammalian TGA conservation, this would represent a unique example where gBGC unambiguously acts in opposition of selection (promoting TGA, not TAA, stop codon usage).

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:			
Sequence conservation need not imply purifying selection: evidence from mammalian stop codon usage			
Publication status (tick one)			
Draft manuscript	<input type="checkbox"/>	Submitted	<input type="checkbox"/>
In review	<input checked="" type="checkbox"/>	Accepted	<input type="checkbox"/>
Published	<input type="checkbox"/>		
Publication details (reference)	N/A		
Copyright status (tick the appropriate statement)			
I hold the copyright for this material	<input checked="" type="checkbox"/>	Copyright is retained by the publisher, but I have been given permission to replicate the material here	<input type="checkbox"/>
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	<p>The candidate contributed to / considerably contributed to / predominantly executed the...</p> <p>Formulation of ideas: 100%</p> <p>Design of methodology: 100%</p> <p>Bioinformatic analyses: 100%</p> <p>Experimental work: N/a</p> <p>Presentation of data in journal format: 100%</p>		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed		Date	03/03/2022

RESEARCH ARTICLE:

Full title:

**Sequence conservation need not imply purifying selection: evidence from
mammalian stop codon usage**

Short title:

Sequence conservation need not imply purifying selection

Alexander T. Ho^{1*} and Laurence D. Hurst¹

1. Milner Centre for Evolution, University of Bath, Bath, UK

Alexander T. Ho*, E-mail: a.t.ho@bath.ac.uk

[Laurence D. Hurst, E-mail: l.d.hurst@bath.ac.uk](mailto:l.d.hurst@bath.ac.uk)

*Author for correspondence

Key words:

biased gene conversion, genetics, molecular evolution, stop codons, genome
degeneration

Abstract

The assumption that conservation of sequence implies the action of purifying selection is central to diverse methodologies to infer functional importance. In mammals, however, GC-biased gene conversion (gBGC), a meiotic mismatch repair bias strongly favouring GC over AT, can in principle mimic the action of selection. As mutation is GC→AT biased, to demonstrate that gBGC does indeed cause false signals requires confidence 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 and elsewhere (in UTR sequences), (ii) that higher TGA usage and higher TAA→TGA substitution rates are predicted by high recombination rate and (iii) across species the difference in TAA ↔ TGA 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 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 discussion see Ponting (2017). It is explicit in, for

example, molecular evolutionary tests for purifying selection (e.g. Ka/Ks test (Nielsen and Yang 1998; Yang and Bielawski 2000; Hurst 2002; Pond and Frost 2005)), attempts to identify sites prone to disease-causing mutations (Cooper, et al. 2010; Sun and Yu 2019), and estimates of the proportion of DNA within a genome that is “functional” (Ponting 2008).

These methods assume, however, that no force other than selection can deterministically act to alter the frequency of extant alleles. Over the past two decades GC-biased gene conversion has been established as a potentially important influence on allele frequencies (Lachance and Tishkoff 2014), mimicking selection (Galtier, et al. 2001; Duret and Galtier 2009; Galtier, et al. 2009). 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 (Brown and Jiricny 1988, 1989). In humans, at non-crossover gene conversion events 67.6% of GC:AT mismatches favour the GC allele (Halldorsson, et al. 2016). 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 (Bernardi, et al. 1985), have genomes with large (> 300Mb) blocks of relatively homogeneous higher or lower GC content (isochores) (Eyre-Walker and Hurst 2001; Galtier, et al. 2001; Duret and Galtier 2009). Importantly, assuming consistency of local recombination rates over evolutionary time and a correlation between crossover rates and non-crossover rates (Williams, et al. 2015), gBGC also can explain the relatively strong correlation between GC content of these blocks and local recombination rates in mammals (Eyre-Walker 1993; Fullerton, et al. 2001; Duret and Arndt 2008; Pessia, et al. 2012) (but see also Marsolier-Kergoat and Yeramian 2009; Kiktev, et al. 2018). That the correlation is stronger with male meiotic events than female ones is taken as evidence that the trends cannot be explained by selection with reduced Hill-Robertson interference in domains of high recombination (Duret and Arndt 2008). Consistent with such models, SNP analysis reveals the predicted fixation bias for AT→GC mutations in GC rich domains, even after allowing for non-equilibrium GC content (Duret, et al. 2002; Lercher, et al. 2002).

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 et al. (Williams, et al. 2015) estimate a 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 et al. (Glemin, et al. 2015) estimate that the net effect on substitutions is on average in the nearly-neutral area. However, as recombination occurs primarily within recombination hotspots ~2% of the human genome is subject to strong gBGC in any generation (Glemin, et al. 2015). 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 (Glemin, et al. 2015). Galtier (2021) estimates that ~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 (Mancera, et al. 2008; Duret and Galtier 2009). Further analysis report a lesser bias (Liu, et al. 2019), with a further large study reporting weak bias in the opposite direction (Liu, et al. 2018). 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 (Liu, et al. 2018). To date strong conversion has been observed in only a few taxa (Liu, et al. 2018), mammals (Duret and Galtier 2009) and birds (Weber, et al. 2014; Rousselle, et al. 2019) being the two well-described exceptions, though weaker and non-regionalised gBGC is suspected in many taxa (Pessia, et al. 2012).

In terms of the population genetics influence, the action of gBGC is directly comparable to meiotic drive (alias segregation distortion) (Gutz and Leslie 1976). In this sense gBGC may be said to “drive” alleles. In turn, such drive can mimic positive selection (Nagylaki 1983). 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 (Dreszer, et al. 2007; Berglund, et al. 2009; Galtier, et al. 2009; Ratnakumar, et al. 2010; Corcoran, et al. 2017; Bolivar, et al. 2018). However, as is implicit in all such models (Harrison

and Charlesworth 2011), 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 (Dreszer, et al. 2007; Berglund, et al. 2009; Galtier, et al. 2009; Ratnakumar, et al. 2010; Corcoran, et al. 2017; Bolivar, et al. 2018). 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 (Smith and Eyre-Walker 2001; Lynch, et al. 2008; Hershberg and Petrov 2010; Long, et al. 2018; Smith, et al. 2018), 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 (Bengtsson 1985), 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 G or C (Vicario, et al. 2007). 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 (Harrison and Charlesworth 2011)). Thus, the core difficulty to establish gBGC as a cause of false signals of purifying selection and cause 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 three stop codons, TAA, TGA and TAG, are not used equally (Belinky, et al. 2018), with TAA being commonly, if not universally, selectively favoured (Belinky, et al. 2018). This is probably owing in large part to selective avoidance of translational read-through (TR). During TR, the stop codon is

missed by its cognate release factor (Rodnina, et al. 2020) due to the mis-binding of a near-cognate tRNA (Roy, et al. 2015; Beznoskova, et al. 2016), leading to the erroneous translation of the 3' UTR and the generation of potentially-deleterious protein products (Rodnina 2016). Each stop codon has a distinct intrinsic error rate such that TGA>TAG>TAA in bacteria (Roth 1970; Strigini and Brickman 1973; Ryden and Isaksson 1984; Parker 1989; Meng, et al. 1995; Sanchez, et al. 1998) and eukaryotes (Geller and Rich 1980; Parker 1989) (including humans (Cridge, et al. 2018)). 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 (Arribere, et al. 2016; Li and Zhang 2019). This is likely as the formation of C-terminal extensions cause energetic wastage (Wagner 2005) as well as problems with protein stability (Clegg, et al. 1971; Namy, et al. 2002; Pang, et al. 2002), aggregation (Vidal, et al. 1999; Vidal, et al. 2000), and localisation (Falini, et al. 2005; Hollingsworth and Gross 2013). Alternatively, in the absence of another 3' in-frame stop codon, both the read-through transcript and nascent protein are likely to be degraded when the translational machinery reaches the polyA⁺ tail (Dimitrova, et al. 2009; Klauer and van Hoof 2012). 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 (Tuller, et al. 2010) and it is robust to two mistranscription events (TAA→TGA, TAA→TAG) while the two other stop codons are resilient to just one (TGA→TAA, TAG→TAA).

It is then noteworthy that stop codon usage in mammals is different to that seen elsewhere (McCaughan, et al. 1995; Belinky, et al. 2018): TGA is more often conserved than TAA (Seoighe, et al. 2020) and, unusually, the substitution rate of TAA→TGA is higher than the reverse (Belinky, et al. 2018). Despite the fact that in humans TAA is disproportionately employed in highly expressed genes (Trotta 2016), this signal of conservation has been interpreted as evidence that purifying selection is operating to preserve TGA in mammals (Seoighe, et al. 2020). 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 (Halldorsson, et al. 2016), could thus resolve the exceptionalism of TGA conservation in mammals.

Here we evaluate this suggestion. Duret and Galtier (2009) 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 to examine 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 confidence that TAA is the optimal stop codon. We consider several tests all of which support this. Finally, we resolve 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 (as with 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 non-specific 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 is 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 isochores GC, covaries with stop codon usage at the focal termination codon. Consistent with the observations of Seoighe et al. (Seoighe, et al. 2020) and Belinky et al. (Belinky, et al. 2018), we find TGA to be the most common stop in the primate lineage (Fig 1). Not only is TGA the most common stop, it also significantly and positively covaries with intronic GC content in humans when both metrics are calculated in 10% percentile bins ($n \sim 1000$ 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 non-coding 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 non-independent; 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.

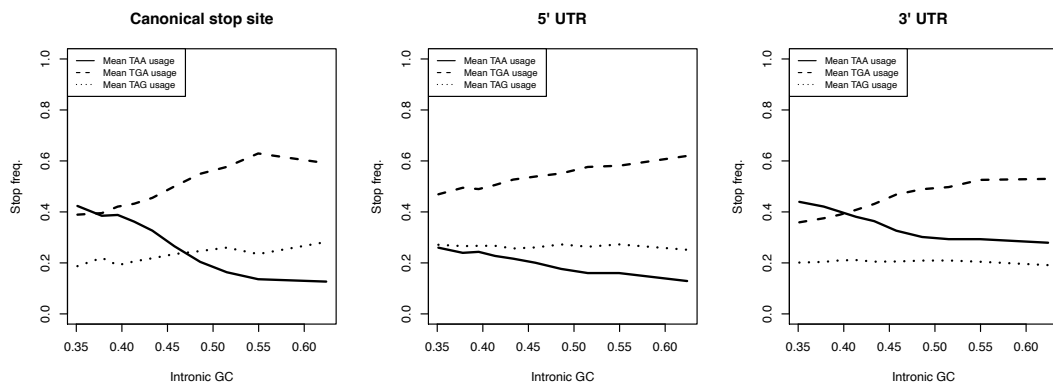


Fig 1. Stop codon frequencies (relative to the usage of all stops) at the canonical stop site, in the 5' UTR, and in the 3' UTR at ten equal sized bins of various intronic GC

contents in the genome. TAA frequency is negatively correlated with intronic GC content in all three 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 three 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; $p = 0.61$, $\rho = 0.19$, $n = 10$). Standard error bars calculated within each bin are extremely small and hence not shown.

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 (Meunier and Duret 2004; Duret and Arndt 2008), GC* here being the predicted fixation bias determined equilibrium value rather than a non-equilibrium 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 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 (Blaschke and Rappold). 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 one candidate transcript per gene annotated in this region) compared to the genome wide average (52.4%). Statistical comparison of TGA usage between these two 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 (Pessia, et al. 2012). 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 (Williams, et al. 2015). 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 non-CDS 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 (Belinky, et al. 2018), 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 two partitions (high GC, 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 formulae proposed by Long, et al. (2018) to calculate TGA content from these flux rates alone, pTGA (see methods). This provides a single metric of the relative substitution rate between the two 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 one 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 non-mammalian species trios (birds, nematodes, flies, and plants) (see <https://github.com/ath32/gBGC> for species lists). To assay the extent of pTGA deviation we calculate (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 two null groups of genes according to the net genomic TAA→TGA and TGA→TAA rates (see methods).

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 four mammalian lineages. The difference between the gene groups is greater than expected by chance in all four cases (primates: $p = 0.014$, dog: $p = 0.040$, cow: $p < 0.0001$, mouse: $p < 0.0001$). Of the non-mammalian 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. Probably due to the selfing biology of *Arabidopsis* (Marais, et al. 2004), pTGA is lower in GC-rich genes than GC-poor genes (nevertheless, $p = 1$ using the same test as the other lineages, Fig 2h).

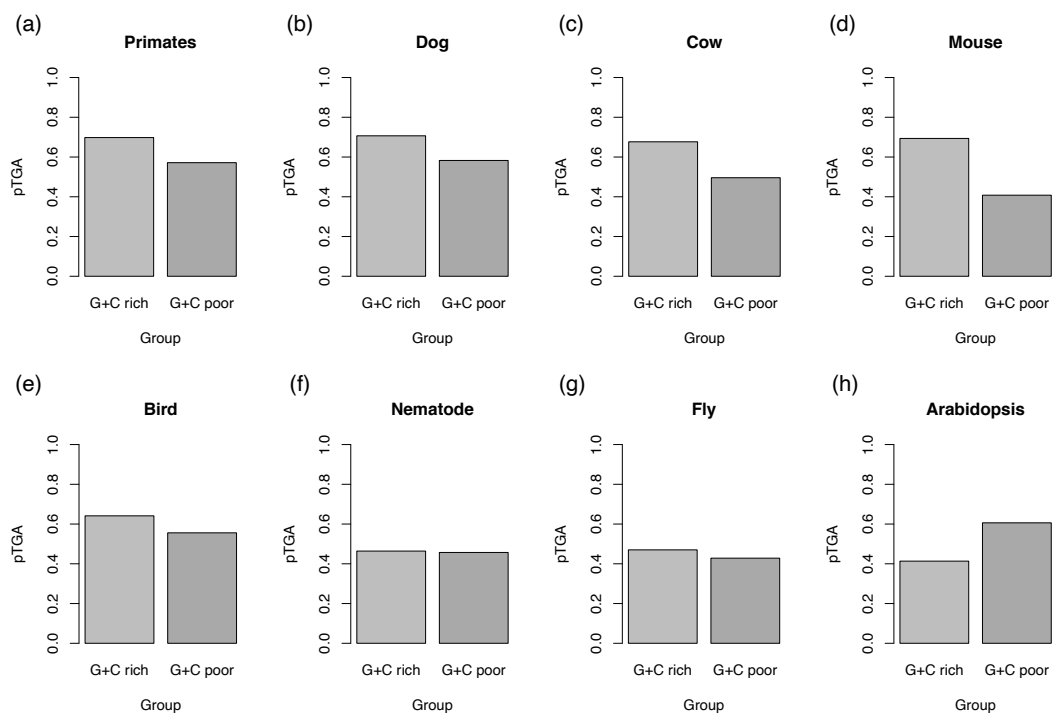


Fig 2. Predicted TGA usage (pTGA) derived from TAA→TGA and TGA→TAA flux for the top 50% of genes by GC content and bottom 50% of genes by GC content in four mammalian (a-d) and four non-mammalian (e-h) lineages. pTGA is calculated as $1/(1+(TGA→TAA/TAA→TGA))$ and hence represents the balance between the two dominant stop codon flux events. Bootstrapped 95% confidence intervals are miniscule and hence not shown.

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 two sets of genes analysed in Fig 2 and call this Δ GC. We also estimate the variance in GC3 between all genes. Consistent with the gBGC hypothesis for explaining TGA usage trends, analysing our eight lineages we find pTGA deviation is significantly correlated with both Δ 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).

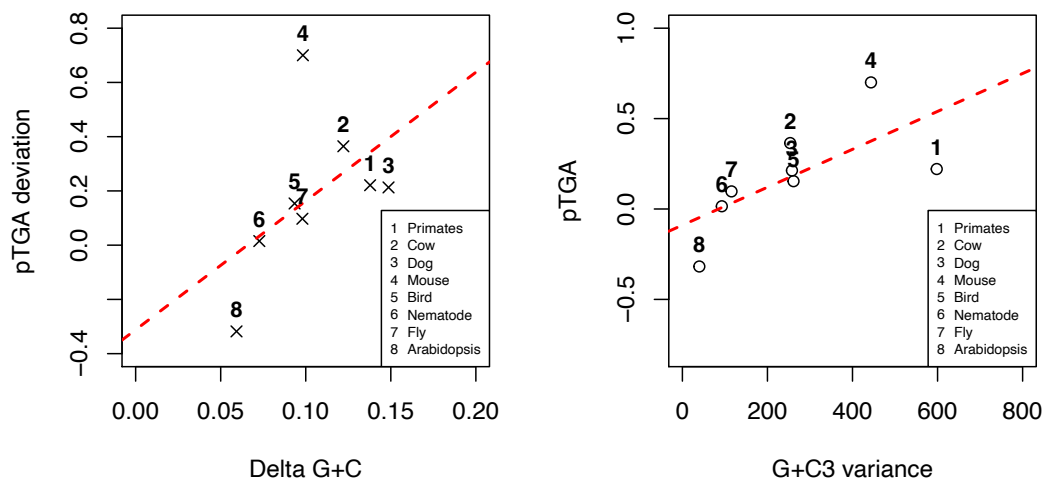


Fig 3. Predicted TGA usage (pTGA) deviation between the top 50% and bottom 50% of genes by GC content as a function of (a) the difference in GC content between the two gene bins, “delta GC”, and (b) coding sequence GC3 content variance across a sample of four mammalian and four non-mammalian lineages. pTGA is calculated as $1/(1+(TGA \rightarrow TAA/TAA \rightarrow TGA))$ and hence represents the balance between the two 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; $p = 0.028$, $\rho = 0.79$, $n = 8$). Bootstrapped error bars on the x and y axes are miniscule and not shown.

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 (Duret and Galtier 2009; Smeds, et al. 2016). 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 (Harrison and Charlesworth 2011; Robinson, et al. 2014; Liu, et al. 2018) or less regionalised effects. Arabidopsis being an almost obligate inbreeder is expected to be most affected by mutation bias and least affected by gBGC (Marais, et al. 2004).

(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).

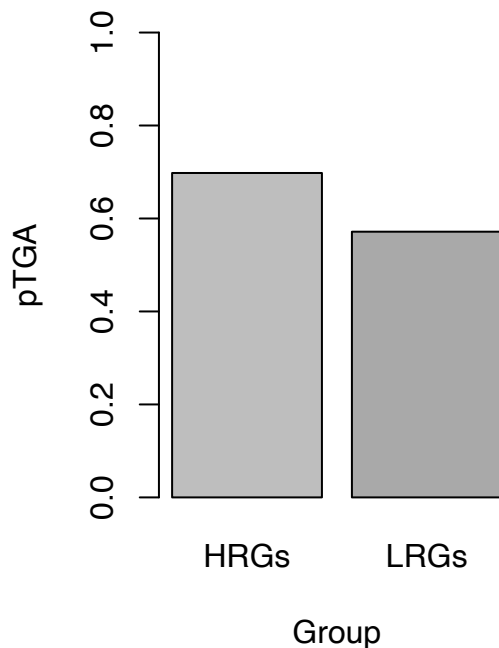


Fig 4. Predicted TGA usage (pTGA) derived from TAA→TGA and TGA→TAA flux for the top 50% of genes by recombination rate (HRGs) and bottom 50% of genes by recombination rate (LRGs) in the human genome. pTGA is calculated as $1/(1+(TGA\rightarrow TAA/TAA\rightarrow TGA))$ and hence represents the balance between the two dominant stop codon flux events. Bootstrapped confidence intervals are miniscule and hence now shown.

No evidence to support TGA optimality in eukaryotes

The evidence from non-termination sites 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 highly expressed genes (Xu and Zhang 2018; Li and Zhang 2019; Xu, et al. 2019) as selection is strongest in highly expressed genes. Over-usage of “optimal” codons in highly expressed genes is a case in point (Duret 2002; de Oliveira, et al. 2021). In the current context, the opportunity for deleterious read-through (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 expression level. Prior data appeared to contradict this, suggesting that human highly expressed genes (HEGs, the opposite being LEGs) preferentially use TAA stop codons (Trotta 2016). However, possible covariation between expression level and GC content (Lercher, et al. 2003; Kudla, et al. 2006; Mordstein, et al. 2020) could disturb the ability to make correct inference. We ask whether TAA or TGA are over employed in highly expressed genes controlling for GC content.

Second, the efficiency of both selection (Ohta 1992; Lynch 2007) and gBGC (Weber, et al. 2014; Galtier 2021) 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 (Galtier 2021), 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 (Galtier, et al. 2018)). 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 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 (Ho and Hurst 2020). However, the possibility remains that for lowly expressed genes 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 three 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 (as a proxy for gene expression, 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).

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

Consistent with prior observations of stop codon covariance with GC content (Korkmaz, et al. 2014; Trotta 2016), we find TAA usage to be negatively (indicated by the sign of the coefficient), and TGA to be positively, correlated with GC3 in all four species trios tested. By the same coefficient analysis, we find that high TAA stop codon usage is predicted by high expression level in all four mammalian lineages (Trotta 2016), 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 usage in our three mammalian lineages. In 8/8 models, the coefficients of protein abundance are consistent with TAA preference over TGA in highly expressed genes. 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 highly expressed genes 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 (Weber, et al. 2014).

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 (Smith and Eyre-Walker 2001; Lynch, et al. 2008; Hershberg and Petrov 2010; Long, et al. 2018; Smith, et al. 2018). Nonetheless the possibility remains either that some more complex k -mer bias might exist or that mutation bias varies by isochores. Indeed, nucleotide pools can vary through the cell cycle potentially altering local mutation bias (Wolfe, et al. 1989). Moreover, CpG to TpG rates are high in humans (Duncan and Miller 1980; Sved and Bird 1990; Fryxell and Moon 2005;

Roberts and Gordenin 2014) 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 pre-existing 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 (Fryxell and Moon 2005) also renders this an unlikely explanation.

We consider the relative rates of human germline *de novo* mutations derived from family trio data (Jonsson, et al. 2017). 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 (Long, et al. 2018). However, equilibrium status is disputed (Sun, et al. 2019) and the predicted equilibrium can vary with complexity of the mutational model (mono-nucleotide, di-nucleotide etc).

Consistent with previous analyses (Smith and Eyre-Walker 2001; Lynch, et al. 2008; Hershberg and Petrov 2010; Long, et al. 2018; Smith, et al. 2018), 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 et al. Smith, et al. (2018)). For each mononucleotide change, the local GC content (10kb window) was calculated. Mutations were then ordered by GC and split into 10% percentile bins of equal size (~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 et al. Smith, et al. (2018). 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; $p = 0.035$, $\rho = -0.68$, $n = 10$).

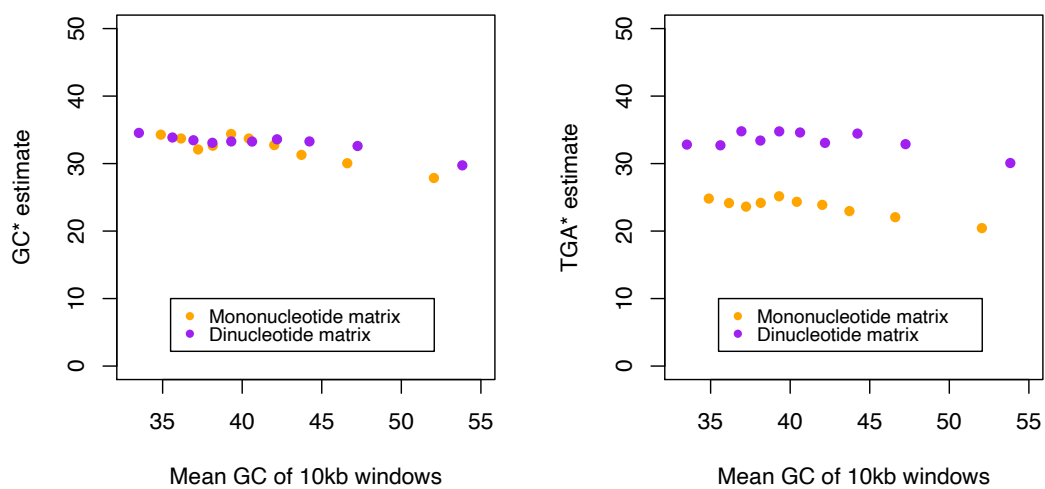


Fig 5. Predicted GC equilibrium (GC*) and relative TGA equilibrium (TGA*) frequencies across isochore GC contents derived from mononucleotide (orange) and dinucleotide (purple) mutational matrices. Standard deviations for the datapoints are minuscule and hence error bars are not shown ($\sim 0.5\%$ for mononucleotide estimates of TGA* and GC*, $\sim 0.5\%$ for dinucleotide estimates of TGA* and ~ 0.1 for dinucleotide estimates of GC*).

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 (Duncan and Miller 1980; Sved and Bird 1990; Fryxell and Moon 2005; Roberts and Gordenin 2014) that are likely to affect the mutation-drift equilibrium frequency of TGA, we expand our analysis to consider the 16x16 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 three stop codons in such sequence and how they vary by 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 10kb GC content) for all sequence that isn't CDS ($r^2 > 0.9$; Figure 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 two, going some way to explain why TAG and TGA behave differently.

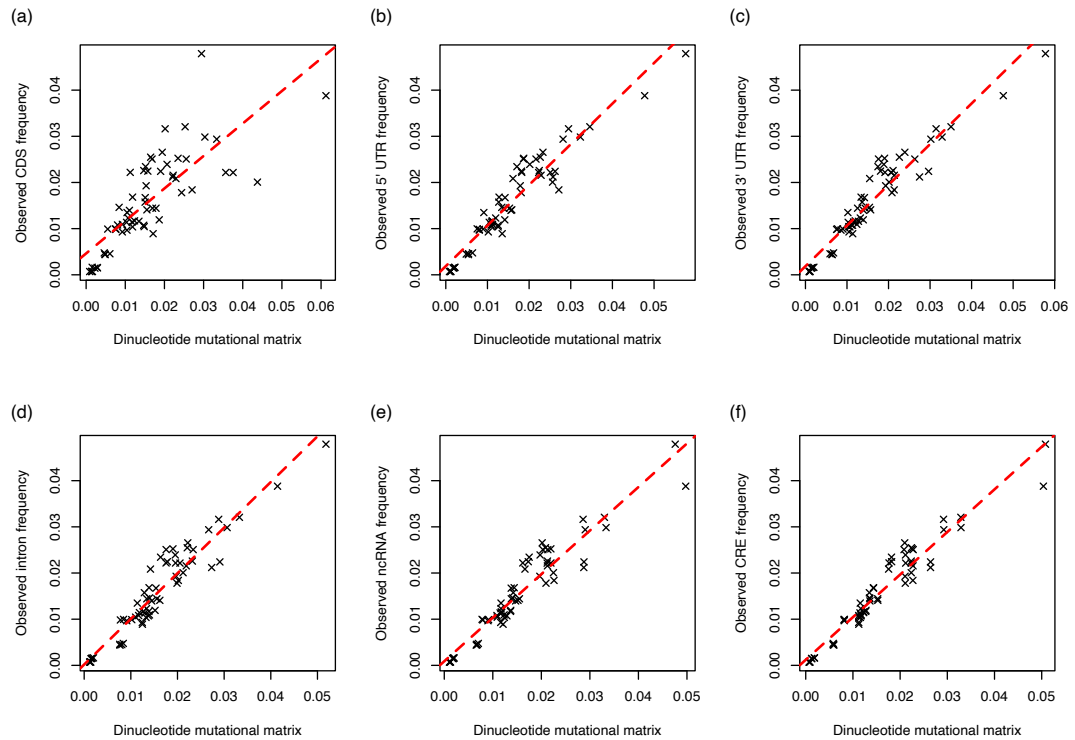


Figure 6. Observed (a) CDS, (b) 5' UTR, (c) 3' UTR, (d) intronic, (e) ncRNA, (f) cis regulatory element (CRE) trinucleotide frequencies as a function of the expected frequencies of the same trinucleotides derived from a dinucleotide mutational matrix. Expected frequencies were calculated simulated DNA 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 $r^2 = 0.93$).

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 (Figure 6). By contrast, in GC rich domains, there exists a

substitutional bias to TGA that is incompatible with mutation bias alone (Figure 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 (Warnecke and Hurst 2007), 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 (Observed-Expected)/Expected 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 10kb 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* 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 one of four classes by GC% (0%, 33%, 66%, 100%). Within the zero 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 Figure 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 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 two-fold synonymous codon pairs, all ending G:A or C:T, in genes of increasing recombination rate. Considering all two-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 can't in these examples be sure which (if either) is the selectively optimal state. The two 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.

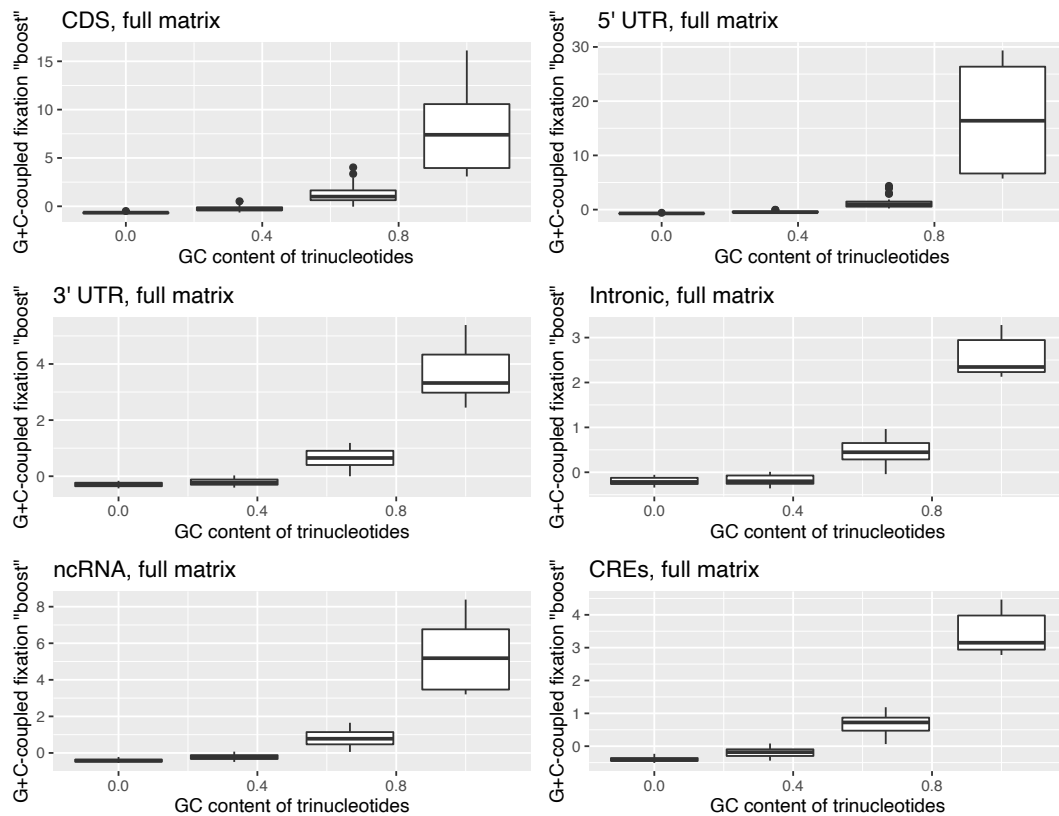


Figure 7. Deviation scores, $(O-E)/E$, describing the difference in GC-coupled fixation “boost” for the four GC classes of trinucleotides. Deviation between fixed and mutational equilibrium frequencies for each trinucleotide in the top 20% of sequences by GC content, D1, was calculated as $(\text{Observed}-\text{Expected})/\text{Expected}$, 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 $(\text{Observed}-\text{Expected})/\text{Expected}$, 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).

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 two 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 we find the ranking within each GC class to be remarkably consistent between sequence types (5' UTR, 3' UTR, 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 classes GC classes (33% and 66%), ranks are significantly 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 non-expressed CRE versus intron comparison gives exceptionally high repeatability indicating that transcription coupled repair/mutation probably does not explain these trends.

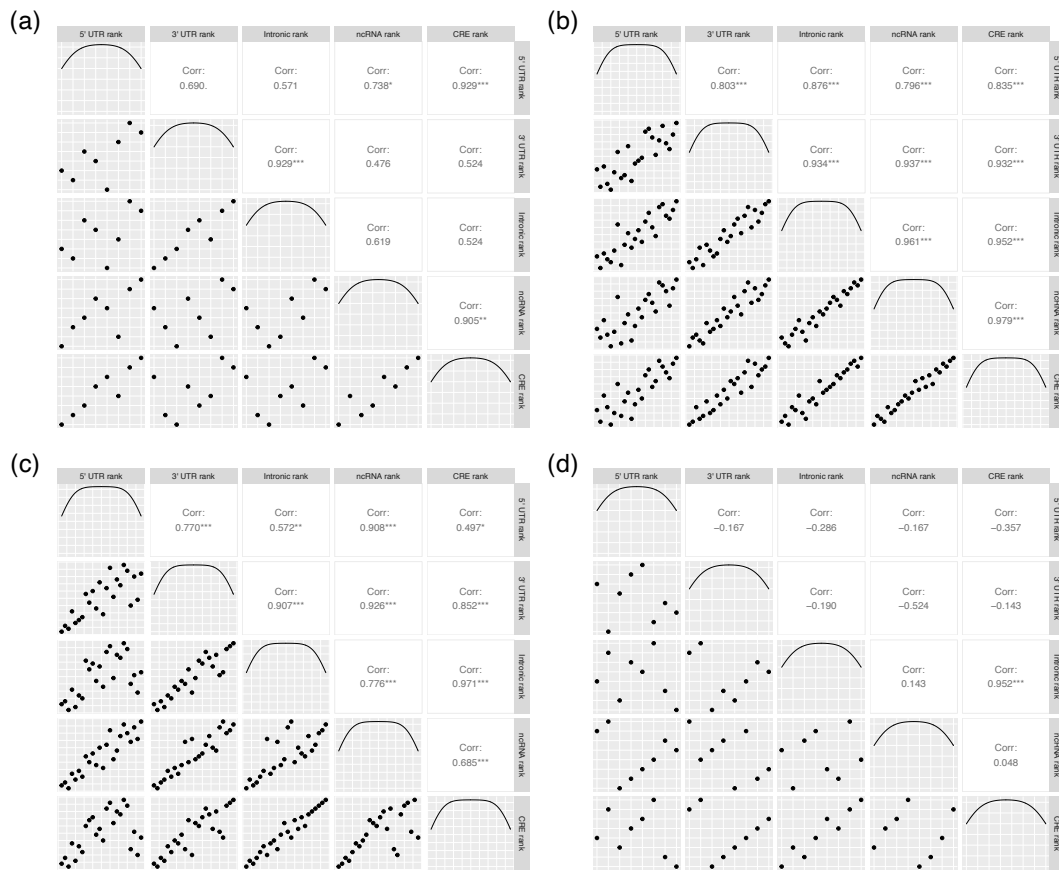


Fig 8. Correlation analysis of trinucleotide ranks (by their gBGC “boost” metric) within the four GC classes (a) 0%, (b) 33%, (c) 66%, (d) 100%. Within the 33% and 66% GC classes, ranks are significantly correlated in all comparisons ($p < 0.01$). This is not true of the 0% and 100% GC classes, correlation 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.

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 two-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 (Nagylaki 1983; Dreszer, et al. 2007; Berglund, et al. 2009; Galtier, et al. 2009; Ratnakumar, et al. 2010; Corcoran, et al. 2017; Bolivar, et al. 2018), there has been less attention paid to the problem that it might also explain sequence conservation, despite this being a logical necessity (Harrison and Charlesworth 2011). 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 non-coding 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 (2009) 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 non-optimal codon, a trend that can be partly overcome by stronger selection for optimality in highly expressed genes.

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 (Hill and Robertson 1966) 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 (Basham, et al. 1995; Vinogradov 2001; Babbitt and Schulze 2012) or on nucleosome positioning (Warnecke, et al. 2008; Babbitt and Cotter 2011; Prendergast and Semple 2011; Langley, et al. 2014). 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, that the GC biasing effect correlates with male not female recombination rates (Duret and Arndt 2008), suggests that the effects are not mediated by reduced Hill-Robertson interference (Duret and Arndt 2008).

Third, 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 (Karlin and Mrazek 1997; Mrazek and Karls 2019). 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 two that are core to TATA box, are consistently the two 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 (Hahn, et al. 2003). If then *Escherichia coli*'s N_e is of the order of 10^8 (Berg 1996), then s must be $\sim -0.09/10^8 = -9 \times 10^{-10}$. For a mutation to be under selection in humans $s \sim 1/2 N_e$ must hold. In a species with $N_e \sim 10,000$ (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, highly expressed genes over-employ 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 (Cridge, et al. 2018), we don't 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 (Dunn, et al. 2013). Indeed, a few examples of functional read-through have been described (Jungreis, et al. 2011; Schueren and Thoms 2016), though the commonality of this in mammals is unknown. Alternatively, read-through may be part of a gene regulatory mechanism (Yordanova, et al. 2018; Seoighe, et al. 2020). 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 two stops and eventually infringing upon the 3' end of the coding sequence itself. At this point, translation of this mRNA molecule is blocked (Yordanova, et al. 2018). 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 non-canonical 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 isn't seen (Ho and Hurst 2020). Instead TAA usage correlates positively with N_e (Ho and Hurst 2020), 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 (Wu and Hurst 2015)).

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 (Lassalle, et al. 2015). 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] (Chen and MacCarthy 2017; Seplyarskiy, et al. 2017). More specifically, several DNA repair processes are known to be affected by local sequence context (Cai, et al. 2010)

including, at least in bacteria, mismatch repair (Mazurek, et al. 2009), the process underpinning gBGC. Here, sequence contexts that enhance localised DNA flexibility are associated with mismatch repair activation (Mazurek, et al. 2009) (see also: Isaacs, et al. 2002; Wang, et al. 2003). Similar evidence for a role of local DNA flexibility has been found in yeast (Isaacs, et al. 2002; Li, et al. 2019). 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 (Bacolla, et al. 1997). Evidence in humans for more effective repair of flexible DNA owing to local sequence context (Ruzicka, et al. 2019) 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 three 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.

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 three 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 (Katoh, et al. 2005).

Rather than using parsimony as done previously (Rogozin, et al. 2016; Belinky, et al. 2018), stop codon switches were reconstructed using a maximum likelihood approach. For each species triplet, ancestral nucleotide states for the internal node between the two ingroups were inferred by maximum likelihood using IQTree v2.1.2 with the -asr flag (Nguyen, et al. 2015; Minh, et al. 2020). 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).

Predicting equilibrium TGA content using flux data

The predicted TGA usage for a given lineage, pTGA, was calculated by adapting the formulae outlined by Long, et al. (2018). 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:

$$P_n = 1 + \frac{1}{m}$$

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:

$$P_{TGA} = 1 + \frac{1}{s}$$

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

The difference in pTGA observed between two gene groups (“A and B”, GC rich and GC poor genes, or highly recombining and lowly recombining genes) 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 one pTGA score from each of the two 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.

Intronic GC as a proxy for isochore GC content

Under the assumption that intronic GC reflects isochore GC content, intronic nucleotide sequences were extracted from one 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 spectrum of intronic GC contents, 10% percentiles were calculated, and genes were binned accordingly. The stop switch method described above was applied to each bin to measure changes in stop switch frequencies across intronic GC contents.

This binning method is effective to segregate genes evenly across intragenomic GC contents but does not allow comparisons between eukaryotic groups. To plot the stop codon switches of multiple different species on the same axis requires a GC-matching methodology. To achieve this, genes were binned at 5% intronic GC content intervals (e.g. genes of GC content between 27.5% and 32.5% would be allocated to the 30% bin). As this method does not use percentiles, the resulting bins are not pre-designated to be equal in size. Bins of insufficient size ($n < 50$) were discarded. As before, the stop switch method was then applied to the GC-matched bins to measure changes in stop switch frequencies.

Calculating mutational equilibria

The equilibrium content of all four nucleotides (indicated N^*) may be estimated using the full mutational spectrum (Charneski, et al. 2011; Rice, et al. 2020). A full spectrum of 108,778 *de novo* mutations (from 1,548 Icelandic human family trios) was downloaded from the supplementary material of Jonsson, et al. (2017). 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 Rice et al. (2020). The same theory can be applied to the three stop codons to predict their equilibrium frequencies as follows, where TAA' indicates the frequency of TAA after some period of time:

$$TAA' = TAA (1 - TAA \rightarrow TGA - TAA \rightarrow TAG) + TGA (TGA \rightarrow TAA) + TAG (TAG \rightarrow TAA)$$

$$TGA' = TGA (1 - TGA \rightarrow TAA - TGA \rightarrow TAG) + TAA (TAA \rightarrow TGA) + TAG (TAG \rightarrow TGA)$$

$$TAG' = TAG (1 - TAG \rightarrow TAA - TAG \rightarrow TGA) + TAA (TAA \rightarrow TAG) + TGA (TGA \rightarrow TAG)$$

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

$$TAA (TAA \rightarrow TGA + TAA \rightarrow TAG) = TGA (TGA \rightarrow TAA) + TAG (TAG \rightarrow TAA)$$

$$TGA (TGA \rightarrow TAA + TGA \rightarrow TAG) = TAA (TAA \rightarrow TGA) + TAG (TAG \rightarrow TGA)$$

$$TAG (TAG \rightarrow TAA + TAG \rightarrow TGA) = TAA (TAA \rightarrow TAG) + TGA (TGA \rightarrow TAG)$$

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 one arbitrarily chosen stop codon frequency with 1 – the sum of the other two. While this would be achieved most accurately using precise mutational flux data between TAA, TGA, and TAG this is not captured within the Jonsson (Jonsson, et al. 2017) dataset. Instead, we estimate flux between the three stops using null frequencies proposed by Belinky et al. (Belinky, et al. 2018). 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 x A>G x G>A.

The full spectrum of 108,778 *de novo* mutations may also be analysed using a 16x16 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 (Wang, et al. 2015). 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 (Frazer, et al. 2007) 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 10kb 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 10kb band it falls within. The data were downloaded from <https://www.decode.com/addendum/> (last accessed 14 September 2020) (Kong, et al. 2010).

Coordinates of the two 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 one ‘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 lowly expressed genes (“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 (Ho and Hurst 2020). Pagel’s lambda (λ) was predicted by maximum likelihood. Species used in this analysis were the same as published in our previous analysis (Ho and Hurst 2020), with the input phylogenetic trees generated using TimeTree (Kumar et al. 2017) and available in our GitHub repository along with the data required to repeat this analysis. TGA enrichment scores in LEGs were calculated such that:

$$TGA\ enrichment = \frac{TGA\ usage\ at\ the\ primary\ stop - mean(TGA\ usage\ downstream)}{mean(TGA\ usage\ downstream)}$$

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:

$$TGA\ usage = \frac{TGA\ frequency}{TAA\ frequency + TGA\ frequency + TAG\ frequency}$$

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, et al. (2020). 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 (2019). 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.

Data access

Raw data used are all publicly available and accessible as outlined in the methods section. All data manipulation was performed using bespoke Python 3.6 scripts. Statistical analyses and data visualisations were performed using R 3.3.3. Scripts required for replication of the described analyses can be found at <https://github.com/ath32/gBGC>.

Competing interests

The authors have no conflicts of interest to disclose.

Acknowledgements

This work was supported by the European Research Council (grant EvoGenMed ERC-2014-ADG 669207 to L.D.H.).

References

- Arribere JA, Cenik ES, Jain N, Hess GT, Lee CH, Bassik MC, Fire AZ. 2016. Translation readthrough mitigation. *Nature*. 534(7609): 719-723.
- Babbitt GA, Cotter CR. 2011. Functional conservation of nucleosome formation selectively biases presumably neutral molecular variation in yeast genomes. *Genome Biol. Evol.* 3(1): 15-22.
- Babbitt GA, Schulze KV. 2012. Codons support the maintenance of intrinsic DNA polymer flexibility over evolutionary timescales. *Genome Biol. Evol.* 4(9): 954-965.
- Bacolla A, Gellibolian R, Shimizu M, Amirhaeri S, Kang S, Ohshima K, Larson JE, Harvey SC, Stollar BD, Wells RD. 1997. Flexible DNA: Genetically unstable CTG center dot CAG and CGG center dot CCG from human hereditary neuromuscular disease genes. *J. Biol. Chem.* 272(27): 16783-16792.

- Basham B, Schroth GP, Ho PS. 1995. An A-DNA triplet code - Thermodynamic rules for predicting A-DNA and B-DNA. *Proc. Natl. Acad. Sci. U.S.A.* 92(14): 6464-6468.
- Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci. Rep.* 8(1): 9260.
- Bengtsson BO. 1985. Biased conversion as the primary function of recombination. *Genet. Res.* 47(1): 77-80.
- Berg OG. 1996. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics.* 142(4): 1379-1382.
- Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol.* 7(1): 45-62.
- Bernardi G, Olofsson B, Filipinski J, Zerial M, Salinas J, Cuny G, Meunierrotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science.* 228(4702): 953-958.
- Beznoskova P, Gunisova S, Valasek LS. 2016. Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. *RNA.* 22(3): 456-466.
- Blaschke RJ, Rappold G. 2006. The pseudoautosomal regions, SHOX and disease. *Curr. Opin. Genet. Dev.* 16(3): 233-239.
- Bolivar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, Ellegren H. 2018. Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for It. *Mol. Biol. Evol.* 35(10): 2475-2486.
- Brown TC, Jiricny J. 1988. Different base base mispairs are corrected with different efficiencies and specificities in monkey kidney-cells. *Cell.* 54(5): 705-711.
- Brown TC, Jiricny J. 1989. Repair of base base mismatches in simian and human-cells. *Genome.* 31(2): 578-583.
- Cai Y, Patel DJ, Broyde S, Geacintov NE. 2010. Base sequence context effects on nucleotide excision repair. *J. Nucleic Acids.* 2010(1): 174252.

Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical at skew in firmicute genomes results from selection and not from mutation. *PLoS Genet.* 7(9): e1002283.

Chen J, MacCarthy T. 2017. 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.* 13(3): e1005471.

Clegg JB, Weatherall DJ, Milner PF. 1971. Haemoglobin constant spring - a chain termination mutant? *Nature.* 234(5328): 337-340.

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. 2010. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods.* 7(4): 250-251.

Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. *Genome Biol. Evol.* 9(11): 2987-3007.

Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 46(4): 1927-1944.

de Oliveira JL, Morales AC, Hurst LD, Urrutia AO, Thompson CRL, Wolf JB. 2021. Inferring adaptive codon preference to understand sources of selection shaping codon usage bias. *Mol. Biol. Evol.* 38(8): 3247-3266.

Dimitrova LN, Kuroha K, Tatematsu T, Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* 284(16): 10343-10352.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. *Genome Res.* 17(10): 1420-1430.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature.* 287(5782): 560-561.

- Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Elife*. 2(1): e01179.
- Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12(6): 640-649.
- Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4(5): e1000071.
- Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genom. Hum. Genet.* 10(1): 285-311.
- Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics*. 162(4): 1837-1847.
- Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B.* 252(1335): 237-243.
- Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat. Rev. Genet.* 2(7): 549-555.
- Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L, La Starza R, Diverio D, Colombo E, Santucci A, et al. 2005. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *New Engl. J. Med.* 352(3): 254-266.
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 449(7164): 851–861.
- Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* 22(3): 650-658.
- Fullerton SM, Bernardo Carvalho A, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* 18(6): 1139-1142.

- Galtier N. 2021. Fine-scale quantification of GC-biased gene conversion intensity in mammals. *bioRxiv*: doi: 10.1101/2021.1105.1105.442789.
- Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25(1): 1-5.
- Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics.* 159(2): 907-911.
- Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glemin S, Bierne N, Duret L. 2018. Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Mol. Biol. Evol.* 35(5): 1092-1103.
- Geller AI, Rich A. 1980. A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature.* 283(5742): 41-46.
- Glemin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res.* 25(8): 1215-1228.
- Gutz H, Leslie JF. 1976. Gene conversion: a hitherto overlooked parameter in population genetics. *Genetics.* 83(4): 861-866.
- Hahn MW, Stajich JE, Wray GA. 2003. The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20(6): 901-906.
- Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir A, Jonasdottir A, Sulem P, et al. 2016. The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* 48(11): 1377-1384.
- Harrison RJ, Charlesworth B. 2011. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu stricto* group of yeasts. *Mol. Biol. Evol.* 28(1): 117-129.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9): e1001115.

- Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* 8(3): 269-294.
- Ho AT, Hurst LD. 2020. Effective population size predicts local rates but not local mitigation of read-through errors in eukaryotic genes. *Mol. Biol. Evol.* 38(1): 244–262.
- Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. *PLoS Genet.* 15(9): e1008386.
- Hollingsworth TJ, Gross AK. 2013. The severe autosomal dominant retinitis pigmentosa rhodopsin mutant Ter349Glu mislocalizes and induces rapid rod cell death. *J. Biol. Chem.* 288(40): 29047-29055.
- Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18(9): 486-487.
- Isaacs RJ, Rayens WS, Spielmann HP. 2002. 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.* 319(1): 191-207.
- Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature.* 549(7673): 519–522.
- Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res.* 21(12): 2096-2113.
- Karlin S, Mrazek J. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.* 94(19): 10227-10232.
- Katoh K, Kuma K-i, Miyata T, Toh H. 2005. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform.* 16(1): 22-33.

- Kiktev DA, Sheng ZW, Lobachev KS, Petes TD. 2018. GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 115(30): E7109-E7118.
- Klauer AA, van Hoof A. 2012. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *WIREs RNA*. 3(5): 649-660.
- Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*. 467(7319): 1099-1103.
- Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* 289(44): 30334-30342.
- Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4(6): 933-942.
- Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am. J. Hum. Genet.* 95(4): 408-420.
- Langley SA, Karpen GH, Langley CH. 2014. Nucleosomes shape DNA polymorphism and divergence. *PLoS Genet.* 10(7): e1004457.
- Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. *PLoS Genet.* 11(2): e1004941.
- Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores: evidence from SNP frequency distributions. *Genetics*. 162(4): 1805-1810.
- Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* 12(19): 2411-2415.
- Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15(5): e1008141.

- Li Y, Lombardo Z, Joshi M, Hingorani MM, Mukerji I. 2019. Mismatch recognition by *Saccharomyces cerevisiae* Msh2-Msh6: Role of structure and dynamics. *Int. J. Mol. Sci.* 20(17): 4271.
- Liu HX, Huang J, Sun XG, Li J, Hu YW, Yu LY, Liti GN, Tian DC, Hurst LD, Yang SH. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat. Ecol. Evol.* 2(1): 164-173.
- Liu HX, Maclean CJ, Zhang JZ. 2019. Evolution of the yeast recombination landscape. *Mol. Biol. Evol.* 36(2): 412-422.
- Long HA, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo WF, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* 2(2): 237–240.
- Lynch M. 2007. *The origins of genome architecture*. Sunderland, MA.: Sinauer Associates Inc.
- Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 105(27): 9272-9277.
- Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature.* 454(7203): 479-485.
- Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome Biol.* 5(7): R45.
- Marsolier-Kergoat MC, Yeramian E. 2009. GC content and recombination: Reassessing the causal effects for the *Saccharomyces cerevisiae* genome. *Genetics.* 183(1): 31-38.
- Mazurek A, Johnson CN, Germann MW, Fishel R. 2009. Sequence context effect for hMSH2-hMSH6 mismatch-dependent activation. *Proc. Natl. Acad. Sci. U.S.A.* 106(11): 4177-4182.

- McCaughan KK, Brown CM, Dalphin ME, Berry MJ, Tate WP. 1995. Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. U.S.A.* 92(12): 5431-5435.
- Meng SY, Hui JO, Haniu M, Tsai LB. 1995. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 211(1): 40-48.
- Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* 21(6): 984-990.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37(5): 1530-1534.
- Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2020. Codon usage and splicing jointly influence mRNA localization. *Cell Sys.* 10(4): 351-362.
- Mrazek J, Karls AC. 2019. In silico simulations of occurrence of transcription factor binding sites in bacterial genomes. *BMC Evol. Biol.* 19(1): 67.
- Nagylaki T. 1983. Evolution of a large population under gene conversion. *Proc. Natl. Acad. Sci. U.S.A.* 80(19): 5941-5945.
- Namy O, Duchateau-Nguyen G, Rousset JP. 2002. Translational readthrough of the PDE2 stop codon modulates cAMP levels in *Saccharomyces cerevisiae*. *Mol. Microbiol.* 43(3): 641-652.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32(1): 268-274.
- Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics.* 148(3): 929-936.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. System.* 23(1): 263-286.

- Pang SY, Wang WH, Rich B, David R, Chang YT, Carbutaru G, Myers SE, Howie AF, Smillie KJ, Mason JI. 2002. 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.* 87(6): 2556-2563.
- Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.* 53(3): 273-298.
- Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* 4(7): 675-682.
- Pond SLK, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22(5): 1208-1222.
- Ponting CP. 2017. Biological function in the twilight zone of sequence conservation. *BMC Biol.* 15(1): 1-9.
- Ponting CP. 2008. The functional repertoires of metazoan genomes. *Nat. Rev. Genet.* 9(9): 689-698.
- Prendergast JGD, Semple CAM. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res.* 21(11): 1777-1787.
- Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. B.* 365(1552): 2571-2580.
- Rice AM, Morales AC, Ho AT, Mordstein C, Muhlhausen S, Watson S, Cano L, Young B, Kudla G, Hurst LD. 2020. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. *Mol. Biol. Evol.* 38(1): 67-83.
- Roberts SA, Gordenin DA. 2014. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer.* 14(12): 786-800.

- Robinson MC, Stone EA, Singh ND. 2014. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. *Mol. Biol. Evol.* 31(2): 425-433.
- Rodnina MV. 2016. The ribosome in action: Tuning of translational efficiency and protein folding. *Protein Sci.* 25(8): 1390-1406.
- Rodnina MV, Korniy N, Klimova M, Karki P, Peng BZ, Senyushkina T, Belardinelli R, Maracci C, Wohlgemuth I, Samatova E, et al. 2020. Translational recoding: canonical translation mechanisms reinterpreted. *Nucleic Acids Res.* 48(3): 1056-1067.
- Rogozin IB, Belinky F, Pavlenko V, Shabalina SA, Kristensen DM, Koonin EV. 2016. Evolutionary switches between two serine codon sets are driven by selection. *Proc. Natl. Acad. Sci. U.S.A.* 113(46): 13109-13113.
- Roth JR. 1970. UGA nonsense mutations in *Salmonella-typhimurium*. *J. Bacteriol.* 102(2): 467-475.
- Rousselle M, Laverre A, Figuet E, Nabholz B, Galtier N. 2019. Influence of recombination and GC-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. *Mol. Biol. Evol.* 36(3): 458-471.
- Roy B, Leszyk JD, Mangus DA, Jacobson A. 2015. Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *Proc. Natl. Acad. Sci. U.S.A.* 112(10): 3038-3043.
- Ruzicka M, Soucek R, Kulhanek P, Radova L, Fajkusova L, Reblova K. 2019. Bending of DNA duplexes with mutation motifs. *DNA Res.* 26(4): 341-352.
- Ryden SM, Isaksson LA. 1984. 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.* 193(1): 38-45.
- Sanchez JC, Padron G, Santana H, Herrera L. 1998. Elimination of an HuIFN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J. Biotechnol.* 63(3): 179-186.

- Schuere F, Thoms S. 2016. Functional translational readthrough: a systems biology perspective. *PLoS Genet.* 12(8): e1006196.
- Seoighe C, Kiniry SJ, Peters A, Baranov PV, Yang H. 2020. Selection shapes synonymous stop codon use in mammals. *J. Mol. Evol.* 88(7): 549-561.
- Seplyarskiy VB, Andrianova MA, Bazykin GA. 2017. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Res.* 27(2): 175-184.
- Smeds L, Mugal CF, Qvarnstrom A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genet.* 12(5): e1006044.
- Smith NGC, Eyre-Walker A. 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. *Mol. Biol. Evol.* 18(6): 982-986.
- Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germline de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet.* 14(3): e1007254.
- Strigini P, Brickman E. 1973. Analysis of specific misreading in *Escherichia coli*. *J. Mol. Biol.* 75(4): 659-672.
- Sun H, Yu GJ. 2019. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci. Rep.* 9(1): 1667.
- Sun JH, Ai SM, Luo HJ, Gao B, Ieee. 2019. Estimation of the equilibrium GC content of human genome. 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology; Hangzhou, China. p. 12-17.
- Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. U.S.A.* 87(12): 4692-4696.
- Trotta E. 2016. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics.* 17(17): 366.

- Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*. 141(2): 344-354.
- Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. *BMC Evol. Biol.* 7(1): 226.
- Vidal R, Frangione B, Rostagno A, Mead S, Revesz T, Plant G, Ghiso J. 1999. A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature*. 399(6738): 776-781.
- Vidal R, Revesz T, Rostagno A, Kim E, Holton JL, Bek T, Bojsen-Moller M, Braendgaard H, Plant G, Ghiso J, et al. 2000. 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.* 97(9): 4920-4925.
- Vinogradov AE. 2001. Bendable genes of warm-blooded vertebrates. *Mol. Biol. Evol.* 18(12): 2195-2200.
- Wagner A. 2005. Robustness, evolvability, and neutrality. *FEBS Lett.* 579(8): 1772-1778.
- Wang H, Yang Y, Schofield MJ, Du CW, Fridman Y, Lee SD, Larson ED, Drummond JT, Alani E, Hsieh P, et al. 2003. DNA bending and unbending by MutS govern mismatch recognition and specificity. *Proc. Natl. Acad. Sci. U.S.A.* 100(25): 14822-14827.
- Wang MC, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 15(18): 3163-3168.
- Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet.* 4(11): e1000250.
- Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol. Biol. Evol.* 24(12): 2755-2762.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15(12): 549.

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R, et al. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife.* 4(1): e04637.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature.* 337(6204): 283-285.

Wu XM, Hurst LD. 2015. 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.* 32(7): 1847-1861.

Xu C, Park JK, Zhang JZ. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biol.* 17(3): e3000197.

Xu C, Zhang JZ. 2018. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. *Cell Sys.* 6(6): 734-742.

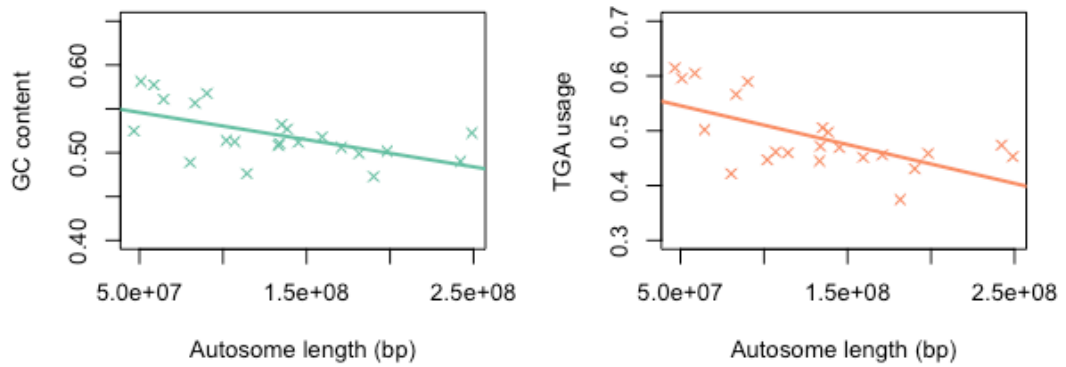
Yang ZH, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15(12): 496-503.

Yordanova MM, Loughran G, Zhdanov AV, Mariotti M, Kiniry SJ, O'Connor PBF, Andreev DE, Tzani I, Saffert P, Michel AM, et al. 2018. AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation. *Nature.* 553(7688): 356-360.

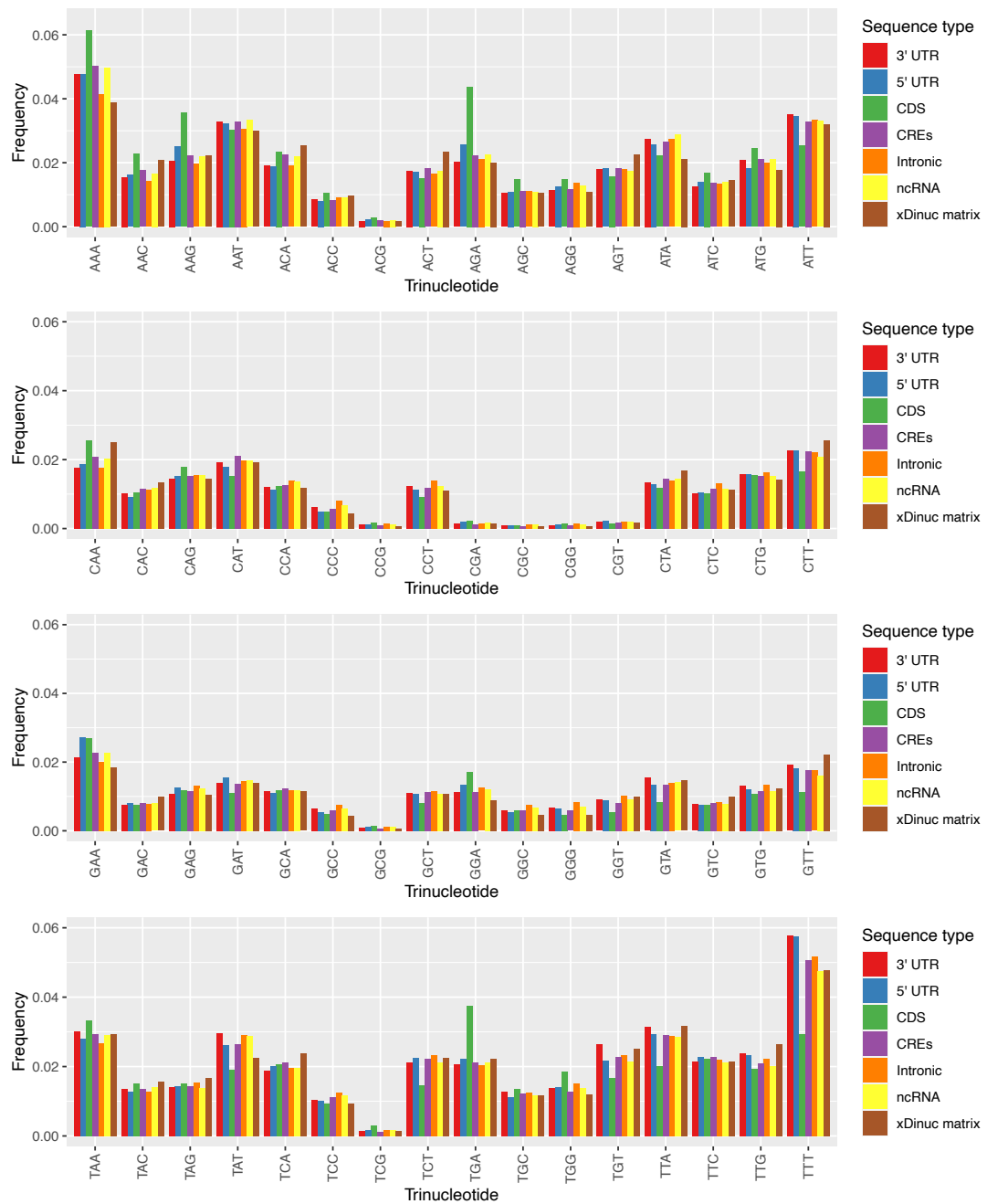
Supplementary information for: Conservation need not imply purifying selection: evidence from mammalian stop codon usage

Alexander T. Ho and Laurence D. Hurst

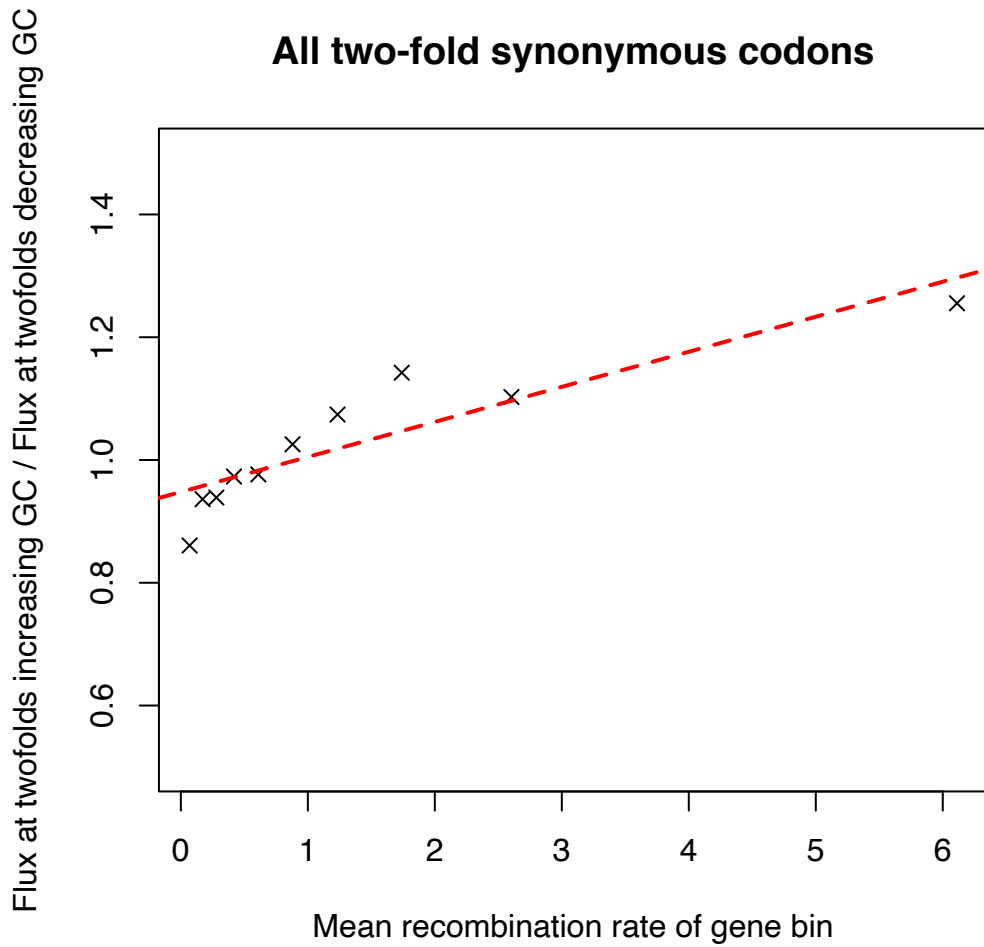
PLoS Biology, (under review)



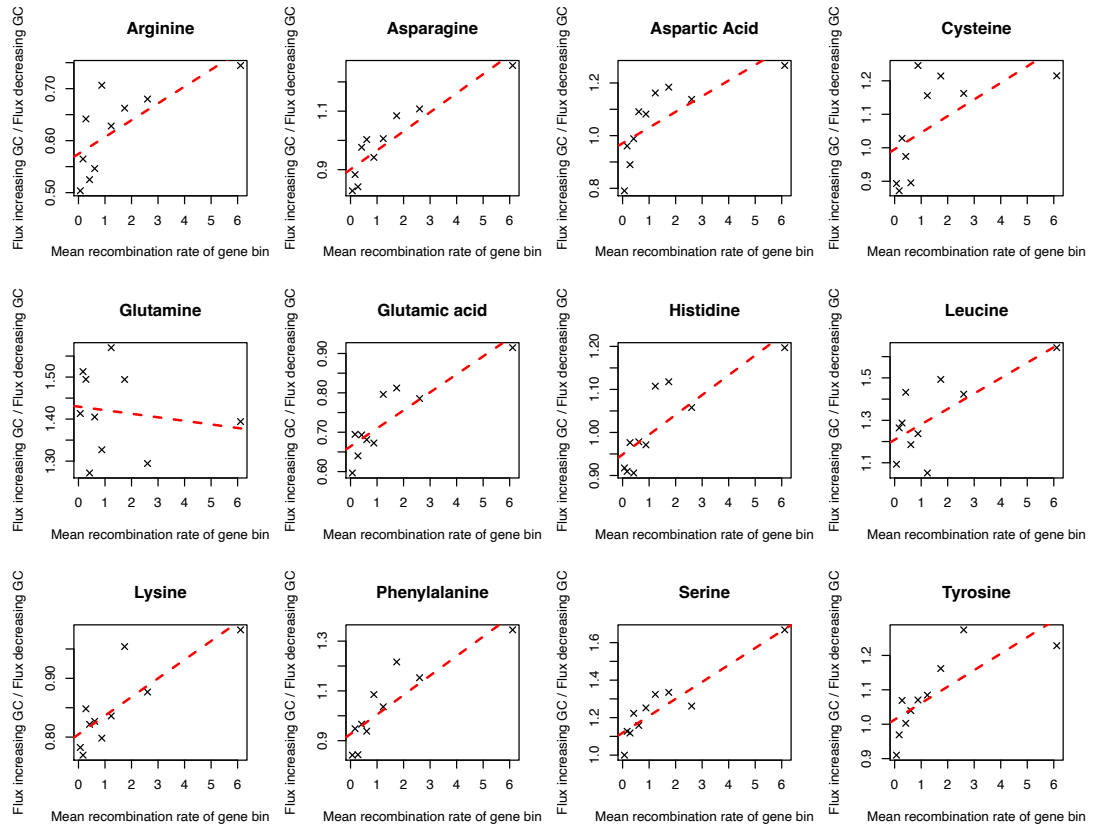
S1 Fig. The relationships of autosome length with GC content and TGA usage in the human genome. Autosomal size (bp length) is negatively associated with G+C content (Spearman's rank; $p = 0.0078$, $\rho = -0.56$, $n = 22$) and TGA usage (Spearman's rank; $p = 0.0094$, $\rho = -0.55$, $n = 22$).



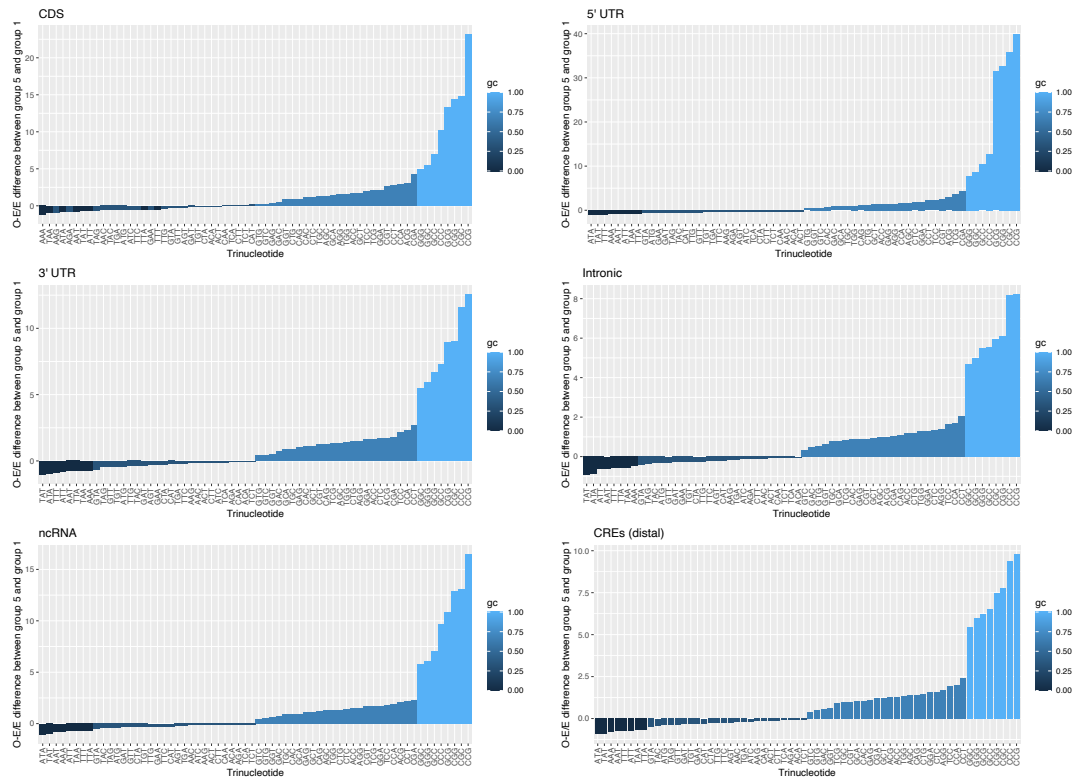
S2 Fig. Trinucleotide frequencies in six sets of different genomic sequences (between 0%-36.31% GC content) compared to dinucleotide matrix-derived equilibrium predictions. The GC range used is the bottom 20% of genes to avoid the possible confounding effects of biased gene conversion. CDS refers to coding sequence, CREs to cis-regulatory elements. “xDinuc matrix” refers to equilibrium estimates of trinucleotide frequencies derived from a dinucleotide mutational matrix.



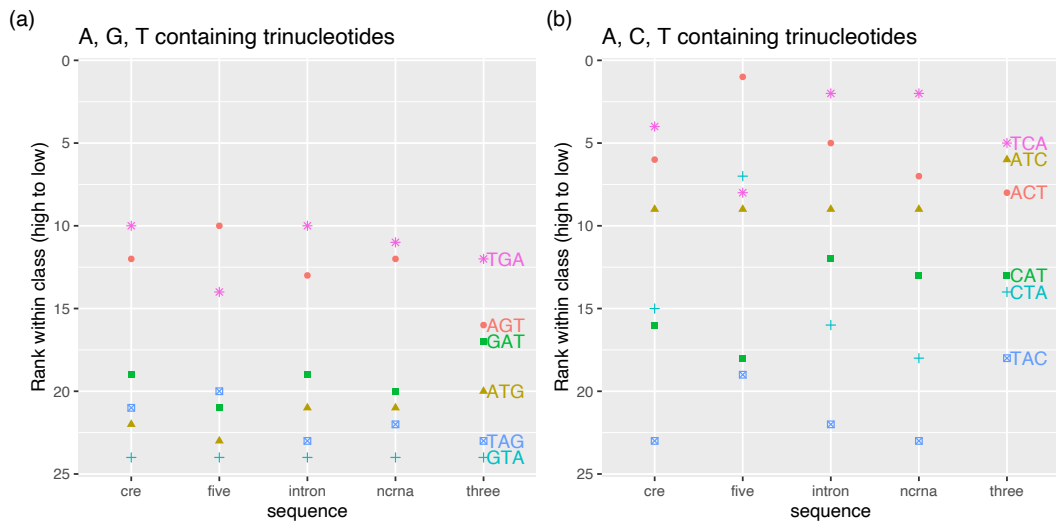
S3 Fig. The rate of flux increasing GC content at twofold degenerate sites divided by the rate of flux decreasing GC content at the same sites across 10 gene bins of increasing recombination rate. Flux to the G+C-rich codons is most strongly favoured at high recombination rates (Spearman's rank; $p 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



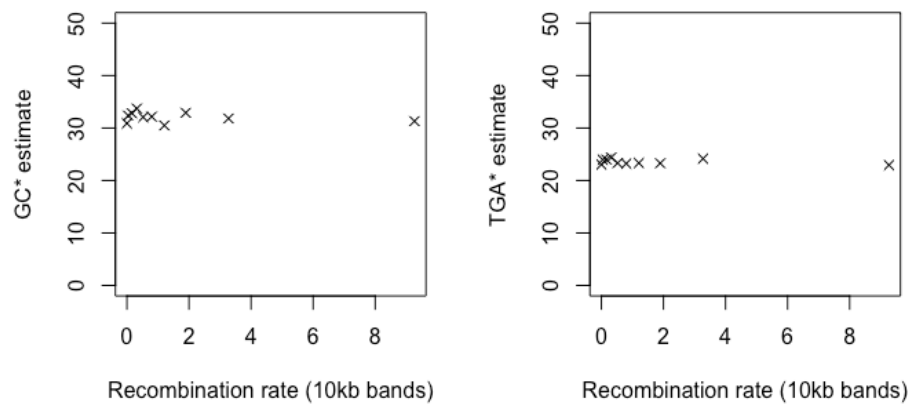
S4 Fig. The rate of flux increasing GC content at twofold degenerate sites divided by the rate of flux decreasing GC content at the same sites across 10 gene bins of increasing recombination rate for each appropriate amino acid. Flux increasing GC content are significantly favoured in regions with higher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



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 *de novo* mutations, DNMs) and fixed trinucleotide frequencies (from 10kb 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.



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 GC-biased gene conversion (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 cis-regulatory elements (cre), 5’ UTR (five), intronic (intron), ncRNA, (ncrna), and 3’ UTR (three). CDS sequences are excluded from this analysis as they are much more prone to selection and other potential fixation biases.



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 10kbp environment. Mutations in non-recombining regions were discarded. The remaining mutations were split into bins of equal size (~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.

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.

Sequence	Model	Estimate	P-value
5' UTR	TAA ~ Intronic G+C	-0.524365	<2e-16
	TGA ~ Intronic G+C	0.544688	<2e-16
	TAG ~ Intronic G+C	-0.02032	0.116
3' UTR	TAA ~ Intronic G+C	-0.688116	<2e-16
	TGA ~ Intronic G+C	0.715928	<2e-16
	TAG ~ Intronic G+C	-0.027812	3.52e-06
Intronic	TAA ~ Coding sequence GC3	-0.015645	<2e-16
	TGA ~ Coding sequence GC3	0.01354	<2e-16
	TAG ~ Coding sequence GC3	0.0007857	0.598

S2 Table. The 4 x 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. 95% confidence intervals (C.I.) were calculated using the `Poisson.test` function in R under the assumption that the observed number of mutations is a Poisson variable.

Reference allele	Estimate	Derived allele			
		A	T	C	G
A	Rate	-	3.87×10^{-6}	4.14×10^{-6}	1.56×10^{-5}
	Upper C.I.	-	4.00×10^{-6}	4.28×10^{-6}	1.58×10^{-5}
	Lower C.I.	-	3.74×10^{-6}	4.01×10^{-6}	1.53×10^{-5}
T	Rate	4.05×10^{-6}	-	1.56×10^{-5}	4.06×10^{-6}
	Upper C.I.	4.18×10^{-6}	-	1.59×10^{-5}	4.19×10^{-6}
	Lower C.I.	3.91×10^{-6}	-	1.54×10^{-5}	3.93×10^{-6}
C	Rate	6.47×10^{-6}	3.55×10^{-5}	-	8.08×10^{-6}
	Upper C.I.	6.67×10^{-6}	3.60×10^{-5}	-	8.31×10^{-6}
	Lower C.I.	6.27×10^{-6}	3.51×10^{-5}	-	7.85×10^{-6}
G	Rate	3.55×10^{-5}	6.32×10^{-6}	8.09×10^{-6}	-
	Upper C.I.	3.59×10^{-5}	6.53×10^{-6}	8.32×10^{-6}	-
	Lower C.I.	3.50×10^{-5}	6.12×10^{-6}	7.87×10^{-6}	-

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

TAG receives less of a “boost” in GC rich domains than TGA. Why is this? As the metric starts by specifying the expected trinucleotide abundance given known mutational profiles, we can eliminate mutation bias, unless it is even more complex than we permitted. Instead, the data support complex fixation biases. Fixation bias may imply selection for or against certain *k*-mers. Why might TAG, and its dinucleotides TA and AG, be selectively avoided in the genome? A parsimonious rationale should explain why TA dinucleotides appear to be under-represented near-universally (Burge, et al. 1992) and why, as we observe, the effect is seen in transcribed and (what we presume to be) untranscribed domains (cis-regulatory elements).

Such generality might point to DNA’s biophysics, for example assumption of A and B forms. However, within the 33% trinucleotide class the two least boosted trinucleotides, TAG and GTA predispose to B and A form respectively (Basham, et al. 1995). Similarly, TAC and TAG predispose to opposite forms but have comparable low fixation bias and, indeed, when we correlate A-DNA propensity energy (APE) against our gBGC “boost” scores for each trinucleotide, we find no significant correlations in 5’ UTR, 3’ UTR, intronic, ncRNA, or cis-regulatory element sequences (Spearman’s rank tests, all $p > 0.05$).

An alternative DNA structural hypothesis is that TA (or certain TA-containing oligonucleotides) might adversely affect chromatin structure (Burge, et al. 1992), probably because AT-rich DNA tends to be concentrated in the nucleosome-free regions associated with transcription start sites. However, bacteria don’t have nucleosomes but nonetheless avoid TAG (Korkmaz, et al. 2014; Ho and Hurst 2019). TA might also be avoided due to selection against UpA motifs on RNA molecules that are targeted by ribonucleases such as RNase L during the antiviral immune response (Floydsmith, et al. 1981; Wreschner, et al. 1981). However, non-transcribed domains show the same trends and why bacteria also avoid TAG is unexplained. A similar problem faces the notion of “AG exclusion zones” that are important for splicing accuracy (Wahl, et al. 2009; Wimmer, et al. 2020). Whether this could explain

genome-wide avoidance of AG dinucleotides seems unlikely given trends in non-transcribed domains. Splicing is also of little relevance in bacteria.

Perhaps the most compelling model is one proposing avoidance of transcription initiation motifs. 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 (Karlin and Mrazek 1997; Mrazek and Karls 2019). Indeed, within the trinucleotides with only A and T, ATA and TAT, the two that are core to TATA box, are consistently the two with the lowest “boost” (Supplementary fig 6). However, a TATA box is classically TATA[A|T]A[A|T]. Why such a motif would select against TAG but not TAA (in the 0% GC class TAA typically has a high boost) is not clear. We need also to be wary of post hoc hypothesising. Indeed, one might also predict selection against CAA or AAT owing to their involvement in CAAT boxes, commonly located about 150 bp 5’ of TATA boxes. We see no evidence for either, CAA indeed being one of the most “boosted” of the 33% GC class and AAT being unexceptional.

Perhaps the most important objection to any such model is that one must suppose efficient selection against a point mutation causing spurious transcription or sequestration of TATA-binding protein which, when population sizes are small (e.g. mammals and birds), seems unlikely. 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 (Hahn, et al. 2003). If then *Escherichia coli*'s N_e is of the order of 10^8 (Berg 1996), then s must be $\sim -0.09/10^8 = -9 \times 10^{-10}$. For a mutation to be under selection $s \sim 1/2 N_e$ must hold. In a species with $N_e \sim 10,000$ (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 spurious binding sites.

More generally, in principle the fixation bias associated with high recombination rates could be compatible with some form of nucleotide level selection that prefers G+C residues. However, *a priori* such selective models are hard to reconcile with

inefficient selection associated with low N_e , so in turn non-selective fixation biases appear more parsimonious.

References

Basham B, Schroth GP, Ho PS. 1995. An A-DNA triplet code - Thermodynamic rules for predicting A-DNA and B-DNA. *Proc. Natl. Acad. Sci. U.S.A.* 92(14): 6464-6468.

Berg OG. 1996. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics*. 142(4): 1379-1382.

Burge C, Campbell AM, Karlin S. 1992. Over-representation and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* 89(4): 1358-1362.

Floydsmith G, Slattery E, Lengyel P. 1981. Interferon action - RNA cleavage pattern of a (2'-5')oligoadenylate-dependent endonuclease. *Science*. 212(4498): 1030-1032.

Hahn MW, Stajich JE, Wray GA. 2003. The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20(6): 901-906.

Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. *PLoS Genet.* 15(9): e1008386.

Karlin S, Mrazek J. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.* 94(19): 10227-10232.

Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* 289(44): 30334-30342.

Mrazek J, Karls AC. 2019. In silico simulations of occurrence of transcription factor binding sites in bacterial genomes. *BMC Evol. Biol.* 19(1): 67.

Wahl MC, Will CL, Luhrmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell*. 136(4): 701-718.

Wimmer K, Schamschula E, Wernstedt A, Traunfellner P, Amberger A, Zschocke J, Kroisel P, Chen YJ, Callens T, Messiaen L. 2020. AG-exclusion zone revisited: Lessons to learn from 91 intronic NF1 3' splice site mutations outside the canonical AG-dinucleotides. *Hum. Mutat.* 41(6): 1145-1156.

Wreschner DH, McCauley JW, Skehel JJ, Kerr IM. 1981. Interferon action-sequence specificity of the ppp(A2'p)nA-dependent ribonuclease. *Nature.* 289(5796): 414-417.

Chapter 6

Discussion (part 1)

Stop codon usage as a window into genome evolution: mutation, selection, biased gene conversion and the TAG paradox

Alexander T. Ho and Laurence D. Hurst

Genome Biology & Evolution (under review)

This chapter contains a draft manuscript written as a commissioned review article for Genome Biology & Evolution. At the time of writing, this manuscript is under review.

Pre-amble

All the preceding chapters provide new insights into the evolution of stop codon usage within bacterial and eukaryotic genomes. In bacteria, stop codon usage appears not to evolve in direct response to the cellular abundance of RF1 and RF2 but as a product of neutral evolutionary processes and selection for reduced translational read-through rates. In eukaryotes there is no RF-linked hypothesis and stop codon usage is expected to be influenced by the same evolutionary forces, with the possible addition of biased gene conversion that might be increasing TGA usage in species that possess GC-biased mismatch repair machinery. TAA appears to be the optimal stop codon in all taxa, even those that are impacted by GC-biased gene conversion, as evidenced by the stop codon preferences of highly expressed genes and comparisons of stop codon usage against dinucleotide-controlled null models.

With regards genomic error-proofing, my results suggest there is a preference to reduce TR rate rather than evolve error mitigation devices at genic level. Bacteria do not enrich their 3' UTR sequences with additional stop codons to act as fail-safe mechanisms should the first stop codon fail, but they do prefer TAA stop codons (the least error prone stop variant) in their highly expressed genes. While some unicellular eukaryotes do appear to be enriched for ASCs, others are not, and ASC enrichment is no more common than chance in multicellular species. Controlling for phylogeny, in eukaryotes TAA enrichment correlates with effective population size (N_e) as predicted by nearly neutral theory, while ASC enrichment does not.

More generally, my results demonstrate the utility of the stop codon exemplar for studying molecular evolution. That TAA, TGA, and TAG possess different TR rates permits the study of selective differences between three ostensibly synonymous codons. That TGA/TAG and TAA differ in GC content allows one to study AT- or GC-favoured mutation or fixation biases. That TGA and TAG are equal in nucleotide content allows us to control for GC content and consider more complex models of mutation or fixation bias. All of this I summarise here, in a commissioned review of stop codon usage for *Genome Biology & Evolution*. As this manuscript places all prior chapters within the stop codon literature, I present this chapter as “discussion part 1”.

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:			
Stop codon usage as a window into genome evolution: mutation, selection, biased gene conversion and the TAG paradox			
Publication status (tick one)			
Draft manuscript	<input type="checkbox"/>	Submitted	<input type="checkbox"/>
In review	<input checked="" type="checkbox"/>	Accepted	<input type="checkbox"/>
Published	<input type="checkbox"/>		
Publication details (reference)	N/A		
Copyright status (tick the appropriate statement)			
I hold the copyright for this material	<input checked="" type="checkbox"/>	Copyright is retained by the publisher, but I have been given permission to replicate the material here	<input type="checkbox"/>
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	<p>The candidate contributed to / considerably contributed to / predominantly executed the...</p> <p>Formulation of ideas: 100%</p> <p>Design of methodology: 100%</p> <p>Bioinformatic analyses: 100%</p> <p>Experimental work: N/a</p> <p>Presentation of data in journal format: 100%</p>		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed		Date	10/03/2022

Issue section: Review

Stop codon usage as a window into genome evolution: mutation, selection, biased gene conversion and the TAG paradox

Alexander T. Ho^{1*} and Laurence D. Hurst¹

1. Milner Centre for Evolution, University of Bath, Bath, UK

*Author for correspondence: a.t.ho@bath.ac.uk

Abstract

Protein coding genes terminate with one of three stop codons (TAA, TGA or TAG) that, like synonymous codons, are not employed equally. With TGA and TAG having identical nucleotide content, analysis of their differential usage provides an unusual window into the forces operating on what are ostensibly functionally identical residues. Across genomes and between isochores within the human genome, TGA usage increases with GC content but, with a universal GC→AT mutation bias, this cannot be explained by mutation bias-drift equilibrium. Increased usage of TGA in GC-rich genomes or genomic regions is also unlikely to reflect selection for the optimal stop codon, as TAA appears to be universally optimal, probably because it has the lowest read-through rate. Despite TAA being favoured by selection and mutation bias, as with codon usage bias GC pressure is the prime determinant of between-species TGA usage trends. In species with strong GC-biased gene conversion (gBGC), such as mammals and birds, the high usage and conservation of TGA is best explained by a AT→GC repair bias. How to explain TGA enrichment in other GC-rich genomes is less clear. Enigmatically, across bacterial and archaeal species and between human isochores TAG usage is mostly unresponsive to GC pressure. This unresponsiveness we dub the TAG paradox as currently no mutational, selective, or gBGC model provides a well-supported explanation. We suggest resolution of the TAG paradox may provide insights into either an unknown but common selective preference (probably at the DNA/RNA level) or an unrecognised complexity to the action of gBGC.

Key words: Stop codon usage, translation termination, translational read-through, stop codon read-through, molecular evolution, genome evolution

Significance statement: Between species and within genomes, codon usage is highly variable due to a complex interplay of evolutionary forces that include mutation bias, selection, and GC pressure. In this review, we consider the influence of each in determining the relative usage of the three stop codons (TAA, TGA, and TAG) for species across the tree of life. In doing so, we not only highlight the significant gaps in our understandings but demonstrate the utility of the stop codon exemplar for studying molecular evolution more generally.

Introduction

There has been extensive consideration of why, within coding sequence, one codon may be used more or less than an alternative codon specifying the same amino acid, this being a cornerstone of the selectionist/neutralist debate (Knight, et al. 2001). Analyses of synonymous codon usage biases have highlighted, amongst other things, the importance of the balance between mutation and selection and the role of translational dynamics in determining codon preferences (Andersson and Kurland 1990; Duret 2002; Chamary, et al. 2006; Hershberg and Petrov 2008; Plotkin and Kudla 2011). Indeed, in many species for each amino acid there exists an optimal codon that commonly reflects the most abundant iso-acceptor tRNA (Sharp and Li 1987; Bulmer 1991; Akashi and Schaeffer 1997; dos Reis, et al. 2004). This optimal codon is also typically enriched in the more highly expressed genes.

Most organisms also have three alternative options for the stop codon (UAA, UGA, and UAG in mRNA or TAA, TGA and TAG in genomic sequence). Like synonymous codons they too share the same “meaning” (Povolotskaya, et al. 2012; Belinky, et al. 2018). As amino-acylated tRNAs are not involved in stop codon recognition (for illustration of the process see Figure 1), it is less obvious why selection might prefer one stop codon over another. Nonetheless, we can ask a series of questions that parallel those asked of codon usage bias. What is the role of mutation bias and neutral, or nearly neutral, evolution in determining within and between species variation in stop codon usage? In any given species is there an optimal stop codon and, if so, why?

While the optimal sense codon for any synonymous group tends to vary between species as tRNA copy numbers vary (Duret 2002), we can also ask whether the same stop codon is optimal in all species. Many potential answers to these questions point to a role for forces that affect nucleotide content beyond the confines of stop codon usage. In this context, trends in TGA and TAG stop codon usage provide an unusual window into genome evolution as, given their identical functionality and nucleotide content, any differences in their usage requires explanation beyond a simple null model.

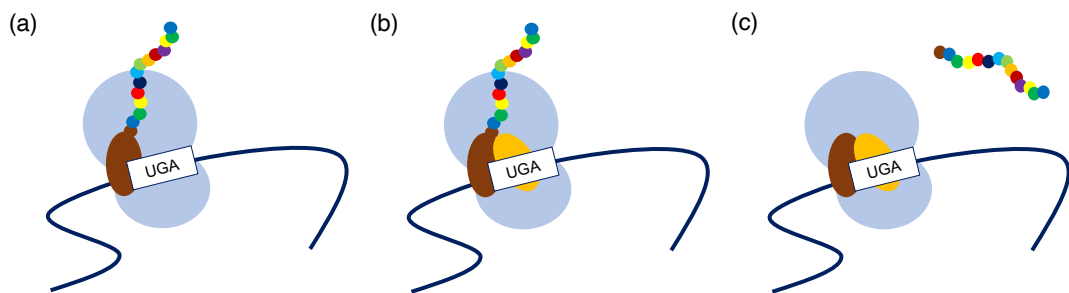


Figure 1. The basic mechanism of stop codon recognition by class I release factors. (a) The translating ribosome decodes coding sequence and recruits cognate amino-acylated tRNAs (brown) to build the growing polypeptide amino acid chain (small, coloured circles). (b) The stop codon (UGA in this example, but typically UAA, UGA or UAG) is recognised by, and becomes bound to, a class I release factor: RF1 or RF2 in bacterial, eRF1 in eukaryotic, or aRF1 in archaeal genomes (orange). (c) The binding between the release factor and stop codon begins a cascade leading to polypeptide release via the action of a class II release factor (not shown). Note that stop codons function in mRNA and hence genomic T (thymine) is replaced by U (uracil).

With direct determination of mutational profiles and extensive genome level analysis permitting analysis of both stop codon substitution rates and usage patterns, there has recently been some progress in understanding the determinants of stop codon usage. Here, we summarize recent advances in understanding the forces operating on stop codon usage emphasising what we do now understand but also the large (and profound) gaps in understandings. We consider two sorts of comparisons. One the one hand we have inter-species variation in usage trends where, with their extreme GC

contents, bacterial genomes are especially informative. On the other we make use of the human genome, where intragenomic extremes of GC content due to its isochores provide of similar utility. Analysis of the human genome is especially useful as we have well resolved parameters, such as the mutational profile and recent recombination rates, along with high quality expression data and ortholog description for closely related species. Intragenomic analysis also controls for possible mechanistic differences between taxa in stop codon recognition and release. In bacteria, for example, there are two class I release factors, RF1 and RF2, that are indispensable for stop codon recognition in all species with the standard genetic code, while in archaea and eukaryotes there is just one (Frolova, et al. 1994; Inagaki and Doolittle 2000; Jackson, et al. 2012; Kobayashi, et al. 2012; Rodnina 2018).

There are numerous issues related to the stop codons that we do not here investigate. For example, there exist species that do not use all three of TAA, TGA, and TAG to terminate translation, such as bacterial genomes decoded by translation table 4 (that don't use TGA) and some ciliates (e.g. *Paramecium tetraurelia* and *Stylonychia mytilis* use only TGA) (Alkalaeva and Mikhailova 2017). Why such species might not use the canonical three stop codons falls outside of our scope. It is also known that selection operates of stop codons outside of the canonical termination context. Additional in-frame stop codons (ASCs), for example, are under positive selection in some eukaryotes (but not bacteria) probably as an error-proofing mechanism to provide a second opportunity for translation to terminate should the primary stop codon be missed (Major, et al. 2002; Liang, et al. 2005; Adachi and Cavalcanti 2009; Korkmaz, et al. 2014; Ho and Hurst 2019). Similarly, out-of-frame stop codons (OSCs) are hypothesised to be selected to mitigate the consequences of frame-shift errors should the reading frame be disrupted (Seligmann and Pollock 2004; Abrahams and Hurst 2018). We do not broach the issues of non-canonical stop codon selection in this review.

The TAG problem: low absolute usage and unresponsiveness to GC pressure

When viewed across species, codon usage has a single strong predictor (Knight, et al. 2001) this being what we here call “GC pressure” so as to not to prejudge its cause. A diagnostic of this is a correlation between GC usage at codon third sites and some

other (hopefully independent) measure of GC content, such as GC of introns, intergene spacer etc. We can ask in turn whether stop codon usage is simply explained by GC pressure. If so, explaining stop codon usage may be simple problem: whatever explains GC pressure explains stop codon usage. If we consider the proportional usage the three stop codons in any given genome and ask how this varies between different bacteria with different GC content, then we see that TAA and TGA behave approximately as expected: TAA usage declines with increasing GC pressure while TGA increases (Fig 2). The enigma is the behaviour of TAG whose usage is both low (~20%) and unchanging with GC pressure, even though TGA has identical nucleotide content (Povolotskaya, et al. 2012; Korkmaz, et al. 2014; Ho and Hurst 2020).

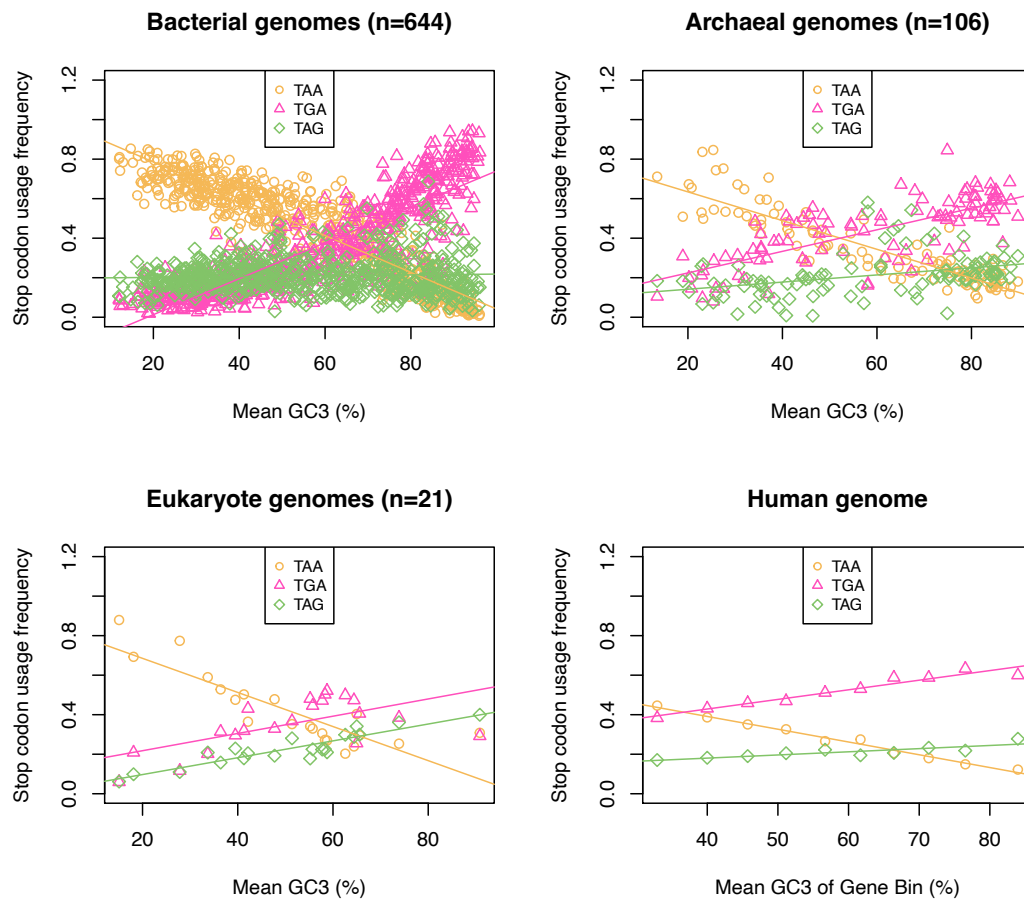


Figure 2. Stop codon usage (a) between 644 bacterial genomes, (b) 106 archaeal genomes, (c) 21 eukaryote genomes and (d) between human isochores. TAA usage is negatively correlated with GC3 content in all four analyses (Spearman’s rank; $p < 2.2 \times 10^{-16}$, $\rho = -0.92$ for bacteria, $\rho = -0.89$ for archaea, $\rho = -0.86$ for eukaryotes, $\rho = -0.99$ within the human genome). TGA usage is positively correlated with GC3

content in all four analyses (Spearman's rank; $p < 2.2 \times 10^{-16}$ for bacteria, archaea and within the human genome, $p = 0.011$ for eukaryotes, $\rho = 0.88$ for bacteria, $\rho = 0.76$ for archaea, $\rho = 0.55$ for eukaryotes, $\rho = 0.98$ within the human genome). TAG usage is uncorrelated with GC3 content in bacteria (Spearman's rank; $p = 0.48$, $\rho = -0.03$). TAG usage is positively correlated with GC3 content, but with lower absolute usage than TGA, in archaea (Spearman's rank; $p = 1.1 \times 10^{-7}$, $\rho = 0.49$), eukaryotes (Spearman's rank; $p = 9.5 \times 10^{-7}$, $\rho = 0.88$), and within the human genome (Spearman's rank; $p = 0.002$, $\rho = 0.88$). Figures adapted from (Ho and Hurst 2021b).

The TAG problem deepens when it is noticed that across archaea and between isochores in the human genome the same three trends in absolute stop codon usage are seen: TAA usage declining, TGA usage increasing and TAG usage either invariant or weakly responding to GC pressure. At first sight, archaea and bacteria look to be slightly different with TAG showing a small GC pressure response (weak positive slope of TAG predicted by GC pressure) in the former but not the latter. However, the bacterial data has more extreme values of GC content and allowing for this (by comparing GC-matched archaeal and bacterial samples) the trends seen in the two are all but identical (Ho and Hurst 2021b).

By considering changes in absolute stop codon usage, one assumes that there is no constraint preventing TAG usage from rising from $x\%$ at 0% GC content, to $x + y\%$ at 100% GC content, just as TGA goes from $z\%$ at 0% GC content to $z + y\%$ at 100% GC content (where $z > x$). By this logic, a car going from 0 to 20 miles per hour in a certain period of time has the same acceleration as a car going from 50 to 70 miles per hour - both would have identical slopes (y). If the mean value of TAG usage is lower than for TGA usage, however, the slopes might nonetheless be affected: a lower mean may be associated with a lower slope, just because the mean is low. An alternative approach that controls for this possibility is to ask about the proportional response, whereby each value is divided by the mean value for the relevant overall set of samples in question. Using this methodology, TAG and TGA usage still have very different slopes in bacteria (Fig 3) and in human 3' and 5' UTR sequences (Fig 4), and in both cases the distribution of TAG usage remains flat. However, in archaea and across human isochores at the focal termination site the slopes for TGA and TAG converge.

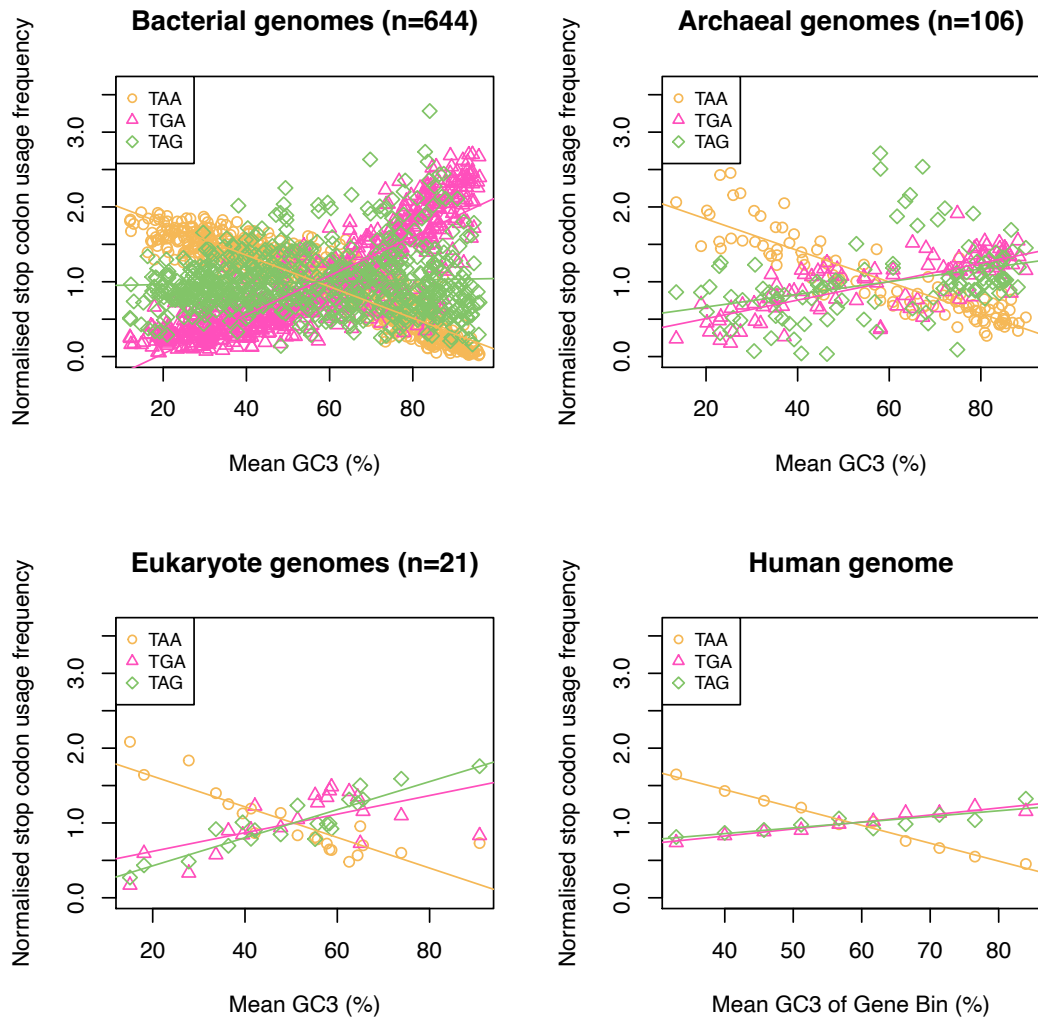


Figure 3. Stop codon usage normalised to the mean (a) between 644 bacterial genomes, (b) 106 archaeal genomes, (c) 21 eukaryote genomes and (d) between human isochores. Normalisation to the mean has no effect on the correlation statistics presented in Fig 2. Normalised TAA usage is negatively correlated with GC3 content in all four analyses (Spearman's rank; $p < 2.2 \times 10^{-16}$, $\rho = -0.92$ for bacteria, $\rho = -0.89$ for archaea, $\rho = -0.86$ for eukaryotes, $\rho = -0.99$ within the human genome). Normalised TGA usage is positively correlated with GC3 content in all four analyses (Spearman's rank; $p < 2.2 \times 10^{-16}$ for bacteria, archaea and within the human genome, $p = 0.011$ for eukaryotes, $\rho = 0.88$ for bacteria, $\rho = 0.76$ for archaea, $\rho = 0.55$ for eukaryotes, $\rho = 0.98$ within the human genome). Normalised TAG usage is uncorrelated with GC3 content in bacteria (Spearman's rank; $p = 0.48$, $\rho = -0.03$). TAG usage is positively correlated with GC3 content, but with lower absolute usage than TGA, in archaea (Spearman's rank; $p = 1.1 \times 10^{-7}$, $\rho = 0.49$), eukaryotes

(Spearman's rank; $p = 9.5 \times 10^{-7}$, $\rho = 0.88$), and within the human genome (Spearman's rank; $p = 0.002$, $\rho = 0.88$).

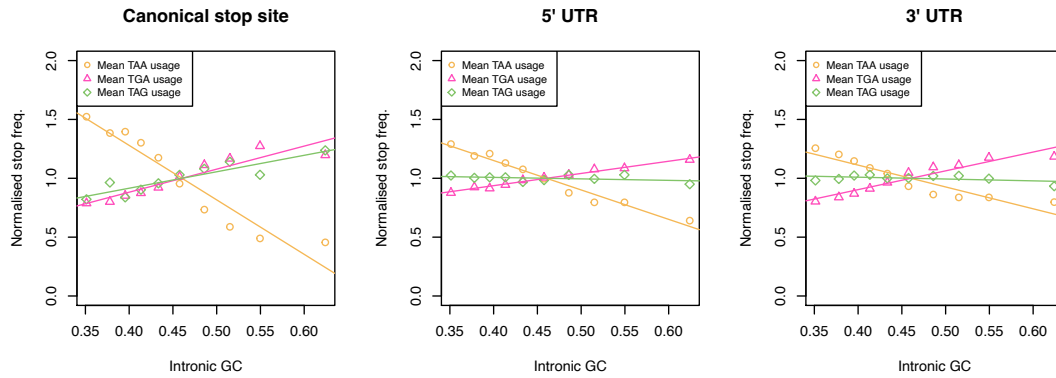


Figure 4. Stop codon frequencies (relative to the usage of all stops) normalised to the mean at the canonical stop site, in the 5' UTR, and in the 3' UTR at 10 equal-sized bins of various intronic GC contents in the genome. Normalised TAA frequency is negatively correlated with intronic GC content in all 3 sequences (Spearman's rank; all $p < 2.2 \times 10^{-16}$, $\rho = -0.99$ at the canonical stop site and in 5' UTR sequences, $\rho = -1$ in 3' UTR sequences). TGA is positively correlated with intronic GC content in all 3 sequences (Spearman's rank; all $p < 2.2 \times 10^{-16}$, $\rho = 0.99$ at the canonical stop site and in 5' UTR sequences, $\rho = 1$ in 3' UTR sequences). TAG usage is positively correlated with intronic GC content at the canonical stop site (Spearman's rank; $p = 0.0014$, $\rho = 0.89$) but is uncorrelated with intronic GC content in both 5' (Spearman's rank; $p = 0.61$, $\rho = 0.19$) and 3' UTR sequences (Spearman's rank; $p = 0.10$, $\rho = 0.55$). Figure adapted from Ho and Hurst (2021a).

It's not immediately obvious which method is most appropriate. For the proportional analysis, there is also a problem. If y were constant – as would be the case if TAG usage went from 20% to 30% and TGA went from 50% to 60% for example – we would conclude that there is instead a TGA problem as TGA would now have the less positive slope, despite the slopes of absolute usage being identical (to refer back to the car analogy, the proportional acceleration of the faster car is lower). Indeed, in such an example where the absolute usage slopes were identical, but the means different, a claim of different slopes might rightly be considered highly questionable. As we are concerned both with the avoidance of TAG and its lack of response to GC pressure,

the absolute response may be the more defensible as the proportional response assumes the under-usage of TAG to be a given, which it need not be.

We therefore conclude that the TAG problem may then be more broadly defined as the enigmatic difference in slope between absolute TGA usage versus GC on the one hand and absolute TAG usage versus GC on the other, the latter being curiously much shallower. That it is seen in three independent contexts adds to the problem. That it is not replicated in analysis across eukaryotes only adds to the perplexity.

Stop codon usage and release factor diversity: a genomic red herring?

A longstanding hypothesis to explain between-species stop codon usage in bacteria stems from the fact that bacterial translation termination at each of the three stop codons requires different molecular machinery. In bacteria, TAG is recognised uniquely by RF1 while TGA is recognised uniquely by RF2 and TAA is recognised either by RF1 or RF2 (Rodnina 2018). Early analysis observed (i) that TAA (with its broad RF binding potential) is the most common stop codon and (ii) that the TAG:TGA usage ratio positively correlated with the RF1:RF2 abundance ratio in a small sample of bacterial genomes, hence it was proposed that release factor abundance was a central driver of bacterial stop codon usage (Sharp and Bulmer 1988). Indeed, subsequent larger multi-species analyses have supported the correlations between RF1:RF2 and TAG:TGA and similarly assumed that RF abundance causes stop codon usage adjustment and not vice versa (Korkmaz, et al. 2014; Wei, et al. 2016). The notion that RF1:RF2 relative abundance determines stop codon usage bears obvious parallels with the idea that synonymous codon usage is determined by iso-acceptor differential tRNA abundance. As Wei, et al. (2016) identified that RF2 is exceptionally low when AT3 content is high across species, RF1:RF2 might also help explain the TAG problem. They argue that in GC poor regions or genomes, mutation bias favours the TAA, the most AT-rich stop codon, over both TAG and TGA. At mid-to-high GC contents, TGA is preferred over TAG as RF2 expression levels become increasingly dominant over that of RF1.

We have since challenged this interpretation of the RF1:RF2 correlation with TAG:TGA, asking why the RF1:RF2 ratio shouldn't instead be moulded to the stop

codon requirements of the genome (Ho and Hurst 2021b). First, we noted that in humans and archaea there is only one release factor. That we see the same TAG problem between human isochores and across archaea (Fig 2) thus indicates that some other forces can give the TAG anomaly. Second, if RF abundance were to cause stop codon usage variation, one might predict that between-species stop codon trends in bacteria (particularly the disconnect between TGA and TAG usage) should not be repeated in non-canonical stop codon contexts where RF recognition is non-important. We however found the relative usage of TGA, TAA and TAG in sequence immediately 3' of genes have the same trends as seen at the canonical stop context (Ho and Hurst 2021b). This is unlikely to be owing to selection for additional stop codons in the 3' non-coding sequence for two reasons. First, while there is evidence of selection for additional 3' in-frame stop codons (ASCs) in some single celled eukaryotes (Ho and Hurst 2019), the same is not seen in multicellular eukaryotes or bacteria (Major, et al. 2002; Korkmaz, et al. 2014; Ho and Hurst 2019). Second, the same trend is also seen if we examine sequence post the first in frame stop codon (Ho and Hurst 2021b). All the above points of evidence strongly suggest that we need to evoke some force other than RF diversity to explain trends in usage of TAA, TGA and TAG.

A further corollary of the above evidence is that the better explanation for the RF1:RF2 correlation with stop codon usage is that RF abundance adapts to stop codon usage and not vice versa. This direction of the causal arrow is parsimonious for several reasons. As we outline in Ho and Hurst (2021b), the moulding of stop codon usage (particularly TGA \leftrightarrow TAG) to respond to the RF environment doesn't make clear evolutionary sense. As the RF hypothesis itself states that TAA is optimal due to its dual recognition by RF1 and RF2 (of which more below), there is no selective need to switch from TAA \rightarrow TGA or TAA \rightarrow TAG. TGA and TAG usage adjustment to match RF1:RF2 hence must theoretically proceed via TGA \leftrightarrow TAG exchanges that require a minimum of two mutational events. This is significant as, as any genome wide shift in usage must involve one step that is opposed by selection. We presume that TAG \rightarrow TGA and TGA \rightarrow TAG cannot occur in one mutational step. Assuming too conservation of stop codon identity then there must be TAA \rightarrow TGA or TAA \rightarrow TAG either of which is deleterious. Under the RF hypothesis, then, it is unclear why selection should favour TGA \leftrightarrow TAG in any scenario.

With the RF hypothesis seemingly unparsimonious to explain between-species within bacteria, and irrelevant when we consider eukaryotes and archaea (which possess just one RF), arguments for stop codon usage trends being driven by RF diversity appear to be a red herring (i.e. a distraction from the main explanation). For the rest of this review, we consider the myriad of factors that likely act to shape the stop codon usage of all species. We consider the roles of mutation bias, selection, and biased gene conversion, discussing how these too might vary between species.

Null mutational models cannot alone explain within- or between-genome variation in stop codon usage

For stop codon usage, as with synonymous codon usage bias, the simplest null would be one of neutral evolution coupled to mutation bias. Originally the variation in GC content between species was indeed assumed to reflect the mutational process, assuming GC content at third site to be approximately neutral and reflective of mutational biases (Knight, et al. 2001). However, now that we can measure mutational biases directly the assumption that GC-rich genomes and genomic regions (such as the GC-rich isochores in humans) are a consequence of mutational bias alone is no longer defensible.

From analysis of mutations seen in parent offspring sequencing, MA lines or rare SNPs, across both eukaryotes and prokaryotes mutation bias appears to be very commonly, if not universally, GC→AT biased (Smith and Eyre-Walker 2001; Lynch, et al. 2008; Hershberg and Petrov 2010; Hildebrand, et al. 2010; Long, et al. 2018). Importantly this universality applies just as much to GC rich genomes as to AT rich ones (Hershberg and Petrov 2010; Hildebrand, et al. 2010). Hence GC rich genomes sit far away from their AT biased mutational equilibrium (Long, et al. 2018). Stop codon usage in part reflects this deviation from mutational null. We would expect under a mutation bias null for TGA and TAG to be universally and equally rare and TAA to be the most abundant. However, the simplest null mutational model fails to explain either within- or between-genome variation. Notably TAA usage, while indeed high in GC-poor bacterial genomes, is low in GC-rich ones (Fig 2) despite the

profile of mutation bias being consistently GC→AT biased (Hershberg and Petrov 2010; Hildebrand, et al. 2010).

Perhaps the best current data come from humans as here, by parent offspring sequencing, we have an exceptional view of the mutational process. We can then for example ask whether in GC rich isochores (with an abundance of TGA) the mutational profile is more AT→GC biased than in the GC poor isochores. Strikingly, a GC→AT mutation bias is approximately invariant to isochore GC content (Smith, et al. 2018; Ho and Hurst 2021a) (Fig 5), but nonetheless TGA usage increases with local GC with TAA decreases (as seen in Fig 2). In this case more complex mutational biases (e.g. high rates of CpG→TpG (Duncan and Miller 1980; Sved and Bird 1990; Roberts and Gordenin 2014) which could generate new TGA stop codons) also cannot account for the decline in TAA usage and increase in TGA usage as local GC content increases (Ho and Hurst 2021a).

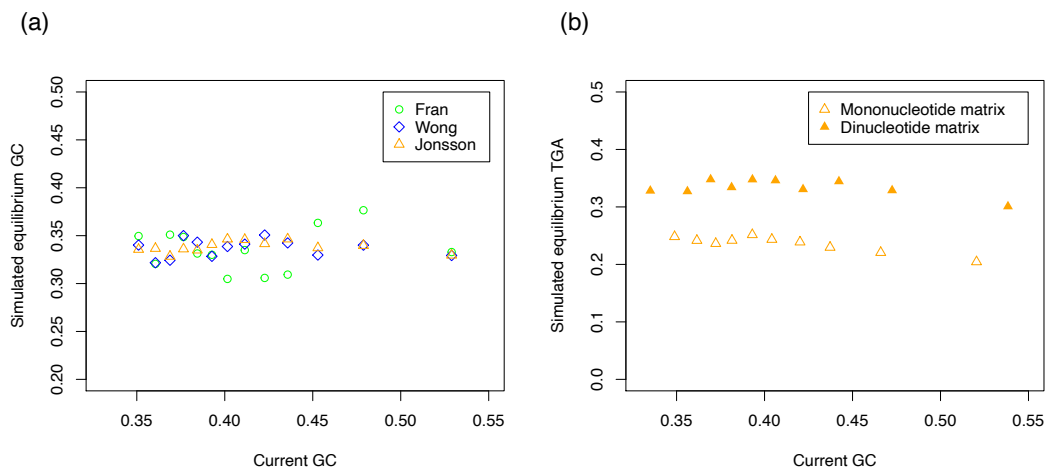


Figure 5. Simulated equilibrium (a) GC content and (b) TGA usage plotted against the current GC content of the windows from which the mutation spectrum was estimated. Panel A is reproduced with permission from Smith, et al. (2018) (the original figure is available open access at: <https://doi.org/10.1371/journal.pgen.1007254.g004>) and shows equilibrium GC estimates from three sources of human *de novo* mutations. Panel B is reproduced with permission from Ho and Hurst (2021a) and illustrates equilibrium TGA usage (relative to TAA and TAG usage) estimated from the Jonsson, et al. (2017) dataset of human *de novo* mutations.

Analysis of the human mutational profile also indicates that trinucleotide frequencies are closer to mutational equilibrium in AT-rich isochores than GC-rich ones (Ho and Hurst 2021a). At AT-rich isochores we can compare equilibrium estimates of TAA, TGA and TAG trinucleotides to their relative usage at the canonical stop site to assess how well the mutational profile predicts what is seen in termination contexts. Deviation at the stop site from the predicted relative frequencies of TAA, TGA, and TAG trinucleotides would indicate the presence of non-mutational forces influencing stop codon usage. Using the same human dinucleotide mutational matrix as in Ho and Hurst (2021a), we estimate the equilibrium relative usage of TAA to be 43.0%, TGA to be 32.5%, and TAG to be 24.5% in the bottom 20% of human genes by GC content. Despite their shared nucleotide content, mutational preferences at the dinucleotide level appears to somewhat discriminate between TGA and TAG, perhaps because CpG→TpG mutations are very common (Duncan and Miller 1980; Sved and Bird 1990; Roberts and Gordenin 2014), and hence could begin to explain the absolute differences in their usage. However, in the same set of AT-rich genes, the stop codon usage at the canonical stop site is 38.4% TAA, 42.0% TGA, and 19.5% TAG. This in turn suggests that in AT rich domains usages are reasonably close to, but distinct from, mutational expectations.

While in AT-rich regions complex mutation bias takes us some way to understanding the lower usage of TAG, mutation bias fails to explain the differing response of TAG and TGA to GC pressure as mutation bias does not covary with GC content mutation. Consequently, the usage of all three stops is far from mutational equilibrium in GC-rich isochores. Using the same dataset, we estimate the equilibrium relative usage of TAA to be 42.6% (compared to an observed usage of 13.6%), TGA to be 32.0% (compared to 63.5%), and TAG to be 25.4% (compared to 22.9%) in the top 20% of human genes by GC content. Coupled with evidence for universality of a GC→AT mutation bias (Smith and Eyre-Walker 2001; Lynch, et al. 2008; Hershberg and Petrov 2010; Hildebrand, et al. 2010; Long, et al. 2018), mutation bias provides no robust explanation of the TAG problem or indeed why organisms differ in the GC content more generally.

The three stop codons are not selectively equivalent

(i) Genomic evidence unanimously supports TAA optimality

Given that a mutational neutral null appears to be insufficient in explaining within- or between-genome variation in stop codon usage, as with synonymous codon usage bias one might suspect that selection has some role in stop codon usage. Several approaches have been taken to determine which stop codon might be optimal. To a first approximation they all concur that TAA is universally optimal.

The first method considers differential usage in highly expressed genes versus lowly expressed genes. This assumes that the costs of translational error are higher in highly expressed genes (see trends in synonymous codon usage). Across bacteria and in the human genome, TAA is relatively enriched stop codon in highly expressed genes suggesting a selective advantage (Korkmaz, et al. 2014; Trotta 2016; Ho and Hurst 2020).

The second method considers enrichment allowing for biases in the usage of dinucleotides within any given genome (note this is observed usage not the mutational profile). Against dinucleotide-controlled null models, it is TAA (and not TGA nor TAG) that is most enriched across bacteria, eukaryotes, and archaea (Ho and Hurst 2021b). The third method consider trends in enrichment compared to nucleotide null as a function of effective population size (N_e), assuming that when N_e is high selection is more efficient and thus enables organisms to be closer to a selectively optimal state (Ohta 1992; Lynch 2007). Such methods come with all the necessary caveats that N_e is hard to estimate (but with mutation rate and polymorphism data, it is now possible). To date this has been done across eukaryotes in a phylogenetically controlled manner with, TAA enrichment correlating positively with N_e (Ho and Hurst 2020).

A final method considers trends in stop codon substitution (i.e. fixation events) using species trios. Such a method can detect differences in relative substitution rates (e.g. TGA→TAA per TGA versus TAA→TGA per TAA) and so infer the conserved state (Belinky, et al. 2018). Note that this is not the same as the mutational analysis as that considers just the rates of origination not the rates of origination and fixation. This method finds TAA conservation near universally (Belinky, et al. 2018). However, a

problem we return to below, is that this method, also reports TGA conservation in mammals (Belinky, et al. 2018).

Almost all methods hence concur on universal TAA optimality. Why then might stop codons have different fitness consequences and is there evidence that any such effects mediate within- or between-genome variation? The dominant models for TAA optimality all point to its resilience in the face of errors as the probable cause, the errors in question being either mutational, mistranscriptional or owing to misreading/misprocessing.

(ii) TAA is more robust to mutation and mistranscription events than TGA and TAG

Perhaps the most immediately noticeable difference between the three stop codons is the differing nucleotide compositions of TAA, TGA, and TAG. This is significant as any selective force that moulds stop codon usage must commonly proceed via stop codon switch events, i.e. TAA \leftrightarrow TGA, TAA \leftrightarrow TAG, and TGA \leftrightarrow TAG, as sense codon intermediate states at the canonical termination site are unlikely to be tolerated. TAA is unique in being robust to two mistranscription (or mutational) events, i.e. TAA \rightarrow TGA or TAA \rightarrow TAG (Ho and Hurst 2020). TGA \rightarrow TAA and TAG \rightarrow TAA switches are similarly resilient to mistranscription or mutation, but TGA \leftrightarrow TAG is not (apart from double mutants which are extremely rare). Not only might TAA be optimal for this reason, but TGA \leftrightarrow TAG switches must proceed via TAA regardless of whether TGA or TAG is optimal. Mutation is probably too rare a process to select for TAA via mutational robustness, however whether much more common mistranscription events could select for TAA is unresolved – see for example the rates in *Escherichia coli* (Lee, et al. 2012; Traverse and Ochman 2016; Meer, et al. 2020) and *Caenorhabditis elegans* (Denver, et al. 2004; Denver, et al. 2009; Gout, et al. 2013; Meer, et al. 2020).

(iii) TAA is the least, while TGA is the most, prone to molecular errors

The selective hypothesis that has garnered the most attention is that stop codons differ in their susceptibility to mistakes during gene expression. With stop codons, the most associated such molecular error is the failure to terminate translation, known as either

translational read-through (TR) or stop codon read-through (SCR). Here we will refer to this phenomenon as TR. When TR occurs the stop codon is missed by the translational machinery, typically due to erroneous misreading of the stop codon by a near-cognate tRNA, leading to unintended translation of the 3' UTR that continues until the next in-frame stop codon or the polyadenylation signal (Fig 6).

TR is most often deleterious for several reasons. At the very least, C-terminal extension is an unnecessary energetic waste (Wagner 2005) and, in more severe cases, might lead to problems with localisation and export (Falini, et al. 2005; Hollingsworth and Gross 2013), aggregation (Vidal, et al. 1999; Vidal, et al. 2000), or stability (Clegg, et al. 1971; Namy, et al. 2002; Pang, et al. 2002) of the final protein product. Should translation reach the polyA⁺ tail, TR can also trigger degradation of both mRNA and protein (Dimitrova, et al. 2009; Klauer and van Hoof 2012). To mitigate these consequences, we expect selection to reduce TR rates. In order of decreasing TR susceptibility, the order appears to be TGA>>TAG>TAA across bacterial (Roth 1970; Strigini and Brickman 1973; Ryden and Isaksson 1984; Parker 1989; Meng, et al. 1995; Sanchez, et al. 1998) and eukaryotic (Geller and Rich 1980; Parker 1989; Cridge, et al. 2018) species. Stop codon switches that lower the TR error rate (TGA→TAG, TGA→TAA, TAG→TAA) could hence be favoured by selection across taxa.

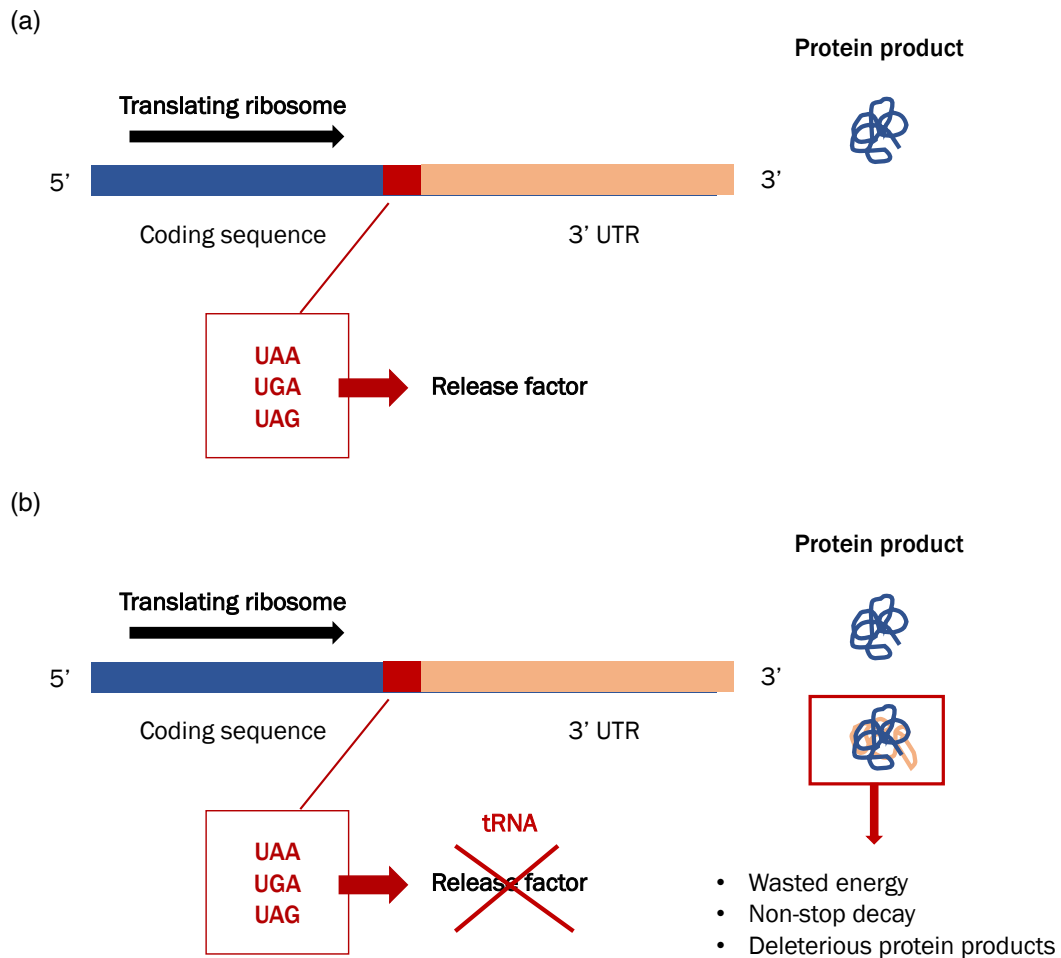


Figure 6. The mechanistic basis of translational read-through. (a) Canonical termination occurs when the stop codon is recognised by its cognate release factor. Only coding sequence is translated to build the polypeptide amino acid chain. (b) Translational read-through occurs when the stop codon is missed by the termination machinery, often due to the erroneous binding of a near-cognate tRNA to the stop codon (Roy, et al. 2015; Beznoskova, et al. 2016). This results in the translation of 3' UTR sequence until the next in-frame stop codon or until the ribosome reaches the polyA⁺ tail, triggering non-stop decay.

That TAA is the least error-prone stop codon variant makes it a strong candidate for optimality. But can we be confident that TR rather than possible other advantages of TAA (such as mistranscriptional robustness) is the core to its selective optimality? A strong clue comes from nucleotide preferences immediately proximal to the stop codon. The sequence involved in modulating TR rate likely extends for at least 6

nucleotides downstream of the stop codon for fine tuning of ribosomal interactions (Bossi and Roth 1980; Namy, et al. 2001; Wei and Xia 2017; Cridge, et al. 2018) and here we find TR-associated nucleotide combinations to be rare in highly expressed genes (Ho and Hurst 2020). It is the +4 nucleotide, however, that is most influential. It is important, therefore, that genes terminating with the most TR-prone context TGAC (Cridge, et al. 2018) are underrepresented in eukaryotic genomes (Cridge, et al. 2006). A second clue comes from sequence conservation. If TR truly does generally result from error, there is no reason why the sequence downstream of the stop codon (and before the first in-frame ASC) should be conserved. Li and Zhang (2019) tested this hypothesis in *Drosophila* and yeast by defining and comparing two regions: region 1 being the sequence between the canonical stop and first ASC, and region 2 being the sequence between the first ASC and second ASC that should be untranslated except for rare events of double TR. In both organisms, they find no evidence to support region 1 sequences being more conserved than region 2 sequences (Li and Zhang 2019).

For the above reasons TAA is thought to be the optimal stop codon for all species for its low relative TR rate. However, one piece of evidence is, in this context, unresolved. In yeast there exist selectively preferred additional stop codons (ASCs) in 3' UTR, enriched at codon site +3 downstream of the canonical termination codon (Liang, et al. 2005). This suggests that read-through happens and selects for a second stop codon. Curiously the conserved second stop codons are enriched for genes terminating TAA (Liang, et al. 2005). *A priori*, TAA is expected to have the lowest read through rates and hence not expected to be associated with conserved additional stop codons. One possible explanation is that these extra stop codons reflect increased read-through following prion upregulation that forces read-through (Wickner, et al. 1995; Liebman and Chernoff 2012). If such read-through were particular to TAA then the circle could be squared. We note too that, while TAA optimality seems universal the mechanistic underpinning of this is not at all clear. As described above, in bacteria this was ascribed to TAA binding both RF1 and RF2 (Sharp and Bulmer 1988), but the universality suggests that this is an unnecessary model.

Just because TAA is generally optimal it does not follow that selection need favour TAA in all cases. There can be occasions when read-through might be employed as

part of a sophisticated mechanism that is favourable, not deleterious. The C-terminal extension of polypeptides by TR for example could theoretically add new signals and domains to proteins to be viewed by natural selection (Dunn, et al. 2013; Schueren and Thoms 2016). Situations such as these are known as functional read-through (FTR) and are described across the tree of life (see Schueren and Thoms (2016) for a thorough review). Perhaps the best example comes from viral genomes that use FTR to improve the coding capacity of their very small genomes (Firth and Brierley 2012). In tobacco mosaic virus, for example, TR of the TAG stop codon of the RNA replicase transcript allows the virus to yield two isoforms from one gene (Pelham 1978). In humans, a well described example of FTR allows a 22 amino acid extension to vascular endothelial growth factor A (VEGFA) to reverse its function from proangiogenic to antiangiogenic (Eswarappa, et al. 2014). The best studied metazoan with regards to FTR is the fruit fly, where ribosomal profiling has estimated ~300-350 candidates in *D. melanogaster* of which 8 were experimentally confirmed (Dunn, et al. 2013). The C-terminal extensions in these cases included transmembrane domains, nuclear localisation signals, a PTS1, and a prenylation signal (Dunn, et al. 2013).

If TR were to be commonly functional, however, one might expect TGA stop codons (the universally leakiest stop) to be selectively preferred. To date there is no evidence to suggest FTR is particularly common in complex organisms, perhaps because FTR is rather unnecessary in larger genomes which are not so constrained in their coding capacity (Schueren and Thoms 2016). Indeed, *in silico* analysis of the stop codon context of 200,000 human transcripts returned only 57 TR candidates (Schueren, et al. 2014). This is not what is expected were TR to be regularly beneficial. Even in *Drosophila*, with its ~300 candidate transcripts for FTR (Dunn, et al. 2013), such numbers are orders of magnitude below what is needed to support genome-wide selection for stop codons that promote TR.

There is a second, if speculative, model that proposes that TR is beneficial for reasons beyond extending protein ends. This states that low-level, but consistent, TR is required for gene regulation and mRNA quality control by controlling ribosomal queuing (Yordanova, et al. 2018). Under this model, translating ribosomes read past the stop site and eventually stall, translation being inhibited when the ribosome queue backs up to the stop codon (Yordanova, et al. 2018). The rate of TR coupled with the

length of sequence to the ribosomal stall site hence might define the number of times the mRNA can be translated. There is evidence for this at the AMD1 locus in humans (Yordanova, et al. 2018), however it remains unknown how widespread a mechanism like this might be. If it is common, it could theoretically affect stop codon usage due to their different TR rates, and hence different ribosomal queuing rates, which could lead to the fine tuning of TAA, TGA and TAG frequencies. It is for this reason that (Seoighe, et al. 2020) consider this model as a potential explanation for the apparent conservation of TGA stop codons in mammals. We return to the issue of mammalian exceptionalism later.

GC-biased gene conversion acts antagonistically to selection and mutation bias to promote TGA usage

Perhaps the most striking conclusion of the above is that, while we can discern TAA optimality, the effects of TAA optimality appear modest: despite TAA optimality its usage at the canonical termination position declines with GC pressure and TAA optimality appears to have little relevance to the TAG problem. Similarly, the enrichment of TAA in highly expressed genes is modest compared with GC pressure. We can detect TAA enrichment across species as a function of N_e but again the effect is quite modest (Ho and Hurst 2020). Perhaps this is most in evidence when comparing TAA abundance across species/isochores to TAA usage in locations when it cannot be employed as a stop codon, the trends being almost identical to those where it does (Ho and Hurst 2019; Ho and Hurst 2021b). This all suggests that TAA optimality is a sideshow (or the icing on a cake) to full understanding of trends in stop codon usage. It also questions what the other forces might be operating that could explain the trends in TAA, TGA and TAG usage.

If TAA truly is universally optimal then there might be lessons to be learned in apparently contradictory examples. In mammals, TGA stop codons are not only high in frequency but appear to be highly conserved, even more so than TAG and, surprisingly, TAA (Belinky, et al. 2018; Seoighe, et al. 2020). Interrogation of stop codon usage and substitution rates has revealed this phenomenon may be primarily driven by highly compartmentalised TAA to TGA bias in domains of high GC content (Ho and Hurst 2021a) (Fig 2). This is particularly interesting given the spatially

structured “isochore” nature of base composition in mammalian genomes (Bernardi 1993; Eyre-Walker and Hurst 2001). The current best explanation for isochore structure is GC-biased gene conversion (gBGC), a process through which mismatches during heteroduplex formation in meiotic recombination are resolved in a GC-favoured manner (Galtier, et al. 2001; Duret and Galtier 2009). As gBGC is tightly coupled to recombination, GC-rich alleles receive the greatest fixation advantage in highly recombining sequences, possibly even when deleterious (Galtier, et al. 2009).

Could mammalian TGA (a GC-rich stop codon) usage and compartmentalised TAA→TGA substitution bias be explained by gBGC? Several pieces of evidence are supportive such as the observation that autosomal size, which correlates negatively with recombination rate and GC content, predicts high TGA usage in smaller, more recombinogenic chromosomes (Ho and Hurst 2021a). TAA→TGA substitution rate also correlates positively with local recombination rate assayed from parent-offspring trios (Ho and Hurst 2021a). Covariance between TAA→TGA substitution rate and GC content also appears somewhat unique to isochore-structured genomes (including birds) consistent with the possibility that they share the same underlying forces (Ho and Hurst 2021a). Indeed, birds and mammals are unique in being known to have a strong AT→GC conversion bias that accords with the gBGC model. In humans for example ~70% of GC:AT mismatches are resolved in favour of the GC residue (Halldorsson, et al. 2016). The gBGC model has no problem explaining why stop codon trends are seen both at the focal stop and in non-coding sequences as it does not depend on termination functionality.

Given that selective and mutational hypotheses for TGA conservation are unparsimonious, for mammals at least it hence appears that gBGC offers the best explanation for TGA conservation and its focus in high GC isochores. We note that this is an unusual case history as TGA is unfavoured by the mutation bias (GC→AT) and selection (most probably for TAA and reduced TR). Consequently, there is only one currently known force that can explain TGA enrichment at the focal stop codon in GC rich domains, this being gBGC. As TGA possesses a higher intrinsic TR error rate than TAA (Cridge, et al. 2018), gBGC appears also to be fixing deleterious mutations.

While gBGC could potentially explain TGA conservation in mammals, what to expect from the gBGC outside of this example is unclear. Is gBGC universal throughout the tree of life? In yeast the best evidence from tetrad sequencing suggests a very weak bias at best possibly even in the opposite direction (Mancera, et al. 2008; Duret and Galtier 2009; Liu, et al. 2018; Liu, et al. 2019). While in humans and birds the bias, per event, is ~60-70%, in yeast the current best estimate is a bias of 50.03, not significantly different from the null of 50% (Liu, et al. 2017). What about bacteria? Could gBGC explain trends seen between-genomes as well as within? Whether gBGC operates in bacteria remains an open issue (Lassalle, et al. 2015), and further work investigating complex gBGC preferences in these groups is needed. Arguing against gBGC is the finding that GC rich bacterial genomes reside above mutation equilibrium even if not recombining (Hildebrand, et al. 2010).

Unravelling the TAG problem: a window into complex k mer trends?

The sequential consideration of mutation bias, selection, and GC pressure in determining stop codon usage primarily focuses on TAA and TGA stop codons. The omission of TAG reflects its non-typical behaviour in response to GC pressure. Any mutational or simple fixation bias (be it gBGC or selection for higher GC content) predicts that TGA and TAG stop codons should be handled the same due to their identical nucleotide content. Across bacterial and archaeal taxa and across isochores within genomes this is not seen, TGA reliably correlates with GC content positively while TAG is underused and unresponsive to GC pressure. (Fig 2) (Korkmaz, et al. 2014; Trotta 2016; Ho and Hurst 2019): How may we attempt to resolve this?

From a mutational perspective we may utilise data from human family trios for scrutiny of more complex mutational profiles (as above). Analysis of a mutational matrix of such *de novo* mutations facilitates the calculation of mutational equilibrium frequencies for any given nucleotide, dinucleotide or trinucleotide which can then be compared against fixed frequencies to elucidate deviations from mutational null. Equilibrium TAG content in humans is indeed lower than TGA content suggesting a more complex mutational bias at least partially explains for its low usage (Ho and Hurst 2021a). Strikingly, however, TAG usage in GC-poor isochores closely resembles equilibrium whereas this is not true in GC-rich domains (Ho and Hurst

2021a). Some kind of fixation bias needs to be evoked. As TAG and TGA have the same mononucleotide content we seem to be left having to evoke, non-monomucleotide (e.g. dinucleotide or trinucleotide or larger) level selection or an added layer of complexity to gBGC that goes beyond a simple AT→GC conversion bias.

In our recent study we investigated the nature of the fixation bias by assigning a fixation bias “boost” score to each trinucleotide based upon the difference between its observed frequency and the predicted mutational equilibrium (derived from a dinucleotide mutational matrix) in GC-rich domains (Ho and Hurst 2021a). We found that TGA consistently receives a higher fixation boost than TAG. Indeed, trinucleotides may be grouped by their GC content such that completely AT-rich trinucleotides such as AAA may be assigned to the 0% GC group, AGA may be assigned to the 33% group, etc. We found that the order of trinucleotides by “GC boost” is highly consistent within each GC class (0%, 33%, 66%, 100%) across different classes of non-coding sequence (Ho and Hurst 2021a). Notably, within the 33% GC content class (trinucleotides with two As or Ts and 1 G or C), fixed TGA frequencies are seen far above its mutational equilibrium in 3' UTR, 5' UTR, introns, enhancers etc, while TAG and TAC are always less affected by whatever fixation bias is at play (Ho and Hurst 2021a). These results support the possibility of a consistent trend for a fixation bias, at least in humans, that can only be evidenced at higher resolution than mononucleotide level.

What might cause such a complex fixation bias? One possibility is some even more complex set of context dependencies of mutational biases not so far considered. However, our dinucleotide model of expected frequencies in domains of low GC very accurately predicts observed frequencies from mutation bias alone (Ho and Hurst 2021a), so this seems unlikely. As regards selection, many possibilities are imaginable but to date none seem particularly compelling. There may for example exist selection against TAG's component dinucleotides (Ho and Hurst 2021a). TA, for example, might be avoided to avoid transcription initiation sites (TATA in eukaryotes and “Pribnow” boxes in prokaryotes). Were this to be important, however, one might expect to see similar selection against, and low abundance of, TAA stop codons too. Other ideas have included more general DNA structural hypotheses such as TA being

avoided to protect chromatin structure as AT-rich DNA is concentrated in nucleosome free regions (Burge, et al. 1992). This, however, cannot explain the stop codon usage trends being the same in bacteria as seen across the human genome as the former don't possess nucleosomes but do avoid TAG. Indeed, while any selective hypothesis must fit many different species (prokaryotic and eukaryotic) it must also involve selective coefficients that are strong enough to explain all trends.

For gBGC to explain the TAG enigma requires significant amendments to the current assumptions. One speculative possibility is that gBGC is better at recognising mismatches at certain residues than others or that the form of the bias is dependent on the *k*-mer context. In the former model, an unrecognised mismatch is resolved in mitosis but with no bias. In the second model, all meiotic mismatches are recognised but the bias differs. Either way, it is possible that net TNA conversion bias would be different from that for TAN. If so gBGC could potentially fix more TAA to TGA mutations than TAA to TAG, for example. While promising, for complex gBGC to explain stop codon trends across taxa more generally this order of trinucleotides would have to be consistent between all organisms showing the TAG problem. Scrutiny of the across eukaryote trends (and the potential lack of TAG problem) may be a means to progress as gBGC seems to be variable in effect across eukaryotes. Examination of the context of gBGC events through tetrad sequencing or sperm typing is a high priority.

All things considered there appears to be something profound about genome evolution that we do not currently understand. From analysis of the low TAG usage at the canonical stop we have identified the TAG problem, a more general low usage and one non-responsive to GC pressure in many comparisons (bacteria, archaea, human, at the focal stop and elsewhere). As we don't have a coherent explanation for this, we suggest that this be considered the TAG paradox. Unravelling the TAG paradox, given its appearance throughout the tree of life, may well provide a window into a previously unrecognised world of unexplained trends in *k*-mer usage that, we suggest, must throw light onto currently not well understood forces behind stop codon usage and genome evolution more generally.

Acknowledgements

We thank Adam Eyre-Walker and colleagues for sharing their GC equilibrium data and providing permission to replot these. This work was supported by the European Research Council (grant EvoGenMed ERC-2014-ADG 669207 to L.D.H.).

References

Abrahams L, Hurst LD. 2018. Refining the ambush hypothesis: Evidence that GC- and AT-rich bacteria employ different frameshift defence strategies. *Genome Biol. Evol.* 10(4): 1153-1173.

Adachi M, Cavalcanti AR. 2009. Tandem stop codons in ciliates that reassign stop codons. *J. Mol. Evol.* 68(4): 424-431.

Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics.* 146(1): 295-307.

Alkalaeva E, Mikhailova T. 2017. Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. *Bioessays.* 39(3): 1600213.

Andersson SGE, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol. Rev.* 54(2): 198-210.

Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci. Rep.* 8(1): 9260.

Bernardi G. 1993. The isochore organization of the human genome and its evolutionary history - a review. *Gene.* 135(1-2): 57-66.

Beznoskova P, Gunisova S, Valasek LS. 2016. Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. *RNA.* 22(3): 456-466.

Bossi L, Roth JR. 1980. The influence of codon context on genetic-code translation. *Nature.* 286(5769): 123-127.

- Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*. 129(3): 897-907.
- Burge C, Campbell AM, Karlin S. 1992. Over-representation and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* 89(4): 1358-1362.
- Chamary J-V, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7(2): 98-108.
- Clegg JB, Weatherall DJ, Milner PF. 1971. Haemoglobin constant spring - a chain termination mutant? *Nature*. 234(5328): 337-340.
- Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 46(4): 1927-1944.
- Cridge AG, Major LL, Mahagaonkar AA, Poole ES, Isaksson LA, Tate WP. 2006. Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. *Nucleic Acids Res.* 34(7): 1959-1973.
- Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledo JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, et al. 2009. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc. Natl. Acad. Sci. U.S.A.* 106(38): 16310-16314.
- Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature*. 430(7000): 679-682.
- Dimitrova LN, Kuroha K, Tatematsu T, Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* 284(16): 10343-10352.
- dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32(17): 5036-5044.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature*. 287(5782): 560-561.

Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Elife*. 2(1): e01179.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12(6): 640-649.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genom. Hum. Genet.* 10(1): 285-311.

Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, Wiliard B, Graham LM, DiCorieto PE, Fox PL. 2014. Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell*. 157(7): 1605-1618.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat. Rev. Genet.* 2(7): 549-555.

Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L, La Starza R, Diverio D, Colombo E, Santucci A, et al. 2005. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *New Engl. J. Med.* 352(3): 254-266.

Firth AE, Brierley I. 2012. Non-canonical translation in RNA viruses. *J. Gen. Virol.* 93(7): 1385-1409.

Frolova L, Legoff X, Rasmussen HH, Cheperegin S, Drugeon G, Kress M, Arman I, Haenni AL, Celis JE, Philippe M, et al. 1994. A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. *Nature*. 372(6507): 701-703.

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

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

- Geller AI, Rich A. 1980. A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature*. 283(5742): 41-46.
- Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. 2013. Large-scale detection of in vivo transcription errors. *Proc. Natl. Acad. Sci. U.S.A.* 110(46): 18584-18589.
- Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir A, Jonasdottir A, Sulem P, et al. 2016. The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* 48(11): 1377-1384.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9): e1001115.
- Hershberg R, Petrov DA. 2008. Selection on codon bias. In. *Annu. Rev. Genet.* p. 287-299.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9): e1001107.
- Ho AT, Hurst LD. 2020. Effective population size predicts local rates but not local mitigation of read-through errors in eukaryotic genes. *Mol. Biol. Evol.* 38(1): 244–262.
- Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. *PLoS Genet.* 15(9): e1008386.
- Ho AT, Hurst LD. 2021a. Sequence conservation need not imply purifying selection: evidence from mammalian stop codon usage. *bioRxiv*: doi.org/10.1101/2022.1103.1102.482615.
- Ho AT, Hurst LD. 2021b. Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage. *Mol. Biol. Evol.* 39(1): msab326.
- Hollingsworth TJ, Gross AK. 2013. The severe autosomal dominant retinitis pigmentosa rhodopsin mutant Ter349Glu mislocalizes and induces rapid rod cell death. *J. Biol. Chem.* 288(40): 29047-29055.

Inagaki Y, Doolittle WF. 2000. Evolution of the eukaryotic translation termination system: Origins of release factors. *Mol. Biol. Evol.* 17(6): 882-889.

Jackson RJ, Hellen CUT, Pestova TV. 2012. Termination and post-termination events in eukaryotic translation. In: Marintchev A, editor. *Advances in Protein Chemistry and Structural Biology, Vol 86: Fidelity and Quality Control in Gene Expression*. p. 45-93.

Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*. 549(7673): 519–522.

Klauer AA, van Hoof A. 2012. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *WIREs RNA*. 3(5): 649-660.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* 2(4): RESEARCH0010.

Kobayashi K, Saito K, Ishitani R, Ito K, Nureki O. 2012. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1 alpha complex. *Nucleic Acids Res.* 40(18): 9319-9328.

Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* 289(44): 30334-30342.

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

Lee H, Popodi E, Tang HX, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 109(41): E2774-E2783.

- Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15(5): e1008141.
- Liang H, Cavalcanti AR, Landweber LF. 2005. Conservation of tandem stop codons in yeasts. *Genome Biol.* 6(4): R31.
- Liebman SW, Chernoff YO. 2012. Prions in yeast. *Genetics.* 191(4): 1041-1072.
- Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, Liti G, Tian D, Hurst LD, Yang S. 2017. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat. Ecol. Evol.* 2(1): 164-173.
- Liu HX, Huang J, Sun XG, Li J, Hu YW, Yu LY, Liti GN, Tian DC, Hurst LD, Yang SH. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat. Ecol. Evol.* 2(1): 164-173.
- Liu HX, Maclean CJ, Zhang JZ. 2019. Evolution of the yeast recombination landscape. *Mol. Biol. Evol.* 36(2): 412-422.
- Long HA, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo WF, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* 2(2): 237–240.
- Lynch M. 2007. *The origins of genome architecture.* Sunderland, MA.: Sinauer Associates Inc.
- Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 105(27): 9272-9277.
- Major LL, Edgar TD, Yee Yip P, Isaksson LA, Tate WP. 2002. Tandem termination signals: myth or reality? *FEBS Lett.* 514(1): 84-89.
- Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature.* 454(7203): 479-485.

- Meer KM, Nelson PG, Xiong K, Masel J. 2020. High transcriptional error rates vary as a function of gene expression level. *Genome Biol. Evol.* 12(1): 3754-3761.
- Meng SY, Hui JO, Haniu M, Tsai LB. 1995. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 211(1): 40-48.
- Namy O, Duchateau-Nguyen G, Rousset JP. 2002. Translational readthrough of the PDE2 stop codon modulates cAMP levels in *Saccharomyces cerevisiae*. *Mol. Microbiol.* 43(3): 641-652.
- Namy O, Hatin I, Rousset JP. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.* 2(9): 787-793.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. System.* 23(1): 263-286.
- Pang SY, Wang WH, Rich B, David R, Chang YT, Carbunaru G, Myers SE, Howie AF, Smillie KJ, Mason JI. 2002. 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.* 87(6): 2556-2563.
- Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev.* 53(3): 273-298.
- Pelham HRB. 1978. Leaky UAG termination codon in tobacco mosaic-virus RNA. *Nature.* 272(5652): 469-471.
- Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12(1): 32-42.
- Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biol. Direct.* 7(1): 30.
- Roberts SA, Gordenin DA. 2014. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer.* 14(12): 786-800.

- Rodnina MV. 2018. Translation in prokaryotes. *Cold Spring Harb. Perspect. Biol.* 10(9): a032664.
- Roth JR. 1970. UGA nonsense mutations in *Salmonella-typhimurium*. *J. Bacteriol.* 102(2): 467-475.
- Roy B, Leszyk JD, Mangus DA, Jacobson A. 2015. Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *Proc. Natl. Acad. Sci. U.S.A.* 112(10): 3038-3043.
- Ryden SM, Isaksson LA. 1984. 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.* 193(1): 38-45.
- Sanchez JC, Padron G, Santana H, Herrera L. 1998. Elimination of an HuIFN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J. Biotechnol.* 63(3): 179-186.
- Schueren F, Lingner T, George R, Hofhuis J, Dickel C, Gartner J, Thoms S. 2014. Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *Elife.* 3(1): e03640.
- Schueren F, Thoms S. 2016. Functional translational readthrough: a systems biology perspective. *PLoS Genet.* 12(8): e1006196.
- Seligmann H, Pollock DD. 2004. The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23(10): 701-705.
- Seoighe C, Kiniry SJ, Peters A, Baranov PV, Yang H. 2020. Selection shapes synonymous stop codon use in mammals. *J. Mol. Evol.* 88(7): 549-561.
- Sharp PM, Bulmer M. 1988. Selective differences among translation termination codons. *Gene.* 63(1): 141-145.
- Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* 4(3): 222-230.

Smith NGC, Eyre-Walker A. 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. *Mol. Biol. Evol.* 18(6): 982-986.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germline de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet.* 14(3): e1007254.

Strigini P, Brickman E. 1973. Analysis of specific misreading in *Escherichia coli*. *J. Mol. Biol.* 75(4): 659-672.

Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. U.S.A.* 87(12): 4692-4696.

Traverse CC, Ochman H. 2016. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci. U.S.A.* 113(12): 3311-3316.

Trotta E. 2016. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics.* 17(17): 366.

Vidal R, Frangione B, Rostagno A, Mead S, Revesz T, Plant G, Ghiso J. 1999. A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature.* 399(6738): 776-781.

Vidal R, Revesz T, Rostagno A, Kim E, Holton JL, Bek T, Bojsen-Moller M, Braendgaard H, Plant G, Ghiso J, et al. 2000. 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.* 97(9): 4920-4925.

Wagner A. 2005. Robustness, evolvability, and neutrality. *FEBS Lett.* 579(8): 1772-1778.

Wei Y, Wang J, Xia X. 2016. Coevolution between stop codon usage and release factors in bacterial species. *Mol. Biol. Evol.* 33(9): 2357-2367.

Wei Y, Xia X. 2017. The role of +4U as an extended translation termination signal in bacteria. *Genetics.* 205(2): 539-549.

Wickner RB, Masison DC, Edskes HK. 1995. PSI and URE3 as yeast prions. *Yeast*. 11(16): 1671-1685.

Yordanova MM, Loughran G, Zhdanov AV, Mariotti M, Kiniry SJ, O'Connor PBF, Andreev DE, Tzani I, Saffert P, Michel AM, et al. 2018. AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation. *Nature*. 553(7688): 356-360.

Chapter 7

Discussion (part 2)

Summary of results

That the multi-step process of gene expression is extremely prone to errors (Drummond and Wilke 2009), these being mostly deleterious (see for example: Xu and Zhang 2014; Liu and Zhang 2018a, b; Jiang and Zhang 2019; Xu and Zhang 2020, 2021), predicts selection to minimise their consequences (Warnecke and Hurst 2011). Such selection may take the form of error prevention or error mitigation, both of which can typically be facilitated by global or local solutions (Rajon and Masel 2011). In this thesis, I use translational read-through (TR) as an exemplar to study the evolutionary dynamics of local error prevention and local error mitigation solutions. How often do we see local evidence for error mitigation against TR? How do local error prevention and local error mitigation solutions to TR co-evolve, the strength of selection for each being dependent on the other? How might we explain the usage of non-optimal stop codons that possess higher intrinsic TR rates?

Stop codons are ideal sequence motifs to understand TR error control because the termination of translation is implicated both in TR prevention and mitigation. TR may be prevented at the local (genic) level by swapping the stop codon to a less error-prone one, TAA being universally the most reliable stop followed by TAG and TGA (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen, et al. 1993; Meng, et al. 1995; Sanchez, et al. 1998; Tate, et al. 1999; Wei, et al. 2016; Cridge, et al. 2018). TR mitigation can be facilitated locally by selection for 3' in-frame additional stop codons (ASCs) which offer a second opportunity for translation to terminate should termination fail at the canonical stop codon site (Liang, et al. 2005; Adachi and Cavalcanti 2009; Fleming and Cavalcanti 2019). In Chapter 2, I show that ASC enrichment above dinucleotide-controlled null is common to unicellular eukaryotes, but not multicellular ones nor bacteria (Ho and Hurst 2019). This result raises two interesting questions. Why might the strength of ASC selection be different between multicells and unicells? Why don't both eukaryotes and bacteria use ASCs for TR error mitigation?

One proposed explanation for the difference in ASC enrichment between unicellular and multicellular eukaryotes is that phenotypic errors are more costly for unicellular

organisms, which do not have the luxury of cell replacement, and hence error proofing selection is generally stronger (Ho and Hurst 2019). Evoking nearly neutral theory, another possibility is that unicellular species more readily evolve local error solutions such as ASCs because they typically possess larger effective population sizes (N_e) (see Lynch, et al. 2016 for some estimates) and thus possess more efficient selection. Any selection for error control is complicated by the fact that selection for error prevention and mitigation solutions are co-dependent. In Chapter 3, I find TAA enrichment above nucleotide expectations (a proxy for reduced TR rate) to correlate positively with N_e , while no such trend exists between ASC enrichment and N_e (Ho and Hurst 2020). While, then, this supports the notion that unicellular eukaryotes can more readily evolve local solutions than their multicellular counterparts, it does not explain why ASCs are under selection in some species but not others. Indeed, this result implies error prevention via modifying stop codon usage is the preferred error control strategy against TR in eukaryotes.

Both the covariance between TAA enrichment and N_e (Ho and Hurst 2020) and the observation that highly expressed genes prefer TAA stop codons (Korkmaz, et al. 2014; Trotta 2016) support TAA being the universal optimal stop codon. How then do non-optimal TGA and TAG stop codons persist in populations? In Chapters 4 and 5, I consider how we might explain between-species TGA and TAG stop codon usage trends. While TGA:TAG usage ratios in bacteria have been assumed to reflect the relative cellular abundance of RF1 and RF2 release factors (Sharp and Bulmer 1988; Korkmaz, et al. 2014; Wei, et al. 2016), in Chapter 4 I show that this need not be the case. Stop codon usage trends against genomic GC content at the canonical stop site are not only consistent when looking in the +1 and +2 reading frame of 3' UTR sequences (Ho and Hurst 2019; Ho and Hurst 2021), but also in eukaryotes and archaea who possess only one universal release factor (Ho and Hurst 2021). RF abundance is thus not needed to explain between-species bacterial stop codon usage trends. In Chapter 5, I consider the enigmatic observation of high TGA abundance and conservation in mammalian genomes (Seoighe, et al. 2020). Consistent with GC-biased gene conversion (gBGC) providing a parsimonious model to explain this, I find TGA usage and TAA→TGA stop codon flux to be well predicted by GC content and recombination rate. Analysis of human *de novo* mutations recovered a GC→AT mutation bias that drives stop codon usage away from, not towards, TGA. The

preferences of highly expressed genes also indicates that TAA, not TGA, is the most optimal stop codon in mammals, ruling out a selective explanation. TGA abundance and conservation in mammals appears to be promoted by gBGC in a manner that mimics purifying selection. This provides a cautionary tale that sequence conservation need not imply purifying selection.

The study of molecular evolution informs transgene design

In the design of any transgene, there are several considerations that can impact its eventual success or failure (van de Sluis and Voncken 2011; Jackson, et al. 2014; Troese-meier, et al. 2019). One might first choose the species that the transgene will originate from. Is the transgene intended for prokaryotic or eukaryotic expression? How diverged is this species from the target organism and how might this affect expression levels of the transgene? Next, one might consider transgene structure. Which regulatory elements should be selected for optimal expression levels? How many introns should be included? These decisions can have significant impacts on transgene expression, as demonstrated by the inclusion of one intron being linked to significant expression boosts (Choi, et al. 1991; Aronow, et al. 1992).

Many of the lessons learned about the optimisation of transgene sequences come from studying molecular evolution. Just as genes may be under selection to match their codon usage to the most abundant tRNAs in the tRNA pool to improve their expression (see Higgs and Ran 2008; Ran and Higgs 2010; Gingold and Pilpel 2011), so too might transgenes be designed with optimal nucleotides at synonymous sites to achieve the same goal (Foster, et al. 2008; Troese-meier, et al. 2019). Theoretically, the same logic could be applied to molecular evolution in response to erroneous protein synthesis. Just as OSCs are under selection in real genomes to protect against frameshifting (Seligmann and Pollock 2004; Abrahams and Hurst 2017; Abrahams and Hurst 2018; Seligmann 2019), transgenic coding sequences could incorporate OSCs proximal to “slippery” codons to minimise the consequences of ribosomal slippage. Indeed, as transgenes are commonly designed for high expression levels, the risks of molecular error may be exceptionally high. Built-in error prevention or error mitigation devices may thus provide great utility with little to no cost.

As the work described in this thesis is focused on translation termination, it is at the canonical stop codon site where I can make several recommendations for transgene design. In order of importance, transgene sequences should:

- (i) terminate using the optimal stop codon, TAA, which is associated with the lowest TR rates (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen, et al. 1993; Meng, et al. 1995; Sanchez, et al. 1998; Tate, et al. 1999; Wei, et al. 2016; Cridge, et al. 2018), preferred by highly expressed genes in bacteria and eukaryotes (Korkmaz, et al. 2014; Trotta 2016), positively correlated with N_e (Ho and Hurst 2020), and consistently the most enriched stop codon above dinucleotide expectations across bacteria, eukaryotes, and archaea (Ho and Hurst 2021).
- (ii) incorporate extended termination motifs that improve termination fidelity, such as +4T in bacteria (Major, et al. 2002; Wei and Xia 2017) and longer motifs in eukaryotes (see Namy, et al. 2001; Cridge, et al. 2018 for example) to reflect those conserved in highly expressed genes.
- (iii) include 3' in-frame additional stop codons as close to the primary stop codon as possible, but downstream of any extended termination motif, when designed for expression in a unicellular eukaryote species (such as yeast) or unicellular expression within a multicellular eukaryote species (such as germline-specific expression) (Ho and Hurst 2019).

Though ASCs do not appear to be under strong selection in multicellular eukaryote genomes, that unicellular eukaryote genomes are often enriched for ASCs supports their utility as a potential fail-safe mechanism in eukaryotic systems (Ho and Hurst 2019). Indeed, unicellular species may be the best model organisms for studying the evolution of phenotypic error solutions. It is in these organisms that the fitness costs of erroneous gene expression might be highest, multicells being able to buffer fitness costs by apoptosis and cell replacement, and if so we can expect error control motifs to be most enriched. Furthermore, unicellular organisms tend to have large N_e and thus can more readily evolve local error solutions that can be detected in genomic analysis. For these reasons I suggest that similar studies of error control within unicellular eukaryote genomes may be influential in elucidating other error prevention or mitigation solutions that might be useful inclusions in transgene design.

That TGA abundance and conservation in mammals cannot be explained by mutational models nor selective hypotheses provides a cautionary tale with regards transgene design. The most abundant codon need not be the most optimal, even when signatures of sequence conservation are present in genome-wide analysis. To include TGA stop codons in human transgenes for example would be detrimental due to its high TR rate, despite it being the most common stop codon in the human genome. The same logic could apply to other genomic contexts where the deleterious allele also happens to be the GC-rich state that is promoted by gBGC and appears conserved. Identifying such examples does not fall under the scope of my thesis, however this could indeed form some important future work.

Both low effective population size and GC-biased gene conversion can be responsible for genomic imperfections

The second broad implication of the work presented relates to understanding the causes of genomic imperfection. Stop codon usage across the tree of life is highly variable and far from perfect, but it isn't the only example of imperfection within genomes. When one mentions genomic imperfection, the first thought probably relates to *de novo* mutations, the resulting alleles being able to circulate in the gene pool for many generations under certain circumstances. While most mutations are only mildly deleterious (Ohta 1992), others contribute to the genetic basis of disease, common examples of this including single gene disorders like cystic fibrosis and cancers with a strong heritable component (e.g. BRCA mutations that predispose an individual to developing breast cancer). These disorders are surprisingly common. In total, an estimated 3.5-5.9% of people suffer with rare diseases, most of which are genetic, equating to ~263-446 million people globally (Wakap, et al. 2020). Indeed, there also exists genomic imperfections that are less obvious. The human genome is extremely "bloated", possessing numerous large introns (Warnecke, et al. 2008), large distances between our genes (Lynch and Conery 2003), and a high transposable element load (Lynch and Conery 2003). How do such genomic features such as these persist despite selection favouring improved fitness? Why aren't genetic diseases purged from our populations? Why aren't our genomes the perfect machine?

My investigations into the forces promoting the usage of non-optimal TGA and TAG stop codons provides a better understanding of the evolutionary processes that lead to the accumulation of deleterious sequences. The results in Chapter 3, for example, lend support to the nearly neutral predictions that the frequency of selectively optimal features of genome architecture should covary positively with N_e (Ohta 1992; Lynch 2007). Within the stop codon exemplar, there is a positive correlation between TAA stop codon enrichment and N_e when controlling for underlying phylogeny (Ho and Hurst 2020). In Chapter 5, I show this result is resilient to whether we consider all genes en masse or restrict analysis to just highly expressed or lowly expressed genes, suggesting TAA is optimal for all genes regardless of expression level. By extension, the nearly neutral theory also predicts non-optimal sequence to covary negatively with N_e . Indeed, in support of this, I find TGA enrichment to be negatively associated with N_e in both highly and lowly expressed genes. N_e and nearly neutral theory hence provide a parsimonious framework to explain much of the between-species variation in stop codon usage observed in eukaryotes: non-optimal TGA and TAG stop codons persist in low N_e species due to the absence of efficient selection to purge them and the increased influence of genetic drift.

While N_e and nearly neutral theory can parsimoniously explain some between-species trends, it provides less insight into within-species trends. The high raw abundance of TGA stop codons in mammals might be attributable to drift, for example, but drift cannot explain the intragenomic covariance between TGA usage and GC content observed in humans (e.g. Trotta 2016). Chapter 5 provides a set of results concordant with gBGC driving TGA to high usage by promoting TAA→TGA flux in GC-rich isochores. As gBGC provides a fixation advantage to GC-rich alleles in a manner coupled tightly to recombination, gBGC correctly predicts (i) a positive correlation between TGA usage and GC content at the canonical stop codon site and outside of the termination context in the 5' UTR and 3' UTR, (ii) a positive correlation between TGA usage and recombination rate, and (iii) increased TAA→TGA flux in GC-rich and highly recombining domains of mammalian and bird genomes. Coupled with evidence supporting TAA optimality, gBGC in the stop codon example appears to promote the fixation of non-optimal sequence in a manner that mimics positive selection, corroborating prior suggestions of its potential to do so (Nagyilaki 1983;

Dreszer, et al. 2007; Berglund, et al. 2009; Ratnakumar, et al. 2010; Corcoran, et al. 2017; Bolivar, et al. 2019).

The mammalian TGA enigma is particularly intriguing because TGA stops are not just highly abundant but highly conserved (Belinky, et al. 2018; Seoighe, et al. 2020). That the flux data, in addition to stop codon usage, supports gBGC as the parsimonious cause hence raises concerns about the reliability of assuming sequence conservation to equal purifying selection, as is common practice in research (see Ponting 2017 for discussion) and medical contexts (Cooper, et al. 2010; Sun and Yu 2019). My results hence advocate control for gBGC in evolutionary studies of sequences where gBGC may be a factor, i.e. in species with a strong AT→GC bias in their mismatch repair and high variance in their recombination rates. This echoes the recommendations of Bolivar, et al. (2019), who suggest that accounting for gBGC is important to accurately estimate the strength of selection. In their example, Bolivar, et al. (2019) demonstrate that control for gBGC in evolutionary analysis may be achieved by calculating dN/dS separately for substitutions that do not alter GC content (i.e. G↔C and A↔T) and are thus unaffected by gBGC. More generally, Duret and Galtier (2009) suggest three key considerations to help distinguish between gBGC and selection. If observed non-neutral genomic patterns are (i) GC-biased, (ii) common to all loci within a given region irrespective of functional status, or (iii) strongest in highly recombining regions, then gBGC should be considered as a possible cause. Duret and Galtier (2009) do note that selection can in certain scenarios favour GC-rich sequences, affect non-coding sequences, and be linked to recombination via Hill-Robertson interference (Hill and Robertson 1966), but nonetheless these questions should help researchers come to reliable conclusions.

Several questions remain

There are several questions that arose during this work and have not yet been answered. Perhaps the most significant of these is the TAG problem. While TAA usage covaries negatively, and TGA usage positively, TAG usage is mostly low and unresponsive to GC pressure in bacteria and archaea despite having an identical nucleotide content to TGA (Povolotskaya, et al. 2012; Korkmaz, et al. 2014; Trotta 2016; Ho and Hurst 2019). In eukaryotes, TAG usage correlates with genomic GC

content but is consistently lower in absolute terms than TGA (Ho and Hurst 2021). Indeed, TGA and TAG also behave differently in the intragenomic analyses of Chapter 5. If gBGC is to explain the observed high abundance and conservation of TGA stop codons, why does not act equally to increase TAG usage? For this question I would like to point to Chapter 6, “Discussion part 1”, where I summarise the salient points from the preceding chapters that relate to the TAG problem highlighting plausible roles for mutation bias, selection, and an extended model of gBGC. Here in “Discussion part 2”, I will instead discuss two other unanswered questions. First, why don’t both prokaryotes and eukaryotes select for ASCs? Second, with high N_e and efficient selection to purge TGA stop codons, why do TGA stop codons dominate in GC-rich bacterial species?

Why don’t prokaryotic genomes employ fail-safe additional stop codons?

That bacterial and eukaryotic genomes promote TAA stop codon usage in their highly expressed genes is evidence that both are under selection to prevent TR (Korkmaz, et al. 2014; Trotta 2016). It is unusual, then, that we do not see evidence of TR mitigation in both groups. ASCs, probably acting as fail-safe mechanisms in the 3’ UTR, are commonly enriched beyond null nucleotide expectations in unicellular eukaryotes but not in bacteria (Ho and Hurst 2019). Why might this be?

One possible explanation concerns differences in N_e . Relative to bacteria (Sung, et al. 2012; Sela, et al. 2016), eukaryotes have typically smaller population sizes and less efficient selection (Ohta 1992) to purge leaky TGA stop codons and mould their stop codon usage in response to TR. With a lack of reliable TR prevention, one could argue that TR mitigation solutions like ASCs might then be under stronger selection in eukaryotes. However, there are several reasons why such a model is unconvincing. First, there is a logic problem. If selection is too weak to select for TAA why should it be strong enough to selection for ASCs? Second, if lower N_e predicts error mitigation instead of error prevention, why do we not see evidence of ASC enrichment in multicellular eukaryotes (Ho and Hurst 2019)? Such species have even smaller N_e than their unicellular relatives and tend to possess lower TAA usage, hence this model predicts error mitigation to be even more important in multicells all else being equal. Third, even if N_e is lower in unicellular eukaryotes than bacteria, it is still large enough

for selection for both high TAA usage (relative to multicellular eukaryotes) and ASC enrichment. With even higher N_e , why should selection in bacteria favour one (TAA) and not the other (ASCs)?

A second possible explanation is that there are differences between prokaryote and eukaryote global error control mechanisms that modify the need for local solutions. To recap the framework of Rajon and Masel (2011), global solutions refer to cellular pathways that prevent or mitigate errors occurring at multiple genomic loci (as opposed to local solutions, like ASCs, which affect just one gene and must hence evolve multiple times for a genome-wide effect). For example, a global TR prevention solution might be a mutation that improves the binding affinity of the class I release factor (RF1 or RF2 in bacteria, eRF1 in eukaryotes) to the stop codon and thus lowers cellular TR rates. RF binding affinity, however, is unlikely to cause the difference between bacteria and eukaryotes as the most recent available data suggests TR rates are comparable between the two groups (Cridge, et al. 2018; Zhang, et al. 2020). Alternatively, global error TR mitigation can be facilitated by NSD pathways that liberate the ribosome should it reach the polyA⁺ tail and lead to degradation of the mRNA transcript (Klauer and van Hoof 2012). Unlike global prevention via RF binding affinity, that global mitigation by NSD degradation is mediated by the exosome (van Hoof, et al. 2002) in eukaryotes and RNase R in bacteria (Richards, et al. 2006) provides a possible difference in efficacy between the prokaryote and eukaryote systems. If NSD degradation were more efficient or less energetically expensive in bacteria than eukaryotes, this could theoretically circumvent the need for local TR mitigation (via ASCs) in bacteria. A similar, speculative, possibility is that there exists an additional global mitigation pathway unique to bacterial systems that removes the need for local mitigation by ASCs. No such pathway however exists to the best of my knowledge.

To further investigate this problem computationally is difficult without more experimental data. Elucidation of TR rates for a wider range of bacteria and eukaryotes would be beneficial to determine whether rates are consistent between the two groups as is currently the expectation, or if TR occurs at a lower rate in bacteria and thus ASCs are not needed for error control in these species. More generally, the preferred error control strategies of both groups might be best examined by experimental

evolution. I imagine one could optimise a transgene for high expression with a particularly leaky translation termination signal, a TGA stop codon with a TR-associated nucleotide context, and express the construct in a model organism for each group (*E. coli* and *S. cerevisiae*). Growth of these organisms over many generations with sequencing at several time points should provide a clue as to which mutations are preferentially fixed in populations for error control. Modifications to the termination motif would indicate a preference towards error prevention whereas mutations increasing ASC density in the 3' UTR would indicate a preference for error mitigation. I note that it's possible selection could instead opt to reduce the expression level of the transcript, rather than directly manage the TR problem, if the transgene is not essential for survival and growth. This might be avoided by modifying the termination motif of an existing essential gene, using CRISPR for example, rather than expressing an exogenous transgene.

How can we explain high TGA usage in GC-rich bacteria?

Stop codon usage between bacterial genomes is highly variable (Korkmaz, et al. 2014; Ho and Hurst 2021). Just as TGA stop codons dominate in some eukaryotes, so do they dominate in some bacteria, not least those with high genomic GC content. Given bacteria possess large N_e (Sung, et al. 2012; Sela, et al. 2016), explaining high TGA usage in these species within the framework of nearly neutral theory is unclear as non-optimal TGA stops should be reliably purged. What, then, are the other hypotheses for TGA abundance in GC-rich bacteria?

As outlined in Chapter 4, the release factor (RF) hypothesis provides one possible framework for explaining TAG (recognised by RF1) and TGA (recognised by RF2) usage trends (Sharp and Bulmer 1988). This received support from observations that the TAG:TGA usage ratio is correlated with RF1:RF2 abundance ratios (Korkmaz, et al. 2014; Wei, et al. 2016). Most notably, Wei, et al. (2016) notice that RF2 abundance is low in species with low GC3 content and thus provide a possible explanation for the positive covariance between TGA usage and genomic GC content. They propose that the unresponsiveness of TAG to GC pressure is caused by a myriad of factors: TAA usage is high at low GC due to the mutation bias, TAA is preferred at mid-GC by selection for low TR, and TGA is highly abundant at high GC due to high RF2 levels

relative to RF1 (Wei, et al. 2016). The results presented in Chapter 4, however, argue RF biology is not needed to explain stop codon usage trends. Bacterial stop codon usage trends against genomic GC content are the same at the canonical stop site as seen (i) downstream in the 3' UTR, (ii) in archaeal (and to a lesser extent eukaryotic) genomes that lack RF diversity, and (iii) between human isochores where the RF environment is the same for all genes (Ho and Hurst 2021).

A few studies to date have proposed gBGC may be widespread across bacterial taxa and this in theory could go some way to explaining variation in TGA usage, not least its positive covariance with GC content. Most evidence for bacterial gBGC comes indirectly from either observed correlations between recombination and GC content or the analysis of intergenic regions that are free from coding constraints (Touchon, et al. 2009; Lassalle, et al. 2015). Across a broad sample of 21 bacterial clades, Lassalle, et al. (2015) observe consistent positive relationships between GC content and recombination rate both within genes and within intergenic regions. This, they argue, is unlikely to be driven by increased selection by Hill-Robertson interference, as there is a depletion of optimal AU-ending codons in highly recombining genes (Lassalle, et al. 2015). Indeed, the observations of Lassalle, et al. (2015) meet all three of criteria of Duret and Galtier (2009) that suggest gBGC should at the very least that be considered a possible cause. Combined with recombination rates being highly variable between bacterial species (Vos and Didelot 2009), gBGC theoretically could explain between-species variation in GC content and TGA usage across the whole group. This might indeed be more parsimonious than evoking selective reasons to explain why many intergenic regions, free of coding or expression constraint, sit at higher GC content than mutational equilibrium (Hershberg and Petrov 2010; Lassalle, et al. 2015).

The notion of widespread gBGC influencing nucleotide composition in bacteria isn't without its controversy, however. Hildebrand, et al. (2010) and Yahara, et al. (2016) for example report only a modest relationship between recombination rate and GC content. More generally, such analysis has its limitations due to the difficulty in accurately estimating recombination rates at high resolution (e.g. at specific sites and not for kilobase windows). To circumvent this issue, Bobay and Ochman (2017) took a site-by-site approach to analyse only alleles for which there is strong direct evidence

for recombination. Contra to the gBGC hypothesis, they found recombinant alleles to be subjected to stronger purifying selection than those resulting from *de novo* mutation and biased towards AT, not GC (Bobay and Ochman 2017). These results clearly do not support that the strength of recombination and gBGC may counteract the AT-biased nucleotide composition introduced to bacterial genomes by the mutation.

At present, it seems the extent to which gBGC can explain between-species bacterial trends in GC content and TGA usage remains an open question. What about the other hypotheses for GC content variation in bacteria? As TGA usage positively correlates with GC content, it's likely that the same forces are responsible for both. With the broad acceptance that the mutation bias is universally in the direction of GC->AT (Smith and Eyre-Walker 2001; Lynch, et al. 2008; Hershberg and Petrov 2010; Hildebrand, et al. 2010; Long, et al. 2018), the alternative hypotheses for bacterial GC content variation are selective. Elucidating the key selection pressures is difficult though, not least because genomic GC content covaries with many intrinsic and extrinsic factors (Foerstner, et al. 2005). While GC content positively correlates with genome size (Heddi, et al. 1998; Moran 2002; Rocha and Danchin 2002), it is also proposed, controversially (Marashi and Ghalanbor 2004), to correlate with growing temperature (Musto, et al. 2004) and the ability to fix nitrogen (McEwan, et al. 1998). If robust, these observations may provide some clues to the underlying selection pressures that influence genomic GC content. The robustness of these trends is however unclear as some groups (including Lassalle, et al. (2015) in their endorsement of bacterial gBGC) find the trends between GC content and ecology to be weak.

How to further investigate this problem is not trivial. That GC content covaries with ecological niche is not only a difficult problem for selective hypotheses, but also for the bacterial gBGC hypothesis. In Chapter 5, I find that GC3 variance within a species' genome to be a good predictor of the strength of gBGC. A simple experiment, then, could test for correlation between GC3 variance and the aforementioned intrinsic and extrinsic factors related to ecological niche. If a correlation were to be recovered, this could be considered new evidence in favour of gBGC causing bacterial GC content and TGA usage variation. I note that no matter what the force is that is shaping bacterial GC content and nucleotide composition, the stop codon exemplar suggests that it must be rather strong. The effect of GC pressure on TGA usage must be

balanced with concurrent selection to minimise TR rates. In AT-rich bacteria, both low GC content and low TR rates are facilitated by AT-rich TAA stop codons. In GC-rich bacteria, however, GC pressure must be greater than selection against TR such that the net effect is towards TGA stop codons, at least outside of highly expressed genes (Korkmaz, et al. 2014).

Summary, limitations, and outlook

In this thesis I have presented a series of manuscripts that develop our understanding of stop codon usage and its role in error control. Through the systematic analysis of prokaryotic, eukaryotic, and archaeal genomes, I have elucidated the stop codon preferences of each group, corroborating prior suggestions that TAA is universally optimal (Povolotskaya, et al. 2012; Korkmaz, et al. 2014; Trotta 2016; Belinky, et al. 2018). Rather than different species having different optimal stop codons, much of the variation in stop codon usage between taxa may instead be explained by differences in the interplay of mutation bias, selection against TR, genetic drift, and gBGC. TR may be locally prevented by mutations that switch the stop codon or locally mitigated by selection for ASCs (Rajon and Masel 2011). As both involve stop codons, this provides a unique framework to study the co-evolution of error prevention and error mitigation strategies. My results suggest that error prevention is the preferred solution to TR in bacteria and eukaryotes. With the publication of these chapters, I hope to provide an interesting account of stop codon evolution throughout the tree of life. With the opportunity to write an invited stop codon usage review at a respected journal, I hope to re-energise interest in the field by highlighting the significant gaps in our understandings.

All the analyses presented in this thesis were conducted computationally. On the one hand, I believe this has demonstrated the robustness of bioinformatic methods and the utility of such methods for improving our understanding of molecular evolution. On the other, I acknowledge that the lack of experimental data may be considered a limitation. Indeed, experimental validation of some of my results would be influential in providing a molecular basis to support them. In Chapter 5, for example, my proposal of an extended gBGC model that allows for context dependency to treat GC-equivalent trinucleotides differently would benefit from tetrad analysis that directly records the

sequence context at gBGC loci as well as the GC bias. This could be influential work with wide implications for molecular evolution, not just for the field of stop codon usage where validation of the extended gBGC model has the potential to explain the TAG problem.

My computational investigations into stop codon usage are also limited by some important assumptions. It is assumed, for example, that TR is common and deleterious in all species. That TR is common enough to be opposed by selection in all species seems reasonable as TR rates are consistent between model species of eukaryotes (e.g. humans (Cridge, et al. 2018)) and prokaryotes (e.g. *Escherichia coli* (Zhang, et al. 2020)). Experimentally derived TR rates are however not available for many other species and the elucidation of TR rates across the tree of life would thus be greatly informative to the design of future bioinformatic analyses. Indeed, any observed TR rates that are significantly deviant from the existing data would oppose the assumption that TR rates may be generalised to large species datasets such as those analysed here. The second assumption that TR is deleterious for all genes and in all species also seems reasonable given the support from the best available data (Li and Zhang 2019). Nevertheless, it should be noted that functional TR (FTR) has now been described in a few species (e.g. Jungreis, et al. 2011; Schueren and Thoms 2016; Zhang, et al. 2020). For most species, the impact of FTR on genome-wide selection against TR should be minimal as FTR is most often unique to a small subset of genes (Jungreis, et al. 2011; Schueren and Thoms 2016) or specific conditions of environmental stress (Zhang, et al. 2020). Some viral genomes, which must maximise the coding capacity of their small genomes (Firth and Brierley 2012), might be exceptions to this rule but are not analysed in any of the above chapters. Viruses do, however, demonstrate the possibility that genome-wide TR can be net beneficial not deleterious for some species. I am not aware of any prokaryotes or eukaryotes where this is the case, but this is nonetheless a limitation of between-species analysis in the absence of more TR rate data.

Related to the above is the methodological limitation of estimating a species' selective preferences by averaging selective effects across all genes in its genome. For example, in Chapter 3, when calculating "TAA enrichment" and "ASC enrichment" metrics I first calculate the genome-wide frequencies of each and compare these to a null model.

Such calculation assumes all genes are equally subjected to the same selection pressures and thus gives equal weighting to each gene. This crucially ignores more localised selection pressures like FTR for individual genes. More generally, the averaging across genes is also sensitive to local nucleotide composition preferences. TAA stop codons may occur commonly in GC-poor sequences while TGA and TAG stop codons might occur in GC-rich sequences due to chance alone. I have attempted to control for such regionalised effects by performing nucleotide or dinucleotide-matched controls, however I acknowledge the limitations of calculating such genome-wide “enrichment” metrics.

I would like to conclude with the sentiment that, while computational methods such as those used in this thesis come with assumptions and limitations, they have been instrumental in developing my understanding of stop codon usage, TR, and molecular evolution more generally. We now find ourselves with a wealth of publicly available genomic datasets waiting for bioinformatic analysis, and extracting the maximum out of these datasets will be integral to science in the 21st century. Not only does computational analysis provide new insights, it also can be used to advise the design of experimental procedures to optimise time and save resources. With sequencing now happening at an unprecedented scale, exemplified by the ongoing sequencing effort in response to the ongoing SARS-CoV-2 pandemic, this opportunity looks set to continue for many years to come.

References

Abrahams L, Hurst LD. 2017. Adenine enrichment at the fourth CDS residue in bacterial genes is consistent with error proofing for +1 frameshifts. *Mol. Biol. Evol.* 34(12): 3064-3080.

Abrahams L, Hurst LD. 2018. Refining the ambush hypothesis: Evidence that GC- and AT-rich bacteria employ different frameshift defence strategies. *Genome Biol. Evol.* 10(4): 1153-1173.

Adachi M, Cavalcanti AR. 2009. Tandem stop codons in ciliates that reassign stop codons. *J. Mol. Evol.* 68(4): 424-431.

- Aronow BJ, Silbiger RN, Dusing MR, Stock JL, Yager KL, Potter SS, Hutton JJ, Wiginton DA. 1992. Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. *Mol. Cell. Biol.* 12(9): 4170-4185.
- Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci. Rep.* 8(1): 9260.
- Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in humans genes. *PLoS Biol.* 7(1): e26.
- Bobay L-M, Ochman H. 2017. Impact of recombination on the base composition of bacteria and archaea. *Mol. Biol. Evol.* 34(10): 2627-2636.
- Bolivar P, Gueguen L, Duret L, Ellegren H, Mugal CF. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol.* 20(1): 5.
- Choi T, Huang M, Gorman C, Jaenisch R. 1991. A generic intron increases gene expression in transgenic mice. *Mol. Cell. Biol.* 11(6): 3070 - 3074.
- Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. 2010. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods.* 7(4): 250-251.
- Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. *Genome Biol. Evol.* 9(11): 2987-3007.
- Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.* 46(4): 1927-1944.
- Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. *Genome Res.* 17(10): 1420-1430.

- Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* 10(10): 715-724.
- Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genom. Hum. Genet.* 10(1): 285-311.
- Firth AE, Brierley I. 2012. Non-canonical translation in RNA viruses. *J. Gen. Virol.* 93(7): 1385-1409.
- Fleming I, Cavalcanti ARO. 2019. Selection for tandem stop codons in ciliate species with reassigned stop codons. *PLoS One.* 14(11): e0225804.
- Foerster KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6(12): 1208-1213.
- Foster H, Sharp PS, Athanasopoulos T, Trollet C, Graham IR, Foster K, Wells DJ, Dickson G. 2008. Codon and mRNA sequence optimization of microdystrophin transgenes improves expression and physiological outcome in dystrophic mdx mice following AAV2/8 gene transfer. *Mol. Ther.* 16(11): 1825-1832.
- Geller AI, Rich A. 1980. A UGA termination suppression tRNA^{Trp} active in rabbit reticulocytes. *Nature.* 283(5742): 41-46.
- Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* 7(1): 481.
- Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P. 1998. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: A peculiar G+C content of an endocytobiotic DNA. *J. Mol. Evol.* 47(1): 52-61.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9): e1001115.
- Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol. Biol. Evol.* 25(11): 2279-2291.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9): e1001107.

- Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* 8(3): 269-294.
- Ho AT, Hurst LD. 2020. Effective population size predicts local rates but not local mitigation of read-through errors in eukaryotic genes. *Mol. Biol. Evol.* 38(1): 244–262.
- Ho AT, Hurst LD. 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3' additional stop codons. *PLoS Genet.* 15(9): e1008386.
- Ho AT, Hurst LD. 2021. Variation in release factor abundance is not needed to explain trends in bacterial stop codon usage. *Mol. Biol. Evol.* 39(1): msab326.
- Jackson MA, Sternes PR, Mudge SR, Graham MW, Birch RG. 2014. Design rules for efficient transgene expression in plants. *Plant Biotechnol. J.* 12(7): 925-933.
- Jiang D, Zhang J. 2019. The preponderance of nonsynonymous A-to-I RNA editing in coleoids is nonadaptive. *Nat. Commun.* 10(1): 5411.
- Jorgensen F, Adamski FM, Tate WP, Kurland CG. 1993. Release factor-dependent false stops are infrequent in *Escherichia coli*. *J. Mol. Biol.* 230(1): 41-50.
- Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res.* 21(12): 2096-2113.
- Klauer AA, van Hoof A. 2012. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *WIREs RNA.* 3(5): 649-660.
- Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* 289(44): 30334-30342.
- Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. *PLoS Genet.* 11(2): e1004941.

- Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15(5): e1008141.
- Liang H, Cavalcanti AR, Landweber LF. 2005. Conservation of tandem stop codons in yeasts. *Genome Biol.* 6(4): R31.
- Liu Z, Zhang J. 2018a. Human C-to-U coding RNA editing is largely nonadaptive. *Mol. Biol. Evol.* 35(4): 963-969.
- Liu Z, Zhang J. 2018b. Most m(6)A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Mol. Biol. Evol.* 35(3): 666-675.
- Long HA, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo WF, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* 2(2): 237–240.
- Lynch M. 2007. *The origins of genome architecture.* Sunderland, MA.: Sinauer Associates Inc.
- Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* 17(11): 704-714.
- Lynch M, Conery JS. 2003. The origins of genome complexity. *Science.* 302(5649): 1401-1404.
- Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 105(27): 9272-9277.
- Major LL, Edgar TD, Yee Yip P, Isaksson LA, Tate WP. 2002. Tandem termination signals: myth or reality? *FEBS Lett.* 514(1): 84-89.
- Marashi SA, Ghalanbor Z. 2004. Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. *Biochem. Biophys. Res. Commun.* 325(2): 381-383.

- McEwan CEA, Gatherer D, McEwan NR. 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas*. 128(2): 173-178.
- Meng SY, Hui JO, Haniu M, Tsai LB. 1995. Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 211(1): 40-48.
- Moran NA. 2002. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell*. 108(5): 583-586.
- Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* 573(1-3): 73-77.
- Nagylaki T. 1983. Evolution of a large population under gene conversion. *Proc. Natl. Acad. Sci. U.S.A.* 80(19): 5941-5945.
- Namy O, Hatin I, Rousset JP. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.* 2(9): 787-793.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. System.* 23(1): 263-286.
- Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev.* 53(3): 273-298.
- Ponting CP. 2017. Biological function in the twilight zone of sequence conservation. *BMC Biol.* 15(1): 1-9.
- Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biol. Direct.* 7(1): 30.
- Rajon E, Masel J. 2011. Evolution of molecular error rates and the consequences for evolvability. *Proc. Natl. Acad. Sci. U.S.A.* 108(3): 1082-1087.
- Ran W, Higgs PG. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol. Biol. Evol.* 27(9): 2129-2140.

- Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. B.* 365(1552): 2571-2580.
- Richards J, Mehta P, Karzai AW. 2006. RNase R degrades non-stop mRNAs selectively in an SmpB-tmRNA-dependent manner. *Mol. Microbiol.* 62(6): 1700-1712.
- Rocha EP, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18(6): 291-294.
- Sanchez JC, Padron G, Santana H, Herrera L. 1998. Elimination of an HufN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J. Biotechnol.* 63(3): 179-186.
- Schueren F, Thoms S. 2016. Functional translational readthrough: a systems biology perspective. *PLoS Genet.* 12(8): e1006196.
- Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 113(41): 11399-11407.
- Seligmann H. 2019. Localized context-dependent effects of the “ambush” hypothesis: More off-frame stop codons downstream of shifty codons. *DNA Cell Biol.* 38(8): 786-795.
- Seligmann H, Pollock DD. 2004. The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23(10): 701-705.
- Seoighe C, Kiniry SJ, Peters A, Baranov PV, Yang H. 2020. Selection shapes synonymous stop codon use in mammals. *J. Mol. Evol.* 88(7): 549-561.
- Sharp PM, Bulmer M. 1988. Selective differences among translation termination codons. *Gene.* 63(1): 141-145.
- Smith NGC, Eyre-Walker A. 2001. Synonymous codon bias is not caused by mutation bias in G+C- rich genes in humans. *Mol. Biol. Evol.* 18(6): 982-986.

- Strigini P, Brickman E. 1973. Analysis of specific misreading in *Escherichia coli*. *J. Mol. Biol.* 75(4): 659-672.
- Sun H, Yu GJ. 2019. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci. Rep.* 9(1): 1667.
- Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. U.S.A.* 109(45): 18488-18492.
- Tate WP, Mansell JB, Mannering SA, Irvine JH, Major LL, Wilson DN. 1999. UGA: a dual signal for 'stop' and for recoding in protein synthesis. *Biochemistry (Mosc).* 64(12): 1342-1353.
- Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5(1): e1000344.
- Troeseemeier J-H, Rudolf S, Loessner H, Hofner B, Reuter A, Schulenburg T, Koch I, Bekeredjian-Ding I, Lipowsky R, Kamp C. 2019. Optimizing the dynamics of protein expression. *Sci. Rep.* 9(1): 7511.
- Trotta E. 2016. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics.* 17(17): 366.
- van de Sluis B, Voncken JW. 2011. Transgene design. In: Hofker MH, VanDeursen JM, editors. *Transgenic Mouse Methods and Protocols, Second Edition*. p. 89-101.
- van Hoof A, Frischmeyer PA, Dietz HC, Parker R. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science.* 295(5563): 2262-2264.
- Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *Isme Journal.* 3(2): 199-208.

- Wakap SN, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Mury D, Le Cam Y, Rath A. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Europ. J. Hum. Genet.* 28(2): 165-173.
- Warnecke T, Hurst LD. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat. Rev. Genet.* 12(12): 875-881.
- Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9(2): r29.
- Wei Y, Xia X. 2017. The role of +4U as an extended translation termination signal in bacteria. *Genetics.* 205(2): 539-549.
- Wei YL, Wang J, Xia XH. 2016. Coevolution between stop codon usage and release factors in bacterial species. *Mol. Biol. Evol.* 33(9): 2357-2367.
- Xu C, Zhang J. 2020. Mammalian alternative translation initiation is mostly nonadaptive. *Mol. Biol. Evol.* 37(7): 2015-2028.
- Xu C, Zhang J. 2021. Mammalian circular RNAs result largely from splicing errors. *Cell Rep.* 36(4): 109439.
- Xu G, Zhang J. 2014. Human coding RNA editing is generally nonadaptive. *Proc. Natl. Acad. Sci. U.S.A.* 111(10): 3769-3774.
- Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, Falush D. 2016. The landscape of realized homologous recombination in pathogenic bacteria. *Mol. Biol. Evol.* 33(2): 456-471.
- Zhang H, Lyu Z, Fan Y, Evans CR, Barber KW, Banerjee K, Igoshin OA, Rinehart J, Ling J. 2020. Metabolic stress promotes stop-codon readthrough and phenotypic heterogeneity. *Proc. Natl. Acad. Sci. U.S.A.* 117(36): 22167-22172.

Appendix 1

Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design

Alan M. Rice, Atahualpa Castillo Morales, Alexander T. Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, Laurence D. Hurst



Molecular Biology & Evolution, 38(1): 67–83.

This chapter contains work published on 20th July 2020 at MBE, the original and sole place of publication. It thus contains analysis of publicly available data using bespoke scripts that are freely available at the locations cited within the paper. The paper is open access and I have permission as the author to include the article in full (https://academic.oup.com/journals/pages/authors/production_and_publication/online_licensing). The latest version of the published article can be found by following the address: <https://doi.org/10.1093/molbev/msaa188>.

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:				
Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design				
Publication status (tick one)				
Draft manuscript <input type="checkbox"/> Submitted <input type="checkbox"/> In review <input type="checkbox"/> Accepted <input type="checkbox"/> Published <input checked="" type="checkbox"/>				
Publication details (reference)	Rice AM, Morales AC, Ho AT, Mordstein C, Muhlhausen S, Watson S, Cano L, Young B, Kudla G, Hurst LD. 2020. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol. Biol. Evol. 38(1): 67-83.			
Copyright status (tick the appropriate statement)				
I hold the copyright for this material <input type="checkbox"/> Copyright is retained by the publisher, but I have been given permission to replicate the material here <input checked="" type="checkbox"/>				
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the... Formulation of ideas: 10% Design of methodology: 10% Bioinformatic analyses: 20% Experimental work: N/a Presentation of data in journal format: 10%			
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.			
Signed	<table border="1" style="width: 100%;"> <tr> <td style="width: 70%;"></td> <td style="width: 10%;">Date</td> <td style="width: 20%;">03/03/2022</td> </tr> </table>		Date	03/03/2022
	Date	03/03/2022		

Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design

Alan M. Rice,^{1,†} Atahualpa Castillo Morales,^{1,†} Alexander T. Ho ^{1,†} Christine Mordstein,^{1,2} Stefanie Mühlhausen,¹ Samir Watson,³ Laura Cano,² Bethan Young,^{1,2} Grzegorz Kudla,^{†,2} and Laurence D. Hurst ^{†,1}

¹The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

²MRC Human Genetics Unit, Institute for Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, United Kingdom

³Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

[†]These authors contributed equally to this work.

[‡]Co-senior authors.

*Corresponding author: E-mail: l.d.hurst@bath.ac.uk

Associate editor: Jeffrey Townsend

Abstract

Large-scale re-engineering of synonymous sites is a promising strategy to generate vaccines either through synthesis of attenuated viruses or via codon-optimized genes in DNA vaccines. Attenuation typically relies on deoptimization of codon pairs and maximization of CpG dinucleotide frequencies. So as to formulate evolutionarily informed attenuation strategies that aim to force nucleotide usage against the direction favored by selection, here, we examine available whole-genome sequences of SARS-CoV-2 to infer patterns of mutation and selection on synonymous sites. Analysis of mutational profiles indicates a strong mutation bias toward U. In turn, analysis of observed synonymous site composition implicates selection against U. Accounting for dinucleotide effects reinforces this conclusion, observed UU content being a quarter of that expected under neutrality. Possible mechanisms of selection against U mutations include selection for higher expression, for high mRNA stability or lower immunogenicity of viral genes. Consistent with gene-specific selection against CpG dinucleotides, we observe systematic differences of CpG content between SARS-CoV-2 genes. We propose an evolutionarily informed approach to attenuation that, unusually, seeks to increase usage of the already most common synonymous codons. Comparable analysis of H1N1 and Ebola finds that GC3 deviated from neutral equilibrium is not a universal feature, cautioning against generalization of results.

Key words: SARS-CoV-2, mutation equilibrium, selection, synonymous mutations, vaccine design, viral attenuation.

Introduction

Multiple strategies toward the development of a SARS-CoV-2 vaccine are being pursued (Thanh Le et al. 2020). These include attenuated or inactivated viruses, replicating and non-replicating viral vectors, proteins, and nucleic acids (reviewed in Thanh Le et al. [2020]). Some of these strategies, notably DNA or RNA vaccines, rely on the expression of viral genes in humans. These and other modes of vaccine development (e.g., to produce high protein titers) might benefit from synonymous site modification (Gustafsson et al. 2004; Coleman et al. 2008; Kudla et al. 2009; Fath et al. 2011; Bentele et al. 2013; Mordstein et al. 2020). Coding sequence optimization methods can be directed to modification of codon usage, codon pair usage, nucleotide and dinucleotide content, and other properties of coding sequences, with the aim of achieving a desired phenotype, such as increased gene expression (Gustafsson et al. 2004; Kudla et al. 2009; Fath et al. 2011;

Bentele et al. 2013; Mordstein et al. 2020), improved immunogenicity (Stachyra et al. 2016), or virus attenuation (Coleman et al. 2008).

The DNA and RNA vaccine design methods that might benefit from synonymous site modification fall broadly into two classes: those that aim to detune the live virus (Coleman et al. 2008; Mueller et al. 2010; Manokaran et al. 2019; Cai et al. 2020) and those that aim to enhance expression of individual genes (Stachyra et al. 2014). As with the expression of any transgene, if one viral gene alone is to be expressed in a vector, for example, as part of a DNA vaccine (Stachyra et al. 2014, 2016), then codon optimization of the gene concerned to enable high gene expression may be desirable, not least because such genes can improve immunogenicity (Stachyra et al. 2016), thereby requiring fewer doses (Wang et al. 2006). Such DNA-based vaccines are regarded as relatively safe as no infective form of the virus is required (Khan 2013).

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Mol. Biol. Evol. doi:10.1093/molbev/msaa188 Advance Access publication July 20, 2020

67

Downloaded from <https://academic.oup.com/mbe/article/38/1/67/587368> by University of Bath user on 03 January 2022

Article

Viral attenuation differs from other coding sequence optimization strategies in that it aims to produce gene sequences with low expression levels, with the assumption that this will lead to the production of intact (or near intact) viruses with low pathogenicity, which can nevertheless induce an immune response in the host (Coleman et al. 2008). Synthesis of a complete attenuated virus with detuned synonymous sites can however result in a virus almost unable to replicate (Coleman et al. 2008) and, as such, a mosaic synthetic virus, with some genes deoptimized some not, can be preferable (Coleman et al. 2008). Attenuation via modification of many synonymous sites has the advantage that any such virus employed as a vaccine will likely need many mutations to acquire wild-type fitness. Such a strategy is thus likely to be robust to virus/vaccine intrahost evolution (Coleman et al. 2008), this being reinforced by the relatively low mutation rate of SARS-CoV-2 (about one mutation every 2 weeks, 26.9 per year; Hill and Rambaut 2020; Nextstrain 2020). Synonymous codon manipulation has thus been proposed as a viable strategy for SARS-CoV-2 attenuation and vaccine production (Kames et al. 2020). A live attenuated codon deoptimized vaccine is being attempted by three groups (as of July 14th Report of World Health Organization 2020). For further consideration of development and safety aspects of SARS-CoV-2 vaccines, see Peeples (2020) and Amanat and Kramer (2020).

Viral attenuation can be achieved by alteration of synonymous sites as a means to modify the pattern of dinucleotides that bridge between successive codons (alias codon pair bias) while retaining the original protein (Karlin et al. 1994; Rima and McFerran 1997; Coleman et al. 2008). This codon pair bias attenuation effect has recently been shown to be largely owing to increased CpG content (Tulloch et al. 2014; Gaunt et al. 2016). This is very likely to relate to the activity of the human zinc antiviral protein (ZAP) as this targets transcripts with high CpG content (Takata et al. 2017; Ficarella et al. 2019), although it is by no means the only antiviral protein (supplementary table 1, Supplementary Material online). As might be expected, ZAP is under positive selection owing to host-parasite coevolution (Kerns et al. 2008). This activity of ZAP suggests a simple attenuation strategy for SARS-CoV-2, that is, to increase CpG content (Kames et al. 2020), this being consistent with the observed low CpG enrichment of the virus as sequenced in the wild (Xia 2020), also seen in cytoplasmic viruses more generally (Simmonds et al. 2013). UpA is commonly considered alongside CpG not least because both are underrepresented in native human transcripts (Simmonds et al. 2013) and UpA is cleaved by RNaseL (Odon et al. 2019). Similarly, viruses lacking CpG also tend not to have UpA and engineering increased CpG and UpA attenuates viruses (Simmonds et al. 2013; Odon et al. 2019). UpA depletion in SARS-CoV-2 is weaker than CpG depletion (see below).

Although codon pair bias and dinucleotide composition have been commonly discussed in the context of virus attenuation, these are not the only coding sequence modification strategies that can conceivably produce attenuated viruses. Recently, codon bias (Radhakrishnan et al. 2016; Wu et al.

2019; Buschauer et al. 2020), nucleotide composition (Kudla et al. 2006; Mordstein et al. 2020), and RNA structure (Mauger et al. 2019) have (re-)emerged as important interrelated determinants of gene expression in mammalian cells. Additionally, viral nucleotide and dinucleotide composition have a known role in the immunogenicity of nucleic acids via TLR-7 (Diebold et al. 2004). As a result, understanding forces that operate on synonymous site composition, and on nucleotide content more generally, are central to evolutionarily informed vaccine design, and to our understanding of the biology of SARS-CoV-2. As codon optimization is commonly informed by synonymous site usage in the host genome, we here focus on the relationship between synonymous site selection in the virus and attenuation but are cognizant that lessons learnt may also apply to the optimization problem. Specifically, we aim to discern how selection acts on synonymous sites with a view to engineering the virus against the direction favored by selection on the virus.

One means to test for selection, or more generally forces causing a fixation bias, is to identify a difference between predicted equilibrium nucleotide composition (or dinucleotide composition) under a neutral-mutation bias model and the values observed in the wild. To perform such a test one requires data on the relative rates of different classes of mutations (e.g., A→U and G→C) and from these rates per occurrence of the nucleotide calculate the equilibrium position, that is, the nucleotide content at which the rate of gain by mutation from other residues is equal to the rate of mutational loss. One can then compare observed and neutral equilibrium predicted values, with any discrepancy implicating a fixation bias.

Such methods have revealed commonplace deviations from null neutral expectations. For example, bacteria show a common GC→AT mutational bias (Hershberg and Petrov 2010), and hence a deviation from equilibrium in GC rich bacteria (Hildebrand et al. 2010). Similarly, nonequilibrium TA nucleotide skews (Charneski et al. 2011) have been identified. A recent large survey indicated that G+C deviating from neutral equilibria is also common within both prokaryotes and eukaryotes (Long et al. 2018). To derive this conclusion Lynch and colleagues extracted, from mutation accumulation (MA) experiments or parent-offspring sequencing, mutational profiles for numerous species and showed that the observed G+C content, even at codon third sites, was commonly higher than expected given the profile of mutational events (Long et al. 2018). The cause of this is unresolved, although GC biased gene conversion is one possible explanation (Long et al. 2018).

Rapid, accurate, and common sequencing of epidemic and pandemic pathogens provide a rich source of data from which to derive the mutational profile (Hershberg and Petrov 2010; Hildebrand et al. 2010; Charneski et al. 2011). It is possible to ascribe both ancestral and derived states and hence infer the full mononucleotide mutational matrix (a 4×4, 12 parameter matrix of all possible mutations from one state to another) and, with enough mutations, the full dinucleotide matrix (a 16×16, 240 parameter matrix of all

Table 1. The 4 × 4 Mutational Matrix for 1,151 Mutations at 4-Fold Synonymous Sites (in *italics*) and from 5,482 Mutations Observed Anywhere in Codons (not *italics*).

Reference allele	Derived Allele			
	A	U	C	G
A	—	0.048780.02204	0.016260.01722	0.129270.10067
U	0.024540.01753	—	0.115280.08912	0.012960.01296
C	0.058420.03545	0.541240.40877	—	0.015460.00896
G	0.239130.12389	0.521740.18060	0.050720.02111	—

NOTE.—Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome at 4-fold third sites (*italics*) or in codons. Note that because of different normalizations, the two sets of numbers are not directly comparable in absolute terms.

possible mutations from one dinucleotide to another). Here, then, we apply this method to SARS-CoV-2.

Under the assumption of selection against CpG (Xia 2020), we predict that observed GC content would be lower than the neutral mutational equilibrium GC content. Under the assumption that synonymous sites are neutrally evolving, we expect the predicted equilibrium distribution of the four nucleotides at 4-fold degenerate sites to be no different to that observed. We find in support of neither hypothesis. Our data suggest, unusually, that the most common third site residue (U) is also the one selected against. Given this, we thus propose the unusual strategy of increasing the usage of the already most highly used residue so as to degrade performance of the virus. Given that prior evidence indicated that selection for reduced CpG content is particular to just immediate early genes (Lin et al. 2020), we also propose a “genespoke” approach (i.e., one tailored to each gene’s characteristics) sensitive to both CpG and putative selection on synonymous site U.

Results

SARS-CoV-2 Mutations Are Heavily GC→U Biased

From the 14,599 genomes, we can identify spontaneous mutations. From these, we derive a mutational matrix and from this, we solve for mutational equilibrium. From 1,151 mutations at 4-fold degenerate third sites, we find a heavily GC→AU biased mutational profile (table 1). From this, we deduce that equilibrium GC (termed GC*) should be 17.13% (95% bootstrap estimates 17.09–17.52). The corresponding number is 17.10% on removing six homoplasies. Specifically, we find: U4* = 65.67%; A4* = 17.20%; C4* = 13.09%; G4* = 4.04%. The striking bias toward U has been recently commented on and considered to be consistent with APOBEC editing (Di Giorgio et al. 2020; Simmonds 2020).

Cognizant that there might be dinucleotide-based mutation biases, we extend the mononucleotide matrix to a 16 × 16 dinucleotide matrix with 240 parameter estimates derived across the coding sequences (fig. 1 and supplementary table 6, Supplementary Material online). With 13,209 dinucleotide switches this represents an average of 55.04 mutations per parameter estimate which is liable to be noisy and potentially weakly influenced by selection on nonsynonymous mutations. With this, we determine equilibrium content for all dinucleotides and in turn all nucleotides (A* = 17.66%, C* = 11.65%, U* = 62.42%, G* = 8.27%). We

thus estimate from this GC* of 19.92% (95% bootstraps 19.87–20.05%) more or less in line with mononucleotide calculations.

Evidence for Selection Acting to Counter a Large Mutation Bias toward U

If selection favors reduced G+C content owing to selection for reduced CpG content, we expect that the observed GC3 should be lower than that predicted under neutrality (17.1%). We find the opposite to be true, observed GC3 being 28% (GC3 at 4-fold sites = 20.2%). All numbers are beyond 95% bootstrap bounds of the predicted equilibrium frequency derived from analysis of mononucleotide profiles at 4-fold degenerate sites (bounds: 17.09–17.52). More specifically, at 4-fold synonymous sites, observed U4 (50.8%) is less than predicted under neutral equilibrium U4* (65.7%), whereas all other bases are higher than expected (A4 = 28.95%, A4* = 17.20%; C4 = 13.70%, C4* = 13.09%; G4 = 6.50%, G4* = 4.04%). A parsimonious explanation is that the sizeable mutation bias toward U generates deleterious mutations, nonoptimal even at synonymous sites, and selection therefore favors reduced U content.

GC of coding sequence is even more removed from the neutral equilibrium at 38%. This deviation suggests selection in favor of nonsynonymous mutations that increase G+C content. Examination of nonequilibrium status by dinucleotide content supports this. It shows one striking effect, namely that UUs predicted equilibrium frequency greatly exceeds what is observed (fig. 2). More generally, U content whether derived from mononucleotides at 4-fold third sites (predicted 65.7%) or mononucleotides across the genes (predicted 60.3%) or from dinucleotides (62.4%) is greatly in excess of U content, this being 32% for the complete viral sequence. The mutational matrix, whether through mono- or dinucleotide analysis, predicts a great enrichment of U which we infer is being opposed by selection at third sites and in gene bodies (unweighted gene body means: U1% = 25.7%, U2% = 36.3%, U3% = 41%). We notice that CpG content is above that expected under neutrality (fig. 2). However, this we suggest is not so much evidence against selection toward high CpG so much as selection against UU, which by necessity increases the observed relative frequency of CpG and most other dinucleotides as frequencies must sum to one.

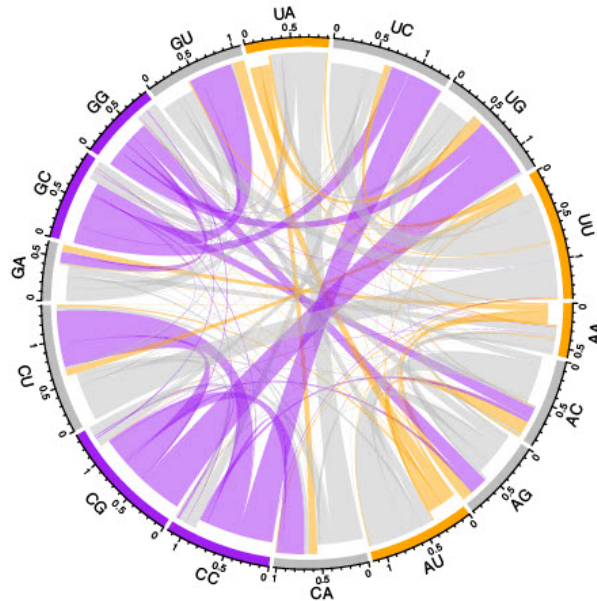


FIG. 1. Chord diagram displaying the rate of flux from one dinucleotide to another in the coding sequence of SARS-CoV-2. For each node, the direction of flux is indicated by the indentation of the connecting links: the outermost layer represents flux into the node and the inner layer represents flux out. The frequency of the flux exchange is represented by the width of any given link where it meets the outer axis. Dinucleotide nodes are colored according to their GC-content. Hence, it is evident that there is high flux away from GC-rich dinucleotides whereas AU-rich dinucleotides are largely conserved.

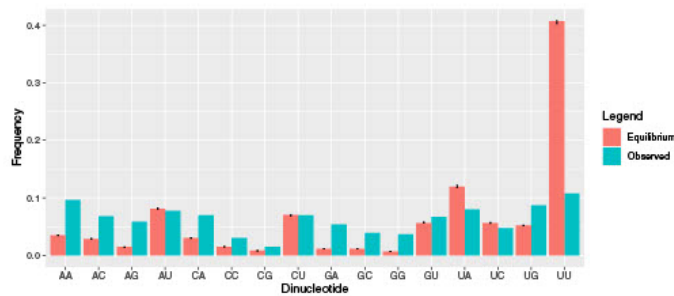


FIG. 2. Comparison of dinucleotide content across SARS-CoV-2 compared with neutral expectations. Error bars represent bootstrapped 95% upper and lower confidence bounds.

Evidence for Contemporaneous Selection against U at Non-4-Fold Redundant Sites

Possibly consistent with a role for selection, using 5,482 mutations that occur anywhere in the coding sequence (table 1), we observe that the G→U flux at 4-fold degenerate sites is

much greater than that observed throughout the sequence. The same is true to a lesser extent for the C→U and A→U fluxes. Assuming the flux rate at 4-fold degenerate sites is more indicative of the true mutational flux, this is consistent with nonsynonymous U mutations being under strong

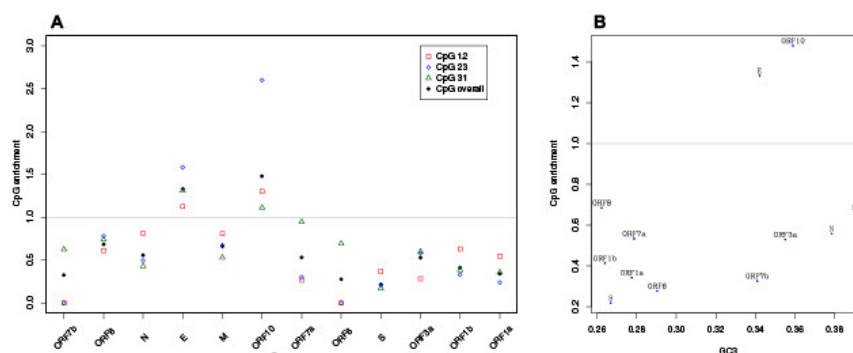


FIG. 3. (a) CpG enrichment across the genes of SARS-CoV-2. Gray line, no enrichment. (b) Relationship between CpG enrichment and GC3.

enough selection to be eliminated prior to sequencing. The predicted bias from this matrix is thus slightly more GC rich than that determined from 4-fold redundant sites (GC^* at 21.37%; 95% bootstrap estimates 21.39–21.60, 21.44% also after excluding homoplasies).

The lower occurrence of mutations generating U at non-4-fold sites would be consistent with contemporary selection on non-4-fold sites opposing mutations toward U, consistent also with the difference between $U4^*$, $U4$, and U content overall. To ask whether the difference between the two equilibria solutions is significantly different, we developed a non-parametric Monte Carlo simulation (see Materials and Methods). We find that the Euclidean distances from the random sampling are the same as, or greater than, the Euclidean distance between 4-folds and non-4-fold sites in just 323/10,000 cases (hence $P = 0.0323$) (repeating using an alternative distance metric, sum of modular difference between equilibria, makes no meaningful difference, $P = 0.0454$). To clarify that it was selection against U, we considered each nucleotide individually (see Materials and Methods). Such analysis indeed provides evidence for significant counter selection of U at non-4-fold sites ($Un4^* = 60.8$, $U4^* = 64.6\%$, $Z = -1.98$). Commensurably, predicted G equilibrium content derived from mutations at non-4-fold sites is higher than that derived from mutations at 4-fold degenerate sites ($Z = 5.34$), whereas A and C content are less affected (Z for A = 0.26, Z for C = -0.56). Thus, not only do we detect deviation away from the predicted neutral equilibrium (at 4-fold sites, third sites generally and through the gene body), we also can detect a signal consistent with selection on SARS-CoV-2 that skews the mutational matrix prior to sequencing.

Significant Heterogeneity in the Degree of CpG Avoidance between Genes

Although selection against U or UU provides a viable model for $GC3 > GC3^*$, might there be other explanations that would be consistent with selection against CpG, to avoid ZAP, but in favor of G+C? One possibility is that we may

be witnessing between-gene heterogeneity (Digard et al. 2020). Imagine that some genes are indeed under selection for low CpG and hence for low GC3, but others are not under selection for low CpG and thus are more free to have selection favoring higher GC3 (for unspecified reasons, but possibly to enable efficient expression; Mordstein et al. 2020). When then considered en masse, we see both selection for CpG and more raised GC3. Recent reports suggest that not all genes are equally subject to selection for low CpG to avoid ZAP, with only “immediate early” genes under such selection (Lin et al. 2020).

Were this the explanation, or part thereof, we would predict that CpG enrichment would be heterogeneous between genes (see also, Digard et al. [2020]) and that those with relatively high CpG enrichment will also be those genes contributing to raised GC3 (i.e., a positive correlation between CpG enrichment and GC3). Note that, although CpG counts are likely to be necessarily higher as GC3 goes up, CpG enrichment is normalized to underlying GC content and so CpG enrichment and high GC3 are not logically coupled (e.g., if at the limit 50% of residues are C and 50% G, so long as CpG usage is random, $CpG = 0.5 \times 0.5$, CpG enrichment will not be seen).

To assay this, we calculated CpG enrichment at codon sites 12, 23, and 31, these providing three measures of CpG enrichment for each gene. We can then perform a Kruskal–Wallis test (KW) for heterogeneity. Even with such limited data, we find that the three measures for the same gene are more similar than expected by chance (KW, $P = 0.019$, $df = 11$: mean $E(CG) = 0.61 \pm 0.4$ SD; fig. 3a). This implies that at all sites CpG is avoided or preferred to the same degree within any given gene. We see however only marginal evidence that genes released from CpG constraint are those with higher GC3 (CpG enrichment vs. GC3, $\rho = 0.41$, $P = 0.19$, Spearman’s test, fig. 3b). Thus, although there is evidence for differential CpG usage between genes, we do not find that this predicts GC3, although trends are in the expected direction and the tests underpowered.

71

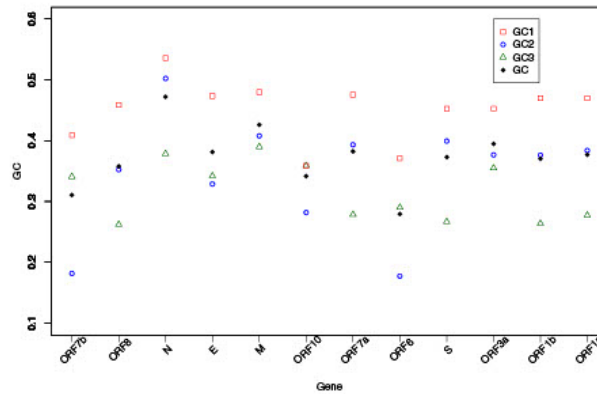


Fig. 4. GC content across genes of SARS-CoV-2 at codon sites 1, 2, 3, and averaged across the gene.

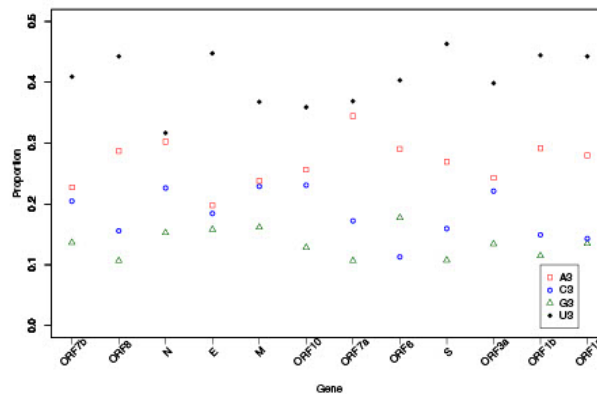


Fig. 5. Base composition at codon third sites across genes of SARS-CoV-2.

More generally, we can ask whether gene body G+C content behaves the same as gene body CpG content with each gene having its own characteristic profile. We assay this by considering GC1, GC2, and GC3 in a manner as above. We find no evidence that genes are more similar in these three measures than expected by chance (KW, $P=0.49$, $df=11$: fig. 4). Similarly, we see no correlation between GC3 and GC12 although the trend is positive ($\rho=0.15$, $P=0.63$, Spearman's rank) (see also, Dilucca et al. [2020]). However, we do observe some regularities. First, GC3 is consistently lower than GC12 (Wilcoxon signed-rank test, $P=0.007$), the mean GC3 being 28%, whereas that of GC12 is 40%, consistent with selection on amino acid content.

The most striking feature of third site nucleotide usage is that all genes have a preponderance of U (fig. 5). As noted

above, this we can attribute only in some part to mutation as the predicted levels, whereas in the rank order as observed (U>A>C>G) are highly deviant from null. Specifically, the predicted numbers are $0.66 > 0.17 > 0.13 > 0.04$, whereas the observed are $0.44 > 0.28 > 0.16 > 0.13$. Approximately, the same predicted equilibrium values are seen employing all mutations ($0.60 > 0.18 > 0.13 > 0.08$). Selection against U seems strong, despite this being the most common nucleotide, as it is heavily reduced from its predicted equilibrium content.

Genes Avoiding CpG Also Avoid UpA

Prior analysis suggests that viruses lacking CpG also tend not to have UpA and that engineering increased CpG and UpA attenuates viruses, possibly because both are

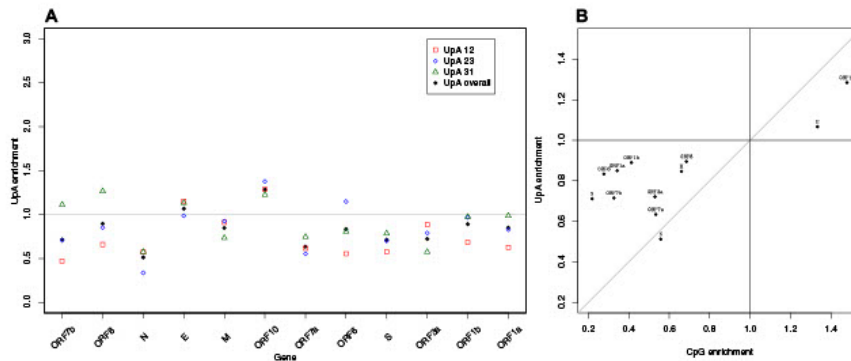


Fig. 6. (a) UpA enrichment across genes of SARS-CoV-2 and (b) correlation with CpG enrichment. Gray line is the line of slope 1 through the origin.

Table 2. Between-Gene Correlations in Dinucleotide Enrichment Scores (Pearson product moment correlation *r* values).

	UpAe	ApUe	CpGe	GpCe
UpAe	—	0.20	<i>0.76**</i>	−0.18
ApUe		—	−0.15	−0.16
CpGe			—	<i>0.007</i>
GpCe				—

Note.—Significant correlations in italics:
***P* < 0.005.

underrepresented in human transcripts (Simmonds et al. 2013). We also observe that UpA enrichment and CpG enrichment tend to positively correlate across viruses ($N = 1,290$, $\rho = 0.165$, $P = 2.68 \times 10^{-9}$; data in supplementary table 2, Supplementary Material online). To understand whether increasing CpG and UpA might be a useful attenuation strategy, we ask whether UpA is also avoided in genes of SARS-CoV-2 and whether it is avoided in the same genes that avoid CpG. We consider not just the CpG enrichment predicting UpA enrichment but also, to control for mononucleotide effects, the two other symmetric nucleotide pairings (ApU and GpC).

On an average, UpA is, like CpG, avoided although not to the same extent as CpG (mean UpA enrichment = 0.83 ± 0.2 SD) (fig. 6b). UpA also shows between gene heterogeneity (KW, $P = 0.04$). We find that exclusively for CpG enrichment and UpA enrichment do we see a correlation between genes (table 2 and fig. 6a). ApU is also avoided (mean enrichment = 0.83 ± 0.14 SD), but there is no evidence for within gene homogeneity (KW test $P = 0.14$) (fig. 7). By contrast, there is no evidence for GpC avoidance (mean GpC enrichment = 1.13 ± 0.34 SD) (fig. 7) and genes do not show gene-specific GpC enrichment (KW, $P = 0.11$, comparing GpC enrichment at sites 12, 23, and 31). We conclude that if CpG enrichment is a viable strategy to attenuate a gene, increasing UpA may also.

Evidence for U Content Predicting Expression Level

The results above are consistent with a model in which CpG content is under selection in some genes to be reduced, whereas GC3 content is above the level expected under neutrality, in no small part because the U mutation bias is so extreme that equilibrium U content (especially UU content) would render the virus much less fit. There are several possible mechanistic explanations for the GC3>GC3* effect. With our recent evidence that intronless low GC genes are barely expressed in human cell lines (Mordstein et al. 2020), selection for raised GC3 (reduced U3) to enable more effective gene expression is a strong contender. In this context, whereas we do not see a GC3 expression correlation ($r = 0.09$, $P = 0.82$), we do observe a GC expression correlation ($r = 0.79$, $P = 0.01$ and fig. 8). Breaking this down by nucleotide, we see that this is owing to a negative correlation with U content and a positive correlation with both C and G content (A freq: $r = 0.33$, $P = 0.83$; C freq: $r = 0.64$, $P = 0.06$; G freq: $r = 0.81$, $P = 0.009$; U freq: $r = -0.88$, $P = 0.0017$). Why this will require considerable experimental manipulation of sequences to understand but we note a correlation between expression level and predicted per nucleotide stability (Pearson's $r = -0.86$, $P = 0.0027$, $df = 7$). It is notable that we observe such an effect with such an underpowered test.

A more broad-brush approach is to consider viral sequences more generally (supplementary table 2, Supplementary Material online). As part of the mechanism by which GC enrichment boosts expression is thought to be intranuclear (e.g., nuclear export) (Mordstein et al. 2020), if selection is operating on gene expression of viruses, we might predict that nuclear viruses might have a higher GC content than cytoplasmic viruses. Using mean GC of all viruses within a taxonomic grouping, we observe this to be the case (Mann-Whitney *U* test $P < 2.2 \times 10^{-16}$, fig. 9). CpG enrichment and UpA enrichment is similarly lower in cytoplasmic viruses (fig. 9). This is a very arms-length result and requires due caution in its interpretation (it could just as well be evidence

Downloaded from https://academic.oup.com/mbe/article/38/1/67/5878882 by University of Bath user on 03 January 2022

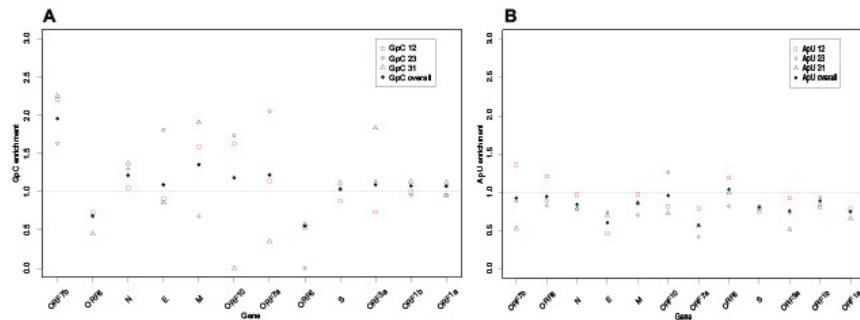


Fig. 7. (a) GpC and (b) ApU enrichment across the genes of SARS-CoV-2.

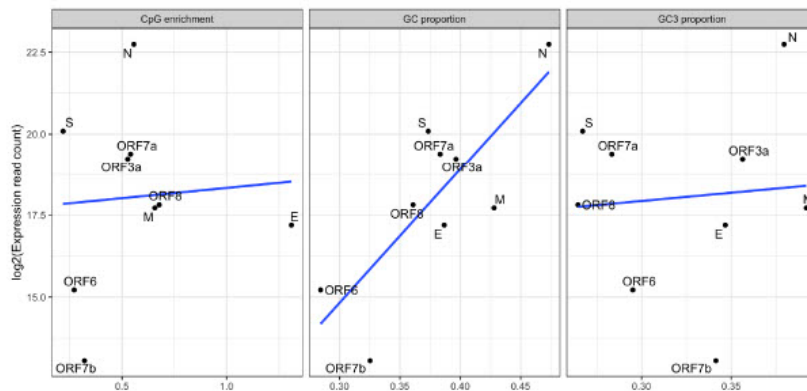


Fig. 8. Correlation between expression level and CpG enrichment, GC content, and GC3.

of different mutational biases). Nonetheless, within the context of our prior result, we suggest that this merits further scrutiny. There is some evidence that if selection might favor reduced CpG content it might also favor reduced UpA content, as, within both groups, those viruses with low CpG enrichment also tend to have low UpA enrichment, but the effect is weak (Spearman's test, cytoplasmic viruses: $\rho = 0.096$, $P = 0.016$; nuclear viruses: $\rho = 0.084$, $P = 0.031$).

Designing the Optimally Attenuated SARS-CoV-2

With the above evidence for selection for G+C at third sites and for heterogeneity between genes in enrichment of CpG and UpA, we suggest that simply increasing CpG by manipulation of synonymous sites need not be the optimal strategy. It may enable recognition by ZAP, but may also favor increased fitness by increasing G+C/reducing U.

As not all genes are under selection for reduced CpG/UpA, reducing their G+C content by increasing U content seems a relatively safe and robust strategy. We thus suggest to classify genes according to the CpG enrichment (>1 or <1). For those in the first category, likely not affected by ZAP (E and ORF10, see also, Digard et al. [2020]), we suggest decreasing their synonymous G+C by increasing where possible U content and forcing them closer to their mutational equilibrium. For those with especially low CpG enrichment and most likely strong targets of ZAP (ORF1a, ORF1b, ORF6, ORF7b, and S), we suggest, raising their CpG, even at the cost of increased G+C. Where possible UpA should also be increased. For the remainder, we suggest increasing CpG content while holding GC3 content static or decreasing if possible. However, with the possibility of synonymous sites also being parts of key motifs, for example, for RNA-binding proteins (Savisaar and Hurst 2017), a simplistic strategy, even if gene-tailored, may

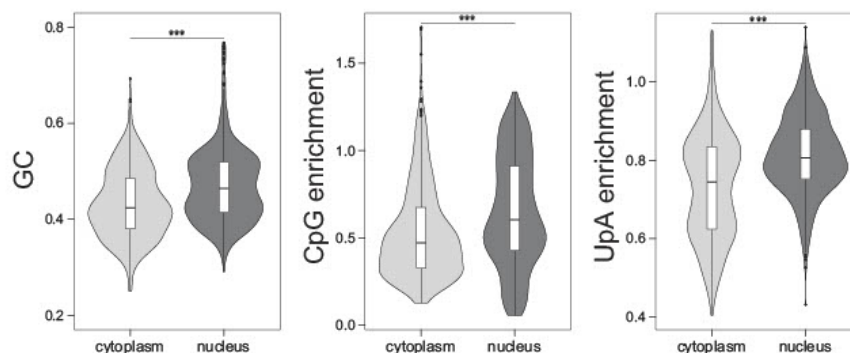


FIG. 9. GC content of cytoplasmic and nuclear viruses. Cytoplasmic viruses have significantly lower values for all three measures (Mann–Whitney U test: GC: $P = 6.42e-18$, CpG enrichment: $P = 1.35e-13$, UpA enrichment: $P = 9.1e-29$).

have deleterious undesirable side consequences. Unlike alternative strategies that permute existing codons (Jorge et al. 2015), the proposed strategy (supplementary table 7, Supplementary Material online) enables deviations in overall nucleotide content. Recognizing, however, that the extreme nucleotide content can cause gene inactivation, as with prior strategies (Jorge et al. 2015), we propose a stochastic methodology to derive a large number of variants modified in the desired direction that could be experimentally tested (for algorithm, see supplement table 7, Supplementary Material online; for variants for each gene, see supplement table 8, Supplementary Material online). We suggest variants for each gene recognizing the most effective attenuated construct may be a mosaic of wild-type and attenuated genes (Coleman et al. 2008).

GC3* > GC3 Is Not a General Property of Viruses

We observed that GC content at third sites was both higher than expected given selection against CpG and higher than expected given the underlying mutational profile. Is a deviation from mutational equilibrium a general property of human viruses? Were this so, this too could have implications for engineering of attenuated forms. To address this, we consider other viruses with rich sequencing from epidemics.

For H1N1 using the same mode of analysis, we observe both a less extreme GC→AT mutation bias (table 3) and an observed GC3 content very close to that predicted. From analysis of third sites, the predicted value is GC3* = 41.8% (bootstrap 95% intervals 41.45–42.04), from all sites the prediction is GC* = 42.8% (bootstrap 95% 42.56–42.96). The observed GC3 is 41.8%, within the bounds of the prediction based on third site mutations. For Ebola (table 4), we find observed GC3 is all but identical to predicted (observed GC3 = 46.4%, expected = 46.7%). We conclude that analysis of SARS-CoV-2 and its nonmutational equilibrium status at synonymous sites does not necessarily hold lessons for other viruses. In contrast to others (Kames et al. 2020), we suggest caution in generalizing vaccine strategies.

Discussion

Mutation bias across all taxa is typically GC→AT biased (Hershberg and Petrov 2010; Hildebrand et al. 2010; Liu et al. 2018) and neutral predicted equilibrium frequencies below GC of 20% (as observed here) are not without precedent (see, e.g., Long et al. [2018]). Broadly the U enrichment at third sites within the genome is then compatible with a large role for mutation bias, possibly mediated by members of the APOBEC gene family (Di Giorgio et al. 2020; Simmonds 2020), known mutators of viruses (Lee et al. 2008) with C→U and UC→UU preferences (Chen and MacCarthy 2017). However, we have shown that nucleotide usage, although skewed in the direction imposed by mutation bias, is nonetheless deviant from it. The difference between observed and expected U3 and UU (fig. 2) proportions are noteworthy. At 4-fold degenerate sites, although C and G usage are close to equilibrium, A is far above and U is far below ($U_4 = 50.8\%$, $U_4^* = 65.67\%$; $A_4 = 28.95\%$, $A_4^* = 17.20\%$; $C_4 = 13.70\%$, $C_4^* = 13.09\%$; $G_4 = 6.50\%$, $G_4^* = 4.04\%$). We propose that a parsimonious explanation is that the sizeable mutation bias toward U generates deleterious mutations, even at synonymous sites, and selection therefore favors reduced U content. However, increasing C or G potentially comes at a cost of increased CpG, so the base most in excess of its equilibrium is A. As a consequence, although CpG avoidance is real in some genes, GC3 is a little higher than predicted from the underlying mutational profile. This thus presents an unusual case in which the most common synonymous codons (those ending in U) are not the selectively advantageous ones.

We have not directly addressed the problem of the causes of any such selection on synonymous mutations. Given G+C preference in human coding genes to enable effective expression (Kudva et al. 2006; Mordstein et al. 2020), the negative correlation between U usage and expression is broadly consistent with evidence for preferential degradation of transcripts with nonoptimal codon usage (Radhakrishnan et al. 2016; Buschauer et al. 2020). Potentially in tandem to such

Table 3. The 4×4 Mutational Matrix for 1,522 Mutations at Synonymous Sites (in italics) and from 2,571 Mutations Observed Anywhere in Codons (not italics) for H1N1.

Reference Allele	Derived Allele			
	A	U	C	G
A	—	0.08710.04597	0.0650.0451	0.42910.25542
U	0.08030.05143	—	0.49450.24889	0.05290.03429
C	0.16910.11426	0.56990.30675	—	0.02510.02607
G	0.60890.32052	0.09480.05027	0.03230.0207	—

NOTE.—Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome at third sites (italics) or in codons.

Table 4. The 4×4 Mutational Matrix for 1,682 Mutations at Synonymous Sites (in italics) and from 3,523 Mutations Observed Anywhere in Codons (not italics) for Ebola.

Reference Allele	Derived Allele			
	A	U	C	G
A	—	0.07390.05077	0.09640.06722	0.21230.14803
U	0.05940.05152	—	0.21450.13429	0.05360.04786
C	0.08450.08086	0.26390.14868	—	0.03940.04845
G	0.26390.16051	0.07510.05139	0.06940.05139	—

NOTE.—Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome at third sites (italics) or in codons.

possible effects high U content may trigger immunogenicity of nucleic acids via TLR-7 (Diebold et al. 2004). Whether a virus with a few more U residues is importantly more immunogenic is, however, uncertain. Alternatively, effects may be mediated by changes in mRNA secondary structure (Mauger et al. 2019). We indeed observe a correlation between expression level and predicted per nucleotide stability. Given this, it could be speculated that RNA stability may explain the thermal intolerance of the virus (Demongeot et al. 2020), although many other mechanisms are imaginable.

On a related note, aside from the possible influence of APOBEC generating the excess of UU mutations, we have not considered the causes of the very different rates for each class of mutation. Indeed, for the high G→U rate, we know of no editing process that has this profile (see supplementary table 1, Supplementary Material online). However, we speculate that this might be owing to oxidation of guanosine that can lead to a G to U transversion. This process may be accelerated by NO via its oxidate species ONOO⁻ (Yermilov et al. 1995; Juedes and Wogan 1996), the former being produced primarily by inducible NO synthase. This enzyme is upregulated in many cells including inflammatory phagocytic cells including macrophages, mediated by proinflammatory cytokines including IFN γ (Zhuang et al. 1998). This has known effects on viral mutation (for review, see Akaïke and Maeda [2000]). It may also be informative to consider the relationship between RNA secondary structure and these mutation biases (Krishnan et al. 2004), although overall proportion of variable 4-fold redundant sites does not covary with stability (Pearson's $r = -0.06$, $P = 0.88$, $df = 7$).

There are, however, at least four problems with our mode of analysis. First, a theoretical alternative explanation for the difference between predicted and observed values is that the virus was at neutral mutational equilibrium in its prior host

(cf. H1N1, Ebola), but since the transfer to humans, the mutational profile has altered. Were this so we may just have identified a lag in viral evolution from one neutral equilibrium to another. In this context, deviation from equilibrium has little if anything to say about either selection or optimal vaccine design. Although evidence for GC→AT biased mutation in related viruses (Simmonds 2020) renders this less parsimonious an explanation, direct examination of mutational profiles of the virus in its ancestral host (whatever that may be) would be valuable. The evidence for subtly but significantly different mutational matrices dependent on the class of site employed provides more direct evidence for contemporary selection on U content throughout gene bodies that cannot be accounted for by a temporal shift in mutational profile.

Second, assuming no change to the mutational matrix, *sensu stricto*, we have observed a force that would cause a fixation bias (Lercher et al. 2002). Evidence for such a force need not necessarily indicate the direction of selection, as selection bias is only one class of fixation bias. In biased gene conversion, for example, the mismatch repair machinery recognizes, during double-strand break repair, heteroduplex GCAT mismatches and corrects these in favor of GC residues (Brown and Jiricny 1988). This causes a meiotic drive like process in which deleterious mutations can be driven to higher frequencies (for further consideration, see Hurst [2019]). Given that single-strand RNA cytoplasmic viruses, such as SARS-CoV-2, are unlikely to be exposed to the nuclear mismatch repair machinery or need double-strand break repair, biased gene conversion is unlikely to explain GC3>GC3* and U3<U3*. We cannot with our data, however, rule out unknown mechanisms causing similar nonselective fixation biases. It is then valuable to provide more direct evidence for an advantageous effect of reduced U3/increased GC3, as suggested by our preliminary analysis on expression level

Experimental manipulation of GC3 content (cf. Kudla et al. 2006; Mordstein et al. 2020) is a high priority.

Third, we have presumed that the mutational spectrum observed at 4-fold degenerate sites is a good reflection of the true mutational profile. Often when applying methodology like this, we presume that the temporal proximity between occurrence and observation of mutations is so small that there has been no time for selection to filter in a manner that distorts the mutational matrix. Nonetheless, we found that although slight, there is a difference between the mutational profile observed at CDS sites that are not 4-fold degenerate and those that are. Although this difference is so slight it cannot explain why U is so deviant from equilibrium levels, and does not question our overall findings, we do nonetheless presume that the 4-fold site matrix itself is unbiased. For strains sequenced hours to days apart to be biased at 4-fold degenerate sites would require strong and biased selection at 4-fold redundant sites. Although not obviously plausible, we have no means to disprove this (and strong selection, albeit associated with splicing, has been identified at synonymous sites in human genes; Savaisaar and Hurst 2018). Nonetheless, any such bias would also force the mutational matrix observed to predict a nucleotide content more closely resembling the observed nucleotide content, rendering the test conservative. Derivation of the mutational profile by *in vivo* analysis (cf. Denison et al. 2011) could enable more direct tests of our findings. Analysis of SARS-CoV (responsible for the 2002 SARS outbreak) with exonuclease activity (which we presume to mimic SARS-CoV-2, it having a nsp14 homolog of the SARS-CoV exonuclease; Pachetti et al. 2020) reports a massive AU→GC bias with eight of 11 reported mutations being in this direction and only two GC→AU (Smith et al. 2013). This implies either a radically different mutation bias in SARS-CoV than in SARS-CoV-2 or great sensitivity of results to experimental conditions, such as cell lines employed and APOBEC activity. We note that the SARS-CoV analysis employed Vero cells in which the interferon response is disabled (Smith et al. 2013), thus likely to have neither ZAP (see, e.g., MacDonald et al. [2007]) nor APOBEC activity (see, e.g., Peng et al. [2006]).

Fourth, we have presumed that, after filtering (see Materials and Methods), all sequences are error free. Although sequencing errors cannot explain a bias as strong as the difference between excess and expected UU or U3, nor can they obviously explain the evidence for contemporary selection against U, it may possibly explain the small difference between predicted and observed nucleotide content at 4-fold sites for G and C (the deviations of A and U from predicted equilibria are relatively large). One suggested means to avoid this is to only employ mutations that have been sequenced more than once (Hildebrand et al. 2010). However, this has been shown to introduce its own bias (Charneski et al. 2011). Using high-quality sequence, it was shown that using mutations that appear once and those that appear twice or more makes a significant difference to the matrix and estimates of equilibria (Charneski et al. 2011). The cause of this is likely to be a selection filter: mutations that persist longer to be sequenced twice or more will be skewed toward

milder effect mutations. This accords with our observation of a slight difference between matrices that restrict just to 4-fold degenerate sites and those that do not. The ideal then is to filter not by regularity of appearance but by sequencing quality (hence our decisions on which sequences to employ: see Materials and Methods). Nonetheless, to err on the side of caution, we considered mutations at 4-fold degenerate third sites that appear more than once (i.e., excluding singletons) and found that GC* is now even lower than previously predicted (GC* = 10.3%, 95% bounds 10.19–10.61). Thus, we are confident that we can exclude sequencing error as an explanation for observed GC3>GC3* and U4<U4* (singleton excluded prediction of U4* = 65.7%). Nonetheless, owing to observation bias and low sample size, we caution against overinterpretation of this result. Given possible biases owing to sequencing platform, we also ask about the expected equilibrium content for Illumina and Nanopore sequencing separately. We find that the predicted equilibrium vectors for 4-fold degenerate sites are no different from each other ($P = 0.62$).

Assuming we have identified the direction of selection (against U, against CpG in some genes) this can inform vaccine design. Unusually, even though U is the most common nucleotide at third sites (by a considerable margin), we propose increasing this even more thereby forcing the viruses against the direction of purifying selection. We predict that raising CpG in the genes that are CpG deficient would be a viable strategy even at a cost of raising GC3/lowering U. By contrast for those few genes with $E(\text{CpG}) > 1$ (i.e., gene E, ORF10, see also, Digard et al. [2020]) CpG manipulation increasing GC3 would be a dangerous strategy, potentially achieving little more than an increase in expression. Increasing their U content would appear to be the antiselection direction. We note however that ORF10s function, if any, remains unclear there being no evidence of transcripts from it, despite it looking like a well-formed ORF (starts ATG stops TAG, multiple of three long). Its GC3 content is also far from neutral equilibrium (GC3 = 36%). In this context, gene E may be a good one to alter synonymous site usage as it appears not to be under selection for CpG or UpA avoidance.

Genes ORF1a, ORF1b, ORF6, ORF7b, and S are good candidates for the raising of CpG content. Gene N is noteworthy in being very highly expressed, long (1,260 bp), GC rich (GC3 = 38%), and with moderate CpG enrichment ($E(\text{CpG}) = 0.56$). Given these characteristics it should be possible to increase CpG by manipulating some third sites (those with C at codon position 2 or G at codon position +1) while reducing GC and increasing U content at other sites. For smaller genes, there is less leeway. In this context S, ORF1a and ORF1b are also very strong candidates being long, with moderate GC3 and low CpG enrichment. A more detailed description of the algorithm for attenuation alongside attenuated variants can be found in [supplementary tables 7 and 8, Supplementary Material](#) online. Although the particular strategy for attenuation reflects the particulars of selection operating on SARS-CoV-2, the more general notion of evolutionarily informed vaccine design, with attenuation achieved by synthesizing variants rich in the compositional

features opposed by selection, is worthy of experimental scrutiny.

Materials and Methods

Gene Locations

We employed NC_045512 to specify the gene sequence to determine observed GC content and CpG content. However, following further annotation of genes (Kim et al. 2020), we modified the gene locations to reflect those specified <https://github.com/hyeshik/sars-cov-2-transcriptome/blob/master/reference/SARS-CoV-2-annotations.gff>. Specifically, to avoid a small codon overlap, we exclude the overlap hence employed annotation:

ORF7a protein 27394.27759→27394.27753

ORF7b protein 27756.27887→27762.27887

To consider ORF1a and ORF1b independently and to avoid overlap, we employ:

ORF1a→266-13465

ORF1b→13471-21552

Estimating Flux Rates from Data

As with parent–offspring sequencing and MA lines, to estimate neutral equilibrium nucleotide content, we require that the mutations observed are an unbiased sample of the mutational profile (Hildebrand et al. 2010; Long et al. 2018). With very common sequencing (in all cases, short-time periods between ancestor and progeny), we can ignore the possibility of multiple sequential hits at the same site (with the first hits going unsequenced) contaminating the mutational matrix. In principle, the method can be misled by strong selection purging, in a nonrandom fashion, mutations prior to their appearing in the population. However, if most selection is weak purifying selection there is then a lag between a deleterious mutation appearing (and being sequenced) and it being purged from a population. Declines in Ka/Ks as time to common ancestry increases in closely related bacteria strains (Rocha et al. 2006) is consistent with such a model. In principle, even if there is strong selection on some mutations this too need not be problematic, so long as strong selection only affects the observed rate of appearance in sequencing data of new mutations but not the relative proportions of the different mutational classes (C→G, A→U, etc.). Moreover, if selection does act in a biased manner it should force the predicted equilibrium to more closely resemble the observed nucleotide content, rendering the test conservative. To be cautious, however, we focus on segregating mutations at 4-fold degenerate synonymous sites as the closest approximation to the underlying mutational profile.

In total, 15,721 SARS-CoV-2 genome assemblies available on 12 May, 2020 were downloaded from the GISAID (Shu and McCauley 2017) Initiative EpiCoV platform. Only assemblies flagged as “complete (>29,000 bp),” “high coverage only,” and from a human isolate were downloaded. Isolates with > 1% of ambiguous base calls (rounded to 298 bases) were removed, leaving 14,855 genomes. Sequences were aligned with MAFFT 7.458 (Katoh and Standley 2013) to Wuhan-Hu-1 reference genome (EPI_ISL_402124). EPI_ISL_402124 was collected

from a retailer at Huanan Seafood Wholesale Market, Wuhan on December 30, 2019. We employed this sequence as not only was it an early sequence but it also matches the consensus generated from all the 19 sequences that were collected prior to December 31. Variant sites were obtained from the MSA using the package SNP-sites (Page et al. 2016) and whole-genome nucleotide flux estimates were obtained by counting the frequency of each type of mutation with respect to the reference genome. Each given mutation at any given site was counted once, regardless of its frequency within the population. Our method should be insensitive to the presence of recombination, not that there is any evidence that SARS-CoV-2 has recombined through its pandemic phase (Wang et al. 2020). For consideration of homoplasies (independent mutations at the same site), see below.

Isolates containing at least one coding sequence of length not divisible by three were excluded, removing 58 strains, resulting in a set of 14,599 sequences. CDSs were then translated using BioPython, realigned using MAFFT, and then reverse translated using TranslatorX (Abascal et al. 2010). MSA of CDSs were concatenated and then, just as with the whole-genome analysis, variant sites were obtained using SNP-sites and flux estimates were obtained by counting the frequency of each type of change with respect to the reference.

Additionally, H1N1 influenza A pdm09 sequences for strains collected between January 2009 and August 2010 that contained segments PB2, PB1, PA, HA, NP, NA, MP, and NS were obtained from GISAID (Shu and McCauley 2017) for four segments: RNA polymerase subunit (PB2), hemagglutinin (HA), nucleoprotein (NP), and neuraminidase (NA). Sequences with length not divisible by three or containing a stop codon when translated were excluded. Remaining sequences were translated by BioPython and aligned to Mexican strain EPI_ISL_66702 using MAFFT, and reverse translated to nucleotides using TranslatorX (Abascal et al. 2010).

Multiple sequence alignment of 1,610 full Ebola virus (EBOV) genomes sampled between March 17, 2014 and October 24, 2015 in West Africa was downloaded from EboIaD database (Carneiro and Pereira 2016). The alignment includes the reference genome NC_002549.1. Genomes with a proportion of > 10% missing sites were discarded. CDSs for each strain were obtained by extracting the coordinates from the reference genome on the alignment. In order to include in the analysis as the largest proportion of the gene ZEBOVgp4, the longest CDS (NP_066246.1) was used, and the shorter, overlapping proteins NP_066247.1 and NP_066248.1 were discarded. Just as in the case of H1N1, sequences with length not divisible by three were excluded. Remaining sequences were translated aligned to the reference strain using MAFFT, and reverse translated to nucleotides using TranslatorX (Abascal et al. 2010).

Estimating Equilibria

In principle, one can estimate neutral GC equilibria knowing relative rates of GC→AT and AT→GC mutations alone (Long et al. 2018). However, we take a fuller approach to estimate the equilibrium content of all nucleotides that

also enables us to capture nucleotide skews (Charneski et al. 2011). This has the advantage of treating all four bases as separate independent states, as is fitting for a single-stranded virus unconstrained by Chargaff's first parity rule (Elsom and Chargaff 1952). Let us denote the frequency of G as G and the frequency of U and C. We shall write that the mutational frequency of G to U will be $g2u$, these being measured per occurrence of the starting base. The frequency of the nucleotides after some period (N) will then be:

$$G' = G (1 - g2u - c2c - g2a) + A (a2g) + U (u2g) + C (c2g)$$

$$C' = C (1 - c2u - c2g - c2a) + A (a2c) + U (u2c) + G (g2c)$$

$$A' = A (1 - a2u - a2c - a2g) + G (g2a) + U (u2a) + C (c2a)$$

$$U' = U (1 - u2g - u2c - u2a) + A (a2u) + G (g2u) + C (c2u)$$

We then solve such that $G' = G$ and $U' = U$. This thus resolves to:

$$G (g2u + g2c + g2a) = A (a2g) + U (u2g) + C (c2g)$$

$$C (c2u + c2g + c2a) = A (a2c) + U (u2c) + G (g2c)$$

$$A (a2u + a2c + a2g) = G (g2a) + U (u2a) + C (c2a)$$

$$U (u2g + u2c + u2a) = A (a2u) + G (g2u) + C (c2u)$$

Note that the left hand of each equation is the rate of loss given current abundance, whereas the right is the rate of gain given current abundances (i.e., we are solving for gain = loss). The 12 flux parameters ($a2u$ and $a2c$) we derive from the mutational profile these being the number of observed changes per relevant occurrence of the nucleotide in the ancestral (premutated) sequence. We then solve these four simultaneous equations. Note that, we replace any one arbitrarily chosen frequency by 1—sum of the other three (e.g., $U = 1 - A - C - G$). These were solved in NumPy. Equilibrium solutions we denote with an asterisk (e.g., G^* and $GC3^*$). $N4^*$ implies nucleotide content of nucleotide N at 4-fold degenerate sites.

To assign bounds on the equilibrium estimates, we perform a bootstrap test in which we resample with replacement M mutations from the set of M mutations. For each sampled vector, we recalculate the predicted equilibria thereby

assigning bounds. We report 95% bootstrap bounds from 100 resamplings.

The same approach applies to the 16×16 dinucleotide matrix with 240 parameters.

Comparing Mutational Matrices

We sought to test whether the predicted equilibria solutions were different between the matrices reflecting mutational profiles at 4-fold degenerate sites and all mutations at other sites (i.e., not 4-fold degenerate), as might be predicted were there contemporaneous selection against mutations that are nonsynonymous. We partitioned all CDS mutations into those at 4-fold redundant sites ($n = 1,151$) and all others ($n = 5,482$). Using these two data sets, we calculated observed equilibrium frequencies for each nucleotide (4^* for 4-folds and $n4^*$ for non-4-folds), representing each as a vector of length four. We then determined the Euclidean distance between the two vectors. To test for significance, we compare the magnitude of this Euclidean distance to that expected by chance employing a nonparametric Monte Carlo simulation. To this end, we randomly extracted without replacement 1,151 mutations from the full set of mutations so as to create a subsample of pseudo "4-folds." The remaining 5,482 mutations we then considered a sample of pseudo "non-4-fold" mutations. For each randomization, we assembled the corresponding mutational matrix, solved for equilibria, and calculated the Euclidean distance between the resulting vectors of predicted equilibrium for the four nucleotides. We repeated this procedure 10,000 times to generate a null distribution of Euclidean distances that controls for sample size differences. Significance was given as $P = n/m$, where n is the number of simulations in which the Euclidean distance is as great or greater than observed in the real data and m is the number of simulations (i.e., 10,000). To check for robustness, we considered an alternative distance metric, namely sum of modular differences (Euclidean distance considers square root of sum of squares of difference).

To consider each nucleotide individually, from the same Monte Carlo sampling, we calculated the difference between predicted equilibria at sampled pseudo "4-folds" and pseudo "non-4-folds" for the 10,000 repeats. This generates four distributions, one for each nucleotide. For each nucleotide, we calculate the mean (~ 0) and SD of these randomizations. The observed difference seen for each nucleotide between the equilibria predicted using mutations at 4-fold sites (their predicted neutral equilibria) compared with that calculated using mutations at non-4-fold site, may then be represented as a Z score ($Z = (\text{observed} - \text{mean of simulations}) / \text{SD of simulations}$), $Z > |1.96|$ indicating significant deviation.

Homoplasy Screen in SARS-CoV-2

Sites can appear as having independently occurring mutations for at least two reasons: the extra mutation may be a sequencing error or it may be a true homoplasy (i.e., the same mutation at the same site occurring more than once independently) (van Dorp et al. 2020). Sequencing errors need to be removed. Knowing how to handle true homoplasies in the

construction of a mutational matrix is not as conceptually simple.

At first sight one might suggest that, as independent mutations, each occurrence of the mutation should be considered. The key question, however, is whether the mutational profile at these sites is representative of activity at other sites. If it is not, then their over inclusion will bias the matrix toward the profile of homoplastic sites away from that of the rest of the genome, which could itself cause a false signal of nonequilibrium status (i.e., where mutationally predicted and observed nucleotide compositions—largely at nonhomoplastic sites—disagree). A priori by virtue of the fact that they are homoplastic we might suppose that mutational activity at these sites is not reflective of the mutational profile elsewhere in the genome and it is the equilibrium properties of other sites that we are interested in. Equally, these may well be sites that are more likely to be under selection (van Dorp et al. 2020) and hence, again, not necessarily reflective of the mutational process. One could then opt to filter out mutations at homoplastic sites considering them possibly unrepresentative. However, we do not know they are unrepresentative and so their removal may be depleting the analysis of information. We also do not know how many of the nonmutated sites had had the property of being homoplastic prior to current sequencing. An alternative, the middle way, is to include them but count all occurrences at any given site as one event, thereby employing the mutations but preventing such sites from overly skewing the matrix and further reducing the impact of possible (missed) sequencing errors.

For analysis of SARS-CoV-2, we opt for the latter “middle way” approach but also check for resilience by removing such sites. Fortunately, as such sites are so rare (6 of 1,151 4-fold degenerate sites), removal of these sites makes no important difference to calculation of GC equilibrium content, nor to estimation of observed nucleotide content. We thus report the homoplastic-excluded results as minor asides.

Phylogenetic tree of 11,204 SARS-CoV-2 isolates was downloaded from the COVID-19 Genomics UK Consortium website (<https://www.cogconsortium.uk/>, version of April 24, 2020). Subsequently, the MSA and the resulting tree were used to identify recurrent mutations (homoplasies) using HomoplasmyFinder (Crispell et al. 2019). All ambiguous sites in the alignment were set to “N.” Sites in the first and last 200bp of the genome alignment were masked to account for the fact that a higher degree of spurious variants that can appear homoplastic tend to locate at the ends of the multiple sequence alignment.

HomoplasmyFinder identified 408 putative homoplasies that were distributed over the SARS-CoV-2 genome. Homoplasies can occur as a result of convergent evolution, recombination, or due to artifacts such as specific combinations of sample preparation, sequencing technology, consensus calling approaches, and sequencing errors. In order to remove spurious homoplastic sites, a particular worry of this data set because a mix of technologies and methods have been employed by different contributing research groups, these were filtered using a set of parameters and thresholds defined in (van Dorp et al. 2020) to obtain a set of high-confidence

homoplasies. Briefly, for each homoplasmy, the proportion of isolates with the homoplasmy where the nearest neighboring isolate in the phylogeny also carried the homoplasmy (pnn) was computed and all homoplasies with pnn < 0.1 were excluded. Furthermore, we also excluded homoplasies that were shared in < 0.1% of the isolates (> 11 isolates). We also required that no isolate had an ambiguous base near the homoplasies (± 5 bp). These filters reduced the number of homoplastic sites to 67. The predicted equilibrium frequency of the four nucleotides at all the homoplastic sites (440 mutations), counting each class of mutation only once, is not different from that at nonhomoplastic sites (Euclidean distance method: $P = 0.61$). The filtered (accepted) homoplasies are also not significantly different from nonhomoplastic mutations (Euclidean distance method: $P = 0.63$) or from the rejected ones (Euclidean distance method: $P = 0.41$). We conclude that the accepted set, with each mutation counted once, presents a defensible balance between inclusion and stringency.

Estimating Dinucleotide Enrichment

For the dinucleotide NpM (e.g., CpG and CpC), we define gene body enrichment ($E(NM)$) as:

$$E(NM) = p(NM) / [p(N) \times p(M)],$$

where $p(NM)$ is the frequency of all dinucleotides within the gene that are NM and $p(N)$ and $p(M)$ are the frequencies of the mononucleotides within the same gene. We then consider site-specific enrichment, that is sites 12, 23, or 31 defined by codon position, 31 being a third site and the codon first site of the following codon. Then at sites xy:

$$E(NM_{xy}) = p(NM_{xy}) / [p(N_x) \times p(M_y)],$$

where NM_{xy} is the relevant dinucleotide initiating with N at site x.

Gene Expression

We employ expression data specified by Kim et al. (2020). We used the highest read count for each subgenomic RNAs in Supplementary Table 3 of Kim et al. (2020) and compared log₂ normalized read counts to gene G+C content, G+C at third sites, and CpG enrichment. As the authors employed nanopore sequencing, read count does not obviously require gene length normalization. Note that subgenomic RNA measures exclude ORF1a and ORF1b. ORF10 is excluded as no reads were identified. We used the shapiro.test function in R to test log₂-transformed read counts for normality.

RNA Stability

The minimum free energy of mRNA secondary structure was calculated for entire SARS-CoV-2 coding sequences (as defined in Gene Locations), using the hybrid-ss-min (UNAFold) program version 3.8 (Markham and Zuker 2008), with default settings (NA = RNA, $t = 37$). The folding energy of each sequence was then divided by the length of the corresponding sequence, to obtain the per nucleotide mRNA stability measure that was used in downstream calculations.

Data Compilation of Vertebrate Viruses

Vertebrate virus sequences were retrieved from the Virosaurus database (Virosaurus databases 2020_4.1, Release April 2020, file: Virosaurus90v_2020_4.1) (Gleizes et al. 2020) (accessed May 7, 2020). In this database, complete sequences were clustered at 90% to remove redundancy. As in this database, herpesviridae and poxviridae are split in genes rather than full genomes, complete sequences for these viruses were retrieved from NCBI RefSeq database (Pruitt et al. 2005). The same was also done for segmented viruses to allow calculation of sequence parameters per species. Genome classification was retrieved from ICTV Virus Metadata Repository, version May 1, 2020; MSL35 (Walker et al. 2019). Annotation for replication compartments was assigned according to ICTV (Walker et al. 2019) and ViralZone (Hulo et al. 2011). CpG and UpA enrichment were calculated as above. For virus sequences obtained from the Virosaurus database, the mean was derived to obtain one value per species. For segmented viruses, segments were first concatenated before calculating sequence parameters. Species information and sequence parameters can be found in [supplementary table 2, Supplementary Material](#) online.

Genome Sources

We acknowledge the sources of the genomes that we employed in [supplementary table 3](#) (for SARS-CoV-2), [supplementary table 4](#) (for H1N1), and [supplementary table 5](#) (for Ebola), [Supplementary Material](#) online.

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the Wellcome Trust (fellowship 207507 to G.K.) and the European Research Council (Advanced Grant No. ERC-2014-ADG 669207 to L.D.H.).

References

- Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38(Suppl 2):W7–W13.
- Akaike T, Maeda H. 2000. Nitric oxide and virus infection. *Immunology* 101(3):300–308.
- Amanat F, Krammer F. 2020. SARS-CoV-2 vaccines status report. *Immunity* 52(4):583–589.
- Bentle K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol.* 9(1):675.
- Brown TC, Jiricny J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney-cells. *Cell* 54(5):705–711.
- Buschauer R, Matsuo Y, Sugiyama T, Chen YH, Alhusaini N, Sweet T, Ikeuchi K, Cheng J, Matsuki Y, Nobuta R, et al. 2020. The Ccr4-Not complex monitors the translating ribosome for codon optimality. *Science* 368(6488):eaay6912.
- Cai Y, Ye C, Cheng B, Nogales A, Iwasaki M, Yu S, Cooper K, Liu DX, Hart R, Adams R, et al. 2020. A Lassa fever live-attenuated vaccine based on codon deoptimization of the viral glycoprotein gene. *Mbio* 11(1):e00039–00020.

- Carneiro J, Pereira F. 2016. EbolaID: an online database of informative genomic regions for Ebola identification and treatment. *PLoS Negl Trop Dis.* 10(7):e0004757.
- Charneski CA, Honfi F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical at skew in firmicute genomes results from selection and not from mutation. *PLoS Genet.* 7(9):e1002283.
- Chen J, MacCarthy T. 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. *PLoS Comput Biol.* 13(3):e1005471.
- Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008. Virus attenuation by genome-scale changes in codon pair bias. *Science* 320(5884):1784–1787.
- Crispell J, Balaz D, Gordon SV. 2019. HomoplasmyFinder: a simple tool to identify homoplasies on a phylogeny. *Microb Genom.* 5:e000245.
- Demongeot J, Flet-Berliac Y, Seligmann H. 2020. Temperature decreases spread parameters of the new Covid-19 case dynamics. *Biology* 9(5):94.
- Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. 2011. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* 8(2):270–279.
- Di Giorgio S, Martignano F, Tordia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv.* 6(25):eabb5813.
- Diebold SS, Kaisho T, Hemmi H, Akira S, Reis e Sousa C. 2004. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science* 303(5663):1529–1531.
- Digard P, Lee H-M, Sharp C, Grey F, Gaunt ER. 2020. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. *bioRxiv* 2020.2005.2008.083816.
- Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. 2020. Codon usage and phenotypic divergences of SARS-CoV-2 genes. *Viruses* 12(5):498.
- Elson D, Chargaff E. 1952. On the desoxyribonucleic acid content of sea urchin gametes. *Experientia* 8(4):143–145.
- Fath S, Bauer AP, Liss M, Spriestersbach A, Maertens B, Hahn P, Ludwig C, Schafer F, Graf M, Wagner R. 2011. Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. *PLoS One* 6(3):e17596.
- Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, Neil SD, Schulz R, Swanson CM. 2019. CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-dependent and -independent mechanisms. *J Virol.* 94(6):e01337–01319.CrossRef
- Gaunt E, Wise HM, Zhang H, Lee LN, Adkinson NJ, Nicol MQ, Highton AJ, Klenerman P, Beard PM, Dutia BM, et al. 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. *Elife* 5:e12735.
- Genomic Epidemiology of Novel Coronavirus—Global Subsampling [Internet]. 2020. Available from: <https://nextstrain.org/ncov/global?i=clock>.
- Gleizes A, Laubscher F, Guex N, Iseli C, Junier T, Cordey S, Fellay J, Xenarios I, Kaiser L, and Le Mercier P. 2020. Virosaurus [Internet]. Available from: <https://viralzone.expasy.org/8676>.
- Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol.* 22(7):346–353.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9):e1001115.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9):e1001107.
- Hill V, Rambaut A. 2020. Phylogenetic analysis of SARS-CoV-2 [Update 2020-03-06]. Available from: <https://virological.org/t/phylogenetic-analysis-of-sars-cov-2-update-2020-03-06/420>.
- Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, Le Mercier P. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* 39(Suppl 1):D576–D582.
- Hurst LD. 2019. A century of bias in genetics and evolution. *Heredity (Edinb)* 123(1):33–43.

- Jorge DM, Mills RE, Lauring AS. 2015. CodonShuffle: a tool for generating and analyzing synonymously mutated sequences. *Virus Evol.* 1(1):vex012.
- Juedes MJ, Wogan GN. 1996. Peroxynitrite-induced mutation spectra of pSP189 following replication in bacteria and in human cells. *Mutat Res.* 349(1):51–61.
- Kames J, Holcomb DD, Kimchi O, DiCuccio M, Hamasaki-Katagiri N, Wang T, Komar AA, Alexaki A, Kimchi-Sarfaty C. 2020. Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. *bioRxiv*. 2020.2003.2030.016832.
- Karlin S, Doerfler W, Cardon LR. 1994. Why is Cpg suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses. *J Virol.* 68(5):2889–2897.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kerns JA, Emerman M, Malik HS. 2008. Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. *PLoS Genet.* 4(1):e21.
- Khan KH. 2013. DNA vaccines: roles against diseases. *Germs* 3(1):26–35.
- Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* 181(4):914–921 e910.
- Krishnan NM, Seligmann H, Raina SZ, Pollock DD. 2004. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA Cell Biol.* 23(10):707–714.
- Kudla G, Lipinski L, Caffin F, Helwak A, Zylcz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4(6):e180.
- Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324(5924):255–258.
- Lee YN, Malim MH, Bieniasz PD. 2008. Hypermutation of an ancient human retrovirus by APOBEC3G. *J Virol.* 82(17):8762–8770.
- Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* 162(4):1805–1810.
- Lin Y-T, Chiwehse S, McCormick D, Raper A, Wickenhagen A, DeFillipis V, Gaunt E, Simmonds P, Wilson SJ, Grey F. 2020. Human cytomegalovirus evades ZAP detection by suppressing CpG dinucleotides in the major immediate early genes. *bioRxiv*: 2020.2001.2007.897132.
- Liu HX, Huang J, Sun XG, Li J, Hu YW, Yu LY, Liu GN, Tian DC, Hurst LD, Yang SH. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat Ecol Evol.* 2(1):164–173.
- Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.* 2(2):237–240.
- MacDonald MR, Machlin ES, Albin OR, Levy DE. 2007. The zinc finger antiviral protein acts synergistically with an interferon-induced factor for maximal activity against alphaviruses. *J Virol.* 81(24):13509–13518.
- Manokaran G, Sujatmoko M, Pherson KG, Simmons CP. 2019. Attenuation of a dengue virus replicon by codon deoptimization of nonstructural genes. *Vaccine* 37:2857–2863.
- Markham NR, Zuker M. 2008. UNAFold software for nucleic acid folding and hybridization. *Methods Mol Biol.* 453:3–31.
- Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, Link K, Khatwani N, Reyniers J, Moore MJ, et al. 2019. mRNA structure regulates protein expression through changes in functional half-life. *Proc Natl Acad Sci U S A.* 116(48):24075–24083.
- Mordstein C, Savaisaar R, Young RS, Bazile J, Talmame L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2020. Codon usage and splicing jointly influence mRNA localization. *Cell Syst.* 10(4):351–362 e358.
- Mueller S, Coleman JR, Papamichail D, Ward CB, Nimnual A, Fletcher B, Skiena S, Wimmer E. 2010. Live attenuated influenza virus vaccines by computer-aided rational design. *Nat Biotechnol.* 28(7):723–726.
- Odon V, Fros J, Goonawardane N, Dietrich I, Ibrahim A, Alshalkhahmed K, Nguyen D, Simmonds P. 2019. The role of ZAP and OAS3/RNASEL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. *Nucleic Acids Res.* 47(15):8061–8083.
- Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angdetti S, Cocozzi M, Gallo RC, et al. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med.* 18(1):179.
- Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom.* 2(4):e000056.
- Peebles L. 2020. News feature avoiding pitfalls in the pursuit of a COVID-19 vaccine. *Proc Natl Acad Sci U S A.* 117(15):8218–8221.
- Peng G, Lei KJ, Jin W, Greenwell-Wild T, Wahl SM. 2006. Induction of APOBEC3 family proteins: a defensive maneuver underlying interferon-induced anti-HIV-1 activity. *J Exp Med.* 203(1):41–46.
- Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33(Database issue):D501–D504.
- Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Collier J. 2016. The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* 167(1):122–132 e129.
- Rima BK, McFerran NV. 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *J Gen Virol.* 78(11):2859–2870.
- Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Fell EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239(2):226–235.
- Savaisaar R, Hurst LD. 2017. Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. *Mol Biol Evol.* 34(5):1110–1126.
- Savaisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* 28(10):1442–1454.
- Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* 22:30494.
- Simmonds P. 2020. Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. *msphere* 5, 10.1128/msphere.00408-20.
- Simmonds P, Xia WJ, Baillie JK, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phylogenetic selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics* 14(1):610.
- Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. 2013. Coronaviruses lacking exonucleotidase activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. *PLoS Pathog.* 9(8):e1003565.
- Stachyra A, Gora-Sochacka A, Sirko A. 2014. DNA vaccines against influenza. *Acta Biochim Pol.* 61(3):515–522.
- Stachyra A, Redkiewicz P, Kosson P, Protasiuk A, Gora-Sochacka A, Kudla G, Sirko A. 2016. Codon optimization of antigen coding sequences improves the immune potential of DNA vaccines against avian influenza virus H5N1 in mice and chickens. *Viral J.* 13(1):143.
- Takata MA, Goncalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* 550(7674):124–127.
- Wang H, Kosakovsky Pond SL, Nekrutenko A and Nielsen R. 2020. Testing Recombination in the Pandemic SARS-CoV-2 Strains [Internet]. 2020. [updated 2020 Jun 17]. Available from: .
- Thanh Le T, Andreadakis Z, Kumar A, Gomez Roman R, Tollefsen S, Saville M, Mayhew S. 2020. The COVID-19 vaccine development landscape. *Nat Rev Drug Discov.* 19(5):305–306.
- Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. 2014. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *Elife* 3:e04531.
- van Dorp L, Acan M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol.* 83:104351.

- Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, Harradh B, Harrison RL, Hendrickson RC, Junglen S, et al. 2019. Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). *Arch Virol*. 164(9):2417–2429.
- Wang S, Taaffe J, Parker C, Solórzano A, Cao H, García-Sastre A, Lu S. 2006. Hemagglutinin (HA) proteins from H1 and H3 serotypes of influenza A viruses require different antigen designs for the induction of optimal protective antibody responses as studied by codon-optimized HA DNA vaccines. *J Virol*. 80(23):11628–11637.
- World Health Organization DRAFT Landscape of COVID-19 Candidate Vaccines—14 July 2020 [Internet]. 2020. Available from: <https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines>
- Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, Wright M, Bazzini AA. 2019. Translation affects mRNA stability in a codon-dependent manner in human cells. *Elife* 8:10–7554.
- Xia X. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Mol Biol Evol*. Advance Access published April 14, 2020, doi:10.1093/molbev/msaa094.
- Yermilov V, Rubio J, Ohshima H. 1995. Formation of 8-nitroguanine in DNA treated with peroxydinitrite in-vitro and its rapid removal from DNA by depurination. *FEBS Lett*. 376(3):207–210.
- Zhuang JC, Lin C, Lin D, Wogan GN. 1998. Mutagenesis associated with nitric oxide production in macrophages. *Proc Natl Acad Sci U S A*. 95(14):8286–8291.

Supplementary information for: Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design

Alan M. Rice, Atahualpa Castillo Morales, Alexander T. Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, Laurence D. Hurst

Molecular Biology & Evolution, 38(1): 67–83.

Supplementary data are available at Molecular Biology and Evolution online: <https://doi.org/10.1093/molbev/msaa188>.

Appendix 2

Causes and Consequences of Purifying Selection on SARS-CoV-2

Atahualpa Castillo Morales, Alan M. Rice, Alexander T. Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, Laurence D. Hurst

Genome Biology & Evolution, 13(10): evab196.

This chapter contains work published on 24th August 2021 at GBE, the original and sole place of publication. It thus contains analysis of publicly available data using bespoke scripts that are freely available at the locations cited within the paper. The paper is open access and I have permission as the author to include the article in full (https://academic.oup.com/journals/pages/authors/production_and_publication/online_licensing). The latest version of the published article can be found by following the address: <https://doi.org/10.1093/gbe/evab196>.

Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:				
Causes and Consequences of Purifying Selection on SARS-CoV-2				
Publication status (tick one)				
Draft manuscript <input type="checkbox"/> Submitted <input type="checkbox"/> In review <input type="checkbox"/> Accepted <input type="checkbox"/> Published <input checked="" type="checkbox"/>				
Publication details (reference)	Morales AC, Rice AM, Ho AT, Mordstein C, Muhlhausen S, Watson S, Cano L, Young B, Kudla G, Hurst LD. 2020. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol. Biol. Evol. 13(10): evab196.			
Copyright status (tick the appropriate statement)				
I hold the copyright for this material <input type="checkbox"/> Copyright is retained by the publisher, but I have been given permission to replicate the material here <input checked="" type="checkbox"/>				
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the... Formulation of ideas: 10% Design of methodology: 10% Bioinformatic analyses: 20% Experimental work: N/a Presentation of data in journal format: 10%			
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.			
Signed	<table border="1" style="width: 100%;"> <tr> <td style="width: 70%;"></td> <td style="width: 10%;">Date</td> <td style="width: 20%;">03/03/2022</td> </tr> </table>		Date	03/03/2022
	Date	03/03/2022		

Causes and Consequences of Purifying Selection on SARS-CoV-2

Atahualpa Castillo Morales^{1,†}, Alan M. Rice^{1,†}, Alexander T. Ho^{1,†}, Christine Mordstein^{1,2,3}, Stefanie Mühlhausen¹, Samir Watson³, Laura Cano², Bethan Young^{1,2}, Grzegorz Kudla^{2,‡}, and Laurence D. Hurst^{1,*,‡}

¹The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, United Kingdom

²MRC Human Genetics Unit, Institute for Genetics and Molecular Medicine, The University of Edinburgh, United Kingdom

³Department of Molecular Biology and Genetics, Aarhus University, Denmark

[†]These authors contributed equally to this work.

[‡]Co-senior authors.

*Corresponding author: E-mail: l.d.hurst@bath.ac.uk

Accepted: 19 August 2021

Abstract

Owing to a lag between a deleterious mutation's appearance and its selective removal, gold-standard methods for mutation rate estimation assume no meaningful loss of mutations between parents and offspring. Indeed, from analysis of closely related lineages, in SARS-CoV-2, the Ka/Ks ratio was previously estimated as 1.008, suggesting no within-host selection. By contrast, we find a higher number of observed SNPs at 4-fold degenerate sites than elsewhere and, allowing for the virus's complex mutational and compositional biases, estimate that the mutation rate is at least 49–67% higher than would be estimated based on the rate of appearance of variants in sampled genomes. Given the high Ka/Ks one might assume that the majority of such intrahost selection is the purging of nonsense mutations. However, we estimate that selection against nonsense mutations accounts for only ~10% of all the "missing" mutations. Instead, classical protein-level selective filters (against chemically disparate amino acids and those predicted to disrupt protein functionality) account for many missing mutations. It is less obvious why for an intracellular parasite, amino acid cost parameters, notably amino acid decay rate, is also significant. Perhaps most surprisingly, we also find evidence for real-time selection against synonymous mutations that move codon usage away from that of humans. We conclude that there is common intrahost selection on SARS-CoV-2 that acts on nonsense, missense, and possibly synonymous mutations. This has implications for methods of mutation rate estimation, for determining times to common ancestry and the potential for intrahost evolution including vaccine escape.

Key words: SARS-CoV-2, mutation rate, purifying selection, codon usage.

Significance

In SARS-CoV-2, we find evidence for common intrahost purifying selection against nonsense, missense, and synonymous mutations, such that the true underlying mutation rate is about 50% higher than would be estimated if one assumes that the mutation rate is the rate of appearance of mutations in the circulating population. This has implications for methods to determine mutation rates, for determining times to common ancestry and the potential for vaccine escape.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Classically purifying selection can be inferred by absence. For example, in the Ka/Ks test, we employ the normalized rate of occurrence of substitutions at synonymous sites (Ks) in a protein coding gene as a measure of the background rate of evolution, comparing this to the normalized rate of nonsynonymous changes (Li et al. 1985; Goldman and Yang 1994). A dearth of the latter compared with the former (Ka/Ks < 1) is taken to imply that protein changing mutations happened but were too deleterious to persist (Li et al. 1985; Goldman and Yang 1994). The method thus implicitly infers the rate of what might be called “missing” mutations.

A consequence of this is that, owing to a lag between mutation appearance and selective removal (Rocha et al. 2006), our ability to resolve purifying selection on recently diverged lineages is weak, few mutations being “missing” (Ponting 2008). Indeed, for this reason, for closely related species Ka/Ks in a pairwise analysis declines as the time to common ancestry increases (Rocha et al. 2006). Consequently, we know relatively little about the activity of purifying selection over the short term (Ponting 2008), let alone what might be called “real time.” Similarly, to estimate the mutation rate (meaning the rate at which new mutations happen, not the rate of lineage evolution), we employ a few generations of mutation accumulation lines (Lynch et al. 2016) under the assumption that the rate of accumulation of changes in DNA/RNA is the mutation rate, as purifying selection is both diminished and will not yet have influenced the fate of mutations. Indeed, parent–offspring trios are now considered a gold standard for mutation rate estimation as such analyses are presumed to be the least affected by the missing mutation problem (Yang et al. 2015).

An ideal examination of real-time selection in the wild would require analysis of massive numbers of full genomes of a relatively fast evolving species sampled continuously in time and place. Such a natural experiment is currently running. Indeed, the volume of genome data for SARS-CoV-2 allows an unparalleled evaluation of the activity of purifying selection in real time. Early analysis, however, suggested that purifying selection was not detectable, Ka/Ks being almost exactly 1 (Bai et al. 2020), that is, there is no distortion from the immediate mutational profile, consistent with assumptions of parent–offspring mutation rate estimation. More recent evidence, by contrast, indicates that such selection is detectable (Dearlove et al. 2020; Shen et al. 2020; Tang et al. 2020; Tonkin-Hill et al. 2021; Lythgoe et al. 2021). Similarly, mutational scanning experiments indicate positions under positive and negative selective constraints in the SARS-CoV-2 receptor-binding domain (Starr et al. 2020).

There are numerous reasons why the study of real-time purifying selection in SARS-CoV-2 in particular might be interesting. For example, the difference between the rate of appearance of new mutations in the population and the

rate at which they actually occur, is indicative of the potential for intrahost evolution. If, for example, there is little disparity (e.g., Ka/Ks = 1) then intrahost selection is not occurring and the nonsynonymous mutations that occur are being transmitted without selection. Conversely, if only a small proportion of actual mutations survive to be transmitted, the adaptive potential, for example, for selection for vaccine escape, must be quite high, there being differential birth and death (i.e., intrahost variance in fitness with the viral clone). Similarly, if we infer the evolutionary rate of a virus by assaying the rate at which RNA changes appear in the population (Duchene et al. 2020; Hill and Rambaut 2020; Nextstrain 2020) and, in turn, assume this to reflect the true underlying rate (much as done with parent–offspring sequencing), then the true underlying rate is likely to be underestimated. Although not necessarily important for inferring the evolutionary rate, allowance for such purifying selection can affect estimation of time to common ancestry (Wertheim and Kosakovsky Pond 2011). Here then, we attempt to estimate the proportion of mutations that occurred but were missing prior to sequencing of circulating variants. From this, we attempt in turn to infer the true mutation rate, more particularly asking whether this is a sizeable correction or not. That Ka/Ks ~ 1, might suggest that no meaningful correction is needed.

Further, the profile of these missing mutations may also contain information as to what selection is acting on. Selection against most nonsense mutations seems inevitable. Indeed, it is possible both that there is purifying selection operating against nonsense mutations and that Ka/Ks = 1, as the later metric does not factor in nonsense mutations. We should then predict fewer nonsense mutations circulating within the sequenced genomes than expected given the underlying mutational profile. Prior sampling of intraindividual variation supports this (Tonkin-Hill et al. 2021), although sequence quality issues may be relevant here (see Nekrutenko [2020]). Indeed, for reasons unknown (see Nekrutenko [2020]), one commonly employed intrahost sequencing project (SRP253798) reports both remarkably high numbers of mutations and that almost all such mutations are C->U. This has the potential to overestimate the rate of generation of nonsense mutations. Given that Ka/Ks (that considers only missense/nonsynonymous changes) is near unity (Bai et al. 2020), one might then suggest that, despite evidence for purifying selection against some missense (nonsynonymous) variants (Dearlove et al. 2020; Lythgoe et al. 2021), the vast majority of purifying selection must be against nonsense mutations. Here, we attempt to assay whether this is so.

We find that there is common purifying selection operating at the protein level (i.e., against nonsynonymous variants). We then ask whether the profile of selection against nonsynonymous variation seen in more distant comparisons can be detected in real time. Classically nonsynonymous mutations are selected against when they disrupt protein function too much. This can be reflected in a dearth of fixed (between two

different species) differences that see an amino acid replaced by one that is chemically very different (Weber and Whelan 2019). We ask whether we can detect such selection operating within hosts. In addition, we might expect, at a higher level of granularity, that a biophysical model of protein functioning might predict which amino acid exchanges are tolerated. We consider spike protein as an exemplar, not least because the model for this protein was not informed by evolutionary constraint data (which would render any analysis circular).

Analyses of longer-term purifying selection suggests that mutations to more biosynthetically costly amino acids are also subject to purifying selection (Richmond 1970; Akashi and Gojobori 2002; Heizer et al. 2006; Hurst et al. 2006; Swire 2007; Chameski et al. 2011). In contrast to the above predictors, we do not necessarily expect this to be detectable, in real time or otherwise, in a virus which may itself not suffer the costs of amino acid synthesis, the ATP costs of amino acid biosynthesis being more obviously suffered by the host than the virus. One might, however, conjecture that what is good for the host might also be good for the virus (fitness covariance) and, as translation imposes the majority of the cost of building a virus, such costs may be under selection (Mahmoudabadi et al. 2017). Indeed, virus-like Gene Transfer Elements integrated in *Alphaproteobacteria* have been suggested to be under positive selection for the reduction of cost (Kogay et al. 2020). However, an integrated element is expected to have stronger fitness covariance with its host than SARS-CoV-2 for whom the host is just a temporary transmission vehicle.

Perhaps the weakest selection we might hope to detect is that of synonymous mutations. Although selection on synonymous sites is likely to be hard to detect, prior evidence suggests viruses might adapt their codon usage to that of their host (Hernandez-Alias et al. 2021), to optimize translational efficiency (Wong et al. 2010; Liu et al. 2011; Fan et al. 2015; Chen et al. 2020; Hernandez-Alias et al. 2021) or avoid certain nucleotide combinations (Shpaer and Mullins 1990; Atkinson et al. 2014; Gaunt et al. 2016; Gu et al. 2019). Some evidence for selection of codon usage in SARS-CoV-2 has been reported (Gu et al. 2020; De Maio et al. 2021; Hernandez-Alias et al. 2021). Our prior analysis reveals that predicted neutral mutational equilibrium content of U at 4-fold degenerate sites (U4*) at 65% is higher than the observed U4, which could indicate purifying selection on U mutations at 4-fold degenerate sites but could also reflect a relatively recent change in mutational profile and lag to mutational equilibrium (Rice et al. 2021).

Here then, in addition to estimating the number of missing mutations, we examine nonsense, missense, and synonymous mutations to test particular hypotheses for the causes of such selection. Although the genomic resources are exceptional, SARS-CoV-2 analysis presents unusual methodological challenges. Site frequency spectra (SFS) approaches have been

applied in an attempt to infer selection on nucleotide composition in SARS-CoV-2 (De Maio et al. 2021). However, broader application of such methods may well be problematic as some methods are advised against in nonrecombining genomes (Bustamante et al. 2001) and inferences can be confounded by effects of demography that can mimic selection. Indeed, SFS methods are more commonly employed to determine demography (Lapierre et al. 2016), analyses that in turn are confounded by their failure to allow for weak selection (Lapierre et al. 2016). Moreover, highly geographically skewed sequencing efforts, including intensive sequencing around outbreak hotspots, will distort the SFS (e.g., a rare mutation in an oversequenced location will appear to be at a relative high net frequency).

Ka/Ks has also been applied to test for selection on SARS-CoV-2 (see, e.g., Bai et al. [2020]). Aside from the fact that the test was designed to be applied to fixed between-species differences (Mugal et al. 2020), this test too has numerous interrelated issues. First, it overlooks nonsense mutations as a source of "missing" mutations. Second, even the best codon-centered models (Goldman and Yang 1994; Wertheim and Kosakovsky Pond 2011) ignore complex mutational effects that bridge between codons, forcing codon pair bias, that is important for viral functioning (Coleman et al. 2008). Third, and related, SARS-CoV-2 has an exceptionally biased and complex mutational profile (Rice et al. 2021; Simmonds 2020b; Graudenzi et al. 2021), with a large bias toward U, especially from CU and GU dinucleotides, that is likely to confound estimation methods. Coupled with differential nucleotide usage at different codon positions, this is likely to interfere with estimation. For example, although one could estimate the true mutation rate by using the rate at 4-fold degenerate codon sites alone (cf. Keightley and Eyre-Walker 2000), as these are much more U biased than the other codon sites (Rice et al. 2021), the rate at 4-fold degenerate sites will not reflect the underlying rate at the other sites, potentially underestimating it as U has a low mutation rate (Rice et al. 2021). Compounded with a short time between mutational occurrence and sampling, these issues may explain why prior Ka/Ks estimation reports a value of 1.008, indicative of no purifying selection (Bai et al. 2020).

To overcome these problems, we apply a variety of methods. Most notably, we estimate rates at 4-fold sites of different nucleotide compositions and use these nucleotide-dependent rates to infer the true underlying mutation rate and hence the rate of missing mutations, given the nucleotide content of all other sites. Similarly, to determine the profile of missing mutations, we define expectations of the rates of amino acids exchanges under a complex null neutral model and examine the predictors of deviations from this. Using related methods, we also attempt to infer the direction of selection on synonymous mutations. These methods have an advantage over direct within-host sampling that they can also estimate rate of mutations so deleterious that they never attain reasonable frequencies within the host. They should

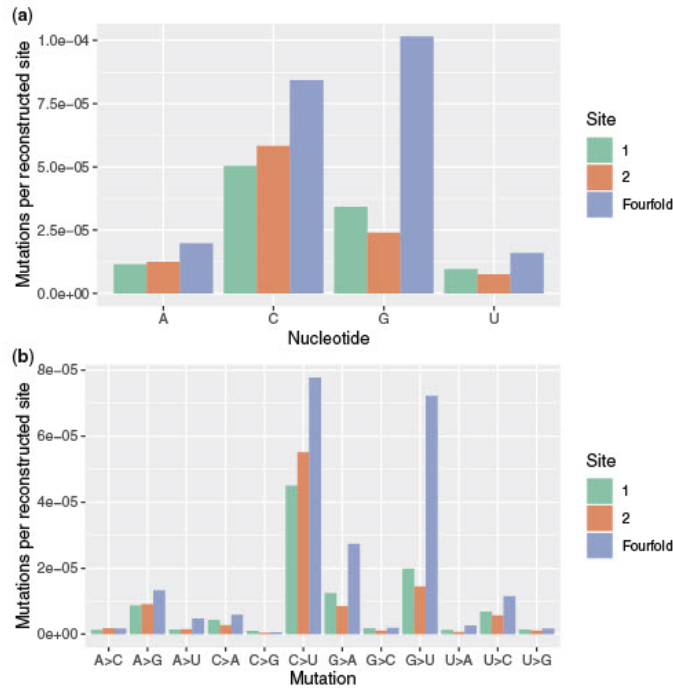


Fig. 1.—Comparisons between 4-fold and codon site 1 and site 2 mutations. (A) Rate of observed mutation per reconstructed (i.e., alignable and qualifying) site in the genome for each base (premutation). (B) The same data as in figure (A) divided by type of mutation given ancestral state. When all 12 mutational types are considered, 4-fold degenerate sites have the highest rate in 22/24 comparisons (binomial test, $P=3.6 \times 10^{-5}$).

also be less subject to sequencing artifact known to affect intrahost sampling (see Nekrutenko [2020]). We also however employ such within individual sequencing to infer selection.

Results

An Excess of Variants at 4-Fold Degenerate Sites Implies Purifying Selection

Were selection ongoing we expect that, per occurrence of a given nucleotide, the number of mutations observed at 4-fold degenerate sites would be higher than at sites 1 and 2 in codons. In all eight independent comparisons (4-fold site vs. site 1, 4-fold site vs. site 2, for four nucleotides), the 4-fold degenerate sites have more mutations per occurrence of the ancestral nucleotide (fig. 1a: binomial test, $P=0.008$). This is consistent with weaker selective constraint on mutations at 4-fold sites detectable even at sites recently sampled (fig. 1a).

We also see that when all 12 mutational types are considered, 4-fold degenerate sites have the highest rate in 22/24 comparisons (fig. 1b: binomial test, $P=3.6 \times 10^{-5}$).

To allow for dinucleotide effects, not considered when performing standard Ka/Ks tests, as performed for SARS-CoV-2 (see, e.g., Lythgoe et al. [2021]), we also consider the incidence rate of mutations centered on a given base at a 4-fold degenerate site in each of the 16 possible dinucleotides (either at sites 2 and 3, denoted “23,” or 3 and 1, “31”) and compare this with observations for the same dinucleotides where the mutations observed are not centered over codon third sites. The finding of a weaker selective constraint at 4-fold degenerate sites is resilient to such control (fig. 2). All four nucleotides are more mutable when situated at a 4-fold position, regardless of dinucleotide (Wilcoxon ranked-sum tests; A: $P=0.0052$, C: $P=9.8 \times 10^{-5}$, G: $P=0.00021$, U: $P=0.00024$).

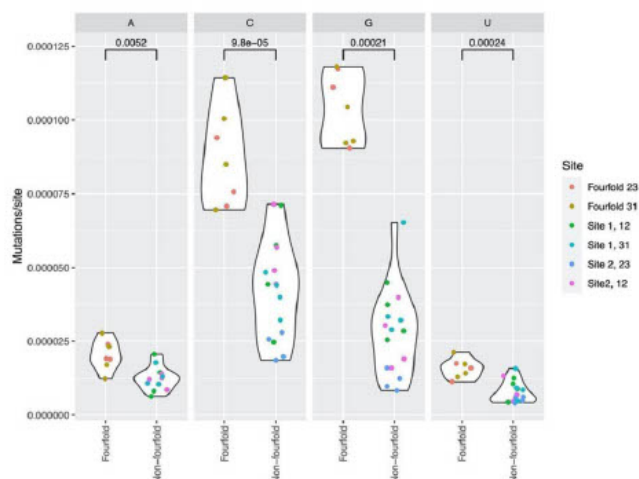


FIG. 2.—Comparisons between 4-fold and non-4-fold mutations at different reconstructed dinucleotide sites. The increased mutability of 4-fold sites is resilient to control for dinucleotide effects.

For Every Ten Variants that we See, around Five Other Mutations Are Not Recovered

The above evidence indicates that there must be some missing mutations derived from codon sites 1 and 2. If x is the number of new mutations seen per unit time down a particular lineage then $x + dx$ must be the true rate, dx being the mutations that happened but disappeared before they were sequenced. How can we estimate dx and hence the true mutation rate, $x + dx$? Under the assumption of no selection on 4-fold degenerate sites, and assuming that most mutations are either neutral or deleterious, then the difference between their rate and that observed elsewhere in the genome (fig. 1) is informing us of the rate of missing mutations. One could, alternatively, estimate the rate at 4-fold sites and assume all other sites have the same rate. However, we have previously identified both strong nucleotide skews at 4-fold sites and strong biases in the mutation rate per occurrence of each of the nucleotides (Rice et al. 2021). Considering that codon sites 1 and 2 are not as skewed in nucleotide content as 4-fold sites (Rice et al. 2021), the optimum approach is to extrapolate from the patterns at 4-fold sites in a manner that is sensitive to differences in nucleotide composition across sites.

dx can be estimated as the number of mutations seen in sequencing data multiplied by the proportion of mutations missing (P_m) (this being the proportion in terms of those observed), which may be estimated by comparing rates at codon sites 1, 2, and 3 to those at 4-folds (see Materials and

Methods for calculation). As $dx = x \cdot P_m$, the true mutation rate $= x [1 + P_m]$ per unit time. We estimate $P_m = 0.672$, that is, we are seeing $1/1.672 = 59.8\%$ of all mutations, missing 40.2% and the true mutation rate is 1.672 times higher than that observed. Most of the mutations missing are at G nucleotides. At A sites, we are seeing 70.0% of mutations and missing 30.0%, this equating to 3,119 mutations lost in the analyzed phylogeny. At C sites, we are missing 22.4% (5,735 mutations), at G sites, we are missing 61.5% (20,974 mutations), and at U sites, we are missing 21.8% (1,969 mutations) of mutations. In total, we estimate there are 31,797 unsequenced mutations missing in total.

Using mutational counts at the dinucleotide level, we may also estimate P_m and dx (and the number of mutations missing for each dinucleotide) by adapting the above method. For example, the mutation rate of A in an AG dinucleotide at site 12 may be compared with the mutation rate of A at AG dinucleotides where A is the 4-fold site. The mutation rate of G in an AG dinucleotide at site 12 is compared with the mutation rate of G at AG dinucleotides where the G is the 4-fold site, and so on. Owing to the structure of the genetic code, there are no 4-fold sites following a second codon position A, hence for these dinucleotides, we use the mutation rates at codon third sites, rather than 4-fold rates, for the comparison. The resulting predicted number of missing mutations is hence likely to be an underestimate. Nevertheless, from our dinucleotide calculations, we estimate $P_m = 0.489$,

that is, we are seeing $1/1.489 = 67.1\%$ of all mutations, missing 32.9% and the true mutation rate is 1.489 times higher than that observed. In terms of raw mutations, this equates to 44,966 missing dinucleotide changes or 22,433 mutations (as each point mutation affects two dinucleotides). Given the probable underestimation, this corroborates the mononucleotide prediction of $\sim 30,000$ missing mutations. Indeed, consistent with most missing mutations being at G sites, our dinucleotide analysis predicts that mutations are most commonly missing from GG (9,963 mutations) and UG (8,867 mutations) dinucleotide sites.

Currently, the rate of SARS-CoV-2 sequence change is estimated from circulating mutations to be about 1 every 2 weeks or $\sim 1 \times 10^{-3}$ per site per year (Duchene et al. 2020; Hill and Rambaut 2020; Nextstrain 2020). We hence suggest the mutation rate to be $\sim 1.5\text{--}1.7 \times 10^{-3}$ per site per year, assuming no selection at 4-fold degenerate sites.

Selection Skews the Mutational Matrix

It is possible that purifying selection acts in a uniform fashion against all sites, in which case all mutations at second sites (none of which can be synonymous) will be equally underrepresented when compared with 4-fold degenerate sites (N.B. a few C \rightarrow U (Leu \rightarrow Leu) and A \rightarrow G (Arg \rightarrow Arg) first site mutations are synonymous). This appears not to be the case with considerable heterogeneity between mutation types. Mutations from G are poorly tolerated at sites 1 and 2 (fig. 1) and in particular G \rightarrow U mutations appear to be commonly counter selected (we presume here that the 4-fold site rate does not indicate positive selection for U at such sites, not least because U4 observed [50.8%] is much less than neutral equilibrium predicted U4 content [65.6%]) (Rice et al. 2021).

To more systematically assess any such skew and the net effect on nucleotide composition, we compare the equilibrium nucleotide contents predicted on knowledge of the mutational profiles. We show using such a method that mutations at 4-fold degenerate sites and those not at 4-fold degenerate sites resulted in significantly different predicted mutational equilibria, with G underrepresented at 4-fold sites ($Z = -8.43$), but still very rare, whereas U is very common but nondeviant between the two sets ($Z = -0.35$). To fully understand the variation between sites, we extend these calculations to consider sites 1-, 2-, and 4-folds separately. This reveals that all three classes of site within a codon are different from all others (table 1). We conclude that selection not only prevents mutations at certain sites from increasing in frequency, but it also skews the mutational matrix with the nature of skew particular to the site concerned.

Evidence for Selection against Nonsense Mutations

Why might selection act differently on different mutations at different sites? We have observed from analysis of 4-fold sites a strong C|G \rightarrow U mutation bias in SARS-CoV-2 (Rice et al.

2021) (fig. 1). The above evidence suggests that at first sites within codons there is especially strong contemporaneous selection to counter this mutation bias. Why might this be? In all genomes, premature stop codons generated by nonsense mutations are commonly under strong purifying selection and there is no reason why this should not apply to SARS-CoV-2. Indeed, intrahost mutation appears to generate nonsense mutations that fail to transmit (Tonkin-Hill et al. 2021).

N \rightarrow U mutations at codons NAA, NGA, and NAG will generate stop codons (where N can be A, C, or G). The nine codons should be at a frequency of $9/61 = 14.75\%$ under unbiased nucleotide content but are at 17.05% with AAA (3.76%) being the second most common codon after GUU (3.9%). Mutations to U at the second site can never generate a stop codon. Consistent with these expectations, the reduction in U seen at non-4-fold sites compared with 4-fold sites is profound at site 1 ($Z = -6.86$) but not seen at site 2 ($Z = 4.72$) (table 1). Similarly, site 1 has much less predicted U content at equilibrium than site 2 ($Z = -12.3$). The raw predicted U content at equilibrium reflects these trends: U1* = 63.2%, U2* = 73.5%, U3* = 70.1%, U4* = 69.5%. More specifically, when considering the full mutational profile of the virus, we find nonsense mutations to be significantly less common than other point mutations (2×2 Chi²; $\chi = 1,942.9$, $df = 1$, $P < 2.2 \times 10^{-16}$). They are also less common when they generate an in-frame stop codon than a +1-frameshifted (2×2 Chi²; $\chi = 1,924.4$, $df = 1$, $P < 2.2 \times 10^{-16}$) or a +2-frameshifted (2×2 Chi²; $\chi = 2,626.1$, $df = 1$, $P < 2.2 \times 10^{-16}$) stop, and are significantly more likely at the first nucleotide position than the second (2×2 Chi²; $\chi = 137.1$, $df = 1$, $P < 2.2 \times 10^{-16}$). The commonality of nonsense mutation at first sites is likely owing to the strong N \rightarrow U mutation bias, all stop codons having U at the first site.

Sequenced isolates deposited in GISAID are usually consensus sequences that discard all but the most frequent base at any position from individual samples, and therefore likely do not fully reflect the diversity of SARS-CoV-2 among infected individuals. To gain insight into within-individual variation, we analyzed variants identified from publicly available SARS-CoV-2 raw sequencing read data to quantify variants within samples. Could this data provide evidence for missing mutations in GISAID sequences and purifying selection being a reason? Counting nonsense mutations present at some frequency in 1,092 samples, there is a mean of 0.23 nonsense mutations per sample. Compared with GISAID isolates, within-individual samples are far more likely to harbor a nonsense mutation (1.3% of GISAID isolates vs. 13.4% of within-individual samples, 2×2 Chi²; $\chi = 1,110.3$, $df = 1$, $P < 2.2 \times 10^{-16}$). Similar to the mutational profile above, for within-individual variation, first nucleotide positions are significantly more likely to generate an in-frame stop codon than second positions (0.9% vs. 0.6%, respectively, 2×2 Chi²; $\chi = 7.0$, $df = 1$, $P = 0.008$).

Table 1
Comparisons between Equilibrium Vectors

Comparisons	P Value	A* 1	A* 2	A: Z Score	C* 1	C* 2	C: Z Score	G* 1	G* 2	G: Z Score	U* 1	U* 2	U: Z Score
4 versus non-4	0.012	0.170	0.142	3.596	0.10	0.099	0.221	0.035	0.060	-8.426	0.695	0.699	-0.348
1 versus 2	<0.001	0.196	0.111	12.939	0.094	0.076	6.858	0.079	0.077	0.550	0.632	0.735	-12.282
1- versus 4-fold	<0.001	0.196	0.170	3.138	0.094	0.10	-1.745	0.079	0.035	14.416	0.632	0.695	-6.860
2- versus 4-fold	<0.001	0.111	0.170	-8.665	0.076	0.10	-8.079	0.077	0.035	12.882	0.735	0.695	4.716

Notes.—*P*s determined by 10,000 simulations (see Materials and Methods). Z score orientation is such that a positive value implies comparative enrichment within the first comparator in the Comparison column. For example, in row 1 (4 versus non-4), the 4-fold degenerate sites are site class 1 and non-4-fold degenerate sites are the non-4-fold degenerate sites (i.e., all others) and are class 2. In this case, C* 1, for example, is then the equilibrium C content of sites of class 1 (4-fold degenerate) and C* 2 the equilibrium C content of sites of class 2.

Nonsense Mutations Account for ~10% of Missing Mutations

A prior observation of $Ka/Ks \sim 1$ (Bai et al. 2020) suggests that nearly all intrahost selection must be against nonsense mutations. Selection against nonsense mutations cannot, however, explain all the observed patterns. Under the assumption that there is no selection to avoid out-of-frame stop codons, we may extrapolate the out-of-frame nonsense mutation rate to estimate how many nonsense mutations are missing in the above trends. Taking the out-of-frame per-trinucleotide nonsense mutation rate as the mean of the +1 and +2 frame-shifted rates, this equals 1.46×10^{-5} mutations per trinucleotide compared with 1.26×10^{-6} in-frame. We are hence missing nonsense mutations at a rate of $1.46 \times 10^{-5} - 1.26 \times 10^{-6} = 1.33 \times 10^{-5}$ per trinucleotide and, scaled to the number of in-frame trinucleotides analyzed that are one point mutation away from a stop, this equates to 3,205 missing nonsense mutations. As we above estimate a total of 31,797 mutations missing from the sequence data, nonsense mutations only account for approximately 10.1% of these.

We also used an alternate method of estimating the expected number of missing nonsense mutations analytically, relying on trinucleotide substitution patterns observed at 4-fold degenerate sites. As we have previously mentioned, 4-fold degenerate sites should evolve in a mostly neutral way and as such, the observed mutation rates on these sites should better reflect mutational bias. For this, we compared the proportion of in-frame nonsense mutations observed in our data set (264 nonsense mutations out of 49,358 trinucleotide changes), to an expected proportion of nonsense mutations derived from distributing this same number of mutations randomly across the sequence at the rate of trinucleotide substitutions of 4-fold degenerate sites (an average of 2,909.362 nonsense mutations out of 49,358 trinucleotide changes, 95% CI lower = 2,908.410, upper = 2,910.314). This comparison equates to approximately 2,645 missing nonsense mutations on an average, accounting for only 8.3% of our 31,797 estimated missing mutations. This is close to the above estimate of ~10%. Given prior evidence that $Ka/Ks = 1$ (Bai et al. 2020), this result is surprising, suggesting that the majority of counter-selected mutations are not nonsense ones.

Reinforcing this result, we also see that when all 12 mutational types are considered, not only do 4-fold degenerate sites have the highest rate in 22/24 comparisons (binomial test, $P = 3.6 \times 10^{-5}$) but the rate is also higher at 4-fold degenerate sites for mutations that could never generate stop codons, for example, G->C, U->C at sites 1 and 2 (fig. 1). Likewise, G->U rates are marginally higher at site 1 rather than site 2, whereas we expect the opposite if all selection is against nonsense mutations.

Although second site nucleotide content is considered the key determiner of the chemical property of the encoded amino acid (Haig and Hurst 1991; Freeland and Hurst 1998; Gills et al. 2001; Schwersensky et al. 2020), only five of 12 first site versus second site comparisons have higher rates at the first site. The same analysis of the 12 mutational types emphasizes the great disparity in G->U, and to a lesser degree C->U, mutation between 4-fold degenerate sites and codon sites 1 and 2, this despite the fact that some (Leu->Leu) first site C->U mutations are synonymous (fig. 1).

What then might predict these trends? We start by considering parameters that might explain why some amino acid exchanges are seen less than expected given the mutational profile. Then we consider in more detail a biophysical model of disruption of a key protein-protein interaction, spike with ACE2.

Amino Acid Cost and Chemical Distance as Predictors

Are there general properties of the missense/nonsynonymous mutations that are underrepresented compared with a mutational null? In order to test this, we first analyzed the relationship between under/overrepresentation of amino acid substitutions and 12 estimators of different biochemical properties of such amino acids (supplementary table 2, Supplementary Material online). However, as mentioned, mutational biases can occur in the context of more than one nucleotide, for example, when responding to codon bias or as a result of nonselective mutational processes, like APOBEC-induced genomic C-to-U deamination (Simmonds 2020b). To account for the effect of multinucleotide mutational biases on amino acid replacements, we first measured the over/underrepresentation of each pair of amino acid replacements,

compared with expectations derived from trinucleotide substitution patterns observed at 4-fold degenerate sites. Then we used a Best subset regression to select an optimal linear model explaining the over/underrepresentation of amino acid substitutions using the 12 estimators of biochemical properties, plus a set of variables measuring the degree of change in U nucleotide, as well as UU and CG dinucleotide content between codons in each pair of amino acids.

The optimal model found includes many parameters indicative of selection against nonsynonymous mutations that break proteins by replacing one amino acid with a chemically dissimilar one. Notably, between pairs of amino acids, predictors include their distance in a BLOSUM100 similarity matrix, differences in polarizability, and residue volume (adjusted $R^2 = 0.3533$, P value = 9.563×10^{-11} , [supplementary table 3, Supplementary Material](#) online). Perhaps more enigmatically, we also observed an enrichment of missense mutations to amino acids with a slower decay, possibly suggesting some selection for reduced metabolic cost of SARS-CoV-2 protein production (NB fast decay means more cost per unit viable amino acid). There is avoidance of UU residues but this is not significant.

Spike-ACE2 Interaction Disruption Predicts Missing Mutations

The above measures are fairly broad brush but suggest, as might be expected, protein disruption to be a source of purifying selection in real time. Using spike protein, for which we also have an underlying biophysical model of its binding (Starr et al. 2020), we can examine the same hypothesis with better granularity. For this, we again compared within-individual variation to GISAID isolates. Firstly, counting observed missense mutations in the receptor-binding domain of the spike (S) gene, we find 212 unique amino acid substitutions in our GISAID alignment compared with the reference sequence and 61 substitutions in the within individual variation. This is especially notable as the number of GISAID isolates in our alignment (83,665 nonreference isolates) with the reference sequence is many times the number of samples with observed variants in the Galaxy Project within-individual variation data set (1,092 samples). Secondly, using a mutational screen of amino acid substitutions in the receptor-binding domain and their measures of relative ACE2-binding activity compared with the reference genome (Starr et al. 2020), we compared the phenotypic effects of the substitutions we observe in GISAID isolates and those from within-individual variation (fig. 3). Substitutions observed within individuals reduce relative ACE2-binding activity more than observed GISAID substitutions (median-binding activity, respectively: -0.27 and -0.08 ; $P = 0.0002$; Wilcoxon ranked-sum test). This provides evidence for unobserved SARS-CoV-2 variation when considering sequenced GISAID isolates only and purifying selection being a possible reason for such variants failing to reach the

most frequent nucleotide at a given position and therefore discarded at the consensus sequence stage.

Synonymous Mutations Degrading Match to the Human Codon Usage Are Counter Selected

Above we have concentrated on what a priori are expected to be relatively large effect mutations. We can also ask whether we can also detect selection at synonymous sites. In order to test this, we compare the proportion of 4-fold synonymous mutations resulting in a codon with an increase, decrease, or with no effect on optimal codon usage. At first sight, one might imagine that such a method could not work as we are attempting to infer selection at synonymous sites using observed mutations at synonymous sites, rendering the analysis circular. However, this need not be true. Consider two amino acids for which the "optimal" codon for each has a different synonymous site. If one amino acid has a U as the optimal synonymous site then common C->U mutations will not be opposed by selection. However, if another amino acid has C as the optimal site then the same mutation will be opposed by selection. If selection is strong enough then both processes will contribute to the net observed mutational matrix. Consequently, although for the two the different synonymous sites with the same nucleotide content the null rate will be the same, we expect to see deviations away from this null in a manner dependent on whether mutation bias and selection are aligned or not. Deviations between expected mutational profiles and observed mutational biases, then have the potential to detect selection on synonymous mutations. The method is flexible to any definition of "optimal" as we can test whether deviation from the observed mutational null tends to act against mutations that are thus defined as nonoptimal. We consider two such definitions.

First, we consider translational efficiency, as measured by the tRNA adaptation index (tAI) (dos Reis et al. 2004) in humans, calculated based on the copy number of tRNA genes and the binding strength between a codon and a tRNA (Yoon et al. 2018) with random expectations derived from simulations taking into account the trinucleotide mutational patterns of 4-fold sites. Using tAI as a measure for selection on translational efficiency has some pitfalls. In multicellular organisms with larger genomes, there is no correlation between codon usage and tRNA, possibly due to a higher tRNA gene redundancy in larger genomes, which would decrease selection for specific codons (dos Reis et al. 2004). Furthermore, tRNA copy numbers do not necessarily reflect the fact that pools of distinct tRNAs are dynamic and can vary considerably in different conditions and tissues (Hernandez-Alias et al. 2021). We observe a significant depletion of 4-fold synonymous mutations increasing codon adaptation (two tailed P value = 0.0022 , [supplementary fig. 1, Supplementary Material](#) online), as well as a small, yet not significant, enrichment of mutations that decrease or do not disrupt tAI (two tailed P

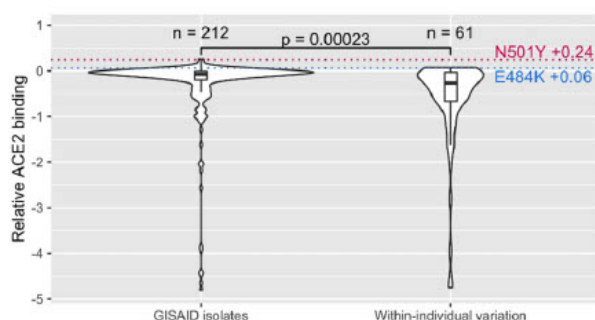


Fig. 3.—Relative effects on ACE2-binding activity for missense mutations in GISAID isolates and within individual variation. Distribution of relative effects on binding activity of unique missense mutations within the receptor-binding domain that are observed one or more times in GISAID isolates and Galaxy Project within individual variation. Change in relative ACE2 binding of two notable amino acid substitutions within the receptor-binding domain of spike observed in variants of concern, N501Y and E484K, are annotated as dotted lines.

value = 0.0984 and 0.326, respectively). These results suggest, if anything, some selection acting against translational efficiency dependent on the tRNA pool.

Second, we compared the number of mutations that caused a switch in the SARS-CoV-2 genome to a codon with a higher relative synonymous codon usage (RSCU) in human. When accounting for the trinucleotide mutational patterns, which would better capture the effects of mutational biases derived from CpG avoidance or APOBEC-induced mutation cytosine deamination, we do observe a significant overrepresentation of mutations that increase RSCU (P value = 1×10^{-4} , [supplementary fig. 2, Supplementary Material online](#)). This result is consistent with selection occurring in SARS-CoV-2 to match the human codon usage profile.

We further ask whether such signatures of selection can be detected within hosts. For this, in a similar manner as with the between host data, we compared the observed tAI and RSCU of 4-fold synonymous positions in the intrahost data set, against random expectations derived from simulations taking into account the trinucleotide mutational patterns of 4-fold sites in this same data. We find a significant depletion of 4-fold increasing codon adaptation (two tailed P value = 0.0044, [supplementary fig. 3, Supplementary Material online](#)), as well as a significant enrichment of mutations that do not disrupt tAI ($P = 0.0354$). This is consistent with the above finding of selection against tRNA-dependent translational efficiency. We also detect overrepresentation of synonymous mutations increasing human RSCU in the SARS-CoV-2 intrahost data but the deviation from null is not significant ([supplementary fig. 4, Supplementary Material online](#)). Although there should be some selection occurring among strains within a host, reflected in differences in allelic frequency, selection on RSCU might not be strong enough to

have a measurable impact at the shorter time scale reflected by intrahost variation.

Discussion

Prior to the genomic age mutation rates were classically estimated by considering substitution rates (between two species) at synonymous sites with assumptions made about generation times and time to common ancestry to provide a per generation per base pair estimate (see, e.g., Keightley and Eyre-Walker [2000]). The restriction to the synonymous sites was a means to reduce the impact of purifying selection depressing the estimate. More recently, this method has been supplanted by MA line or parent–offspring sequencing (Lynch et al. 2016). Such methods assume that there is no important degree of purifying selection between parent and recent descendants and hence that the profile and rate of mutations can be estimated in an unbiased manner. Our finding of common and strong purifying selection detectable in real time affecting mutations prior to their being sequenced strongly suggests that, at least for SARS-CoV-2, this is not the case. In principle our null simulation correction method could also be employed to correct for underestimation in MA and parent–offspring analyses to determine the mutation rate. However, in genomes such as that of humans, where few sites are subject to purifying selection, the correction is probably not important. For more economical genomes (with higher CDS density) it may be more relevant.

Given the evident purifying selection, an estimate of the rate of evolution of the virus is not the mutation rate *sensu stricto*, but rather of the rate at which new mutations appear and are viable enough to be sequenced. The latter measure is sometimes referred to as the substitution rate (van Dorp et al.

2020), the rate of evolution (Hill and Rambaut 2020), or the mutation rate (Zhao et al. 2004; Pathan et al. 2020). Given our results, we advise against the latter usage to avoid confusion. Put differently, if one were to take estimates of rates of sequence change for SARS-CoV-2 that employ observed RNA changes (Duchene et al. 2020; Hill and Rambaut 2020; Nextstrain 2020), and assume that this is the underlying mutation rate, one would be wrong. We indeed find that the discrepancy is not modest (an ~50% correction would be needed).

Although to estimate the true underlying mutation rate, we thus need to control for purifying selection, the discrepancy between the mutation rate (*sensu strictu*) and the evolutionary rate is important in other contexts. If selection on viral escape from vaccines (or antiviral drugs) is in part owing to intrahost selection, then knowing the underlying mutation rate, and the difference between it and the apparent evolutionary rate, is important. Furthermore, claims of higher or lower mutation rates in some lineages would need to control for the possibility of differences in, for example, effective population size (N_e) or sampling depth. Variation in N_e , modulating the strength of selection, could result in conflation of differences in the mutation rate *sensu strictu* with efficacy of selection differences (lower N_e permits more mutations to circulate). Similarly, we would expect that deeper sampling of genomes within an individual will provide evidence for genomes that will be removed by purifying selection but have yet to be removed (as indeed we show). This could also lead to misleading inference of increased mutation rates. To understand how important within-host selection might be, it is important to control for such effects and unbiased sampling of 4-fold degenerate sites is, we suggest, preferable to analysis of sequence classes known to be under purifying selection.

The analysis of missing mutations is, however, of less interest in contexts where we wish to employ the rate of evolution to estimate coalescent times, as in this context the appearance rate (circulating in the population) per unit time is the relevant metric, not the true underlying mutation rate. Nonetheless, in this context understanding whether there are sites subject to purifying selection can be important for determining whether rate estimate correction is needed. As O'Fallon (2010) noted, purifying selection acting at many linked sites can systematically bias genealogical reconstruction but by allowing a class of sites to have a time-dependent rate can enable some degree of correction. Likewise, Wertheim and Kosakovsky Pond (2011) show that, for other viruses, adjusting codon models to allow for purifying selection can lead to estimates of the time to common ancestry longer than those supposed from rates of observed circulating mutations. Our results suggest that such adjustments are then required for SARS-CoV-2.

Our new estimate is likely to be an underestimate. Although we have attempted to control for nucleotide biases

and biases in rates of each class of mutation, we have also assumed that 4-fold sites are themselves free from selection. Our analysis of two specific models of selection on codon usage provided no evidence for selection on codon usage to match tRNA pools (indeed selection appears to be in the opposite direction) but of selection to match human codon usage. The later result was seen unambiguously when testing the circulating genomes for deviation from null, but not statistically significantly replicated with intrahost variation data. However, SARS-CoV-2 has multiple modes of selection on nucleotide content that would not be detected by such methods. These include selection against CpG dinucleotides to avoid ZAP, against UpA to avoid RNAase L and more generally against U, mediated possibly by transcript destabilization and/or expression level (Rice et al. 2021). Just as we observe possible selection against U so we and others have identified possible selection for A (Rice et al. 2021; Kustin and Stern 2021). One possible mechanism of this could indeed reflect the high U content and hence selection for A to enable stable base pairing in RNA stem structures (Ratcliff and Simmonds 2021).

Some of our results on the causes of purifying selection seem fairly simple to interpret. It is not surprising that nonsense mutations are counter-selected, nor that a biophysical model of spike protein function recovers a trace of purifying selection. Similarly, that features like chemical similarity predict amino acid exchange rates make sense, as highly different amino acids are likely to corrupt proteins just as nonsense mutations do. Nonetheless, our results hold a few surprises when considered against the prior literature. Although purifying selection was previously identified (see, e.g., Tonkin-Hill et al. 2021; Lythgoe et al. 2021), given prior Ka/Ks estimates near unity (Bai et al. 2020), seen also for SARS-CoV (Zhao et al. 2004), it might reasonably have been inferred that most of the missing mutations must be nonsense mutations. Our results do not support this. We however consider Ka/Ks an unsuitable tool for analysis of polymorphic data, especially in a context with complex mutation and nucleotide biases (see Introduction).

It is similarly, not so obviously expected that amino acid cost determinants (amino acid decay rate) would factor as predictors of amino acid exchange rates, with selection against more costly ones. The usual logic is that making "costly" amino acids, when cheaper good alternatives are available, causes a fitness cost owing to differential ATP usage. For amino acids with high decay rates, these costs are suffered more as the pool of amino acids needs replenishing faster. However, why a temporary visitor to a cell (the virus) that causes damage regardless, will have selection to use less costly amino acids is not so transparent. Why would it be under selection to use less costly amino acids if the cell making those amino acids will soon be dead anyway? In what sense would the virus benefit from using cheaper amino acids? The key amino acid parameter, decay rate rather than synthesis

cost per se, may point to an alternative cause. There could well be selection for rapid viral replication. A genome that both harms the cell's ability to manufacture new amino acids but needs rapid translation, may be under selection to use those amino acids that have a long half-life, regardless of ATP costs. Usage of those with a short half could leave the virus slowed in translation waiting for ever rarer and diminishing pools of charged tRNAs. We thus suggest that amino acid ATP cost per se is not the key parameter, but rather delay to translation might be. That SARS-CoV-2 interferes with the host's splicing and translational machinery (Banerjee et al. 2020), suggests that amino acid biosynthesis may well be affected.

Similar logic may explain why selection on synonymous sites failed to identify adaptation to the tRNA pool. Our estimation of this pool from tRNA copy numbers may well not reflect the pool of charged tRNAs as certain amino acids, with high decay rates, are limiting. Exactly why matching the human codon usage does matter is less clear, but a direct coupling between GC content and gene expression in both nuclear and cytoplasmic compartments (for reasons unknown) of virus-mimicking intronless transgenes (Mordstein et al. 2020, 2021) could underpin such an effect.

We highlighted several analytic challenges associated with this virus's genome. One we have not fully broached is the problem of potential interactions between genomic location, RNA structure, and both mutation rate and mutation profile. We have controlled for complex mutational biases by consideration of di- and trinucleotide context. We have also attempted to control for rate heterogeneity by exclusion of hypermutagenic sites, much as previously we excluded homoplastic sites (Rice et al. 2021). Hypermutagenic sites are relatively rare (1% of all sites, 1.8% of variable sites, 2.7% of 4-fold sites, 4.2% of variable 4-fold sites) but given that they contribute a disproportionate number of observed mutations they have the potential to lead to false inference if the mutational spectrum at such sites is different from that at non-hypermutagenic sites. Although the sample of hypermutagenic sites is limited, we can compare their trinucleotide context with that of the remaining mutations for four-fold sites (supplementary fig. 5, Supplementary Material online). We find relative enrichment of UCN->UUN consistent with more frequent activity of APOBEC on hypermutatable sites. We also see evidence for enrichment of CGN->CUN. This is suggestive of selection against CG residues, possibly owing to ZAP-mediated attack. However such a model would also predict CGN->C[C]A[U]N which we do not see. A possible combination of mutation bias (toward U) and selection against CG might need to be evoked.

Our method to control for hypermutagenic sites defined sites by reference to the number of independent mutational events seen across all sites, with hypermutagenic being defined by deviation from a negative binomial. This method, however, makes no allowance for position by nucleotide effects. One could suggest that there might be sites that do

not have unusually large numbers of mutations compared with all other sites, but do when considering their ancestral nucleotide state. We have considered such a model treating each of the four nucleotides independently and eliminating, for each, those sites in the alignment with more independent mutational events than expected given a negative binomial distribution parameterized for the nucleotide in question. To assess whether this alternative methodology makes a difference to the final analysis of the residual mutational matrix (i.e., after removal of hypermutagenic sites), we compare the residual matrix from the nucleotide-controlled and -uncontrolled methods. We find no significant difference between the two residual matrices ($P=0.897$: Predicted equilibria for original hypermutable threshold—A: 0.170, C: 0.100, G: 0.035, U: 0.696; Predicted equilibria for nucleotide-controlled thresholds—A: 0.162, C: 0.076, G: 0.025, U: 0.738).

Adding to such complexity is the notion that the rate or profile of any given nucleotide motif may be contingent on its genomic location, for example, in a stem loop or not. Untangling cause and effect in this instance will not be trivial. A low rate of observed SNPs in RNA stem structures (Simmonds 2020a) could, for example, reflect selection against mutations that disrupt RNA stem structures (Simmonds 2020a). Alternatively, it may be owing to a reduced mutation rate if RNA stems protect from mutation, for example, via shielding from APOBEC (Ratdiff and Simmonds 2021). We are unaware of theoretical work that attempts to correct for motif (k mer) by location effects on rates and profiles. This we leave to future work.

Materials and Methods

Creating a Mutational Matrix

Multiple sequence alignment of 106,448 SARS-CoV-2 genome assemblies was downloaded from the GISAID (Shu and McCauley 2017) Initiative EpiCoV platform, these being those available as of September 28, 2020. Isolates with more than 1% of ambiguous base calls or more than 5% of any CDS missing were removed. This left 83,666 genomes. For list of genomes and sources, see supplementary table 1 and data 1, Supplementary Material online.

We employed NCBI Reference Sequence NC_045512.2 to specify CDS coordinates. However, following further annotation of genes (Kim et al. 2020), we modified the gene locations to reflect those specified: <https://github.com/hyeshik/sars-cov-2-transcriptome/blob/master/reference/SARS-CoV-2-annotations.gff>. Specifically, to avoid a small codon overlap, we exclude CDS overlaps, hence employed annotation:

```
ORF7a protein 27394.27759--27394.27753
ORF7b protein 27756.27887--27762.27887
```

To consider ORF1a and ORF1b independently and to avoid overlap, we employ:

ORF1a→266-13465
ORF1b→13471-21552

CDSs for each gene in each strain were extracted from these alignments, and frameshift correction was then applied using the protein sequence of the Wuhan-Hu-1 reference genome (EPI_ISL_402124), sampled from a retailer at Huanan Seafood Wholesale Market, Wuhan on December 30, 2019 as reference, using the DECIPHER R package. This early sequence matches the consensus generated from all of the 19 sequences that were collected prior to December 31. CDSs were then translated, realigned with MAFFT 7.458 (Katoh and Standley 2013), and then reversed translated using TranslatorX (Abascal et al. 2010).

A phylogenetic tree of SARS-CoV-2 isolates (released October 28, 2020; Lanfear 2020) was pruned using DendroPy v4.4.0 (Sukumaran and Holder 2010) to match isolates present in our sequence alignment, and similarly our sequence alignment was filtered to match isolates present in the phylogenetic tree. This left 78,971 genomes present in both. Aligned CDSs were concatenated to create a single coding sequence alignment of length 30,696bp as input for ancestral sequence reconstruction. Ancestral sequence reconstruction at internal nodes of the predefined phylogenetic tree was performed using an empirical Bayesian method with a GTR+G model of substitution in IQTree v2.1.2 (Minh et al. 2020). Inferred bases with a probability of less than 0.99 were masked.

Known problematic sites in the SARS-CoV-2 genome (released December 12, 2020, Available from: https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf) identified and collated at <https://virological.org/masking-strategies-for-sars-cov-2-alignments/> 480 were masked and the number of mutations per site at 4-fold degenerate sites were counted.

Given that some sites appear to be both hypermutable, hence subject to homoplasy (van Dorp et al. 2020), and potentially unrepresentative of the rest of the genome we sought to exclude these sites from more general analysis (we consider their properties separately). To find thresholds for masking hypermutable sites in the genome, a negative binomial distribution, with μ fixed to the median number of mutations per site (median: 1), was fitted to the observed values using the `fitdist` function of the `fitdistrplus` R package (fitted distribution: $\mu = 1$, size estimate = 0.3414126; Delignette-Muller and Dutang 2015). An expected number of hypermutable sites can be estimated from the fitted distribution for a given number of sites. We set a cut-off threshold where we expect no more than one site with that number of mutations and mask the sites above that threshold. For example, for 4,248 4-fold degenerate sites, we expect at least one site with 17 mutations and less than one site with 18 mutations, and therefore mask sites where 18 or more mutations have occurred independently across the tree. For 9,739

first, second, or third codon position sites, we expect at least one site with 19 mutations per site, etc.

We also consider a second approach in which we define (and exclude) hypermutable sites by reference to the number of mutated sites with the same premutation nucleotide. That is to say, for each site, we determine the number of independent mutational events at that site. We then compare these by-site numbers to other sites within the alignment with the same premutation nucleotide. We then calculate the mean number of independent mutational events for all such sites of a given premutation nucleotide. The mean of this distribution then informs an expectation based on a negative binomial. We again set a cut-off threshold where we expect no more than one site with that number of mutations and mask the sites above that threshold. Under the first method 2.7% (116/4,248) of 4-fold sites are hypermutable and 4.2% (116/2,798) of variable 4-fold sites are hypermutable. Under this second, nucleotide-dependent method 0.5% (19/4,248) of 4-fold sites are hypermutable and 0.7% (19/2,798) of variable 4-fold sites are hypermutable. About 17 of these 19 hypermutable sites are considered hypermutable in the prior method too.

Mutations were counted from root to tips of the tree, taking ancestral parent nodes as reference and counting mutations in descendants at each node of the tree. If a mutation occurred at the same site in two descendants at the same position of the tree, this mutation was counted once (similar to De Maio et al. [2021]). When counting variants, known problematic sites within the genome were masked, hypermutable sites above their respective thresholds were masked, and codons containing more than one variant in a single genome compared with its direct ancestor were masked. Whole-genome nucleotide flux estimates were obtained by counting the frequency of each type of mutation and normalizing by the frequency of the nucleotide in the reconstructed ancestral genomes. This resulted in a data set of 51,244 variants.

Estimating the Number of "Missing" Mutations

How many mutations would be expected if all codon sites evolved as if they are 4-fold? To estimate this, and hence how many mutations might be missed in the sequencing data, let us suppose that the number of mutations at ancestral base N ($N = A, C, G, \text{ or } U$) normalized to the number of ancestral N s at 4-fold degenerate sites is N_4 . Likewise, N_1, N_2, N_3 for codon sites 1–3, respectively. The absolute number of missing (M) mutations across the genome is hence:

$$M = \sum_{i=1}^3 F(N_i) \cdot (N_4 - N_i),$$

where $F(N)$ is the absolute number of occurrences of nucleotide N as the ancestral residue at base i across all reconstructed sites in the tree.

The comparable sum of all mutations observed (O) is:

$$O = \sum_{i=1}^{j=3} F(N^i) \cdot N_i.$$

Note here, we use all mutations at third sites because we need to count all mutations. The true total (T) number of mutations then is: $T = O + M$. For every observed mutation the proportion missing (P_m) of those observed is: $P_m = M/O$.

We extend the same method to consideration of dinucleotide-defined mutation bias. There are, however, two complicating factors in such analysis: 1) dinucleotides may mutate at either of their nucleotide sites and 2) any given point mutation will affect two overlapping dinucleotides (a mutation at B in ABC, is both associated with AB and BC). To address problem (2), we calculate missing “dinucleotide changes,” rather than mutations, the total number of which may be halved to estimate the number of missing mutations. To address problem (1), we control for each mutation’s nucleotide position within the dinucleotide in our analysis.

For each of the 16 dinucleotides, we first calculate six position-specific mutation rates: $D_{(1)2}$, $D_{(1)2}$, $D_{(2)3}$, $D_{(2)3}$, $D_{(3)1}$, and $D_{(3)1}$, where the numbers represent dinucleotide position within a codon and brackets indicate the mutation site. These we compare with the position-controlled 4-fold null mutation rates. The number of missing dinucleotide changes (M) for dinucleotide “D” may be hence be calculated at each position (12, 23, or 31) as:

$$M_{12} = \sum F(D^{12}) \cdot (D_{(4)1} - D_{(1)2}) + \sum F(D^{23}) \cdot (D_{(2)4} - D_{(1)2}),$$

$$M_{23} = \sum F(D^{23}) \cdot (D_{(4)1} - D_{(2)3}) + \sum F(D^{23}) \cdot (D_{(2)4} - D_{(2)3}),$$

$$M_{31} = \sum F(D^{31}) \cdot (D_{(4)1} - D_{(3)1}) + \sum F(D^{31}) \cdot (D_{(2)4} - D_{(3)1}),$$

where $F(D^{12})$ is the number of occurrences of dinucleotide D as the ancestral residue at position 12. The total number of missing dinucleotide changes (M) is: $M = M_{12} + M_{23} + M_{31}$.

The comparable sum of all dinucleotide changes observed, O , for dinucleotide “D” can be calculated at each position (12, 23, or 31):

$$O_{12} = \sum F(D^{12}) \cdot (D_{(1)2}) + \sum (D^{12}) \cdot (D_{(1)2}),$$

$$O_{23} = \sum F(D^{23}) \cdot (D_{(2)3}) + \sum (D^{23}) \cdot (D_{(2)3}),$$

$$O_{31} = \sum F(D^{31}) \cdot (D_{(3)1}) + \sum (D^{31}) \cdot (D_{(3)1}).$$

The total number of observed changes for dinucleotide D is calculated as: $O = O_{12} + O_{23} + O_{31}$.

The true total for dinucleotide D is then: $T = O + M$.

The true totals of each dinucleotide may be summed to estimate the true total number of dinucleotide changes. As point mutations affect two dinucleotides, we divide this value by two to predict the true number of mutations.

Calculation of Mutational Equilibria

Given that the mutational profile is strongly U biased, considering solely rates of GC \leftrightarrow AU mutations (Long et al. 2018) is likely to miss important dimensions. The equilibrium content of all four nucleotides we therefore estimate using the full mutational spectrum (Charneski et al. 2011; Rice et al. 2021). We here follow the same methodology as used in our previous publication (see Rice et al. [2021]). Briefly, if the frequency of G is denoted G and the frequency of U is denoted U , etc., mutational flux from G to U, per occurrence of G, is denoted $g2u$, and A to C, per occurrence of A, is denoted $a2c$, and so on (each mutational flux captured by the mutational matrix). Equilibrium is then defined as occurring when the rate of loss of each nucleotide is equal to the rate of gain of the nucleotide, for all nucleotides, with the additional constraint that $A + U + C + G = 1$:

$$G(g2u + g2c + g2a) = A(a2g) + U(u2g) + C(c2g)$$

$$C(c2u + c2g + c2a) = A(a2c) + U(u2c) + G(g2c)$$

$$A(a2u + a2c + a2g) = G(g2a) + U(u2a) + C(c2a)$$

$$U(u2g + u2c + u2a) = A(a2u) + G(g2u) + C(c2u).$$

Comparing Mutational Matrices

For each class of site (e.g., 4-fold degenerate, not 4-fold degenerate, codon first sites, etc.), we determine the absolute number of each of the 12 classes of mutation (A \rightarrow C, A \rightarrow U, etc.), the rate then being this normalized to the frequency of the ancestral base giving the rates ($n2m$) defined above, that is, the rate of $n2m$, per incidence of n . We then analytically solve, using NumPy (Walt et al. 2011), to determine the mutational equilibrium vector (of length 4), this specifying the frequencies of the four bases at mutation-neutral equilibrium.

To compare between pairs of equilibrium values (e.g., for codon first sites and for 4-fold degenerate sites), we determine the Euclidean distance between the resulting vectors and perform randomizations. In these, we randomly reallocate the underlying mutations to pools the same size as contributed to the two vectors in the first instance. From each simulation, we derive the equilibrium predicted values of the two pseudo mutational profiles and calculate the difference between them. From multiple simulations, we determine the null distribution. We express the observed difference in terms of the distance away from the mean of the simulants in standard deviation units derived from the simulants (i.e., a Z score). The method permits both estimation of the

significance of the distance between any two vectors and identification of the nucleotides most deviant (and the significance of each one's deviation).

Analysis of Amino Acid Properties

In order to test the relationship between overrepresentation of particular missense mutations and changes in the biochemical properties of amino acids, we built a generalized linear model. We first started by calculating the bias in missense mutations as a Z score:

$$ZAA_{\text{bias.cod}} = \frac{(O_{\text{cod}} - E_{\text{cod}})}{SD(E_{\text{cod}})},$$

$$ZAA_{\text{bias}} = \frac{\sum ZAA_{\text{bias.cod}}}{n_{\text{cod}}},$$

$ZAA_{\text{bias}} = \sum ZAA_{\text{bias.cod}} / n_{\text{cod}}$, where $ZAA_{\text{bias.cod}}$ is the mean measure of over/underrepresentation of change between codon pairs for each pair of amino acids: O_{cod} is the observed number of single nucleotide substitutions switching from a particular codon to another for that pair of amino acids, and E_{cod} is the expected number of codon changes when accounting for the rate of trinucleotide substitution at trinucleotides centered on 4-fold degenerate sites. E_{cod} and its standard deviation were estimated as the mean of 10,000 simulations distributing 49,358 mutations randomly across the SARS-CoV-2 CDSs at the same rate as the trinucleotide substitution observed at 4-fold degenerate sites. The parameter n_{cod} is the number of codon pairs resulting in a particular amino acid replacement. $ZAA_{\text{bias.cod}}$ values for each pair amino acids were then averaged to obtain a measure of over/underrepresentation of amino acid replacements, ZAA_{bias} . We then used a best subset regression, optimizing for Bayesian information criterion, using the “bestglm” R package, to search for a subset of biochemical properties of amino acids (supplementary table 2, Supplementary Material online, for the full list of tested properties and references) that, on a generalized linear model, would best predict ZAA_{bias} .

Estimate of Expected Nonsense Mutations

We used the same method as above, in order to calculate an estimate of the expected proportion of nonsense mutations. Briefly, in order to obtain the expected number of codon changes into a stop codon, when accounting for the rate of trinucleotide substitution at trinucleotides centered on 4-fold degenerate sites, we performed 10,000 simulations distributing 49,358 mutations randomly across the SARS-CoV-2 CDSs at the trinucleotide substitution rates centered around 4-fold

degenerate sites. We additionally employ a method using the rate of out of frame mutations to UAG, UGA, or UAG.

Analysis of tRNA Adaptation and Codon Usage Bias

To test if there is any evidence of selection on translational efficiency at 4-fold synonymous sites, we measured the difference in human tAI and codon usage bias caused by each of the 4-fold degenerate synonymous mutations identified in our analysis (4,064 variants) when compared against the SARS-CoV-2 reference genome. tAI per codon were obtained from the STADIUM database (Yoon et al. 2018) and codon usage tables were obtained from the CoCoPUTs database (Alexaki et al. 2019). In order to measure if any particular type of change is overrepresented when compared with random expectations, we generated 10,000 simulations of 4,064 variants across all 4-fold degenerate synonymous codons in the SARS-CoV-2 reference genome, at the same rate as the nucleotide substitution observed at 4-fold degenerate sites. *P* values of overrepresentation of each type of mutation were calculated numerically from comparing with the distribution of these simulants.

In order to account for trinucleotide mutational biases, we repeated the simulation process accounting for the rate as the nucleotide substitution observed at 4-fold degenerate sites. We first masked any 4-fold degenerate synonymous variant that was followed by a mutation or an alignment gap in the first site of the next codon in a particular strain or if a hypermutable or problematic site occurred within the codon or the first site of the next codon.

Analysis of selection on translational efficiency on within-individual variation (data described below) was performed in the same way. Briefly, we measured the difference in human tAI and RSCU bias caused by each of the 4-fold degenerate synonymous mutations identified in the within-host data set (1,208 variants), and compared it with random expectation derived from 10,000 simulations aleatorily distributing 1,208 variants across all 4-fold degenerate synonymous codons in the SARS-CoV-2 coding sequence, at the same rate as the nucleotide substitution observed at 4-fold degenerate sites in the within-individual variation data set.

Within-Individual Variation and Receptor-Binding Domain Substitution Analysis

Within-individual variants generated by Galaxy and HyPhy developments Teams (Nekrutenko et al. 2020) as part of the Galaxy Project SARS-CoV-2 data analyses (Available from: <https://covid19.galaxyproject.org/genomics/4-variation/>) were obtained from GitHub (Available from: https://github.com/galaxyproject/SARS-CoV-2/blob/4df1456e65367cf62c011c33d322643e79a9513e/genomics/4-Variation/variant_list.tsv.gz), updated on May 29, 2020 and last accessed on July 21, 2020. Known problematic sites in SARS-CoV-2

sequencing were removed as in section “Creating a Mutational Matrix” and only variants with allele frequency >5% were considered. Samples from the sequencing project with NCBI SRA Study Accession SRP253798 were removed prior to analysis as some samples from this study were noted as being dominated by C->U (>99% variants of some samples C->U, Available from: <https://virological.org/vgained-stops-in-data-from-the-peter-doherty-institute-for-infection-and-immunity/486> last accessed on July 21, 2020). Nonsense mutations were already annotated as “EFF[*].FUNCLASS = NONSENSE” and here were quantified per sample and at which position the mutations occurred in codons. To compare nonsense mutations at first and second nucleotide positions of codons, the number of codons that were one mutation from a stop codon were counted in the reference sequence (for first sites: NAA, NAG, NGA; for second sites: UNA, UNG) and a χ^2 test was performed.

Effects on binding activity of single mutations within the receptor-binding domain of SARS-CoV-2 spike protein were obtained from [supplementary table 2](#) of Starr et al. (2020). The above alignment of GISAID SARS-CoV-2 isolates was used to quantify unique amino acid substitutions at positions within this region. Within-individual variants were filtered for those within the receptor-binding domain and unique amino acid substitutions were quantified. This method has the advantage that the predicted mutational effect is called dependent on biophysics alone, rather than methods that employ sequence conservation and variant frequencies (Dunham et al. 2021) that would render the present analysis circular.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Wellcome Trust (fellowship 207507 to G.K.) and the European Research Council (advanced grant ERC-2014-ADG 669207 to L.D.H.). We acknowledge the providers of all the genomes (see [supplementary table 1](#), [Supplementary Material](#) online).

Data Availability

All genomes used in the analysis are available from GISAID. Genomes employed are listed in [supplementary data 1](#) and [table 1](#), [Supplementary Material](#) online.

Literature Cited

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38(Web Server issue):W7–W13.

- Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A.* 99(6):3695–3700.
- Alexaki A, et al. 2019. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *J Mol Biol.* 431(13):2434–2441.
- Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P. 2014. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Res.* 42(7):4527–4545.
- Bai Y, et al. 2020. Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends. *Int J Infect Dis.* 100:164–173.
- Banerjee AK, et al. 2020. SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. *Cell* 183(5):1325–1339.e1321.
- Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159(4):1779–1788.
- Charneski CA, Hontela F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical at skew in firmicute genomes results from selection and not from mutation. *PLoS Genet.* 7(9):e1002283.
- Chen F, et al. 2020. Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat Ecol Evol.* 4(4):589–600.
- Coleman JR, et al. 2008. Virus attenuation by genome-scale changes in codon pair bias. *Science* 320(5884):1784–1787.
- De Maio N, et al. 2021. Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol Evol.* 13(5). doi:10.1093/gbe/evab087.
- Dearlove B, et al. 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci U S A.* 117(38):23652–23662.
- Delignette-Muller ML, Dutang C. 2015. fitdistrplus: an R package for fitting distributions. *J Stat Soft.* 64(4):1–34.
- dos Reis M, Sawa R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32(17):5036–5044.
- Duchene S, et al. 2020. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* 6(2):veaa061.
- Dunham A, Jang GM, Muralidharan M, Swaney D, Beltrao P. 2021. A missense variant effect prediction and annotation resource for SARS-CoV-2. *bioRxiv*. Available from: 10.1101/2021.02.24.432721
- Fan RLY, et al. 2015. Generation of live attenuated influenza virus by using codon usage bias. *J Virol.* 89(21):10762–10773.
- Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. *J Mol Biol.* 47(3):238–248.
- Gaunt E, et al. 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. *Elife.* 5:e12735.
- Gills D, Massar S, Cerf NJ, Rooman M. 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol.* 2(11):research0049.
- Goldman N, Yang ZH. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.
- Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. 2021. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *IScience* 24(2):102116.
- Gu HG, Chu DKW, Peiris M, Poon LLM. 2020. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *Virus Evol.* 6(1):veaa032.
- Gu HG, Fan RLY, Wang D, Poon LLM. 2019. Dinucleotide evolutionary dynamics in influenza A virus. *Virus Evol.* 5(2):vez038.

- Haig D, Hurst LD. 1991. A quantitative measure of error minimization in the genetic code. *J Mol Evol.* 33(5):412–417.
- Heizer EM Jr, et al. 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Mol Biol Evol.* 23(9):1670–1680.
- Hernandez-Allas X, Benisty H, Schaefer MH, Serrano L. 2021. Translational adaptation of human viruses to the tissues they infect. *Cell Rep.* 34(11):108872.
- Hill VJ, Rambaut A. 2020. Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06. *Virological*. Edinburgh, UK: University of Edinburgh. Available from: <https://virological.org/phylogenetic-analysis-of-sars-cov-2-update-2020-03-06/42>.
- Hurst LD, Feil EJ, Rocha EP. 2006. Protein evolution: causes of trends in amino-acid gain and loss. *Nature* 442(7105):E11–E12; discussion E12.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Keightley PD, Eyre-Walker A. 2000. Deleterious mutations and the evolution of sex. *Science* 290(5490):331–333.
- Kim D, et al. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* 181(4):914–921.e910.
- Kogay R, Wolf YI, Koonin EV, Zhaxybayeva O. 2020. Selection for reducing energy cost of protein production drives the GC content and amino acid composition bias in gene transfer agents. *Mbio* 11:e01206–20.
- Kustin T, Stern A. 2021. Biased mutation and selection in RNA viruses. *Mol Biol Evol.* 38(2):575–588.
- Lanfear R. 2020. A global phylogeny of SARS-CoV-2 sequences from GISAID. *Zenodo*. doi: 10.5281/zenodo.3958883
- Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol Biol Evol.* 33(7):1711–1725.
- LI W-H, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* 2(2):150–174.
- Liu YS, et al. 2011. The characteristics of the synonymous codon usage in enterovirus 71 virus and the effects of host on the virus in codon usage pattern. *Infect Genet Evol.* 11(5):1168–1173.
- Long H, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.* 2(2):237–240.
- Lynch M, et al. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* 17(11):704–714.
- Lythgoe KA, et al. 2021. SARS-CoV-2 within-host diversity and transmission. *Science* 372(6539):eabg0821.
- Mahmoudabadi G, Milo R, Phillips R. 2017. Energetic cost of building a virus. *Proc Natl Acad Sci U S A.* 114(22):E4324–E4333.
- Minh BQ, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.* 37(5):1530–1534.
- Mordstein C, et al. 2020. Codon usage and splicing jointly influence mRNA localization. *Cell Syst.* 10(4):351–362.e358.
- Mordstein C, et al. 2021. Transcription, mRNA export and immune evasion shape the codon usage of viruses. *Genome Biol Evol.* doi:10.1093/gbe/evab106.
- Mugal CF, Kutschera VE, Botero-Castro F, Wolf JBW, Kaj I. 2020. Polymorphism data assist estimation of the nonsynonymous over synonymous fixation rate ratio ω for closely related species. *Mol Biol Evol.* 37(1):260–279.
- Nekrutenko A. 2020. Gained Stops in Data from The Peter Doherty Institute for Infection and Immunity [Internet]. *Virological*. Available from: <https://virological.org/t/gained-stops-in-data-from-the-peter-doherty-institute-for-infection-and-immunity/486>.
- Nekrutenko A, et al. 2020. galaxyproject/SARS-CoV-2: Second Bioniv Release [Internet]. 2020. *Zenodo*. Available from: 10.5281/zenodo.3685264.
- Nextstrain. Genomic Epidemiology of Novel Coronavirus – Global Subsampling [Internet]. 2020. Available from: <https://nextstrain.org/ncov/global?i=clock>.
- O’Fallon BD. 2010. A method to correct for the effects of purifying selection on genealogical inference. *Mol Biol Evol.* 27(10):2406–2416.
- Pathan RK, Biswas M, Khandaker MU. 2020. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. *Chaos Solitons Fractals* 138:110018–110018.
- Ponting CP. 2008. The functional repertoires of metazoan genomes. *Nat Rev Genet.* 9(9):689–698.
- Ratcliff J, Simmonds P. 2021. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* 556:62–72.
- Rice AM, et al. 2021. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. *Mol Biol Evol.* 38(1):67–83.
- Richmond RC. 1970. Non-Darwinian evolution: a critique. *Nature* 225(5237):1025–1028.
- Rocha EP, et al. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239(2):226–235.
- Schwidersky M, Rooman M, Pucci F. 2020. Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness. *BMC Biol.* 18(1):146.
- Shen Z, et al. 2020. Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. *Clin Infect Dis.* 71(15):713–720.
- Shpaer EG, Mullins JI. 1990. Selection against CpG dinucleotides in lentiviral genes – a possible role of methylation in regulation of viral expression. *Nucleic Acids Res.* 18(19):5793–5797.
- Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* 22(13):pii=30494. doi:10.2807/1560-7917.ES.2017.2222.2813.30494.
- Simmonds P. 2020a. Pervasive RNA secondary structure in the genomes of SARS-CoV-2 and other coronaviruses. *Mbio* 11(6):e01661–20.
- Simmonds P. 2020b. Rampant C → U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories. *Mosphere* 5:e00408–20.
- Starr TN, et al. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182(5):1295–1310.e1220.
- Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26(12):1569–1571.
- Swire J. 2007. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol.* 64(5):558–571.
- Tang X, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev.* 7(6):1012–1023.
- Tonkin-Hill G, et al. 2021. Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* 10:e66857.
- van Dorp L, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol.* 83:104351.
- Walt S, Colbert SC, Varoquaux G. 2011. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng.* 13(2):22–30.
- Weber CC, Whelan S. 2019. Physicochemical amino acid properties better describe substitution rates in large populations. *Mol Biol Evol.* 36(4):679–690.
- Wertheim JO, Kosakovsky Pond SL. 2011. Purifying selection can obscure the ancient age of viral lineages. *Mol Biol Evol.* 28(12):3355–3365.

Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM. 2010. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evol Biol.* 10:253.
Yang S, et al. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523(7561): 463–U187.

Yoon J, Chung YJ, Lee M. 2018. STADIUM: species-specific tRNA adaptive index compendium. *Genomics Inform.* 16(4):e28.

Zhao Z, et al. 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol.* 4:21.

Associate editor: Mario dos Reis

Supplementary information for: Causes and Consequences of Purifying Selection on SARS-CoV-2

Atahualpa Castillo Morales, Alan M. Rice, Alexander T. Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, Laurence D. Hurst

Genome Biology & Evolution, 13(10): evab196.

Supplementary data are available at Molecular Biology and Evolution online:

<https://doi.org/10.1093/gbe/evab196>.