Errors and the evolution of genomes: the role of stop codons

Abrahams, Liam

Award date:
2020

Awarding institution:
University of Bath

Errors and the evolution of genomes: the role of stop codons

Liam Richard Abrahams

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

January 2020

Copyright

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 exception of Chapter 3 which contains work that formed part of a collaboration between Rosina Savisaar, Christine Mordstein, Bethan Young and myself (contribution details have been included with the Chapter introduction).

Table of Contents

Acknowledgements

I would first like to thank Laurence for giving me the opportunity to do my PhD. Throughout the four years, he has not only provided me with the skills to perform thorough scientific research, but to also question everything. In particular, he has taught me to trust the data, even if is not what I was expecting. I am immensely grateful to him for humouring my often-ridiculous ideas and persuading me back in the right direction. It has made the four years of my research thoroughly enjoyable and go far too quickly.

I would also like to thank everyone who has been part of the lab – Alan, Alex, Atahualpa, Becky, Dana, Lucy, Joëlle, Rosina and Stefanie – and the people in the Milner Centre more generally for making my time in Bath so enjoyable.

I would like to thank the European Research Council for the funding provided to Laurence so that I could do my PhD and for funding the various meeting and conferences I have attended.

I would like to give a special thanks to Pete, Nicola, Lexi and Matilda for providing me with a bed in Bath and who without I do not think I would have been able to survive the hundreds of miles driving down the M4. I will miss the regular eating of chocolate, dodgy DIY, playing computer games, unimpressed morning conversations with Lexi but also just regularly having a good time with close friends.

Finally, I would like to thank my family for their support throughout. A special thanks goes to my fiancé Emily for encouraging me to do my PhD and for putting up with me for the last four years. For some reason you have agreed to put up with me for much longer!

Summary

The multi-step process of converting information stored in DNA to functional molecules is inherently error-prone. As we move towards an era of precision medicine, understanding why such errors occur is essential both for accurate diagnosis and therapeutic design. However, also interesting is how genomes have evolved to prevent or mitigate the deleterious consequences of such errors. In this thesis, I use stop codons as an exemplar, as their presence/absence in sequence outside of translation termination may be indicative of function. I ask two broader questions: first, are vital components that ensure accurate splicing, exonic splice enhancers (ESEs), constrained by often residing in coding sequence? If so, do these constraints apply to other sequences? I show stop codons are depleted in ESEs and this depletion is most parsimonious with functioning in CDS. Consequently, stop codons in long intergenic noncoding RNAs (lincRNAs) are also unexpectedly depleted, attributable to the presence of ESEs. This depletion appears to result in a susceptibility to nonsense mutational errors, resulting in nonsense-associated altered splicing (NAS). I find ≈6% of genome-wide nonsense mutations in healthy individuals result in exon skipping, but such an effect is probably stronger when disease-associated. Given ESE use in the human genome, I turned my attention to bacterial genomes to ask a second question: are stop codons employed as a direct error-proofing mechanism? I find bacterial genomes appear select for out of frame stop codons to terminate frameshifts based on their probability of frameshifting, and not downstream costs. Interestingly, I also show that in bacterial genes, a stop codon appears to be selected for immediately following the start codon, hypothesising that this helps the ribosome correctly initiate translation initiation. Stop codons are therefore implicated genome-wide in both preventing errors and making genes more susceptible to errors.

Abbreviations

A	adenine
bp	base pair
C	cytosine
CAI	codon adaptation index
CDS	coding sequence
CLS	capture long seq
DMD	Duchenne muscular dystrophy
eIF	eukaryotic initiation factor
EJC	exon junction complex
EMBL	European Molecular Biology Laboratory
EOR	expected observation rate
ESE	exonic splice enhancer
ESS	exonic splice silencer
FDR	false discovery rate
FE	fold-enrichment
G	guanine
GTEx	Genotype-Tissue Expression
ISE	intronic splice enhancer
ISS	intronic splice silencer
LFTR	leucyl/phenylalanyl-tRNA-protein transferase
lincRNA	long intergenic noncoding RNA
lncRNA	long noncoding RNA
loess	local regression model
MFC	multi factor complex
miRNA	microRNA
MPS VI	Mucopolysaccharidosis type VI
mRNA	messenger RNA
NAS	nonsense-associated altered splicing
NCBI	National Center for Biotechnology Information
NMD	nonsense-mediated decay

NME	N-terminal methionine excision
NMH	normalised per motif per 1,000 base pair hits
ORF	open reading frame
OSC	out of frame stop codon
PABP	polyA-binding protein
PGP-UK	Personal Genome Project - United Kingdom
PhyloCSF	phylogenetic codon substation frequency
pORF	pseudo-open reading frame
pPSI	pseudo-PSI
pPTC	pseudo-PTC
pRPMskip	pseudo-RPMskip
PSI	percentage spliced in
PTC	premature termination (stop) codon
RBP	RNA-binding protein
RF	release factor
RPMinclude	reads per million supporting exon inclusion
RPMskip	reads per million supporting exon skipping
SCD	stop codon density
SD	Shine-Dalgarno
SNP	single nucleotide polymorphism
SR protein	serine/arginine-rich protein
T	thymine
TCGA	The Cancer Genome Atlas
TF	transcription factor
TPM	tags per million
tRNA	transfer RNA
TSS	transcription start site
U	uracil
UTR	untranslated region
wt	wild type

Chapter 1

Introduction

Dual coding information driving genome evolution

On completion of the first whole genome sequence (Sanger et al. 1977), two genes of the bacteriophage φX174 were found to be encoded by the same stretch of DNA in two different reading frames. Thus, even though only one primary DNA sequence exists, this provided the first evidence that genomes can encode overlapping, or *dual coding*, layers of information. Although initially thought to characterise compacted viral genomes (Weisbeek et al. 1977; Barrell et al. 1978; Yen and Webster 1981; Belshaw et al. 2007; Chirico et al. 2010; Pavesi et al. 2018), open reading frame (ORF) overlaps are now known not to be unique to viruses (see Dan et al. 2002; Rogozin et al. 2002; Johnson and Chisholm 2004; Veeramachaneni et al. 2004; Makalowska et al. 2005; Steigele and Nieselt 2005; David et al. 2006; Sabath et al. 2008; Huvet and Stumpf 2014; Rosikiewicz et al. 2018) with estimates suggesting ≈20-40% of genes in human and mouse overlap another gene on the opposite strand (Chen et al. 2004; Zhang et al. 2006). The presence of overlapping ORFs therefore provided the first suggestion that the selective constraints acting on genes can extend beyond the need to preserve the primary peptide sequence and that genomes are not linear strings of genetic information.

It is now known that other forms of dual coding information, many facilitating gene expression, are ubiquitous among all genomes. For example, synonymous codon usage is thought to be under selection to facilitate fast and accurate translation (Ikemura 1985; Dix and Thompson 1989; Akashi 1994; Drummond et al. 2006; Stoletzki and Eyre-Walker 2007; Behura and Severson 2011; Gingold and Pilpel 2011; Doherty and McInerney 2013; Ma et al. 2014; Brandis and Hughes 2016; Frumkin et al. 2018; LaBella et al. 2019). Similarly, there is strong selection to maintain specific mRNA secondary structures (Carlini et al. 2001; Chamary and Hurst 2005b; Meyer and Miklos 2005; Kudla et al. 2006; Shabalina et al. 2006; Gu et al. 2010b; Tuller et al. 2010; Smith et al. 2013; Tuller and Zur 2015; Jacobson and Clark 2016; Gebert et al. 2019), each with the potential to have substantial effects on gene expression. The need to appropriately position nucleosomes is thought to be responsible for greater conservation of nucleosome-free "linker" sequence both at synonymous and nonsynonymous sites (Warnecke et al. 2008; Warnecke et al. 2009) and sequences

more generally (Cohanim and Haran 2009; Dai et al. 2011; Prendergast and Semple 2011; Quintales et al. 2015), in turn mediating transcription start site (TSS) choice (Dreos et al. 2016). There is also considerable support for the selection of functional microRNA (miRNA) target sites and surrounding sequence constraining the CDS in order to maintain efficient binding sites (Hurst 2006; Forman et al. 2008; Guo et al. 2008; Fang and Rajewsky 2011; Gu et al. 2012; Hausser et al. 2013; Liu et al. 2015), highlighted by the contribution of synonymous and non-synonymous single nucleotide polymorphisms (SNPs) in miRNAs to human disease (Wang et al. 2015).

This ever-growing body of evidence is therefore slowly unpicking the intricate nature of how information is incorporated and selected for within sequences. Indeed, the ability to incorporate additional information is thought to have been a driver for the structure of the genetic code itself (Freeland and Hurst 1998; Itzkovitz and Alon 2007; Itzkovitz et al. 2010). Cases where the dual coding information overlaps with the CDS are perhaps the most intriguing, as this suggests proteins may be constrained by selection pressures other than one ensuring the correct peptide sequence. As we move into an era of precision and personalised medicine, understanding the exact nature of how selection operates on genes, which may not be obvious due to dual coding information, becomes particularly pertinent for therapies to be effective.

In this regard, understanding the role that genomic errors play in the evolution of genomes, regarding both the selection of dual coding information to prevent errors and as sources of genetic novelty, is important and motivates this thesis. For the remainder of the introduction, I will introduce and provide a snapshot of the current literature concerning error-related dual coding constraints and then dive deeper into how selection for stop codons, in terms of errors specifically, constrains genes. I will then provide a summary of the work I have conducted, detailing how stop codons constrain both coding and noncoding sequence and are implicated in disease.

Error-proofing the genome

The multi-step procedure by which DNA is decoded and processed lends itself highly susceptible to errors (Drummond and Wilke 2009; Warnecke and Hurst 2011). The

initiation of gene expression, for example, might occur at the wrong time or at rates too high or low for proper functioning. Once initiation of expression has occurred, the gene might be mistranscribed or misspliced. If a coding gene, a translation error may occur or, if noncoding, the transcript may be accidentally translated. Gene products may misfold, not fold at all, interact promiscuously with other molecules or not be correctly degraded. Both proteins and non-coding RNAs be mislocated within the cell. Consequently, the deleterious consequences of errors are diverse. Misfolded proteins are frequently toxic and aggregate or interact inappropriately (Stefani and Dobson 2003) and are often implicated in neurodegenerative diseases (see Chiti and Dobson 2017; Sweeney et al. 2017). Aberrant proteins resulting from mistranslation events may also themselves be toxic, but also incur process costs due to unproductive ribosome use and clean up that may be rate-limiting for gene expression (Stoebel et al. 2008; Shachrai et al. 2010; Shah et al. 2013). Inaccuracies in splicing, whether attributable to mutations or not, are also often implicated in disease (Lopez-Bigas et al. 2005; Baralle et al. 2009; Lim et al. 2011b; Sterne-Weiler et al. 2011; Wu and Hurst 2016; Soemedi et al. 2017).

Such is the scope for errors to occur, preventing or mitigating the effects at each stage of expression is potentially important for ensuring cellular fitness. From an evolutionary perspective, although understanding why these errors occur is important, perhaps the more interesting question is how genomes have evolved to cope with them (Drummond and Wilke 2009; Warnecke and Hurst 2011). One hypothesis is that is selection is weak and leads to a "bloated" genome (as a result of small chance insertions) and thus prone to errors. As a consequence, error-proofing selection should be strong (e.g. see Wu and Hurst (2015)). Current evidence argues that such error-related selection pressures are indeed amongst the strongest drivers of genome evolution (Drummond and Wilke 2009; Warnecke and Hurst 2011). For example, the spatial ordering and orientation of genes (genome architecture) in prokaryotes and eukaryotes is highly non-random (Tamames 2001; Hurst et al. 2004; Li et al. 2006; Semon and Duret 2006; Warnecke and Hurst 2011). Such ordering is parsimoniously explained as a result of selection to limit the effects of highly deleterious fluctuations in the expression of noise-sensitive essential genes (eukaryotes) (Batada and Hurst 2007) due to intrinsically random "bursts" of expression (Raser and O'Shea 2005; Chubb et al. 2006; Raj et al. 2006) or organisation of genes into operons containing

genes that form part of the same pathway or protein complex to be expressed simultaneously (bacteria) (Wolf et al. 2001; de Daruvar et al. 2002; Rocha 2008; Saenz-Lahoya et al. 2019) or boost expression (Lim et al. 2011a). Large-scale chromosomal duplications and deletions have also been shown to increase robustness to translational errors (Kalapis et al. 2015), although this strategy also incurs considerable fitness costs.

However, a significant burden experienced during gene expression likely occurs at the transcript level on a case-by-case basis. Thus, individual sequences are under selection to limit the rate at which errors occur (error prevention) or reduced the impact of and increase robustness to them when they do occur (error mitigation) (Warnecke and Hurst 2011). Although proteins that interact with the transcript can operate with high accuracy (e.g. the spliceosome (Fox-Walsh and Hertel 2009) or ribosome codon verification (Ieong et al. 2016)), the ability for sequences to be dual coding lends itself ideal for incorporating additional regulatory error-proofing signals.

Perhaps the most often cited example of error-related dual coding is codon usage bias aiding translational accuracy. Translational errors resulting from the incorporation of the wrong transfer RNA (tRNA) typically occur at codons corresponding to rarer tRNAs (Baranov et al. 2004; Kramer and Farabaugh 2007; Laine et al. 2008; Kramer et al. 2010; Shah and Gilchrist 2010). Consequently, codon usage tends to overrepresent the codons, particularly in highly expressed genes, that more accurately match tRNA abundances (Ikemura 1985; Eyre-Walker 1996; Duret 2000; Rocha 2004; Plotkin and Kudla 2010; LaBella et al. 2019), with optimal codons often associated with conserved amino acid sites (Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008; Zhou et al. 2009). The inherent redundancy of codon usage in the genetic code also permits selection for codons robust to large-effect mistranslation errors (Archetti 2004, 2006). Furthermore, robustness against mistranslation induced misfolding or unfolding of proteins is also thought to be a dominant selective force acting on sequences and codon usage (Drummond and Wilke 2008; Zhou et al. 2009; Yang et al. 2010) with selection driving the use of translationally optimal codons at structurally sensitive sites (Zhou et al. 2009; Warnecke and Hurst 2010).

However, sequence-specific constraints also reflect selection to maintain functional regulatory binding information allowing effector molecules to act with higher fidelity and efficiency. For example, the binding of particular miRNAs to the CDS negatively regulates expression (Krek et al. 2005; Xie et al. 2005; Filipowicz et al. 2008; Bartel 2009; Reczko et al. 2012; Marin et al. 2013) and could be viewed as selection for regulating gene expression. Such functional miRNA target sites in the coding regions appear to be under purifying selection (Hurst 2006; Forman et al. 2008; Guo et al. 2008; Fang and Rajewsky 2011; Gu et al. 2012; Hausser et al. 2013; Liu et al. 2015). Similarly, it has been argued transcription factor (TF) binding sites (Stergachis et al. 2013; Birnbaum et al. 2014; Reyna-Llorens et al. 2018) in CDS are under selection, although whether such selection does really exist given similar levels of conservation between TF-bound and TF-depleted codons is debatable (Xing and He 2015; Agoglia and Fraser 2016).

More recently, splice-related dual coding constraints have received a larger focus of attention, as ensuring accurate splicing is vital given the number of genes that undergo splicing (≈97% in human Grzybowska (2012)) and as splice disruption-related errors are often associated with disease (Faustino and Cooper 2003; Baralle and Baralle 2005; Wang and Cooper 2007; Anna and Monika 2018). Exon-intron boundary discrimination was previously thought to be defined by conserved nucleotide usage at the splice sites as the splice acceptor AG and splice donor GT dinucleotides, along with the branch sites, define the majority of exon-intron boundaries (Burset et al. 2000; Black 2003). Yet, by themselves, these motifs are often not sufficient in ensuring accurate splicing (Lim and Burge 2001). Instead, exons frequently contain important dual coding regulatory elements in the form of short *cis*-acting RNA motifs either promoting (exonic splice enhancers, (ESEs)) (Blencowe 2000; Fairbrother et al. 2004b; Zhang and Chasin 2004; Goren et al. 2006; Ke et al. 2011; Caceres and Hurst 2013; Lee and Rio 2015) or inhibiting (exonic splice silencers, (ESSs)) (Wang et al. 2004; Zhang and Chasin 2004; Wang et al. 2006; Lee and Rio 2015) the inclusion of exons in the mature transcript, especially if the splice site in question is considered "weak" (Fairbrother et al. 2002; Caceres and Hurst 2013).

The constraints these motifs, ESEs in particular, impose on the CDS and sequences more generally underline the necessity of error-free splicing and champion the

hypothesis that many synonymous sites are not neutrally evolving (Chamary et al. 2006). ESEs are, for example, enriched in exons identified by exon definition (Fairbrother et al. 2004a; Zhang and Chasin 2004), whereas ESSs are more variable and define alternative splice sites (Wang et al. 2006). ESEs have been shown to be under strong purifying selection (Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley et al. 2006; Parmley and Hurst 2007; Ke et al. 2008; Sterne-Weiler et al. 2011; Savisaar and Hurst 2016) constraining both amino acid and synonymous site choice (Willie and Majewski 2004; Chamary and Hurst 2005a; Carlini and Genut 2006; Parmley et al. 2006; Caceres and Hurst 2013). Nowhere is this constraint more evident, even in noncoding sequences (Schuler et al. 2014; Haerty and Ponting 2015), than towards the ends of exons where ESE density is typically highest (Fairbrother et al. 2002; Fairbrother et al. 2004a; Fairbrother et al. 2004b). Furthermore, exons flanked by longer introns tend to be those hardest to consistently splice accurately (Bell et al. 1998; Fox-Walsh et al. 2005), contrasting with the more tightly regulated splicing of short introns (Pickrell et al. 2010). Consequently, ESE density in humans is frequently greater when the flanking intron is large (Caceres and Hurst 2013; Wu and Hurst 2015), thought to be a result of selection for stronger splice site reinforcement as the probability of encoding potential decoy splice sites, and therefore splice-related errors, is increased (Wu and Hurst 2015). This necessity for accurate splice definition has resulted in a strong selection pressure to preserve ESE motifs that significantly constrains ≈15-20% of fourfold-degenerate sites (Savisaar and Hurst 2018). Further, such is the need to ensure accurate splicing and thus the strength of selection on ESEs that the proportion of exonic sequence devoted to splicing is as strong a predictor of the rate of human protein evolution (Parmley et al. 2007).

Whilst the above cases highlight prominent examples of selection to maintain dual coding regulatory information, selective pressures to avoid particular motifs may be equally as important. In CDSs, the distribution of conservation P-values (that is, the enrichment over expected in dinucleotide matched control motifs) of RNA-binding motifs (RBPs) more generally (including, but not restricted to ESEs) is bimodal (Savisaar and Hurst 2017). This, most probably, is a result of the location binding-specifics (CDS or non-CDS) of each RBP, such that the depletion of classes of RBP motifs that do not function in the CDS is likely due to purifying selection to avoid

inappropriate binding. Similarly, bacterial CDSs are depleted of motifs that resemble the Shine-Dalgarno (SD) sequence, likely reflecting selection to prevent inappropriate ribosome binding or stalling within the CDS (Li et al. 2012; Diwan and Agashe 2016; Yang et al. 2016). The CDSs of *Saccharomyces cerevisiae*, *Escherichia coli* and *Caenorhabditis elegans* are also depleted in mononucleotide repeats (Ackermann and Chao 2006; Gu et al. 2010a) prone to transcriptional slippage (Wagner et al. 1990).

The above examples demonstrate dual coding information is crucial to error-proofing genomes and under selection. How then, can we further elucidate this phenomenon? One might argue that searching for constraints and patterns within genes such as those above without a clear direction would be like searching for a needle in a haystack. Yet, stop codons are also dual coding, having previously been shown to have error-related roles. Thus, any patterns of stop codon selection beyond the end of CDSs may provide opportunities to identify cases of error-proofing selection as they should be operating for reasons other than the expected termination of translation.

Error-related stop codon selection

The primary role of stop codons is well understood (see Nakamura and Ito 1998; Bertram et al. 2001; Dever and Green 2012). Specific release factors recognise a stop codon (TAA, TAG, TGA) entering the ribosome but, although all three stop codons are employed across the domains of life, the release factors that recognise each stop codon vary. Eukaryotes, for example, employ a single class I release factor (RF), eRF1, to identify all three stop codons (Frolova et al. 1999; Frolova et al. 2000; Song et al. 2000; Schmeing and Ramakrishnan 2009; Kryuchkova et al. 2013; Beissel et al. 2019) with a second class II release factor, eRF3, stimulating eRF1 activity in the presence of GTP (Zhouravleva et al. 1995; Salas-Marco and Bedwell 2004; Alkalaeva et al. 2006). Prokaryotes instead employ three RFs - RF1 recognising TAA and TAG, RF2 recognising TAA and TGA and RF3 facilitating RF1/RF2 dissociation (Scolnick et al. 1968; Freistroffer et al. 1997). These translation termination systems are well conserved, except for the occasional stop codon reassignments in, for example, the mitochondria of several eukaryotes (Barrell et al. 1979; Campbell et al. 2013; Ivanova et al. 2014), bacteria (Inamine et al. 1990; Tate et al. 1999; Campbell et al. 2013) and

ciliates (Lozupone et al. 2001; Sánchez-Silva et al. 2003). Surprisingly, some ciliates lack dedicated stop codons altogether (Swart et al. 2016).

Although the majority of translation events are terminated as expected, ribosomal readthrough errors of stop codons occur with regularity. For example, experimental estimates range from an error in one in 10^{-2} translation events to one in 10^{-5} in bacteria (Sambrook et al. 1967; Roth 1970; Strigini and Brickman 1973; Bossi 1983; Cridge et al. 2018; Li and Zhang 2019)) and are thus at orders of magnitude higher than estimated mutation rates (mutation rate estimates range from of 10^{-9} per base per year in mammals (Kumar and Subramanian 2002), 10^{-4} to 10^{-10} per base per generation in the yeast *S. cerevisiae* (Lang and Murray 2008; Zhu et al. 2014) and 10^{-10} per base per generation in the bacteria *E. coli* (Drake 1991; Garibyan et al. 2003; Lee et al. 2012; Jee et al. 2016)). As selection against mutations is a significant driver of genome evolution, it is logical to assume selection also acts to minimise errors associated with stop codons themselves. Indeed, stop codons in bacteria are not selectively equivalent (Povolotskaya et al. 2012) with selection for TAA (Belinky et al. 2018) particularly in highly expressed genes (Korkmaz et al. 2014; Trotta 2016; Wei et al. 2016). A localised nucleotide preference for +4U (+1 being the first base of the stop codon) also appears to improve termination efficiency (Brown et al. 1990; Poole et al. 1998; Wei and Xia 2017).

However, while selection acts on the stop codons themselves to minimise canonical termination errors, other examples are dual coding signals. For example, evidence from *S. cerevisiae* (Liang et al. 2005), *Arabidopsis thaliana* (Kochetov et al. 2011) and two ciliate species (Adachi and Cavalcanti 2009) suggests additional stop codons are selected for in 3' untranslated regions (UTRs). These stop codons are thought to act as "backstop" stop codons, providing a second chance to terminate translation after readthrough or mutation to the canonical stop codon. However, a lack of selection in prokaryotes (Major et al. 2002; Ho and Hurst 2019) leaves it unclear to what extent backstop codons are under selection more generally. Such sweeping generalisations also assume readthrough has a particularly detrimental fitness cost. Similarly stop codons in any reading frame in 5' UTRs immediately upstream of canonical start codons are thought to terminate translation events where the ribosome has yet to reach the recognised start codon (Seligmann 2007). Such codons thereby increase translation

efficiency and robustness to incorrectly initiation translation events. Hidden off-frame stop codons within the CDS are also thought to be selected for in the event the ribosome ends up translating an incorrect reading frame following a frameshift, termed the "ambush hypothesis" (Seligmann and Pollock 2004; Singh and Pardasani 2009; Tse et al. 2010). However, more recent evidence suggests this may be attributable to GC biases (Morgens et al. 2013).

The above examples mitigate the effects of potentially deleterious translation events. Yet, further cases are subtler and more intricate. Premature termination codons (PTCs), in-frame stop codons located before the canonical stop codon, may arise from heritable germline nonsense mutations or be created by errors in transcription or splicing. During the pioneer round of translation, components of the nonsense-mediated decay (NMD) pathway recognises PTCs in eukaryotes and target the offending transcript for degradation (Maquat and Carmichael 2001), catching and preventing the synthesis of potentially toxic truncated peptides. Thus, stop codons provides a signal to highlight mistranscribed or misspliced transcripts. Curiously, *Paramecium tetraurelia* is depleted in introns with length divisible by three (3n) when compared with 3n + 1 and 3n + 2 introns (18.7%, 42.3% and 39.0% of the total respectively) (Jaillon et al. 2008). However, unlike 3n + 1 and 3n + 2 introns that would likely introduce a PTC in the downstream exons by a frameshift, a 3n intron would retain the ORF during the translation of the intron-containing mature transcript. Not only are 3n introns avoided, but there also appears to be selection to avoid 3n introns with no stops or, put differently, selection appears to favour introns if retained would be detected by NMD or induce a frameshift. Evidence from *A. thaliana*, *Homo. sapiens*, *C. elegans*, *Drosophila melanogaster*, *Schizosaccharomyces pombe* (Jaillon et al. 2008) and *Yarrowia lipolytica* (Mekouar et al. 2010) all demonstrate a significant deficit of 3n introns lacking in-frame stop codons, suggesting a strongly conserved selective pressure concerning stop codons to prevent translation of intron-retained transcripts.

The effectiveness of NMD in catching erroneous transcripts, however, is highly organism-specific. In mammalian transcripts, PTCs located further than ≈50-55 bases upstream of exon junction complexes (EJCs) are typically detected via the more

efficient intron-dependent pathway (reviewed in Chang et al. (2007); Isken and Maquat (2007); Brogna and Wen (2009); Lykke-Andersen and Jensen (2015); Kurosaki et al. (2019)). However, in mammals single-exon genes appear to be insensitive to NMD (Maquat and Li 2001; Brocke et al. 2002) and the last exons of genes only employ the less-efficient intron-independent pathway (Buhler et al. 2006; Matsuda et al. 2007; Metze et al. 2013). How do single-exon human genes compensate? The answer is transcriptional robustness – single-exon genes have ≈8% decrease in the use of codons in close one-step proximity to a stop codon (Cusack et al. 2011). This robustness also applies to last-exons (≈7%) and histone genes invisible to NMD (32%). Thus, rather than selection for stop codons, selection to avoid stop codons in instances when they cannot be caught also shapes the evolution of genomes.

The above examples demonstrate how stop codons can provide ideal markers in the vast quantity of genetic information to establish patterns of selection. Searching for constraints related to stop codons may therefore help answer questions about how error-proofing shapes genome evolution. Given ESEs also impart strong selective constraints, they too might provide further insight. In this thesis, I therefore ask two broad questions. The first examines the interplay between ESEs and stop codons. As ESEs are key components in preventing deleterious splicing but also function in CDS, is their composition constrained by the need for CDSs to avoid stop codons in at least one reading frame? If so, are there consequences? Second, can I find evidence within genomes for stop codon selection consistent with being error-proofing mechanisms?

Unexpected depletions of stop codons in lincRNA sequences as a result of SR proteins binding both coding and noncoding sequence

In eukaryotes, and particularly humans, the role ESEs play in ensuring splicing accuracy is well documented (Blencowe 2000; Fairbrother et al. 2004b; Zhang and Chasin 2004; Goren et al. 2006; Ke et al. 2011; Caceres and Hurst 2013; Lee and Rio 2015). Such is this requirement, sequences containing ESEs are subject to strong purifying selection (Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley et al. 2006; Parmley and Hurst 2007; Ke et al. 2008; Sterne-Weiler et al. 2011; Savisaar and Hurst 2016, 2018). When we consider sequence evolution in the context of such

motifs, we typically focus on the constraints imposed by the RNA-binding protein (RBP) to ensure appropriate binding - in the case of ESEs, serine/arginine-rich (SR) proteins. This is particularly noteworthy in light of the work by Savisaar and Hurst (2017) who demonstrate that sequences are selectively constrained by the need to both maintain desired binding but to restrict binding of the unwanted RBPs. Yet, this perspective only portrays a part of the picture. Two components are involved in RBP binding: the binding motif in the target sequence and the RNA-recognition motif in the RBP protein itself. Thus, while selection may act on the target sequence to preserve the binding site, there may exist selective constraints on RBPs to preferentially bind motifs characteristic of the sequence to which they are required to bind. ESEs, for example, are dual coding motifs and thus have to able to correctly function within CDS to define the amino acid sequences and therefore subject to protein-coding constraints. The composition of ESE motifs is therefore likely to be an amalgamation of multiple competing selective constraints.

This raises an interesting question. Are ESEs constrained by the requirement to also function within CDS? For example, one could expect that SR proteins have evolved to recognise motifs depleted in stop codons as the sequences in which they reside, by definition, cannot contain stop codons in at least one of the three reading frames (although in principle can also apply to other motifs of CDS-binding proteins). I therefore hypothesise that, as a consequence of being dual coding, ESEs may be depleted in stop codons. This reasonably assumes that as ESEs can function in any reading frame, stop codons should be avoided in all reading frames of ESEs.

I start by asking whether the set of INT3 ESEs (Caceres and Hurst 2013) is depleted in stop codons when compared with all other remaining hexamers. I find this to be the case (≈ 43% depletion). This depletion is supported by a significant reduction in stop codon density (SCD) in the real ESE motifs when compared with sets of dinucleotide-matched pseudo motifs. I find depletion not a result of nucleotide biases, but consistent with their exonic (and CDS) location. I therefore conclude the depletion in ESEs appears to be stop codon specific, but not ESE specific, likely as a result of being dual coding and subject to the protein-coding constraint. I was intrigued as to whether this constraint has more extensive consequences. Noncoding intron-containing transcripts

are spliced like their protein-coding counterparts (reviewed in Will and Luhrmann 2011; De Conti et al. 2013; Quinn and Chang 2016), with a key selective constraint of long intergenic noncoding RNA (lincRNA) sequences is to preserve ESEs (Schuler et al. 2014; Haerty and Ponting 2015). Are noncoding sequences, as a result of containing ESEs, also depleted in stop codons? If so, is this due to ESEs? Without the context of the stop codon depletion in ESEs, this hypothesis might seem counterintuitive, as with no protein-coding constraint there should be no reason for stop codons in lincRNA to be avoided.

As predicted, I find a significant depletion of stop codons in multi-exon lincRNA sequences with this depletion specific to exonic and not intronic regions of lincRNAs. This depletion is reduced in single-exon genes where selection to include splice-related ESEs should be weaker. Importantly, I find that stop codon density is lowest in the exonic regions where ESE density is highest and that the lincRNA sequence not predicted to be ESE is significantly increased in stop codons. The selective constraint imposed upon ESEs as a result of having to function within CDS therefore appears to transfer an unexpected constraint on lincRNA sequences, which I have termed "transfer selection". I find this has interesting implications for sequence annotation, notably that almost 10% of lincRNA sequences exceed the typically used 300 bp ORF threshold used for identifying noncoding transcripts for reasons beyond chance.

This work therefore provides an important case study in how dual coding motifs are constrained beyond their role to facilitate binding and how RBPs evolve their binding preferences to the sequences they bind. It also provides an example of a novel mode of selection, one I have termed *transfer selection*, by which one class of genes are indirectly constrained by the constraints imposed on another class of genes. The depletion of stop codons in noncoding sequence may also be significant when considering of the evolution of *de novo* genes that recently has gained attention as a source of genetic novelty (Tautz and Domazet-Loso 2011; McLysaght and Guerzoni 2015; McLysaght and Hurst 2016).

Stop codons as a source of splice and variant classification errors

The depletion of stop codons in ESEs found in Chapter 2 may, however, have important practical implications. As stop codons are found infrequently in ESEs, a mutation creating a stop codon in an ESE is likely to break the binding utility of the motif. Thus, as a result of ESEs being dual coding and the density of ESEs required in sequences to ensure accurate splicing, sequences may be susceptible to mutations creating stop codons and disruption of splicing. However, what if a mutation occurs that creates a stop codon occurs in frame?

Typically, nonsense mutations would be classified as having deleterious effects due to protein truncation or NMD of the offending transcript (Maquat 2005; Brogna and Wen 2009). Indeed, such PTCs are the cause of a variety of diseases (see Linde and Kerem 2008 for an overview). Yet, if a nonsense mutation also happens to disrupt an ESE, the deleterious effect may instead be splice-related. In the literature, this phenomenon is known as nonsense-associated alternative splicing (NAS) (Gibson et al. 1993; Dietz and Kendzior 1994; Hull et al. 1994; Endo et al. 1995; Messiaen et al. 1997; Shiga et al. 1997; Valentine and Heflich 1997; Hoffmeyer et al. 1998; Mazoyer et al. 1998; Melis et al. 1998; Valentine 1998; Gersappe and Pintel 1999; Ars et al. 2000; Wimmer et al. 2000; Di Blasi et al. 2001; Caputi et al. 2002; Li et al. 2002; Wang et al. 2002a; Wang et al. 2002b; Pagani et al. 2003; Pasmooij et al. 2004; Vuoristo et al. 2004; Zatkova et al. 2004; Mendive et al. 2005; Stasia et al. 2005; Disset et al. 2006; Aznarez et al. 2007; Laimer et al. 2008; Chemin et al. 2010; Littink et al. 2010; Lenassi et al. 2014; Peterlongo et al. 2015; Barny et al. 2018; Meldau et al. 2018).

Many of the published examples above describe single-gene cases of NAS. However, given ESEs are employed genome-wide, it is unlikely NAS is restricted to particular genes. Furthermore, given the evolutionary importance placed on accurate splicing, NAS may be a common source of pathogenicity. To date, there is no study investigating the genome-wide frequency of NAS. In the context of the work in Chapter 2, in which the rarity of stop codons in ESEs is evidenced, such a survey is overdue.

This model that is implicit above is dependent on the disruption of splice motifs. In the literature, however, there are two competing models to describe the mechanism of NAS: a splice-motif disruption model (Shiga et al. 1997; Valentine 1998; Liu et al.

2001; Caputi et al. 2002; Pagani et al. 2003; Zatkova et al. 2004; Aznarez et al. 2007; Peterlongo et al. 2015) and a nuclear scanning model (Dietz and Kendzior 1994; Gersappe and Pintel 1999; Mendell and Dietz 2001; Li et al. 2002; Wang et al. 2002a; Wang et al. 2002b; Shi et al. 2015). If NAS is a genome-wide phenomenon, contributed to by the depletion of stop codons in ESEs as hypothesised, evidence would need to be in favour the first of these hypotheses.

Chapter 3 therefore presents the work of a collaborative project predominantly with Rosina Savisaar. We conducted the first genome-wide NAS investigation in humans employing polymorphism and transcriptomic data from 462 individuals from the 1000 Genomes Project (Lappalainen et al. 2013; The 1000 Genomes Project Consortium 2015). We ask whether we can discriminate between the two competing models, not only to provide evidence either supporting or against our splice disruption hypothesis but to help consolidate the field. In addition to the draft manuscript that I have prepared and presented here, Rosina presented a preliminary version as part of her thesis (Savisaar 2018).

We find that premature termination codons (PTCs) are significantly associated with exon skipping, but in the opposite direction to that expected (i.e. increased skipping in the genes of non-PTC-containing individuals). I find that this is likely a statistical artefact of many differences in the levels of exon inclusion between the PTC-containing (PTC-/+) and non-PTC-containing (PTC-/-) individuals (ΔPSI) being very small. However, when plotting the distributions of the differences according to whether the presence of the PTC causes an increase or decrease of exon skipping separately, the majority of large effect cases and those likely to be biologically meaningful are consistent with NAS. After controlling for the nucleotide composition of mutations that generate PTCs, we also find a significant increase in exon skipping. This skipping is not a result of NMD continually degrading PTC-containing transcripts. By only considering PTCs for which there is both a relative (percentage spliced in, PSI) and absolute (reads per million skipped, RPMskip) increase in exon skipping when associated with a PTC is greater than 5%, we find 30/541 (5.55%) of PTCs affect splicing. Comparisons with mutations that create out of frame stop codons suggest that not only is the effect reading frame independent, but that stop codon

creating mutations out of frame also frequently disrupt splicing. This result is therefore most parsimonious with the disruption of splice motifs and lends evidence against the nuclear scanning model for which the effect would be reading frame dependent. I have collaborated with Christine Mordstein and Grzegorz Kudla from the University of Edinburgh to provide experimental validation of our top NAS candidate.

This work therefore provides a practical application and consequence of the evolutionary question addressed in Chapter 2. An important caveat with this work, however, is that it was performed on a population of healthy individuals in which PTCs are unlikely to segregate at high levels as if the PTCs were particularly pathogenic the individuals would be unlikely to survive. Perhaps of greater interest is whether pathogenic nonsense mutations also disrupt splicing. Being able to quantify how often nonsense mutations disrupt splicing, and thus how frequently such mutations may be misclassified as classical nonsense mutations, is of clinical importance. I have therefore conducted an analysis using data from a curated source of pathological mutations (ClinVar, Landrum et al. (2018)) to provide an estimate of how frequently pathogenic NAS may occur. I estimate that in the region of up to ≈33% of pathogenic nonsense mutations could affect splicing. Moreover, using the neural network algorithm MMsplice (Cheng et al. 2019), ≈80% of these pathogenic nonsense mutations are predicted to have a negative effect on how frequently an exon is included. While purely computational, taken together with the experimental validation of our computational findings, at the very least these results demonstrate more in-depth analyses must be undertaken to ensure accurate mutation classification.

A refined dual-strategy mode of selection for out of frame stop codons in bacteria to correct for translation errors

The work in Chapter 2 and Chapter 3 suggests that selection on genes to optimise appropriate and accurate splicing minimises the number of stop codons found in any frame of multi-exon sequences, although this has the unfortunate consequence that mutations that create stop codons do disrupt splicing at a non-negligible rate. These results therefore argue against the conventional understanding of the ambush hypothesis (Seligmann and Pollock 2004) - that genes should be optimised for

incorporating out of frame stop codons to minimise translational frameshift costs. Upon reflection, this is not surprising given the fitness advantage of ensuring accurate splicing of every transcript is almost certain to outweigh the fitness cost associated with catching ribosomal frameshifts.

The avoidance of stop codons in human sequences, however, does not necessarily provide evidence against the ambush hypothesis, but instead suggests it may take a more refined form and be relatively organism-specific. The ambush hypothesis was derived from studies of bacteria that lack more complex processes such as splicing. In such genomes, the ability to select for out of frame stop codons (OSCs) may provide a significant fitness advantage, particularly given translational frameshift errors are estimated to occur once at every 3.3×10^{-5} codons (Parker 1989; Farabaugh and Bjork 1999). The complexities of different genomes may therefore dictate the strategy by which the most accurate gene expression can be achieved and consequently, whether or not stop codons are selected for. If so, it may help explain why Singh and Pardasani (2009) find very few positive correlations between genome codon usage and the codons' contribution to OSCs in the genomes of vertebrates, primates, rodents or other mammals in either the nuclear or mitochondrial genes but found results consistent with Seligmann and Pollock (2004) in bacteria.

Thus, to address the second question as to whether I can find evidence of stop codon selection, the best genomes to use may be bacterial genomes. However, upon reviewing the literature, there appeared no concise and confirmatory evidence in bacteria either way as to the extent of selection for OSCs. For example, several studies claim codon usage biases are consistent with OSC selection (Seligmann and Pollock 2004; Singh and Pardasani 2009), but their supporting evidence is weak (only 37% and 7% of genomes demonstrate such biases in the studies respectively). Critically, a fundamental limitation of these studies is that they do not interrogate the actual frequencies of OSCs, especially given the relationship between the ability of a codon to form an OSC and its genome-wide usage can be explained almost entirely by GC content (Morgens et al. 2013). The current state of the literature therefore appears to be somewhat confusing and contradictory. Furthermore, an important limitation I raise is whether the previous studies appropriately model bacterial sequence constraints. For

example, as OSCs are dual coding they must ensure correct protein-coding functionality, yet the Markov models employed do not preserve amino acid content. Is there then, a real effect where OSCs are selected for in bacteria, or are the results in the current literature reflective of the models used and the biases they impose? I therefore sought to provide clarity on whether bacterial genomes do select for OSCs (Chapter 4) as a result of not having to specify splice regulatory information, by taking advantage of a larger number of bacterial genomes.

I proposed and tested three alternative simulation models, each of which has its limitations, but maintaining more properties of the real sequences. Each model returns largely similar results – that there are genomes with OSC excesses, but OSC excess is often significantly negatively correlated with GC content. Many genomes also frequently demonstrate depletions of OSCs. Furthermore, excesses of TAA and TGA are often more limited to the AT-rich genomes and reduced when compared with sense codons of similar nucleotide composition. This somewhat contradicts previous thinking, where one might expect stronger selection in GC-rich genomes where the occurrence of an OSC by chance is reduced. A further test negating any biases due to simulation design uses repeats of two isoleucine (ATA, ATC and ATT codons) and two valine (GTA, GTC, GTG and GTT codons) codons. I hypothesise that, if there is selection for OSCs, the synonymous site of the first codon should be under selection to use a nucleotide that encodes a +1 OSC. I find limited evidence consistent with this, but results may be biased by localised mutation biases.

These results therefore appear to only add to the confusion of previous work. One curiosity, however, remains. Why is it the AT-rich genomes, in which OSCs would be more likely to occur by chance, are those with the apparent selection for OSCs? The ambush hypothesis, arguing for stronger selection in GC-rich genomes because off-frame translation would continue for longer in GC-rich genomes by chance, only considers the processivity costs of such errors. However, to consider the cost dynamics of a frameshift error, one must consider not only the processivity costs after committing an error but also how likely it is that the frameshift error occurs. I adopted the "process cost of accidental frameshift" model from Warnecke et al. (2010) that incorporates information about the genome tRNA repertoire to calculate the possibility of frameshifting, on the basis that the tRNA repertoire is important in determining

translational accuracy (Baranov et al. 2004; Shah and Gilchrist 2010; Warnecke et al. 2010) and that enrichment of the tRNA repertoire is correlated with reduced frameshift susceptibility (Warnecke et al. 2010). I find that not only are sequences in AT-rich genomes more susceptible to frameshifting, but that this susceptibility is significantly positively correlated with the extent of OSC excess. Thus, the excess and apparent selection of OSCs in AT-rich genomes may be explained by an increased propensity to frameshift, rather than an increased cost post frameshift as classically assumed. This means that, unexpectedly, the evolution of bacterial CDSs to combat frameshift errors appears to reflect one of two largely GC-dependent strategies – either reduce the likelihood of frameshifting (GC-rich genomes) or, select for the dual coding, error mitigation OSCs.

Stop codon selection to prevent ribosomal false starts

The questions addressed in the preceding chapters consider stop codons and errors at a genome-wide scale but also while considering genes as a whole, i.e. the analyses have searched all regions of genes, exons or CDS. The work in Chapter 4 demonstrates that at the genome-wide scale, we can detect signals of OSC selection but only in AT-rich bacteria. However, it is unlikely that strong signals of selection would arise through such analyses as any signal may be diluted by noisy signals elsewhere. An alternative approach is to consider a more localised approach and search regions of genes that would most benefit from preventing or mitigating errors. For example, the presence of U directly after the CDS stop codon in bacteria provides an immediate signal in the noncoding sequence to improve termination efficiency (Brown et al. 1990; Poole et al. 1998; Wei and Xia 2017) and is entirely site-specific. The work in Chapter 5 therefore adopts this more localised approach and asks whether there is any evidence of site-specific selection for OSCs.

Previous work suggests ribosomes do not always initiate correctly at the start codon (Seligmann 2007). A logical position for an OSC is therefore immediately following the start codon as if translation initiation begins on an incorrect reading frame it would be immediately terminated. The first indication of potential selection is that in the CDSs of bacteria from 646 bacterial species, each sampled from a unique genus, the

fourth CDS reside (i.e. following the start codon, typically ATGN, and the first nucleotide of the second CDS codon) is more often than expected an adenine (A) nucleotide. In the most extreme case in the *Polaribacter* sp., 63.26% of CDSs have A at the fourth site. Even in the most GC-rich genome (*Streptobacillus* sp., GC = 0.113), for which using A at the fourth site would be most difficult, 35.00% of sequences have fourth site A. Almost all genomes (640/646 = 99.07%) have an excess of A at this site compared with what is expected - astonishing given that under the simplest null model, only 25% of fourth sites should be an A. A simple explanation is a bias due to an oversampling of GC-rich genomes. However, when I compare the ratio of fourth site A use to genome A-codon use for each genome, I find ratios are significantly negatively correlated with GC3 content arguing against this hypothesis.

I therefore consider several alternative hypotheses. Neither a preference of amino acids that have A-starting codons (Stenstrom et al. 2001; Bivona et al. 2010; Shemesh et al. 2010) nor simply selection for the destabilisation of mRNA secondary structure (Qing et al. 2003; Kudla et al. 2009; Gu et al. 2010b; Bentele et al. 2013; Goodman et al. 2013) are consistent with such a strong fourth site A bias. Yet, as start codons are almost exclusively of the form NTG, I hypothesised that the fourth site A might indeed provide a mechanism by which frameshifts can be immediately corrected, as the sequence would read NTGA and encode an OSC in the +1 reading frame immediately following the start codon. As stated above, this hypothesis makes biological sense. I show that in the subset of bacterial genomes for which TGA instead encodes tryptophan, enrichment at the fourth site is only at levels consistent with reducing RNA stability and indicating the A is likely to function to create a stop codon. Consistent with a regulatory role, fourth site A usage is reduced in sequences with a SD translation initiation regulatory signal known to increase initiation efficiency (Shine and Dalgarno 1974; Di Giacco et al. 2008). I further speculate more specifically as to how the OSC might improve translation initiation. I conclude that two models, immediately terminating translation after initiation or as a regulatory “stop and adjust” mechanism where the ribosome is assisting in locating the correct initiation codon, are most parsimonious.

References

Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. *PLoS Genet.* 2:e22.

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

Agoglia RM, Fraser HB. 2016. Disentangling Sources of Selection on Exonic Transcriptional Enhancers. *Mol. Biol. Evol.* 33:585-590.

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

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* 125:1125-1136.

Anna A, Monika G. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* 59:253-268.

Archetti M. 2004. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *J. Mol. Evol.* 59:258-266.

Archetti M. 2006. Genetic robustness and selection at the protein level for synonymous codons. *J. Evol. Biol.* 19:353-365.

Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X. 2000. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* 9:237-247.

Aznarez I, Zielenski J, Rommens JM, Blencowe BJ, Tsui LC. 2007. Exon skipping through the creation of a putative exonic splicing silencer as a consequence of the cystic fibrosis mutation R553X. *J. Med. Genet.* 44:341-346.

Baralle D, Baralle M. 2005. Splicing in action: assessing disease causing sequence changes. *J. Med. Genet.* 42:737-748.

Baralle D, Lucassen A, Buratti E. 2009. Missed threads. The impact of pre-mRNA splicing defects on clinical practice. *EMBO Rep* 10:810-816.

Baranov PV, Gesteland RF, Atkins JF. 2004. P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA* 10:221-230.

Barny I, Perrault I, Michel C, Soussan M, Goudin N, Rio M, Thomas S, Attie-Bitach T, Hamel C, Dollfus H, et al. 2018. Basal exon skipping and nonsense-associated altered splicing allows bypassing complete CEP290 loss-of-function in individuals with unusually mild retinal disease. *Hum. Mol. Genet.* 27:2689-2702.

Barrell BG, Bankier AT, Drouin J. 1979. A different genetic code in human mitochondria. *Nature* 282:189-194.

Barrell BG, Shaw DC, Walker JE, Northrop FD, Godson GN, Fiddes JC. 1978. Overlapping genes in bacteriophages phiX174 and G4. *Biochem. Soc. Trans.* 6:63-67.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* 136:215-233.

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* 39:945-949.

Behura SK, Severson DW. 2011. Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in Aedes aegypti and Anopheles gambiae. *Insect Mol. Biol.* 20:177-187.

Beissel C, Neumann B, Uhse S, Hampe I, Karki P, Krebber H. 2019. Translation termination depends on the sequential ribosomal entry of eRF1 and eRF3. *Nucleic Acids Res.* 47:4798-4813.
Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci Rep* 8:9260.
Bell MV, Cowper AE, Lefranc MP, Bell JI, Screaton GR. 1998. Influence of intron length on alternative splicing of CD44. *Mol. Cell. Biol.* 18:5930-5941.
Belshaw R, Pybus OG, Rambaut A. 2007. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.* 17:1496-1504.
Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* 9:675.
Bertram G, Innes S, Minella O, Richardson JP, Stansfield I. 2001. Endless possibilities: translation termination and stop codon recognition. *Microbiology* 147:255-269.
Birnbaum RY, Patwardhan RP, Kim MJ, Findlay GM, Martin B, Zhao J, Bell RJ, Smith RP, Ku AA, Shendure J, et al. 2014. Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genet.* 10:e1004592.
Bivona L, Zou Z, Stutzman N, Sun PD. 2010. Influence of the second amino acid on recombinant protein expression. *Protein Expr Purif* 74:248-256.
Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72:291-336.
Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 25:106-110.
Bossi L. 1983. Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. *J. Mol. Biol.* 164:73-87.
Brandis G, Hughes D. 2016. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLoS Genet.* 12:e1005926.
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.* 11:331-335.
Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* 16:107-113.
Brown CM, Stockwell PA, Trotman CN, Tate WP. 1990. The signal for the termination of protein synthesis in procaryotes. *Nucleic Acids Res.* 18:2079-2086.
Buhler M, Steiner S, Mohn F, Paillusson A, Muhlemann O. 2006. EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length. *Nat. Struct. Mol. Biol.* 13:462-464.
Burset M, Seledtsov IA, Solovyev VV. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28:4364-4375.
Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.
Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Soll D, Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci U S A* 110:5540-5545.
Caputi M, Kendzior RJ, Jr., Beemon KL. 2002. A nonsense mutation in the fibrillin-1 gene of a Marfan syndrome patient induces NMD and disrupts an exonic splicing enhancer. *Genes Dev.* 16:1754-1759.

Carlini DB, Chen Y, Stephan W. 2001. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics* 159:623-633.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* 62:89-98.

Chamary JV, Hurst LD. 2005a. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.* 21:256-259.

Chamary JV, Hurst LD. 2005b. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 6:R75.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7:98-108.

Chang YF, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* 76:51-74.

Chemin G, Tinguely A, Sirac C, Lechouane F, Duchez S, Cogne M, Delpy L. 2010. Multiple RNA surveillance mechanisms cooperate to reduce the amount of nonfunctional Ig kappa transcripts. *J. Immunol.* 184:5009-5017.

Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* 32:4812-4820.

Cheng J, Nguyen TYD, Cygan KJ, Celik MH, Fairbrother WG, Avsec Z, Gagneur J. 2019. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* 20:48.

Chirico N, Vianelli A, Belshaw R. 2010. Why genes overlap in viruses. *Proc Biol Sci* 277:3809-3817.

Chiti F, Dobson CM. 2017. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu Rev Biochem* 86:27-68.

Chubb JR, Trcek T, Shenoy SM, Singer RH. 2006. Transcriptional pulsing of a developmental gene. *Curr. Biol.* 16:1018-1025.

Cohanim AB, Haran TE. 2009. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res.* 37:6466-6476.

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.* 46:1927-1944.

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

Dai Z, Dai X, Xiang Q. 2011. Genome-wide DNA sequence polymorphisms facilitate nucleosome positioning in yeast. *Bioinformatics* 27:1758-1764.

Dan I, Watanabe NM, Kajikawa E, Ishida T, Pandey A, Kusumi A. 2002. Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution. *Nucleic Acids Res.* 30:2906-2910.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* 103:5320-5325.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* 4:49-60.

de Daruvar A, Collado-Vides J, Valencia A. 2002. Analysis of the cellular functions of Escherichia coli operons and their conservation in Bacillus subtilis. *J. Mol. Evol.* 55:211-221.

Dever TE, Green R. 2012. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb Perspect Biol* 4:a013706.

Di Blasi C, He Y, Morandi L, Cornelio F, Guicheney P, Mora M. 2001. Mild muscular dystrophy due to a nonsense mutation in the LAMA2 gene resulting in exon skipping. *Brain* 124:698-704.

Di Giacco V, Marquez V, Qin Y, Pech M, Triana-Alonso FJ, Wilson DN, Nierhaus KH. 2008. Shine-Dalgarno interaction prevents incorporation of noncognate amino acids at the codon following the AUG. *Proc Natl Acad Sci U S A* 105:10715-10720.

diCenzo GC, Finan TM. 2017. The Divided Bacterial Genome: Structure, Function, and Evolution. *Microbiol. Mol. Biol. Rev.* 81:e00019-00017.

Dietz HC, Kendzior RJ, Jr. 1994. Maintenance of an open reading frame as an additional level of scrutiny during splice site selection. *Nat Genet* 8:183-188.

Disset A, Bourgeois CF, Benmalek N, Claustres M, Stevenin J, Tuffery-Giraud S. 2006. An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. *Hum. Mol. Genet.* 15:999-1013.

Diwan GD, Agashe D. 2016. The Frequency of Internal Shine-Dalgarno-like Motifs in Prokaryotes. *Genome Biol Evol* 8:1722-1733.

Dix DB, Thompson RC. 1989. Codon choice and gene expression: synonymous codons differ in translational accuracy. *Proc Natl Acad Sci U S A* 86:6888-6892.

Doherty A, McInerney JO. 2013. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Mol. Biol. Evol.* 30:2263-2267.

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

Dreos R, Ambrosini G, Bucher P. 2016. Influence of Rotational Nucleosome Positioning on Transcription Start Site Selection in Animal Promoters. *PLoS Comput Biol* 12:e1005144.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* 23:327-337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341-352.

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

Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16:287-289.

Endo F, Awata H, Katoh H, Matsuda I. 1995. A nonsense mutation in the 4-hydroxyphenylpyruvic acid dioxygenase gene (Hpd) causes skipping of the constitutive exon and hypertyrosinemia in mouse strain III. *Genomics* 25:164-169.

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? *Mol. Biol. Evol.* 13:864-872.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007-1013.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.* 32:W187-190.

Fang Z, Rajewsky N. 2011. The impact of miRNA target sites in coding sequences and in 3'UTRs. *PLoS One* 6:e18067.

Farabaugh PJ, Bjork GR. 1999. How translational accuracy influences reading frame maintenance. *EMBO J.* 18:1427-1434.

Faustino NA, Cooper TA. 2003. Pre-mRNA splicing and human disease. *Genes Dev.* 17:419-437.

Filipowicz W, Bhattacharyya SN, Sonenberg N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 9:102-114.

Forman JJ, Legesse-Miller A, Coller HA. 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci U S A* 105:14879-14884.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A* 102:16176-16181.

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

Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. *J. Mol. Evol.* 47:238-248.

Freistroffer DV, Pavlov MY, MacDougall J, Buckingham RH, Ehrenberg M. 1997. 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.* 16:4126-4133.

Frolova LY, Merkulova TI, Kisselev LL. 2000. Translation termination in eukaryotes: Polypeptide release factor eRF1 is composed of functionally and structurally distinct domains. *Rna-a Publication of the Rna Society* 6:381-390.

Frolova LY, Tsivkovskii RY, Sivolobova GF, Oparina NY, Serpinsky OI, Blinov VM, Tatkov SI, Kisselev LL. 1999. Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. *RNA* 5:1014-1020.

Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci U S A* 115:E4940-E4949.

Garibyan L, Huang T, Kim M, Wolff E, Nguyen A, Nguyen T, Diep A, Hu K, Iverson A, Yang H, et al. 2003. Use of the rpoB gene to determine the specificity of base substitution mutations on the Escherichia coli chromosome. *DNA Repair (Amst)* 2:593-608.

Gebert D, Jehn J, Rosenkranz D. 2019. Widespread selection for extremely high and low levels of secondary structure in coding sequences across all domains of life. *Open Biol* 9:190020.

Gersappe A, Pintel DJ. 1999. A premature termination codon interferes with the nuclear function of an exon splicing enhancer in an open reading frame-dependent manner. *Mol. Cell. Biol.* 19:1640-1650.

Gibson RA, Hajianpour A, Murer-Orlando M, Buchwald M, Mathew CG. 1993. A nonsense mutation and exon skipping in the Fanconi anaemia group C gene. *Hum. Mol. Genet.* 2:797-799.

Gilbert W. 1978. Why genes in pieces? *Nature* 271:501.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* 7:481.

Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342:475-479.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol. Cell* 22:769-781.

Grzybowska EA. 2012. Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem. Biophys. Res. Commun.* 424:1-6.

Gu T, Tan S, Gou X, Araki H, Tian D. 2010a. Avoidance of long mononucleotide repeats in codon pair usage. *Genetics* 186:1077-1084.

Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. *Mol. Biol. Evol.* 29:3037-3044.

Gu W, Zhou T, Wilke CO. 2010b. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* 6:e1000664.

Guo X, Gui Y, Wang Y, Zhu QH, Helliwell C, Fan L. 2008. Selection and mutation on microRNA target sequences during rice evolution. *BMC Genomics* 9:454.

Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* 21:333-346.

Harrison PW, Lower RP, Kim NK, Young JP. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol.* 18:141-148.

Hausser J, Syed AP, Bilen B, Zavolan M. 2013. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* 23:604-615.

Hidalgo O, Pellicer J, Christenhusz MJM, Schneider H, Leitch IJ. 2017. Genomic gigantism in the whisk-fern family (Psilotaceae): Tmesipteris obliqua challenges record holder Paris japonica. *Bot. J. Linn. Soc.* 183:509-514.

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:e1008386.

Hoffmeyer S, Nurnberg P, Ritter H, Fahsold R, Leistner W, Kaufmann D, Krone W. 1998. Nearby stop codons in exons of the neurofibromatosis type 1 gene are disparate splice effectors. *Am. J. Hum. Genet.* 62:269-277.

Hull J, Shackleton S, Harris A. 1994. The stop mutation R553X in the CFTR gene results in exon skipping. *Genomics* 19:362-364.

Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J. Mol. Evol.* 63:174-182.

Hurst LD, Pal C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5:299-310.

Huvet M, Stumpf MP. 2014. Overlapping genes: a window on gene evolvability. *BMC Genomics* 15:721.

Ieong KW, Uzun U, Selmer M, Ehrenberg M. 2016. Two proofreading steps amplify the accuracy of genetic code translation. *Proc Natl Acad Sci U S A* 113:13744-13749.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2:13-34.

Inamine JM, Ho KC, Loechel S, Hu PC. 1990. Evidence that UGA is read as a tryptophan codon rather than as a stop codon by Mycoplasma pneumoniae, Mycoplasma genitalium, and Mycoplasma gallisepticum. *J. Bacteriol.* 172:504-506.

Isken O, Maquat LE. 2007. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev.* 21:1833-1856.

Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* 17:405-412.

Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. *Genome Res.* 20:1582-1589.

Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. 2014. Stop codon reassignments in the wild. *Science* 344:909-913.

Jacobson GN, Clark PL. 2016. Quality over quantity: optimizing co-translational protein folding with non-'optimal' synonymous codons. *Curr. Opin. Struct. Biol.* 38:102-110.

Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Segurens B, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* 451:359-362.

Jee J, Rasouly A, Shamovsky I, Akivis Y, Steinman SR, Mishra B, Nudler E. 2016. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* 534:693-696.

Johnson ZI, Chisholm SW. 2004. Properties of overlapping genes are conserved across microbial genomes. *Genome Res.* 14:2268-2272.

Kalapis D, Bezerra AR, Farkas Z, Horvath P, Bodi Z, Daraba A, Szamecz B, Gut I, Bayes M, Santos MA, et al. 2015. Evolution of Robustness to Protein Mistranslation by Accelerated Protein Turnover. *PLoS Biol.* 13:e1002291.

Katsir L, Zhepu R, Piasezky A, Jiang J, Sela N, Freilich S, Bahar O. 2018. Genome Sequence of "Candidatus Carsonella ruddii" Strain BT from the Psyllid Bactericera trigonica. *Genome Announc* 6:e01466-01417.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360-1374.

Ke S, Zhang XH, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res.* 18:533-543.

Kinniburgh AJ, Mertz JE, Ross J. 1978. The precursor of mouse β-globin messenger RNA contains two intervening RNA sequences. *Cell* 14:681-693.

Kochetov AV, Volkova OA, Poliakov A, Dubchak I, Rogozin IB. 2011. Tandem termination signal in plant mRNAs. *Gene* 481:1-6.

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.* 289:30334-30342.

Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* 13:87-96.

Kramer EB, Vallabhaneni H, Mayer LM, Farabaugh PJ. 2010. A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA* 16:1797-1808.

Krawiec S, Riley M. 1990. Organization of the bacterial chromosome. *Microbiol Rev* 54:502-539.

Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, et al. 2005. Combinatorial microRNA target predictions. *Nat Genet* 37:495-500.

Kryuchkova P, Grishin A, Eliseev B, Karyagina A, Frolova L, Alkalaeva E. 2013. Two-step model of stop codon recognition by eukaryotic release factor eRF1. *Nucleic Acids Res.* 41:4573-4586.

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:e180.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255-258.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* 99:803-808.

Kurosaki T, Popp MW, Maquat LE. 2019. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat. Rev. Mol. Cell Biol.* 20:406-420.

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2019. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet.* 15:e1008304.

Laimer M, Onder K, Schlager P, Lanschuetzer CM, Emberger M, Selhofer S, Hintner H, Bauer JW. 2008. Nonsense-associated altered splicing of the Patched gene fails to suppress carcinogenesis in Gorlin syndrome. *Br. J. Dermatol.* 159:222-227.

Laine S, Thouard A, Komar AA, Rossignol JM. 2008. Ribosome can resume the translation in both +1 or -1 frames after encountering an AGA cluster in Escherichia coli. *Gene* 412:95-101.

Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T. 2007. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.* 8:104-115.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46:D1062-D1067.

Lang GI, Murray AW. 2008. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics* 178:67-82.

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506-511.

Lee H, Popodi E, Tang H, 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:E2774-2783.

Lee Y, Rio DC. 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem* 84:291-323.

Lenassi E, Saihan Z, Bitner-Glindzicz M, Webster AR. 2014. The effect of the common c.2299delG mutation in USH2A on RNA splicing. *Exp Eye Res* 122:9-12.

Li B, Wachtel C, Miriami E, Yahalom G, Friedlander G, Sharon G, Sperling R, Sperling J. 2002. Stop codons affect 5' splice site selection by surveillance of splicing. *Proc Natl Acad Sci U S A* 99:5277-5282.
Li C, Zhang J. 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS Genet.* 15:e1008141.
Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484:538-541.
Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX. 2006. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol* 2:e74.
Liang H, Cavalcanti AR, Landweber LF. 2005. Conservation of tandem stop codons in yeasts. *Genome Biol* 6:R31.
Lim HN, Lee Y, Hussein R. 2011a. Fundamental relationship between operon organization and gene expression. *Proc Natl Acad Sci U S A* 108:10626-10631.
Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011b. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A* 108:11093-11098.
Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* 98:11193-11198.
Linde L, Kerem B. 2008. Introducing sense into nonsense in treatments of human genetic diseases. *Trends Genet.* 24:552-563.
Littink KW, Pott JW, Collin RW, Kroes HY, Verheij JB, Blokland EA, de Castro Miro M, Hoyng CB, Klaver CC, Koenekoop RK, et al. 2010. A novel nonsense mutation in CEP290 induces exon skipping and leads to a relatively mild retinal phenotype. *Investigative ophthalmology & visual science* 51:3646-3652.
Liu G, Zhang R, Xu J, Wu CI, Lu X. 2015. Functional conservation of both CDS- and 3'-UTR-located microRNA binding sites between species. *Mol. Biol. Evol.* 32:623-628.
Liu HX, Cartegni L, Zhang MQ, Krainer AR. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* 27:55-58.
Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. 2005. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* 579:1900-1903.
Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol.* 11:65-74.
Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* 16:665-677.
Ma L, Cui P, Zhu J, Zhang Z, Zhang Z. 2014. Translational selection in human: more pronounced in housekeeping genes. *Biol Direct* 9:17.
Major LL, Edgar TD, Yee Yip P, Isaksson LA, Tate WP. 2002. Tandem termination signals: myth or reality? *FEBS Lett.* 514:84-89.
Makalowska I, Lin CF, Makalowski W. 2005. Overlapping genes in vertebrate genomes. *Comput. Biol. Chem.* 29:1-12.
Maquat LE. 2005. Nonsense-mediated mRNA decay in mammals. *J. Cell Sci.* 118:1773-1776.
Maquat LE, Carmichael GG. 2001. Quality control of mRNA function. *Cell* 104:173-176.
Maquat LE, Li X. 2001. Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. *RNA* 7:445-456.

Marin RM, Sulc M, Vanicek J. 2013. Searching the coding region for microRNA targets. *RNA* 19:467-474.
Matsuda D, Hosoda N, Kim YK, Maquat LE. 2007. Failsafe nonsense-mediated mRNA decay does not detectably target eIF4E-bound mRNA. *Nat. Struct. Mol. Biol.* 14:974.
Mazoyer S, Puget N, Perrin-Vidoz L, Lynch HT, Serova-Sinilnikova OM, Lenoir GM. 1998. A BRCA1 nonsense mutation causes exon skipping. *Am. J. Hum. Genet.* 62:713-715.
McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* 370:20140332.
McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17:567-578.
Mekouar M, Blanc-Lenfle I, Ozanne C, Da Silva C, Cruaud C, Wincker P, Gaillardin C, Neuveglise C. 2010. Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biol* 11:R65.
Meldau S, De Lacy RJ, Riordan GTM, Goddard EA, Pillay K, Fieggen KJ, Marais AD, Van der Watt GF. 2018. Identification of a single MPV17 nonsense-associated altered splice variant in 24 South African infants with mitochondrial neurohepatopathy. *Clin. Genet.* 93:1093-1096.
Melis MA, Muntoni F, Cau M, Loi D, Puddu A, Boccone L, Mateddu A, Cianchetti C, Cao A. 1998. Novel nonsense mutation (C-->A nt 10512) in exon 72 of dystrophin gene leading to exon skipping in a patient with a mild dystrophinopathy. *Hum. Mutat.* Suppl 1:S137-138.
Mendell JT, Dietz HC. 2001. When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell* 107:411-414.
Mendive FM, Rivolta CM, Gonzalez-Sarmiento R, Medeiros-Neto G, Targovnik HM. 2005. Nonsense-associated alternative splicing of the human thyroglobulin gene. *Mol Diagn* 9:143-149.
Messiaen L, Callens T, De Paepe A, Craen M, Mortier G. 1997. Characterisation of two different nonsense mutations, C6792A and C6792G, causing skipping of exon 37 in the NF1 gene. *Hum. Genet.* 101:75-80.
Metze S, Herzog VA, Ruepp MD, Muhlemann O. 2013. Comparison of EJC-enhanced and EJC-independent NMD in human cells reveals two partially redundant degradation pathways. *RNA* 19:1432-1448.
Meyer IM, Miklos I. 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.* 33:6338-6348.
Morgens DW, Chang CH, Cavalcanti AR. 2013. Ambushing the Ambush Hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. *BMC Genomics* 14:418.
Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* 314:267.
Nakamura Y, Ito K. 1998. How protein reads the stop codon and terminates translation. *Genes Cells* 3:265-278.
Pagani F, Buratti E, Stuani C, Baralle FE. 2003. Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. *J. Biol. Chem.* 278:26580-26588.

Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev* 53:273-298.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* 23:301-309.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* 24:1600-1603.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.
Pasmooij AM, van Zalen S, Nijenhuis AM, Kloosterhuis AJ, Zuiderveen J, Jonkman MF, Pas HH. 2004. A very mild form of non-Herlitz junctional epidermolysis bullosa: BP180 rescue by outsplicing of mutated exon 30 coding for the COL15 domain. *Experimental dermatology* 13:125-128.
Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, Belshaw R, Firth A, Karlin D. 2018. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS One* 13:e0202513.
Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* 164:10-15.
Peterlongo P, Catucci I, Colombo M, Caleca L, Mucaki E, Bogliolo M, Marin M, Damiola F, Bernard L, Pensotti V, et al. 2015. FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. *Hum. Mol. Genet.* 24:5345-5355.
Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 6:e1001236.
Plotkin JB, Kudla G. 2010. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*
Poole ES, Major LL, Mannering SA, Tate WP. 1998. 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.* 26:954-960.
Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biol Direct* 7:30.
Prendergast JG, Semple CA. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res.* 21:1777-1787.
Qing G, Xia B, Inouye M. 2003. Enhancement of translation initiation by A/T-rich sequences downstream of the initiation codon in Escherichia coli. *J. Mol. Microbiol. Biotechnol.* 6:133-144.
Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17:47-62.
Quintales L, Soriano I, Vazquez E, Segurado M, Antequera F. 2015. A species-specific nucleosomal signature defines a periodic distribution of amino acids in proteins. *Open Biol* 5:140218.
Rada-Iglesias A, Grosveld FG, Papantonis A. 2018. Forces driving the three-dimensional folding of eukaryotic genomes. *Mol. Syst. Biol.* 14:e8214.
Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. 2006. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4:e309.
Raser JM, O'Shea EK. 2005. Noise in gene expression: origins, consequences, and control. *Science* 309:2010-2013.

Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. 2012. Functional microRNA targets in protein coding sequences. *Bioinformatics* 28:771-776.
Reyna-Llorens I, Burgess SJ, Reeves G, Singh P, Stevenson SR, Williams BP, Stanley S, Hibberd JM. 2018. Ancient duons may underpin spatial patterning of gene expression in C4 leaves. *Proc Natl Acad Sci U S A* 115:1931-1936.
Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* 14:2279-2286.
Rocha EP. 2008. The organization of the bacterial genome. *Annu. Rev. Genet.* 42:211-233.
Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV. 2002. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.* 18:228-232.
Rosikiewicz W, Suzuki Y, Makalowska I. 2018. OverGeneDB: a database of 5' end protein coding overlapping genes in human and mouse genomes. *Nucleic Acids Res.* 46:D186-D193.
Roth JR. 1970. Uga Nonsense Mutations in Salmonella-Typhimurium. *J. Bacteriol.* 102:467-475.
Sabath N, Graur D, Landan G. 2008. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. *Biol Direct* 3:36.
Saenz-Lahoya S, Bitarte N, Garcia B, Burgui S, Vergara-Irigaray M, Valle J, Solano C, Toledo-Arana A, Lasa I. 2019. Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. *Proc Natl Acad Sci U S A* 116:1733-1738.
Salas-Marco J, Bedwell DM. 2004. GTP hydrolysis by eRF3 facilitates stop codon decoding during eukaryotic translation termination. *Mol. Cell. Biol.* 24:7769-7778.
Sambrook JF, Fan DP, Brenner S. 1967. A strong suppressor specific for UGA. *Nature* 214:452-453.
Sánchez-Silva Ro, Villalobo E, Morin Lc, Torres A. 2003. A New Noncanonical Nuclear Genetic Code. *Curr. Biol.* 13:442-447.
Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687-695.
Savisaar R. 2018. The dual coding of RNA and protein level information within open reading frames. [[University of Bath]: University of Bath.
Savisaar R, Hurst LD. 2016. Purifying Selection on Exonic Splice Enhancers in Intronless Genes. *Mol. Biol. Evol.* 33:1396-1418.
Savisaar R, Hurst LD. 2017. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Mol. Biol. Evol.* 34:1110-1126.
Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* 28:1442-1454.
Schmeing TM, Ramakrishnan V. 2009. What recent ribosome structures have revealed about the mechanism of translation. *Nature* 461:1234-1242.
Schuler A, Ghanbarian AT, Hurst LD. 2014. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* 31:3164-3183.
Scolnick E, Tompkins R, Caskey T, Nirenberg M. 1968. Release factors differing in specificity for terminator codons. *Proc Natl Acad Sci U S A* 61:768-774.

Seligmann H. 2007. Cost minimization of ribosomal frameshifts. *J. Theor. Biol.* 249:162-167.
Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23:701-705.
Semon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* 23:1715-1723.
Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148:458-472.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* 34:2428-2437.
Shachrai I, Zaslaver A, Alon U, Dekel E. 2010. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. *Mol. Cell* 38:758-767.
Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. *Cell* 153:1589-1601.
Shah P, Gilchrist MA. 2010. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet.* 6:e1001128.
Shemesh R, Novik A, Cohen Y. 2010. Follow the leader: preference for specific amino acids directly following the initial methionine in proteins of different organisms. *Genomics Proteomics Bioinformatics* 8:180-189.
Shi M, Zhang H, Wang LT, Zhu CL, Sheng K, Du YH, Wang K, Dias A, Chen S, Whitman M, et al. 2015. Premature termination codons are recognized in the nucleus in a reading-frame-dependent manner. *Cell Discovery* 1:15001.
Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M. 1997. Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. *J. Clin. Invest.* 100:2204-2210.
Shine J, Dalgarno L. 1974. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* 71:1342-1346.
Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: Evidences for phylogenetic trends. *Comput. Biol. Chem.* 33:239-244.
Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* 41:8220-8236.
Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, Fairbrother WG. 2017. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet* 49:848-855.
Song H, Mugnier P, Das AK, Webb HM, Evans DR, Tuite MF, Hemmings BA, Barford D. 2000. The crystal structure of human eukaryotic release factor eRF1--mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. *Cell* 100:311-321.
Stasia MJ, Bordigoni P, Floret D, Brion JP, Bost-Bru C, Michel G, Gatel P, Durant-Vital D, Voelckel MA, Li XJ, et al. 2005. Characterization of six novel mutations in the CYBB gene leading to different sub-types of X-linked chronic granulomatous disease. *Hum. Genet.* 116:72-82.
Stefani M, Dobson CM. 2003. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J Mol Med (Berl)* 81:678-699.

Steigele S, Nieselt K. 2005. Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. *Nucleic Acids Res.* 33:5034-5044.

Stenstrom CM, Jin H, Major LL, Tate WP, Isaksson LA. 2001. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in Escherichia coli. *Gene* 263:273-284.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 342:1367-1372.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* 21:1563-1571.

Stoebel DM, Dean AM, Dykhuizen DE. 2008. The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products. *Genetics* 178:1653-1660.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol. Biol. Evol.* 24:374-381.

Strigini P, Brickman E. 1973. Analysis of specific misreading in Escherichia coli. *J. Mol. Biol.* 75:659-672.

Suwanto A, Kaplan S. 1989. Physical and genetic mapping of the Rhodobacter sphaeroides 2.4.1 genome: presence of two unique circular chromosomes. *J. Bacteriol.* 171:5850-5859.

Swart EC, Serra V, Petroni G, Nowacki M. 2016. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. *Cell* 166:691-702.

Sweeney P, Park H, Baumann M, Dunlop J, Frydman J, Kopito R, McCampbell A, Leblanc G, Venkateswaran A, Nurmi A, et al. 2017. Protein misfolding in neurodegenerative diseases: implications and strategies. *Transl Neurodegener* 6:6.

Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol* 2:research0020.

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. Biokhimiia* 64:1342-1353.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692-702.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68-74.

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

Trucksis M, Michalski J, Deng YK, Kaper JB. 1998. The Vibrio cholerae genome contains two unique circular chromosomes. *Proc Natl Acad Sci U S A* 95:14464-14469.

Tse H, Cai JJ, Tsoi HW, Lam EP, Yuen KY. 2010. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics* 11:491.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A* 107:3645-3650.

Tuller T, Zur H. 2015. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.* 43:13-28.
Valentine CR. 1998. The association of nonsense codons with exon skipping. *Mutat. Res.* 411:87-117.
Valentine CR, Heflich RH. 1997. The association of nonsense mutation with exon-skipping in hprt mRNA of Chinese hamster ovary cells results from an artifact of RT-PCR. *RNA* 3:660-676.
Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I. 2004. Mammalian overlapping genes: the comparative perspective. *Genome Res.* 14:280-286.
Vuoristo MM, Pappas JG, Jansen V, Ala-Kokko L. 2004. A stop codon mutation in COL11A2 induces exon skipping and leads to non-ocular Stickler syndrome. *Am. J. Med. Genet. A* 130A:160-164.
Wagner LA, Weiss RB, Driscoll R, Dunn DS, Gesteland RF. 1990. Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. *Nucleic Acids Res.* 18:3529-3535.
Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 8:749-761.
Wang J, Chang YF, Hamilton JI, Wilkinson MF. 2002a. Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol. Cell* 10:951-957.
Wang J, Hamilton JI, Carter MS, Li S, Wilkinson MF. 2002b. Alternatively spliced TCR mRNA induced by disruption of reading frame. *Science* 297:108-110.
Wang Y, Qiu C, Cui Q. 2015. A Large-Scale Analysis of the Relationship of Synonymous SNPs Changing MicroRNA Regulation with Functionality and Disease. *Int J Mol Sci* 16:23545-23555.
Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* 119:831-845.
Wang Z, Xiao X, Van Nostrand E, Burge CB. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell* 23:61-70.
Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet.* 4:e1000250.
Warnecke T, Huang Y, Przytycka TM, Hurst LD. 2010. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. *Genome Biol Evol* 2:636-645.
Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. *Mol. Syst. Biol.* 6:340.
Warnecke T, Hurst LD. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat. Rev. Genet.* 12:875-881.
Warnecke T, Weber CC, Hurst LD. 2009. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem. Soc. Trans.* 37:756-761.
Wei Y, Wang J, Xia X. 2016. Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. *Mol. Biol. Evol.* 33:2357-2367.
Wei Y, Xia X. 2017. The Role of +4U as an Extended Translation Termination Signal in Bacteria. *Genetics* 205:539-549.
Weisbeek PJ, Borrias WE, Langeveld SA, Baas PD, Van Arkel GA. 1977. Bacteriophage phiX174: gene A overlaps gene B. *Proc Natl Acad Sci U S A* 74:2504-2508.

Will CL, Luhrmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 3:a003707.
Willie E, Majewski J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* 20:534-538.
Wimmer K, Eckart M, Stadler PF, Rehder H, Fonatsch C. 2000. Three different premature stop codons lead to skipping of exon 7 in neurofibromatosis type I patients. *Hum. Mutat.* 16:90-91.
Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11:356-372.
Wu X, 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.* 32:1847-1861.
Wu X, Hurst LD. 2016. Determinants of the Usage of Splice-Associated cis-Motifs Predict the Distribution of Human Pathogenic SNPs. *Mol. Biol. Evol.* 33:518-529.
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434:338-345.
Xing K, He X. 2015. Reassessing the "duon" hypothesis of protein evolution. *Mol. Biol. Evol.* 32:1056-1062.
Yang C, Hockenberry AJ, Jewett MC, Amaral LAN. 2016. Depletion of Shine-Dalgarno Sequences Within Bacterial Coding Regions Is Expression Dependent. *G3 (Bethesda, Md.)* 6:3467-3474.
Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* 6:421.
Yen TS, Webster RE. 1981. Bacteriophage f1 gene II and X proteins. Isolation and characterization of the products of two overlapping genes. *J. Biol. Chem.* 256:11259-11265.
Zatkova A, Messiaen L, Vandenbroucke I, Wieser R, Fonatsch C, Krainer AR, Wimmer K. 2004. Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1. *Hum. Mutat.* 24:491-501.
Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46:D754-D761.
Zhang XH, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* 18:1241-1250.
Zhang Y, Liu XS, Liu QR, Wei L. 2006. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.* 34:3465-3475.
Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol. Biol. Evol.* 26:1571-1580.
Zhouravleva G, Frolova L, Le Goff X, Le Guellec R, Inge-Vechtomov S, Kisselev L, Philippe M. 1995. Termination of translation in eukaryotes is governed by two interacting polypeptide chain release factors, eRF1 and eRF3. *EMBO J.* 14:4065-4072.
Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A* 111:E2310-2318.

Zuerner RL, Herrmann JL, Saint Girons I. 1993. Comparison of genetic maps for two Leptospira interrogans serovars provides evidence for two chromosomes and intraspecies heterogeneity. *J. Bacteriol.* 175:5445-5451.

Chapter 2

A depletion of stop codons in lincRNA is owing to transfer of selective constraint from coding sequences

Liam Abrahams and Laurence D. Hurst

Molecular Biology and Evolution, in press. msz299.

This chapter contains work accepted for publication on 16th December 2019 at Molecular Biology and Evolution, the original and sole place of publication. The latest version of the published article can be found by following the address: https://academic.oup.com/mbe/advance-article-abstract/doi/10.1093/molbev/msz299/5678793.

Chapter 2 contains content from the published article, but reformatted for this thesis. This chapter contains analysis of publicly available data. The data and custom scripts 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://academic.oup.com/journals/pages/access_purchase/rights_and_permissions/publication_rights).

This declaration concerns the article entitled:			
A depletion of stop codons in lincRNA is owing to transfer of selective constraint from coding sequences			
Publication status (tick one) **Draft manuscript** [ ] **Submitted** [ ] **In review** [ ] **Accepted** [ ] **Published** [X]			
Publication details (reference)	Liam Abrahams, Laurence D Hurst, A depletion of stop codons in lincRNA is owing to transfer of selective constraint from coding sequences, *Molecular Biology and Evolution*, msz299. In press.		
Copyright status (tick the appropriate statement) I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [X]			
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% Bioinformatics 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**	

Abstract

While the constraints on a gene's sequence are often assumed to reflect the functioning of that gene, here we propose *transfer selection*, a constraint operating on one class of genes transferred to another, mediated by shared binding factors. We show that such transfer can explain an otherwise paradoxical depletion of stop codons in long intergenic noncoding RNAs (lincRNAs). Serine/arginine-rich (SR) proteins direct the splicing machinery by binding exonic splice enhancers (ESEs) in immature mRNA. As coding exons cannot contain stop codons in one reading frame, stop codons should be rare within ESEs. We confirm that the stop codon density (SCD) in ESE motifs is low, even accounting for nucleotide biases. Given that SR proteins binding ESEs also facilitate lincRNA splicing, a low SCD could transfer to lincRNAs. As predicted, multi-exon lincRNA exons are depleted in stop codons, a result not explained by open reading frame (ORF) contamination. Consistent with transfer selection, stop codon depletion in lincRNAs is most acute in exonic regions with the highest ESE density, disappears when ESEs are masked, is consistent with stop codon usage skews in ESEs and is diminished in both single-exon lincRNAs and introns. Owing to low SCD, the maximum lengths of pseudo-ORFs (pORFs) frequently exceed null expectations. This has implications for ORF annotation and the evolution of *de novo* protein-coding genes from lincRNAs. We conclude that not all constraints operating on genes need be explained by the functioning of the gene but may instead be transferred owing to shared binding factors.

Introduction

When considering the evolution of a gene or protein we assume, often implicitly, that sequence constraints within that gene are important in terms of the functioning of its RNA/protein products. For example, when we observe constraint on a protein domain within any given protein, we trivially assume it to be a result of the domain being important for the function of that protein. The same logic extends beyond protein motifs to RNA level features such as microRNA pairing sites. The assumption that features of genes or proteins exist to enable the functioning of that gene or protein appears so self-evidently correct that it is difficult to comprehend that there may be selectively constrained features of genes that do not reflect the functioning of the gene in question, except for overlapping genes. In this paper, we suggest that compositional patterns observed in some genes may instead be explained by a transfer of a selective constraint from one class of gene to another. We present an exemplar theoretical instance and show that it makes correct predictions of otherwise paradoxical sequence features.

Our exemplar considers the stop codon density (SCD) in long intergenic noncoding RNAs (lincRNAs). We define codon density as the number of nucleotide positions constituted by the codon in question in any frame of a given sequence, divided by the total number of nucleotides in the sequence. For example, in the sequence AGATAGGGGA, the GGA codon (AGATAGG**GGA**) has a density of 0.3. By counting each nucleotide within the queried sequence only once, the density is bound by the limits zero and one (for example, the density of the codon GGG in the same sequence AGATA**GGGG**A is 0.4). We can extend our density calculation to codon sets, by considering groupings of more than one codon whose density we calculate together as per single codon cases. For example, the di-codon set [GAT, GGG] defines 7/10 positions (A**GAT**A**GGGG**A) and has a density of 0.7. Thus, we define SCD as the positions comprised of the tri-codon set [TAA, TAG, TGA]. The sequence GG**TGATAA**CA, for example, has SCD equal to 0.6.

Unlike coding sequence (CDS) that is constrained to one in-frame stop codon per sequence, lincRNAs have no comparable constraint. The SCD in lincRNAs should

therefore be predictable from underlying nucleotide content. However, we argue that a particular mode of selection, which might be termed *transfer selection*, would result in lower stop codon usage than expected. Our argument is simple. Exonic splice enhancers (ESEs), typically short hexameric motifs occurring towards exon ends (within ≈70 bp of the splice site) (Berget 1995; Fairbrother et al. 2002; Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley et al. 2006; Parmley et al. 2007; Woolfe et al. 2010; Caceres and Hurst 2013) act as binding sites in the immature mRNA for serine/arginine-rich (SR) proteins to help direct the splice machinery. As ESEs overlap CDS, they cannot introduce an in-frame stop codon. Consequently, it seems highly likely that ESEs functioning in CDS are under selection to contain no or few stop codons. If SR proteins bind the same or similar ESEs in multi-exon coding and noncoding transcripts, the need to employ ESEs in lincRNAs should mean a depletion of stop codons in CDS ESEs transfers to lincRNA ESEs, despite stop codons in lincRNA having no translational function. In short, the binding preferences of SR proteins in CDS may transfer a necessary constraint operating on CDS to an unnecessary and otherwise paradoxical sequence constraint operating in noncoding sequences.

Many of the assumptions of our model are robust. First, lincRNA transcripts containing introns are processed similarly to protein-coding pre-mRNA transcripts (reviewed in Will and Luhrmann 2011; De Conti et al. 2013). Although SR protein binding is reported to be ≈30% less efficient in lincRNA than in protein-coding exons, evidence suggests the same SR proteins bind both gene classes as the binding of SR proteins SRSF2, SRSF5, and SRSF6 in lincRNA all improve splicing efficiency (Krchnakova et al. 2019). Second, ESEs are under purifying selection in both CDS and lincRNA, indicative of functionality. In CDS, this is illustrated by decreased rates of evolution at both synonymous (Fairbrother et al. 2002; Carlini and Genut 2006; Chamary et al. 2006; Parmley et al. 2006; Parmley and Hurst 2007; Sterne-Weiler et al. 2011; Caceres and Hurst 2013; Savisaar and Hurst 2018) and nonsynonymous sites (Parmley et al. 2006; Parmley et al. 2007) and the relative lack of single nucleotide polymorphisms (SNPs) (Majewski and Ott 2002; Fairbrother et al. 2004a; Carlini and Genut 2006; Caceres and Hurst 2013) within ESEs. This selection is not modest and, indeed, the proportion of exonic sequence devoted to governing splicing,

predominantly moderated by selection for ESEs, predicts the rate of human protein evolution as well as the amount a gene is expressed, the phylogenetically universal best predictor (Parmley et al. 2007). Similarly, purifying selection on ESEs is thought to explain most lincRNA constraint (Schuler et al. 2014; Haerty and Ponting 2015).

To test this model of transfer selection, we start by asking whether the SCD in ESE motifs is unusually low. We find this to be the case, even when controlling for the nucleotide composition of ESEs. We then ask whether, in contrast to *a priori* expectation, lincRNA sequences are also relatively depleted in stop codons and, if so, whether ESEs are the cause. We show that lincRNAs do contain fewer stop codons than expected given their nucleotide content. We provide several lines of evidence to support the hypothesis that this is due to the presence of ESEs and not open reading frame (ORF) sequence contamination.

Selective avoidance of stop codons could, at first sight, be misinterpreted as evidence that any given lincRNA is an unrecognized coding gene. As the low density of stop codons in lincRNAs ensures that the longest possible ORF is longer than expected under null models, our finding has ramifications for transcript annotation. We show that the typically used threshold of minimal ORF size (300 bp) causes a high (≈10%) false-positive rate if used in isolation. While the dearth of stop codons could confuse annotation, it might also have consequences for *de novo* gene origination via erroneous translation of noncoding RNA as accidental peptides can be longer than expected.

Results

If our model of transfer selection has validity, results must be consistent with several predictions. First, for any motif that functions within CDS, the protein-coding constraint requires it to contain no stop codons in one of the three reading frames. Thus, stop codons should be relatively rare in ESE motifs that have to reside in CDS. The same need not be true of motifs that function exclusively in introns or non-coding exons. Second, any rarity should be specific to the set of stop codons and not peculiarities resulting from motif set choice or motif functionality. Third, stop codons

should also be depleted in lincRNA sequences after accounting for their nucleotide content, this depletion being attributable to ESEs. We test each of these predictions.

ESEs are depleted in stop codons

To address the first prediction, we first consider the SCD in the "gold-standard" (low false-positive) INT3 ESE motif set (N = 84 hexamers), for which each motif was identified in at least three of four high-throughput ESE datasets (Caceres and Hurst 2013) (see Materials and Methods for an overview of how each ESE set was derived). The raw INT3 SCD is 0.054, lower than the SCD of 0.094 for the 4,012 possible hexamers not found in the INT3 set. This low SCD in the INT3 set is significantly lower than SCDs of 10,000 iterations of 84 hexamers randomly sampled from the pool of all possible 4,096 hexamers ($P \approx 0.034$, one-tailed empirical P-value). Thus, to a first approximation, stop codons appear depleted in the true ESE motifs.

ESEs are significantly depleted in stop codons after controlling for nucleotide content

The above result is *prima facie* evidence that ESE motifs are unusual in having a low SCD. However, it could also be owing to underlying nucleotide biases within the set of ESEs. If so, ESEs should also be depleted of codons of similar nucleotide content to the stop codons. To address whether the low SCD of ESEs reflects an avoidance specific to the stop codons, we have to control for both the nucleotide content of the stop codons and nucleotide content of the ESE motifs.

To control for the nucleotide content of the stop codons, we compiled codon sets that are compositionally-matched to the stop codon set (see Figure 1A, 1B, 1C, Materials and Methods). We start by considering the 2,879 GC-matched tri-codon sets (i.e. with GC content = 0.222, the same as the stop codon set). To test whether the stop codons specifically are under-employed in ESEs, we also have to control for ESE nucleotide content. We therefore generated 10,000 dinucleotide matched pseudo-ESE motif sets (N = 84 pseudo-motifs per iteration matching the number of INT3 motifs). For any given codon set, we can then calculate a fold-enrichment (FE) score (see Materials and

Methods) that gives the relative enrichment of a given codon set in the true ESEs while accounting for underlying ESE nucleotide content. FE > 0 implies enrichment, FE < 0 implies depletion and FE ≈ 0 reflects null.

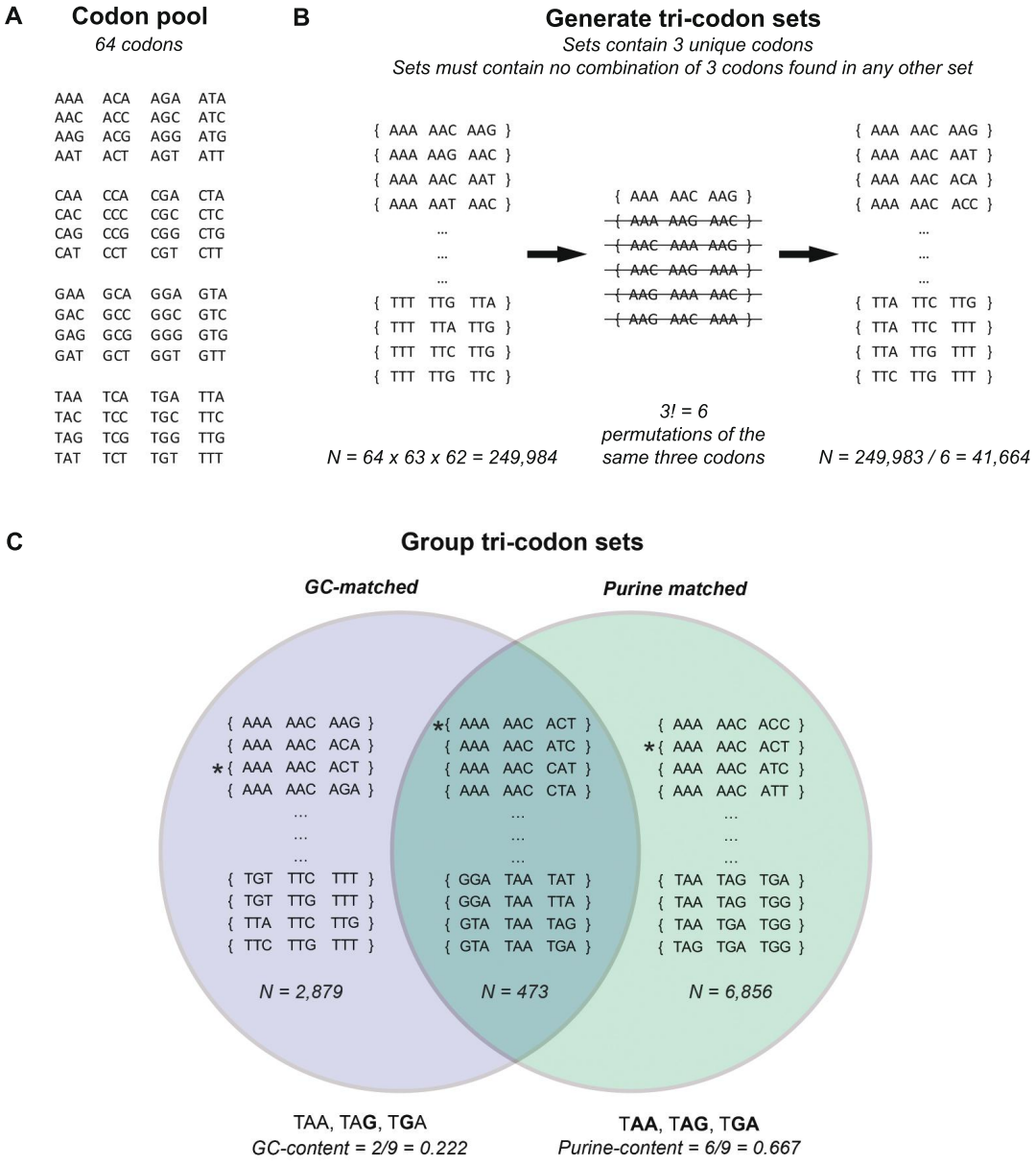


Figure 1: Overview of how the tri-codon sets were derived. (A) Every codon was considered (N = 64). (B) Every possible permutation of three codons was generated, ensuring each permutation contained three unique codons, leaving N = 64 × 63 × 62 = 249,984 sets. For each grouping of three unique codons, there exists 3! = 6 possible permutations of the three codons. Codon sets with the same three codons, just in a different order, were considered to be the same codon set, and so duplicates were removed leaving N = 249,984 / 6 = 41,664 codon sets. (C) The codon sets from (B) were then grouped. The first set contains codon sets with identical

net GC content to the stop codons (GC = 0.222, N = 2,879). A second contained codon sets with identical net purine content as the stop codons (purine = 0.667, N = 6,856). Finally, a set comprising the intersection of both GC- and purine-matched sets was generated (N = 473). The example codon set [AAA, AAC, ACT] has both equal GC and purine content to the stop codons and is highlighted by the *.

If stop codons are depleted in the true set of ESEs, because they are stop codons, their FE should be lower than the FE of the GC-matched control codon sets. Conversely, if stop codons are depleted in ESEs because of the nucleotide content of stop codons and ESEs, their FE should be no lower than the FE of the GC-matched control codon sets. We find 2,018/2,879 = 70.09% of the GC-matched codon sets have a higher FE than for the stop codon set ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5, Figure 2A), consistent with a depletion due to being stop codons.

A particular curiosity of the ESE motifs (and of INT3 ESEs more specifically) is that they are purine-rich (mean number of purine nucleotides in an INT3 motif = 4.702/6, minimum = 2/6, maximum = 6/6) (Xu et al. 1993; Dirksen et al. 1994; Tanaka et al. 1994; Gersappe and Pintel 1999; Fairbrother et al. 2002; Caceres and Hurst 2013). As stop codons are also purine-rich (6/9 nucleotides in the stop codons are purines), the INT3 motifs should be more conducive to including stop codons. Thus, distorted purine content within both ESEs and stop codons is unlikely to explain why stop codons are, in absolute terms, under-employed in ESE motifs. Nonetheless, we can ask whether after controlling for purine content the stop codons are specifically under-employed as our transfer selection model predicts.

To examine this, we identified the 6,856 tri-codon sets that exactly match the purine content of the stop codon set (Figure 1C). The majority of these purine-matched sets (5,497/6,856 = 80.18%) have a higher FE than for the stop codon set ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5). This implies the stop codon depletion in ESEs is specific to stop codons and not explained by purine content. Neither this result nor that for GC-matched sets above can be explained by allowing

stop codons to exist in the matched codon sets or by the inability for stop codons to overlap one another (Supplementary Text 1).

We can also control for both parameters simultaneously by considering tri-codon sets that have both GC- and purine content exactly matching the stop codon set (N = 413, Figure 1C) (e.g. the set [AAA, AAC, ACT]). We find that significantly more of these GC-purine-matched codon sets have greater FE than the stop codon set (317/413 = 67.02%, P = 5.484×10^{-14}, one-tailed exact binomial test, null probability of success = 0.5, Figure 2B). In sum, we conclude that the depletion of stop codons in ESEs is relatively specific to the stop codons themselves, rather than being owing to the peculiarities in nucleotide content of ESEs and stop codons.

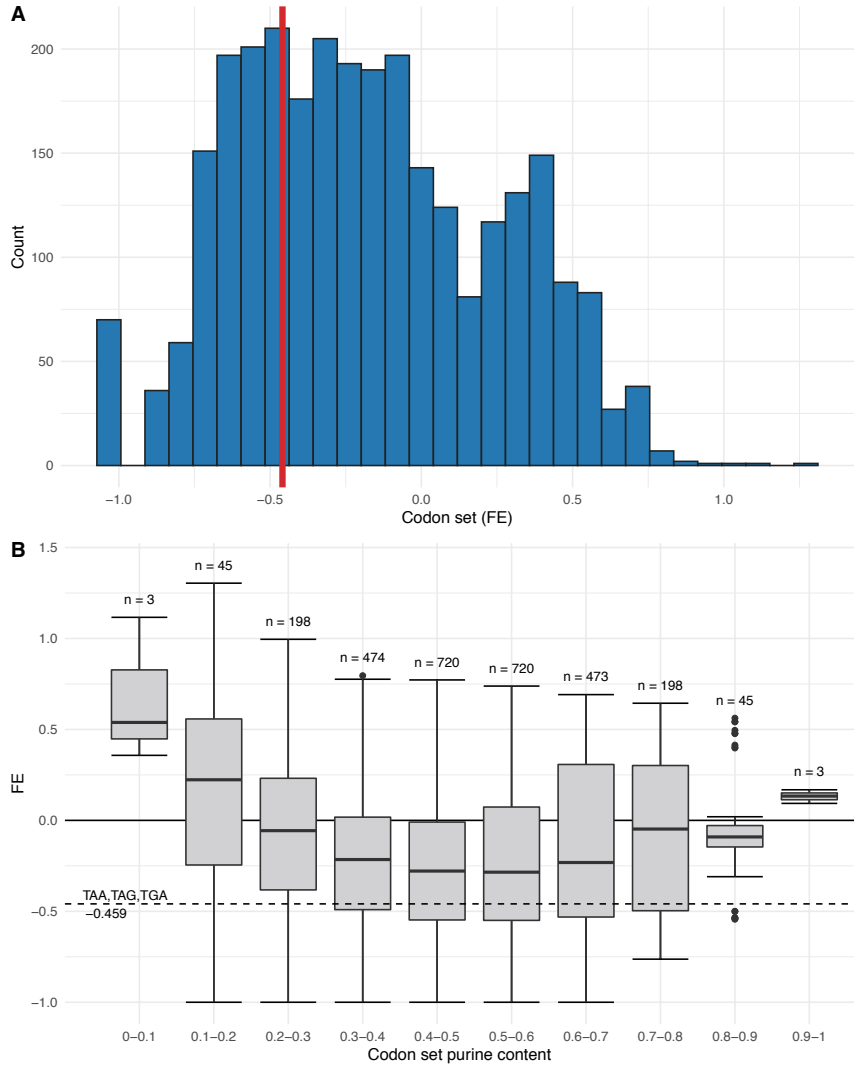


Figure 2: Comparisons of fold-enrichment (FE) scores of the stop codon set. (A) Histogram showing the FE scores of codon sets containing three unique codons with identical GC content to the stop codon set (GC = 0.222) in INT3 ESEs. The stop codon set highlighted by the vertical line. When controlling for the dinucleotide-content of ESEs, the FE of the stop codon set is highly depleted compared with GC-matched codon sets and falls towards the lower tail of the distribution of FE scores. (B) Boxplots of FE scores for tri-codon sets with GC content equal to that of the set of stop codons, grouped by purine content. Not only is the FE of the stop codon set (dotted horizontal line) reduced when compared with GC-matched codon sets, it is significantly reduced ($P = 5.484 \times 10^{-14}$, one-tailed exact binomial test) when compared with sets also containing identical purine content (purine content grouping 0.6-0.7, N = 473).

The stop codon depletion is a general property of ESE motifs defined within coding sequence

Another possibility that may explain the above depletion is a peculiarity of the motifs contained in INT3 set. To address this and ask whether the stop codon depletion applies to ESEs more generally, we calculated the FE score for the stop codon set in several ESE collections derived from analyses of coding exons. As expected, stop codons are significantly depleted in all ESE sets (Table 1, Figure 3A; note, the INT3 is not fully independent of the RESCUE-ESE, ESR and Ke400 sets). This result also confirms the INT3 set is representative of ESE sets more generally. Stop codons in the Ke400 set (Ke et al. 2011), unexpectedly enriched in exon cores and under positive selection (Caceres and Hurst 2013; Savisaar and Hurst 2018), are also significantly depleted ($P \approx 0.001$, one-tailed empirical P-value) consistent with depletions due to functioning within coding regions. These results also argue against the depletion in the INT3 set being a result of motif ascertainment biases resulting from the methods used to identify any particular set of ESEs (see section "Motif sets" in the Materials and Methods for an overview of how each set was derived).

To avoid covariance with CDS parameters (such as codon usage), the PESE set (Zhang and Chasin 2004) was derived from comparisons of constitutively spliced noncoding exons, unspliced pseudoexons and 5' untranslated regions (UTRs) of intronless genes. Motifs in this set are therefore not subject to protein-coding constraints and should provide an exception to the rule. For this set, the SCD (0.084) is higher (P = 0.001,

one-tailed one-sample t-test) and FE (-0.122) negative but higher (P = 0.008, one-tailed one-sample t-test) than for other ESE sets (Table 1). This result is in the direction we expect and consistent with our model. That the FE is not zero is likely a result of ESEs in this set also featuring in the ESE sets derived from CDS exons (Caceres and Hurst 2013), suggesting that some of these ESEs are likely functional in CDS and subject to protein-coding constraint.

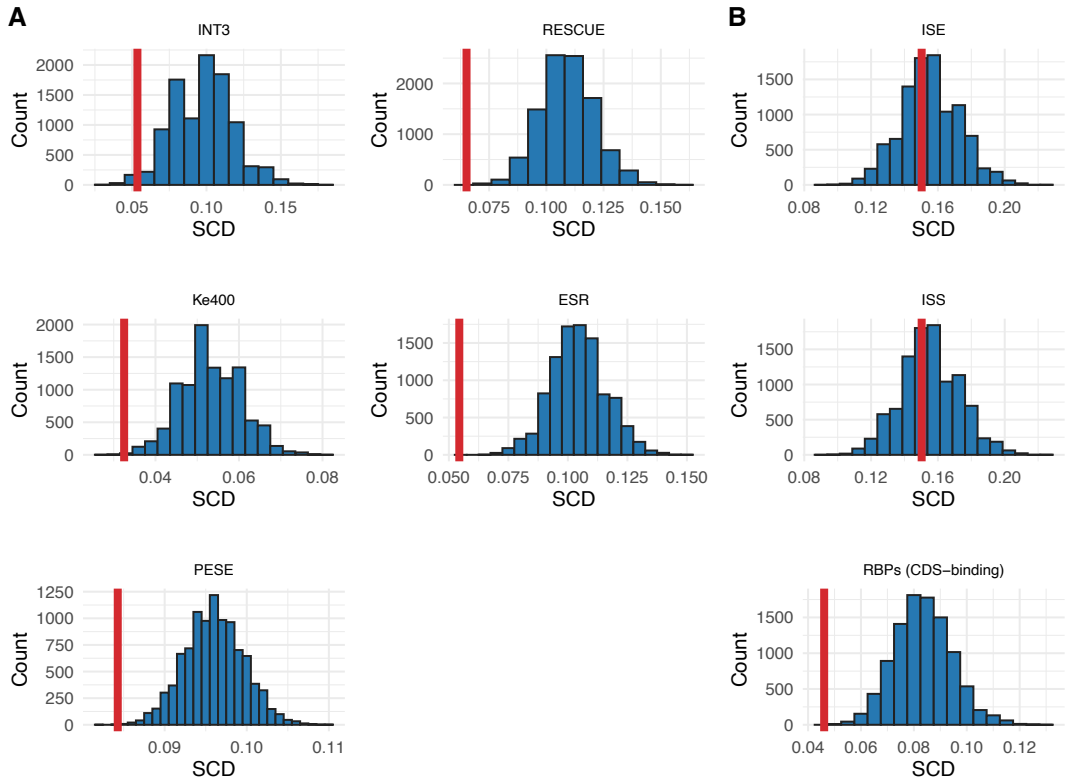


Figure 3: Histograms of the stop codon densities (SCDs) in 10,000 sets of dinucleotide-matched null pseudo-motif sets. The SCD in the real motifs of each set is shown by the vertical line. (A) Each exonic splice enhancer (ESE) motif set demonstrates a significant stop codon depletion. (B) SCDs in the motifs of intronic splice enhancers (ISEs), intronic splice silencer (ISSs) and CDS-binding RNA-binding proteins (RBPs). These depletions accord with their locations - intronic motifs not avoiding stop codons, CDS exonic motifs avoiding stop codons.

The stop codon depletion is a general property of motifs that function in coding sequence

These results could, however, also be explained if there is a general avoidance of stop codons in all splice-related or RNA binding protein (RBP) motifs whether they bind

CDS or not. By contrast, our hypothesis predicts that motifs that do not function in CDS should not have a significant depletion. To ask whether the constraint is specific to motifs that do function in CDS, we consider FE for CDS-binding RBPs more generally and motifs associated with intronic binding.

The general set of RBP motifs thought to be CDS-binding (Savisaar and Hurst 2017) (compiled from RBPDB (Cook et al. 2011), RBPmap (Paz et al. 2014), SFmap (Paz et al. 2010) and CISBP-RNA (Ray et al. 2013)) demonstrates a similar significant stop codon depletion ($P \approx 9.999 \times 10^{-5}$, one-tailed empirical P-value). For the non-CDS motifs, both intronic splice enhancers (ISEs) ($P \approx 0.417$, one-tailed empirical P-value) and intronic splice silencers (ISSs) ($P \approx 0.307$, one-tailed empirical P-value) have no avoidance of stop codons (Figure 3B). These results therefore argue that the depletion of stop codons in motifs functioning in exonic sequence is not ESE specific, splice-specific nor a result of being an RBP-binding motif, but rather a peculiarity associated with being located in exonic CDS. Further, we find no evidence that stop codon containing ESE motifs are avoided in protein-coding sequences and cannot be discounted as being suboptimal (Supplementary Texts 2-5).

Multi-exon lincRNA sequences are significantly depleted in stop codons

Does the lack of stop codons within ESEs transfer to and constrain lincRNA sequences as we propose? To test this, we employed the set of lincRNA sequences identified by Cabili et al. (2011). In this set, potential protein-coding transcripts were removed (see Materials and Methods for details) and so this set should contain a minimised number of lincRNAs with potential protein-coding ORFs that would contaminate our results. After our filtering, we employ 1,919 multi-exon lincRNAs (53 from multi-gene families, 1,866 from singleton families, see Materials and Methods).

To eliminate the possible effects of nucleotide bias of the lincRNA sequences, we ask whether mature lincRNA transcripts are depleted for stop codons given the underlying nucleotide content of each sequence. We shuffled the nucleotides within every lincRNA and calculated the SCD for that iteration of 1,919 shuffled "pseudo-lincRNAs". After repeating this for 1,000 iterations to generate a null distribution, we

find no simulated iteration with overall SCD as low or lower than the SCD in the real 1,919 lincRNA sequences (true SCD = 0.130, FE = -0.162, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value, Table 2). As the absence of a potential ORF was used to classify RNA species as lincRNA (rather than mRNA), this is a potentially conservative estimate.

To confirm that this low SCD in lincRNAs is specific to the stop codons and not the GC content of the stop codons, we consider densities of GC-matched tri-codon sets within the real and simulated sets of lincRNA sequences (converting codon set SCD values to FE values). The FE of the stop codon set is significantly lower than the FE of the majority of GC matched control tri-codon sets both when stop codons are permitted in the GC-matched codon sets (codon sets with FE > stop codon set FE = 2,315/2,879 = 80.41%, $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5, Table 2) and when they are excluded (codon sets with FE > stop codon set FE = 1,771/2,121 = 83.50%, $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5). We conclude that the low SCD in lincRNA cannot be explained by the low GC content of the stop codons.

We find this stop codon depletion is also robust to pairwise analysis (i.e. each gene versus randomisations of that same gene) (Table 3), with 79.62% (1,528/1,919) of sequences having FE less than zero ($P \approx 0$, one-tailed exact binomial test, null probability of success = 0.5). Of these, 493/1,919 have a significant depletion ($P = 1.23 \times 10^{-200}$, one-tailed exact binomial test, null probability of success = 0.05). Results are not affected by the choice of sequences from paralogous families (Supplementary Text 6). Results are also quantitively similar using a second independent set of sequences (GENCODE RNA Capture Long Seq (CLS) annotated sequences, Lagarde et al. (2017)) (Supplementary Text 7, Table 3). This trend is unlikely to result from hidden ORF contamination as the sequences 5' of the most 5' ATG, and therefore lacking protein-coding potential, also have reduced SCD (Supplementary Text 8).

Is this depletion specific to exonic lincRNA sequence as predicted by our transfer selection model? We compared the SCD in exons and introns of lincRNA sequences

in a pairwise manner. This test is potentially conservative as some "intronic" sequence may well be hidden exon derived from unannotated alternative splice forms. However, we find that in 68.79% (1,320/1,919) of genes the SCD of the exons is less than the SCD of the introns ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5, Table 2). Thus, the depletion appears to be more specific to exonic sequences, consistent with our model.

Exons of multi-exon lincRNAs demonstrate significantly reduced stop codon densities when compared with single-exon lincRNA exons

The above results are all consistent with our model of transfer selection. If we are to attribute this depletion to the presence of ESEs, the magnitude of the depletion in exons of single-exon lincRNAs should not be as great as that for multi-exon lincRNAs, assuming single-exon genes do not need to contain ESEs to bind splicing factors. As the filtered Cabili et al. (2011) dataset contained only 12 single-exon sequences in total, we performed this analysis on the GENCODE lincRNA sequences (Lagarde et al. 2017).

As expected, the SCDs of single-exon sequence exons (N = 877 exons) are significantly higher than the SCDs of the exons of multi-exon sequences (N = 1,417 exons, N = 456 sequences) (median single-exon SCD = 0.139, median multi-exon SCD = 0.122, $P = 8.878 \times 10^{-14}$, Wilcoxon rank sum test). However, given the compositional difference between single-exon and multi-exon transcripts (median single-exon GC = 0.456, median multi-exon GC = 0.477, $P = 4.938 \times 10^{-12}$, Wilcoxon rank sum test), it is important to control for the compositional differences of the exons for each class. We therefore calculated FE scores of single-exon sequence exons and multi-exon sequence exons by simulating each exon sequence individually. Consistent with the above result and our expectations, FE scores for single-exon lincRNA are negative but significantly higher than for multi-exon lincRNA (median single-exon FE = -0.148, median multi-exon FE = -0.167, P = 0.027, one-tailed Wilcoxon rank sum test). That the single-exon genes also have a negative FE is not unexpected, as they are likely to be frequently bound by RNA-binding proteins that also bind in CDS and contain ESEs that have splice-independent roles (Savisaar and Hurst 2016). In accord

with the reduced SCD in single-exon lincRNAs, we also find a lower ESE density (median single-exon sequence exon ESE density = 0.127, median multi-exon sequence ESE density = 0.155, $P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon rank sum test). Confirming that the FE metric controls for GC differences, the slope on the line of FE predicted by GC is not significantly different from zero (P = 0.334).

All else being equal, 5' UTRs of protein-coding genes should have a lower SCD in multi-exon transcripts than in single-exon ones, not least because the first intron is often close to the ATG and hence to the UTR. We find that there is a lower SCD in 5' UTRs of single-exon protein-coding genes than in multi-exon protein-coding genes, although this is not robust to nucleotide control (see Supplementary Text 9). For reasons unknown, the 5' UTRs of single-exon protein-coding genes have higher ESE densities than for those of multi-exon genes (see Supplementary Text 9), which both runs counter to *a priori* expectations and conflates the above test.

Stop codon density is lowest in regions where ESE density is highest

While the above is consistent with reduced SCD in lincRNAs (compared with a nucleotide controlled null) as we predict, can we attribute this to ESEs and hence argue that the depletion is a result of CDS-imposed constraints on ESEs? If so, we expect SCD to be lowest in the regions in which ESEs typically reside. Despite selection on ESEs in protein-coding genes being most pronounced at exon ends (Berget 1995; Fairbrother et al. 2002; Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley et al. 2006; Parmley et al. 2007; Caceres and Hurst 2013), in lincRNA the proportion of sequence within 70 bp of an exon junction is not significantly correlated with evolutionary rate (Schuler et al. 2014), probably because in lincRNA ESEs function at the 5' end more profoundly than at the 3' end (Krchnakova et al. 2019). The depletion of stop codons and enrichment of ESEs should therefore be strongest at the 5' end of lincRNA exons.

For each lincRNA gene, we divided each exon longer than 207 nucleotides into the 5' flank (nucleotides 3-69), the equivalent 3' flank and exon core (67 nucleotides centred about the exon midpoint), such that each region from each exon contained 67

nucleotides. We then calculated both ESE density and SCD for each region within each exon. As predicted, ESEs are enriched in 5' flanking regions, while SCDs in this region are closer to zero than either the core or 3' regions (Figure 4). In accord with the notion that 3' ends are not such key SR protein interaction domains, ESE densities in 3' flanks are lower than in 5' flanks and have higher SCDs. Similarly, exon cores have lower ESE densities and higher SCDs than 5' flanks.

SCDs in the various regions differ significantly from null expectation ($\chi^2 = 160.822$, $P = 1.20 \times 10^{-35}$, chi-squared test), with the 5' region observed/expected frequency (O/E) lowest of all regions (5' flank O/E = 0.917, core O/E = 0.960, 3' flank O/E = 1.1253). Further, when simulating each region separately, the stop codon FE for the 5' flank (-0.185) is more negative than both the core (-0.142) and 3' flank (-0.121) (all FE scores with empirical P-values < 0.05). This broad-scale data is therefore consistent the lowest SCD being in the region where ESEs are most frequent.

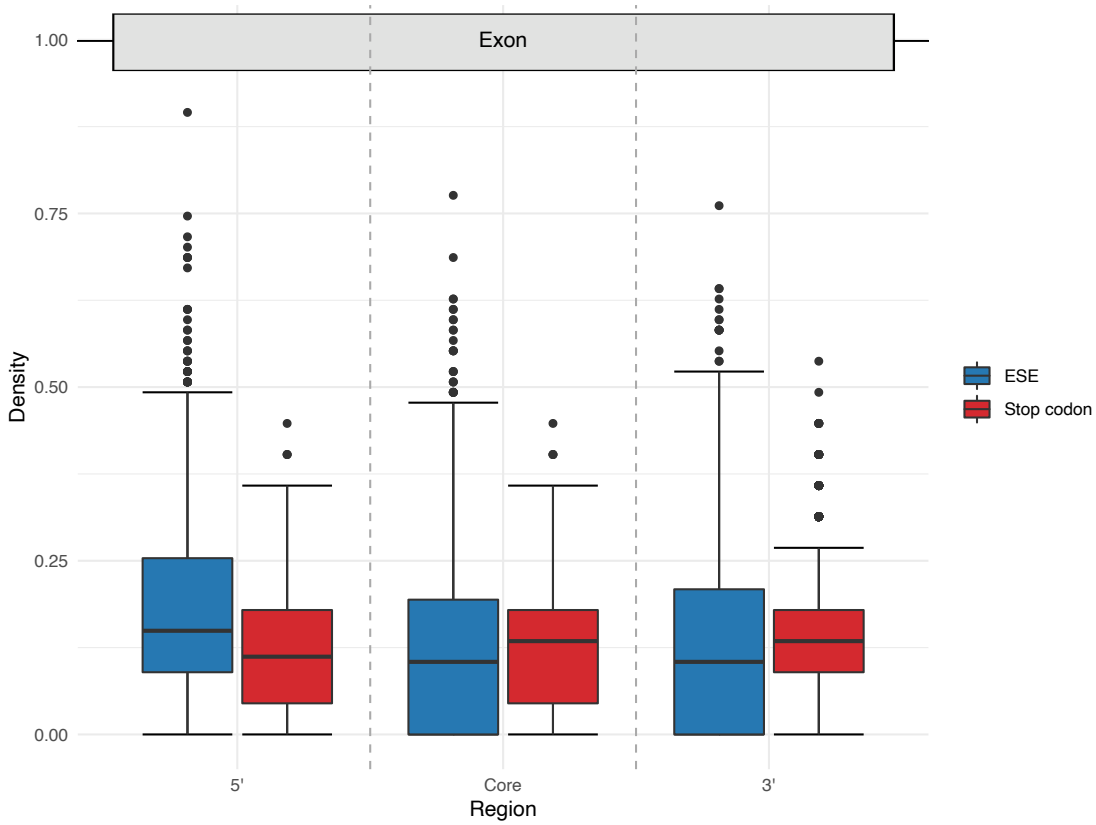


Figure 4: Densities of ESEs and stop codons in separate regions of lincRNA exon sequences longer than 207 nucleotides. 5' flanks contain nucleotides 3-69 and 3' flanks the corresponding nucleotides at the other exon terminus. Core regions are the 67 nucleotides centred about the

exon midpoint. In the 5' flank region with higher ESE density, the stop codon density is reduced. In both the core and 3' flank where ESE density is much reduced, stop codon density is increased. These trends are consistent with the presence of ESEs reducing stop codon density.

Reduced stop codon densities in lincRNA sequences are attributable to the presence of predicted ESE motifs

Can we attribute the depleted SCD in lincRNAs to ESEs directly? We compiled a consensus set of motifs from the non-redundant union of all ESE motif sets (2582 motifs, 468 hexamers and 2,060 octamers), excluding the Ke400 set as motifs in this set demonstrate positive selection and enrichment in exon cores over flanks (Caceres and Hurst 2013; Savisaar and Hurst 2018) despite splice mutations being enriched at exon ends (Woolfe et al. 2010). By excluding any sequence that matches a motif within the consensus set after predicting hits to all motifs to recover overlapping motifs, the influence of ESEs on SCD is eliminated. If ESEs are driving the depletion, the remaining (unmatched) sequence should have SCD similar to that predicted by its underlying nucleotide content.

We predicted hits to the consensus ESE motifs in each lincRNA and retained only the unmatched sequence. After randomly shuffling the remaining nucleotides we observe that the real non-ESE sequence has a higher SCD than null (FE = 0.159, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value, Supplementary Table 1) indicating the overall depletion of stop codons is owing to ESE motifs. This result further argues against the net depletion of stop codons in lincRNAs being an artefact of hidden protein-coding ORFs, as such a model predicts stop codon depletion both within and outside of ESEs. Why the remaining sequence is enriched in stop codons is unknown, but could be the result of selection on the remaining non-ESE sequence to "appear" less like ESE to SR proteins to prevent inappropriate binding (e.g. see Savisaar and Hurst (2017)) (Supplementary Text 10). We also find that the depletion of stop codons is not a result of lincRNA sequences avoiding the use of those ESE motifs that contain stop codons (Supplementary Texts 11-12).

Skewed stop codon usage in ESEs reflects skewed stop codon usage in lincRNA

Above we have treated the stop codons as a single set. However, in all ESE sets (including INT3 and the consensus set), TGA is more abundant than TAA or TAG (Table 4). This provides us with a further test of our transfer selection model. If the stop codon avoidance in lincRNAs is owing to ESEs avoiding stop codons, the avoidance of TAA and TAG in ESEs should be reflected in the usage of each stop codon within lincRNAs.

Using the Cabili et al. (2011) set of lincRNAs, we find the stop codons in lincRNA are not used at similar frequencies, with TGA the most abundant (TAA density = 0.043, TAG density = 0.027, TGA density = 0.060). When compared with null randomised shuffled lincRNA sequences, both TAA and TAG are significantly depleted (TAA: FE = -0.292, $P \approx 0.001$; TAG: FE = -0.439, $P \approx 0.001$, one-tailed empirical P-values) while TGA is significantly enriched (FE = 0.247, $P \approx 0.001$, one-tailed empirical P-value). Thus, the stop codons most avoided in ESEs are those most avoided in lincRNA.

To attribute this directly to the presence of ESE motifs, we also ask whether the significant depletion of TAA and TAG occurs when ESEs are not present. If the TAA and TAG depletions remain in non-ESE sequence, this would argue for depletion due to reasons other than ESEs. As before we considered the lincRNA sequence that remains after the removal of sequence matching motifs in the consensus ESE set. After removal, both TAA and TAG are now found significantly more frequently than expected (FE = 0.590, $P \approx 0.001$ and FE = 0.117, $P \approx 0.001$ respectively, one-tailed empirical P-values) while TGA is depleted (FE = -0.164, $P \approx 0.001$, one-tailed empirical P-value). We conclude that the depletion of both TAA and TAG in ESEs appears to force lincRNA sequences to also under-employ these two stop codons, consistent with our model.

The majority of lincRNAs contain permissible pseudo-ORFs longer than expected by chance

Taken together the above results are consistent with our model, transfer selection forcing a low density of stop codons in lincRNAs. Might this impact gene annotation? To distinguish noncoding RNA from protein-coding sequence, computational annotation approaches often consider the lengths of potential ORFs (Frith et al. 2006a; Clamp et al. 2007; Dinger et al. 2008). To reduce the likelihood of falsely categorising noncoding RNAs, putative noncoding RNAs are considered as those lacking ORFs longer than 300 bp as the majority (>95%) of annotated eukaryotic proteins are thought to be longer than 100 amino acids (Frith et al. 2006a; Clamp et al. 2007; Dinger et al. 2008).

Our results above, however, have implications for any potential lincRNA pseudo-"ORF" (pORF) lengths. If the net depletion of stop codons constrains lincRNA sequences as we suggest, lengths of potentially tolerated lincRNA pORFs should be longer than expected. Indeed, Niazi and Valadkhan (2012) show a non-negligible proportion of "functional" long noncoding RNAs (lncRNAs), although not intergenic, have an "ORF" length greater than 300 nucleotides (e.g., the *Xist* gene encodes a functional ≈15 kb transcript in mouse (Prasanth and Spector 2007) with a potential 592 nucleotide ORF (Brockdorff et al. 1992)).

To address the likely extent to which true noncoding lincRNAs present long pORFs by chance, we generated 1,000 sets of simulated lincRNA sequences by shuffling the full multi-exon lincRNA transcripts. For each real and simulant sequence, we determined the length of the longest pORF, assuming pORFs start ATG and terminate with a stop codon in the same reading frame. Seven real sequences had no complete pORF in any frame and were excluded. We calculated the Z score for each sequence, with a positive Z indicating an increased maximum pORF length compared with null sequences.

We find robust evidence that pORFs are commonly longer than expected, with 62.33% (1,227/1,912) having $Z > 0$ ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5, median longest pORF length: real = 159, simulants = 129; maximum longest pORF length: real = 2,202, simulants = 450). Further, more sequences than expected by chance also have a significantly positive Z (13.13% =

251/1,912, $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.05). Only one sequence had a significantly shorter pORF than expected ($P \approx 1$, one-tailed exact binomial test, null probability of success = 0.05). These differences in pORF length are greatest when AT-content is highest i.e. when stop codons are more likely to occur by chance ($\rho = -0.083$, $P = 3.00 \times 10^{-4}$, Spearman's rank correlation between sequence GC content and sequence pORF length, Figure 5A). Thus, it would appear that not only are permissible pORFs longer than expected, but there exists greater deviation from expected pORF lengths (measured in standard deviation units) when stop codons should be more frequent.

Almost 10% of sequences would be misannotated if categorised on open reading frame length alone

Do longer than expected pORFs have implications for lincRNA sequence identification? While the 300 bp ORF lower limit is applied to reduce false-positive rates (Frith et al. 2006a; Clamp et al. 2007; Dinger et al. 2008), sequences are also annotated based upon their level of sequence conservation as noncoding RNAs demonstrate conservation but below that of protein-coding genes. However, a conservation approach is *a priori* poor at identifying young ORFs. Given pORF lengths are increased owing to stop codon avoidance, we ask what a safe length threshold might be.

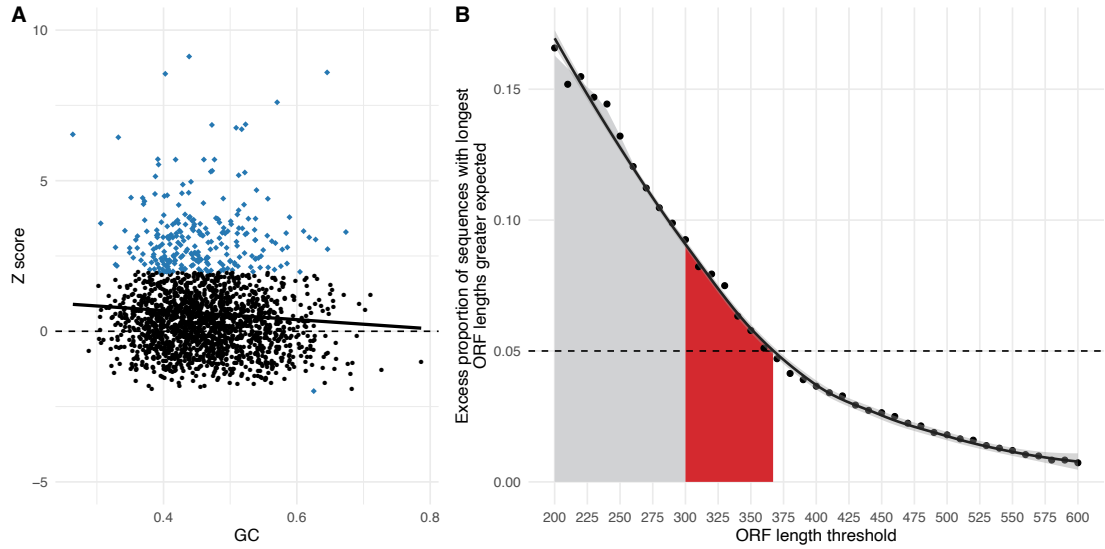


Figure 5: Analyses of potential ORFs in lincRNA sequences. (A) Z scores for the longest permissible ORF lengths when compared with randomly shuffled simulated sequences are negatively correlated with GC content (ρ = -0.082, P = 2.976 × 10^{-4}, Spearman's rank correlation). Moreover, data points demonstrating significant deviations (blue diamond) from expected ORF are all positive, except one. One sequence with Z = 32.519 has been removed from the figure for visual purposes. (B) The excess proportion of sequences with maximum ORF lengths longer than expected decreases with increasing ORF length thresholds. A threshold of 368 bp is required such that there is less than a 5% excess (dotted line). Results for thresholds within the region highlighted in red demonstrate areas of ORF lengths that could be ambiguous if used as the sole determinant of coding capability, extending beyond thresholds that are up to and including the commonly used 300 bp threshold (grey).

We find 11.57% (222/1,919) of the total sequences meet or exceed the 300 bp threshold in our data (taking the median length for sequences grouped into gene families). However, is this number biologically relevant? For example, random sequences of equal length to lincRNA sequences may also contain pORFs longer than 300 bp. To test this, we concatenated all exons from all sequences, randomly shuffled the concatenation and extracted randomised sequences with lengths matching the real mature transcript sequences, thereby generating 1,000 randomised null sets of sequences with equal overall transcript length and nucleotide content. For each iteration of randomised sequences, we then calculated the number of sequences with a pORF exceeding 300 bp. We find the number of real lincRNA sequences exceeding the threshold (222) is almost significantly greater for the null sets (mean number

exceeding = 44.434, standard deviation = 171.770, P ≈ 0.051, one-tailed empirical P-value). Further, no randomised set had a pORF longer than the longest pORF seen in the true lincRNAs (P ≈ 9.99×10^{-4}, one-tailed empirical P-value, maximum true = 2202, maximum simulant = 594). Using the mean number of simulant sequences exceeding the 300 bp threshold as the expected number to exceed the threshold, this suggests $\frac{222-44.434}{1,919} \approx 9.25\%$ of real sequences could be misclassified based upon ORF length alone beyond that expected by chance.

The above results suggest that owing to transfer selection, achieving a 5% false-positive rate requires a threshold longer than 300 bp. How long might this cut-off be? Using 10 nucleotide threshold intervals between 200 to 600, we calculated percentage excesses over null as above and fitted a local regression model. This model predicts a threshold of 368 bp is required so that only 5% of sequences exceed the threshold (Figure 5B). However, while a longer threshold reduces the false-positive rate, we note that there likely exists an abundance of functional protein-coding genes that encode short proteins (Oyama et al. 2004; Frith et al. 2006b; Andrews and Rothnagel 2014; Slavoff et al. 2014). Thus, a longer threshold will also increase the false-negative rate. Given this, bioinformatic approaches should be coupled with experimental validation (Kashi et al. 2016) whenever possible.

Discussion

Whilst much is known about the selective pressures acting on coding sequences, those in noncoding sequences are less well understood. Human lincRNA are under weaker purifying selection than protein-coding genes (Marques and Ponting 2009; Cabili et al. 2011; Haerty and Ponting 2013) and contain fewer conserved regions (Pang et al. 2006). However, ESE motifs that are under strong purifying selection in protein-coding genes (Parmley and Hurst 2007; Parmley et al. 2007; Warnecke et al. 2008; Smithers et al. 2015; Savisaar and Hurst 2018) are also under purifying selection in lincRNA sequences, suggesting splicing of multi-exonic lincRNA transcripts is also important for function (Schuler et al. 2014; Haerty and Ponting 2015). With both coding and noncoding sequence thought to undergo the same splicing process by the

same splice machinery (Will and Luhrmann 2011; De Conti et al. 2013; Krchnakova et al. 2019), we hypothesised that the same constraints should apply to both types of sequence.

Here we have provided evidence consistent with a depletion of stop codons found in ESE motifs that, allowing for nucleotide content, is specific to the stop codons. That both ESEs and stop codons are purine-rich makes the depletion of stop codons particularly noteworthy (indeed the high purine content may be a defining feature of ESEs to discriminate exon ends from other sequences, see Supplementary Text 13). The evidence that we have presented suggests that this stop codon depletion of ESEs that function in CDS transfers to lincRNA sequences. As a consequence, and contrary to null expectations, lincRNAs too are significantly depleted in stop codons. Multiple lines of evidence, including a significant increase in stop codons found after removing ESEs from lincRNA sequences, suggests that ESEs are the origin of this depletion (or at least a major contributor). Thus, constraints imposed on motif composition in protein-coding sequences can transfer to noncoding sequence.

One could argue that the most obvious alternative explanation for the depletion of stop codons in lincRNA is the contamination of the dataset with true, but unrecognised, protein-coding sequences. However, several pieces of evidence argue against this. First, if a lack of ORF is used to classify RNA species as lincRNA, rather than mRNA, we expect an enrichment of stop codons in lincRNA. Our tests comparing SCDs are thus conservative. Second, we observe similar depletions from two independent datasets, in which both take measures to exclude sequences demonstrating evidence of protein-coding potential. That there is also a depletion of stop codons in exon regions with the highest ESE density, yet no depletion in exonic sequence after the removal of ESEs (and indeed an enrichment), suggests lincRNA sequences are not depauperate in stop codons in their entirety but biased by the presence of ESEs. Furthermore, the sequence upstream of the first ATG is stop codon depleted, despite no influence of any ORF on densities (Supplementary Text 8).

We also question whether ORF contamination could explain the magnitude of the observed reduction in SCD. Any contamination by real hidden protein-coding ORFs would also have to be substantial, particularly given the pairwise analysis of SCDs

against randomisations of each gene indicates that 79.62% of sequences have a stop codon depletion. Given the filters on the original sequences, it seems unlikely that true ORFs, common enough to provide such contamination, would have gone unrecognised.

In principle, lincRNA sequences may be depleted in stop codons if they overlap unannotated protein-coding genes on the same strand. Unless there is a rich source of unannotated overlapping ORFs this is not parsimonious to explain the commonality of stop codon depletion. Moreover, in neither the hidden ORF nor the unannotated overlapping ORF model is the specificity of stop codon depletion to ESEs and exon 5' ends (where ESEs are most abundant) explained. That the stop codons depleted in lincRNA (TAA and TAG) accord with the stop codons depleted in ESEs also supports transfer selection above ORF contamination. In sum, transfer selection therefore provides the most parsimonious explanation of our observations.

We have also assumed that as SR proteins must bind coding exons there is a constraint transferred to noncoding exons. Might there be a transfer in the opposite direction? If we consider CDS alone, a theoretical set of motifs with most utility (most likely to hit exons exclusively) would be one that avoids stop codons entirely. Over evolutionary time, selection might therefore be expected to eliminate ESEs with stop codons as potential binding motifs. Yet stop codon containing motifs persist. However, RBPs and binding motifs are thought to coevolve which has been exploited to predict RBP binding domains (Yang et al. 2018). If these stop codon containing motifs can be more easily employed in noncoding sequence whilst also providing the adequate binding capability, then they might still provide enough splicing functionality to be selected for. Given only a minority of transcribed sequence is protein-coding (The Encode Project Consortium 2012), the relative frequency of noncoding RNA splicing may render such motifs selectable. In turn, they may also then be useful motifs in protein-coding sequence where splice specificity is less important, but the nucleotide composition of the sequence allows their usage. Thus, if motifs that include stop codons are of utility and can be frequently used within lincRNA, it may be that RBPs and stop codon containing motifs coevolve such that they persist as functioning motifs. A suggestion of this is found in our result that the ESEs that feature stop codons are if anything overused on a per motif basis in CDS (Supplementary Text 2).

Stop codon depletions and the origin of *de novo* genes

The stop codon depletion in lincRNA and ESEs more generally might modulate the evolution of new genes. The origin of new genes receives much attention (overviews in Long et al. 2003; Kaessmann 2010; Tautz and Domazet-Loso 2011; McLysaght and Hurst 2016). While duplication and rearrangement (Ohno 1970; Jacob 1977; Zhang 2003; Ciccarelli et al. 2005; Innan and Kondrashov 2010; Magadum et al. 2013; Van Oss and Carvunis 2019) are known to be important processes that adapt and reuse functional sequence, the creation of *de novo* protein-coding genes from previously non-functional or noncoding sequences is increasingly being recognised as a source of novelty (Tautz and Domazet-Loso 2011; McLysaght and Guerzoni 2015; McLysaght and Hurst 2016).

Two important steps are required to give rise to and allow fixation of functional proteins from noncoding sequence: acquisition of uninterrupted ORFs and regulatory transcriptional signals. The order of these events is not clear nor necessarily uniform, with two models proposed each arguing for the respective events occurring first (McLysaght and Guerzoni 2015; Schlotterer 2015). In the "RNA-first" scenario, the abundance of long noncoding RNAs that are transcribed and, possibly accidentally, associated with ribosomes (Wilson and Masel 2011; Ruiz-Orera et al. 2015) makes it possible that many unintended peptides are actively translated, thereby becoming proto-genes. In an "ORF-first" scenario, if an ORF is already present within the sequence mutations in *cis* regions could induce expression of the ORF (Kaessmann 2010; Zhao et al. 2014).

LincRNAs containing longer than expected ORFs owing to stop codon depletion are relevant to the RNA-first model. What is unknown is how the length of the pORF of a protogene relates to the probability of evolving from proto to functional protein-coding gene. If longer sequences are more likely to find immediate utility, rather than be toxic (Boyer et al. 2004; Levine et al. 2006), then this should exaggerate any putative tendency for *de novo* genes to originate in GC-rich sequence. Although ORF lengths in AT-rich regions have a greater deviation from expected (Figure 5A), the raw ORF lengths are longer in GC rich domains (correlations between GC raw ORF lengths are

significantly positive, $\rho = 0.219$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation). Further, GC-rich regions are more transcriptionally active (Lercher et al. 2003) with transcription factor binding sites being GC-rich (Wang et al. 2012a), and therefore more likely to give rise to lincRNA expression.

Stop codon avoidance is seen for other RNA-binding protein motifs

While above we have considered ESEs and show that they contain few stop codons, in principle, these are only one exemplar of CDS exonic motifs subject to stop codon depletion and hence subject to transfer selection. An expectation of stop codon depletion should then not be limited to ESEs but should also apply to other RBP binding motifs that function within coding regions. We indeed find a broader set of such motifs (compiled by Savisaar and Hurst (2017)) has a significant depletion of stop codons ($P \approx 9.999 \times 10^{-5}$, one-tailed empirical P-value, Table 1, Figure 3). We caution that only conservative conclusions should be drawn from this result as the quality of motifs used in the set is thought to vary (Savisaar and Hurst 2017). Nonetheless, we suggest that any peculiarities of sequence content necessitated by binding within CDS could have multiple transfer modes. It remains to be seen to what extent the compositional properties of lincRNAs are a consequence of carryover of binding preferences of RBPs shared with CDSs.

It is also the case that transfer selection should not be considered restricted to RBPs but may apply in other contexts and not limited to stop codons. Here we consider the comparison between coding and noncoding sequence, yet in theory similar logic could be applied to anything that interacts with two different sequence types. For example, there may be proteins that interact within different regions at the DNA level and transfer constraints between them.

Materials and Methods

General

Analyses were conducted using custom Python 3.6.4 scripts (available at https://github.com/la466/lincrna_stops_repo) using standard, readily available Python libraries. R version 3.5.1 (R Core Team 2018) was used for statistical testing and plotting of figures. BEDTools version 2.27.1 (Quinlan and Hall 2010) was used for operations performed on sequence coordinate data. For motif simulations, 10,000 iterations were run. For all other simulations, 1,000 iterations were run unless specified.

Retrieval and filtering of lincRNA sequences

LincRNA sequence coordinates were downloaded from the Supplementary Data Set 2 "TraitTable" sheet of Cabili et al. (2011). Sequences identified by Cabili et al. (2011) were done so via four key steps: 1) transcriptome reconstruction from RNA-seq data using two transcript assemblers (Cufflinks and Scripture); 2) compilation of all noncoding and unclassified transcripts previously annotated; 3) determination of unique isoforms from each transcript locus by integrating RNA-seq reconstructions with all annotation resources (Cuffcompare) and 4) processing of transcripts to identify those reliably expressed, large, multi-exonic, noncoding, and intergenic. Of these, the lowly expressed transcripts were removed using a learned read coverage threshold. Noncoding transcripts were filtered from novel potential protein-coding transcripts by removing those with evolutionary constraint to preserve amino acid content in any of the three reading frames (those with a positive phylogenetic codon substitution frequency (PhyloCSF) metric (Lin et al. 2011)) and by excluding transcripts matching a protein-coding domain present in the Pfam database (Finn et al. 2010).

From the Supplementary Data, only entries with the "ConservativeSet" flag set to 1 were retained, to leave 4,662 data points. These sequences are those with no evidence of protein-coding potential and that can be reconstructed in at least two different tissues or reconstructed by two assemblers in the same tissue. As such, transcripts with insufficient coverage should also have been removed. This sequence set should therefore contain a minimised number of potential protein-coding transcripts. Sequences containing non-canonical nucleotides and those containing only one exon were removed, leaving 4,646 multi-exon sequence data points.

To limit the effects of retaining genes with similar composition from our results, genes were clustered into paralogous families. The sequences were BLASTed all against all (Nucleotide-nucleotide BLAST 2.4.0+ (Camacho et al. 2009)). Starting with a randomly selected sequence, all sequences that had a significant hit were grouped as part of the same family and considered a single data point for the analyses. After grouping into paralogous families, 1,919 data points remained. For analyses, either the median value for sequences that are members of the same family was taken, or one member selected at random to represent the family. Where one member was selected at random, the analysis was repeated multiple times to avoid biases resulting from the random family member chosen.

Intergenic GENCODE lncRNA sequences reannotated by RNA Capture Long Seq (CLS) from heart, testes, liver, brain, human K562 and human HeLa cells were also used (Lagarde et al. 2017). Sequences IDs corresponding strictly and exclusively to lncRNA were obtained from Supplementary Dataset 1 of Lagarde et al. (2017) (although annotated as lncRNA, these sequences are intergenic and therefore appropriate). A processed bed file containing only entries for full-length transcripts whose 5' end is supported by FANTOM5 CAGE transcription start site (TSS) data and 3' is polyadenylated (cage+polyASupported) was downloaded from the GEO database accession GSE93848 (last accessed May 24, 2019). Only entries corresponding to the exclusive lncRNA IDs were retained. From these, the full-length multi-exonic transcripts containing only canonical nucleotides were built, retaining only those longer than 200 nucleotides to leave 11,083 transcript sequences. These were then subject to clustering into paralogous families as before, leaving 456 multi-exon data points for analyses. No sequences were identical to sequences from Cabili et al. (2011). The exons of single-exon lincRNAs were also extracted (N = 2,972) and clustered into paralogous families, leaving 877 single-exon data points.

Retrieval and filtering of protein-coding sequences

Protein-coding sequences were retrieved using similar protocols to Savisaar and Hurst (2016). To extract genome features, both the genome sequence and genome features were downloaded from the Ensembl database (Zerbino et al. (2018), Release 94, ftp://ftp.ensembl.org/pub/release-94/, last accessed October 25, 2018). The genome features were queried and only those labelled as "CDS" and "protein-coding" were

retained. From these features, the full CDS was constructed leaving 98,382 CDSs in the dataset. This dataset was filtered to remove CDSs that contained noncanonical bases, were not of a length divisible by three, did not start with ATG, did not end with a stop codon or contained in-frame stop codons. If more than one transcript per gene was present, the longest was retained; if two with the same length per gene were present, the first to be queried was retained.

The genome sequence and features for the *Macaca mulatta* genome were also obtained from the Ensembl database (Zerbino et al. (2018), Release 94, ftp://ftp.ensembl.org/pub/release-94/, last accessed November 05, 2018). Orthologs for all human genes remaining after the filtering steps described above were obtained via an Ensembl Biomart query using the Pybiomart Python package (https://github.com/jrderuiter/pybiomart). The orthologous CDSs of *M. mulatta* that corresponded to the remaining filtered human genes were extracted in the same process as for human CDSs and filtered according to the previous criteria. Both the human and macaque CDSs were translated to protein sequences and aligned using MUSCLE v3.8.31 (Edgar 2004) via the Biopython wrapper. Once aligned, the sequences were converted back to the corresponding DNA sequences. The d_S and d_N /d_S scores of the human/macaque alignments were calculated using PAML codeml (Yang 2007) using the Bio.Pyhlo module (Talevich et al. 2012) from the Biopython wrapper, with the settings *seqtype = 1*, *runmode = 0*, *model = 0*, *Nsites = []* and an arbitrary tree. Only CDSs that produced a d_S score of less than 0.2 or a d_N /d_S score of less than 0.5 were retained to minimise the risk of pseudogene contamination (Savisaar and Hurst 2016). After this filtering, 13,187 multi-exon sequences remained.

Sequences were then grouped as before into paralogous families. 1,036 single-exon sequences were also extracted and grouped into paralogous families. 5' UTR sequences of both multi- and single-exon sequences were obtained by constructing the full-length mature transcript and querying for the index for where the CDS starts. The 5' UTR was defined as all nucleotides up to this index point. Introns of the sequences were extracted from the genome sequence using the coordinates from the relevant exon entries.

Motif sets

The INT3 motif was downloaded from the supplement of Caceres and Hurst (2013). Other ESE motif sets except Ke400 were obtained as described in Caceres and Hurst (2013) and Savisaar and Hurst (2018). ISEs were obtained from the supplement of Wang et al. (2012b). ISS motifs were obtained from the supplement of Wang et al. (2012c). RBP motifs were obtained from the supplement of Savisaar and Hurst (2017). For RBP motifs, those that had significant enrichment P-values were considered CDS-binding and those with significant depletion P-values were considered non-CDS-binding. We provide a brief overview of each ESE motif set below:

RESCUE: Motifs were derived computationally (Fairbrother et al. 2002; Fairbrother et al. 2004b), on the assumption that ESEs should be enriched in constitutively spliced exons and avoided in flanking introns and be more frequently when splice sites are weak. Internal exons and flanking introns were queried. Results were experimentally validated and compared with prior data.

Ke400: A systematic experimental analysis (Ke et al. 2011) where all 4,096 hexamers were substituted at five positions in two internal exons in mini-gene constructs. These constructs were transfected to human cells with the splice promoting ability of each motif reported. The top 400 most potent splice modifying hexamers were retained for the Ke400 dataset.

ESR: Motifs were derived computationally (Goren et al. 2006), searching human-mouse orthologous exons with the same lengths, shorter than 250 nucleotides and with classical GT-AG splice sites. Two expected metrics were used to query di-codon frequencies, assuming the two codons appear independently. The first, expected conservation rate (ECR), multiplied the probability of codon 1 to be conserved between human and mouse, the probability of codon 2 to be conserved between human and mouse and the number of times the di-codon appeared conserved between human and mouse. For each di-codon, this reflects the expected frequency of observing a conserved human-mouse di-codon. The second, expected observation rate (EOR), multiplied the number of times the pair of amino acids encoded by the di-codon was detected in the data. These numbers were compared with the real frequency of conserved and occurred di-codons. Only di-codons that were statistically significantly overrepresented and highly conserved at synonymous sites were considered.

PESE: Computationally derived motifs (Zhang and Chasin 2004) comparing frequencies of octamers overrepresented in constitutively spliced noncoding exons versus unspliced pseudoexons and 5' UTRs of intronless genes, assuming ESEs are not frequently in pseudoexons and UTRs are devoid of ESE activity. Experimental confirmation of many ESEs subsequently provided (Zhang et al. 2005).

INT3: The motifs that appear in at least three of the RESCUE-ESE, Ke400, ESR and PESE datasets (Caceres and Hurst 2013). Considered a "gold standard" set and designed to have a low false-positive rate.

Generating compositionally-matched codon sets

All permutations of three unique codons were generated (N = 64 × 63 × 62 = 249,984), including stop codons (Figure 1A). However, 3! = 6 permutations of the same three codons exist (for example, the set [ATC, GAC, TCA] is equivalent to [GAC, TCA, ATC]) and so redundant sets were removed, leaving N = 249,984 / 6 = 41,664 codon sets (Figure 1B). To control for the net GC content of stop codons (Figure 1C), we filtered the remaining codon sets to retain only those with identical net GC content as the stop codon set, GC content of a codon set being defined as the sum of the number of G and C residues of the three codons divided by 9 (the number of nucleotides). For example, the tri-codon set [AGT, AAT, GAT] has GC content of 0.222, the same as the stop codon set. There are 2,879 tri-codon sets with net G and C content identical to the stop codon set.

A purine-matched subset (N = 6,856) was also derived by taking all sets with identical purine content as the set of stop codons (net purine content = 0.667). Note the size of the GC- and purine-matched codon sets differs as a result of the GC content (0.222) being more extreme than the purine content (0.666), with the smallest groupings of codon sets being those with the most extreme content, following binomial principles.

The intersection of these two groupings of tri-codon sets contained those tri-codon sets with both equal GC- and purine content to the stop codons (N = 473). We performed

further restrictions to generate sets with identical GC content but that contained no stop codons (N = 2,121) and with identical GC content but in which no codon could overlap with any others from the same set (N = 131).

Generating dinucleotide-matched motif sets

Sequences within a motif set (e.g. INT3) were scanned for every dinucleotide in both reading frames (e.g. the motif GAAGTA contains the dinucleotides GA, AG, TA, AA, GT). The frequencies for each dinucleotide were totalled for all motifs in the dataset. Then, for each simulation iteration, for each real motif (typically six nucleotides) a pseudo-motif of the same length was generated by randomly sampling dinucleotides with probabilities defined by the true dinucleotide frequencies calculated (i.e. for a real motif of length six, three dinucleotides were randomly sampled). If a motif was not of even length, a random nucleotide was sampled using the distribution of nucleotides in the true motif set and appended pseudo-motif. If the new pseudo-motif had already been generated in that iteration, it was removed and the process restarted. Each simulation iteration therefore contained an identical number of pseudo-motifs as the number in the true set.

Density calculations

We calculated density as outlined in the introduction. If a query motif overlapped another, overlapping nucleotides were only counted once. For example, for the query motif set [CCT, GGG] in the sequence TGATA**GGGG**A we only consider the 4 nucleotides that match the query motifs.

While we refer to this metric as "codon density" or "motif density", this term can be slightly misleading as we count the number of nucleotides matching the motif/codon, not the number of matching motifs/codons per se. This density metric, however, does enable us to control for varying query motif or queried sequence lengths. The metric therefore describes how much of a particular sequence is comprised by the query motifs (a codon, ESE etc) and therefore has a minimum of zero (the sequence contains no nucleotides matching the query motifs) and maximum of one (all nucleotides in the sequences match one or more of the query motifs).

Calculating fold-enrichment scores

We employ a FE metric in several cases to describe the deviations of true measures of abundance from that expected given underlying nucleotide distributions. Null expectations were obtained via iterated simulations. We provide an example below of calculating codon set density in ESE hexamers, but the same method is applied to calculating SCD in full-length lincRNA sequences, SCD in individual exons and pORF lengths in the lincRNA sequences.

To calculate the FE of any given codon set in the INT3 ESEs, we first calculated the raw density of the codon set (as detailed above) within the true INT3 ESE hexamers. Second, having generated randomised sets of pseudo-ESE motifs (as detailed above; for other calculations, these are sets of randomised shuffled sequences), we calculated the density of the codon set in each iteration of the randomised pseudo-ESE motif sets. This provided us with a density score for the codon set in the real ESE motifs and distribution of density scores from the simulated motifs. FE was then calculated using the formula $FE = \frac{O-E}{E}$, where O = observed density of the codon set motifs in true motifs and E = mean density of the codon set motifs in the simulant motifs.

FE as a metric has the benefit that FE < 0 implies a relative depletion given underlying nucleotide content, FE > 0 a relative enrichment and FE ≈ 0 as expected given underlying nucleotide content.

Calculating Z scores

Z scores for pORF lengths were calculated similarly to FE scores. First, the longest pORF in any frame in the true sequence was calculated. Then the longest pORF was calculated for each randomisation of the lincRNA sequences. The Z score for each sequence was then defined as the real longest pORF length minus the mean of the group of simulated longest pORFs, divided by the standard deviation of the simulated longest pORFs, taking the median Z score for sequences that are members of a paralogous family.

Predicting hits to motifs in sequences

Regular expressions were used to predict hits to motifs in sequences using the standard built-in Python package. For each sequence, the indices for the hits to each motif were

stored. Any subsequent hits to a motif were appended to this list. For each sequence, the list of indices was then filtered such that each index could only appear once. In this way, if two motifs overlapped, we would only consider the nucleotides that matched both only once in our calculations.

Removal of sequence matching ESEs

To interrogate sequence that featured no ESEs given a particular ESE set, hits were first predicted to the ESEs for each query sequence. The index of each nucleotide hit that overlapped an ESE for each sequence was stored, and only once all motifs had been queried were these indices further considered. In this way, all overlapping motifs were identified. For each sequence, the positions corresponding to indices that were not stored were calculated and the corresponding sequence parts extracted. Sequence parts interrupted by a predicted ESE were treated as separate sequence parts. This prevents unexpected motifs being generated by concatenating the remaining sequence. For example, querying the sequence ACTAC**TTTTT**AGA for the motif TTT would have resulted in two unmatched parts, ACTAC and AGA. Analyses were then performed on these remaining sequences individually.

Identifying potential open reading frames

Potential ORFs were identified by scanning each sequence for every ATG in every frame. For each ATG, downstream codons in the matching frame were then queried in order until a stop codon was identified. The nucleotide distance to the stop codon was stored and once all ATG's had been queried, the longest ORF was retained. Seven of the lincRNA sequences contained no potential ORF and so were excluded from the analysis.

Calculating empirical P-values

Empirical P-values were calculated using outputs from the simulations using the formula $P \approx \frac{m+1}{n+1}$ where m = the total number of simulants scoring less than or equal to the real value and n = the total number of simulants. If the direction of the one-tailed test was in the opposite direction, m = the total number of simulants scoring greater than or equal to the real value.

Acknowledgements

We would like to extend our thanks to our colleagues Rosina Savisaar and Alex Ho for reading and providing useful feedback on our manuscript. This work was supported by the European Research Council (Advanced Grant ERC-2014-ADG 669207 to L.D.H.). We also thank the reviewers for their useful comments.

Author Contributions

L.A. and L.D.H. conceived the project. L.A. performed the analyses. L.A. and L.D.H. wrote the manuscript. All authors have read and approved the manuscript.

Tables

Table 1: Stop codon densities (SCD) and fold-enrichment (FE) scores calculated from dinucleotide-matched controls for various RNA-binding protein motif sets. The PESE motif set marked * was derived from analysis of constitutively spliced noncoding exons, unspliced pseudoexons and 5' untranslated regions of intronless genes. The motif sets marked § indicate those not located within coding sequence.

Motif set	Number of motifs	Proportion containing stop codons	SCD	FE	P-value [a]
INT3 ESE	84	0.107	0.054	-0.459	0.020
RESCUE ESE	238	0.126	0.065	-0.404	9.999×10^{-5}
Ke400 ESE	400	0.063	0.033	-0.391	0.001
ESR ESE	285	0.109	0.054	-0.479	9.999×10^{-5}
PESE ESE*	2,069	0.222	0.084	-0.122	5.000×10^{-4}
ISE§	110	0.436	0.150	-0.034	0.417

ISS[§]	103	0.427	0.146	-0.068	0.307
RBP motifs (CDS)	232	0.103	0.046	-0.450	9.999×10^{-5}

[a] One-tailed empirical P-value asking whether the real set of motifs have significantly less stop codons than simulated motif sets.

Table 2: A summary of various tests of sequence composition for sequences in the two lincRNA data sets.

	Sequence set	
	Cabili et al. (2011)	**Lagarde et al. (2017)**
Number of sequences	1,919	456
SCD	0.130	0.128
FE [a]	-0.162 $P \approx 9.99 \times 10^{-4}$	-0.169 $P \approx 9.99 \times 10^{-4}$
Number of GC-matched codon sets with FE > stop codon set FE [b]	2,315/2,879 (80.41%) $P < 2.2 \times 10^{-16}$	2,300/2,879 (79.89%) $P < 2.2 \times 10^{-16}$
Number of GC-matched codon sets excluding stop codons with FE > stop codon set FE [b]	1,771/2,121 (83.50%) $P < 2.2 \times 10^{-16}$	1,751/2,121 (82.56%) $P < 2.2 \times 10^{-16}$
Number of sequences with exonic SCD < intronic SCD [b]	1,320/1,919 (68.79%) $P < 2.2 \times 10^{-16}$	325/456 (71.27%) $P < 2.2 \times 10^{-16}$
Median single-exon sequence exon SCD	n/a	0.139 [c]
Median multi-exon sequence exon SCD	n/a	0.122 [c]
Median single-exon sequence exon FE	n/a	-0.148 [d]
Median multi-exon sequence exon FE	n/a	-0.162 [d]

[a] One-tailed empirical P-value

[b] One-tailed binomial P-value, null probability of success = 0.5

[c] $P = 8.878 \times 10^{-14}$, Wilcoxon rank sum test between the SCD of each gene's exons and introns

[d] $P = 8.878 \times 10^{-14}$, Wilcoxon rank sum test between the FE scores of each gene's exons and introns

Table 3: A summary of individual sequence fold-enrichment (FE) scores after comparisons with randomised simulations of the same gene.

	Sequence set	
	Cabili et al. (2011)	**Lagarde et al. (2017)**
Sequences with FE < 0 [a]	1,528 (79.62%) $P \approx 0$	416 (91.23%) $P \approx 0$
Sequences with FE < 0, empirical P < 0.05 [b]	493 (25.69%) $P = 1.23 \times 10^{-200}$	206 (45.18%) $P = 2.33 \times 10^{-139}$

[a] One-tailed binomial P-value, null probability of success = 0.5

[b] One-tailed binomial P-value, null probability of success = 0.05

Table 4: Density of each the three stop codons in each of the ESE motif sets.

	Codon density in motif set		
Motif set	**TAA**	**TAG**	**TGA**
INT3	0	0	0.054
ESR	0.005	0.011	0.039
Ke400	0	0	0.033
PESE	0.005	0.007	0.071
RESCUE	0	0	0.065
Combined	0.005	0.007	0.068

References

Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15:193-204.

Berget SM. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* 270:2411-2414.

Boyer J, Badis G, Fairhead C, Talla E, Hantraye F, Fabre E, Fischer G, Hennequin C, Koszul R, Lafontaine I, et al. 2004. Large-scale exploration of growth inhibition caused by overexpression of genomic fragments in Saccharomyces cerevisiae. *Genome Biol* 5:R72.

Brockdorff N, Ashworth A, Kay GF, Mccabe VM, Norris DP, Cooper PJ, Swift S, Rastan S. 1992. The Product of the Mouse Xist Gene Is a 15 Kb Inactive X-Specific Transcript Containing No Conserved Orf and Located in the Nucleus. *Cell* 71:515-526.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25:1915-1927.

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* 62:89-98.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7:98-108.

Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* 15:343-351.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 104:19428-19433.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* 39:D301-308.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* 4:49-60.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* 4:e1000176.

Dirksen WP, Hampson RK, Sun Q, Rottman FM. 1994. A purine-rich exon sequence enhances alternative splicing of bovine growth hormone pre-mRNA. *J. Biol. Chem.* 269:6431-6436.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007-1013.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.* 32:W187-190.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211-222.

Frith MC, Bailey TL, Kasukawa T, Mignone F, Kummerfeld SK, Madera M, Sunkara S, Furuno M, Bult CJ, Quackenbush J, et al. 2006a. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol* 3:40-48.

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM. 2006b. The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2:e52.

Gersappe A, Pintel DJ. 1999. CA- and purine-rich elements form a novel bipartite exon enhancer which governs inclusion of the minute virus of mice NS2-specific exon in both singly and doubly spliced mRNAs. *Mol. Cell. Biol.* 19:364-375.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol. Cell* 22:769-781.

Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* 14:R49.

Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* 21:333-346.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11:97-108.

Jacob F. 1977. Evolution and tinkering. *Science* 196:1161-1166.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313-1326.

Kashi K, Henderson L, Bonetti A, Carninci P. 2016. Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochim Biophys Acta* 1859:3-15.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360-1374.

Krchnakova Z, Thakur PK, Krausova M, Bieberstein N, Haberman N, Muller-McNicoll M, Stanek D. 2019. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Res.* 47:911-928.

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Perez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* 49:1731-1740.

Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* 12:2411-2415.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A* 103:9935-9939.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27:i275-282.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4:865-875.

Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. *J Genet* 92:155-161.
Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12:1827-1836.
Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* 10:R124.
McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* 370:20140332.
McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17:567-578.
Niazi F, Valadkhan S. 2012. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA* 18:825-843.
Ohno S. 1970. Evolution by Gene Duplication. Springer.
Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, Isobe T, Sugano S. 2004. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res.* 14:2048-2052.
Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22:1-5.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* 23:301-309.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* 24:1600-1603.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.
Paz I, Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y. 2010. SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res.* 38:W281-285.
Paz I, Kosti I, Ares M, Jr., Cline M, Mandel-Gutfreund Y. 2014. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* 42:W361-367.
Prasanth KV, Spector DL. 2007. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.* 21:11-42.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Vienna, Austria: R Foundation for Statistical Computing.
Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499:172-177.
Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R, Marques-Bonet T, Alba MM. 2015. Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genet.* 11:e1005721.
Savisaar R, Hurst LD. 2016. Purifying Selection on Exonic Splice Enhancers in Intronless Genes. *Mol. Biol. Evol.* 33:1396-1418.

Savisaar R, Hurst LD. 2017. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Mol. Biol. Evol.* 34:1110-1126.
Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* 28:1442-1454.
Schlotterer C. 2015. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet.* 31:215-219.
Schuler A, Ghanbarian AT, Hurst LD. 2014. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* 31:3164-3183.
Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A. 2014. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* 289:10950-10957.
Smithers B, Oates ME, Gough J. 2015. Splice junctions are constrained by protein disorder. *Nucleic Acids Res.* 43:4814-4822.
Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* 21:1563-1571.
Talevich E, Invergo BM, Cock PJ, Chapman BA. 2012. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* 13:209.
Tanaka K, Watakabe A, Shimura Y. 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol. Cell. Biol.* 14:1347-1354.
Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692-702.
The Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-74.
Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet.* 15:e1008160.
Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012a. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22:1798-1812.
Wang Y, Ma M, Xiao X, Wang Z. 2012b. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* 19:1044-1052.
Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z. 2012c. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat. Struct. Mol. Biol.* 20:36-45.
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:R29.
Will CL, Luhrmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 3:a003707.
Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol* 3:1245-1252.
Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. *Genome Biol* 11:R20.
Xu R, Teng J, Cooper TA. 1993. The cardiac troponin T alternative exon contains a novel purine-rich positive splicing element. *Mol. Cell. Biol.* 13:3660-3674.

Yang S, Wang J, Ng RT. 2018. Inferring RNA sequence preferences for poorly studied RNA-binding proteins based on co-evolution. *BMC Bioinformatics* 19:96.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586-1591.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46:D754-D761.

Zhang JZ. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18:292-298.

Zhang XH, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* 18:1241-1250.

Zhang XH, Kangsamaksin T, Chao MS, Banerjee JK, Chasin LA. 2005. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.* 25:7323-7332.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* 343:769-772.

Supplement to Chapter 2

The Supplementary Tables for Chapter 2 can be found on the attached CD.

Supplementary Texts

The Supplementary Texts and Supplementary Text Figures presented below can also be found accompanying the published paper at the link described in the introduction to Chapter 2. These have been reformatted for this thesis.

Supplementary Text 1

The presence of stop codons within the tri-codon codon sets nor the inability for the stop codons to overlap can explain the stop codon depletion in the INT3 ESE motifs when compared with compositionally matched codon sets

We find that when controlling for both GC content and purine content, the stop codon set is depleted in the INT3 ESE motifs. However, are these results due to a subset of GC-matched codon sets also containing stop codons?

To eliminate this potential bias, we performed the same analysis as in the main text using tri-codon sets containing no stop codons and compared them with the stop codon set. N = 2,121 sets have identical GC content and contained none of the three stop codons. We find a similar result as before – 1,578/2,121 (74.40%) have higher FE than the stop codon set ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5). When matching by purine-content, the removal of codon sets containing stop codons left N = 5,587 purine-matched codon sets. Again, 4,457/5,587 (81.56%) have higher FE than the stop codon set, a significant number ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5). The high proportion of codon sets with a greater FE than for the stop codon set is therefore not a consequence of some stop codon sets containing stop codons.

It is also the case that no stop codon can overlap with another stop codon. If any given motif contains a stop codon, there is therefore less chance it will contain another stop

codon. For example, in the motif G**TAA**AA, no second stop codon can exist without destroying the motif. Yet, if the query codon set is [AAA], as AAA can overlap (e.g. GT**AAAA)** it is more likely to have a higher density. We therefore restricted the codon sets again, but only considered codon sets where no codon could overlap another (for example the sets [AAA, CCC, TAT] and [TTA, TTC, TTG]). For GC-matched codon sets with no overlapping codons (N = 131), 91/131 (69.47%) have greater FE than the stop codons ($P = 4.898 \times 10^{-6}$, one-tailed exact binomial test, null probability of success = 0.5). For the purine-matched non-overlapping sets (N = 712), 565/712 (79.35%) have a greater FE than for the stop codon set, again a significant number ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5).

That some stop codons appear in the tri-codon sets and that stop codons are unable to overlap with themselves do not therefore explain their depletion in ESEs.

Supplementary Text 2

Stop codon containing ESEs are per motif more frequent in exonic sequence than ESEs that contain no stop codons, although differences can be attributed to differences in dinucleotide content

One possibility that might explain why stop codons are depleted in ESEs is that the motifs containing stop codons are for some reason poorer functioning ESEs in any context. Evidence in *Drosophila sp.* and *C. elegans* suggests individual SR proteins have functional differences (Ring and Lis 1994; Hoffman and Lis 2000; Kawano et al. 2000; Longman et al. 2000; Kim et al. 2003) and by association, the motifs they bind may therefore be used at different frequencies. If so, there could be a selective pressure against using stop codon containing motifs. This may have a direct consequence in that they are used less frequently, but also an indirect consequence in that they are then less likely to be identified using computational approaches. In this scenario, there would be no reason to suppose any CDS-imposed constraint is transferred.

To address this, we ask whether ESEs that contain stop codons are suboptimal and hence underused given their relative frequency within the set of ESEs. If stop codon containing motif frequencies deviate from neutral expectations, it could indicate that selective pressures are acting against the usage of such motifs. We focus analyses on the "gold-standard" INT3 set of motifs of highly constrained core motifs likely to be true functional binding motifs.

If functional differences are leading to depletion, it is expected that the stop codon containing motifs are used less frequently per motif in exonic sequences. Hits to the INT3 motifs were predicted in human protein-coding coding exons, picking one transcript sequence per paralogous family at random (N = 6,045 family data points). The density of stop codon containing ESEs is much reduced (0.026) when compared with the density of the remaining ESEs (0.169). However, two factors could explain this difference. First, only 9/84 of INT3 motifs contain a stop codon. Second, stop codon containing ESEs cannot be incorporated into one of the three reading frames. This raw difference therefore has little meaning and these two factors must be controlled for.

To control for the reading frame effect, hits to the stop codon containing motifs were predicted separately in each reading frame and the number of predicted hits scaled by the number of motifs that can function in that frame. This scaling is not uniform (i.e. TGAAGA, TGAAGC, TGAAGG, TGAGAA cannot function in the +0 frame, AATGAC, AGTGAC, GATGAA in the +1 frame or CTGAAG, GTGAAG in the +2 frame; frames denote motif start position relative to the ORF). In the +0 frame, only 5/9 motifs contribute to hits and we are therefore in effect only sampling 5/9 of potential hits that a motif not containing a stop codon could contribute. The raw hit count was therefore multiplied by 9/5 (and similarly 9/6, 9/7 for the +1, +2 frames).

Second, to control for the number of motifs per class (stop codon containing motifs or motifs not containing stop codons), the frame-normalised total hits in each class were divided by the number motifs in each class (N = 9 for stop codon containing and N = 75 for motifs that contain no stop codons). This provided the total hits per frame per motif for both classes of motif. This total was further normalised to give the number of hits per 1,000 bp of protein-coding exonic sequence to give a normalised per motif per 1,000 bp hits (NMH) for all coding exons. An NMH = 2 for stop codon containing motifs, for example, means that on average there exist two hits to each stop codon containing motif per 1,000 bp of lincRNA sequence.

We find the stop codon containing INT3 ESE NMH = 0.928 is significantly greater than that for 1,000 sets of dinucleotide-matched and stop-codon matched (that is, the same number of motifs contain a stop codon for each iteration) pseudo-ESE motifs (median simulant NMH = 0.647, $P \approx 0.009$, one-tailed empirical P-value). Whilst motifs that do not contain stop codons are also found significantly more frequently than dinucleotide matched pseudo-motifs (NMH = 0.596, median simulant NMH = 0.465, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value), the difference between the two NMH values for the real INT3 ESE motifs argues that, if anything, stop codon containing motifs are more frequent and not avoided.

Despite this greater per motif use of the stop codon containing motifs, this could be explained if the stop codon containing motifs better match the nucleotide composition

of protein-coding exons. Is the difference therefore greater than expected? We calculated the ratio between NMH value of stop codon containing motifs to that of motifs that contain no stop codons (0.928/0.596 ≈ 1.559) and asked whether this is greater than expected by chance. To define chance, we considered the 1,000 sets of simulated pseudo-ESEs and calculated the same ratio. The ratio for real ESEs is not significantly greater than the equivalent ratio for the null pseudo-ESE motif sets (median simulant NMH ratio = 1.399, P ≈ 0.257, one-tailed empirical P-value).

Thus, although stop codon containing ESEs are per motif more frequent, this result suggests the increased usage between the two classes if not significantly greater than expected. Repeating the analysis for a total of 10 runs to control for paralogous family member choice, per motif stop codon containing ESE enrichment remains significantly greater than controls (median P ≈ 0.009, one-tailed empirical P-value) but not significantly greater than per motif usage of the remaining ESEs (median P ≈ 0.257, one-tailed empirical P-value, Supplementary Table 2) and results are therefore not biased to sequences interrogated.

If stop codon containing ESEs were of lesser quality and under weaker selection or selected against, there should be a depletion of the stop codon containing motifs relative to motifs not containing stop codons. We find no evidence this is the case and no conclusive evidence to argue the depletion of stop codons is a result of avoiding employing stop motifs as a consequence of being less functional.

Supplementary Text 3

The density of high quality, low false-positive stop codon containing ESE motifs increases as flanking intron size increases

Results suggest the relative usage of stop codon containing motifs per motif is not significantly greater than for motifs of similar dinucleotide content. However, the per motif frequency may not be the most informative measure if the quantity of splice information incorporated is important for ensuring accurate splicing. For example, by having more contributing motifs, the combined quantity of splice information encoded by ESEs not containing stop codons may be important. Thus, total ESE frequency (or density) rather than per motif frequency could be more informative.

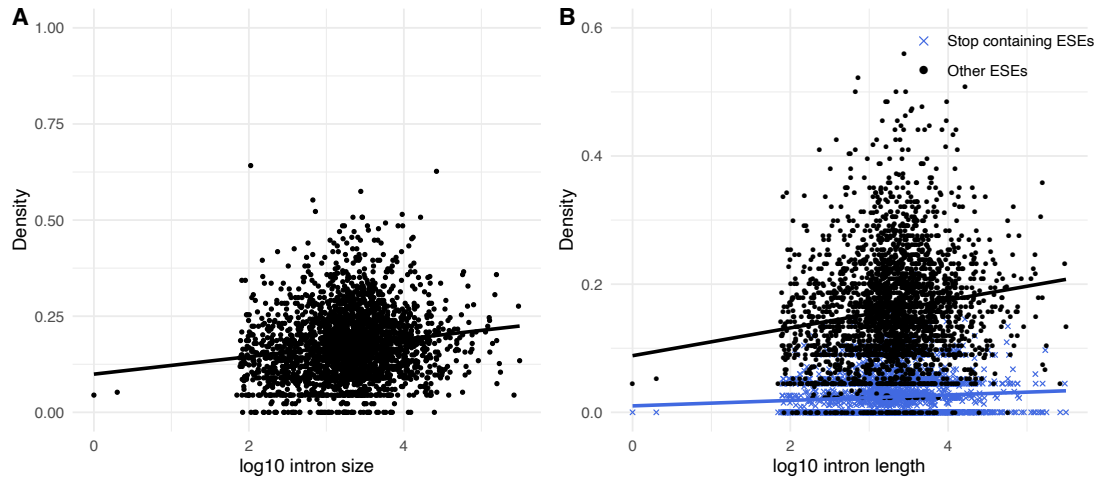
ESE density is known to be positively correlated with intron size in protein-coding genes (Dewey et al. 2006; Caceres and Hurst 2013; Wu and Hurst 2015), thought to be a result of reinforcement selection increasing the quantity of splice information to distract SR proteins away from possible cryptic splice sites (Wu and Hurst 2015). Any differences how stop codon containing motifs and the remaining motifs are employed as the flanking intron size increases may therefore provide further insight.

We first check whether ESE density in general increases as flanking intron size increases in our protein-coding exon dataset after grouping paralogous family members (Dewey et al. 2006; Caceres and Hurst 2013; Wu and Hurst 2015). We find this to be the case ($\rho = 0.198$, $P = 1.04 \times 10^{-24}$, Spearman's rank correlation, Supplementary Texts Figure 1A). As ESEs typically reside in exon flanks in protein-coding genes and therefore more likely to be splice related, we limited sequences to only the 5' and 3' flanks (nucleotides 2-69 from exon boundary) of sequences greater than 207 nucleotides in length (to restrict sequences to those with both 5'/3' flanks and core regions). By doing this, we also control for the quantity of sequence in which the ESEs can reside, and thus ESE density directly reflects how many ESEs are incorporated. Again, in the exon flanks, we find significant positive correlations with intron length ($\rho = 0.136$, $P = 2.79 \times 10^{-12}$, Spearman's rank correlation). Results are not subject to biases as a result of picking one member at random from each paralogous

family grouping as we find a similar significant positive correlation when using all sequences ($\rho = 0.158$, $P = 2.05 \times 10^{-61}$, Spearman's rank correlation).

If stop codon containing ESEs are less functional and avoided, it would be expected that they are used less frequently as flanking intron size increases. We asked whether the usage of stop codon containing ESEs and ESEs containing no stop codons differs as intron size increases. ESEs were grouped into stop codon containing motifs and others and densities calculated for the two classes. To control for the number of motifs in each class and the reading frame restrictions on the stop codon containing ESEs, we only compare the correlations between densities and intron size for each the two ESE motif classes rather than directly comparing densities themselves. We find correlations are significantly positive for both stop codon containing ($\rho = 0.139$, $P = 1.07 \times 10^{-12}$, Spearman's rank correlation) and the remaining ESEs ($\rho = 0.190$, $P = 2.17 \times 10^{-23}$, Spearman's rank correlation) (Supplementary Texts Figure 1B), suggesting the usage of both classes of ESE increases as greater quantities of splice information are required.

Is the density of stop codon containing motifs significantly positively correlated when restricted only to exon flanks? Again, correlations of both stop codon containing ($\rho = 0.078$, $P = 4.586 \times 10^{-5}$, Spearman's rank correlation) and remaining ESEs ($\rho = 0.132$, $P = 1.04 \times 10^{-11}$, Spearman's rank correlation) with intron length are significantly positive. Interestingly, the density of stop codon containing motifs does increase significantly slower than for the remaining ESEs ($Z = 5.176$, $P = 2.26 \times 10^{-7}$, two-tailed Z-tests of equivalency in exon flanks) meaning it could be the case that the ESEs not containing stop codons are more specialist (they have, for example, greater binding affinities), or stop codon containing ESEs have additional constraints (which could be the case due to reading frame requirements). Importantly, evidence suggests stop codon containing ESEs are not avoided as splicing becomes more difficult.



Supplementary Texts Figure 1: The log10 median lengths for introns versus the density of INT3 ESE motifs for the flanking 2-69 nucleotides of (A) all ESEs combined and (B) ESEs grouped as stop codon containing and the remaining ESEs. In all cases, correlations between flanking intron size and densities are significantly positively correlated ($P < 0.01$, Spearman's rank correlations).

Supplementary Text 4

Stop codon containing ESE motifs can overlap with fewer motifs than the remaining ESEs, but not less than dinucleotide matched controls

The result in Supplementary Text 3 suggests that while stop codon containing motifs are not avoided as splicing becomes more difficult, motifs not containing stop codons are increasingly used over those containing stop codons. Are there constraints at the sequence level such that ESEs containing no stop codons are used more frequently? If it is the case, stop codon containing motifs may still increase within increasing flanking intron size because they provide splice function, but usage may be restricted. In such a case, it could again help to explain the stop codon depletion beyond protein-coding constraints predicted under our transfer selection model.

We note that we observe an increase in ESE density for both stop codon containing motifs and others even when for exon flank sequences the size of the sequence scanned is controlled. In this scenario, for ESE density to increase there is an increased probability that motifs will have to overlap to include the splice information. We calculated the number of overlapping motifs and non-overlapping motifs in each exon (both classes of motifs combined), overlapping defined as at least one nucleotide shared between two or more motifs. We find a weak yet significant correlation between both the raw number of overlapping motifs ($\rho = 0.065$, $P = 8.127 \times 10^{-4}$, Spearman's rank correlation) and the proportion of overlapping motifs within a sequence ($\rho = 0.077$, $P = 1.069 \times 10^{-4}$, Spearman's rank correlation) as flanking intron size increases, suggesting more overlapping does occur as more splice information is required.

However, to overlap in protein-coding sequence the motifs in addition to functioning as RBP-binding sites must, at a minimum, also encode the correct amino acids (although other constraints e.g. RNA secondary structures are likely to be imposed). In protein-coding sequence, this is more difficult if a motif contains a stop codon. For example, suppose there exists the sequence NNA|AGA|T**T**A|AGA. A **T** → **G** mutation creates two ESE motifs GATGAA and TGAAGA but also creates a premature stop

codon. However, a **T** → **C** mutation does not alter amino acid content nor generate an in-frame stop codon, whilst also creating ESE motifs AAGATC and TCAAGA that could serve as potential functional SR protein binding sites. If there exists a selection pressure to maximise splice information, the only useful motifs would be those containing no stop codon (the probability the mutation reaches fixation would then be determined by whether the new binding motifs provide sufficient functional benefit). At the sequence level, stop codon containing ESEs may therefore be less preferable at higher ESE densities not because they are poorer functioning motifs (e.g. reduced ability to ensure accurate splicing), but because they are more difficult to include.

To establish whether this can explain the depletion of stop codons in ESEs, we must first ask whether stop codon containing motifs themselves are inherently more difficult to overlap – that is, are stop codon containing motifs less likely to be used because their dinucleotide content prevents them from overlapping with other motifs as frequently.

For each INT3 motif, we calculated the number of other motifs (including the focal motif) that overlap the focal motif by at least two nucleotides. For example, the two motifs AT**GTAA** and **GTAA**TA share a four-nucleotide overlap. The mean number of motifs a stop codon containing motif can overlap with (28.667) is fewer than for the motifs containing no stop codons (36.800). Is this attributable to the dinucleotide content of the motifs? Calculating the mean number of overlaps for the 1,000 sets of dinucleotide- and number of stop codon-matched null motif sets (stop codon-matched such that an equal number of pseudo ESE motifs per iteration contain stop codons), the real stop codon containing motifs do not overlap with significantly fewer motifs when compared with the controls ($P \approx 0.732$, one-tailed empirical P-value). A fold-enrichment overlap score (FEO) of 0.083 (calculated as per FE using the real stop codon containing mean overlap and simulant ESE mean overlap) confirms the insignificant difference.

Motifs containing no stop codons, however, can overlap with a significantly greater number of motifs than expected (FEO = 0.333, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value). Moreover, this difference between the number of overlaps between the stop

codon containing and remaining ESEs is also greater than expected ($P \approx 0.031$, one-tailed empirical P-value). Using the RESCUE set of ESEs, the number of motifs a motif can overlap with is significantly greater than expected for both stop codon containing (75.833, $P \approx 0.003$, one-tailed empirical P-value) and the remaining motifs (77.390, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value), with this difference not significantly greater than the control motifs ($P \approx 0.148$, one-tailed empirical P-value), suggesting the ability to overlap is not subject to motif set bias.

Taken together, rather than the dinucleotide content of stop codon containing motifs making it more difficult to overlap, motifs containing no stop codons can consistently overlap more frequently with other motifs. Thus, the relative increase in usage of ESEs containing no stop codons with intron size is unlikely to be a result of difficulties including a stop codon containing motif where two are required to overlap.

Supplementary Text 5

A greater frequency of ESE motifs containing stop codons when required to overlap another ESE motif from that expected argues against a stop codon depletion in the ESEs due to an inability to be combined with other motifs

The result in Supplementary Text 4 raises an interesting question. Could the composition of real ESE motif set be due to an enrichment of motifs containing no stop codons that are more readily able to overlap and as a consequence depleted for stop codon containing motifs, independent of any protein-coding constraints? If so, we expect stop codon containing motifs to be used less frequently when overlapping another motif in exonic sequences.

We therefore predicted hits to all motifs within protein-coding exons and for each motif asked whether it overlapped another. For both overlapping motifs and non-overlapping motifs, we then asked what proportion contained stop codons. In this way, we can establish whether stop codon containing motifs are used less frequently when involved in an overlap, controlling for the fact these motifs are fewer in number. For the INT3 set, we find the opposite to be true – a slightly higher proportion of motifs contain a stop codon if overlapping another motif (0.106) than if not overlapping (0.098). Although small, this is a highly significant increase ($\chi^2 = 49.863$, $P = 1.649 \times 10^{-12}$, chi-square test of raw frequencies using non-overlap frequency as the expected frequency). A similar significant increase is observed for the RESCUE set (overlap stop proportion: 0.143, non-overlap stop proportion: 0.119, $\chi^2 = 1041.418$, $P = 1.649 \times 10^{-12}$, chi-square test).

Thus, although stop codon containing motifs are intrinsically more difficult to overlap (see Supplementary Text 4), we instead find a significant increase in usage from that expected if a stop codon containing motif overlaps another motif. This argues against the hypothesis that stop codon containing motifs are harder to include in overlapping motifs in the real sequences. The depletion of stop codons in ESEs is unlikely to result from stop codon containing motifs being of less utility in forming part of a larger binding motif. This result would also suggest that much of the increase in ESE density

as intron size increases is likely due to singular motifs. Together, these results argue against stop codon containing ESEs being of poorer function.

Supplementary Text 6

The depletion of stop codons in lincRNA sequences is not a result of biases due to the sequences chosen paralogous family groupings

To dismiss the possibility that the depletion of stop codons in lincRNA is due to biases as a result of the random paralogous family member chosen, we ran the analyses shuffling the Cabili et al. (2011) lincRNA sequences 10 times in total. The depletion of stop codons remains significant in all cases (median $P \approx 9.99 \times 10^{-4}$; median real SCD = 0.130, median of median simulated SCDs = 0.155, median FE = -0.162, Supplementary Table 3). The same control applied to the pairwise test in the main text also holds (median all depletions $P \approx 0$, median significant depletions $P \approx 7.09 \times 10^{-201}$, Supplementary Table 4). The depletion also exists if all sequences are considered and not grouped by family ($P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value, N = 4,646, SCD = 0.135, median simulated SCD = 0.158, FE = -0.149, Supplementary Table 5). We can therefore eliminate biases due to the individual sequences used as the reason for the depletion of stop codons seen in lincRNAs.

Supplementary Text 7

The depletion of stop codons in lincRNA sequences is robust in the second set of independently derived lincRNA sequences

To verify the depletion of stop codons in lincRNA sequences by eliminating biases due to the total set of lincRNA sequences chosen, we repeated the analysis using a more recently derived set of GENCODE sequences reannotated by RNA Capture Long Seq (CLS) (Lagarde et al. 2017). CLS enables manual-quality full-length annotations at high throughput levels and enables a quality assessment of protein-coding potential (Lagarde et al. 2017). A minority of these sequences had protein-coding potential, but none had peptide-based evidence of translation. This set therefore provides a second "clean" dataset with minimal protein-coding contamination.

Results are qualitatively similar to those when using the Cabili et al. (2011) sequences. In the dataset as a whole, stop codons are depleted when compared with shuffled versions (FE = -0.169, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value, Supplementary Table 6). Similarly, 91.23% (416/456) of the sequences have a stop codon depletion when compared with randomisations for the same gene, again a significant excess above null ($P \approx 0$, one-tailed exact binomial test, null probability of success = 0.5, Supplementary Table 7). Of these, 206/456 (45.18%) have a significant depletion ($P = 2.33 \times 10^{-139}$, one-tailed exact binomial test, null probability of success = 0.05). These results, as per the Cabili et al. (2011) set of sequences, are not biased by sequence chosen (Supplementary Tables 6-7). Further, the SCD in exons is less than in introns of the same gene for 325/456 (71.27%) genes ($P = 2.33 \times 10^{-139}$, one-tailed exact binomial test, null probability of success = 0.5).

Supplementary Text 8

The stop codon depletion in lincRNAs is not owing to hidden ORF contamination

Could the apparent lack of stop codons in lincRNAs be biased by a subset of sequences being under strong selection to avoid stop codons, as these sequences contain true but unrecognised ORFs?

The initial quality control of the lincRNA datasets argues against this. However, we also consider only the sequence upstream of the first annotated ATG (in any frame) in lincRNAs which should be devoid of protein-coding potential (although we cannot eliminate cases where transcription start sites are 5' to the annotated sequence). Again, we find a depletion of stop codons compared with the randomly shuffled nulls of these upstream sequences, robust to differing lengths of sequence before the ATG (median $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value, median real densities = 0.089, median simulant densities = 0.114, Supplementary Table 8). The stop codon depletion in lincRNA is therefore not parsimoniously explained as simple annotation artefact.

Supplementary Text 9

The stop codon density in multi-exon 5' untranslated sequences is significantly greater than in single-exon sequences, but not after controlling for nucleotide composition

Under our model, it could be expected that the 5' untranslated regions (UTRs) should behave similarly to lincRNA sequences as these should have no underlying coding potential. Also, the first intron is often close to the ATG and hence to the UTR. Therefore, we hypothesise that those UTR sequences for multi-exon sequences should have significantly lower SCD than those for single-exon sequences. Considering only 5'UTRs of length greater than 50 nucleotides and after picking one sequence per paralogous family, this is what we find ($P = 2.218 \times 10^{-5}$, Wilcoxon rank sum test, median single-exon 5' UTR SCD = 0.074, median multi-exon 5' UTR SCD = 0.066).

However, when comparing the FE scores after generating null sequences by randomly shuffling the nucleotides of the UTR sequences, we find no significant difference between FE scores of the real lincRNA and the simulants ($P = 0.203$, Wilcoxon rank sum test). Thus, it appears that the 5' UTR sequences of multi-exon sequences are less conducive to incorporating stop codons than expected. Surprisingly, if anything it is the UTRs of single-exon sequences that are more deviated from the null (median single-exon 5' UTR FE = -0.150, median multi-exon 5' UTR FE = -0.141).

At first sight, this appears to contradict our hypothesis. However, upon closer inspection, it is the single-exon UTRs with significantly higher ESE density ($P = 3.667 \times 10^{-9}$, Wilcoxon rank sum test, median single-exon 5' ESE density = 0.113, median multi-exon 5' ESE density = 0.092). Consistent with our model, it is therefore the sequences with greater ESE density that have greater negative deviations from expected in SCD. Why the 5' UTR sequences of single-exon sequences contain more ESEs than those of spliced sequences is unanswered but could be due to the additional functional roles of ESEs beyond splicing (Savisaar and Hurst 2016).

Supplementary Text 10

Enrichment of stop codons in non-ESE lincRNA sequence may prevent inappropriate SR-protein binding

Is the enrichment of stop codons in lincRNA sequence that is not predicted to be ESE genuine? If not, and a consequence of the remaining sequence being conducive to generating stop codons, the SCD of the remaining sequence should be similar to a sequence in which motifs with similar dinucleotide content have been removed. However, if it is, the remaining sequence should have a greater SCD after removal of dinucleotide-matched controls. In this test, the dinucleotide-matched control motifs must also be matched in stop codon frequency otherwise the simulated remaining sequence may retain greater/fewer stop codons as there are more/fewer motifs with stop codons to potentially remove. After removal of motifs from the combined ESE set, the remaining sequence has a higher SCD than after removing sequence matching the dinucleotide- and stop codon-number matched control motifs (FE = 0.116, $P \approx 0.003$, one-tailed empirical P-value). This result is not affected by the randomly paralogous family member sequences chosen (Supplementary Table 9). This result is therefore consistent with a genuine increase in stop codons in lincRNA outside of ESE motifs.

Why then might the remaining sequence be enriched for stop codons? One explanation could be that there exists a selective pressure to include more stop codons in non-ESE sequence at exon cores or 3' flanks such that incorrect SR protein binding is less likely as the remaining sequence "appears" less like ESE. If such a selection pressure exists, the difference in SCD between ESE hits and the remaining sequence should be greater than similar comparisons when predicting hits to control motifs. This is what we find. The SCD in ESE hits is 0.087 and the remaining sequence 0.113. The ratio of SCD of ESEs to non-ESEs is $0.087/0.113 \approx 0.770$, significantly smaller than the equivalent ratio for the control motifs (median control motif proportion = 0.908, $P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value), indicating that there is indeed a larger difference in SCD in real sequence than expected. Whilst indicative, this result is also compatible with selection to incorporate stop codons that become premature termination codons

(PTCs), thereby making the transcript subject to nonsense-mediated decay (NMD) if erroneously recruited to the translation machinery (Niazi and Valadkhan 2012). We make no further inferences as to reasons for the increased SCD, but simply suggest there may be several regulatory mechanisms that would benefit from such selection.

Supplementary Text 11

The stop codon depletion is not a result of lincRNA sequences avoiding the use of stop codon containing ESEs

Our results suggest that despite the lack of translational constraint, stop codons are found less frequently than expected in lincRNA. While this is consistent with our transfer selection model, it could also be explained if, for some unknown reason and unlike protein-coding sequences, stop codon containing ESEs are less functional or avoided in lincRNAs. If this is the case, then stop codon containing motifs should be found less frequently per motif than the other ESE motifs and should be under-employed compared with sets of nucleotide composition-matched controls.

To address this, we calculated the raw number of hits within the lincRNA sequences to both stop codon containing ESE motifs and the remaining ESE motifs from the INT3 data set as per protein-coding sequences in Supplementary Text 2. From this, we calculated the normalised number of hits per motif per 1,000 bp (NMH) in lincRNA exons. With no reading frame constraints in lincRNA, the first part of the normalisation divided the total raw hits in each class by the total number of motifs contributing to each class (N = 9 stop codon containing ESE motifs, N = 75 other ESE motifs), to give the number of hits per ESE in each class. This value was further normalised to give the number of hits per 1,000 bp of exonic sequence.

We find that the stop codon containing INT3 ESEs have more hits within lincRNAs (NMH = 0.464) than for dinucleotide-matched simulant sets of pseudo-ESE motifs (see Supplementary Table 10), although not significantly so (median simulant NMH = 0.409 hits per stop codon containing motif, $P \approx 0.113$, one-tailed empirical P-value). We therefore conclude that the stop codon containing ESE motifs are not avoided when compared to null expectations. The NMH for the real INT3 motifs not containing a stop codon is also greater than for the pseudo-ESE motifs that do not contain stop codons (NMH = 0.489), but again not significantly so (median simulant NMH = 0.450 hits per motif for motifs not containing stop codons, $P \approx 0.055$, one-tailed empirical P-value).

While these results indicate no avoidance of stop codon containing ESE motifs in lincRNA, they are together surprising as we expect ESE motifs to be enriched in lincRNA compared to nucleotide-matched controls. When both the stop codon containing motifs and those containing no stop codons are combined, we do find the real ESE motifs are found more frequently than the simulant motifs in lincRNA ($P \approx 0.039$, one-tailed empirical P-value).

While neither of the sub-groups (those hexamers with a stop codon and those without) is significantly enriched in isolation, the data suggests a possible greater enrichment of the motifs that do not contain stop codons (these motifs are borderline significantly enriched). Might this indicate possible preferential usage of the motifs that do not contain a stop codon? To address we take the ratio between the hits per motif seen in the two groups (NMH for stop codon containing motifs = 0.464, for motifs without a stop codon = 0.489, ratio = 0.949) and ask whether this is lower than expected by chance.

To define chance, we consider sets of simulated ESEs in which the dinucleotide content of the real set is maintained but ensuring the total number of stop codons within the set of simulated ESEs equals that in the real ESEs. For each simulant set of ESEs we again split the motifs into two groups, those with stop codons and those without. For each class of pseudo-ESE, we determine the hits per motif in lincRNA and the ratio of average hits per motif within each group (those with a stop codon to those without). By repeating multiple times, we can then define the null distribution of values of the above ratio controlling for dinucleotide content of ESEs. From this, we determine that the observed ratio is not significantly lower than expected by chance, even employing a one-tailed test ($P \approx 0.656$ one-tailed empirical P-value, median simulant ratio = 0.905).

We conclude that there is no evidence stop codon containing ESEs are underemployed, given both the commonality of such ESEs and the dinucleotide content of ESEs. Results are not subject to biases due to the random sequences chosen (Supplementary Table 10). Furthermore, this is also not a result of the ESEs used. Using the RESCUE

ESE dataset containing more motifs (Supplementary Table 11), we find both hits to stop codon containing motifs (NMH = 0.444, median simulant NMH = 0.388, P ≈ 0.004, one-tailed empirical P-value) and hits to those motifs not containing stop codons (NMH = 0.413, median simulant NMH = 0.376, P ≈ 0.003, one-tailed empirical P-value) are significantly greater than matched pseudo motifs, while the ratio between the NMH values is not significantly greater than expected (ratio = 1.076, median simulant ratio = 1.030, P ≈ 0.747, one-tailed empirical P-value). This suggests that both the ESE motifs containing stop codons and those not containing stop codons are enriched in lincRNAs and that neither class is significantly more or less enriched than the other. The depletion of stop codons in lincRNA is unlikely to be a result of underuse of specific ESEs or avoidance of stop codon containing ESEs in lincRNA.

Supplementary Text 12

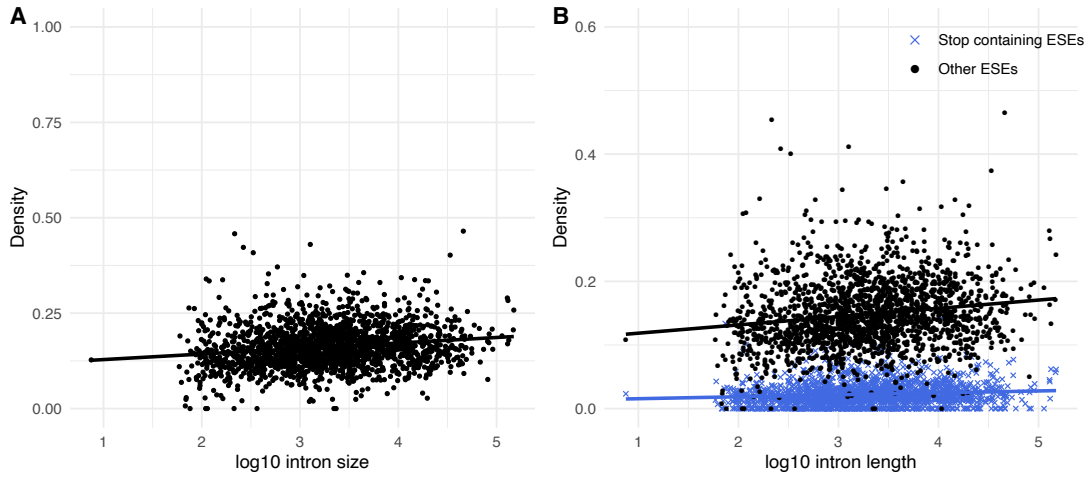
Stop codon containing ESEs are not avoided in lincRNA as intron size increases

Are stop codon containing ESEs functional in lincRNA sequences? As with protein-coding sequences (see Supplementary Text 3), one indication of their functionality would be an increased density as flanking intron size increases. Correlations between ESE density and intron size have previously been documented by Schuler et al. (2014) using the RESCUE set of motifs (Fairbrother et al. 2004). We find a similar trend for both full sequences ($\rho = 0.190$, $P = 4.37 \times 10^{-17}$, Spearman's rank correlation) and when restricted to exon flanks ($\rho = 0.010$, $P = 7.26 \times 10^{-5}$, Spearman's rank correlation) for our dataset (N = 1,919 lincRNA sequences).

Does the use of stop codon containing ESEs increase with intron size? As with protein-coding sequences, we again find a significant positive correlation in full sequences ($\rho = 0.126$, $P = 4.37 \times 10^{-17}$, Spearman's rank correlation) and flanking regions ($\rho = 0.049$, $P = 0.049$, Spearman's rank correlation, Supplementary Texts Figure 2), arguing that the motifs are indeed employed and functional. Although the signal is much weaker when restricted to flanks, only the density of motifs rather than exon location appears important in lincRNA (Schuler et al. 2014). Thus, with both the flanking regions and full sequences both displaying significantly positive trends, these results argue that stop motifs are likely functional. As per protein-coding genes, the density of ESEs containing no stop codons is also significantly positively correlated within flanking intron size (all sequence: $\rho = 0.180$, $P = 1.94 \times 10^{-15}$; flanks: $\rho = 0.087$, $P = 4.28 \times 10^{-4}$ Spearman's rank correlations).

With evidence arguing the enrichment of ESEs near exon ends is indicative of functionality in CDS (Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006; Parmley et al. 2007; Ke et al. 2011; Sterne-Weiler et al. 2011; Caceres and Hurst 2013; Ramalho et al. 2013; Savisaar and Hurst 2018) and lincRNAs (Schuler et al. 2014; Haerty and Ponting 2015), this is consistent with stop codon containing ESEs

likely being functional and therefore the interpretation of low SCD is the need to preserve ESEs, in which stop codons are depleted.



Supplementary Texts Figure 2: The log10 median lengths for introns versus the density of INT3 ESE motifs for the whole exon when (A) all ESEs are combined and (B) ESEs are grouped as stop codon containing and those containing no stop codons. In all cases, correlations between flanking intron size and densities are significantly positively correlated (Spearman's rank correlations). Correlations of densities in whole exons rather than exon flanks are shown as ESE location is not considered as such an important predictor of evolutionary rate in lincRNA (Schuler et al. 2014).

Supplementary Text 13

The high purine content of ESEs is consistent with a model in which ESEs are highly non-intronic, making the depletion of purine-rich stop codons particularly noteworthy

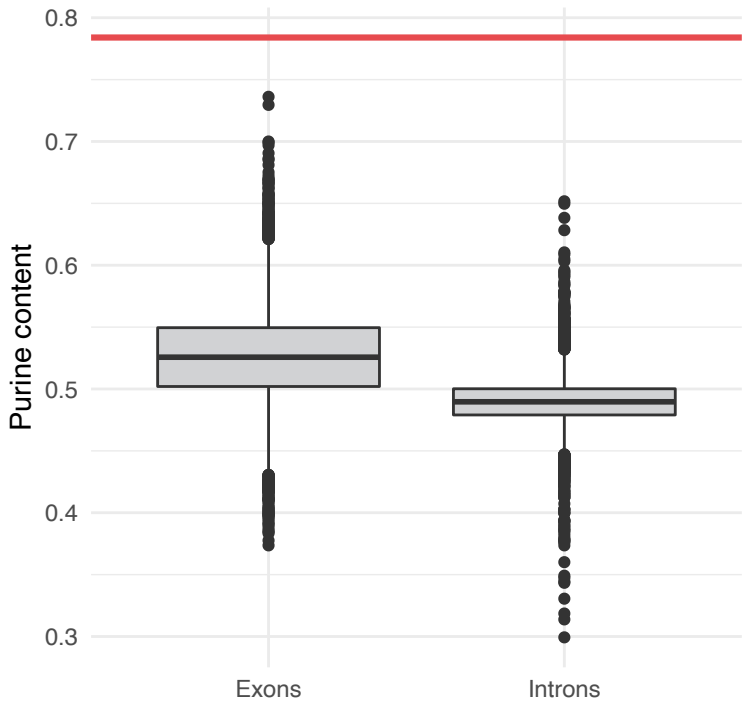
That both ESEs (Xu et al. 1993; Dirksen et al. 1994; Tanaka et al. 1994; Gersappe and Pintel 1999; Fairbrother et al. 2002; Caceres and Hurst 2013) and stop codons are purine-rich makes the depletion of stop codons in ESEs more noteworthy. With skewed nucleotide usage at exon ends in both protein-coding and lincRNA sequences, is the high purine content of ESEs in itself a defining feature of ESEs? The presence of uracil residues within polypurine sequences has been shown to reduce splicing ability (Tanaka et al. 1994). Thus, purine-richness may simply be a requirement of SR protein binding - the binding sites of the SR protein SF2/ASF, for example, are 80% purine (Graveley 2000).

As vertebrate genes are characterised by short exons dispersed between longer introns (Zhang 1998; Sakharkar et al. 2005), the more distinguished a motif is within its surrounding pre-mRNA transcript sequence, the less erroneous binding to intronic sequence should occur. An alternative, but not necessarily mutually exclusive model, is therefore one in which the purine content helps to distinguish ESEs. This model makes several basic predictions. First, exonic sequences should differ in purine content to introns. In other words, if purine content helps to define ESEs as exonic, in the first instance one would expect that exons themselves have increased purine content when compared with intronic sequence. Second, random intronic motifs should have reduced purine content compared with the real motifs, minimising the chance of inappropriate SR protein binding to introns.

We therefore calculated the purine content for coding exons and their corresponding introns, with genes considered as part of the same paralogous family as a single data point (N = 5,620). As expected, the purine content for exons (median purine content = 0.523) is significantly greater than that of introns (median purine content = 0.490) ($P < 2.2 \times 10^{-16}$, paired Wilcoxon signed-rank test). In general, exons are more purine-

rich than introns (Supplementary Texts Figure 3). This differentiation suggests that by being purine-rich, ESEs can be seen as non-intronic sequence. However, this increased exonic purine content may simply be a consequence of the ESEs situated within the exons. After removing all possible motifs using the combined ESE set exonic purine content remains significantly greater ($P < 2.2 \times 10^{-16}$, paired Wilcoxon signed-rank test, median exon purine content = 0.456, median intron purine content = 0.438), suggesting even without ESEs exons are different in terms of purine content.

Are then, ESEs also differentiated further from exons because of their purine content? Using the terminal 50 nucleotides of exon sequences, Caceres and Hurst (2013) find ESEs to have significantly higher purine content. This result holds in our dataset when employing exons longer than 100 nucleotides (N = 5,032 data points) ($P < 2.2 \times 10^{-16}$, paired Wilcoxon signed-rank test). However, this differentiation should not be limited simply to exon ends - a difference in purine content should extend throughout the whole exon as ESEs need to differentiate from surrounding exonic sequence. Using the INT3 set of ESEs, we find that the purine content of sequence that overlaps an ESE is significantly greater than that of non-ESE exonic sequence ($P < 2.2 \times 10^{-16}$, paired Wilcoxon signed-rank test). Thus, whilst the purine content of exons tends to be greater than that of introns, ESEs are further differentiated from the surrounding sequence. This result is therefore consistent with the ESE purine content defining ESEs as not only non-intronic, but highly non-intronic and may therefore act as a marker to differentiate key binding sites.



Supplementary Texts Figure 3: The purine content of both exon and introns, with the purine content of INT3 ESEs shown by the horizontal line. Exons in general tend to have higher purine than introns. The purine content of ESEs differentiates them from both surrounding exonic and intronic sequence.

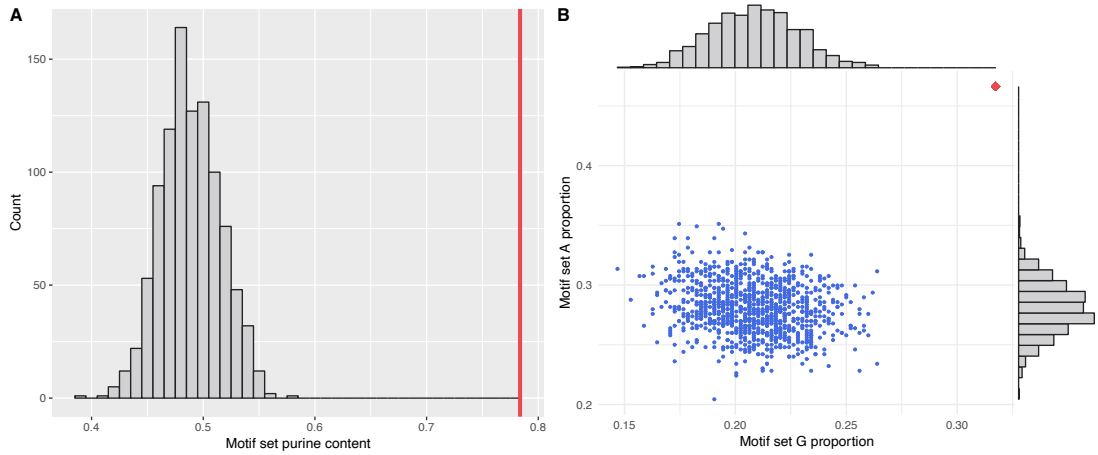
This purine-richness should therefore make ESEs all the rarer within the intronic sequence, possibly such that ESEs are not frequently found within introns to prevent inappropriate binding. This argument is logical – ESEs, as motifs functioning within exons, should be less abundant in intronic sequence. However, it has been documented that ESEs may have functional roles in introns; increases in ESE density have been documented in introns with weak donor sites, suggesting ESEs that help to splice weak donor sites may exist in introns rather than exons (Wu et al. 2005), whilst ESEs in introns have also been shown to have repressor abilities (Kanopka et al. 1996; McNally and McNally 1998). Does then, the purine content of ESEs differentiate them from random intronic motifs for the probability of an SR protein inappropriately binding within an intron to be reduced?

We generated 1,000 sets of random hexamers from intronic sequence, picking one gene at random from those considered as part of a paralogous family. We then asked

whether the purine content of these random motifs differed significantly from the real ESE motifs. We find this is to be the case ($P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-value) – no set of hexamers has purine content close to that of ESEs (Supplementary Texts Figure 4A). Thus, hexamers with the nucleotide content found within ESEs are highly unlikely to be frequently found within introns, making them ideal candidates to ensure the SR proteins correctly locate and bind exclusively exonic sequence in proximity to splice sites.

Given the above result, it is interesting to ask whether there is a particular bias of A/G nucleotides within ESEs that discriminates them from intronic hexamers. We therefore calculated the proportions of each nucleotide in ESEs and the random hexamers derived from intronic sequence. We find a striking difference in both A and G content when compared with these motif sets (Supplementary Texts Figure 4B), finding no sets of random intronic hexamers with either A or G content greater than that of the real ESEs ($P \approx 9.99 \times 10^{-4}$, one-tailed empirical P-values). Thus, it is the combination of both A and G nucleotides found less frequently within introns that differentiates ESEs from intronic sequence.

By having high purine content, ESEs therefore differ significantly from both intronic and exonic sequence, making it less likely SR proteins will bind off-target motifs. Purine content may then act to differentiate such motifs. Despite this, ESEs may have functional repressive roles if present in introns (Kanopka et al. 1996) and may complicate this issue. However, this model still predicts that the purine content would help differentiate such motifs from surrounding intronic sequence and therefore still preventing inappropriate binding. We make no further attempts to address this issue, but simply conclude that the purine content is consistent with helping ESEs look different from the surrounding sequence.



Supplementary Texts Figure 4: (A) The purine content of INT3 ESEs (vertical red line) is significantly greater than the purine content of random sets of intronic hexamers of equal number. (B) Both the A and G content of ESEs is greater than that of the random intronic hexamer sets. We find no set with both A and G content higher than that of the real INT3 ESE set.

References

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25:1915-1927.

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* 62:89-98.

Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.

Dirksen WP, Hampson RK, Sun Q, Rottman FM. 1994. A purine-rich exon sequence enhances alternative splicing of bovine growth hormone pre-mRNA. *J. Biol. Chem.* 269:6431-6436.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007-1013.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.* 32:W187-190.

Gersappe A, Pintel DJ. 1999. CA- and purine-rich elements form a novel bipartite exon enhancer which governs inclusion of the minute virus of mice NS2-specific exon in both singly and doubly spliced mRNAs. *Mol. Cell. Biol.* 19:364-375.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* 6:1197-1211.

Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* 21:333-346.
Hoffman BE, Lis JT. 2000. Pre-mRNA splicing by the essential Drosophila protein B52: tissue and target specificity. *Mol. Cell. Biol.* 20:181-186.
Kanopka A, Muhlemann O, Akusjarvi G. 1996. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* 381:535-538.
Kawano T, Fujita M, Sakamoto H. 2000. Unique and redundant functions of SR proteins, a conserved family of splicing factors, in Caenorhabditis elegans development. *Mech Dev* 95:67-76.
Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360-1374.
Kim S, Shi H, Lee DK, Lis JT. 2003. Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nucleic Acids Res.* 31:1955-1961.
Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Perez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* 49:1731-1740.
Longman D, Johnstone IL, Caceres JF. 2000. Functional characterization of SR and SR-related genes in Caenorhabditis elegans. *EMBO J.* 19:1625-1637.
McNally LM, McNally MT. 1998. An RNA splicing enhancer-like sequence is a component of a splicing inhibitor element from Rous sarcoma virus. *Mol. Cell. Biol.* 18:3103-3111.
Niazi F, Valadkhan S. 2012. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA* 18:825-843.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* 23:301-309.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.
Ramalho RF, Gelfman S, de Souza JE, Ast G, de Souza SJ, Meyer D. 2013. Testing for natural selection in human exonic splicing regulators associated with evolutionary rate shifts. *J. Mol. Evol.* 76:228-239.
Ring HZ, Lis JT. 1994. The SR protein B52/SRp55 is essential for Drosophila development. *Mol. Cell. Biol.* 14:7499-7506.
Sakharkar MK, Perumal BS, Sakharkar KR, Kangueane P. 2005. An analysis on gene architecture in human and mouse genomes. *In silico biology* 5:347-365.
Savisaar R, Hurst LD. 2016. Purifying Selection on Exonic Splice Enhancers in Intronless Genes. *Mol. Biol. Evol.* 33:1396-1418.
Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* 28:1442-1454.
Schuler A, Ghanbarian AT, Hurst LD. 2014. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* 31:3164-3183.
Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* 21:1563-1571.
Tanaka K, Watakabe A, Shimura Y. 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol. Cell. Biol.* 14:1347-1354.

Wu X, 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.* 32:1847-1861.
Wu Y, Zhang Y, Zhang J. 2005. Distribution of exonic splicing enhancer elements in human genes. *Genomics* 86:329-336.
Xu R, Teng J, Cooper TA. 1993. The cardiac troponin T alternative exon contains a novel purine-rich positive splicing element. *Mol. Cell. Biol.* 13:3660-3674.
Zhang MQ. 1998. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* 7:919-932.

Chapter 3

Rarity of stop codons in exonic motifs cause nonsense mutations to disrupt splicing in disease and non-disease genes

This chapter contains a reformatted version of the draft manuscript from a collaboration primarily between Rosina Savisaar and myself. The manuscript was previously submitted to Genome Biology and I am due to submit a version of this revised manuscript for an invited second round of peer-review imminently.

Rosina Savisaar and myself reviewed the nonsense-associated altered splicing literature, determined the splice-quantification analysis specifics, implemented the pipeline and analysed the data. I implemented and analysed the disease-associated data. I have also collaborated with Christine Mordstein, Bethan Young and Grzegorz Kudla at the MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, who performed the experimental analysis. Rosina Savisaar presented a preliminary account of these results in her thesis. Since then, I have reanalysed the data, updated results and performed new analyses in order to produce the manuscript here.

This chapter contains analysis of publicly available data. The data and custom scripts are freely available at locations cited within the paper.

This declaration concerns the article entitled:			
Rarity of stop codons in exonic motifs cause nonsense mutations to disrupt splicing in disease and non-disease genes			
Publication status (tick one) **Draft manuscript** [X] **Submitted** [ ] **In review** [ ] **Accepted** [ ] **Published** [ ]			
Publication details (reference)	N/A		
Copyright status (tick the appropriate statement) I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [ ]			
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: *Review of NAS literature - 50%* Design of methodology: *Splice quantification pipeline design - 50%* *Splice quantification pipeline implementation and analysis - 50%* *Disease analysis pipeline design – 100%* *Disease analysis implementation and analysis – 100%* *TPM expression analysis – 0%* Experimental work: *Liaison with collaborators about experiment design specifics – 100%* *Experimental analysis – 0%* 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**	

Abstract

Background

It is often assumed that transcripts containing premature termination/stop codons (PTCs) are degraded by nonsense-mediated decay (NMD) or produce truncated proteins. Nonsense-associated altered splicing (NAS), shown in a few genes and mechanistically unresolved, is however, a further possibility. Here we provide a genome-wide estimate for NAS prevalence in non-disease-associated and disease-associated contexts and test predictions discriminating between two NAS mechanistic models, motif disruption and scanning.

Results

Using 1000 Genomes project data with associated RNA-seq data, we identify genome-wide associations between PTCs and exon skipping, with 30 prime candidates, conservatively estimating that ≈6% of nonsense mutations disrupt splicing in non-disease-associated contexts. We experimentally validate our top NAS candidate. Disease-associated nonsense mutations are *in silico* predicted to commonly disrupt splicing and are enriched at exon ends (where the density of splicing information is highest). From such enrichment, we estimate ≈33% of disease-associated nonsense mutations may affect splicing. That disease-associated nonsense mutations disproportionately hit exonic splice enhancer (ESE) motifs and that out of frame stop codons also disrupt splicing, supports the motif-disruption NAS model.

Conclusions

Genome-wide NAS is rare in non-disease-associated contexts but likely common in disease-associated ones. NAS mostly likely occurs owing to splice motif disruption. Indeed, the results accord with a model in which, given that within ESEs stop codons are heavily depleted, mutations to nonsense occurring in ESEs are especially likely to disrupt splicing. The realization that NAS may underpin many instances of nonsense-mediated pathogenesis has implications for genetics-based therapeutics.

Background

A fundamental component of genetic-based medicine is correctly understanding the molecular mechanisms that underpin genetic diseases (e.g. see (Price et al. 2015; Ginsburg and Phillips 2018; Jackson et al. 2018)). Here we consider how nonsense mutations generating premature termination codons (PTCs) - in-frame stop codons in the mature transcript – might elicit their deleterious effects. PTCs are often implicated in disease (Holbrook et al. 2004b; Mort et al. 2008), with approximately 11.5% of all described mutations causing human inherited diseases being the result of nonsense mutations (Mort et al. 2008).

PTC pathogenicity is often ascribed to one of two well-described mechanisms. First, a PTC results in the synthesis of a truncated protein with potentially problematic loss of function or gain of toxicity (Holbrook et al. 2004a; Drummond and Wilke 2008; Karam et al. 2008; Chung et al. 2018). Alternatively, eukaryotes also possess nonsense-mediated decay (NMD) (Maquat 2005; Brogna and Wen 2009), a system that recognises and targets for degradation (during translation) some PTC-containing transcripts (possibly owing to such toxicity). Such PTCs targeted by NMD may be mutational in origin or result from transcriptional errors. The importance of NMD in dealing with PTC-containing transcripts is highlighted by an evolved robustness in cases where NMD cannot function. For example, the use of particular codons minimises PTCs resulting from mistranscription events in intronless transcripts, in the last exon or less than 50 nucleotides from the last EJC (Cusack et al. 2011) as these are all largely hidden from NMD (Maquat and Li 2001; Brocke et al. 2002) and hence cannot be degraded.

There is, however, at least one further possibility, nonsense-associated altered splicing (NAS). PTCs have been observed to alter splicing patterns where the PTC-containing exons have been spliced out. The PTC in question is hence not subject to NMD and instead unexpected splice isoforms are produced (Gibson et al. 1993; Dietz and Kendzior 1994; Hull et al. 1994; Endo et al. 1995; Messiaen et al. 1997; Shiga et al. 1997; Valentine and Heflich 1997; Hoffmeyer et al. 1998; Mazoyer et al. 1998; Melis et al. 1998; Valentine 1998; Gersappe and Pintel 1999; Ars et al. 2000; Wimmer et al.

2000; Di Blasi et al. 2001; Caputi et al. 2002; Li et al. 2002; Wang et al. 2002a; Wang et al. 2002c; Pagani et al. 2003; Pasmooij et al. 2004; Vuoristo et al. 2004; Zatkova et al. 2004; Mendive et al. 2005; Stasia et al. 2005; Disset et al. 2006; Aznarez et al. 2007; Chang et al. 2007; Laimer et al. 2008; Sperling and Sperling 2008; Chemin et al. 2010; Littink et al. 2010; Lenassi et al. 2014; Peterlongo et al. 2015; Barny et al. 2018; Meldau et al. 2018). Skipping of exons has been associated with pathogenicity (Shiga et al. 1997; Lorson et al. 1999; Moseley et al. 2002; Helderman-van den Enden et al. 2010; Xu et al. 2014), most likely due to the deletion of important peptide or structural information and so NAS is likely to be a source of pathogenicity. Exons skipped as a result of NAS also have the added potential consequence of introducing downstream PTCs if the exon is not of length three. This can result in the truncation of the exon-skipped proteins or targeting the resultant transcript for NMD.

To date, despite NAS being a potential source of pathogenicity, no genome-wide study of the prevalence of NAS has been performed in either a healthy or pathogenic context. Instead, to date studies reporting the effects of NAS are performed on a case by case basis (see previous references for examples). Here we aim to provide the first such genome-wide estimates. We consider both the non-disease-associated context (via 1000 Genomes data (The 1000 Genomes Project Consortium 2015)) and the disease-associated context (via ClinVar data (Landrum et al. 2018)). We expect the two to provide different estimates. As splice disruption is expected to be highly damaging, we expect that the frequency of nonsense mutations resulting in NAS to be higher in disease-associated contexts than in non-disease-associated contexts (i.e. nonsense mutations circulating in the human population) as splice disruption can be highly deleterious (Lopez-Bigas et al. 2005; Baralle et al. 2009; Lim et al. 2011; Sterne-Weiler et al. 2011; Wu and Hurst 2016).

The mechanism of NAS also remains unsolved. Two models have been proposed (reviewed in Cartegni et al. 2002; Maquat 2002). The first “motif disruption” model suggests that a nonsense mutation could disrupt important regulatory splice motifs such as exonic splice enhancers (ESEs). ESEs are especially abundant at exons ends (the terminal ≈70bp) and function by binding serine-arginine rich (SR) proteins that in turn direct the splicing machinery to the splice junction and facilitate the assembly of

the spliceosome (Blencowe 2000). ESEs and their disruption could then result in incorrect splicing of the transcript (Shiga et al. 1997; Valentine 1998; Liu et al. 2001; Caputi et al. 2002; Pagani et al. 2003; Zatkova et al. 2004; Aznarez et al. 2007; Peterlongo et al. 2015). Alternative motif centred mechanisms are imaginable, such as the mutation creating an exonic splicing silencer (ESS), *cis*-regulatory elements that inhibit the use of adjacent splice sites (Wang et al. 2004).

The motif disruption model has recently been rendered especially parsimonious by the observation that, as a consequence of overlapping with coding sequence (CDS), exonic splice motifs, including ESEs, have an especially low density of stop codons (Abrahams and Hurst in press), despite purine enrichment of both ESEs and stop codons. Given the overall rarity of the trinucleotides TAA, TAG and TGA in ESEs, any mutation creating a stop codon in an ESE may therefore inhibit the binding ability of the motif with potential implications for splicing of that exon. Furthermore, given the prevalence and selective pressures imposed upon and by ESEs (Savisaar and Hurst 2017, 2018) (the proportion of exonic sequence governing splicing moderated ESE selection is as strong a predictor of human protein evolution as the amount a gene is expressed (Parmley et al. 2007)), it would be *a priori* expected that NAS is likely to occur at a non-negligible level genome-wide. Furthermore, at least some of the time, this disruption would be expected to be associated with motif disruption.

The second "nuclear scanning" model (Dietz and Kendzior 1994; Gersappe and Pintel 1999; Mendell and Dietz 2001; Li et al. 2002; Wang et al. 2002a; Wang et al. 2002c; Shi et al. 2015) instead requires the detection of the PTC via a translation-like scanning mechanism. In this model, if a PTC is detected, information is fed back to modulate alternative splicing of transcripts that are subsequently transcribed from the same locus and upregulate the synthesis of transcripts skipping the PTC-containing exon. A defined start codon and by implication reading frame are prerequisites (Shi et al. 2015). Whether the disruption to the ORF is recognised in the cytoplasm or nucleus is unclear (Mendell et al. 2002; Wang et al. 2002c; Chang et al. 2007), with some splice variants also suppressed in a process decoupled from NMD (Pan et al. 2006). In the case of nonsense mutations in exon 51 of the *FBN1* gene, all of the three nonsense variants disrupt splicing which, in turn, is restored by introducing upstream frameshifts of the nonsense variant (Dietz and Kendzior 1994). Such a scanning mechanism is not likely

to operate directly on the mutated transcript itself, but by instead passing information to the site of transcription whereby splicing can be modified for subsequent transcripts.

The two models make different predictions about out of frame stop codon mutations. The scanning mechanism predicts only in-frame stops in CDS elicit a response, while the motif disruption model predicts that the mutations generating the trinucleotides TAA, TGA or TGA in any frame in CDS could have an effect if they disrupt splicing motifs. We therefore test whether out of frame stop codons also initiate splice disruption.

We begin by asking whether it is possible to detect genome-wide associations between PTCs and exon skipping. Exploiting the publicly available polymorphism data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) and associated publicly available RNA-seq data from the Geuvadis project (Lappalainen et al. 2013), we identify a set of 30 prime candidate PTCs associated with large increases in exon skipping consistent with NAS. Accordingly, we estimate that approximately 6% of polymorphic nonsense mutations have a non-negligible effect on splicing. We provide experimental data for our computationally derived top NAS candidate using a minigene construct. Evidence is consistent with NAS as the source of exon skipping. Given we also find an unexpected splice variant with partial intron retention, we argue that the most parsimonious model is one evoking the disruption of regulatory motifs. Similarly, we find that out of frame stop codons have similar effects to in-frame stop codons, an effect not predicted by the nuclear scanning model.

That both *in silico* and experiment data are consistent with the motif disruption model enables us to estimate the prevalence of disease-associated nonsense mutations associated with disrupted splicing. Specifically, we ask whether known disease-associated PTCs (irrespective of molecular function) are more prevalent towards the flanks of exons. Exon flanks (≈70 bp) are regions rich in splice controlling elements such that known splice disrupting mutations are reported to be especially abundant at exon ends (Woolfe et al. 2010). We find that disease-associated PTCs do preferentially locate in exon flanks. From the degree of enrichment, we estimate that ≈33% of disease-associated nonsense mutations might affect splicing. In addition, disease-

associated PTCs disrupt ESEs significantly more frequently than expected by chance. This enrichment in ESEs further supports the motif disruption model for NAS and underpins the role of splice disruption in pathogenicity. Taken together, the results accord with a model in which, given that stop codons are heavily depleted in ESEs, nonsense mutations in ESEs are especially likely to disrupt splicing. This insight has potential medical implications as it highlights the importance of accurately classifying mutations when trying to ensure the effectiveness of genetic and drug therapeutic strategies.

Results

PTCs associated with increased exon skipping have a ≈107-fold stronger effect than those associated with increased exon inclusion

To obtain an estimate for the rate of genome-wide PTC-associated exon skipping, we employed DNA polymorphism data obtained from the 1000 Genomes project (The 1000 Genomes Project Consortium 2015) and associated RNA-seq data for 462 individuals from the Geuvadis RNA-sequencing project (Lappalainen et al. 2013) for matching individuals. This dataset was chosen for two reasons. First, it contains data from enough individuals to uncover any potential associations. Second, and perhaps more importantly, it is readily publicly available.

A high-quality set of internal coding exons was assembled. Of these exons, we retained 541 PTC-containing exons in which we could quantify splicing (see Methods). For each exon, the median percentage spliced in (PSI) was calculated (see Methods) for each of the three genotypes - homozygous non-PTC (PTC-/-), heterozygous PTC (PTC-/+) and homozygous PTC (PTC+/+). Our analyses focus on comparisons between PTC-/- and PTC-/+ variants as only 24/541 (4.37%) exons had an individual with a PTC +/+ variant (and of which only 14 have a change in PSI from the PTC-/- variant).

We first asked whether there are detectable differences in exon inclusion for the same exon whether containing or not containing a PTC. If PTCs are responsible for exon skipping, we expect the median PSI for PTC-/+ variants to be significantly lower for than for PTC-/- variants. We do find a significant difference between two genotypes ($P = 3.294 \times 10^{-4}$, paired Wilcoxon signed-rank test), although unexpectedly it is the PTC-/- variants that are significantly reduced ($P = 1.647 \times 10^{-4}$, one-tailed paired Wilcoxon signed-rank test). If anything, this result argues against an increase in exon skipping associated with PTCs.

Why might the PTC-/+ variants have significantly higher PSI? We introduce the term ΔPSI ($PSI_{PTC-/+}$ - $PSI_{PTC-/-}$) that describes the PSI difference between genotypes for each exon. If there is less exon inclusion when the PTC is present, ΔPSI is negative. We find that almost half of exons exhibit no difference in PSI between genotypes (ΔPSI = 0, N = 239, 44.18%). In many cases, the presence of a PTC appears to have no effect on exon inclusion. Of the remaining exons where ΔPSI is not equal to zero (N = 302), only 68 have ΔPSI < 0 (12.57%) with reduced exon inclusion for the PTC variant. For the majority of cases (N = 234), the PTC containing variant therefore actually has greater exon inclusion and may explain the previous result. This contradicts a hypothesis of exon skipping being a common consequence of nonsense mutations.

However, upon closer inspection both the median (-9.023) and standard deviation (19.926) ΔPSI scores for exons with ΔPSI < 0 are further from zero than ΔPSI > 0 exons (median = 0.084, standard deviation = 2.442 respectively) (Figure 6A). The median absolute effect on exon inclusion is ≈107-fold greater when the PTC increases exon skipping than when it is associated with inclusion. Using the absolute ΔPSI values for the two groupings ΔPSI > 0 and ΔPSI < 0, we find the absolute ΔPSI < 0 scores are significantly greater than the absolute ΔPSI > 0 scores ($P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon rank sum test). That the initial test demonstrates that the opposite effect appears to be an side effect of the statistical test wherein many variants with small and likely negligible effects mask the less frequent, but larger and more disruptive differences in the direction consistent with NAS (see Supplementary Figure 1A, Supplementary Figures). When considering absolute quantitative effect of PTCs

on exon inclusion, those variants associated with increased skipping and consistent with NAS have a significantly greater impact than those increasing exon inclusion.

PTCs are associated with significant increases in exon skipping after controlling for the nucleotide composition of mutations that generate PTCs

While PTCs appear to have large effects consistent with NAS, it is also the case that the relative proportions of the different mutation classes generating PTCs (e.g. A→T, C→T) are significantly different than for mutations creating synonymous and nonsynonymous variants in the same set of exons (N = 1,458 non-nonsense mutations, χ^2 = 481.192, P = 5.680 × 10^{-104}, chi-squared test, Supplementary Spreadsheet 1, Supplementary Spreadsheets). This is robust to excluding N→C mutations (N = 216) that cannot generate a PTC (χ^2 = 329.757, P = 2.480 × 10^{-72}, chi-squared test, Supplementary Spreadsheet 1, Supplementary Spreadsheets). The strength of the effect in the previous result may not be a result of the PTC per se, but more general of nucleotides involved in the mutations that generate PTCs.

To account for any such nucleotide biases, we asked whether the PTCs are associated with increased exon skipping beyond that expected given their nucleotide composition. To do this, for each PTC we simulated 100 pseudo-PTCs (pPTCs) by replacing each real PTC with a randomly sampled missense mutation matched by ancestral allele, variant allele and variant allele frequency (e.g. if the total PTC count on both alleles was 6/300, the matched mutation allele frequency was ≈ 0.2). To quantify any difference, we calculated a Z score for each PTC, defined as the ΔPSI for the true PTC minus the mean of pseudo-ΔPSIs (ΔpPSI), divided by the standard deviation of ΔpPSI. Thus, if real PTCs have a more negative effect on PSI than the matched pPTCs as a result of being a PTC and not the nucleotides involved, Z scores will be negative. Indeed, we find 308/541 (56.93%) PTCs with a Z score less than zero, a significant proportion (P = 7.213× 10^{-4}, one-tailed exact binomial test, Figure 1B). Results are more consistent with increased skipping of the PTC variant when the distance of the pPTC to the exon boundary is also controlled (by sampling a matched mutation from within the 10 bp window around true nonsense variant location) (342/541 exons with Z < 0, P = 4.125 × 10^{-10}, one-tailed exact binomial test). PTCs

therefore have a significantly stronger tendency to promote exon skipping than expected given their nucleotide composition.

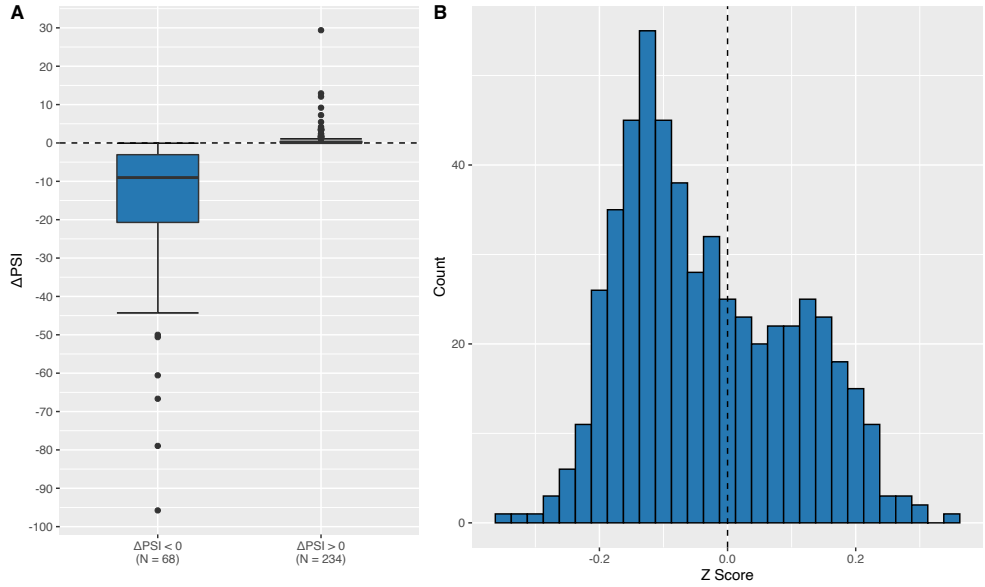


Figure 6: Differences in relative exon skipping levels.

(A) The differences in PSI scores between PTC-/+ and PTC-/- variants (ΔPSI) for exons with non-zero differences in PSI between the two genotypes. ΔPSI scores corresponding to exons for which the PTC is associated with increased exon inclusion (ΔPSI > 0) are typically small. ΔPSI scores for variants associated with increased exon skipping (ΔPSI < 0) and consistent with NAS are typically have a much larger effect. (B) Z scores comparing the ΔPSI score for each PTC with ΔPSI scores for 100 missense mutations matched by ancestral allele, variant allele, allele frequency and distance to exon boundary. The distribution of Z scores is skewed to the left of zero, indicating that the variants increase exon skipping more than expected when compared with the matched simulants.

The first result associating PTCs with increased exon inclusion is therefore highly misleading. In many cases, the effect of a PTC on exon inclusion is small and unlikely to be biologically significant. However, for cases with more substantial variations in ΔPSI, they are consistent with the PTC presence increasing exon skipping, with these much more likely to be biologically meaningful, and not a result of nucleotide biases that may promote greater skipping. Taken together, these results indicate there does

exist a direct association between the presence of a PTC and significant increases in relative levels of exon skipping beyond that expected, consistent with NAS.

NMD cannot account for many cases of increased exon skipping associated with a PTC

Despite the above results, it remains that NMD could explain these differences. By definition, the PSI metric is dependent on the number of reads for the full-length isoform. Yet, if full-length transcripts containing a PTC are removed at a particular rate by NMD, the relative proportion of reads demonstrating exon inclusion (PSI) would decrease despite no increase in the absolute number of skipped reads (see Supplementary Spreadsheet 2, Supplementary Spreadsheets for an example). Thus, although we might observe a relative change in the rate of exon skipping between variants, there might be no absolute change in the number of exons skipped.

It is therefore vital to eliminate any PSI variations we observe that are also consistent with NMD. To do this, we used absolute read counts supporting exon skipping or inclusion, normalised to the number of total reads per million to control for differing read depths between samples. We introduce the metrics reads per million included (RPMinclude) and reads per million skipped (RPMskip) that define read counts supporting exon inclusion or skipping, respectively. Accordingly, ΔRPMincl ($RPMinclude_{PTC-/+}$ - $RPMinclude_{PTC-/-}$) and ΔRPMskip ($RPMskip_{PTC-/+}$ - $RPMskip_{PTC-/-}$) then describe the differences between PTC-/+ and PTC-/- variants for RPMinclude and RPMskip, respectively.

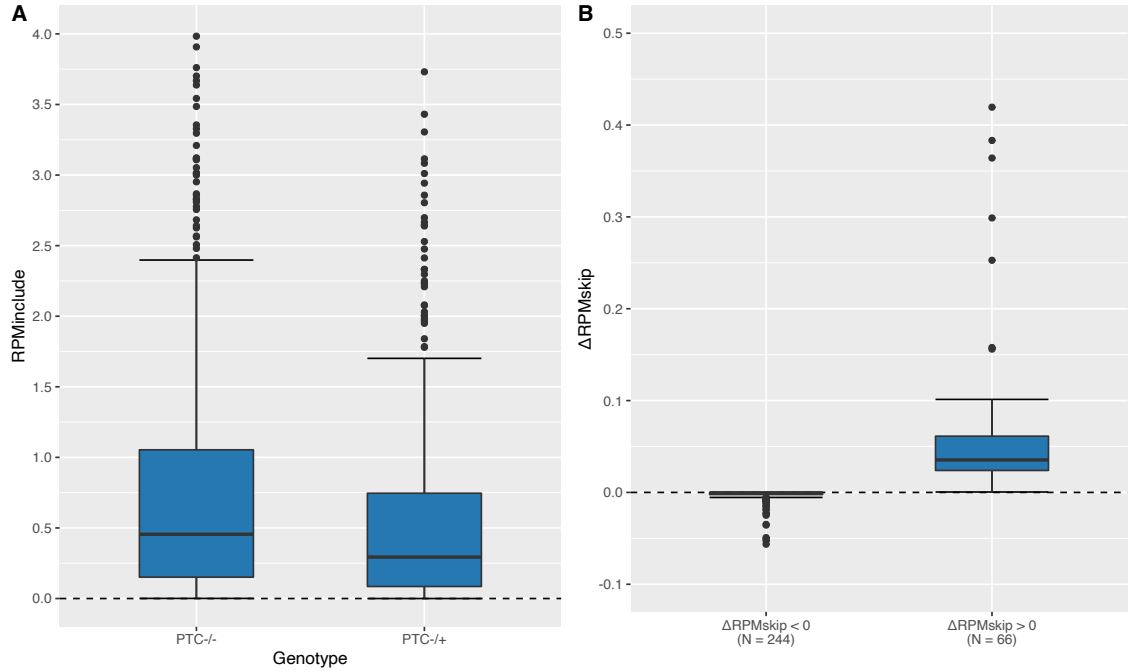


Figure 7: Differences in absolute exon skipping levels.

(A) Raw read counts per million reads supporting exon inclusion (RPMinclude) for non-PTC and PTC variants. 50 outlier data points are removed for visualisation purposes. (B) Differences in the raw read counts supporting exon skipping between variants (ΔRPMskip) are consistent with those in the direction consistent with NAS having a larger and biologically relevant effect. The median negative ΔRPMskip is -4.794×10^{-4} arguing that when the PTC is associated with reduced exon skipping, the effect is almost negligible. One data point for ΔRPMskip < 0 at y = -0.759 and two outlier data points at for ΔRPMskip > 0 at y = 1.242, y = 4.778 are omitted for visualisation purposes.

We first asked whether we could detect NMD. If so, the raw number of reads supporting inclusion, RPMinclude, should be significantly higher in PTC-/- than for PTC-/+ variants, i.e. ΔRPMincl < 0 as full-length transcripts containing a PTC are subject to NMD at some rate. Our results suggest this is the case ($P < 2.2 \times 10^{-16}$, one-tailed paired Wilcoxon signed-rank test, Figure 7A), with the median RPMinclude for PTC-/+ variants (0.328) almost one third less than the median RPMinclude for PTC-/- variants (0.502). Thus, NMD is detectable in our samples.

We next asked whether the raw reads provide evidence consistent with NAS. If PTCs are associated with exon skipping, then RPMskip should be greater for PTC-/+ than

for PTC-/- variants. However, as with PSI, PTC-/- variants have significantly higher RPMskip values ($P = 4.829 \times 10^{-5}$, one-tailed paired Wilcoxon signed-rank test). As per PSI, many cases have ΔRPMskip = 0 (N = 239). Further, relatively few exons have ΔRPMskip scores greater than 0 (N = 64 with increased skipping for the PTC-/+ variant) (Supplementary Figure 1B, Supplementary Figures). However, for the positive ΔRPMskip scores both the median (0.045) and standard deviation (0.627) suggest a distribution further from zero than those in the opposite direction (median = 4.667×10^{-4}, deviation = 0.050) (Figure 7B). Again, as per PSI, absolute ΔRPMskip values in the direction consistent with NAS are significantly greater than absolute values of those in the opposite direction ($P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon rank sum test). The likely meaningful associations are therefore consistent with NAS.

Is the number of PTCs with raw read counts supporting greater skipping for the PTC variant also higher than expected given the nucleotide composition of PTC mutations? We reanalysed the set of 100 matched missense simulants and asked how many PTCs differ in ΔRPMskip when compared with simulant pseudo-ΔRPMskip (ΔpRPMskip) values. A significant number, 339/541 (62.66%), have positive Z scores (P = 0.004, one-tailed exact binomial test; note here a positive Z score indicates increases in RPMskip over the simulants). This result is robust to missense mutations being matched by their distance to the exon boundary (381/557, $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test).

These results are therefore consistent with an association between the PTC and an increase in the absolute read count supporting exon skipping beyond that attributable to NMD. While we cannot discount the contribution of NMD entirely, these results argue against an overall systematic involvement of NMD increasing skipping and therefore is unlikely to fully explain the associations observed between the PTCs and exon skipping.

6% of nonsense mutations result in non-negligible levels of exon skipping consistent with NAS

The above results provide, to the best of our knowledge, the first evidence of genome-wide associations between PTCs and exon skipping. However, as alluded to, in most cases the effect is minimal (median absolute ΔPSI ≈ 0.001% for all exons with ΔPSI not equal to zero). In many cases, it is therefore likely that skipping of the exon has little phenotypic consequence. For example, for this median effect only 1 of 100,000 transcripts will have an altered splicing pattern. However, for those cases where the effect is substantial and by assumption biologically meaningful, we find both the relative increase in exon-skipped isoforms and absolute change in the numbers of isoforms with a skipped exon are significantly higher for those variants consistent with NAS.

Thus, to provide a more reliable estimate quantifying the genome-wide extent of NAS in this population, we consider only the large-effect cases where differences in exon skipping are likely to be phenotypically important. Using ±5% as our ΔPSI thresholds (see Supplementary Text 1 for justification of the 5% threshold), we find 50/541 (9.24%) exons have ΔPSI exceeding these thresholds (either positively or negatively). Of these, 44/50 (88.00%) exons have ΔPSI < 0, many with a difference that far exceeds 5% (Figure 8A). The direction of this enrichment is highly significant ($P = 1.662 \times 10^{-8}$, one-tailed exact binomial test). Thus, a significant majority of PTCs that result in a substantial effect on exon inclusion are consistent with increased exon skipping.

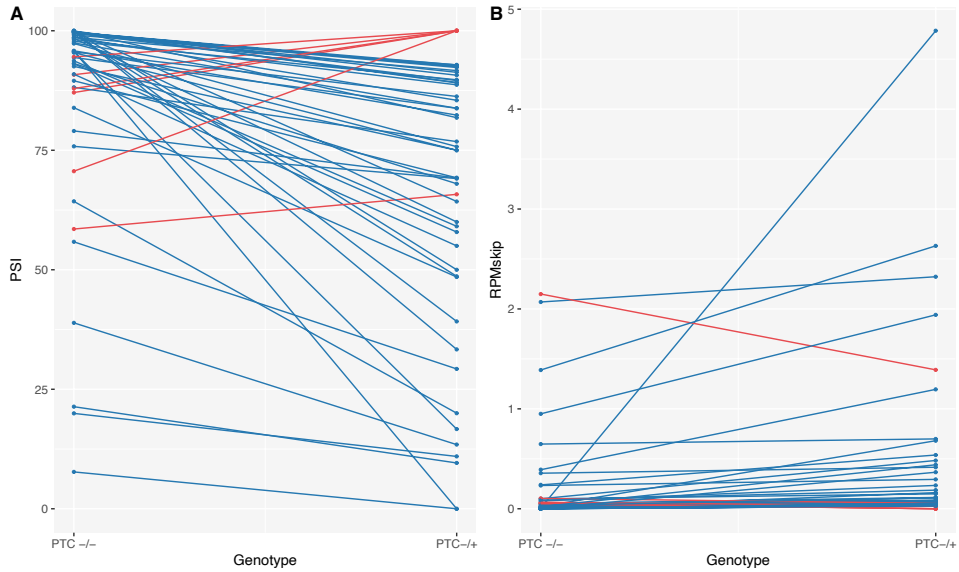


Figure 8: Individual large effect cases for both PSI and RPMskip.

Large differences between PTC-/- and PTC-/+ genotypes for (A) PSI and (B) RPMskip. Variants with changes between genotypes consistent with NAS are in blue, those in the opposite direction in red. For both PSI and RPMskip, the number of large effect variants in the direction consistent with NAS is significantly greater than by expected by chance.

Is this trend similar for RPMskip, and thus attributable to NAS and not NMD? To match the number of large-effect PSI cases (50), we set the RPMskip threshold to 0.026. Of these, we find 43/50 (86.00%) have ΔRPMskip > 0 with increased exon skipping for the PTC-/+ variant (Figure 8B). Again, this proportion is significantly higher than expected by chance ($P = 1.049 \times 10^{-7}$, one-tailed exact binomial test). Thus, large-effect cases for which exon skipping cannot be attributed solely to NMD are consistent with NAS.

However, although the RPMskip metric helps us to understand the contribution of NMD and NAS to exon skipping, PTC-/- and PTC-/+ variants may have genetic differences beyond the presence of a PTC. For example, differing rates of transcript initiation may lead to differing levels of mRNA being produced for the two genotypes. In this scenario, ΔRPMskip may vary because the expression of the PTC-/- isoform is higher and not due to increased exon skipping, downregulation or degradation of the PTC-/+ transcript. PSI does not suffer from this issue because it is normalised to the total number of reads analysed. Thus, although we do not expect such differences to lead to higher expression of the PTC-/+ genotype consistently, it is important to not base inferences solely on RPMskip.

Thus, with caveats to both PSI and RPMskip metrics, the most robust evidence for candidate exons being valid NAS targets with meaningful phenotypic consequences in this dataset are those overlapping both large-effect ΔPSI and large-effect ΔRPMskip groups. 30 exons appear in both groupings in directions consistent with NAS (Table 1). To ask whether this overlap is significant, we performed 10,000 simulations picking 44 and 43 exons from the full set of PTC-containing exons (44 and 43 correspond to the number of exons with large effect sizes for PSI and RPMskip in the correct direction respectively). We find no simulation iteration has an overlap as large

as the real overlap ($P \approx 9.999 \times 10^{-5}$, one-tailed empirical P-value, maximum simulant overlap = 10).

These PTCs and associated exons are therefore prime candidates for cases in which a single nucleotide polymorphism (SNP) generating a PTC causes potentially detrimental exon skipping via NAS. Although we acknowledge that NAS may occur more regularly at lower frequencies, NAS does appear to be a genome-wide phenomenon. We estimate that ≈6% (30/541) of annotated nonsense mutations are likely to have their effect via NAS.

Large-effect PTCs are predicted to have larger increases in exon skipping when compared with the other PTCs

We also predicted changes in PSI for each PTC using MMSplice (Cheng et al. 2019), a neural network model that outperforms other splicing variant scoring models (HAL (Rosenberg et al. 2015), SPANR (Xiong et al. 2015) and the baseline predictor model MaxEntScan (Yeo and Burge 2004)). MMsplice reports the effect of a variant on PSI on the logistic scale ($\Delta\text{logit}\Psi$), with $\Delta\text{logit}\Psi < 0$ indicating a predicted increase in exon skipping associated with the variant. We find 25/30 (83.33%) of our large-effect candidates are predicted to increase skipping in this model, significantly more than expected by chance ($P = 3.249 \times 10^{-4}$, two-tailed exact binomial test). Further, these differences in predicted skipping are significantly greater than for the remaining 511 variants (median large-effect PTC $\Delta\text{logit}\Psi = -0.185$, median other PTC $\Delta\text{logit}\Psi = -0.140$, $P = 0.242$, one-tailed Wilcoxon rank sum test).

Experimental validation of the gene with the largest ΔRPMskip, *ACP1*, confirms exon skipping consistent with NAS

Having computationally identified potential NAS candidates, we sought to validate our results experimentally. A minigene construct for the prime candidate exon from the *ACP1* gene with the greatest ΔRPMskip (ENST00000272065.5, Table 1) was constructed and expressed as described in the Methods (Figure 9A). In HeLa cells, we find a significant difference in PSI between the wildtype (wt) and PTC-containing

constructs ($P = 2.226 \times 10^{-5}$, two sample t-test, Figure 9A and B), with skipping almost exclusively restricted to the PTC-containing construct. Consistent with skipping resulting from NAS and not NMD, this difference in PSI remains after knockdown of the core NMD factor Upf1 ($P = 2.783 \times 10^{-8}$, two sample t-Test, Figure 9A and B). That there is a small but significant difference between the PTC variants both with and without NMD ($P = 0.002$, two sample t-Test) suggests a minor proportion of the exon skipping between the wt and PTC variants can be attributed to NMD for cells in which NMD functions.

We also asked whether RPMskip levels also significantly differ as expected were NAS is the underlying cause. We find an increase in RPMskip with inclusion of the PTC ($P = 1.804 \times 10^{-4}$, two sample t-Test, Figure 9C) and again when NMD is knocked down ($P = 2.741 \times 10^{-4}$, two sample t-Test, Figure 9C), suggesting that the presence of the PTC results in an increase in the absolute number of reads supporting skipping regardless of NMD. Consistent with this notion, RPMskip for PTC variants does not significantly differ between cells where NMD is both present and knocked down ($P = 0.302$, two sample t-test, Figure 4C). We infer that NMD cannot explain the exon skipping associated with the PTC.

To confirm that NMD was depleted and results not subject to any unexpected NMD, levels of Upf1 mRNA were quantified. We find that in cells with the wt and PTC-containing constructs Upf1 mRNA levels are significantly lower in the Upf1 knockdowns ($P < 0.001$, two sample t-Tests, Figure 9D). Further, confirming that the NMD depletion and depleted Upf1 mRNA levels correlate with protein depletion, wt mRNA levels of the T-cell receptor (*TCR*) reporter gene (a well-established NMD reporter gene) are significantly depleted in the Upf1 knockdown when compared with no knockdown ($P < 0.05$), but levels are significantly increased and stabilised when the PTC is present in Upf1 knockdown cells when compared with no knockdown. We conclude that the NMD knockdown was effective.

Interestingly, there exists an unexpected third band for the PTC mutants (Figure 9A, black triangle). Given that its size is greater than the size of the correctly spliced variants, this appears to be a splice variant in which the focal exon is included but with

partial intron retention. As the PTC variants appear to produce two splice variants (one expected and one additional) that the wt variants do not, the PTC seems to be disrupting splicing signals. While we cannot discount subsequent downregulation of these isoforms, to have partial intron retention in this scenario there must first have been a disruption to splicing. The most parsimonious explanation is that the PTC disrupts a splice motif and elicits aberrant splicing. We also note that experimental PSI and RPMskip levels are broadly consistent with our computational PSI calculations for the 1000 Genomes samples. We find similar patterns in a set of Hek293T cells (Supplementary Figure 2) thus demonstrating PTC-associated exon skipping independent of cell type.

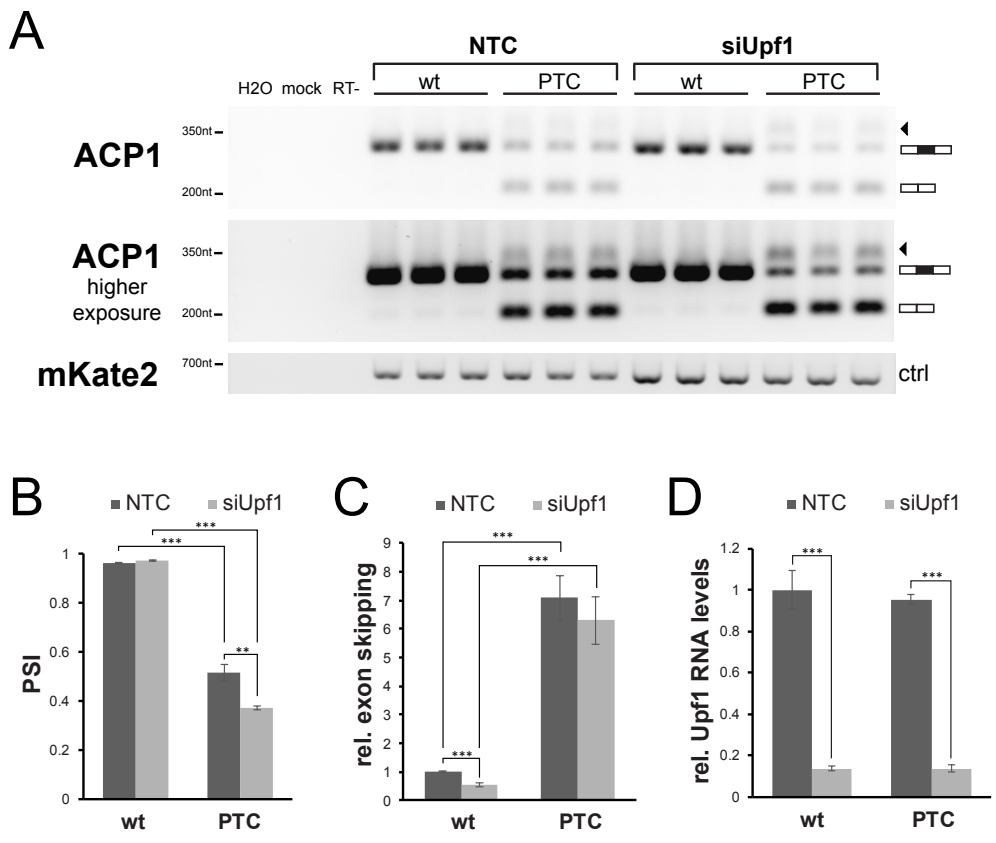


Figure 9: Experimental validation of the top NAS candidate located in the ACP1 gene.

(A) Gel electrophoresis of the ACP1 variants in cells with both the non-targeting siRNA pool control (NTC) and cells with Upf1 knockdown (siUpf1) in HeLa cells. (B) PSI levels for wt and PTC-containing variants. (C) Relative levels of raw reads supporting exon skipping, normalised to the average number of reads supporting skipping exon skipping in the wt NTC

cells. (D) Relative Upf1 mRNA levels are consistent with inhibition of NMD for both wt and PTC-containing variants.

PTC-associated exon skipping is reading frame independent, supporting the splice motif disruption model whilst providing further evidence against NMD as the source of differential exon inclusion

The above results indicate that associations between PTCs and exon skipping are detectable at the genome-wide level and can be reproduced experimentally. Notably, the experimental result also provides evidence consistent with the PTC disrupting the regular splicing pattern, lending support to the motif disruption model. Can our computational data provide further evidence for or against either of the competing mechanistic NAS models?

Importantly, the distinction between the two mechanisms provides us with a method to test the models. The motif disruption model predicts exon skipping is due to the disruption of important splice control motifs, regardless of the reading frame and that NAS is simply a consequence of a particular mutation that happens to generate an in-frame stop. This mechanism should also apply to mutations that create stop codons out of frame, whereas the scanning model is reading frame-dependent. However, the scanning mechanism requires knowledge of the start codon and reading frame. Thus, to distinguish between these models, we analysed PTCs that are out of frame by one nucleotide (mutations creating out of frame stops, shiftPTCs) and asked if there were similar associations.

We performed similar analyses to those above. We find the absolute differences in shiftPTC PSI scores between the variants (ΔshiftPSI) significantly greater for those with differences consistent with NAS ($P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon rank sum test, median ΔshiftPSI > 0 = 0.051, median ΔshiftPSI < 0 = -1.101). Further, we find the same is true for RPMskip for the shiftPTCs (ΔshiftRPMskip) ($P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon rank sum test, median (ΔshiftRPMskip > 0 = 0.012, median (ΔshiftRPMskip < 0 = -3.518×10^{-4}).

Focusing on exons with large differences in exon skipping as a result of shiftPTCs, we set the threshold for absolute ΔshiftPSI to 5% to match the original analysis. 116/2,949 exons meet this criterion, with a significantly greater number showing decreased PSI in the shiftPTC-/+ than for the shiftPTC-/- variants (94/116, $P = 4.328 \times 10^{-12}$, one-tailed exact binomial test, see Supplementary Figure 3A, Supplementary Figures). Further, 93/116 (80.17%) cases have ΔshiftRPMskip > 0 consistent with NAS (using the threshold of 0.0375 to match the 116 PSI variants), also a significant number ($P = 1.799 \times 10^{-11}$, one-tailed exact binomial test, see Supplementary Figure 3B, Supplementary Figures). 55/116 (1.87% of total shiftPTCs) of these shifted PTCs have greater than 5% difference for both PSI and RPMskip. The number of large-effect cases is significantly lower than for those in-frame ($\chi^2 = 26.056$, $P = 3.316 \times 10^{-7}$, chi-squared test), however, simulations picking 94 and 93 cases at random suggest the 55 large-effect cases with both PSI and RPMskip in the direction consistent with NAS is more than expected by chance ($P \approx 9.999 \times 10^{-5}$, one-tailed empirical P-value).

As observations suggest large-effect cases also associate stop codons with increased exon skipping, the association between the in-frame PTCs and exon skipping is unlikely to be a result of a reading frame-dependant mechanism. These results strengthen the case for the splice motif disruption model. In this context, the mutations disrupting splicing appear to be nonsense mutations that happen to be in-frame, rather than mutations that exert their effects via other pathways.

Importantly, as out of frame PTCs would not be subjected to NMD, the above result provides further evidence that NMD cannot explain the differences in exon inclusion we observe between genotypes.

Five large-effect PTCs have documented associations with disease

That we can identify genome-wide associations between PTCs and potentially disease-associated transcripts is noteworthy. However, the nature of the 1000 Genomes dataset means PTCs are segregating in a healthy population with the circulating frequency of nonsense mutations low, as any with large detrimental phenotypic consequences would likely result in the individual not surviving. Thus, while the 1000 Genomes

dataset is of utility in the context of this study to establish an initial base level of the genome-wide occurrence of NAS and to further elucidate the mechanism, it is limited due to the relative lack of nonsense mutations. More informative are likely to be those PTCs occurring in non-healthy individuals. Thus, we examine NAS indirectly but in a disease-related context.

Are there associations between disease and our large-effect PTCs identified or the genes in which they reside? Several examples of casual relationships with the genes in which the mutations reside are found in the literature. For example, the transcript with the largest PSI difference, ENST00000409520, is encoded by the *TRABD2A* gene associated with negative regulation of the Wnt signalling pathway, itself heavily implicated in cancers (Polakis 2000; Taipale and Beachy 2001; Reya and Clevers 2005; Klaus and Birchmeier 2008; Polakis 2012; Zhang et al. 2012; Zhan et al. 2016). Further, mutations in the *NDUFV2* gene producing the transcript ENST00000400033 have been associated with Parkinson's (Hattori et al. 1998) and Leigh syndrome (Cameron et al. 2015). Mutations in our prime candidate *ACP1* (ENST00000272065) have been associated with diabetes (Gloria-Bottini et al. 1996; Stanford et al. 2017).

We also find associations between the PTC mutations themselves and disease. Five large-effect cases overlap with disease-associated mutations in the ClinVar database (Landrum et al. 2018) archiving relationships between medically important variants and phenotypes (*rs62624965, rs202001274,* rs148458820, *rs200355697*, and *rs74103423* (Table 1 bold, Table 2)).

Pathogenic nonsense mutations are enriched at exon ends and hit ESEs more frequently than expected

As our evidence is consistent with NAS likely to be a result of disrupting important regulatory splice motifs, we ask whether known disease-associated nonsense mutations are frequently located and enriched towards exons ends where ESEs typically reside (Graveley et al. 1998; Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006; Parmley and Hurst 2007; Caceres and Hurst 2013)?

Using the annotated disease-associated mutations from the ClinVar dataset, we established a set of 7,429 nonsense mutations that were not identified in the 1000 Genomes dataset and labelled exclusively either "pathogenic" (N = 6,354) or "likely pathogenic" (N = 1,075). Of the pathogenic mutations, 68.49% (4,352/6,354) are located in the exon-flank nucleotides where ESEs are typically located (nucleotides 3-69), despite the flanking regions only accounting for 55.66% (422,017/758,223) of total coding nucleotides. This frequency is significantly more than expected when comparing locations of SNPs in the splice site nucleotides (nucleotides 1-2), flanking nucleotides (nucleotides 3-69) and exon cores (remaining nucleotides) that result in nonsense codons ($\chi 2 = 578.140$, $P = 2.870 \times 10^{-126}$, chi-squared test, Supplementary Spreadsheet 3, Supplementary Spreadsheets), indicating that pathogenic nonsense mutations typically reside in regions critical for splice regulation. This is also true of the likely-pathogenic mutations (787/1,075 (73.21%) occur in the 111,002/264,391 (41.98%) flanking nucleotides, $\chi^2 = 539.590$, $P = 6.750 \times 10^{-118}$, chi-squared test, Supplementary Spreadsheet 3, Supplementary Spreadsheets). These results are robust to removing "short" exons (those with a length shorter than 138 nucleotides, so the remaining exons include both splice sites, exon flanks and an exon core region), as short exons could be defined as all "exon flank" (Supplementary Spreadsheet 3, Supplementary Spreadsheets). A similar skew towards exon ends has been seen for missense mutations (Wu and Hurst 2016), while the opposite is observed for SNPs that disrupt ESEs circulating in the population (Fairbrother et al. 2004; Carlini and Genut 2006; Caceres and Hurst 2013), consistent with selection against mutations that disrupt splicing.

Although suggestive, this exon flank bias may be a result of a nucleotide-related mutational bias towards exon ends (Chamary and Hurst 2005). To control for this, we performed 10,000 simulations in which every real nonsense mutation was replaced with a randomly chosen nucleotide from the same exon that matched the reference allele of the nonsense mutation, ensuring that each matched nucleotide is not also a disease-associated SNP. For each randomised set of simulated mutations, we then asked where in exons they were located. The real number of pathogenic nonsense mutations in exon flanks is significantly higher when compared with the number expected from the simulations ($Z = 2.217$, $P \approx 0.013$, one-tailed empirical P-value,

Supplementary Figure 4). However, the effect is less pronounced and non-significant for the likely-pathogenic variants ($Z = 0.756$, $P \approx 0.238$, one-tailed empirical P-value). As expected, for the 1000 Genomes PTC mutations we find a significant depletion of mutations in exon flanks ($Z = -2.039$, $P \approx 0.019$, one-tailed empirical P-value). Thus, nonsense mutations that occur in the exon flanks, and are therefore more prone to disrupt splicing, are typically pathogenic.

Despite this biased distribution, pathogenic nonsense mutations may not disrupt splicing. However, if pathogenic nonsense mutations hit ESEs more frequently than expected, this could implicate NAS as a prevalent cause of disease. We asked how often each pathogenic nonsense mutation hit one of the "gold-standard" INT3 ESEs (Caceres and Hurst 2013), expecting ESEs in the 3-69 exon flanking regions to be hit more frequently than by the reference allele-matched simulants. We find this to be the case ($Z = 9.555$, $P \approx 9.99 \times 10^{-5}$, one-tailed empirical P-value, Figure 10). Likely-pathogenic mutations also hit ESEs within the exon flanks more frequently than expected ($Z = 5.877$, $P \approx 9.99 \times 10^{-5}$, one-tailed empirical P-value) although the effect is weaker than for pathogenic variants.

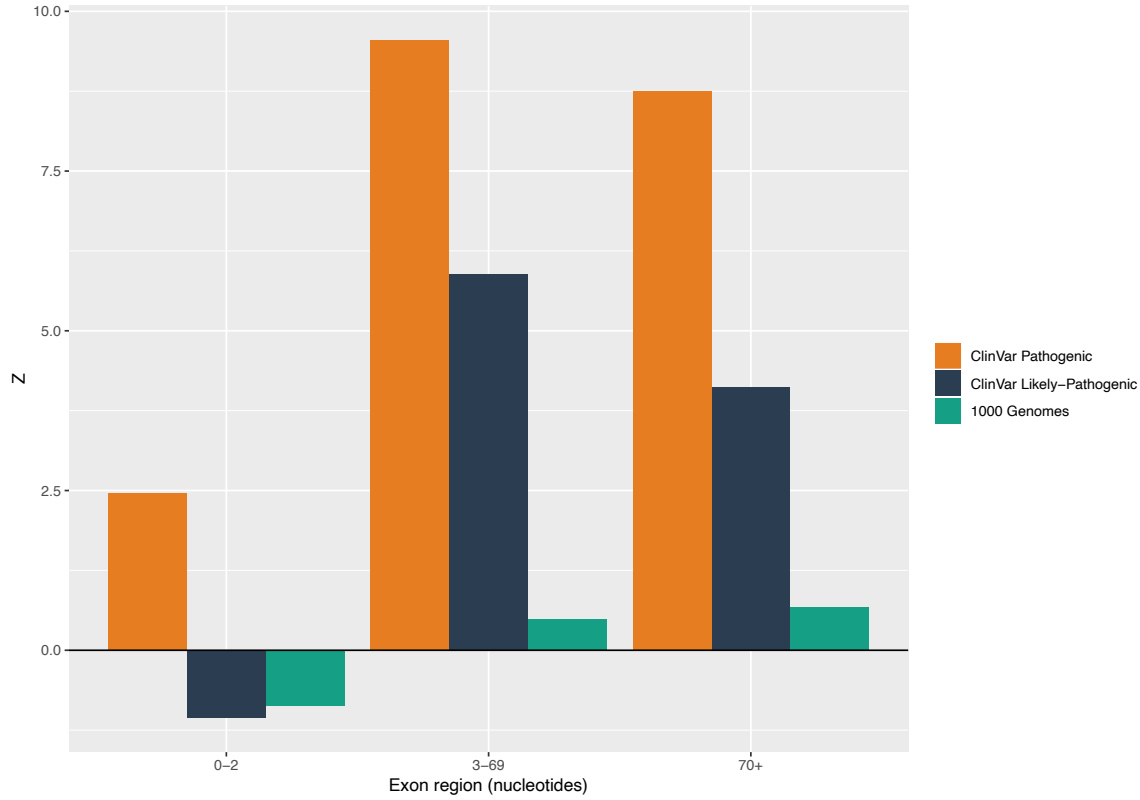


Figure 10: Frequencies of ESE nucleotides hit by PTCs in respective exon regions.

Z scores comparing how frequently pathogenic and likely pathogenic variants from the ClinVar data set and variants in the 1000 Genomes dataset hit ESE motifs when compared with 10,000 randomly sampled nucleotides matching the reference-allele of the nonsense mutation SNP variant. Pathogenic and likely-pathogenic variants hit ESEs significantly more frequently than expected ($P \approx 9.99 \times 10^{-5}$ in both cases), although the enrichment over expected is stronger for pathogenic variants. Consistent with the non-exceptionality of the 1000 Genomes variants, these do not hit ESEs more frequently than expected in any region when compared with the randomly sampled variants.

However, given that the simulated mutations occur less frequently in the 3-69 nucleotide region (see above), for each pathogenic nonsense mutation in the 3-69 nucleotide exon region we randomly picked a nucleotide-matched pseudo-nonsense mutation (not generating a PTC in the ClinVar dataset) also from within the 3-69 nucleotide region. Again, the real nonsense disease-associated mutations hit ESEs more frequently than expected ($Z = 1.920$, $P \approx 0.030$, one-tailed empirical P-value).

These results indicate that disease-associated nonsense mutations are distributed non-randomly in exons and hit ESEs more frequently than expected after control for mutational frequency (between exons and across the same exon), underlying nucleotide content of the 3-69 nucleotide region and relative expression (simulations are within the same exon). Thus, the indications are that disease-associated nonsense mutations are likely to be heavily involved in splice disruption.

We find no significant difference ($\chi^2 = 0.243$, $P = 0.621$, chi-squared test) in the number of exons of length 3n hit by pathogenic nonsense variants (1,440/3,572 = 40.31%) than for other variants (1,565/3,826 = 40.90%). This result suggests that exons not of length 3n (those that would not allow partial rescue of the transcript if skipped) are not more prone to nonsense mutations. Interestingly, exons affected by pathogenic nonsense mutations are longer than those exons affected exclusively by pathogenic synonymous and nonsynonymous variants ($P < 2.2 \times 10^{-16}$, Wilcoxon rank sum test). Consistent with this, these exons have a significantly greater number of codons per 100 bp of exon sequence that are one nucleotide away from a stop codon (median pathogenic nonsense exon one away rate = 10.853, median pathogenic non-nonsense one away rate = 10.120, $P < 2.2 \times 10^{-16}$, Wilcoxon rank sum test). Further, exons containing pathogenic nonsense variants also have a significantly greater ESE density (number of nucleotides contributing to an ESE per bp of sequence) than for the non-nonsense variants (median pathogenic nonsense exon ESE density = 0.171, median pathogenic non-nonsense exon ESE density = 0.159, $P = 2.716 \times 10^{-8}$, Wilcoxon rank sum test). Together, these results suggest that pathogenic nonsense mutations occur more frequently in exons that are more prone to nonsense mutations both in terms of nucleotide content and ESE frequency. Thus, pathogenic nonsense mutations are likely to disrupt splicing frequently.

33% of pathogenic nonsense mutations may have their effect via splicing

Can we estimate the proportion of pathogenic nonsense mutations that may affect splicing (similar to (Wu and Hurst 2016))? We assume exonic core pathogenic nonsense mutations (beyond both the 5' and 3' terminal 69 nucleotides) not to have an effect on splicing. Their rate provides us with a background non-splicing rate (although

is likely conservative as splice-affecting mutations also occur in exon cores (Woolfe et al. 2010)). Any excess of pathogenic nonsense mutations above this core level we then assume to be splice-related. This test also controls for the rate at which NMD degrades transcripts containing nonsense mutations.

Taking the coding exons in which pathogenic nonsense mutations occur (N = 3,572), we define the exon core as any nucleotide beyond the terminal exon 69 nucleotides. We observe 1,804 nonsense mutations in the 321,918 nucleotides of exon cores at a rate of 0.0056 mutations per nucleotide. Thus, assuming exon flanks behave like exon cores we expect ≈2,365 nonsense mutations in the 422,017 exon flank nucleotides. Instead, we observe 4,447, an excess of 2,082 (32.77%) of mutations (see Supplementary Spreadsheet 4, Supplementary Spreadsheets). This suggests that in regions with an increased density of splice information, pathogenic nonsense mutations occur much more frequently than expected. Interestingly, the effect of likely pathogenic mutations appears stronger with an excess of 58.49% of nonsense mutations in exon flanks (see Supplementary Spreadsheet 4, Supplementary Spreadsheets).

Despite the excess, even if pathogenic PTCs are located in regions typically associated with splicing, it is unclear as to whether their association with disease is splice related. We find that a striking 5,258/6,354 (82.75%) of variants had a negative effect on the computationally predicted PSI, a significantly greater number than expected simply by chance ($P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5). This effect is slightly more pronounced in exon flanks (3,722/4,447 (83.70%), $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5) but not significantly so ($\chi^2 = 1.603$, $P = 0.206$, chi-squared test), suggestive that pathogenic nonsense mutations in exon cores also frequently disrupt splicing. This is, however, confounded by the fact that "short" (those less than 138 bp) exons are all exon flank. When restricting the analysis to only those exons longer than 138 bp, although pathogenic nonsense mutations reduce PSI more than expected in both the exon flanks (1,938/2,327 (83.28%), $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5) and exon cores (1,454/1,804 (80.59%), $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5), the difference in the

relative number of mutations decreasing PSI between the two regions is significant (χ^2 = 4.805, P = 0.028, chi-squared test). Thus, our previous estimate of the number of pathogenic nonsense mutations disrupting splicing based on the core rate is likely conservative. We find a similar number of likely-pathogenic mutations decrease PSI (906/1,075 (84.28%), $P < 2.2 \times 10^{-16}$, one-tailed exact binomial test, null probability of success = 0.5).

Taken together with the excess in ESE flanking regions, it is reasonable to assume that splice disruption and exon skipping attributable to PTCs is likely to be a relatively frequent source of pathogenicity. That a conservative estimate suggests nearly one-third of PTCs may affect splicing further highlights the baseline 6% estimation of NAS, we have observed in a healthy population likely underestimates the significance and implications of NAS in disease.

Discussion

Often it is assumed that the pathological consequences of nonsense mutations are due to either the downregulation of mutated isoforms via NMD or to truncation of the regular protein. However, key regulatory splice motifs, ESEs, despite being purine-rich, contain few stop codons (Abrahams and Hurst in press) and are therefore particularly sensitive to mutations creating stop codons disrupting their binding with SR proteins. It could therefore be the case that nonsense mutations exert their detrimental effects by via disruption of splicing at the processing level via NAS. Given the use of and strength of selection on ESEs throughout the genome and that NAS has only been studied in single-/few-gene studies, we have performed a genome-wide study to examine the prevalence of NAS.

Using publicly available RNA-seq data from the Geuvadis RNA-sequencing project (Lappalainen et al. 2013) and DNA polymorphism data from the 1000 Genomes database (The 1000 Genomes Project Consortium 2015), we find significant genome-wide associations between nonsense mutations and exon skipping after accounting for the nucleotide content of the nonsense mutations. In many cases, however, the

biological effects of such increased skipping are likely to be negligible. For instances in which levels of exon inclusion do differ, the effects are significantly greater when in the direction consistent with NAS. Focusing on cases where the effects on exon inclusion are large enough that the consequences are unlikely to be negligible, we find that the proportion of cases where exon skipping is increased with the PTC variant is highly significant. Similarly, we find a significant association between an increase in the absolute number of reads supporting exon skipping and PTCs after controlling for nucleotide content. This suggests that degradation effect of NMD cannot explain the association between PTCs and increased exon skipping. Although not a widespread phenomenon, we established a set of 30 strong candidate large-effect cases for which both the relative (PSI) and the absolute rate (RPMskip) of exon skipping is are substantially increased in the PTC variant and that are likely strong candidates for NAS. We confirmed experimentally that our computational NAS top candidate has quantifiable levels of exon skipping that cannot be explained by NMD. We estimate that ≈6% of annotated nonsense mutations may have significant implications at the mRNA processing level by disrupting splicing.

Despite five of our prime candidate PTC mutations being associated with disease, many of the genome-wide effects on exon skipping that we detect tend to be extremely small. Due to the nature of the 1000 Genomes dataset, the circulating frequency of nonsense mutations is likely low as PTCs with large phenotypic consequences would not be tolerated (and hence unlikely to be seen as segregating polymorphisms). As expected, when asking whether nonsense mutations from the 1000 Genomes data are enriched in the exon flanks based upon nucleotide content alone, we witness a significant depletion. However, for nonsense mutations in the ClinVar dataset, we see a significant excess (see Supplementary Figure 4, Supplementary Figures). This is consistent with purifying selection acting especially strongly on mutations at exons ends (within 69bp of the junction) and argues against a null of differential mutation rates across the exon. Given the reduced rate in 1000 Genomes, likely owing to purifying selection, our 6% estimate is likely conservative and the rate of NAS for disease causing PTCs may be significantly greater. Considering disease-associated nonsense mutations, our results suggest that at least ≈33% might have an effect on splicing. Although the primary aim of this study was to provide a baseline estimate for

NAS based on our observation of ESE composition, our analyses would no doubt benefit from further study using RNA-seq data from both healthy and diseased tissue from the same individual.

Evidence is most parsimonious with PTCs causing exon skipping via splice motif disruption

One of the aims of this study was to distinguish between the two mechanistic models proposed in the literature to explain NAS. Given that we find similar results when comparing the rates of exon skipping when considering out of frame with in-frame stop codons, results are most parsimonious with the splice motif disruption model. For a nuclear scanning mechanism, knowledge of the CDS reading frame post splicing is required to verify the integrity of the transcript and no increase in skipping for out of frame PTCs should be observed. This is contrary to our findings. The fact that we only observe a significant effect in a minority of exons provides further evidence against a nuclear scanning mechanism, as effects could be expected to have a more systematic effect like NMD. That the experimental results demonstrate two different splice isoforms when expressed in HeLa cells, one of which is indicative of partial intron retention, further indicates PTC disruption of splice motifs as this effect is not predicted by the nuclear scanning model.

Further, we find that many disease-associated nonsense mutations are located in regions towards exon ends in which ESEs reside more often than expected, whilst hitting ESEs more frequently than expected by chance within these regions. If the effect was a result of a reading-frame dependant mechanism, there should be no location bias beyond that expected by the underlying nucleotide content. Thus, we can largely discount the nuclear scanning mechanism with our data being more parsimonious with exon skipping via the disruption of ESEs. The RPMskip data, analysis of out of frame stop codons and experimental data examining skipping in the absence of NMD, all argue against NMD as the cause of the effects that we see.

It is, however, interesting to note that without looking at the absolute ΔPSI values or correcting for nucleotide composition, we did not observe PTCs to associate with

lower PSI. This is surprising given that such an association would be expected purely because of NMD downregulation of the full-length isoform, even if no NAS is occurring. One explanation could be that NMD is very weak in our samples. However, the large and highly significant decrease in RPMinclude in PTC-/+ samples argues against this scenario. Alternatively, it is possible that the mutations either create ESEs or disrupt ESSs leading to a slight increase in exon inclusion. A more likely explanation is that, at least for the exons being considered, splicing is very precise and, in most cases, no detectable exon skipping is observed. Indeed, the median PSI overall for PTC-/- samples is ≈99.994% and median ΔPSI ≈0.

NAS is unlikely to be an evolutionarily conserved error-proofing mechanism to rescue PTC-containing transcripts

NMD is commonly thought of as an evolved mechanism to protect against "unwanted" transcripts by recognizing that they contain premature stop codons. However, NMD might itself be the source of problems by reducing the dosage. Could NAS be an evolved quality control mechanism to prevent NMD operating on a particular subclass of genes?

In many cases, the phenotypic consequences of splicing out an exon containing what would be a PTC should be less harmful than either degradation of the transcript or truncation of the protein, particularly if exon skipping maintains reading frame integrity. In this scenario, there could conceivably be a selective evolutionary pressure to preserve ESEs whose nucleotide content allows the formation of a PTC upon a single mutation to protect transcripts that frequently contain nonsense mutations from complete degradation effects of NMD. For example, nonsense mutations in the dystrophin-encoding *DMD* gene result in loss of functional protein (Aartsma-Rus et al. 2016) resulting in Duchenne muscular dystrophy (DMD). However, in Becker muscular dystrophy (BMD), which has a less severe phenotype (Shiga et al. 1997; Carsana et al. 2005; Helderman-van den Enden et al. 2010; Flanigan et al. 2011; Anthony et al. 2014; Bello et al. 2016; Moore et al. 2017), the PTC results in NAS encoding a shortened transcript but retaining the reading frame, restoring partial protein functionality. Similarly, the ability to express functional, yet shortened

isoforms, such as *CEP290* exon-skipped isoforms, is correlated with disease severity (Melis et al. 1998; Di Blasi et al. 2001; Pasmooij et al. 2004; Littink et al. 2010).

However, we find no evidence to suggest PTCs associated with NAS occur in exons of length 3n more frequently than expected, or suggestions that the exons in which the PTCs occur are particularly exceptional (see Supplementary Text 2, Supplementary Texts). Given the relative rates of large-effect NAS, and that much of the variation in splicing associated with other PTCs is very small and likely a reflection of stochastic variation in exon inclusion, it seems unlikely that NAS is a genome-wide error-proofing mechanism under selection to rescue transcripts from NMD. Further, fitness benefits associated with the small variations in exon skipping PTCs are unlikely to be selectable. If PTC-containing transcripts derived from inherited mutations are particularly costly to fitness, the PTC-containing allele would likely be eliminated via purifying selection (although in rare and very specific cases variants are advantageous (North et al. 1999; Yang et al. 2003; Hawn et al. 2005)). Thus, NAS is unlikely to be an evolutionarily conserved adaptive mechanism but rather occurs as a consequence of ESE-binding proteins having to recognise a set of motifs that, due to being located within exons, by definition have a depletion of stop codons.

It could also be questioned why our prime candidates are not seen to hit ESEs more frequently. Several reasons may provide some explanation. First, even if the PTC hits a motif that resembles an ESE, it is hard to know without further analyses whether that motif functions as a splice enhancer in that context, given that many ESE hits are likely to be false positives (Savisaar and Hurst 2018). Alternatively, it is possible that other motifs also function as splice enhancers but are not included in our set of motifs. From a motif analysis alone, it is hard to draw further conclusions.

The importance of accurate classification of nonsense mutations and their roles in therapeutics

Our results demonstrate the importance of understanding the broader implications for the classification of mutations. Even our conservative estimate suggests that the pathogenic effects of a significant proportion of nonsense mutations could be

misunderstood. This data provides further evidence to suggest mutations in general, but SNPs in particular, should be routinely analysed at the mRNA level prior to classification as mutations with seemingly no functional significance can be deleterious (Fackenthal et al. 2002; Pfarr et al. 2005). This is particularly applicable to synonymous mutations, whose pathogenic significance might otherwise be overlooked - such mutations may disrupt ESEs or even create cryptic splice sites that result in a diseased phenotype (Rice et al. 2013; Sheikh et al. 2013; Austin et al. 2017) despite having no direct effect on the peptide sequence.

The consequences of correct classification of nonsense mutations might be best contextualised when considering therapeutic approaches to disease. A variety of therapies targeting nonsense mutations have been shown to restore protein function (Keeling et al. 2014; Dabrowski et al. 2018), however, these therapies are only effective if the PTC is present in the mature transcript. For example, a variety of diseases including Mucopolysaccharidosis type VI (MPS VI) (Bartolomeo et al. 2013), Usher syndrome (Goldmann et al. 2011; Goldmann et al. 2012) and DMD (Yukihara et al. 2011; Finkel et al. 2013) are treated using strategies involving PTC124. This is thought to suppress translation termination at PTCs but not natural stop codons (Welch et al. 2007) and is therefore only effective if substantial levels of mRNA are available containing the PTC. However, if the PTC disrupts splicing and leads to exon skipping the therapy is unlikely to be effective. Furthermore, for cases where a PTC resulting in exon skipping and not of length 3n, such a therapeutic might suppress any downstream out of frame stop codons and allow the synthesis of unknown proteins which may have further detrimental consequences.

Our results demonstrate the first genome-wide association between nonsense mutations and exon skipping. The ability to predict at which sites we would expect such mutations, or mutations in general, to disrupt splicing would be of great future utility for disease prediction, diagnosis and treatment.

Methods

Data sources

All analyses were performed using the reference genome sequence and annotations for GRCh37, Ensembl release 87 (Zerbino et al. 2018) (http://ftp.ensembl.org/; last accessed 25 January 2018). Polymorphism data was retrieved from the EBI 1000 Genomes FTP site (The 1000 Genomes Project Consortium 2015) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/, last accessed 24 January 2018). BAM files containing RNA-seq data for individuals from the 1000 Genomes project were retrieved from the EBI FTP site (Lappalainen et al. 2013) (http://ftp.ebi.ac.uk/, last accessed 8 February 2018). Only samples present in both datasets were retained. Protein family data was downloaded from Ensembl Biomart (Kinsella et al. 2011) (http://grch37.ensembl.org/biomart, last accessed 12 February 2018). ClinVar data containing information regarding disease associated mutations was downloaded from the NCBI FTP site (Landrum et al. 2018) (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/, last accessed 11 May 2018). INT3 ESE motifs were retrieved from the supplementary data to Caceres and Hurst (2013).

General Methods

Custom Python 3.6.4 scripts were used for all data handling and are available at http://github.com/rosinaSav/NAS_code, including the use of standard Python modules. Data plotting and statistical analyses were performed using R v3.2.1 (R Core Team 2017). BEDTools v2.27.1 was used for operations on genome coordinates (Quinlan and Hall 2010). SAMTools v1.7 was used for BAM file manipulation (Li et al. 2009). VCFtools v0.1.15 (Danecek et al. 2011) and tabix v0.2.5 (Li 2011) were used to perform operations on SNP data.

Compilation of protein-coding coding exon set

The main open reading frame (ORF) for protein-coding genes was extracted from the genome annotations. Sequences were filtered to include only those that had canonical start and stop codons, only contained canonical nucleotides, were of a length that is a multiple of three and did not include premature stop codons. Only the transcript isoform with the longest ORF was retained for each of the genes. In order to preserve data independence, only a single gene was retained from each Ensembl protein family. Finally, the internal fully coding exons that did not overlap other annotated exons were extracted. This filtered set of exons were used for all analyses.

SNP filtering

SNPs for individuals were intersected with the set of coding exons to obtain all SNPs within the samples. From these, their relative positions within the exon and CDS were calculated. The mutation status of each SNP was manually determined using this positional data using the reference and variant alleles, with only SNPs that could be verified as PTCs retained. Note that if multiple PTCs were identified in any given exon, only one PTC was kept, leaving 1,180 PTCs.

Quantification of splice isoforms

Reads from the Geuvadis BAM files were subject to quality filtering as per (Lappalainen et al. 2013). Reads were filtered to uniquely mapped reads with a base mapping quality scale between 251 and 255 or 175 and 181 inclusive. Further, only reads with no more than 6 mismatches were included. These reads were then mapped to the exon-exon junctions that flank the exons in our dataset.

For each exon and each individual, we counted the number of reads that support inclusion by counting those that overlapped the focal exon and either of the two flanking exons as defined by Ensembl annotations. Similarly, we counted reads supporting skipping by counting the number of reads that map to the junction between the two flanking exons of the focal gene. The number of reads supporting exon skipping were multiplied by two as these reads can only map to a single exon-exon junction whereas reads that support exon inclusion can overlap either of two exon-exon junctions.

Read counts were then used to calculate several metrics for each exon in each sample: PSI, RPMinclude and RPMskip. PSI is defined as the number of reads containing the exon, divided by the number of reads containing the exon plus the number of reads where the exon is skipped. RPMinclude is defined as the number of reads containing the exon divided by the total number of reads in the sample. RPMskip is defined as the number of reads without the exon divided by the total number of reads in the sample.

For RPMinclude and RPMskip, the total number of reads after quality filtering of the BAM files is required to account for read quality between samples and differences in sequencing depth. However, it is computationally impractical to quality filter complete BAM files. Therefore, we first determined the total read count. We then filtered the BAM file to only contain reads overlapping our exon-exon junctions. We performed the quality filtering on these exon-exon junction reads and sampled the read count. The proportional decrease between the non-quality filtered exon-exon junction reads and quality filtered exon-exon junction reads was then used to scale the initial read count to estimate the number of total reads after quality filtering. We find no significant difference ($P = 0.188$, paired Wilcoxon signed-rank test) between the proportion of reads retained after filtering the full BAM file and after filtering after intersection with exon-exon junctions, arguing that applying the proportional decrease for exon-exon junctions to the full read count is unbiased and appropriate (see Supplementary Figure 5, Supplementary Figures).

Detecting splicing of exons

Having identified a set of PTC-containing exons and calculated the metrics, we filtered our list of 1,180 PTC-containing exons to exclude those for which none of the individuals in which we could quantify splicing contained the PTC. The number of individuals with quantifiable splicing from the exon with the maximum number of individuals with quantifiable splicing was retained. We then excluded PTC-containing exons in which the number of individuals with quantifiable splicing was less than half of the maximum. Finally, to remove exons that might be alternatively spliced only exons included in all transcript isoforms from the same locus were retained. This left $N = 541$ PTC-containing exons.

Missense mutation simulations

We performed 100 simulations in which each of the real PTCs was randomly matched to a missense mutation. For each PTC, the missense mutation was sampled in order to match the PTCs ancestral allele identify, variant allele identity and variant allele frequency (within a threshold of 0.05). The same analyses were then performed on the sets of pPTCs. To further control for distance to exon boundary, a window of five nucleotides to either side of the relative position of the PTC was defined, with any appropriate pPTCs selected. If none were available, this window was increased by one

nucleotide until a suitable simulant was identified or 10 window expansions had occurred, whichever was the former.

Minigene constructs

A minigene construct for *ACP1* (ENST00000272065) was ordered from GeneArt as a double-stranded DNA string subcloned into the Gateway-entry vector *pENTR221*. The minigene consisted of the 5' flanking exon, 5' flanking intron, focal exon, 3' flanking intron and 3' flanking exon (see Supplementary Spreadsheet 5, Supplementary Spreadsheets for sequence information). Two versions were designed: one in which the wild-type sequence of the focal exon is preserved (wt) and one containing the PTC-causing mutation (mut). To allow these genes to be translated, a start codon (ATG) was added at the 5' end of all sequences. The 3' flanking exon is the final exon and therefore already contains a stop codon (TGA). All minigenes were subcloned into *pCM3*, a Gateway-compatible CMV-driven mammalian expression vector (described in (Mordstein et al. 2019)), using Gateway LR Clonase II enzyme mix (Thermo Fisher) according to manufacturer's instructions. *pCM3* additionally also drives the constitutive expression of *mKate2* from an independent expression cassette which allows to correct for technical variability in transfection efficiency. The control NMD reporter constructs of human *TCR*-β have been previously described (Wang et al. 2002b).

Plasmid and siRNA transfections

HeLa and Hek293T cells were maintained in DMEM (Gibco) supplemented with 10% fetal calf serum (FCS) at 37°C, 5%CO_2. NMD knockdown experiments were performed by two rounds of consecutive transfections with siRNA targeting *Upf1* (sihUPF1-I: GAGAAUCGCCUACUUCACU (+UU) and sihUPF1-II: GAUGCAGUUCCGCUCCAUU (+UU), Dharmacon, mixed in equimolar ratio). As a negative control, cells were transfected with a non-targeting control siRNA (ON-TARGETplus Non-targeting Control Pool, Dharmacon). In brief, cells were grown to 40% confluency in 12-well plates before transfecting with 1.25ul of 20uM siRNA stocks using 5ul Dharmafect1 transfection reagent (Dharmacon). After 48hrs, the siRNA transfection was repeated using Lipofectamine2000 transfection reagent instead (Thermo Fisher) and with the addition of 100ng of *pCM3* plasmid carrying the minigenes. Cells were grown for a further 48hrs before harvesting.

RNA extraction and RT-PCR analysis

RNA from transfected cells was extracted using the Qiagen RNAeasy kit according to manufacturer's instructions, including the on-column DNase digest step. cDNA synthesis was performed using SuperScript III Reverse Transcriptase (Thermo Fisher) with 1ug of RNA and using 500ng anchored oligo(dT)20 primers (Thermo Fisher). cDNA was further treated with 5U RNAse H (NEB) before diluting with 30ul nuclease-free water. 2ul of each cDNA dilution were used as template in PCR reactions using either AccuPrime *Pfx* DNA polymerase (Life Technologies; *ACP1* and *mKate2* for HeLa samples) or *Taq* DNA polymerase (Life Technologies; *ACP1* and *mKate2* for Hek293T samples) following manufacturer's recommendations and 0.3uM of gene-specific primers (for primer sequences see Supplementary Spreadsheet 6, Supplementary Spreadsheets), ensuring amplification is within the exponential range. For quantitative Real-time PCR measurements of *Upf1* and *TCR* expression, samples were analysed in triplicate reactions on a Roche LightCycler480 using Roche LightCycler480 SYBR Green I Master Mix. Relative expression levels were determined using the Comparative Ct method (Livak and Schmittgen 2001) and normalised against *GAPDH* levels. *ACP1* and *mKate2* PCR products were resolved on 1.5% agarose in TBE gels stained with Ethidiumbromide and imaged on a Syngene U:Genius 3 gel imager. Bands were quantified via densitometry with background subtraction using Image Studio Lite (v5.2). The resulting signals from *ACP1* bands were further normalised to the signal of *mKate2* bands from the same respective cDNA to account for technical variability in transfection efficiency. PSI was calculated as before, using the normalised signal of full-length transcript divided by the normalised signal of full-length transcript plus the normalised signal of transcript with skipped exon.

Out of frame PTCs analysis

For the out of frame PTC analysis, when determining the SNP type (synonymous, missense, nonsense) we shifted the reading frame forwards by one nucleotide. As a result, if the three nucleotides starting from the second position of a codon encoded a stop codon, we called this a PTC. For the last codon of the ORF, the reading frame was instead shifted backwards one nucleotide. We then repeated the pipeline with shifted PTCs.

ClinVar analyses

Disease-associated mutations were downloaded from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/, last accessed May 11 2018; (Landrum et al. 2016)) and intersected with the filtered exon set to leave only SNPs that occurred in our coding exons (N = 156,730). We then verified the status of the disease-associated mutations, retaining only those labelled "pathogenic" or "likely-pathogenic". The mutation status of each SNP was then verified. N = 13,959 synonymous and nonsynonymous variants were retained for ensuring these variants were not used in the reference allele-matched simulations and for determining the exons in which they reside for exon comparisons. Nonsense mutations were intersected with the 1000 Genomes dataset and only the N = 7,429 non-overlapping variants retained.

Splice variant prediction

PTC variants were analysed using MMsplice (Cheng et al. 2019), a neural network model trained on large-scale genomics datasets to predict the effects of variants on exon skipping, splice site choice, splicing efficiency and pathogenicity. Variants were compiled into a single VCF file with effects predicted using the model default parameters (exon_cut_l = 0, exon_cut_r = 0, acceptor_intron_cut = 6, donor_intron_cut = 6, acceptor_intron_len = 50, acceptor_exon_len = 3, donor_exon_len = 5, donor_intron_len = 13, split_seq = False). Changes in exon inclusion are reported as mmsplice_dlogitPsi values, with negative values indicating a predicted increase in exon skipping (lower PSI) and positive values indicating a predicted decrease in exon skipping (greater PSI) due to the variant.

Expression analysis

We used FANTOM5 data (The Fantom Consortium et al. 2014) to estimate expression parameters independently of the Geuvadis RNA-seq data that was used to analyse splice isoforms. We retrieved the phase 1 and 2 combined normalized .osc file from the FANTOM5 website (http://fantom.gsc.riken.jp/5/datafiles; last accessed 11 February 2016). We only retained samples where the name contained the string *adult*, *pool1*. All brain tissues except for the full brain sample and the retinal sample were removed to avoid redundancy. For each gene included in our analysis, we defined a region of 1001 base pairs centred on the start coordinate of the Ensembl transcript

annotation as the promoter and associated all peaks that overlapped that promoter to that peak. If several peaks were associated to a single transcript, we summed the tags per million (TPM) within each sample across the peaks. A gene was considered to be expressed in a given tissue if TPM > 5.

Tables

Table 5: 30 prime NAS candidates

	PSI			RPMskip			
Exon ID	-/+	-/-	Δ	-/+	-/-	Δ	ΔlogitΨ
ENST00000272065.5	39.19	99.76	-60.57	4.784	0.007	4.777	-0.520
ENST00000325083.24	29.26	55.84	-26.59	2.631	1.389	1.242	-0.266
ENST00000271324.6	88.69	98.20	-9.51	1.195	0.391	0.804	-0.166
ENST00000400033.8	16.67	95.62	-78.96	0.681	0.022	0.659	-0.963
ENST00000216027.4	59.09	93.71	-34.62	0.483	0.100	0.383	-0.432
ENST00000359028.47	64.29	99.83	-35.55	0.366	0.001	0.364	-0.465
ENST00000367409.18	**69.08**	**75.82**	**-6.74**	**0.538**	**0.239**	**0.299**	**0.005**
ENST00000267430.22	48.63	99.24	-50.61	0.162	0.004	0.158	0.560
ENST00000288050.18	76.81	88.17	-11.36	0.234	0.078	0.156	-0.163
ENST00000456763.12	**75.76**	**97.33**	**-21.57**	**0.111**	**0.011**	**0.100**	**-0.029**
ENST00000255409.8	57.89	93.08	-35.18	0.111	0.018	0.093	-0.144
ENST00000272252.4	89.09	99.88	-10.79	0.079	0.001	0.078	-4.478
ENST00000222800.4	68.00	93.15	-25.15	0.110	0.032	0.078	-0.090
ENST00000382977.11	33.33	100.00	-66.67	0.073	0.000	0.073	-0.804
ENST00000389175.23	20.00	64.30	-44.30	0.113	0.052	0.061	-0.151
ENST00000265316.3	**83.78**	**97.45**	**-13.67**	**0.079**	**0.018**	**0.061**	**0.053**
ENST00000355774.3	89.19	99.93	-10.75	0.054	0.000	0.054	-0.379
ENST00000398141.8	10.95	19.95	-9.00	0.699	0.648	0.052	-0.640
ENST00000357115.15	90.70	99.74	-9.04	0.053	0.003	0.051	-0.570
ENST00000487270.3	**92.59**	**99.68**	**-7.08**	**0.052**	**0.002**	**0.050**	**-0.251**
ENST00000216294.2	92.31	99.55	-7.24	0.054	0.003	0.050	-0.150
ENST00000338382.7	91.30	99.77	-8.47	0.053	0.003	0.050	-0.559
ENST00000331493.9	9.58	21.35	-11.77	0.149	0.099	0.050	-0.203
ENST00000328867.14	69.23	89.54	-20.31	0.055	0.014	0.041	-1.237
ENST00000376811.6	89.58	99.50	-9.91	0.041	0.003	0.037	-0.092
ENST00000535273.7	83.78	94.41	-10.62	0.081	0.045	0.036	0.116
ENST00000370132.6	**85.45**	**98.61**	**-13.16**	**0.041**	**0.008**	**0.033**	**-0.221**
ENST00000542534.16	50.00	100.00	-50.00	0.027	0.000	0.027	-0.167
ENST00000354366.10	81.82	99.98	-18.16	0.027	0.000	0.027	-0.153
ENST00000238561.9	82.35	95.81	-13.45	0.041	0.015	0.026	0.018

The 30 prime NAS candidates are those supporting an association between the PTC and exon increased relative exon skipping (ΔPSI < -5) and absolute exon skipping (ΔRPMskip > 0.026), sorted by decreasing ΔRPMskip. Exon ID is defined as "ensembl_transcript_id.exon_number" where the exon number is incremented in the

direction of transcription. ΔlogitΨ scores are those predicted by MMSplice. PTCs that also appear in the ClinVar dataset are shown in bold.

Table 6: Further information regarding the five prime NAS candidates overlapping ClinVar variants.

PTC ID	Exon ID	Mutation	Information
rs62624965	ENST00000367409.18	T > G	• *ASPM* gene. • Benign mutation (Landrum et al. 2018). • *ASPM* produces two isoforms, one with exon 18 skipped, in both human and mouse and therefore may encode two proteins with different functions (Kouprina et al. 2005), thus skipping of exon 18 may not be as detrimental.
rs202001274	ENST00000456763.12	C > T	• *MAPKBP1* gene. • Associated with Nephronophthisis 20 (Macia et al. 2017). • Homozygous PTC Individual produced full-length and exon-skipped isoforms. • Thought to affect binding of serine-arginine rich (SR) protein SF2/ASF binding leading to exon skipped isoforms.
rs148458820	ENST00000265316.3	G > A	• *ABCB6* gene. • Mitochondrial porphyrin transporter essential for heme biosynthesis. • Associated with Langereis blood group (Helias et al. 2012). • May have implications in blood transfusions and drug therapies (Boswell-Casteel et al. 2017). • ABCB6 also thought to contribute to anticancer

			drug resistance (Kelter et al. 2007).
rs200355697	ENST00000487270.3	C > T	• *RAD51B* gene. • Encodes a DNA repair protein. • Uncertain significance for hereditary cancer-predisposing syndrome. • *RAD51B* splice mutations leading to exon skipping have been associated with cancer (Golmard et al. 2013).
rs74103423	ENST00000370132.6	G > T	• *Dihydrolipoamide branched chain transacylase E2* gene. • Associated with maple syrup urine disease (MSUD) (Fisher et al. 1993). Truncated and exon skipped isoforms found.

Exon ID is defined as “ensembl_transcript_id.exon_number” where the exon number is incremented in the direction of transcription.

References

Aartsma-Rus A, Ginjaar IB, Bushby K. 2016. The importance of genetic diagnosis for Duchenne muscular dystrophy. *J. Med. Genet.* 53:145-151.

Abrahams L, Hurst LD. in press. A depletion of stop codons in lincRNA is owing to transfer of selective constraint from coding sequences. *Mol. Biol. Evol.*

Anthony K, Arechavala-Gomeza V, Ricotti V, Torelli S, Feng L, Janghra N, Tasca G, Guglieri M, Barresi R, Armaroli A, et al. 2014. Biochemical characterization of patients with in-frame or out-of-frame DMD deletions pertinent to exon 44 or 45 skipping. *JAMA neurology* 71:32-40.

Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X. 2000. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* 9:237-247.

Austin F, Oyarbide U, Massey G, Grimes M, Corey SJ. 2017. Synonymous mutation in TP53 results in a cryptic splice site affecting its DNA-binding site in an adolescent with two primary sarcomas. *Pediatric blood & cancer* 64.

Aznarez I, Zielenski J, Rommens JM, Blencowe BJ, Tsui LC. 2007. Exon skipping through the creation of a putative exonic splicing silencer as a consequence of the cystic fibrosis mutation R553X. *J. Med. Genet.* 44:341-346.

Baralle D, Lucassen A, Buratti E. 2009. Missed threads. The impact of pre-mRNA splicing defects on clinical practice. *EMBO Rep* 10:810-816.

Barny I, Perrault I, Michel C, Soussan M, Goudin N, Rio M, Thomas S, Attie-Bitach T, Hamel C, Dollfus H, et al. 2018. Basal exon skipping and nonsense-associated altered splicing allows bypassing complete CEP290 loss-of-function in individuals with unusually mild retinal disease. *Hum. Mol. Genet.* 27:2689-2702.

Bartolomeo R, Polishchuk EV, Volpi N, Polishchuk RS, Auricchio A. 2013. Pharmacological read-through of nonsense ARSB mutations as a potential therapeutic approach for mucopolysaccharidosis VI. *J. Inherited Metab. Dis.* 36:363-371.

Bello L, Campadello P, Barp A, Fanin M, Semplicini C, Soraru G, Caumo L, Calore C, Angelini C, Pegoraro E. 2016. Functional changes in Becker muscular dystrophy: implications for clinical trials in dystrophinopathies. *Sci Rep* 6:32439.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 25:106-110.

Boswell-Casteel RC, Fukuda Y, Schuetz JD. 2017. ABCB6, an ABC Transporter Impacting Drug Response and Disease. *The AAPS Journal* 20:8.

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.* 11:331-335.

Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* 16:107-113.

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Cameron JM, MacKay N, Feigenbaum A, Tarnopolsky M, Blaser S, Robinson BH, Schulze A. 2015. Exome sequencing identifies complex I NDUFV2 mutations as a novel cause of Leigh syndrome. *European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society* 19:525-532.

Caputi M, Kendzior RJ, Jr., Beemon KL. 2002. A nonsense mutation in the fibrillin-1 gene of a Marfan syndrome patient induces NMD and disrupts an exonic splicing enhancer. *Genes Dev.* 16:1754-1759.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* 62:89-98.

Carsana A, Frisso G, Tremolaterra MR, Lanzillo R, Vitale DF, Santoro L, Salvatore F. 2005. Analysis of dystrophin gene deletions indicates that the hinge III region of the protein correlates with disease severity. *Ann Hum Genet* 69:253-259.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* 3:285-298.

Chamary JV, Hurst LD. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.* 21:256-259.

Chang YF, Chan WK, Imam JS, Wilkinson MF. 2007. Alternatively spliced T-cell receptor transcripts are up-regulated in response to disruption of either splicing elements or reading frame. *J. Biol. Chem.* 282:29738-29747.

Chemin G, Tinguely A, Sirac C, Lechouane F, Duchez S, Cogne M, Delpy L. 2010. Multiple RNA surveillance mechanisms cooperate to reduce the amount of nonfunctional Ig kappa transcripts. *J. Immunol.* 184:5009-5017.

Cheng J, Nguyen TYD, Cygan KJ, Celik MH, Fairbrother WG, Avsec Z, Gagneur J. 2019. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* 20:48.

Chung CG, Lee H, Lee SB. 2018. Mechanisms of protein toxicity in neurodegenerative diseases. *Cell. Mol. Life Sci.* 75:3159-3180.

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

Dabrowski M, Bukowy-Bieryllo Z, Zietkiewicz E. 2018. Advances in therapeutic use of a drug-stimulated translational readthrough of premature termination codons. *Mol Med* 24:25.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.

Di Blasi C, He Y, Morandi L, Cornelio F, Guicheney P, Mora M. 2001. Mild muscular dystrophy due to a nonsense mutation in the LAMA2 gene resulting in exon skipping. *Brain* 124:698-704.

Dietz HC, Kendzior RJ, Jr. 1994. Maintenance of an open reading frame as an additional level of scrutiny during splice site selection. *Nat Genet* 8:183-188.

Disset A, Bourgeois CF, Benmalek N, Claustres M, Stevenin J, Tuffery-Giraud S. 2006. An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. *Hum. Mol. Genet.* 15:999-1013.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341-352.

Endo F, Awata H, Katoh H, Matsuda I. 1995. A nonsense mutation in the 4-hydroxyphenylpyruvic acid dioxygenase gene (Hpd) causes skipping of the constitutive exon and hypertyrosinemia in mouse strain III. *Genomics* 25:164-169.

Fackenthal JD, Cartegni L, Krainer AR, Olopade OI. 2002. BRCA2 T2722R is a deleterious allele that causes exon skipping. *Am. J. Hum. Genet.* 71:625-631.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268.

Finkel RS, Flanigan KM, Wong B, Bonnemann C, Sampson J, Sweeney HL, Reha A, Northcutt VJ, Elfring G, Barth J, et al. 2013. Phase 2a study of ataluren-mediated dystrophin production in patients with nonsense mutation Duchenne muscular dystrophy. *PLoS One* 8:e81302.

Fisher CW, Fisher CR, Chuang JL, Lau KS, Chuang DT, Cox RP. 1993. Occurrence of a 2-bp (AT) deletion allele and a nonsense (G-to-T) mutant allele at the E2 (DBT) locus of six patients with maple syrup urine disease: multiple-exon skipping as a secondary effect of the mutations. *Am. J. Hum. Genet.* 52:414-424.

Flanigan KM, Dunn DM, von Niederhausern A, Soltanzadeh P, Howard MT, Sampson JB, Swoboda KJ, Bromberg MB, Mendell JR, Taylor LE, et al. 2011. Nonsense mutation-associated Becker muscular dystrophy: interplay between exon definition and splicing regulatory elements within the DMD gene. *Hum. Mutat.* 32:299-308.

Gersappe A, Pintel DJ. 1999. A premature termination codon interferes with the nuclear function of an exon splicing enhancer in an open reading frame-dependent manner. *Mol. Cell. Biol.* 19:1640-1650.

Gibson RA, Hajianpour A, Murer-Orlando M, Buchwald M, Mathew CG. 1993. A nonsense mutation and exon skipping in the Fanconi anaemia group C gene. *Hum. Mol. Genet.* 2:797-799.

Ginsburg GS, Phillips KA. 2018. Precision Medicine: From Science To Value. *Health Aff (Millwood)* 37:694-701.

Gloria-Bottini F, Gerlini G, Lucarini N, Borgiani P, Amante A, La Torre M, Antonacci E, Bottini E. 1996. Phosphotyrosine protein phosphatases and diabetic pregnancy: an association between low molecular weight acid phosphatase and degree of glycemic control. *Experientia* 52:340-343.

Goldmann T, Overlack N, Möller F, Belakhov V, van Wyk M, Baasov T, Wolfrum U, Nagel-Wolfrum K. 2012. A comparative evaluation of NB30, NB54 and PTC124 in translational read-through efficacy for treatment of an USH1C nonsense mutation. *EMBO Molecular Medicine* 4:1186-1199.

Goldmann T, Overlack N, Wolfrum U, Nagel-Wolfrum K. 2011. PTC124-mediated translational readthrough of a nonsense mutation causing Usher syndrome type 1C. *Hum. Gene Ther.* 22:537-547.

Golmard L, Caux-Moncoutier V, Davy G, Al Ageeli E, Poirot B, Tirapo C, Michaux D, Barbaroux C, d'Enghien CD, Nicolas A, et al. 2013. Germline mutation in the RAD51B gene confers predisposition to breast cancer. *BMC Cancer* 13:484.

Graveley BR, Hertel KJ, Maniatis T. 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *The EMBO Journal* 17:6747-6756.

Hattori N, Yoshino H, Tanaka M, Suzuki H, Mizuno Y. 1998. Genotype in the 24-kDa subunit gene (NDUFV2) of mitochondrial complex I and susceptibility to Parkinson disease. *Genomics* 49:52-58.

Hawn TR, Wu H, Grossman JM, Hahn BH, Tsao BP, Aderem A. 2005. A stop codon polymorphism of Toll-like receptor 5 is associated with resistance to systemic lupus erythematosus. *Proc Natl Acad Sci U S A* 102:10593-10597.

Helderman-van den Enden AT, Straathof CS, Aartsma-Rus A, den Dunnen JT, Verbist BM, Bakker E, Verschuuren JJ, Ginjaar HB. 2010. Becker muscular dystrophy

patients with deletions around exon 51; a promising outlook for exon skipping therapy in Duchenne patients. *Neuromuscular disorders : NMD* 20:251-254.
Helias V, Saison C, Ballif BA, Peyrard T, Takahashi J, Takahashi H, Tanaka M, Deybach J-C, Puy H, Le Gall M, et al. 2012. ABCB6 is dispensable for erythropoiesis and specifies the new blood group system Langereis. *Nature Genetics* 44:170.
Hoffmeyer S, Nurnberg P, Ritter H, Fahsold R, Leistner W, Kaufmann D, Krone W. 1998. Nearby stop codons in exons of the neurofibromatosis type 1 gene are disparate splice effectors. *Am. J. Hum. Genet.* 62:269-277.
Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. 2004a. Nonsense-mediated decay approaches the clinic. *Nat Genet* 36:801-808.
Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. 2004b. Nonsense-mediated decay approaches the clinic. *Nature Genetics* 36:801-808.
Hull J, Shackleton S, Harris A. 1994. The stop mutation R553X in the CFTR gene results in exon skipping. *Genomics* 19:362-364.
Jackson M, Marks L, May GHW, Wilson JB. 2018. The genetic basis of disease. *Essays Biochem* 62:643-723.
Karam R, Carvalho J, Bruno I, Graziadio C, Senz J, Huntsman D, Carneiro F, Seruca R, Wilkinson MF, Oliveira C. 2008. The NMD mRNA surveillance pathway downregulates aberrant E-cadherin transcripts in gastric cancer cells and in CDH1 mutation carriers. *Oncogene* 27:4255-4260.
Keeling KM, Xue X, Gunn G, Bedwell DM. 2014. Therapeutics based on stop codon readthrough. *Annu Rev Genomics Hum Genet* 15:371-394.
Kelter G, Steinbach D, Konkimalla VB, Tahara T, Taketani S, Fiebig HH, Efferth T. 2007. Role of transferrin receptor and the ABC transporters ABCB6 and ABCB7 for resistance and differentiation of tumor cells towards artesunate. *PLoS One* 2:e798.
Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation* 2011:bar030.
Klaus A, Birchmeier W. 2008. Wnt signalling and its impact on development and cancer. *Nature Reviews Cancer* 8:387-398.
Kouprina N, Pavlicek A, Collins NK, Nakano M, Noskov VN, Ohzeki J-I, Mochida GH, Risinger JI, Goldsmith P, Gunsior M, et al. 2005. The microcephaly ASPM gene is expressed in proliferating tissues and encodes for a mitotic spindle protein. *Hum. Mol. Genet.* 14:2155-2165.
Laimer M, Onder K, Schlager P, Lanschuetzer CM, Emberger M, Selhofer S, Hintner H, Bauer JW. 2008. Nonsense-associated altered splicing of the Patched gene fails to suppress carcinogenesis in Gorlin syndrome. *Br. J. Dermatol.* 159:222-227.
Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44:D862-868.
Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46:D1062-D1067.
Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013.

Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506-511.

Lenassi E, Saihan Z, Bitner-Glindzicz M, Webster AR. 2014. The effect of the common c.2299delG mutation in USH2A on RNA splicing. *Exp Eye Res* 122:9-12.

Li B, Wachtel C, Miriami E, Yahalom G, Friedlander G, Sharon G, Sperling R, Sperling J. 2002. Stop codons affect 5' splice site selection by surveillance of splicing. *Proc Natl Acad Sci U S A* 99:5277-5282.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27:718-719.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A* 108:11093-11098.

Littink KW, Pott JW, Collin RW, Kroes HY, Verheij JB, Blokland EA, de Castro Miro M, Hoyng CB, Klaver CC, Koenekoop RK, et al. 2010. A novel nonsense mutation in CEP290 induces exon skipping and leads to a relatively mild retinal phenotype. *Investigative ophthalmology & visual science* 51:3646-3652.

Liu HX, Cartegni L, Zhang MQ, Krainer AR. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* 27:55-58.

Livak KJ, Schmittgen TD. 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* 25:402-408.

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. 2005. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* 579:1900-1903.

Lorson CL, Hahnen E, Androphy EJ, Wirth B. 1999. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci U S A* 96:6307-6311.

Macia MS, Halbritter J, Delous M, Bredrup C, Gutter A, Filhol E, Mellgren AEC, Leh S, Bizet A, Braun DA, et al. 2017. Mutations in MAPKBP1 Cause Juvenile or Late-Onset Cilia-Independent Nephronophthisis. *Am. J. Hum. Genet.* 100:323-333.

Maquat LE. 2002. NASty effects on fibrillin pre-mRNA splicing: another case of ESE does it, but proposals for translation-dependent splice site choice live on. *Genes Dev.* 16:1743-1753.

Maquat LE. 2005. Nonsense-mediated mRNA decay in mammals. *J. Cell Sci.* 118:1773-1776.

Maquat LE, Li X. 2001. Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. *RNA* 7:445-456.

Mazoyer S, Puget N, Perrin-Vidoz L, Lynch HT, Serova-Sinilnikova OM, Lenoir GM. 1998. A BRCA1 nonsense mutation causes exon skipping. *Am. J. Hum. Genet.* 62:713-715.

Meldau S, De Lacy RJ, Riordan GTM, Goddard EA, Pillay K, Fieggen KJ, Marais AD, Van der Watt GF. 2018. Identification of a single MPV17 nonsense-associated altered splice variant in 24 South African infants with mitochondrial neurohepatopathy. *Clin. Genet.* 93:1093-1096.

Melis MA, Muntoni F, Cau M, Loi D, Puddu A, Boccone L, Mateddu A, Cianchetti C, Cao A. 1998. Novel nonsense mutation (C-->A nt 10512) in exon 72 of dystrophin gene leading to exon skipping in a patient with a mild dystrophinopathy. *Hum. Mutat.* Suppl 1:S137-138.

Mendell JT, ap Rhys CM, Dietz HC. 2002. Separable roles for rent1/hUpf1 in altered splicing and decay of nonsense transcripts. *Science* 298:419-422.

Mendell JT, Dietz HC. 2001. When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell* 107:411-414.

Mendive FM, Rivolta CM, Gonzalez-Sarmiento R, Medeiros-Neto G, Targovnik HM. 2005. Nonsense-associated alternative splicing of the human thyroglobulin gene. *Mol Diagn* 9:143-149.

Messiaen L, Callens T, De Paepe A, Craen M, Mortier G. 1997. Characterisation of two different nonsense mutations, C6792A and C6792G, causing skipping of exon 37 in the NF1 gene. *Hum. Genet.* 101:75-80.

Moore RS, Tirupathi S, Herron B, Sands A, Morrison PJ. 2017. Dystrophin Exon 29 Nonsense Mutations Cause a Variably Mild Phenotype. *Ulster Med J* 86:185-188.

Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2019. Splicing buffers suboptimal codon usage in human cells. *bioRxiv*:527440.

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. *Hum. Mutat.* 29:1037-1047.

Moseley CT, Mullis PE, Prince MA, Phillips JA, III. 2002. An Exon Splice Enhancer Mutation Causes Autosomal Dominant GH Deficiency. *The Journal of Clinical Endocrinology & Metabolism* 87:847-852.

North KN, Yang N, Wattanasirichaigoon D, Mills M, Easteal S, Beggs AH. 1999. A common nonsense mutation results in α-actinin-3 deficiency in the general population. *Nature Genetics* 21:353-354.

Pagani F, Buratti E, Stuani C, Baralle FE. 2003. Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. *J. Biol. Chem.* 278:26580-26588.

Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* 20:153-158.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* 23:301-309.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* 24:1600-1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.

Pasmooij AM, van Zalen S, Nijenhuis AM, Kloosterhuis AJ, Zuiderveen J, Jonkman MF, Pas HH. 2004. A very mild form of non-Herlitz junctional epidermolysis bullosa: BP180 rescue by outsplicing of mutated exon 30 coding for the COL15 domain. *Experimental dermatology* 13:125-128.

Peterlongo P, Catucci I, Colombo M, Caleca L, Mucaki E, Bogliolo M, Marin M, Damiola F, Bernard L, Pensotti V, et al. 2015. FANCM c.5791C>T nonsense

mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. *Hum. Mol. Genet.* 24:5345-5355.
Pfarr N, Prawitt D, Kirschfink M, Schroff C, Knuf M, Habermehl P, Mannhardt W, Zepp F, Fairbrother WG, Loos M, et al. 2005. Linking C5 deficiency to an exonic splicing enhancer mutation. *J. Immunol.* 174:4172-4177.
Polakis P. 2000. Wnt signaling and cancer. *Genes Dev.* 14:1837-1851.
Polakis P. 2012. Wnt Signaling in Cancer. *Cold Spring Harbor Perspectives in Biology* 4:a008052.
Price AL, Spencer CCA, Donnelly P. 2015. Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences* 282:20151684.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
Reya T, Clevers H. 2005. Wnt signalling in stem cells and cancer. *Nature* 434:843-850.
Rice GI, Reijns MA, Coffin SR, Forte GM, Anderson BH, Szynkiewicz M, Gornall H, Gent D, Leitch A, Botella MP, et al. 2013. Synonymous mutations in RNASEH2A create cryptic splice sites impairing RNase H2 enzyme function in Aicardi-Goutieres syndrome. *Hum. Mutat.* 34:1066-1070.
Rosenberg Alexander B, Patwardhan Rupali P, Shendure J, Seelig G. 2015. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* 163:698-711.
Savisaar R, Hurst LD. 2017. Estimating the prevalence of functional exonic splice regulatory information. *Hum. Genet.* 136:1059-1078.
Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* 28:1442-1454.
Sheikh TI, Mittal K, Willis MJ, Vincent JB. 2013. A synonymous change, p.Gly16Gly in MECP2 Exon 1, causes a cryptic splice event in a Rett syndrome patient. *Orphanet Journal of Rare Diseases* 8:108.
Shi M, Zhang H, Wang LT, Zhu CL, Sheng K, Du YH, Wang K, Dias A, Chen S, Whitman M, et al. 2015. Premature termination codons are recognized in the nucleus in a reading-frame-dependent manner. *Cell Discovery* 1:15001.
Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M. 1997. Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. *J. Clin. Invest.* 100:2204-2210.
Sperling J, Sperling R. 2008. Nuclear surveillance of RNA polymerase II transcripts. *RNA Biol* 5:220-224.
Stanford SM, Aleshin AE, Zhang V, Ardecky RJ, Hedrick MP, Zou J, Ganji SR, Bliss MR, Yamamoto F, Bobkov AA, et al. 2017. Diabetes reversal by inhibition of the low-molecular-weight tyrosine phosphatase. *Nat. Chem. Biol.* 13:624-632.
Stasia MJ, Bordigoni P, Floret D, Brion JP, Bost-Bru C, Michel G, Gatel P, Durant-Vital D, Voelckel MA, Li XJ, et al. 2005. Characterization of six novel mutations in the CYBB gene leading to different sub-types of X-linked chronic granulomatous disease. *Hum. Genet.* 116:72-82.
Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* 21:1563-1571.
Taipale J, Beachy PA. 2001. The Hedgehog and Wnt signalling pathways in cancer. *Nature* 411:349-354.

Team RC. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 4.3.2. Vienna, Austria: R Foundation for Statistical Computing.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68-74.

The Fantom Consortium, the RP, Clst, Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, Lassmann T, et al. 2014. A promoter-level mammalian expression atlas. *Nature* 507:462-470.

Valentine CR. 1998. The association of nonsense codons with exon skipping. *Mutat. Res.* 411:87-117.

Valentine CR, Heflich RH. 1997. The association of nonsense mutation with exon-skipping in hprt mRNA of Chinese hamster ovary cells results from an artifact of RT-PCR. *RNA* 3:660-676.

Vuoristo MM, Pappas JG, Jansen V, Ala-Kokko L. 2004. A stop codon mutation in COL11A2 induces exon skipping and leads to non-ocular Stickler syndrome. *Am. J. Med. Genet. A* 130A:160-164.

Wang J, Chang YF, Hamilton JI, Wilkinson MF. 2002a. Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol. Cell* 10:951-957.

Wang J, Gudikote JP, Olivas OR, Wilkinson MF. 2002b. Boundary-independent polar nonsense-mediated decay. *EMBO reports* 3:274-279.

Wang J, Hamilton JI, Carter MS, Li S, Wilkinson MF. 2002c. Alternatively spliced TCR mRNA induced by disruption of reading frame. *Science* 297:108-110.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* 119:831-845.

Welch EM, Barton ER, Zhuo J, Tomizawa Y, Friesen WJ, Trifillis P, Paushkin S, Patel M, Trotta CR, Hwang S, et al. 2007. PTC124 targets genetic disorders caused by nonsense mutations. *Nature* 447-:87.

Wimmer K, Eckart M, Stadler PF, Rehder H, Fonatsch C. 2000. Three different premature stop codons lead to skipping of exon 7 in neurofibromatosis type I patients. *Hum. Mutat.* 16:90-91.

Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. *Genome Biol* 11:R20.

Wu X, Hurst LD. 2016. Determinants of the Usage of Splice-Associated cis-Motifs Predict the Distribution of Human Pathogenic SNPs. *Mol. Biol. Evol.* 33:518-529.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806.

Xu W, Yang X, Hu X, Li S. 2014. Fifty-four novel mutations in the NF1 gene and integrated analyses of the mutations that modulate splicing. *Int. J. Mol. Med.* 34:53-60.

Yang N, MacArthur DG, Gulbin JP, Hahn AG, Beggs AH, Easteal S, North K. 2003. ACTN3 genotype is associated with human elite athletic performance. *Am. J. Hum. Genet.* 73:627-631.

Yeo G, Burge CB. 2004. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* 11:377-394.

Yukihara M, Ito K, Tanoue O, Goto K, Matsushita T, Matsumoto Y, Masuda M, Kimura S, Ueoka R. 2011. Effective drug delivery system for duchenne

muscular dystrophy using hybrid liposomes including gentamicin along with reduced toxicity. *Biol Pharm Bull* 34:712-716.

Zatkova A, Messiaen L, Vandenbroucke I, Wieser R, Fonatsch C, Krainer AR, Wimmer K. 2004. Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1. *Hum. Mutat.* 24:491-501.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46:D754-D761.

Zhan T, Rindtorff N, Boutros M. 2016. Wnt signaling in cancer. *Oncogene* 36:1461-1473.

Zhang X, Abreu JG, Yokota C, MacDonald BT, Singh S, Coburn KLA, Cheong S-M, Zhang MM, Ye Q-Z, Hang HC, et al. 2012. Tiki1 is required for head formation via Wnt cleavage-oxidation and inactivation. *Cell* 149:1565-1577.

Supplement to Chapter 3

The Supplementary Spreadsheets for Chapter 3 can be found on the attached CD.

Supplementary Texts

Supplementary Text 1: Defining the large-effect threshold

We find many differences in PSI (ΔPSI) small, with differences in the overall degree to which a particular exon is skipped unlikely to have any phenotypically meaningful impact. In order to filter cases where there is likely to be an effect, we need to apply a lower bound ΔPSI threshold for which cases with differences above the threshold are considered to be meaningful in terms of exon skipping. A 5% difference threshold was chosen. Note, the result of large effect cases being consistent with increased exon skipping if the PTC is present (by one-tailed exact Binomial test) is robust to threshold choice until reaching a lower-limit threshold of ≈ 0.7% (Supplementary Figure 6). Further, the significance of the result at 5% lies close to the threshold with strongest significance and minimal P-values (5.5% - 6.2%).

Thus, despite the threshold being arbitrarily defined, a significant P-value in the direction consistent with NAS is not simply an artefact of limiting results to a high threshold generating a significant result. Second, this threshold eliminates smaller-effect cases that when included contribute to a significant result in the direction consistent with NAS. By setting the threshold at 5% as a lower bound, we also the include all cases where the effect is strongest at slightly higher thresholds (5.5% - 6.2%) which would be excluded if set even 1.5% higher. Therefore, although user-defined, the 5% threshold is therefore appropriate for identifying cases whereby the effects of the PTC on exon skipping are likely to have important effects.

Supplementary Text 2: NAS is not an adaptive mechanism to save transcripts from NMD, whilst exons susceptible to NAS appear to be unremarkable

If exon skipping is an adaptive mechanism to save transcripts from NMD, we might expect the large effect exons to be more frequently of length three the exons for the other PTCs. We find this not to be the case when sampling 30 random exons from the non-large-effect PTCs 10,000 times ($P \approx 0.186$, one-tailed empirical P-value). Equally, if adaptive, we might expect the large effect cases to occur in shorter exons so if skipped it would have a smaller impact on the resulting protein, but we find no significant difference when comparing the median length with median length in the simulation exon sets ($P \approx 0.662$, one-tailed empirical P-value).

Additional examples of possible exceptionalism all suggest these exons have no particular defining characteristic. For example, one would expect ESEs to be more frequently disrupted by the set of PTCs in the large effect cases than those with little difference between PTC-/- and PTC-/+ variants. Using the INT3 ESE set (Caceres and Hurst 2013), we find 6/30 PTCs hit a motif that resembles an ESE, however this number of hits is not significantly more than expected when taking 10,000 random sets of 30 of the remaining PTCs and asking how many have an equal or greater number of ESE hits than the real hits ($P \approx 0.299$, one-tailed empirical P-value). This result, however, is dependent on ESEs being functional in these exons. An alternative approach is to look where in exons the PTCs are located. Are those in the prime set found more frequently in the ESE hotspot region? 23/30 are in the 3-69bp region, although this is not significantly different to the number found in the region when comparing with 10,000 randomly chosen sets of 30 of the remaining PTCs ($P \approx 0.425$, one-tailed empirical P-value). Neither do we find a bias for PTCs located at a particular end of the exon (15 and 15 at the 5' and 3' ends respectively), again not significantly different to 10,000 simulants ($P \approx 0.703$, one-tailed empirical P-value).

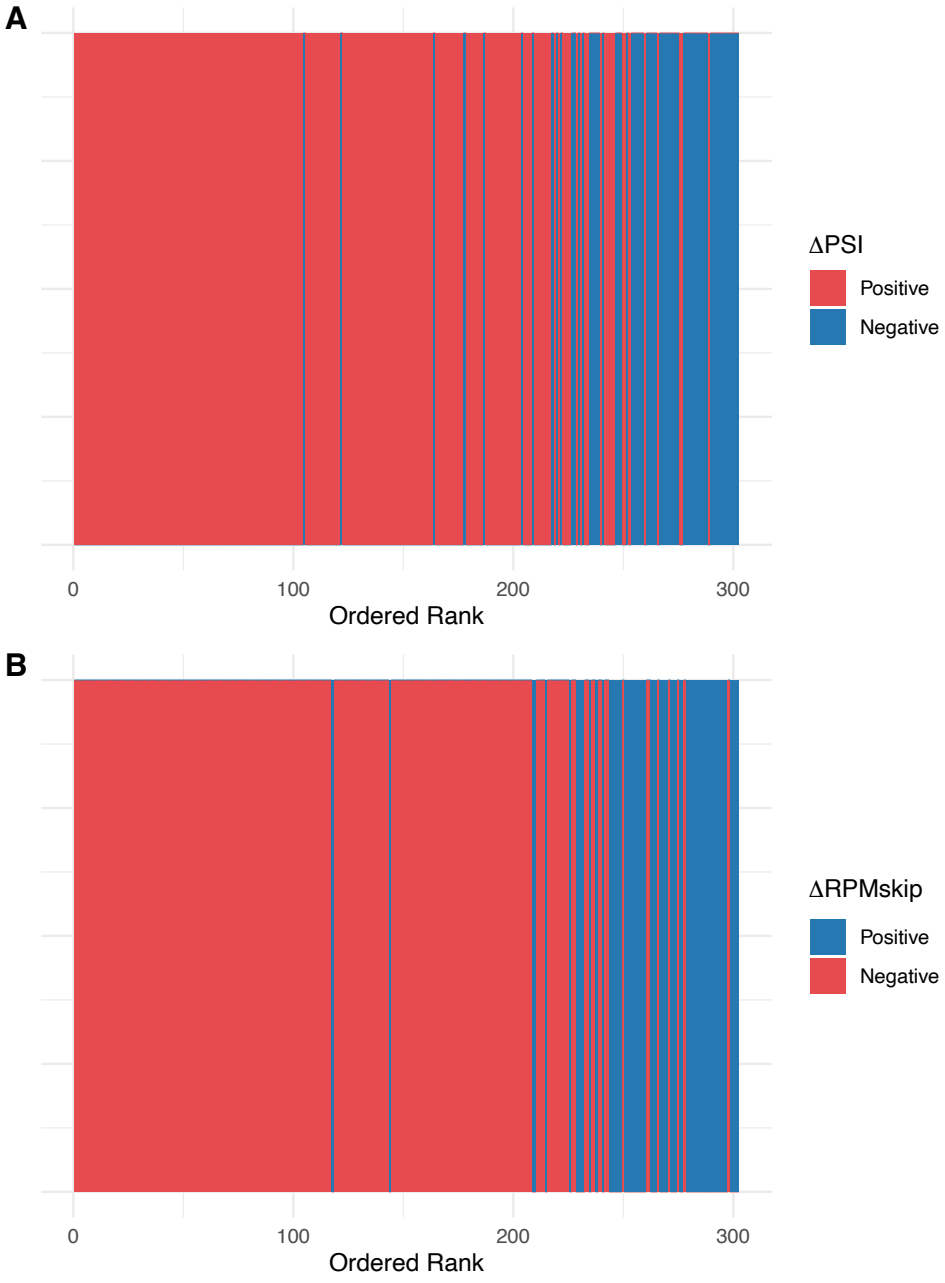
Thus, it appears that the exons in which we observe large increases in exon skipping when associated with the PTC are not exceptional. However, as a proportion of mutations are tolerated in genes that are either lowly expressed or are of less importance, it could be the case that these mutations simply occur in exons of lowly

expressed genes or those in which there is little phenotypic consequence. Indeed, the genes that contain PTCs tend to be more tissue specific than the genes containing pPTCs (see Supplementary Spreadsheet 7, Supplementary Spreadsheets).

References

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

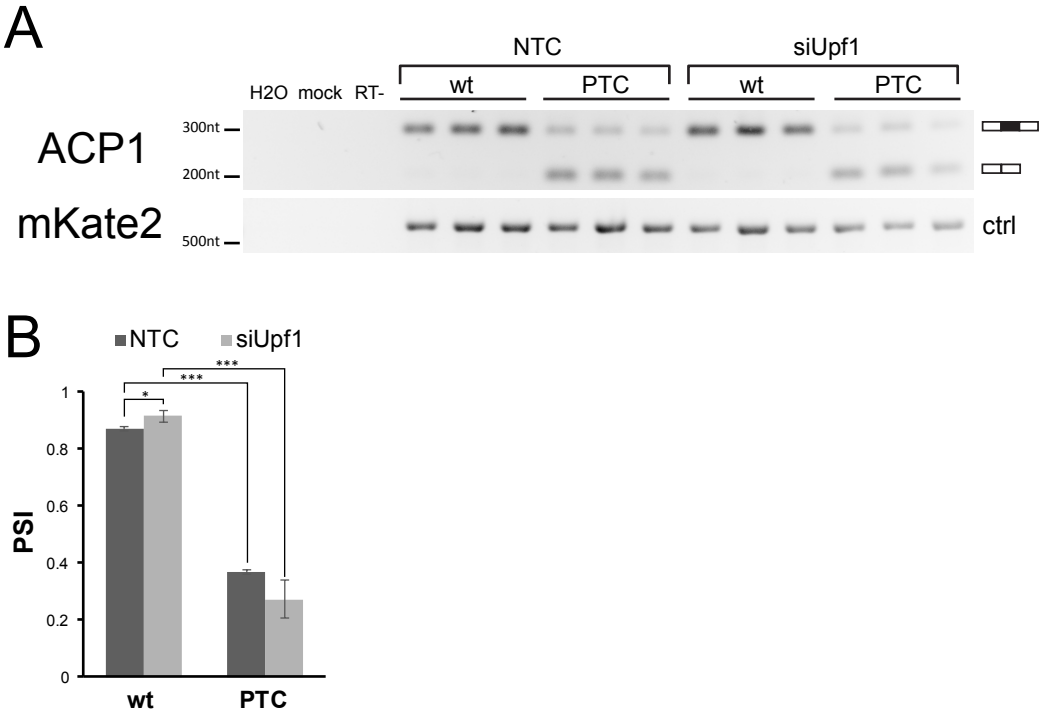
Supplementary Figures



Supplementary Figure 1: Ordered rank scores for ΔPSI and ΔRPMskip, with rankings shown in blue consistent with NAS and those shown in red in the opposite direction.

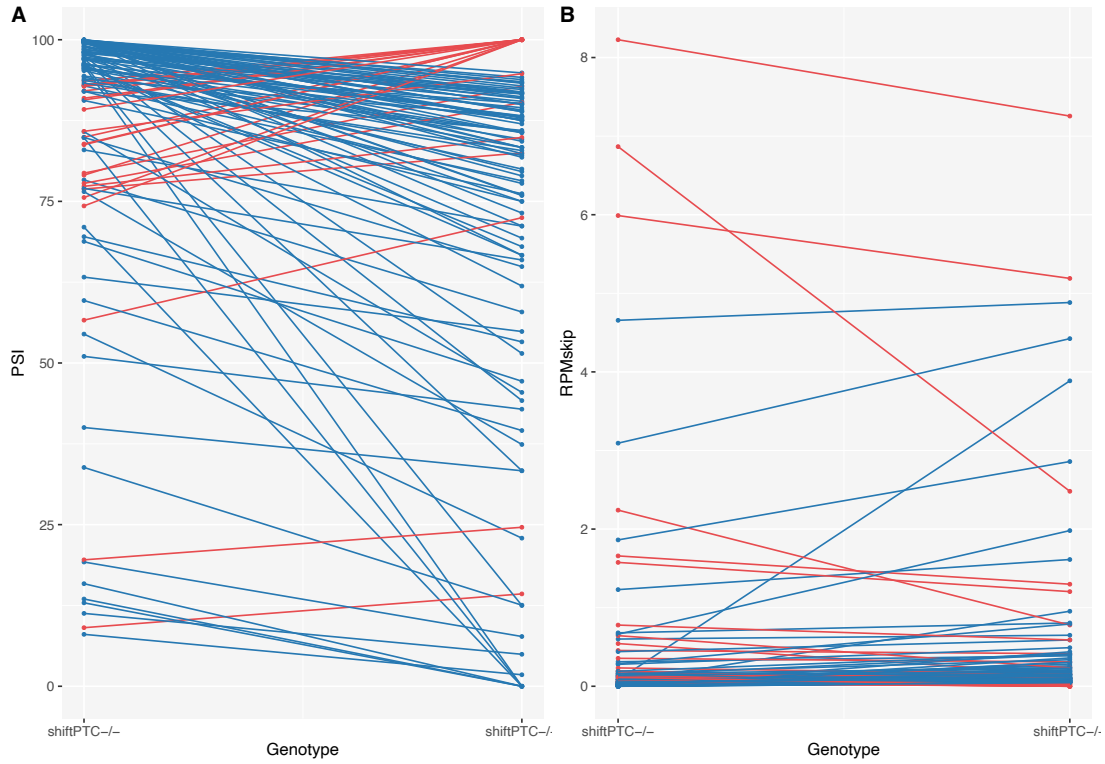
(A) Ordered ranks of ascending absolute ΔPSI scores for the N = 302 exons with ΔPSI not equal to zero. Exons with increased exon skipping consistent with NAS are distributed towards the higher ordered ranks. However, due to the relatively few data points (N = 68) the absolute

sum of ordered ranks (17,442) for ΔPSI < 0 scores is less than that for the larger group (N = 239) of ΔPSI > 0 scores (28,331), (B) Ordered ranks of ascending absolute ΔRPMskip scores for the N = 310 exons where ΔRPMskip is not equal to zero. Exons with ΔRPMskip consistent with NAS also rank high, but again the sum of absolute ranks for positive ΔRPMskip scores (16,704) is less than the sum of absolute ranks for negative ΔRPMskip scores (29,049).



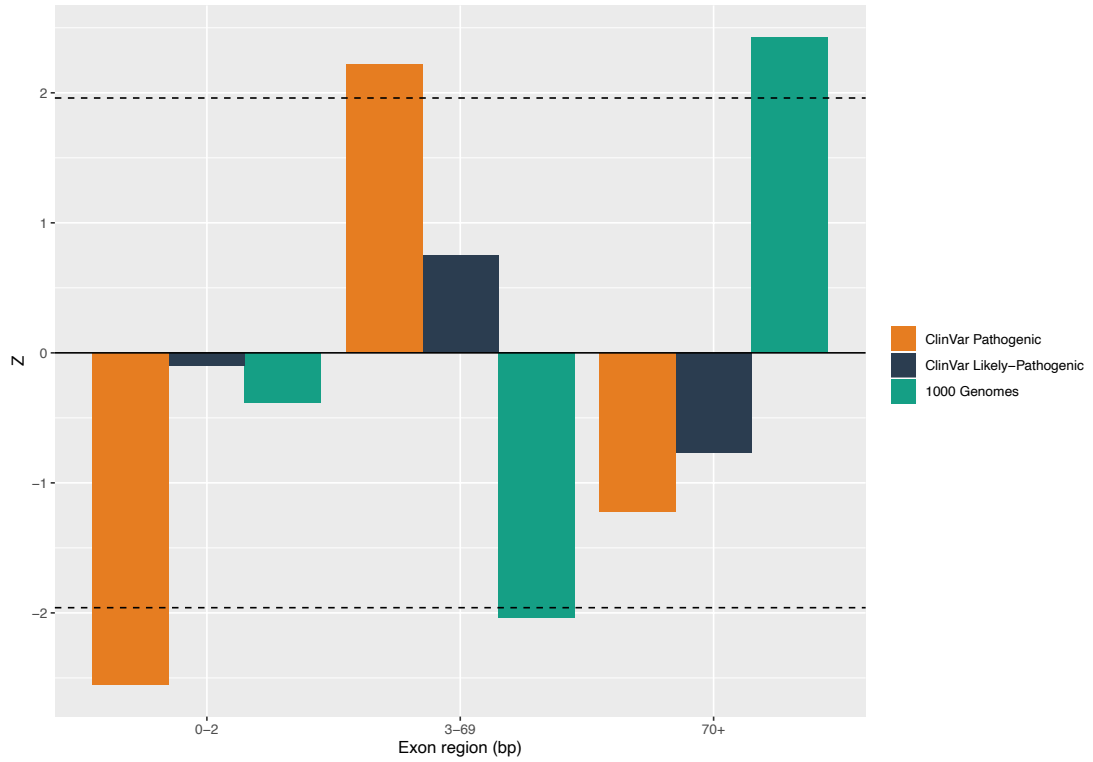
Supplementary Figure 2: Experimental expression of *ACP1* minigene constructs in Hek293T cells.

(A) Gel electrophoresis of the *ACP1* variants in cells with both the non-targeting siRNA pool control (NTC) and cells with Upf1 knockdown (siUpf1) in Hek293T cells. (B) PSI levels for wt and PTC-containing variants.



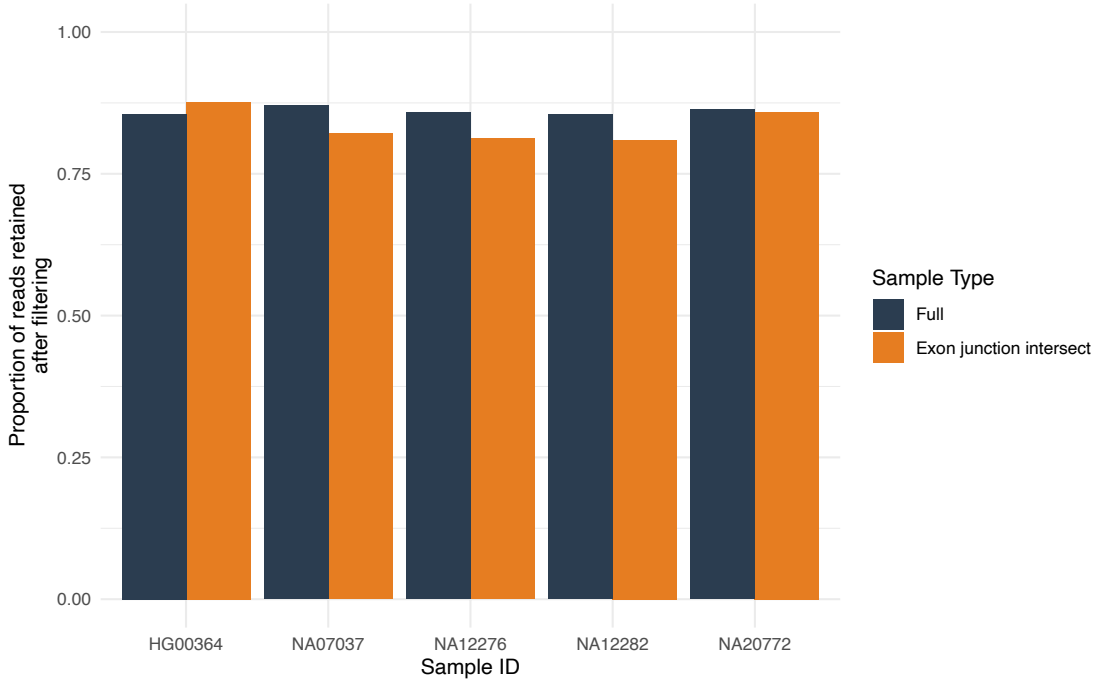
Supplementary Figure 3: Changes in PSI for large effect variants (>5%) for PTCs found off frame by one nucleotide.

(A) 94/116 of the large effect off-frame cases demonstrate a decrease in PSI for the PTC containing isoform, a significant number ($P = 4.328 \times 10^{-12}$, one-tailed exact Binomial test), (B) 93/116 of the shifted variants an increase in RPMskip associated with the shiftPTC, again a significant number ($P = 1.799 \times 10^{-11}$, one-tailed exact Binomial test). 56/120 large effect shifted PTCs have both PSI and RPMskip in the direction consistent with NAS, arguing against a reading frame dependant mechanism of skipping.



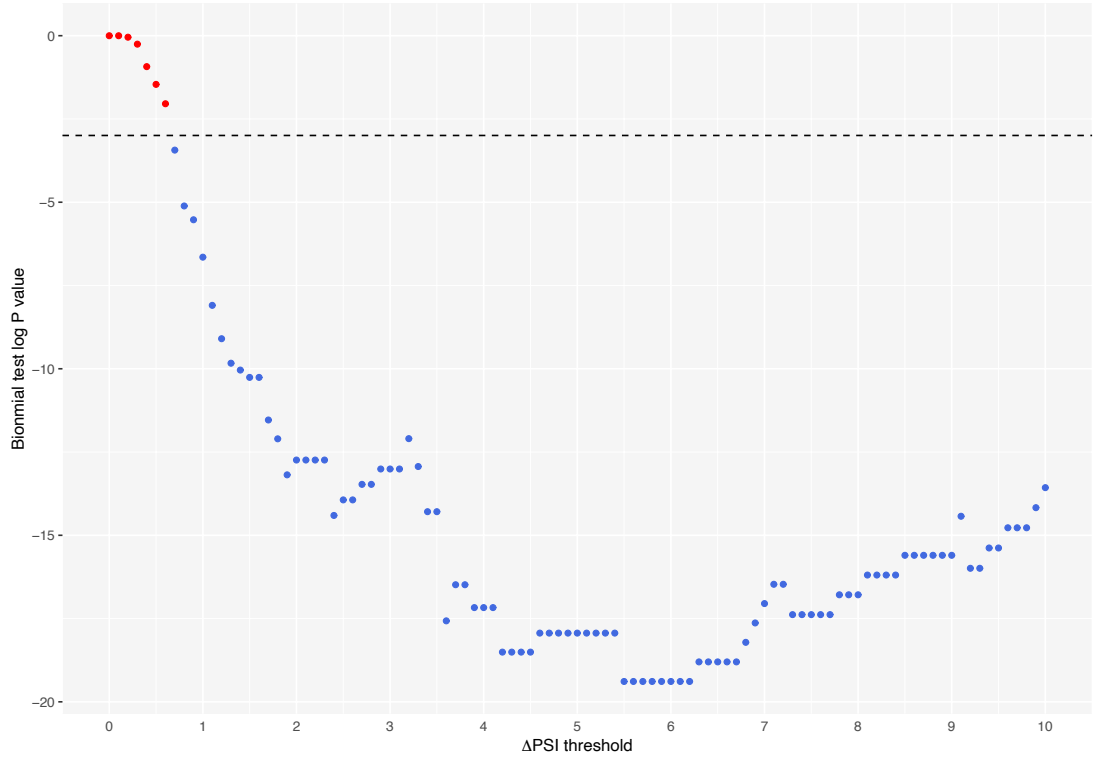
Supplementary Figure 4: Z scores for the number of nonsense mutations located in each exon region for the 1000 Genomes and ClinVar datasets when compared with randomly sampled nucleotide-matched simulants.

Z scores for the number of nonsense mutations located in each of the exonic regions when compared with 10,000 reference-allele nucleotide matched simulants for the pathogenic and likely-pathogenic ClinVar and the 1000 Genomes variants. The dotted line represents Z = ±1.96 where P ≈ 0.05. Only the pathogenic variants are significantly enriched in the exon flank regions consistent with splice disruption being a consistent source of disease. The significantly negative 1000 Genomes Z (-2.039) is also consistent with these mutations segregating in a healthy population.



Supplementary Figure 5: Proportion of reads retained after BAM filtering for both full samples and samples filtered to retain only exon junction reads.

The proportion of reads retained following quality filtering of the BAM files for 5 randomly sampled files. All files retain similar levels of reads after each filtering with differences not significant (P = 0.188, paired Wilcoxon signed-rank test), suggesting the approach used to approximate the number of reads retained following filtering of the intersect file can be appropriately applied to estimate the number of reads retained after filtering of the whole file.



Supplementary Figure 6: Determining the large-effect ΔPSI threshold

P values for one-tailed exact binomial tests asking whether the number of PTCs for showing ΔPSI above increasing ΔPSI threshold in the direction consistent with NAS is significant. Thresholds above ≈0.7% demonstrate this effect.

Chapter 4

Refining the Ambush Hypothesis: Evidence That GC- and AT-Rich Bacteria Employ Different Frameshift Defence Strategies

Liam Abrahams and Laurence D. Hurst

Genome Biology and Evolution (2018) 10(4):1153-1173

This chapter contains analysis of publicly available data. The data and custom scripts 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://academic.oup.com/journals/pages/access_purchase/rights_and_permissions/publication_rights).

This declaration concerns the article entitled:			
Refining the Ambush Hypothesis: Evidence That GC- and AT-Rich Bacteria Employ Different Frameshift Defence Strategies			
Publication status (tick one) **Draft manuscript** [ ] **Submitted** [ ] **In review** [ ] **Accepted** [ ] **Published** [X]			
Publication details (reference)	Liam Abrahams, Laurence D Hurst, Refining the Ambush Hypothesis: Evidence That GC- and AT-Rich Bacteria Employ Different Frameshift Defence Strategies, Genome Biology and Evolution, Volume 10, Issue 4, April 2018, Pages 1153–1173, https://doi.org/10.1093/gbe/evy075		
Copyright status (tick the appropriate statement) I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [X]			
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% Bioinformatics 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**	

Refining the Ambush Hypothesis: Evidence That GC- and AT-Rich Bacteria Employ Different Frameshift Defence Strategies

Liam Abrahams* and Laurence D. Hurst

Department of Biology and Biochemistry, The Milner Centre for Evolution, University of Bath, United Kingdom

*Corresponding author: E-mail: l.abrahams@bath.ac.uk.

Accepted: March 30, 2018

Data deposition: We have submitted all of the raw sequences used in this study to the Sequence Read Archive under the BioProject accession number PRJNA378178 to be released upon publication.

Abstract

Stop codons are frequently selected for beyond their regular termination function for error control. The "ambush hypothesis" proposes out-of-frame stop codons (OSCs) terminating frameshifted translations are selected for. Although early indirect evidence was partially supportive, recent evidence suggests OSC frequencies are not exceptional when considering underlying nucleotide content. However, prior null tests fail to control amino acid/codon usages or possible local mutational biases. We therefore return to the issue using bacterial genomes, considering several tests defining and testing against a null. We employ simulation approaches preserving amino acid order but shuffling synonymous codons or preserving codons while shuffling amino acid order. Additionally, we compare codon usage in amino acid pairs, where one codon can but the next, otherwise identical codon, cannot encode an OSC. OSC frequencies exceed expectations typically in AT-rich genomes, the +1 frame and for TGA/TAA but not TAG. With this complex evidence, simply rejecting or accepting the ambush hypothesis is not warranted. We propose a refined post hoc model, whereby AT-rich genomes have more accidental frameshifts, handled by RF2–RF3 complexes (associated with TGA/TAA) and are mostly +1 (or −2) slips. Supporting this, excesses positively correlate with in silico predicted frameshift probabilities. Thus, we propose a more viable framework, whereby genomes broadly adopt one of the two strategies to combat frameshifts: preventing frameshifting (GC-rich) or permitting frameshifts but minimizing impacts when most are caught early (AT-rich). Our refined framework holds promise yet some features, such as the bias of out-of-frame sense codons, remain unexplained.

Key words: out-of-frame stop codon, dual coding, sequence evolution, ambush hypothesis, frameshift.

Introduction

DNA sequences have the ability to carry multiple overlapping layers of noncoding, yet critical "dual-coding" information. Examples are widespread (Itzkovitz et al. 2010; Lin et al. 2011; Shabalina et al. 2013; Pancsa and Tompa 2016) often preventing or mitigating the cellular costs of transcriptional or translational errors (Drummond and Wilke 2009; Warnecke and Hurst 2011). The highly diverse nature of errors means signatures of dual-coding error control mechanisms are also varied. For instance, codon and amino acid usage is biased toward exon ends as purifying selection acts at synonymous and nonsynonymous sites of exonic splice enhancers (ESEs; Parmley et al. 2006,2007; Wu and Hurst 2015) to minimize mis-splicing rates (Blencowe 2000; Fairbrother et al. 2004; Wu et al. 2005; Caceres and Hurst 2013). Similarly, codon usage biases are thought to minimize translational missense errors (Drummond and Wilke 2008; Zhou et al. 2009; Serohijos et al. 2012), while synonymous and nonsynonymous site evolution in nucleosome linker sequences governs correct nucleosome positioning (Warnecke et al. 2008). Furthermore, synonymous codon selection surrounding micro-RNA (miRNA) binding sites ensures efficient miRNA binding (Gu et al. 2012).

Alternatively, avoiding particular sequences or motifs may be of equal importance. Selection acts to prevent mutations that cause inappropriate binding of RNA-binding proteins' binding within coding sequences (CDSs; Savisaar and Hurst 2017), to avoid intra-CDS Shine-Dalgarno (SD) motifs (Shine and Dalgarno 1974) that limit synthesis rates and promote incorrect folding inducing undesired frameshifting (Betney et al. 2010; Li et al. 2012; Diwan and Agashe 2016), or to avoid mononucleotide repeats or sequences prone to

ribosomal slippage (Ackermann and Chao 2006; Gurvich et al. 2005; Gu et al. 2010a).

Beyond their principle termination function, stop codons are repeatedly implicated in error control. In-frame stop codons located in introns are under selection (He et al. 1993; Jaillon et al. 2008; Farlow et al. 2010; Mekouar et al. 2010) to allow nonsense-mediated decay (NMD) to selectively degrade incorrectly spliced transcripts. In CDS regions where NMD is unable to operate, codons in close nucleotide space proximity to a stop codon are selectively avoided as a robustness to mistranscription errors (Cusack, et al. 2011). Stop codons found 5′ to recognized translation initiation sites increase protein activity, suggesting unwanted or incorrect translation initiations prior to the recognized start codon are terminated. (Seligmann 2007).

Despite selection to mitigate translational errors, the tradeoff between optimal decoding accuracy and translational speed (Wohlgemuth et al. 2010) permits ribosomal frameshifts errors, synthesizing peptides never intended. Robustness to such errors is thought to drive selection on transport RNA (tRNA) repertoires in genomes where frameshifts may be more costly (Warnecke et al. 2010) and may direct ribosome evolution (Atkins and Bjork 2009). Further, the ability to correct frameshift errors is thought to explain why three stop codons exist (Itzkovitz and Alon 2007). Out-of-frame stop codons (OSCs) prematurely terminate frameshifted translation events, minimizing process and cytotoxic costs associated with synthesizing an incorrect peptide from the incorrect reading frame (cellular resources, unproductive ribosomal demand, and toxic aggregation; Gingold and Pilpel 2011).

Recently, we identified a strong site-specific signature of selection for one OSC (Abrahams and Hurst 2017), finding a significant excess of A at CDSs fourth sites in nearly all bacterial genomes. Translation initiation on an ATG (and more generally, NTG) that becomes +1 out of frame thus encounters TGA, providing the potential ability for immediate ribosome correction. The "ambush hypothesis" (Seligmann and Pollock 2004), however, proposes that OSCs should be selectively favored throughout the gene body to reduce genome-wide frameshift costs. Several studies examine usage of codons that could, but don't necessarily, constitute an OSC and claim codon usage biases are consistent with such OSC selection (Seligmann and Pollock 2004; Singh and Pardasani 2009). However, with few genomes demonstrating biases (38.00%/6.23% of total genomes, 36.96%/7.07% of bacterial genomes for the two studies respectively), evidence is underwhelming. Moreover, these codon usage biases might be explained almost entirely by GC content (Morgens et al. 2013)—GC3 and GC1 content are the strongest determinants of OSC frequency in the +1 and +2 frames, respectively (Wong et al. 2008). Importantly, this method does not examine actual OSC frequencies. Thus, initial evidence supporting the ambush hypothesis is weak, speculative, and not robust to compositional controls to account for the high AT-content of stop codons.

An alternative approach compares real sequences with a distribution of null sequences simulating real CDSs, for which compositional biases can be controlled. Using Markov chain models, a remarkable 99.1% and 93.3% of prokaryotic genomes exhibit OSC excesses using second-order and fifth-order models that control for GC content and dinucleotide or pentanucleotide frequencies (Tse, et al. 2010), although numbers are reduced slightly for Morgens, et al. (2013) (83% and 85% respectively). Critically, these models directly interrogate OSC densities, although they do not preserve amino acid or codon usages.

While results from these models are consistent with OSCs exerting a near-universal selection pressure constraining CDS evolution, it is important to consider the wider biological context of these excesses. If the ambush hypothesis correctly predicts selection, prima facie it has been argued that selection to incorporate OSCs should be stronger in GC-rich genomes, as codon usage biases restrict chance dicodons yielding OSCs (note stop codons are AT-rich) (Tse et al. 2010; Morgens et al. 2013). Significant positive correlations between genome GC content and extent of excess suggests this is the case (Tse et al. 2010; Morgens et al. 2013). Yet, these excesses are attributable predominantly to TGA and not TAA or TAG (Morgens et al. 2013). Furthermore, out-of-frame sense TGN codons have similar, if not greater, number of genomes with excess and positive correlations with GC content (Morgens et al. 2013). These issues raise several potential caveats that may also apply to previous studies. First, when considered together, any excess may, for reasons unknown, only reflect TGA excesses, highlighting the need to consider each stop codon separately. Second, any excesses of OSCs might be an artifact of selection for codons with similar nucleotide composition and not selection directly for OSCs themselves, with OSC frequencies not exceeding expectations given underlying nucleotide composition.

The current status of the ambush hypothesis could therefore be considered as confused and uncertain with contradictory (i.e., some supportive and some unsupportive) evidence. Although the Markov models by Tse et al. (2010) and Morgens et al. (2013) improve on initial methods, are the results limited by the model design? As reported earlier, it is essential that GC content is controlled. Equally, as protein coding sequences are being simulated, the requirement for specific amino acids in specific orders might need to be retained. While the Markov models do provide some compositional bias control (GC content, higher order biases, e.g., dinucleotide frequencies), the stepwise addition of nucleotides does not preserve codon or amino acid identities, amino acid sequence ordering likely essential for protein function, nor small mutational or motif biases. Thus, the flexibility allowed by Markov models may not appropriately reflect real biological coding constraints that underpin OSC frequencies.

In this study, we therefore return to this issue concerning OSC selection. We first confirm previous results using Markov models (in part to ascertain whether our data set can mimic prior results). We then propose and test a series of simulation models that attempt to control for these compositional biases to varying degrees. While it is easy to criticize the Markov models, we acknowledge that our models also do not control completely for all competing selection pressures and biases.

In addition to the above mentioned problems, there is also the issue in quantifying deviation from null. We suppose a Z-score metric (deviation in standard deviation units) enables a more biologically valuable metric, as this enables us to quantify and compare excesses between models while accounting for genome variability. As +1 and −2 and +2 and −1 frameshifts incur equal costs (except for immediately at the start codon), for simulation models we consider only +1 and +2 frameshifts.

We find a complex pattern of results that provides neither a clear rejection nor acceptance of the ambush hypothesis. In this context, we motivate a post hoc refined version of the hypothesis, which broadly proposes that GC- and AT-rich genomes handle the problems associated with frameshifts differently, that +1 frameshifts are the dominant form of accidental slippage, and that frameshifts are predominantly resolved via a release factor (RF) 2/RF3 mechanism (which does not apply to TAG). In silico evidence supports the first tenet of the refined model, but we highlight several features that still defy clear explanation.

Materials and Methods

General Methods

All analyses were performed using custom Python 3.6 scripts with standard NumPy 1.8.0, SciPy 0.13, and Biopython 1.66 (Cock et al. 2009) libraries. Statistical analyses and data visualizations were performed using R 3.3.3 (R Core Team 2015). Scripts can be found at (https://github.com/la466/oscs).

Genome Downloads and Filtering

Whole-genome sequences for 3,860 bacterial genomes were downloaded from the European Molecular Biology Laboratory (EBML) database (http://www.ebi.ac.uk/Tools/dbfetch/emblfetch?db=embl, last accessed January 19, 2017). Genomes were filtered to include only one genome per genus larger than 500,000 base pairs (the remaining genomes were not considered in the analysis) in order to minimize any biases attributable to phylogenetic non-independence, leaving 694 genomes. Of these genomes, 690 use National Centre for Biotechnology Information (NCBI) translation tables 11 and 4 use NCBI translation table 4.

Coding Sequence Filtering

Each coding sequence was subjected to filtering in order to ensure the integrity of the sequences analyzed. Sequences were limited to those that contained a multiple of three nucleotides, contained only A, C, G, or T nucleotides, contained no in-frame stop codons, and had a correctly defined stop codon according to the NCBI translation table, TAA, TAG, or TGA for table 11 genomes or TAA or TAG for table 4 genomes.

General Modeling

All simulations were repeated 200 times for each bacterial genome. Increasing the number of simulations had minimal impact on OSC density variance (see supplementary fig. 1, Supplementary Material online, for an example of the variation in *Escherichia coli* OSC densities in the codon shuffle model). We define codon excesses using the standard Z score to compare how the real OSC densities differ beyond those expected by simulation between genomes while accounting for genome coding properties. *P* values were calculated by extrapolating directly from genome Z scores and corrected for multiple comparisons using the Benjamini-Hochberg False Discovery Rate (FDR) correction method, with one *P* value reported per genome. Where we report N/694 genomes with significant excesses, these are N different genomes with both genome $Z > 0$ and $P < 0.05$. OSC densities were calculated per 100 codons.

Markov Models

For each genome, we built Markov models similar to Tse et al. (2010) and Morgens et al. (2013). For each CDS in the genome, start and stop codons were discounted. For second-order models, the first two nucleotides of the remaining sequence and their position in the codon were defined. The third nucleotide, given the previous two nucleotides and their codon positions, was then sampled. After each sample, the two seed nucleotides and codon positions were shifted one nucleotide and resampled until all nucleotides in all CDSs had been accounted for. For fifth-order models, samples were based on the previous five nucleotides. Each real CDS was simulated using the start codon and two or five seed nucleotides using the transition probabilities previously calculated until the simulated sequence was of the same length as the real CDS minus the stop codon, which was then appended.

Codon Shuffle Model

For each CDS within the genome, the start and stop codons were removed. The codons of the CDS were isolated and randomly shuffled before being concatenated to form the simulated sequence.

Synonymous Site Model

For each genome, nucleotide frequencies at synonymous sites of codons within each coding block were calculated and normalized within coding blocks. In contrast to the

synonymous codon model, only synonyms within the same coding block were allowed to vary, and thus it is only the synonymous site that this model is questioning (e.g., serine AGC and AGT and TCA, TCC, TCG, and TCT are considered separately). Each codon in the real CDS had genome, amino acid, and coding block specific probabilities during simulation. For each CDS, each codon was in turn simulated using these coding probabilities.

Synonymous Codon Model

For each genome, codon frequencies were calculated and normalized as the probability of encoding an amino acid. Codons from multiple coding blocks that encode the same amino acid were considered together. For each CDS, each codon was in turn simulated using these probabilities. This test therefore asks whether CDSs using preferentially uses synonymous codons that generate OSCs.

Comparison between Table 11 and Table 4 Genomes

A local regression model (loess) for the specific codon and reading frame was fit between GC content and OSC density per 100 codons that included all table 11 and table 4 genomes in order to account for variation in GC content between the genomes. Residuals from this model for table 11 and table 4 genomes were then compared using Kruskal–Wallis tests. To increase the sample size, genomes of 89 additional table 4 genomes discarded during the original phylogenetic filter (irrespective of genome size) were considered for further comparison of OSC densities (see supplementary table 1, Supplementary Material online, for breakdown). These genomes were subjected to CDS filtering as before. We also restricted this table 4 genome data set by ranking *Mycoplasma* genomes by Z scores of +1 TGA for simulations using the synonymous site simulation and including only the nine genomes with highest Z score (matching the number of *Spiroplasma*, the next most common genus). Thus, this restriction should include only *Mycoplasma* genomes with the weakest negative TGA selection.

Calculating Frameshift Costs and Probabilities

Information regarding tRNA isoacceptor copy number and diversity was downloaded from the tRNADB-CE (Abe 2011; last accessed October 30, 2017). Of our 694 genomes, tRNA copy number and diversity information was available for 281 genomes. As in Warnecke et al. (2010), only genomes in which each codon could be decoded by the tRNA repertoire were considered, resulting in a final set of 231 genomes.

The "genomic cost of processing model" (Warnecke et al. 2010, equation 1) was used to calculated the cost of accidental frameshifting. This model is nested to allow the calculation of the probability of individual codons frameshifting using equation 2 (Warnecke et al. 2010). We inherit the assumption that tRNA copy numbers are reasonable proxies for cellular tRNA concentrations (Dong et al. 1996; Kanaya et al. 1999; Cognat et al. 2008). Further, anticodon–codon matching strategies were derived using the Supplementary Methods from Warnecke et al. (2010) originally proposed by Grosjean et al. (2010).

Codon Adaptation Index Calculations

Bacterial codon use is nonrandom. Highly expressed genes often prefer to use codons that are decoded by the most abundant tRNA (Rocha 2004). The Codon Adaptation Index (CAI) (Sharp and Li 1987) quantifies codon bias with high CAI values correlating with high expression in several organisms including *E coli* (dos Reis et al. 2003). CAI is therefore used as a gene expression proxy.

For each genome, a reference set of 20 genes from *rplA/1—rplF/6*, *rplI/9—rplU/21* and *rpsB/2—rpsU/21* were identified as highly expressed. The first 30 nucleotides were removed from the CDS (the 5′ CDS is biased to facilitate ribosome binding), and the first half of the CDS in this highly expressed set was used to calculate CAI indices using CodonW v1.4.4 (https://sourceforge.net/projects/codonw/; last accessed March 22, 2016) with the arguments "-coa_cu -coa_num 100%" to include all sequences in calculating indices. CAI values for the first half (minus the first 30 nucleotides) of the remaining CDS in the genome were calculated with the "-all_indices" argument using the generated fop_file, cai_file, and cbi_file. OSC densities were subsequently calculated using the second half of the CDS to prevent resampling of the same sequence for two measures for which codon usage is being measured and maximizing the independence of the data.

Results

Markov Models Replicate Prior OSC Excesses

To establish that our set of genomes is comparable with prior efforts, we first simulated sequences using Markov models in order to replicate prior results. Results demonstrate similar distributions of excesses to Morgens et al. (2013) (supplementary result 1, Supplementary Material online). The conclusions of prior results are repeatable, not consistent with ambush hypothesis predictions and that our sample of genomes are able to mimic prior efforts. Further discrepancies are therefore unlikely to be owing to the employment of a different set of genomes.

Genomes with Significant OSC Excesses Are Predominantly AT-Rich in a Model in Which Real Codon Combinations Are Shuffled

It is potentially important that the amino acid content of the protein coding sequences is maintained during simulations. Assuming selection on nonsynonymous sites is stronger than on synonymous sites (Hurst 2009), the principle determinant

Table 1

The Number of Genomes with Significant Out-of-Frame Excesses in Alternative Reading Frames When Coding Sequences Have Been Simulated by Shuffling the Codons within the Coding Sequence. Spearman's rank correlations between genome GC content and OSC excess, defined by the standard Z score, are also shown.

Codon	Reading Frame	# With Excess	% With Excess	ρ	P
All stops	Both	124	17.88	−0.178	2.328×10^{-6}
All stops	+1	367	52.88	−0.295	2.664×10^{-15}
All stops	+2	101	14.55	−0.144	1.489×10^{-4}
TAA	Both	98	14.12	−0.427	$<2.2 \times 10^{-16}$
TAC	Both	168	24.21	−0.113	0.003
TAG	Both	118	17.00	−0.352	$<2.2 \times 10^{-16}$
TAT	Both	186	26.80	−0.343	$<2.2 \times 10^{-16}$
TGA	Both	353	50.86	−0.091	0.017
TGC	Both	599	86.31	0.498	$<2.2 \times 10^{-16}$
TGG	Both	281	40.49	−0.431	$<2.2 \times 10^{-16}$
TGT	Both	165	23.78	−0.308	1.436×10^{-16}
TAA	+1	296	42.65	−0.417	$<2.2 \times 10^{-16}$
TAC	+1	361	52.02	0.572	$<2.2 \times 10^{-16}$
TAG	+1	190	27.38	−0.385	$<2.2 \times 10^{-16}$
TAT	+1	391	56.34	0.408	$<2.2 \times 10^{-16}$
TGA	+1	370	53.31	0.036	0.348
TGC	+1	575	82.85	0.465	$<2.2 \times 10^{-16}$
TGG	+1	256	36.89	−0.406	$<2.2 \times 10^{-16}$
TGT	+1	52	7.49	−0.063	0.099
TAA	+2	80	11.53	−0.231	8.587×10^{-10}
TAC	+2	148	21.33	−0.404	$<2.2 \times 10^{-16}$
TAG	+2	44	6.34	−0.178	2.336×10^{-6}
TAT	+2	176	25.36	−0.471	$<2.2 \times 10^{-16}$
TGA	+2	344	49.57	−0.169	7.508×10^{-6}
TGC	+2	531	76.51	0.233	5.600×10^{-10}
TGG	+2	299	43.08	−0.206	4.950×10^{-8}
TGT	+2	362	52.16	−0.352	$<2.2 \times 10^{-16}$

of any codon is likely the amino acid it encodes. However, not all sense codons can yield an OSC; in order to generate an OSC, two conducive codons must combine in the correct order. A proportion of OSCs will be incorporated irrespective of OSC selection, given some chance dicodon pairs always yield an OSC. For example, any A-starting codon following a methionine codon generates A +1 TGA. Can the OSC frequency be explained by random (no selection for OSCs) dicodon pairings? To test this hypothesis, we randomized codon order within each CDS to disrupt codon combinations that generate OSCs. This simulation controls for GC content exactly while preserving exact amino and codon identities and interactions between codon second and third sites. Amino acid order is not constrained.

We find that 124/694 (17.88%) of genomes have a significant excess of OSCs after randomization ($P < 0.05$, false discovery rate [FDR] correction), much reduced when compared with the Markov models both here and in the previous studies (Tse et al. 2010; Morgens et al. 2013). When each reading frame is considered independently, 367/694 (52.88%, $P < 0.05$, FDR correction) genomes have significant excesses in the +1 frame but many fewer, 101/694 (14.55%, $P < 0.05$, FDR correction) genomes, have significant excess in the +2 frame.

While this evidence is suggestive of OSC selection in the +1 frame in some genomes, several unexpected features are notable. First, correlations between GC content and OSC excesses are significantly negative (Table 1). As post-frameshift runs are longer in GC-rich genomes, the opposite correlation might have been a more obvious prediction (and previously employed as a prediction by Tse et al. 2010 and Morgens et al. 2013). Second, we observe many genomes with significant negative excesses of OSCs (fig. 1), suggesting selection for OSCs is not ubiquitous and often avoided. Furthermore, positive excesses are predominantly limited to the +1 reading frame (fig. 1). Whether this reflects a possible preponderance and susceptibility to +1 frameshift events is unknown.

Excesses of OSCs are also not uniformly distributed between the three stop codons. Only TGA has excesses in over 50% of genomes for any reading frame. This is also perhaps unexpected as TGA is thought to be the weakest of the stop codons (Povolotskaya et al. 2012; Korkmaz

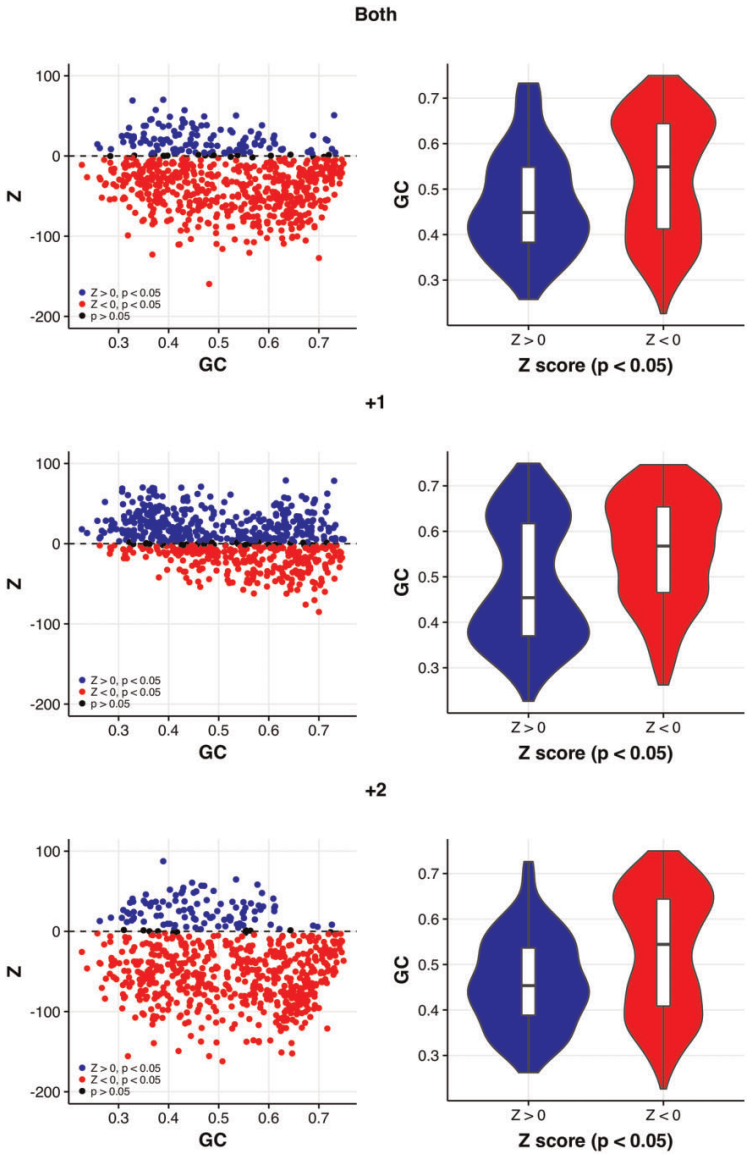


FIG. 1.—Correlations between GC content and out-of-frame stop codon excess ($Z > 0$), when all stop codons are considered together, are significantly negative in each reading frame for coding sequences simulated by random codon shuffling within the CDS. Violin plots emphasize that excesses are biased toward AT-rich genomes.

et al. 2014; Wei et al. 2016). TAA and TAG are often preferred and TGA avoided in highly expressed genes (Wei et al. 2016) while replacing TGA abolishes termination readthrough (Meng et al. 1995), implicating TGA as the least efficient terminator. A TGA preference was also observed by Morgens et al. (2013).

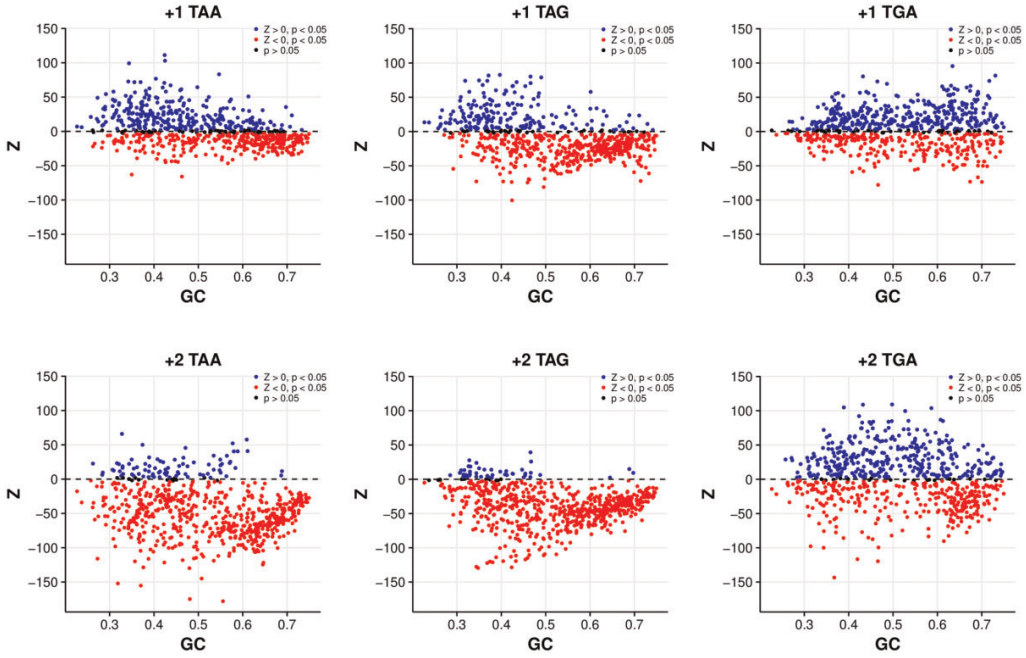


Fig. 2.—Correlations between GC content and genome excess of out-of-frame stop codons ($Z > 0$) are significantly negative ($P < 0.01$, Spearman's rank correlation) for all stop codons, in both reading frames, except for +1 TGA ($P = 0.348$) for the codon shuffle model. Excesses of TAA and TAG are heavily biased toward AT-rich genomes.

Genomes with significant excesses tend to be AT-rich, although significant TGA excesses do extend to some extremely GC-rich genomes, particularly in the +1 frame (fig. 2, Supplementary fig. 2, Supplementary Material online). Intriguingly, excesses of TAA and TAG are more highly restricted to AT-rich genomes, despite the identical GC content of TAG and TGA.

The observations of an excess of OSCs in some genomes in itself need not be evidence for selection for OSCs. Under the ambush hypothesis, we might also expect stronger selection for OSCs when compared with sense codons of similar nucleotide composition (Morgens et al. 2013). However, both TAC and TAT have a greater number of genomes with excesses when compared with TAA or TAG in both reading frames and excesses have significant positive correlations with GC content (Table 1). Excesses of +1 TGC have the strongest correlation and occur in the greatest percentage of genomes when compared with other TGN codons. By contrast, the number of genomes with excesses is greater for TGA than for either TGG or TGT in the +1 frame although only TGG in the +2 frame. Thus, as suggested by Morgens et al. (2013), OSC excesses may simply reflect complex compositional requirements resulting in an overrepresentation of out-of-frame TAN or TGN codons as opposed to selection for OSCs themselves.

OSC Excesses Are Also Seen in a Null Model Where Synonymous Sites Are Randomized

The above mentioned model provided some evidence for an excess of OSCs, especially in AT-rich genomes, although this evidence is by no means unambiguous. There are, however, limitations with the form of the null model used above. Disruptive changes to amino acid sequences would fundamentally alter protein function and not be permitted during sequence evolution. Such disruption would also break up larger motifs. Similar to the Markov models, this model cannot account for site-specific amino acid selection. Indeed, changes to sensitive amino acids can induce conformational changes in protein structure, altering protein stability or robustness to mutational errors (Yutani et al. 1977, Hormoz 2013) and are therefore essential to protein function. Moreover, amino acids that may carry site-specific functional information, for example, the second amino acid that is under

Table 2

The Number of Genomes with Significant Out-of-Frame Excesses for Different Codons When Coding Sequences Have Been Simulated by Randomizing Synonymous Sites within Coding Blocks. Spearman's rank correlations between genome GC content and OSC excess, defined by the standard Z score, are also shown

Codon	Reading Frame	# With Excess	% With Excess	ρ	P
All stops	Both	87	12.54	−0.444	$<2.2 \times 10^{-16}$
All stops	+1	272	39.19	−0.443	$<2.2 \times 10^{-16}$
All stops	+2	103	14.84	−0.260	4.046×10^{-12}
TAA	Both	118	17.00	−0.508	$<2.2 \times 10^{-16}$
TAC	Both	145	20.89	−0.067	0.077
TAG	Both	101	14.55	−0.282	4.371×10^{-14}
TAT	Both	194	27.95	−0.382	$<2.2 \times 10^{-16}$
TGA	Both	288	41.50	−0.326	$< 2.2 \times 10^{-16}$
TGC	Both	636	91.64	0.589	$<2.2 \times 10^{-16}$
TGG	Both	265	38.18	−0.404	$<2.2 \times 10^{-16}$
TGT	Both	252	36.31	−0.403	$<2.2 \times 10^{-16}$
TAA	+1	298	42.94	−0.444	$<2.2 \times 10^{-16}$
TAC	+1	330	47.55	0.595	$<2.2 \times 10^{-16}$
TAG	+1	155	22.33	−0.334	$<2.2 \times 10^{-16}$
TAT	+1	439	63.26	0.403	$<2.2 \times 10^{-16}$
TGA	+1	256	36.89	−0.135	3.729×10^{-4}
TGC	+1	599	86.31	0.625	$<2.2 \times 10^{-16}$
TGG	+1	271	39.05	−0.365	$<2.2 \times 10^{-16}$
TGT	+1	98	14.12	−0.218	7.287×10^{-9}
TAA	+2	93	13.40	−0.321	$<2.2 \times 10^{-16}$
TAC	+2	146	21.04	−0.389	$<2.2 \times 10^{-16}$
TAG	+2	42	6.05	−0.140	2.270×10^{-4}
TAT	+2	185	26.66	−0.500	$<2.2 \times 10^{-16}$
TGA	+2	365	52.59	−0.261	3.777×10^{-12}
TGC	+2	557	80.26	0.178	2.523×10^{-6}
TGG	+2	271	39.05	−0.214	1.386×10^{-8}
TGT	+2	384	55.33	−0.409	$<2.2 \times 10^{-16}$

strong selection to promote methionine cleavage (Liao et al. 2004; Frottin et al. 2006; Ouidir, et al. 2015), are not retained.

A possibly more realistic scenario might be strong selection for synonymous mutations that generate OSCs. To consider this, we simulated synonymous nucleotide frequencies in accordance with genome codon usage frequencies preserving amino acid identities, amino acid order, and net genome codon usage frequencies. For these simulations, we permitted synonymous codon changes from strictly within the same codon block, i.e., codons from the 2-fold and 4-fold blocks of the three 6-fold degenerate amino acids were not interchanged. A similar but less stringent codon simulation model where this codon block restriction is relaxed (i.e., allowing the interchange of all members within 6-fold degenerate blocks) yields similar results (supplementary result 2, Supplementary Material online).

With higher level constraints controlled, if OSCs enforce a strong enough selection pressure, we expect a bias toward nucleotides generating OSCs if the following codon permits. For example, if the amino acid sequence dictates isoleucine-glutamic acid, we expect a bias toward ATA isoleucine codons to encode a +1 TAG. OSCs arising from 1-fold degenerates are not considered as synonymous site selection has no effect.

Perhaps significantly, much like the previous model, the number of genomes with significant excesses is low and predominantly in the +1 frame (272/694, 39.19%, $P < 0.05$, FDR correction) (table 2). The lack of excesses in the +2 frame is particularly surprising for this model, given T is strictly required at the synonymous site for OSCs. When all OSCs are considered together, excesses in each reading frame are significantly negatively correlated with GC content (table 2) and heavily biased toward AT-rich genomes (fig. 3).

This lack of significant excess extends to the individual OSCs. When both frames are considered together, TGA again demonstrates the greatest deviations from null sequences (288/694, 41.50%, $P < 0.05$, FDR correction). Excesses of TAA are lower (118/694, 17.00%, $P < 0.05$, FDR correction) and TAG lower still (101/694, 14.55%, $P < 0.05$, FDR correction). All OSC excesses are limited predominantly to AT-rich genomes (supplementary fig. 3, Supplementary Material online).

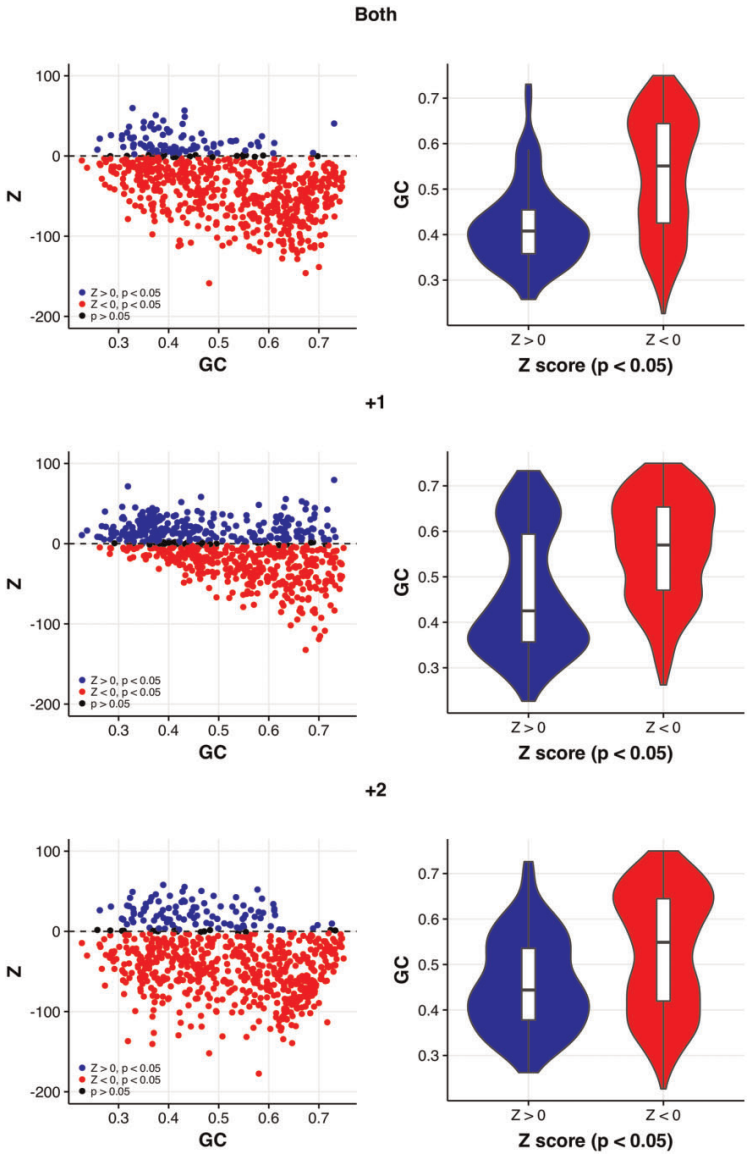


FIG. 3.—Correlations between GC content and out-of-frame stop codon excess ($Z > 0$), when all stop codons are considered together, are significantly negative ($P < 0.01$, Spearman's rank correlation) in each alternative reading frame for coding sequences where synonymous sites are randomized. Violin plots again emphasize a bias towards significant excesses in the AT-rich genomes.

Again, excesses appear more acute in the +1 frame. Unlike the previous model, +1 TAA is now the stop with the greatest number of genomes with excesses (298/694, 42.94%, $P < 0.05$, FDR correction) and greater than +1 TGA (256/694, 36.88%, $P < 0.05$, FDR correction). These +1 TAA excesses are highly restricted to the AT-rich genome and

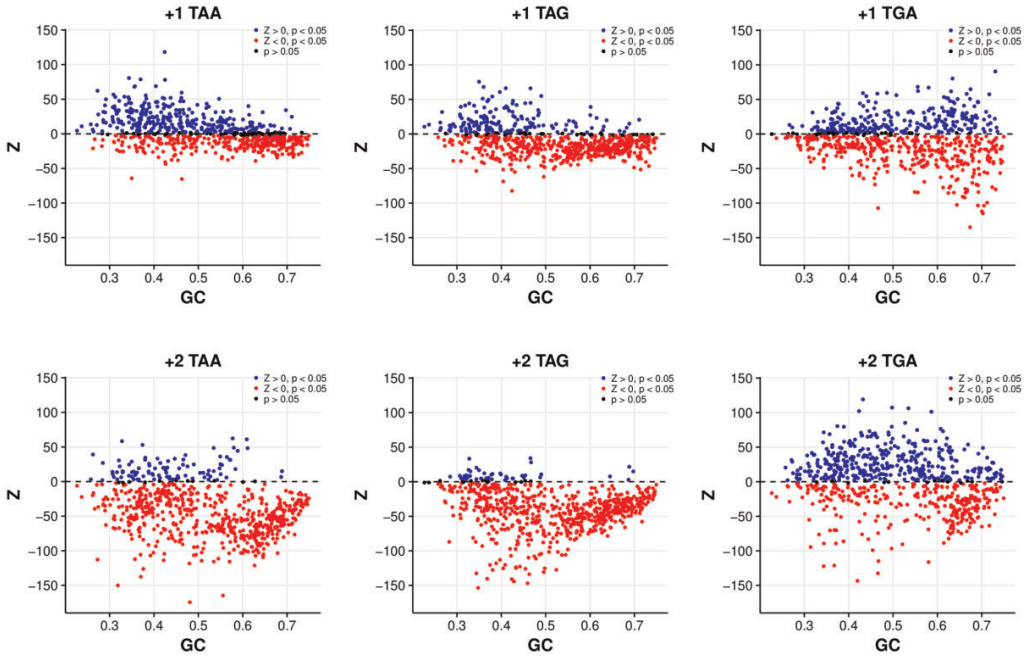


Fig. 4.—Correlations between GC content and genome excess of out-of-frame stop codons ($Z > 0$) are significantly negative ($P < 0.01$, Spearman's rank correlation) for all stop codons in both alternative reading frames for the synonymous site randomisation model. Excesses of TAA and TAG are heavily biased toward AT-rich genomes, with few genomes exhibiting excesses in the +2 frame.

more generally have a significant negative correlation with GC content ($\rho = -0.444$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) (fig. 4, supplementary fig. 3, Supplementary Material online). In contrast, the number of genomes with significant excesses of +1 TAG (155/694, 22.33%, $P < 0.05$, FDR correction), +2 TAA (93/694, 13.40%, $P < 0.05$, FDR correction), and +2 TAG (42/694, 6.05%, $P < 0.05$, FDR correction) are remarkably low. Thus, +1 seems to be the dominant signal, and signals for the most part are not associated with TAG.

It is again unclear whether the excesses reflect stop codon functionality. When compared with off-frame sense codons, both TAA and TAG have fewer genomes with significant excesses than either TAC or TAT. Excesses of TGC (+1: 599/694, 86.31%; +2: 557/694, 80.26%, $P < 0.05$, FDR correction) are the greatest of any TGN codon in either reading frame. Excesses of +1 TGG (271/694, 39.04%, $P < 0.05$, FDR correction) and +2 TGT (384/694, 55.33%, $P < 0.05$, FDR correction) are also greater than TGA in the respective frames.

+1 TAA Demonstrates Evidence of OSC Selection at Synonymous Sites for Amino Acid Repeats Whose Codons Present the Opportunity to Encode an OSC

Results of the above simulation, which is arguably the most realistic determination of the null model, are suggestive but come with caveats, given the excess of OSCs. However, this null model also has limitations. First, we have to make presumptions about the realism of synonymous site selection. For example, if there are subtle location-specific codon usage biases or context-dependent mutational biases, these are likely to overcome any selection for OSCs. The model does not respect differential codon usage biases throughout the CDS nor motif or domain-specific codon usage biases, for example, the bias toward A to disrupt messenger RNA (mRNA) stability at 5′ ends (Gu et al. 2010b; Kudla et al. 2009; Bentele et al. 2013). Furthermore, in assuming each synonymous site is under selection for OSCs, this model assumes selection pressures are of equal strength at all synonymous sites, which is unlikely to be the case.

Given the these issues, we propose a further test that might better control for amino acid order, codon usage

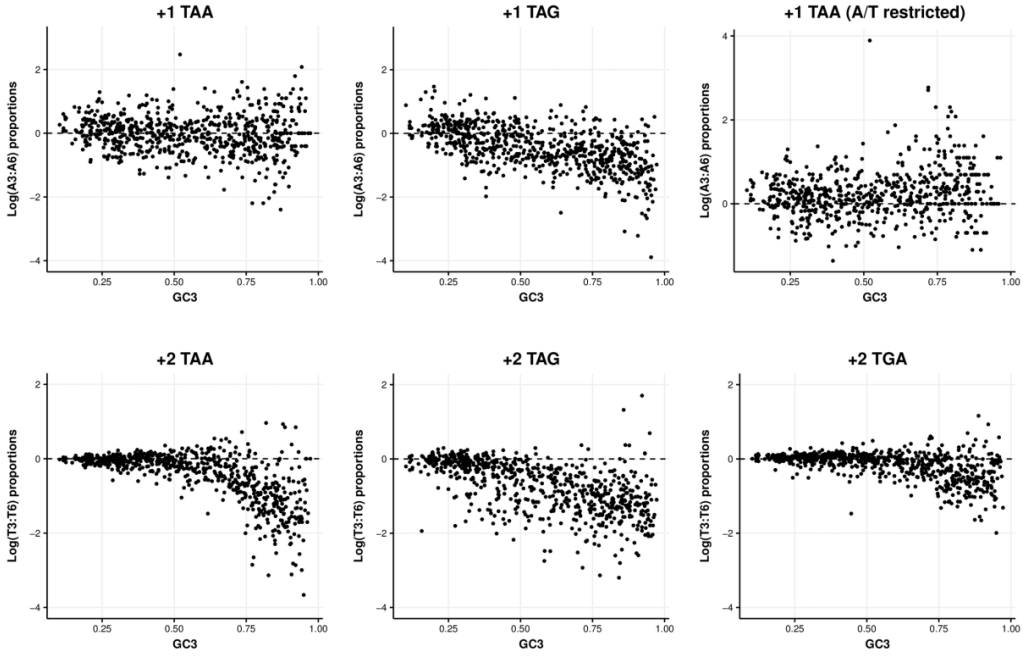


Fig. 5.—Log ratios between the A use at synonymous sites of amino acids whose codons when repeated can generate an OSC. Correlations are significantly negative in each case ($P < 0.05$, Spearman's rank correlations), suggesting A use at the third site decreases compared with the sixth, as GC mutational biases make encoding OSCs more difficult. When codons are restricted to only A/T ending synonyms, +1 TAA demonstrates a significant positive correlation with GC content ($\rho = 0.160$, $P = 4.827 \times 10^{-5}$, Spearman's rank correlation).

biases, and highly regionalized effects, but one that has a more limited sample size. We can ask whether the synonymous codons used in localized sequence contexts encode OSCs when given the opportunity. We isolated any repeat of two isoleucine (codons ATA, ATC, ATT) or valine (codon GTA, GTC, GTG, GTT) amino acids, followed by amino acids whose codon starts with either C or T. In this way, we isolate sequences in which the first codon always has the opportunity to yield an OSC, followed by a second codon, encoding an identical amino acid that strictly cannot. Any regionalized biases are thus minimized while ensuring the amino acid requirement and hence direction of codon usage bias remains identical. If OSC selection constrains codon choice, we predict a stronger bias toward A-ending synonyms for the first codon of the repeat than the second. For example, A use in the sequence 5′-ATH | ATH | YNN-3′ should be greater at site 3 than 6 to encode +1 TAA. ATG has no synonyms, and therefore +1 TGA cannot be examined. We perform paired tests between usage within each genome to control for intragenome localized mutational biases but also to negate effects of intergenome compositional biases. We cannot control the mutational bias (or motif selection) owing to interactions between sites 3 and 4 and sites 6 and 7, but otherwise all other context features are preserved.

Again, the signals are ambiguous. We find no significant difference between the use of A at sites 3 and 6 for +1 TAA encoding sequences ($P = 0.215$, paired Wilcoxon signed rank test). If synonymous sites are being selected for to preserve OSCs, we expect site 3 to be more resistant to mutational pressures than site 6. Thus, as GC3 content increases, we expect relatively little change in A3 but a reduction in A6 giving a positive correlation between A3: A6 and GC content. This is not the case—correlations are significantly negative for possible +1 TAA encoding sequences ($\rho = -0.097$, $P = 0.012$, Spearman's rank correlation) (fig. 5).

This negative correlation might imply that the uncontrolled mutation bias difference (A3: A4 versus A6: A7, difference) is not to be overlooked. However, for this test, GC3 content is not consistent and allows comparisons between ATA and ATC. When GC3 content is controlled by only considering codons using A/T at their synonymous site, A3 use is significantly greater than A6 use ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test, mean proportion of sequences with A: site

3 = 0.278; site 6 = 0.208). Individually, 475/694 (68.44%) genomes have greater A3 use. Furthermore, the correlation between GC3 and A3/A6 correlations is now significantly positive ($\rho = 0.160$, $P = 4.827 \times 10^{-5}$, Spearman's rank correlation). Thus, synonymous codon usage is consistent with +1 TAA selection after GC control.

We apply the same test to valine repeats that have the potential to encode +1 TAG. Unlike +1 TAA-encoding sequences, we find A3 use significantly reduced when all valine codons are considered ($P < 2.2 \times 10^{-16}$, paired Wilcoxon signed rank test, mean proportion of sequences with A: site 3 = 0.137; site 6 = 0.156) and when only GTA and GTT are considered ($P = 6.129 \times 10^{-5}$, paired Wilcoxon signed rank test, mean proportion of sequences with A: site 3 = 0.313; site 6 = 0.329). Correlations are significantly negative between GC3 content and A3: A6 usage in both cases (All codons: $\rho = -0.585$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation; GTA/GTT: $\rho = -0.143$, $P = 1.77 \times 10^{-4}$, Spearman's rank correlation).

Thus, it appears synonymous codon usage is consistent with OSC selection in the specific case of +1 TAA, although motif effects and subtle mutational biases are hard to eliminate as explanations. Employing similar tests for T use for all +2 OSC encoding sequences provides no evidence consistent with OSC selection, nor does a general hypothesis that considers all stop codons and frames together (supplementary result 3, Supplementary Material online).

+1 TGA Densities Are Significantly Reduced in Genomes Where TGA Does Not Function as a Stop Codon, However Both +1 TAA and +1 TAG Densities Are Also Reduced

Although our models present excesses of OSCs in some instances, can we attribute them to stop codon function? The excess of off-frame sense codons suggests that simply looking for an excess of OSCs may be naive. An alternative approach is to consider the subset of prokaryotes (*Entomoplasmatales* and *Mycoplasmatales*) in which TGA is recoded to tryptophan, eliminating stop functionality (Bove 1993). If excesses are due to termination functionality, any off-frame TGA selection should be weaker in these genomes. Further, if terminating frameshift events is of such cellular importance, this recoding should result in compensatory increases of TAA and TAG due to the impaired termination ability. We refer to recoded genomes as "table 4" genomes and those using the standard genetic code as "table 11" genomes using National Centre for Biotechnology Information (NCBI) naming convention. Indeed, there would appear to be weaker +1 TGA selection (supplementary fig. 4, Supplementary Material online) with most table 4 genomes demonstrating negative excesses in our simulations. It is, however, important to compare actual OSC frequencies between genomes using alternative translation tables. Any differences attributable to GC mutational biases (i.e., AT-rich table 4 genomes are likely to have increased OSC densities by chance) are minimized by performing loess regressions and comparing residuals between the two genetic codes.

The OSC densities of stop codons combined are significantly reduced for table 4 genomes when +1 and +2 frames are considered together ($P = 5.572 \times 10^{-4}$, Kruskal–Wallis rank sum test of residuals; table 4 mean residual (MR) = −5.487, table 11 MR = 0.046). Results are similar when reading frames are considered separately (+1: $P = 5.624 \times 10^{-4}$, Kruskal-Wallis rank sum test of residuals, table 4 MR = −2.617, table 11 MR = 0.029; +2: $P = 8.406 \times 10^{-4}$, Kruskal–Wallis rank sum test of residuals, table 4 MR = −2.870, table 11 MR = 0.017). This is not entirely unexpected—even if TAA and TAG are somewhat increased there may not be full compensation for the loss of TGA.

Are these reduced OSC densities attributable to loss of TGA stop functionality? Contrary to expectation, off-frame TGA densities are significantly increased in the +2 frame ($P = 0.002$, Kruskal–Wallis rank sum test of residuals; table 4 MR = 1.113, table 11 MR = −0.011) supporting the excesses in simulation models (supplementary fig. 4, Supplementary Material online). Despite reduced mean residuals, +1 TGA densities are not significantly reduced ($P = 0.125$, Kruskal–Wallis rank sum test of residuals; table 4 MR = −0.233, table 11 MR = −0.001). However, given negative excesses from simulation models (supplementary fig. 4, Supplementary Material online) and these reduced residuals, the lack of table 4 genomes may be limiting. To provide a richer data set, we therefore incorporated all table 4 genomes from our initial data set prior to phylogenetic filtering, increasing the table 4 sample to 93 genomes. We accept that this introduces a degree of nonindependence and bias by including many *Mycoplasmas* (see supplementary table 1, Supplementary Material online, for breakdown of genomes).

With this increased data set, combined OSC densities in table 4 genomes remain significantly reduced when +1 and +2 frames are considered together ($P < 2.2 \times 10^{-16}$, Kruskal–Wallis rank sum test of residuals; table 4 MR = −2.509, table 11 MR = 0.363), in the +1 frame ($P < 2.2 \times 10^{-16}$, Kruskal–Wallis rank sum test of residuals, table 4 MR = −1.171, table 11 MR = 0.176) and the +2 frame ($P < 2.2 \times 10^{-16}$, Kruskal–Wallis rank sum test of residuals, table 4 MR = −1.337, table 11 MR = 0.187) (fig. 6). Specifically, although +2 TGA use remains significantly increased ($P = 1.57 \times 10^{-9}$, Kruskal–Wallis rank sum test of residuals; table 4 MR = 0.328, table 11 MR = −0.055), +1 TGA densities are significantly reduced ($P = 2.091 \times 10^{-7}$, Kruskal–Wallis rank sum test of residuals; table 4 MR = −0.174, table 11 MR = 0.023). Thus, consistent with previous results, any selection for OSCs is likely to be operating predominantly in the +1 frame and +1 TGA use appears to be reduced in table 4 genomes.

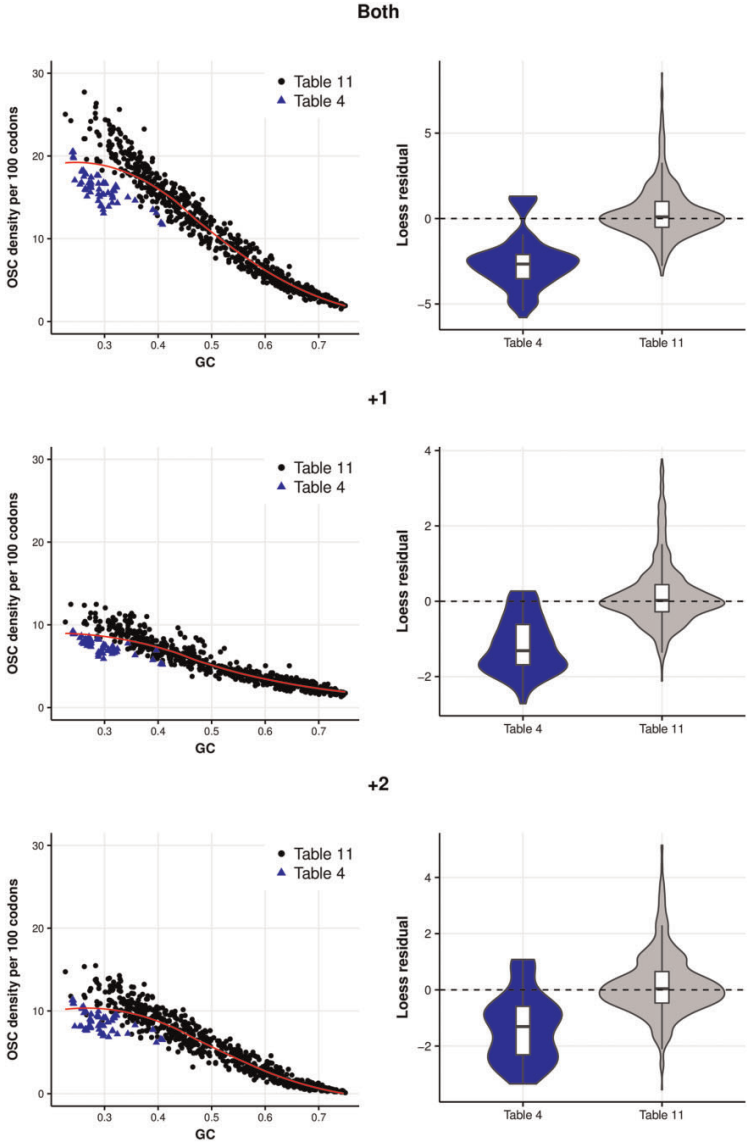


Fig. 6.—OSC densities are reduced in table 4 genomes when compared with table 11 genomes in each alternative reading frame. Violin plots of the loess regression residuals highlight the reduced residuals for OSC densities in table 4 genomes.

Without considering the context of this reduced excess, it is difficult to determine whether this is related to lost termination function. Are TAA and TAG densities increased to compensate? Results indicate this is not the case. Both +1 TAA ($P = 3.107 \times 10^{-6}$, Kruskal–Wallis rank sum test of residuals; table 4 MR = −0.149, table 11 MR = 0.028) and +1 TAG ($P = 4.355 \times 10^{-8}$, Kruskal-Wallis rank sum test of residuals; table 4 MR = −0.284, table 11 MR = 0.048) densities are

significantly reduced in table 4 genomes (fig. 7). When restricting table 4 genomes to include only 9 *Mycoplasma* genomes (matching the total for the next most common genus *Spiroplasma* to reduced bias, see Methods) in which selection against TGA should be weakest, we obtain similar results (supplementary table 2, Supplementary Material online).

Given densities of other OSCs are not increased, we ask whether off-frame TAN and TGN densities are more generally reduced. Both +1 TAC ($P = 1.14 \times 10^{-12}$, Kruskal-Wallis rank sum test of residuals; table 4 MR = −0.170, table 11 MR = 0.020) and +1 TAT ($P = 1.582 \times 10^{-13}$, Kruskal–Wallis rank sum test of residuals; table 4 MR = −0.277, table 11 MR = 0.038) densities are also significantly reduced in table 4 genomes. Results are similar using the restricted *Mycoplasma* data set (supplementary table 2, Supplementary Material online). Thus, reduced TAA and TAG densities may not be termination-function related but rather a consequence of weakened selection for alternative constraints that affects all off-frame TAN codons. Alternatively, table 4 genomes may not exploit OSCs as a frameshift termination mechanism to the same degree, given termination capacity is reduced. These reduced densities dismiss the notion of increased compensatory selection.

For TGN codons, while there is no significant difference between +1 TGC densities ($P = 0.101$, Kruskal–Wallis rank sum test of residuals) or +1 TGT densities ($P = 0.290$, Kruskal–Wallis rank sum test of residuals), +1 TGG densities are significantly reduced ($P = 0.003$, Kruskal–Wallis rank sum test of residuals; table 4 MR = −0.137, table 11 MR = 0.015). For +1 TGC and +TGT, results using the restricted *Mycoplasma* data set are similar (supplementary table 2, Supplementary Material online) although +1 TGG densities are not significantly different ($P = 0.257$, Kruskal-Wallis rank sum test of residuals). Unlike TAN codons, it would be difficult to conclude that reduced TGA densities are attributable to reduced TGN densities but rather toward possible reduced TGR densities or reduced exploitation of OSCs in general. Differences between +1 TGG results when only *Mycoplasma* genomes with reduced negative TGA selection are included and when all are included could suggest that as +1 TGG densities are increasingly affected by the selection against +1 TGA (for codons encoding +1 TGA, G is the nucleotide most likely under selection, which also exists at the second position of +1 TGG). If +1 TGA has been selected against for sufficiently long, it is possible that +1 TGA and +1 TGG reach an equilibrium, whereby densities of both are reduced despite only TGA function being lost.

A Refined Version of the Ambush Hypothesis

One might reasonably suggest that the above evidence only adds to the uncertainty of data related to the ambush hypothesis and highlights the sensitivity of the tests to small assumptions about how to test against a null. What is clear is that the ambush hypothesis cannot unambiguously explain OSC usage in all bacterial genomes. However, the data are such that we also cannot easily dismiss the hypothesis that no genome selects for OSCs. Importantly, there is a considerable overlap in the number of genomes with significant +1 excesses for both the codon shuffle model and synonymous site randomization model (+1 TAA: 90.60%, +1 TAG: 76.84%, +1: TGA: 67.84%, percentages of genomes in the model with most excesses that also have significant excesses in the model with fewer excesses), suggesting the signals we observe for both models are genuine. Prima facie these results appear to contradict the ambush hypothesis, as frameshift tracts should on average be shorter in AT-rich genomes (Warnecke et al. 2010; fig. 2). Thus, if there were to be a refined version of the hypothesis, it would need to explain why AT-rich genomes appear to be more associated with an excess. There is a possible (post hoc) refined version of the hypothesis that we suggest is worth considering and that makes some testable predictions.

AT-Rich Genomes Have Higher Frameshift Rates, Consistent with the Refined Model

We (and others) (Tse et al. 2010 and Morgens et al. 2013) have assumed that the ambush hypothesis predicts greater excess from null in GC-rich genomes, as post-frameshift tract lengths in these genomes will be longer. However, this is only half of the equation. The other critical component is the rate at which frameshifts occur. If the rate of frameshifting is higher in AT-rich genomes, selection for OSCs could be higher, refining our model to predict absolutely higher rates, per base pair, in AT-rich genomes. We can test whether AT-rich genomes have higher rates of frameshifting in silico.

Previous evidence suggests that the composition of the tRNA repertoire is important in determining translational accuracy (Baranov et al. 2004; Shah and Gilchrist 2010; Warnecke, et al. 2010), with frameshift-susceptible codons decoded by rarer tRNAs (Curran and Yarus 1989; Sipley and Goldman 1993; Lainé, et al. 2008) and potentially struggling to meet stringent proofreading demands (Ieong et al. 2016). Enriching the tRNA repertoire correlates with reduced frameshift susceptibility (Warnecke et al. 2010). The susceptibility and cost of frameshifting, associated with tRNA abundance and diversity, may therefore be important in determining OSC frequency. The "process cost of accidental frameshift" model (Warnecke et al. 2010) incorporates tRNA information to calculate the susceptibility and cost of frameshifting.

We find the distribution of correlations between median CDS frameshift cost and OSC density approximately even around 0 (supplementary fig. 5*A*, Supplementary Material online). However, genomes where these correlations are positive are typically AT-rich ($\rho = -0.353$, $P < 1.618 \times 10^{-8}$,

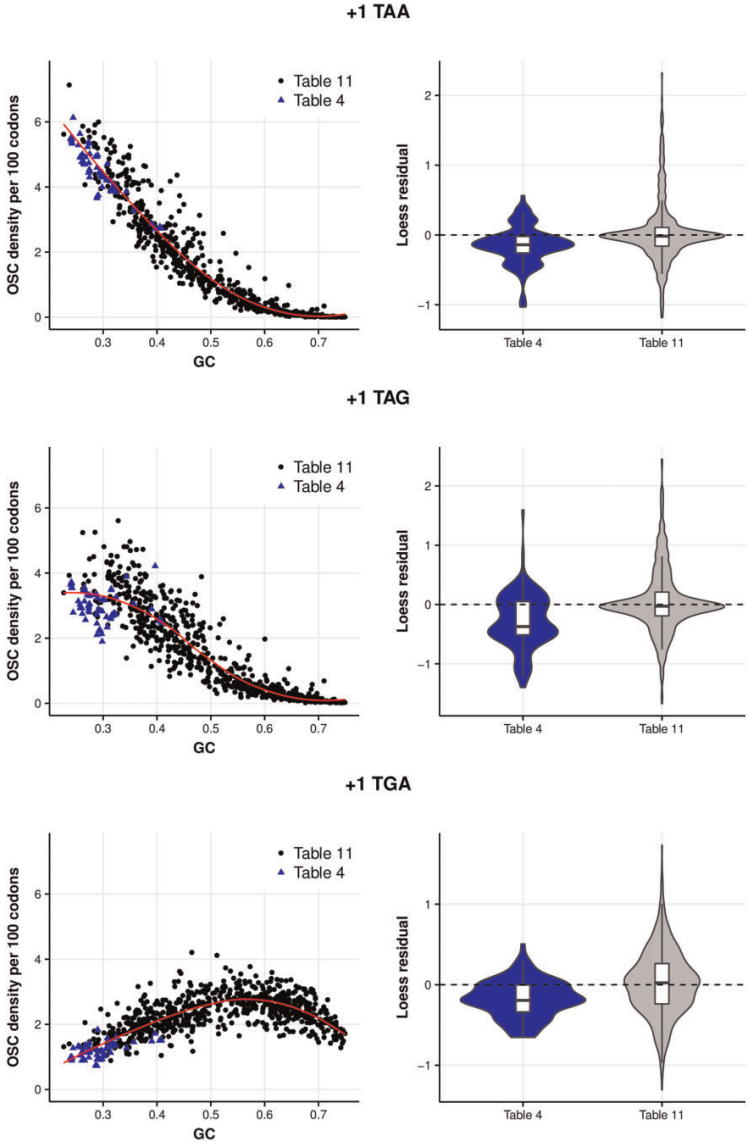


Fig. 7.—OSC densities for table 4 genomes are reduced for each of the stop codons in the +1 frame. Violin plots of the loess regression residuals confirm the reduced densities of each OSC.

Spearman's rank correlation). Thus, despite the on average reduced pre- and post-frameshift tract lengths (Warnecke et al. 2010; fig. 2), frameshifting cost appears to correlate with OSC density.

Are these increased OSC densities compensating for increased costs due to an increased propensity to frameshift? This appears to be the case, as AT-rich genomes seem more susceptible to frameshifting ($\rho = -0.660$, $P < 2.2 \times 10^{-16}$,

Spearman's rank correlation) (fig. 8*A*). Deviations from null (Z scores) are positively correlated with the susceptibility to frameshifting (codon shuffle: estimate: 0.150, $P < 2.2 \times 10^{-16}$; synonymous site simulation: estimate: 0.176, $P < 2.2 \times 10^{-16}$, Spearman's partial correlations) (fig. 8*B*) and not a result of GC-content biases that may increase both frameshift susceptibility and OSC excess. This suggests that our explanation for the connection between AT-richness and OSC excess as a signal of selection in the refined model may have some virtue. In short, in genomes where frameshifting rates are high, tract lengths are typically short and OSCs in excess. Where tract lengths are long, an alternative general strategy to reduce frameshifting rates is the better strategy.

We note that a significant problem faced with this type of analysis is that we must make generalizations in order to compare between genomes. For instance, Warnecke et al. (2010) outline that codon–anticodon interactions are invariably generalizations, as tRNA decoding capacity cannot be predicted from sequence information alone. Furthermore, the effects of modifications to anticodon residues and tRNAs on decoding capacity (Cochella and Green 2005; Daviter, et al. 2006; Grosjean, et al. 2010) are likely to be genome specific. Thus, although results establish a relationship between signatures of OSC selection and frameshift probability, more in-depth conclusions regarding the extent to which OSCs are under selection should be considered in the knowledge of these limitations.

A Refined Model Still Leaves Observations Unexplained

Given the above result, we suggest that the refined model may have some validity. However, although it is to a large degree a post hoc model, it fails to explain everything. Two results post the most obvious problems. First, why do we see so many biases of sense codons with similar nucleotide composition out of frame? Second, why is there a dearth of all off-frame stop codons in the table 4 genomes that do not employ TGA?

Regarding the second of these, had we observed an excess of +1 TAA but not TAG, this would have been consistent with the refined model, but we do not. However, the refined model makes no pretense to suppose all genomes cope with frameshifts by use of OSCs. By virtue of using a different code, table 4 genomes can be automatically considered to be somewhat exceptional. Indeed, selection pressures experienced by these organisms associated with their particular ecological niches (Bove 1993) may also be unusual. Another possibility is the weakened purifying selection attributable to smaller effective populations (N_e) of table 4 genomes. However, if a universal GC to AT mutation bias exists (Lind and Andersson 2008; Hershberg and Petrov 2010), GC content should act as a reasonable proxy for low N_e (many AT-rich bacterial genomes likely have low N_e). Thus, although reduced

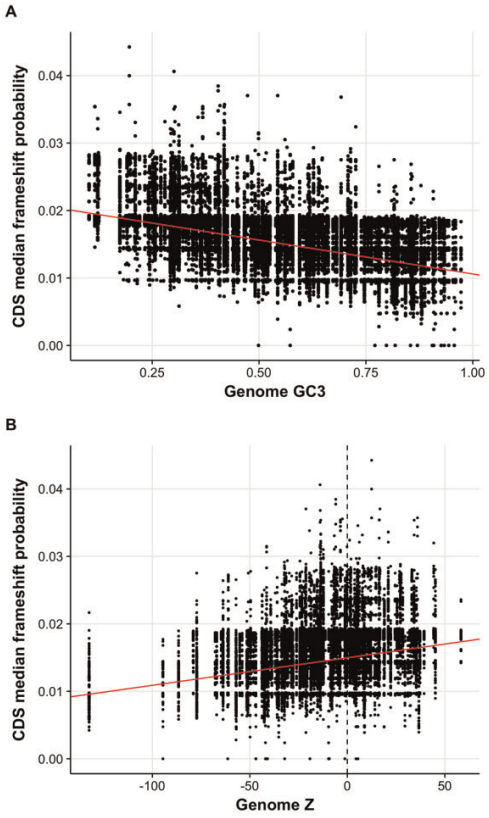


FIG. 8.—*(a)* The median probability of frameshifting decreases with increasing GC3 content. *(b)* Genomes with excesses of OSCs for the synonymous site model tend to have higher +1 frameshift probabilities, suggesting the frequency of OSCs and susceptibility of frameshifting are linked.

N_e may contribute, it is unlikely to explain the overall trends we observe.

Interestingly, we notice both TGA and TGG have similar numbers of genomes with off-frame excesses in our simulation models. Coupled with the results of table 4 genomes, this suggests excesses of TGA may not be related to termination function. In the refined model, increased densities of +2 TGA in the table 4 genomes support this notion, suggesting that some excesses are not associated with stop functionality but either reflect chance or missing layers of complexity not accounted for in our simulations. There may, for example, be constraints on protein-level motifs, or at the DNA or RNA level, coupled to localized selection for optimal codon usage that distorts out-of-frame usage as an incidental side

consequence. For this reason, we remain skeptical that the ambush hypothesis, even in its refined form, commands any strong support at present. This being said, the fact some sense codons are enriched out of frame does not itself demonstrate that stop codon enrichment out of frame is not owing to stop functionality, but rather there might be an alternative unknown explanation. Thus, while both of these unexplained features are not obviously consistent with the refined model, neither are they lethal to it.

Discussion

The notion that OSC selection should constrain sequence evolution to compensate for frameshift errors is logical. More recently, Morgens et al. (2013) demonstrated the initial result on which the ambush hypothesis was founded (Seligmann and Pollock 2004) is not robust to compositional control. Furthermore, this initial evidence only weakens after multiple correction testing (supplementary result 4, Supplementary Material online). However, an alternative approach using simulated sequences from Markov models identifies many genomes with an excess of OSCs (Tse et al. 2010; Morgens et al. 2013). An underlying issue with these models is their inability to strictly maintain amino acid frequencies, amino acid order, and codon usage frequencies. Under real evolutionary constraints, such flexibility is unlikely to be permitted and not realistic. Thus, the motivation of this paper was to establish the extent, if any, to which OSCs drive sequence evolution in a more realistic simulation framework and when microscale position effects are controlled.

We proposed and tested a series of simulation approaches, none of which control for all possible biases, but with each reaching similar conclusions (see supplementary table 2, Supplementary Material online, for summaries), the numbers of genomes with significant excesses are modest, often under 50%; genomes with an excess of OSCs tend to be AT-rich; and not all stop codons nor reading frames are equally affected. A post hoc model makes sense of these observations, but the predictions of this model regarding different handling of TGA and TAA compared with TAG and the preponderance of +1 frameshifts remain to be tested.

An important consequence of the refined model is that naively assuming GC-rich genomes bear greater frameshift costs does not account for more complex frameshift dynamics. Citing a positive correlation between GC content and any excess of OSCs as evidence consistent with OSC selection as in previous studies (Tse et al. 2010; Morgens et al. 2013), even if further analyses are not consistent with selection, is likely to be too simplistic. To more comprehensively quantify the cost of both frameshift errors and errors in general, it is important to consider complex relationships between error frequency and the selective constraints imposed to mitigate any costs.

The structure of the refined model more broadly considers frameshift control in a framework, whereby two distinct strategies have evolved and have different usage in different genomes. In one case, frameshifts are, on average, very damaging due to long frameshift tract lengths (GC-rich genomes). In this instance, a general reduced frameshifting rate is selectively advantageous which in turn reduces the selective pressure to incorporate any given OSC (although downstream of particularly frameshift-prone sites might be an exception). At the limit, if the frameshift rate could be reduced to zero, there would be no requirement for or selection for OSCs. Conversely, in other genomes (AT-rich), the average frameshift has little cost as tract lengths are naturally short. Here, selection cannot act to generally reduce frameshift rates, as there is likely be little return on investment of such a reduction for a given cost. However, even in these genomes, there will remain sites where by chance, tract lengths are long. In these sites, there could then be selection—given the high frameshifting rates—for OSCs. Thus, in this two-mode framework, we might expect more OSC excesses in AT-rich genomes and not as usually asserted in GC-rich genomes, although strategies are likely to be highly genome specific (as evidenced by negative excesses in many genomes).

One interesting notion arising from this framework is the coevolution of frameshift rates and OSCs. Whether proposed frameshift rate increases are due to weakened purifying selection in genomes with reduced N_e (assuming GC-rich genomes have larger effective population sizes), or whether the nucleotide content of AT-rich genomes naturally encoding greater numbers of OSCs means frameshifts are less costly, the ability to prevent frameshifting itself appears to be relaxed in AT-rich genomes. Parenthetically, error frequency may be the principal determinant of the strength of selection for OSCs in these genomes with this framework providing another possible example, whereby selection may be stronger in response to increased error rates when populations are small (Wu and Hurst 2015). In genomes where this frameshift error rate is reduced, or alternative pressures exert stronger selection on the CDS, the ability to maintain OSCs within CDSs may be significantly reduced and not a viable frameshift control strategy leading to significant depletions of OSCs. Indeed, other selective pressures, such as those imposed by environmental constraints (the ability to incorporate new DNA via off-frame recombination in metabolically versatile bacteria, or prevent recombination in more stable symbionts may be imperative to genetic adaptation; Wong et al. 2008), may also be important in determining the degree of OSC selection.

We also question why genomes tend to use TGA and TAA as OSCs. While TGA is the weakest of the stops (and prone to read-through) (Meng et al. 1995; Wei et al. 2016), TGA and TAA are unique in the specificity of release factors (RFs) decoding the stop codons: RF2 decodes both TAA and TGA (Kisselev 2002). RF2 in combination with RF3 is implicated in post peptidyl transfer quality control, ensuring more efficient termination at tRNA/mRNA mismatch complexes and proposed to participate in ribosome rescue (Zaher and Green

2009; Vivanco-Domínguez et al. 2012; Petropoulos et al. 2014). Specific capabilities of RF2 may therefore make TAA and TGA more suitable to frameshift termination, rather than the efficiency of termination of the stop codons themselves and predicts that captured frameshifts are more likely processed by the RF2/RF3 complex. In addition, minimal TAG excesses may possibly reflect avoidance of complementary GATC DNA motifs found frequently in nonrandom clusters on the bacterial chromosome (Touzain, et al. 2011).

One consistency is the bias toward excesses seen for +1 but not for the +2 frame. Here we can only conjecture that frameshifting, by accident, occurs predominantly in the +1 slippage mode. We can speculate that as translation occurs in the 5′ to 3′ direction, the molecular mechanics required to halt and reverse the direction of translation to the first nucleotide of a −1 frameshift, already held in the P-site, are likely to be more complex and require greater energy than for a ribosome to skip to the +1 frame in the same direction. Thus, accidental +1 frameshifts may be more frequent and require greater OSC control, although this is only speculation without comprehensive frameshift rate data and would no doubt benefit from molecular frameshift data. This should be experimentally testable. Our refined model is therefore one in which the genomes, stop codons, and reading frames are important factors in OSC selection.

Problems Defining the Null

One of the lessons of the analysis presented here is that the meaning of a deviation from null is hard to interpret, not least because the results are dependent upon the definition of the null. Aside from the issue of which model is the most appropriate, we have looked for deviations at the genome level and not at the gene level. As OSC selection is likely to be sequence and context specific, it is also worth considering whether investigating OSC selection at the genome level is the most appropriate. For instance, Bertrand et al. (2015) have demonstrated no evidence consistent with OSC selection in the polyketide synthase (PKS) gene in fungi. Furthermore, sequences with differing levels of frameshifting are commonplace in coding regions of *E. coli* (Gurvich et al. 2003). As the information-carrying capacity of CDSs is limited, competing selection pressures providing more beneficial and selectable fitness advantages will be favored. Any selection for OSCs is likely to be one of several competing pressures, with OSC selection therefore potentially undetectable at whole genome scales.

Equally, a more appropriate approach may be to consider the single gene level, as selection may be stronger and more detectable in subsets of genes and avoided in others. For example, one might, at first sight, expect stronger selection in highly expressed genes. This hypothesis, however, has the caveat that highly expressed genes are likely to be composed of codons less susceptible to frameshifting (i.e., matching common tRNAs) and therefore not require OSC selection. The latter case, at least for +1 frameshifts for which this framework is most applicable, seems appropriate (supplementary fig. 6, Supplementary Material online). Alternatively, for genes overly susceptible to frameshifting, such as those incorporating mononucleotide repeats (Coenye and Vandamme 2005), OSCs provide an attractive strategy which tRNA selection is unable to regulate. Extending research to determine whether OSCs have important evolutionary implications at a single gene scale would help to inform us whether OSCs have useful applications in, for example, transgene design.

We also highlight two further limitations of our approach. First, an assumption of our models is that OSCs are indeed selected for. However, it is also known that organisms in all kingdoms utilize frameshifting to increase coding capacity to translate multiple proteins from the same CDS, for example the gag-pol protein (Jacks et al. 1988; Dulude et al. 2002) or in autoregulatory feedback systems (Baranov et al. 2002; Betney et al. 2010) via programmed frameshifting (Farabaugh 1996; Dinman 2012; Ketteler 2012). In such instances, the null expectation should not be selection for OSCs but rather strong avoidance selection. Even with the knowledge of well-annotated programmed frameshifts, it would be difficult to define how a null sequence with no selection should be composed. Our analyses cannot account for such programmed frameshifting without first removing CDSs where these frameshifts occur. The highly site-, context-, and CDS-specific nature of programmed frameshifts are, however, unlikely to greatly influence our conclusions.

Second, we assume that regardless of sequence context an OSC can function as a stop codon. Put differently, our null deviations are defined with respect to OSC number rather than OSC efficiency. There are, however, likely to be many alternative factors influencing the efficiency of terminations both for regular stop codons and for OSCs. For example, we assume that upon entering the ribosome A-site, an OSC functions as regular stop codon and has the same ability to recruit release factors. The nucleotide context surrounding stop codons, particularly the nucleotide following the stop codon, is also an important determinant of termination efficiency and read through (Poole et al. 1995; Tate et al. 1996; Mottagui-Tabar and Isaksson 1997; Namy et al. 2001; Cridge et al. 2006; Wei and Xia 2017). An initial analysis of the nucleotide 3′ of OSCs indicates no such bias (supplementary fig. 7, Supplementary Material online). In *E. coli*, the cooperation of chemical properties to the penultimate two amino acids in the nascent peptide to form secondary structures can also determine termination efficiencies (Mottagui-Tabar et al. 1994; Björnsson et al. 1996). Any analyses that can further establish the extent to which the sequence context surrounding stop codons has on termination efficiency and the implications for OSCs may provide useful.

In summary, we propose that for the ambush hypothesis to be considered as having any validity, care is required in

defining null expectations and that a more appropriate framework is one that considers not all genomes, not all stops, and not all alternative frames as equally relevant. Our modified framework holds promise, given its ability to predict higher frameshifting rates in genomes with high OSC excess but comes with unexplained features and caveats.

Supplementary Material

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

Acknowledgments

This work was supported by the European Research Council (Advanced grant ERC-2014-ADG 669207 to L.D.H.) and the Medical Research Council (grant number MR/L007215/1 to L.D.H.).

Literature Cited

Abe T. 2011. tRNADB-CE 2011: tRNA gene database curated manually by experts. Nucleic Acids Res. 39(Database issue):D210–D213.
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.
Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. PLoS Genet. 2(2):e22.
Atkins JF, Bjork GR. 2009. A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. Microbiol Mol Biol Rev. 73(1):178–210.
Baranov PV, Gesteland RF, Atkins JF. 2002. Release factor 2 frameshifting sites in different bacteria. EMBO Rep. 3(4):373–377.
Baranov PV, Gesteland RF, Atkins JF. 2004. P-site tRNA is a crucial initiator of ribosomal frameshifting. RNA 10(2):221–230.
Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 9:675.
Bertrand RL, Abdel-Hameed M, Sorensen JL. 2015. Limitations of the 'ambush hypothesis' at the single-gene scale: what codon biases are to blame? Mol Genet Genomics. 290(2):493–504.
Betney R, de Silva E, Krishnan J, Stansfield I. 2010. Autoregulatory systems controlling translation factor expression: thermostat-like control of translational accuracy. RNA 16(4):655–663.
Björnsson A, Mottagui-Tabar S, Isaksson LA. 1996. Structure of the C-terminal end of the nascent peptide influences translation termination. EMBO J. 15(7):1696–1704.
Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci. 25(3):106–110.
Bove JM. 1993. Molecular features of mollicutes. Clin Infect Dis. 17(Suppl 1):S10–S31.
Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. Genome Biol. 14(12):R143.
Cochella L, Green R. 2005. An active role for tRNA in decoding beyond codon: anticodon pairing. Science 308(5725):1178–1180.
Cock PJA, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11):1422–1423.
Coenye T, Vandamme P. 2005. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res. 12(4):221–233.
Cognat V, et al. 2008. On the evolution and expression of Chlamydomonas reinhardtii nucleus-encoded transfer RNA genes. Genetics 179(1):113–123.
Cridge AG, et al. 2006. Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. Nucleic Acids Res. 34(7):1959–1973.
Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol. 209(1):65–77.
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.
Daviter T, Gromadski KB, Rodnina MV. 2006. The ribosome's response to codon-anticodon mismatches. Biochimie 88(8):1001–1011.
Dinman JD. 2012. Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip Rev RNA. 3(5):661–673.
Diwan GD, Agashe D. 2016. The frequency of internal Shine-Dalgarno-like motifs in prokaryotes. Genome Biol Evol. 8(6):1722–1733.
Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J Mol Biol. 260(5):649–663.
dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res. 31(23):6976–6985.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10(10):715–724.
Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2):341–352.
Dulude D, Baril M, Brakier-Gingras L. 2002. Characterization of the frameshift stimulatory signal controlling a programmed-1 ribosomal frameshift in the human immunodeficiency virus type 1. Nucleic Acids Res. 30(23):5094–5102.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism–based validation of exonic splicing enhancers. PLoS Biol. 2(9):e268.
Farabaugh PJ. 1996. Programmed translational frameshifting. Annu Rev Genet. 30:507–528.
Farlow A, Meduri E, Dolezal M, Hua L, Schlötterer C. 2010. Nonsense-mediated decay enables intron gain in drosophila. PLoS Genet. 6(1):e1000819.
Frottin F, et al. 2006. The Proteomics of N-terminal Methionine Cleavage. Mol Cell Proteomics. 5:2336–2349.
Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. Mol Syst Biol. 7:481.
Grosjean H, de Crecy-Lagard V, Marck C. 2010. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett. 584(2):252–264.
Gu T, Tan S, Gou X, Araki H, Tian D. 2010. Avoidance of long mononucleotide repeats in codon pair usage. Genetics 186(3):1077–1084.
Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Mol Biol Evol. 29(10):3037–3044.
Gu W, Zhou T, Wilke CO. 2010. A Universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 6(2):e1000664.
Gurvich OL, Baranov PV, Gesteland RF, Atkins JF. 2005. Expression levels influence ribosomal frameshifting at the tandem rare arginine codons AGG_AGG and AGA_AGA in *Escherichia coli*. J Bacteriol. 187(12):4023–4032.
Gurvich OL, et al. 2003. Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*. EMBO J. 22(21):5941–5950.

He F, Peltz SW, Donahue JL, Rosbash M, Jacobson A. 1993. Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1-mutant. Proc Natl Acad Sci USA. 90(15):7034–7038.
Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6(9):e1001115.
Hormoz S. 2013. Amino acid composition of proteins reduces deleterious impact of mutations. Sci Rep. 3:2919.
Hurst LD. 2009. Genetics and the understanding of selection. Nat Rev Genet. 10(2):83–93.
Ieong KW, Uzun U, Selmer M, Ehrenberg M. 2016. Two proofreading steps amplify the accuracy of genetic code translation. Proc Natl Acad Sci USA. 113(48):13744–13749.
Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 17(4):405–412.
Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. Genome Res. 20(11):1582–1589.
Jacks T, et al. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature 331(6153):280–283.
Jaillon O, et al. 2008. Translational control of intron splicing in eukaryotes. Nature 451(7176):359–362.
Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238(1):143–155.
Ketteler R. 2012. On programmed ribosomal frameshifting: the alternative proteomes. Front. Genet. 3:242.
Kisselev L. 2002. Polypeptide release factors in prokaryotes and eukaryotes: same Function, Different Structure. Structure 10(1):8–9.
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, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324(5924):255–258.
Lainé S, Thouard A, Komar AA, Rossignol J-M. 2008. Ribosome can resume the translation in both +1 or −1 frames after encountering an AGA cluster in *Escherichia coli*. Gene 412(1–2):95–101.
Li G-W, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484(7395):538–541.
Liao Y-D, Jeng J-C, Wang C-F, Wang S-C, Chang S-T. 2004. Removal of N-terminal methionine from recombinant proteins by engineered E. coli methionine aminopeptidase. Protein Sci. 13:1802–1810.
Lin MF, et al. 2011. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. Genome Res. 21(11):1916–1928.
Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci USA. 105(46):17878–17883.
Mekouar M, et al. 2010. Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. Genome Biol. 11(6):R65.
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.
Morgens DW, Chang CH, Cavalcanti ARO. 2013. Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. BMC Genomics 14(1):418.
Mottagui-Tabar S, Björnsson A, Isaksson LA. 1994. The second to last amino acid in the nascent peptide as a codon context determinant. EMBO J. 13(1):249–257.
Mottagui-Tabar S, Isaksson LA. 1997. Only the last amino acids in the nascent peptide influence translation termination in *Escherichia coli* genes. FEBS Lett. 414(1):165–170.
Namy O, Hatin I, Rousset J-P. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep. 2(9):787–793.
Ouidir T, Jarnier F, Cosette P, Jouenne T, Hardouin J. 2015. Characterization of N-terminal protein modifications in Pseudomonas aeruginosa PA14. J Proteomics. 114:214–225.
Pancsa R, Tompa P. 2016. Coding Regions of Intrinsic Disorder Accommodate Parallel Functions. Trends Biochem Sci. 41(11):898–906.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23(2):301–309.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5(2):e14.
Petropoulos AD, McDonald ME, Green R, Zaher HS. 2014. Distinct roles for release factor 1 and release factor 2 in translational quality control. J Biol Chem. 289(25):17589–17596.
Poole ES, Brown CM, Tate WP. 1995. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. EMBO J. 14(1):151–158.
Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. Biol Direct. 7(1):30.
R Core Team. 2015. R: A language and environment for statistical computing. Version 4.3.2. Vienna, Austria: R Foundation for Statistical Computing.
Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14(11):2279–2286.
Savisaar R, Hurst LD. 2017. Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. Mol Biol Evol. 34(5):1110.
Seligmann H. 2007. Cost minimization of ribosomal frameshifts. J Theor Biol. 249(1):162–167.
Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol. 23(10):701–705.
Serohijos AWR, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2(2):249–256.
Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 41(4):2073–2094.
Shah P, Gilchrist MA. 2010. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. PLoS Genet. 6(9):e1001128.
Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3):1281–1295.
Shine J, Dalgarno L. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci USA. 71(4):1342–1346.
Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: evidences for phylogenetic trends. Comput Biol Chem. 33(3):239–244.
Sipley J, Goldman E. 1993. Increased ribosomal accuracy increases a programmed translational frameshift in *Escherichia coli*. Proc Natl Acad Sci USA. 90(6):2315–2319.
Tate WP, et al. 1996. The translational stop signal: codon with a context, or extended factor recognition element? Biochimie 78(11–12):945–952.
Touzain F, Petit M-A, Schbath S, Karoui ME. 2011. DNA motifs that sculpt the bacterial chromosome. Nat Rev Microbiol. 9(1):15–26.

Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y. 2010. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics 11(1):491–413.
Vivanco-Domínguez S, et al. 2012. Protein synthesis factors (RF1, RF2, RF3, RRF, and tmRNA) and peptidyl-tRNA hydrolase rescue stalled ribosomes at sense codons. J Mol Biol. 417(5):425–439.
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, Huang Y, Przytycka TM, Hurst LD. 2010. Unique cost dynamics elucidate the role of frame-shifting errors in promoting translational robustness. Genome Biol Evol. 2(0):636–645.
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, 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.
Wohlgemuth I, Pohl C, Rodnina MV. 2010. Optimization of speed and accuracy of decoding in translation. EMBO J. 29(21):3701–3709.
Wong T-Y, et al. 2008. Role of premature stop codons in bacterial evolution. J Bacteriol. 190(20):6718–6725.
Wu X, 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.
Wu Y, Zhang Y, Zhang J. 2005. Distribution of exonic splicing enhancer elements in human genes. Genomics 86(3):329–336.
Yutani K, Ogasahara K, Sugino Y, Matsushiro A. 1977. Effect of a single amino acid substitution on stability of conformation of a protein. Nature 267(5608):274–275.
Zaher HS, Green R. 2009. Quality control by the ribosome following peptide bond formation. Nature 457(7226):161.
Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 26(7):1571–1580.

Associate editor: Mary O'Connell

Supplement to Chapter 4

Supplementary Results

The Supplementary Results, Supplementary Figures and Supplementary Tables presented below can also be found accompanying the published paper. These have been reformatted for this thesis.

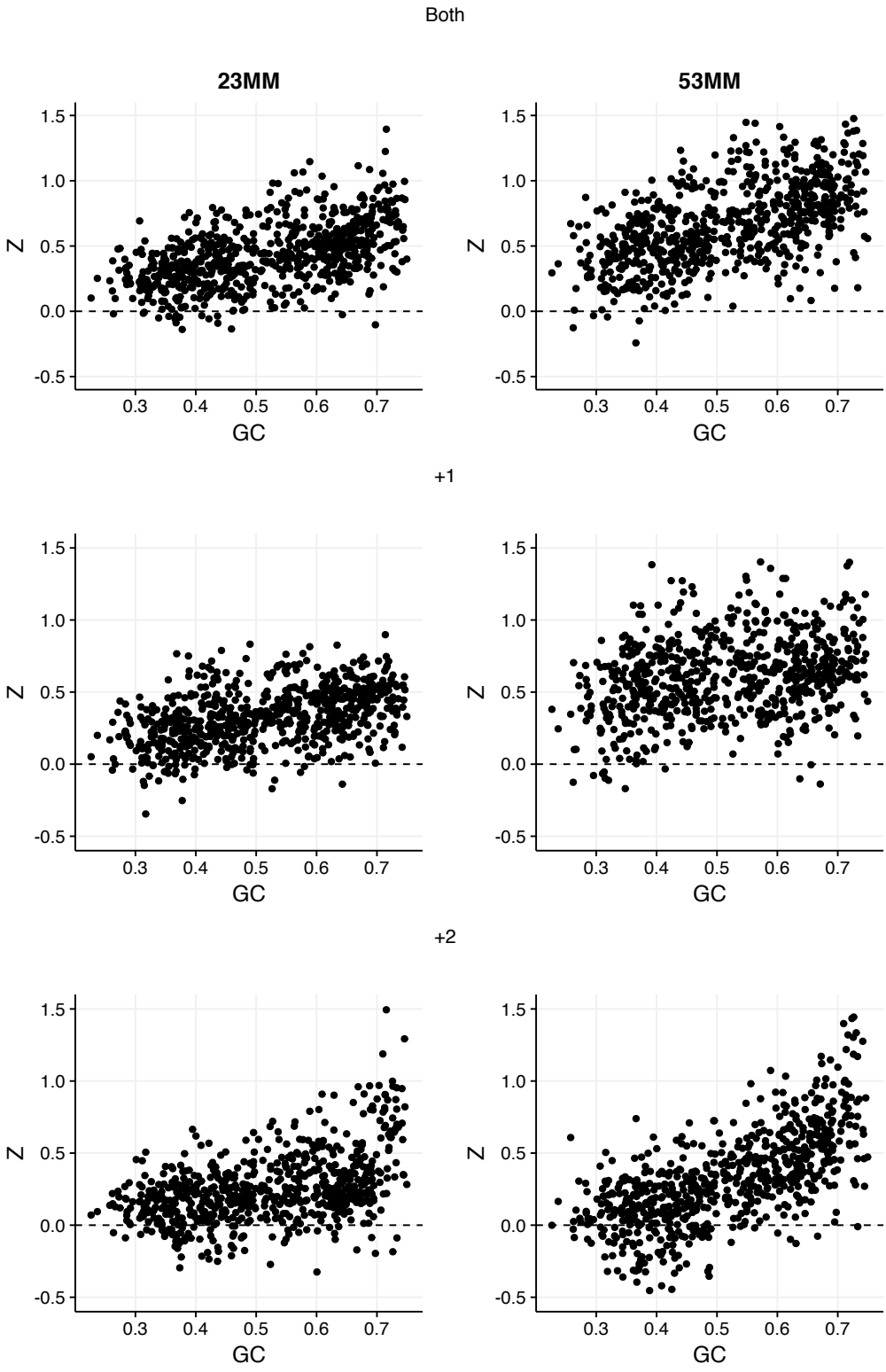
Supplementary Result 1

Markov modelling demonstrates significant positive correlations between OSC excesses and GC content consistent with previous studies

We performed Markov modelling similar to that performed by Tse et al. (2010) and Morgens et al. (2013) in order to ascertain whether we could replicate previous excesses with our dataset. Tse et al. (2010) report 99.1% of genomes with OSC excesses under the second-order model and 93.3% under the fifth-order model. Morgens et al. (2013) report excesses in 83% of genomes analysed for both models. Our simulations report similar distributions of results, however using the Z metric we find no genomes with significant excess. We find OSC excesses in 677/694 (97.55%) genomes under second-order models and 689/694 (99.28%) under fifth-order models (Supplementary Result 1 Table 1).

When reading frames are considered individually, we find 661/694 (95.24%) and 683/694 (98.41%) genomes with significant excesses in the +1 frame using the second-order and fifth-order models respectively. In the +2 frame, 621/694 (89.48%) and 591/694 (85.16%) genomes exhibit significant excesses. Correlations between GC content and OSC excess are significant and positive for each model in each reading frame (second-order - both: $\rho = 0.529$, $P < 2.2 \times 10^{-16}$; +1: $\rho = 0.450$, $P < 2.2 \times 10^{-16}$;

+2: $\rho = 0.443$, $P < 2.2 \times 10^{-16}$; fifth-order – both: $\rho = 0.581$, $P < 2.2 \times 10^{-16}$; +1: $\rho = 0.279$, $P = 8.667 \times 10^{-14}$; +2: $\rho = 0.687$, $P < 2.2 \times 10^{-16}$; Spearman's rank correlations) (Supplementary Result 1 Figure 1) and are consistent with OSC selection as predicted by the ambush hypothesis and previously discussed (Tse et al. 2010; Morgens et al. 2013).



Supplementary Result 1 Figure 1: Correlations between GC content and OSC excesses (standard Z score) after CDS simulation using second-order three-periodic Markov models. Each reading frame, including when both are considered together, demonstrates significant positive correlations ($P < 0.05$, Supplementary Result 1 Table 1) with GC content, for both Markov models.

Do the sense codons demonstrate significant excesses greater than for OSCs, with significant positive correlations with GC content, as identified by Morgens et al. (2013)?

Under the second-order model, +1 TAT (624/694, 89.91%) has a greater number of genomes with excess than +1 TAG (503/694, 72.48%) with a positive correlation between +1 TAA excesses and GC content ($\rho = 0.596$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation), not displayed by either +1 TAA or +1 TAG. For the TGN codons, +1 TGA has the most genomes with excesses (513/694, 73.92%), although the correlation with GC content ($\rho = 0.629$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) is weaker than both +1 TGC ($\rho = 0.736$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) and +1 TGG ($\rho = 0.728$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation). In the +2 frame, TAT (490/694, 70.61%) and TAC (298/694, 42.92%) have more excesses than TAG (284/694, 40.92%). The correlation between GC content and excesses is not significant for +2 TAA ($P = 0.07$, Spearman's rank correlation), whilst both +2 TAC ($\rho = 0.573$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) and +2 TAT ($\rho = 0.381$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) have stronger significant positive correlations that +2 TAG ($\rho = 0.313$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation). +2 TGA has the greatest number of excesses (666/694, 95.97%) compared with +2 TGC (570/694, 82.13%), +2 TGG (545/694, 78.53%) and +2 TGT (24/694, 3.46%).

Under the fifth-order model we find similar trends. +1 TAA (662/694, 95.38%) has greater number of genomes with excesses than other TAN codons, although the correlation with GC content is significantly negative ($\rho = -0.355$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) unlike the remaining TAN codons. +1 TAG has the fewest excesses of TAN codons (527/694, 75.94%). For +1 TGA codons, +1 TGA has the most excesses (603/694, 86.89%) and strongest positive correlation ($\rho = 0.733$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation). For +2 TAN codons, TAA again has more excesses (682/694, 98.27%) with +2 TAT (664/694, 95.68%) having more than +2 TAG (657/694, 94.67%). +2 TGA also has greater excesses than any other +2 TGN codon (207/694, 38.90%).

Our results report distributions of excesses to Morgens et al. (2013) with stop codons often with fewer genomes with excesses and less strongly positively correlated (negatively correlated) with GC content than for sense codons, advocating our choice of genomes. However, results should be interpreted with caution. For example, when both frames are considered together, TAC has significant positive correlations between GC content and OSC excesses and 97.12% and 96.54% of genomes with excesses for second-order and fifth-order models respectively. TAA, despite significant negative correlations (second-order model: ρ = -0.254, P = 1.305 × 10^{-11}, Spearman's rank correlation; fifth-order model: ρ = -0.355, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) has excesses in 98.99% and 98.85% genomes. Results would therefore indicate stronger selection for TAA given the much-increased number of genomes with excesses. Thus, whilst it is important to consider other observation, the first consideration must be whether OSC frequencies deviate from the null frequencies.

Supplementary Result 1 Table 1: Summary of the Markov model simulation genome excesses for OSCs when considered together and individually for each reading. Sense codons are provided for comparison.

	Model	**2nd order Markov model**		**5th order Markov model**	
Codon	**Reading frame**	**# with excess**	**% with excess**	**# with excess**	**% with excess**
All stops	**Both**	677	97.55	689	99.28
All stops	**+1**	661	95.24	683	98.41
All stops	**+2**	621	97.55	591	85.16
TAA	**Both**	687	98.99	686	98.85
TAC	**Both**	674	97.12	670	96.54
TAG	**Both**	665	95.82	606	97.32
TAT	**Both**	684	98.56	644	92.80
TGA	**Both**	309	44.52	400	57.64
TGC	**Both**	327	47.12	203	29.25
TGG	**Both**	333	47.98	57	8.21
TGT	**Both**	321	46.25	10	1.44
TAA	**+1**	678	97.69	662	95.39
TAC	**+1**	652	93.95	627	90.35

TAG	**+1**	503	72.48	527	75.94
TAT	**+1**	624	89.91	609	87.75
TGA	**+1**	513	73.92	603	86.89
TGC	**+1**	444	63.98	252	36.31
TGG	**+1**	452	65.12	69	9.94
TGT	**+1**	11	1.59	23	3.31
TAA	**+2**	527	75.94	682	98.27
TAC	**+2**	298	42.92	647	93.23
TAG	**+2**	284	40.92	657	94.67
TAT	**+2**	490	70.61	664	95.68
TGA	**+2**	666	95.97	270	38.90
TGC	**+2**	570	82.13	181	26.08
TGG	**+2**	545	78.53	113	16.28
TGT	**+2**	24	3.46	243	35.01

Supplementary Result 2

Genomes exhibit minimal OSC excesses when given the flexibility in codon choice between multiple coding blocks permits a choice between synonymous codons that can and can't encode an OSC

We consider a third simulation model, similar to our model which randomises synonymous sites. If we further permit changes between coding blocks, we can ask whether selection favours codons that encode OSCs if given the choice between codons that do and do not. For example, suppose the peptide sequence necessitates a valine followed by serine. If OSCs exert a strong enough selection pressure, we would expect preferential use of GTA or GTG valine codons followed by AGC or AGT serine codons as opposed to GTC or GTT and the T-starting serine codons to encode a +1 OSC. To consider selection to this effect, we randomised the use of synonymous codons throughout the genome, accounting for genome specific codon usage frequencies, controlling amino acid sequences and GC content whilst disrupting site-specific synonymous codon choice. OSCs generated from one-fold degenerate codons are not considered as randomisation has no effect on the identity of these codons.

Similar to the other models, evidence is not consistent with OSC selection. Only 84/694 (12.10%) genomes have significant excesses of OSCs ($P < 0.05$, FDR correction). This result is however, strongly influenced by the reduced excess in the +2 frame; only 107/694 (15.42%) genomes demonstrate significant excesses ($P < 0.05$, FDR correction) compared with 262/694 (37.75%) in the +1 frame ($P < 0.05$, FDR correction). Correlations between GC content and excesses are significantly negative for each reading frame (Supplementary Result 2 Table 1). The evidence to suggest CDSs favour codons that generate an OSC is weak and limited predominantly to the +1 frame, with significant excesses highly restricted to the AT-rich genomes (Supplementary Result 2 Figure 1).

Supplementary Result 2 Table 1: The number of genomes with significant out-of-frame excesses for different codons in the various reading frames when synonymous codons have

been randomised. Spearman's rank correlations between GC content and OSC excess, defined by the standard Z score are also shown.

Codon	Reading frame	# with excess	% with excess	ρ	*P*
All stops	**Both**	84	12.10	-0.444	$< 2.2 \times 10^{-16}$
All stops	**+1**	262	37.75	-0.458	$< 2.2 \times 10^{-16}$
All stops	**+2**	107	15.42	-0.234	4.781×10^{-10}
TAA	**Both**	116	16.71	-0.513	$< 2.2 \times 10^{-16}$
TAC	**Both**	160	23.05	-0.051	0.176
TAG	**Both**	90	12.97	-0.273	3.407×10^{-13}
TAT	**Both**	194	27.95	-0.364	$< 2.2 \times 10^{-16}$
TGA	**Both**	281	40.49	-0.336	$< 2.2 \times 10^{-16}$
TGC	**Both**	629	90.63	0.595	$< 2.2 \times 10^{-16}$
TGG	**Both**	264	38.04	-0.416	$< 2.2 \times 10^{-16}$
TGT	**Both**	252	36.31	-0.345	$< 2.2 \times 10^{-16}$
TAA	**+1**	296	42.65	-0.437	$< 2.2 \times 10^{-16}$
TAC	**+1**	366	52.74	0.581	$< 2.2 \times 10^{-16}$
TAG	**+1**	157	22.62	-0.322	$< 2.2 \times 10^{-16}$
TAT	**+1**	432	62.25	0.404	$< 2.2 \times 10^{-16}$
TGA	**+1**	252	36.31	-0.169	7.942×10^{-6}
TGC	**+1**	596	85.88	0.623	$< 2.2 \times 10^{-16}$
TGG	**+1**	269	38.76	-0.383	$< 2.2 \times 10^{-16}$
TGT	**+1**	105	15.13	-0.131	5.497×10^{-4}
TAA	**+2**	95	13.69	-0.308	1.496×10^{-16}
TAC	**+2**	146	21.04	-0.379	$< 2.2 \times 10^{-16}$
TAG	**+2**	43	6.20	-0.151	6.600×10^{-5}
TAT	**+2**	183	26.37	-0.491	$< 2.2 \times 10^{-16}$
TGA	**+2**	361	52.02	-0.249	3.367×10^{-11}
TGC	**+2**	557	80.26	0.185	9.933×10^{-7}
TGG	**+2**	265	38.18	-0.209	3.156×10^{-8}
TGT	**+2**	381	54.90	-0.391	$< 2.2 \times 10^{-16}$

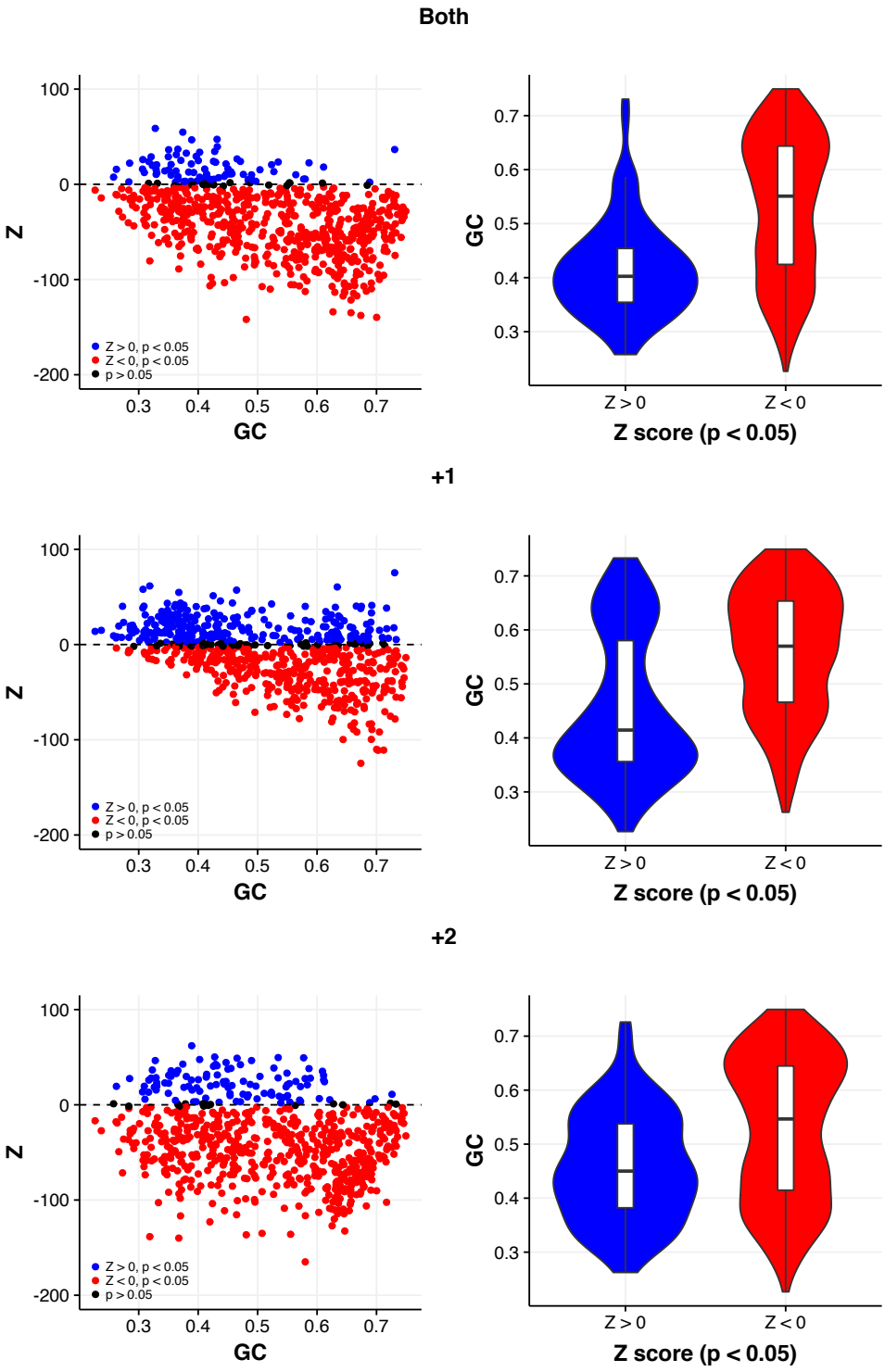
Individually TGA demonstrates the greatest excesses when considered in both reading frames (Supplementary Result 2 Table 1). Few genomes have an excess of TAG in

any frame. Each OSC, in each reading frame, demonstrates significant negative correlations with GC content (Supplementary Result 2 Figure 2, Supplementary Result 2 Figure 3). The number of genomes with significant excesses in the +1 frame is greatest for TAA (296/694, 42.65%, $P < 0.05$, FDR correction). In comparison, TAA use in the +2 frame is extremely reduced (95/694, 13.69%, $P < 0.05$, FDR correction), with +2 TGA having the highest number of genomes with excesses (361/694, 52.02%, $P < 0.05$, FDR correction).

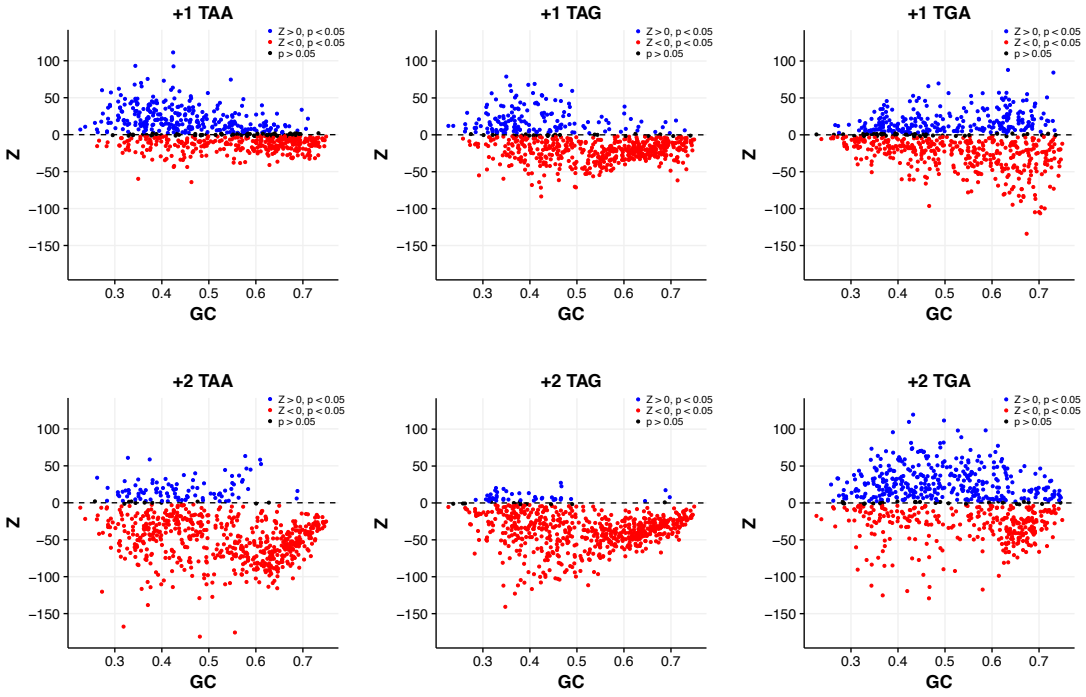
For off-frame sense codons, TGC has an extremely high number of genomes with significant positive excesses in each frame (both: 629/694, 90.63%; +1: 596/694, 85.88%; +2: 557: 80.26%, $P < 0.05$, FDR correction) and is greater than TGA in each reading frame. Both TAC and TAT have greater excesses than TAA or TAG in any reading frame and show significant positive correlations with GC content in the +1 frame. Neither TAA nor TAG demonstrates significant positive correlations in any reading frame.

Given the flexibility of the model to allow for synonymous codon interchange within coding blocks for arginine and serine, we would have expected greater excesses of TAA and TGA in the +1 frame, or TAG in the +2 frame, given the ability of real coding sequences to encode an OSC simply by using the A-starting synonyms. This is not the case. For +2 TAG in particular, where the second codon in the encoding dicodon can only be either an AGR arginine or AGY serine codon, we find extremely low number of genomes with significant excesses (43/694, 6.20%).

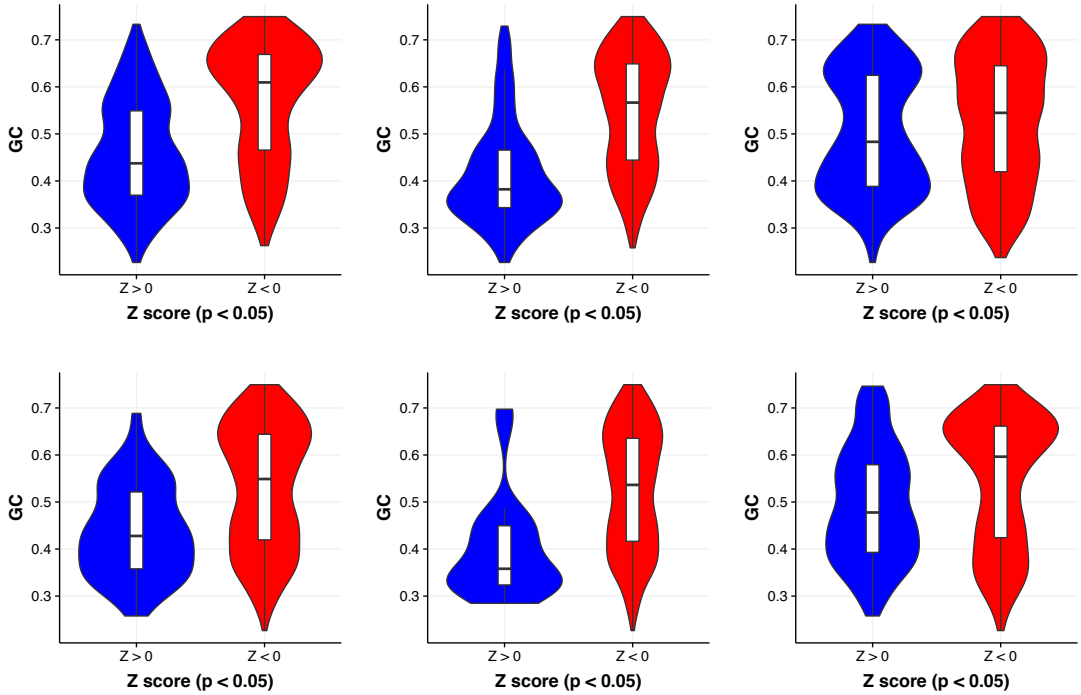
As with other models, we have to recognise several limitations to this model. Selection pressures on the CDS resulting in local synonymous codon biases, for example to reduce 5' mRNA stability (Qing et al. 2003; Kudla et al. 2009; Gu et al. 2010; Bentele et al. 2013; Goodman et al. 2013) are likely to be stronger than for including an OSC. Selection for synonymous codons that encode OSCs is likely to be limited to sequence sites without additional requirements. This model also assumes sequences permit flexibility between synonyms from two coding blocks, which is unlikely to occur given a codon change of this type requires mutations at two positions of the codon. However, evidence from this model is not consistent with predictions for OSC selection.



Supplementary Result 2 Figure 1: Correlations between OSC excesses (Z) and GC content, when all genome stop codons are considered together, are significantly negative for each reading frame for a model in which synonymous codons are randomly simulated. Violin plots emphasise that genomes with significant excesses are typically AT-rich.



Supplementary Result 2 Figure 2: Correlations between genomes excesses (Z) and GC content are significantly negative for all stop codons in each reading frame when coding sequences are simulated by randomising synonymous codons and permitting changes between codon blocks.



Supplementary Result 2 Figure 3: Violin plots for OSC excesses in the each of the reading frames for the synonymous codon model. GC content of genomes with significant positive excesses are similar to those found in the codon shuffle model. Unlike the codon shuffle model, the GC content of genomes with significant positive excesses of TGA are more biased towards the AT-rich genomes, with only a subset of GC-rich genomes have excesses in the +1 frame.

Supplementary Result 3

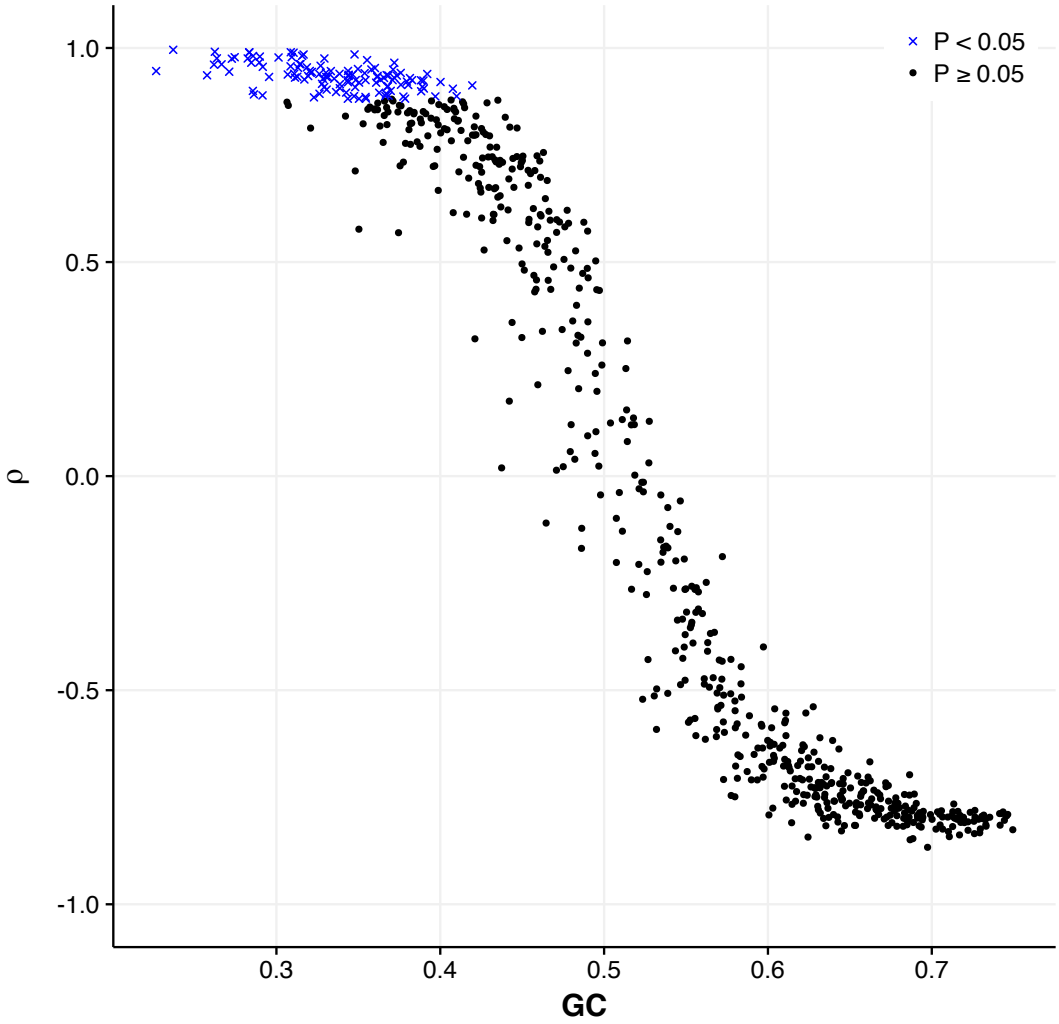
We can ask similar questions concerning localised synonymous site selection for OSCs using the sequences of amino acids repeats that provide the opportunity to encode +2 OSC, using asparagine for TAA (AAT, AAC), serine for TAG (AGT, AGC) and aspartic acid TGA (GAT, GAC). In these cases, site 3 T usages should be increased, although we are unable to control for GC3 content. However, T is used significantly less at site 3 in all cases (+2 TAA: $P < 2.2 \times 10^{-16}$; +2 TAG: $P < 2.2 \times 10^{-16}$; +2 TGA: $P < 5.911 \times 10^{-9}$, paired Wilcoxon rank sum tests). Moreover, correlations between GC3 content and log T3:T6 ratios are significantly negative in each case (+2 TAA: $\rho = -0.642$, $P < 2.2 \times 10^{-16}$; +2 TAG: $\rho = -0.636$, $P < 2.2 \times 10^{-16}$; +2 TGA: $\rho = -0.513$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlations).

We can test an overall hypothesis that synonymous codon usage is biased towards codons that generate OSCs if the following codon will allow by considering one-tailed tests all +1 and +2 contexts. This hypothesis is not supported ($P \approx 1$, Fisher's method combining one-tailed paired Wilcoxon rank sum tests). Thus, after minimising the potential effects that localised contexts may have had on our models, our evidence provides little support for any consistent genome wide OSC selection pressure, restricted to +1 TAA contexts.

Supplementary Result 4

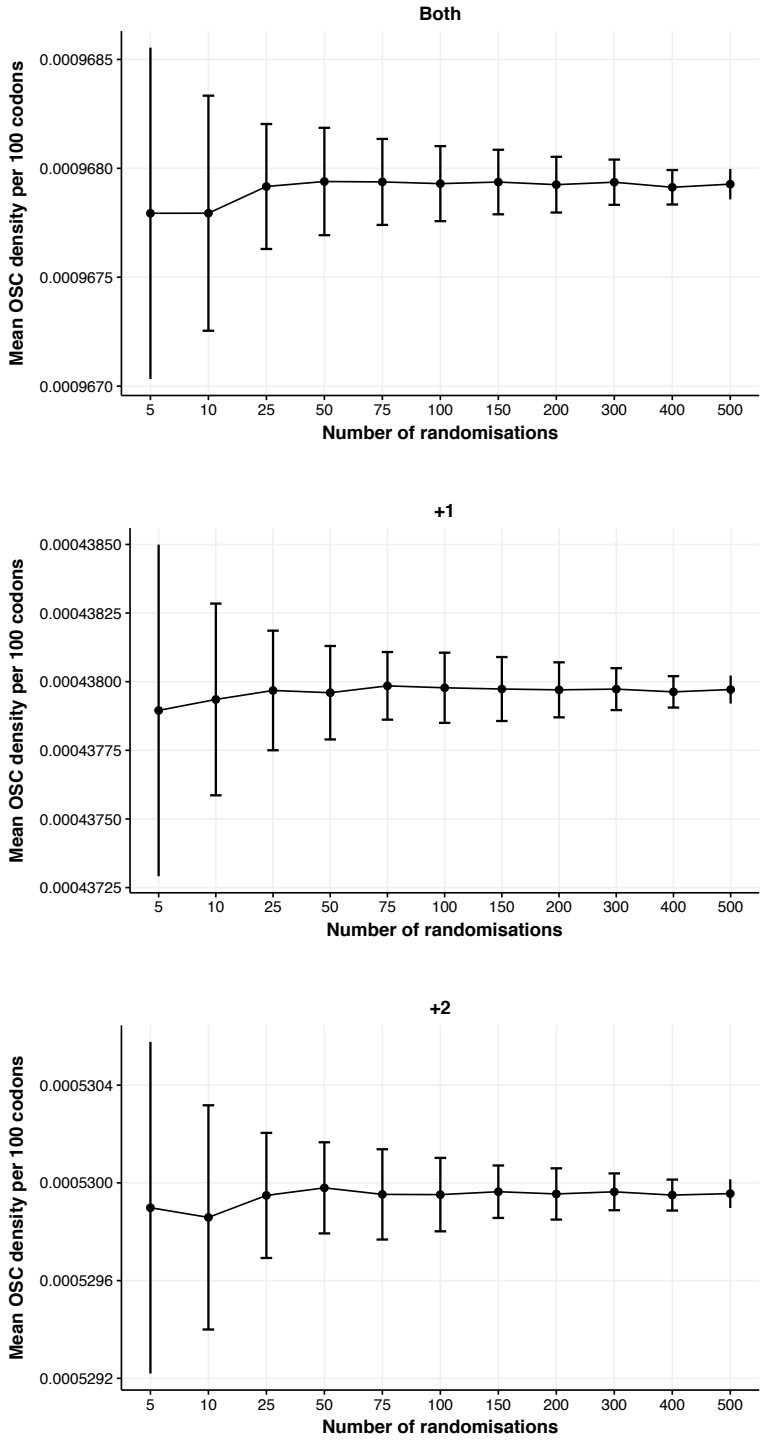
Significant positive correlations between codon contribution to hidden stops are limited to AT-rich genomes following multiple correction testing and predicted by genome GC content

Although it is has been shown that the method Seligmann and Pollock (2004) use for detecting OSC excesses is not appropriate (Morgens et al. 2013), we replicated the analysis with our genome sample. Positive correlations were identified in 341/694 (49.13%) genomes of which 201 (28.96%) are significant, although we also find 141 (20.32%) significant negative correlations. However, Seligmann and Pollock (2004) make no mention of correction for multiple comparisons. When we perform such correction, we find only 121 (17.44%, FDR correction) genomes maintain significant positive correlations (Supplementary Result 4 Figure 1). Thus, the original evidence underpinning the ambush hypothesis is limited and weakened further after such control. Results therefore suggest that not only is the evidence for selection for OSCs determining codon usage weak (positive correlation in less than half of genomes) and inappropriate, but the strength of results are further weakened by further statistical analyses.

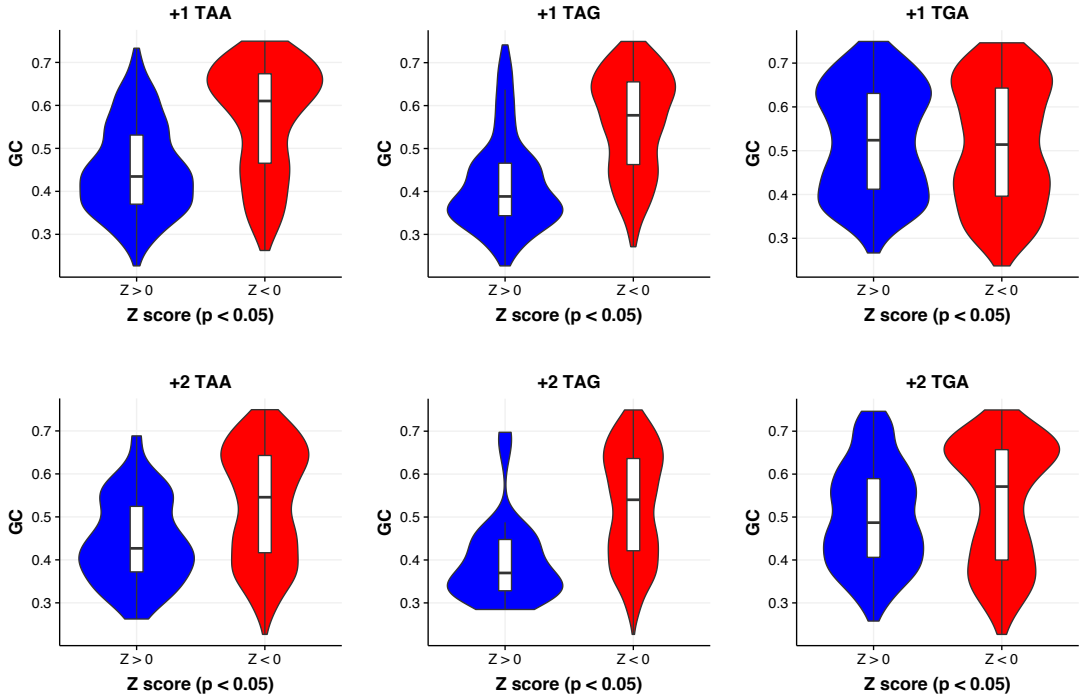


Supplementary Result 4 Figure 1: Significant positive Spearman's rank correlations between genome codon usage and codon contribution to OSCs are highly restricted to AT-rich genomes after correction for multiple comparisons.

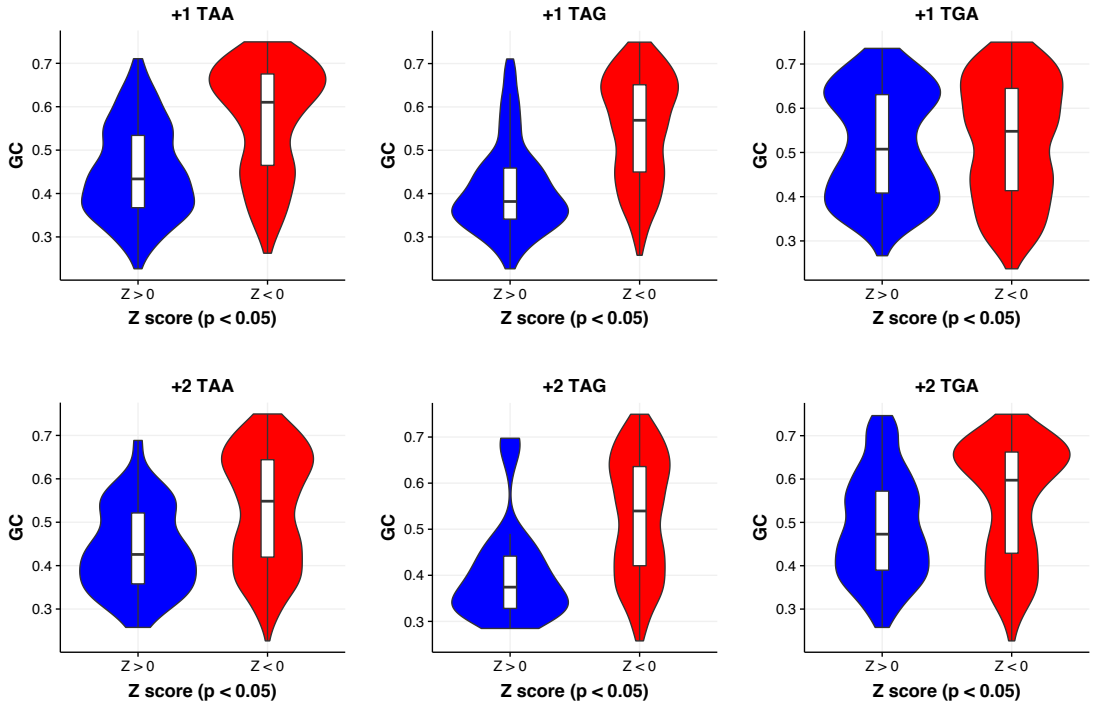
Supplementary Figures



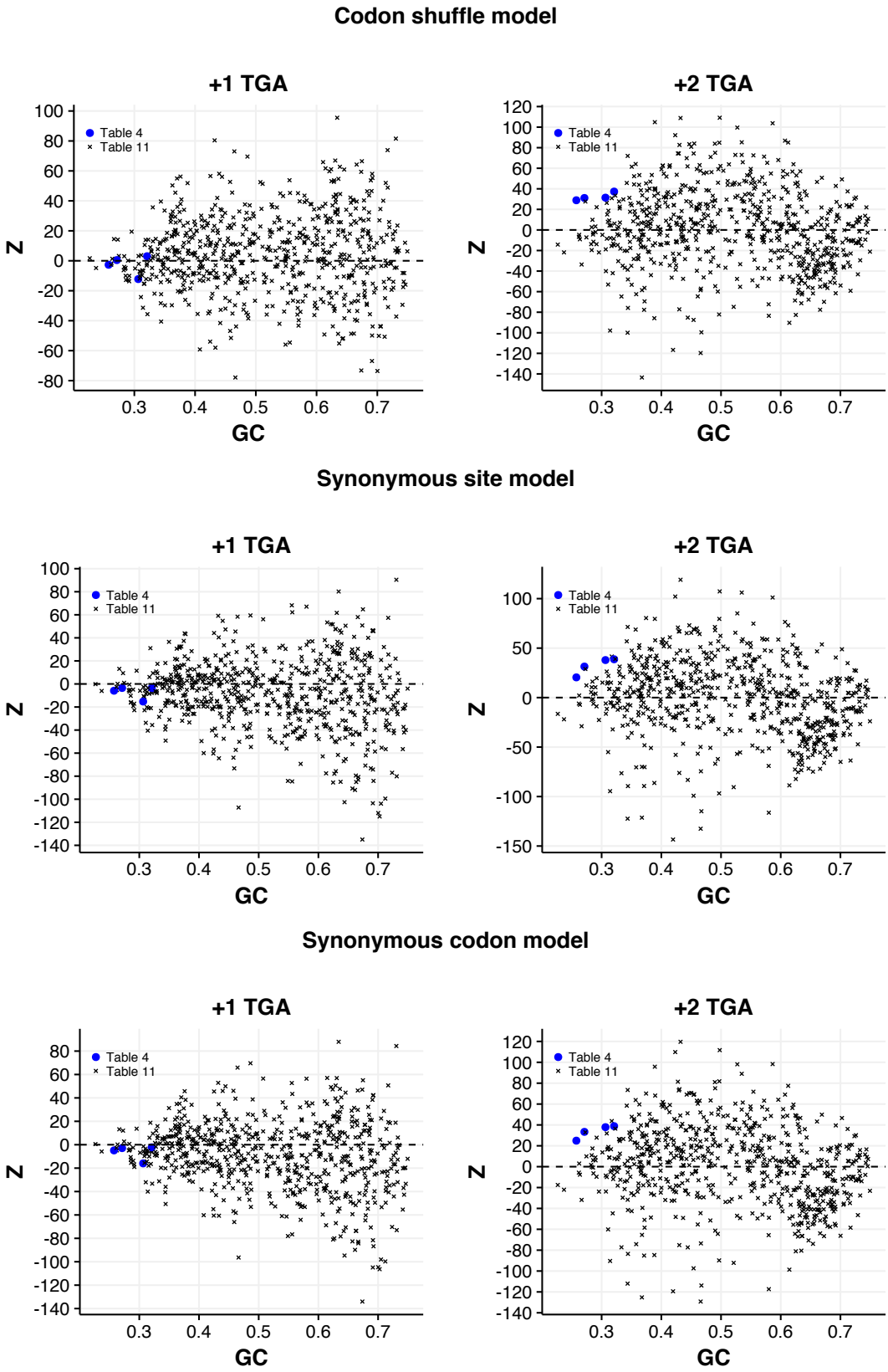
Supplementary Figure 2: Mean OSC densities for *E*. coli in the codon shuffle model. Mean densities vary little beyond 100 repeats.



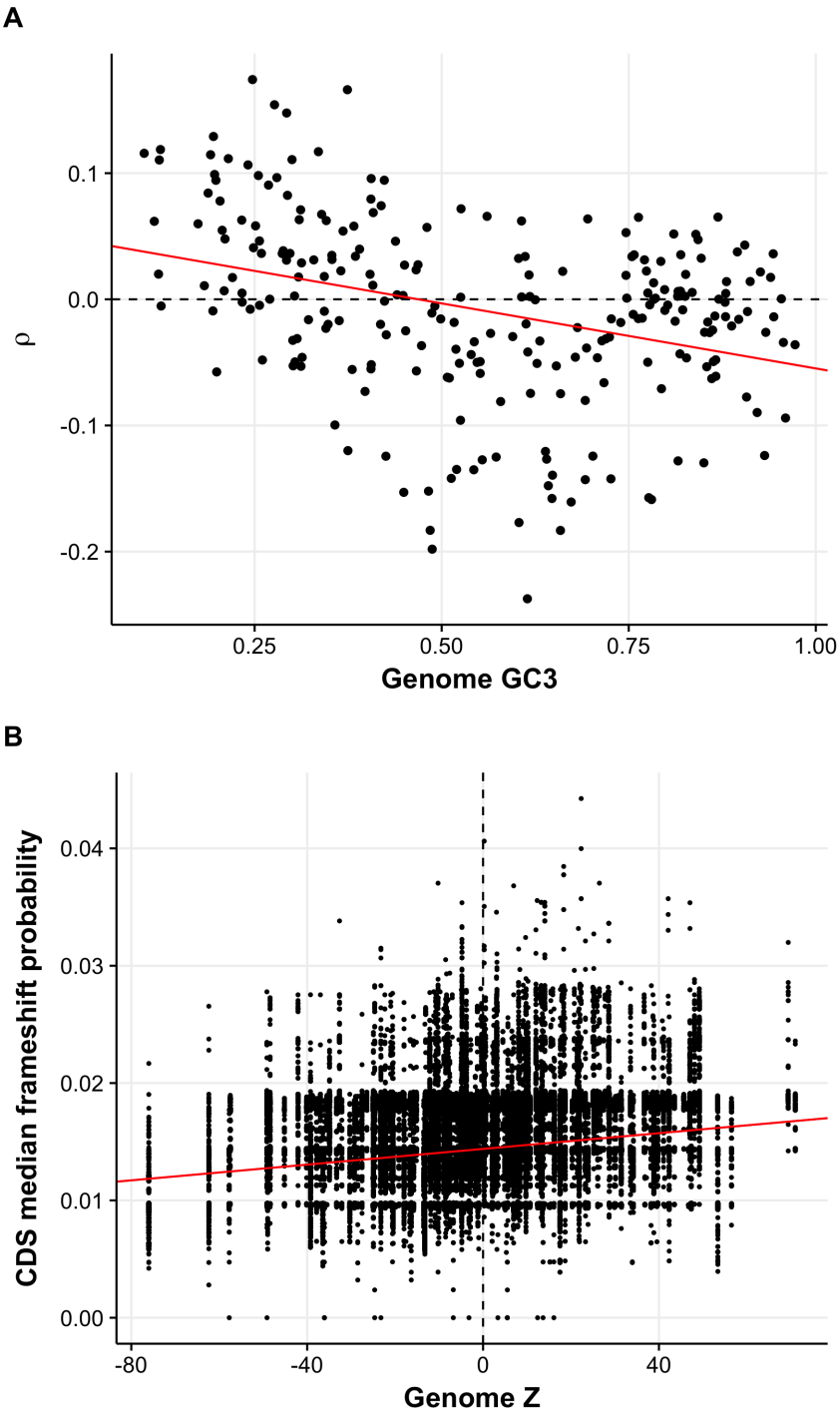
Supplementary Figure 2: Violin plots for OSC excesses in the each of the reading frames for the codon shuffle model. Genomes with significant positive excesses are typically AT-rich, particularly for TAG in all reading frames. Interestingly, the GC content of genomes with significant excesses of TGA are more similar to those without significant excess, suggesting that selection to incorporate off-frame TGA can overcome the restrictions of reduced AT-rich codons that make up OSCs in GC-rich genomes.



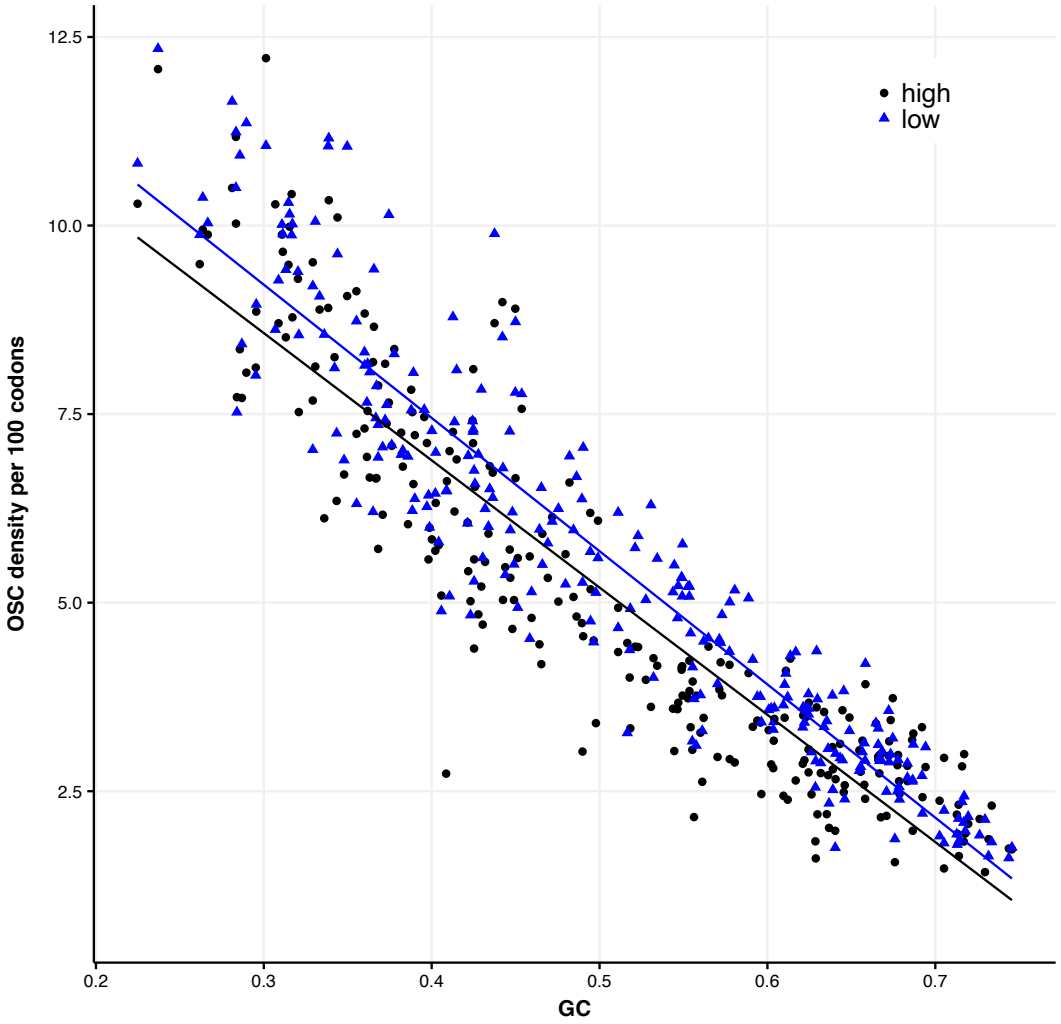
Supplementary Figure 3: Violin plots for OSC excesses in the each of the reading frames for the synonymous site model. GC content of genomes with significant positive excesses are extremely similar to those found for the codon shuffle model. Significant positive excesses are skewed towards the AT-rich genomes.



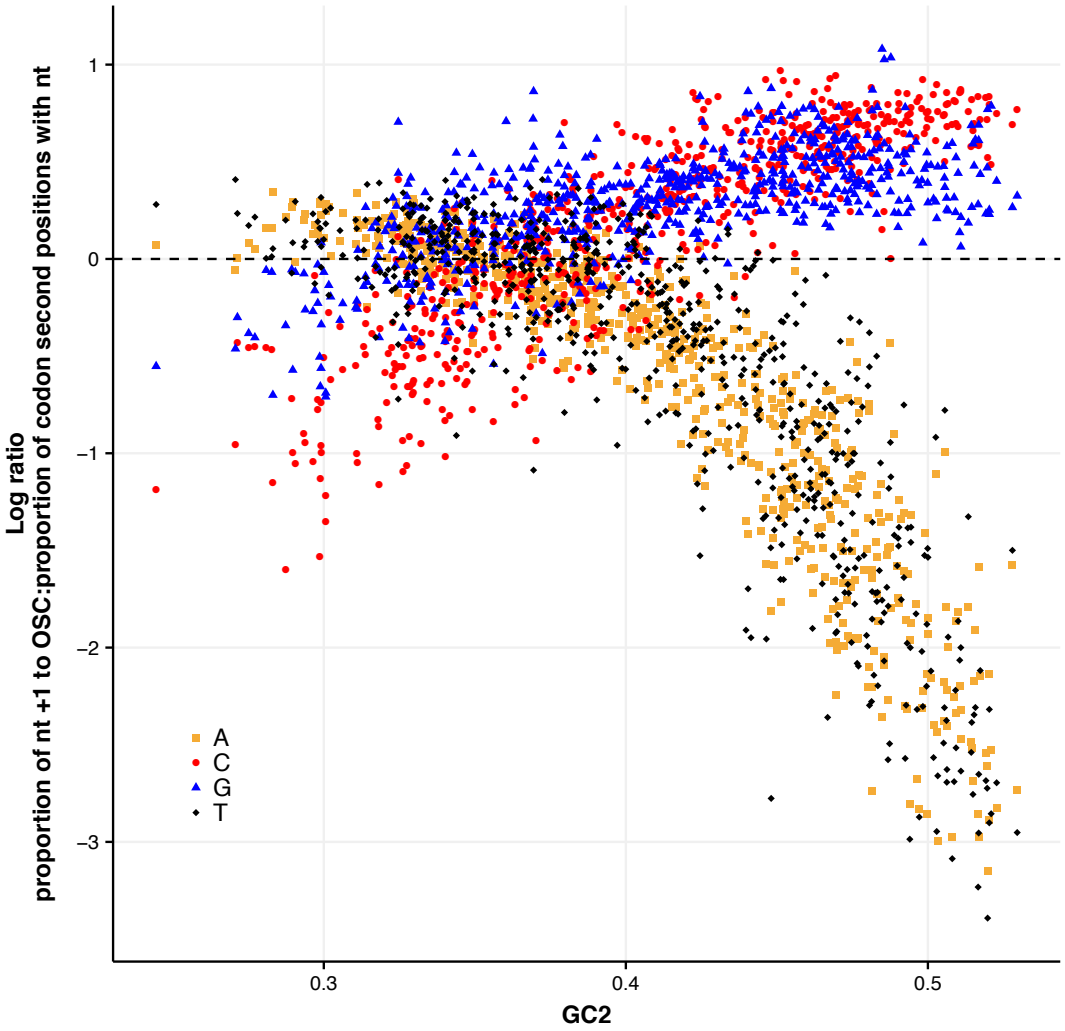
Supplementary Figure 4: Off-frame TGA densities of table 4 genomes for each simulation model. Table 4 genomes appear to have less +1 TGA than expected in the +1 frame.



Supplementary Figure 5: A) Correlation between genome GC3 content and the genome correlation between the median cost of frameshifting and OSC density for each CDS. Genomes with a positive correlation between OSC density and frameshift costs are typically AT-rich, B) Genomes with excesses of OSCs for the codon shuffle model tend to have higher +1 frameshifts probabilities.



Supplementary Figure 6: OSC densities in the +1 frame for genes with higher CAI (as a proxy for gene expression) are lower than for genes with lower CAI. These differences are significantly different ($P = 1.214 \times 10^{-9}$, Kruskal-Wallis rank sum test of loess regression residuals). This supports the hypothesis that highly expressed genes are less susceptible to frameshifting and therefore requiring less OSCs.



Supplementary Figure 7: Log ratios comparing nucleotide use after a +1 OSC to nucleotide use in second codon position reveals no nucleotide bias.

Supplementary Tables

Supplementary Table 1: A summary of the extended set of table 4 genomes

Genus	Count	Percentage
Mycoplasma	76	81.72
Spiroplasma	9	9.68
Ureaplasma	4	4.30
Mesoplasma	2	2.15
Candidatus Mycoplasma	2	2.15

Supplementary Table 2: *P*-values for the Kruskal-Wallis rank sum test of residuals from a loess regression comparing codon densities in genomes using translation table 4 and table 11 when *Mycoplasma* genomes have been restricted. Mean residuals (MR) for the two translation tables are also shown.

Codon(s)	**Kruskal-Wallis *P*-value**	**Table 4 MR**	**Table 11 MR**
Both frames, combined OSCs	$< 2.2 \times 10^{-16}$	-4.083	0.197
+1, combined OSCs	$< 2.2 \times 10^{-16}$	-1.704	0.091
+2, combined OSCs	3.972×10^{-16}	-2.379	0.107
+1 TAA	2.848×10^{-4}	-0.271	0.017
+1 TAC	1.786×10^{-6}	-0.192	0.008
+1 TAG	5.839×10^{-7}	-0.440	0.033
+1 TAT	4.073×10^{-9}	-0.316	0.013
+1 TGA	0.032	-0.133	0.003
+1 TGC	0.249	-0.097	0.001
+1 TGG	0.257	-0.125	-0.002
+1 TGT	0.196	-0.049	-0.001
+2 TGA	7.77×10^{-5}	0.493	-0.030

Supplementary Table 2: Summary of expectations and results for each model.

Model	Expectations	Results
Codon randomisation within CDS	• Significant excess of OSCs when compared with the null (OSCs present due to chance dicodons) • Positive correlations between genome OSC excesses and GC content • Greater positive deviations from the null for OSCs when compared with sense codons • OSC excesses biased towards more efficient stop codons	• Number of genomes with significant excesses: max = 53.31% (+1 TGA), min = 6.34% (+2 TAG) • Significant excesses predominantly in the +1 frame • All OSCs with significant negative correlations with GC except +1 TGA ($\rho = 0.036$, $P = 0.348$) • Strong AT-bias for genomes with significant excesses, for each OSC • Greater excesses of TAC, TAT, TGC (+1, +2) and TGT (+2) • Excesses rank TGA > TAA > TAG (+1, +2)
Synonymous site randomisation within coding blocks	• Significant excesses of OSCs when compared with null (synonymous sites are not under selection to encode OSCs) • Positive correlations between genome OSC excesses and GC content • Greater positive deviations from the null for OSCs when compared with sense codons • OSC excesses biased towards more efficient stop codons	• Number of genomes with significant excesses: max = 52.59% (+2 TGA), min = 6.05% (+2 TAG) • Significant excesses predominately in the +1 frame • All OSCs with significant negative correlations with GC • Strong AT-bias for genomes with significant excesses, for each OSC • Greater excesses of TAC, TAT, TGC +1, +2), TGG (+1) and TGT (+2) • Excesses rank TAA > TGA > TAG (+1), TGA > TAA > TAG (+2)
Synonymous codon	• Significant excesses of OSCs when compared	• Number of genomes with significant excesses: max =

randomisation permitting interchange between codon blocks	with null (synonymous codon use is not determined by the ability to encode an OSC) • Positive correlations between genome OSC excesses and GC content • Greater positive deviations from the null for OSCs when compared with sense codons • OSC excesses biased towards more efficient stop codons	52.02% (+2 TGA), min = 6.20% (+2 TAG) • Significant excesses predominantly in the +1 frame (Supplementary Result 2) • All OSCs with significant negative correlations with GC • Strong AT-bias for genomes with significant excesses, for each OSC • Greater excesses of TAC, TAT, TGC (+1, +2), TGG (+1), TGT (+2) • Excesses rank TAA > TGA > TAG (+1), TGA > TAA > TAG (+2)
OSC encoding amino acid repeats	• Significant increase in use of synonyms that encode OSCs for the first codon when compared with the second (which strictly cannot encode an OSC) • Positive correlations between GC content and the site 3:site 6 ratio of use of the OSC facilitating nucleotide	• Only in the case of +1 TAA (isoleucine repeat) and only when synonymous site restricted to A/T • Only +1 TAA with a significant positive correlation after restriction to only A/T at synonymous sites
Table 4 genome comparison	• Reduced off-frame TGA densities in table 4 genomes • Possible increased compensatory TAA and TAG off-frame densities in table 4 genomes	• Reduced +1 TGA, TAA and TAG densities in table 4 genomes • Reduced densities of all +1 TAN codons in table 4 genomes • Reduced densities of +1 TGR codons in table 4 genomes • Increased +2 TAA, TAG and TGA densities in table 4 genomes

References

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* 9:675.

Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342:475-479.

Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* 6:e1000664.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255-258.

Morgens DW, Chang CH, Cavalcanti AR. 2013. Ambushing the Ambush Hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. *BMC Genomics* 14:418.

Qing G, Xia B, Inouye M. 2003. Enhancement of translation initiation by A/T-rich sequences downstream of the initiation codon in Escherichia coli. *J. Mol. Microbiol. Biotechnol.* 6:133-144.

Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23:701-705.

Tse H, Cai JJ, Tsoi HW, Lam EP, Yuen KY. 2010. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics* 11:491.

Chapter 5

Adenine Enrichment at the Fourth CDS Residue in Bacterial Genes Is Consistent with Error Proofing for +1 Frameshifts

Liam Abrahams and Laurence D. Hurst

Molecular Biology and Evolution (2017) 34(12):3064-3080

This chapter contains analysis of publicly available data. The data and custom scripts 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://academic.oup.com/journals/pages/access_purchase/rights_and_permissions/publication_rights).

This declaration concerns the article entitled:			
Adenine Enrichment at the Fourth CDS Residue in Bacterial Genes Is Consistent with Error Proofing for +1 Frameshifts			
Publication status (tick one) **Draft manuscript** [ ] **Submitted** [ ] **In review** [ ] **Accepted** [ ] **Published** [X]			
Publication details (reference)	Liam Abrahams, Laurence D Hurst, Adenine Enrichment at the Fourth CDS Residue in Bacterial Genes Is Consistent with Error Proofing for +1 Frameshifts, Molecular Biology and Evolution, Volume 34, Issue 12, December 2017, Pages 3064–3080, https://doi.org/10.1093/molbev/msx223		
Copyright status (tick the appropriate statement) I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [X]			
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% Bioinformatics 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**	

Adenine Enrichment at the Fourth CDS Residue in Bacterial Genes Is Consistent with Error Proofing for +1 Frameshifts

Liam Abrahams*[,1] and Laurence D. Hurst[1]
[1]Department of Biology and Biochemistry, The Milner Centre for Evolution, University of Bath, Bath, United Kingdom
*Corresponding author: E-mail: l.abrahams@bath.ac.uk.
Associate editor: Claus Wilke

Abstract

Beyond selection for optimal protein functioning, coding sequences (CDSs) are under selection at the RNA and DNA levels. Here, we identify a possible signature of "dual-coding," namely extensive adenine (*A*) enrichment at bacterial CDS fourth sites. In 99.07% of studied bacterial genomes, fourth site *A* use is greater than expected given genomic *A*-starting codon use. Arguing for nucleotide level selection, *A*-starting serine and arginine second codons are heavily utilized when compared with their non-*A* starting synonyms. Several models have the ability to explain some of this trend. In part, *A*-enrichment likely reduces 5′ mRNA stability, promoting translation initiation. However *T*/*U*, which may also reduce stability, is avoided. Further, +1 frameshifts on the initiating ATG encode a stop codon (TGA) provided *A* is the fourth residue, acting either as a frameshift "catch and destroy" or a frameshift stop and adjust mechanism and hence implicated in translation initiation. Consistent with both, genomes lacking TGA stop codons exhibit weaker fourth site *A*-enrichment. Sequences lacking a Shine–Dalgarno sequence and those without upstream leader genes, that may be more error prone during initiation, have greater utilization of *A*, again suggesting a role in initiation. The frameshift correction model is consistent with the notion that many genomic features are error-mitigation factors and provides the first evidence for site-specific out of frame stop codon selection. We conjecture that the NTG universal start codon may have evolved as a consequence of TGA being a stop codon and the ability of NTGA to rapidly terminate or adjust a ribosome.

Key words: **frameshift, error mitigation, dual coding, fourth site, translation initiation.**

Article

Introduction

A simplistic model of protein-coding gene evolution assumes that amino acid composition is a reflection of selection optimizing the biochemical function of the encoded protein. Consistent with such a model, domains or individual positions critical to protein function are under strong purifying selection (Guo et al. 2004; Furlong and Yang 2008; Gray and Kumar 2011; McFerrin and Stone 2011). Such is the strength of selection on particular amino acids that methods predicting protein domain function from amino acid content are of great utility (Al-Shahib et al. 2007; Sankararaman et al. 2009).

We are becoming increasingly aware of selection pressures beyond those specifying the amino acid sequence acting on coding sequence (CDS) composition. For example, eukaryotic exonic splice enhancers (ESEs) are purine-rich binding-site motifs found at exon ends assisting recruitment of the splicing machinery by regulatory proteins (Blencowe 2000; Graveley 2000; Cartegni et al. 2002; Zhou and Fu 2013). Consequently, codon and amino acid content toward exon ends is biased (Willie and Majewski 2004; Chamary and Hurst 2005a; Parmley and Hurst 2007; Caceres and Hurst 2013) with nonsynonymous and synonymous mutations in ESEs under purifying selection (Fairbrother et al. 2004; Xing and Lee 2005; Carlini and Genut 2006; Parmley et al. 2006; Wu and Hurst 2015). More generally, RNA binding proteins of all flavors impose purifying selection on CDSs (Savisaar and Hurst 2017). There are claims that the CDS is under selection to bind transcription factors (Stergachis et al. 2013), although these are contested (Xing and He 2015; Agoglia and Fraser 2016). Selection might be for avoidance of, rather than selection for, certain motifs, such as intra-CDS Shine–Dalgarno (SD)-like sequences (Diwan and Agashe 2016; Yang et al. 2016), or motifs for RNA binding proteins that bind to introns are avoided within CDSs (Savisaar and Hurst 2017).

A common fingerprint of additional CDS functionality is biased codon usage. Aside from selection for ESEs, codon choice is thought to be affected by, for example, translational selection (Behura and Severson 2011; Doherty and McInerney 2013; Ma et al. 2014), the positioning of nucleosomes (Warnecke et al. 2008; Cohanim and Haran 2009; Prendergast and Semple 2011) and cotranslational protein folding (Zhang et al. 2009; Yu et al. 2015; Buhr et al. 2016). Both RNA and protein structural effects may influence the selection for differential nucleotide content (Chamary and Hurst 2005b; Meyer and Miklós 2005; Shabalina et al. 2006; Gu et al. 2010; Smith et al. 2013; Babbitt et al. 2014). Additionally, intra-CDS microRNA (miRNA) pairing can also impose purifying selection on synonymous mutations in miRNA target sites but, given the span of such binding sites, it is likely they affect nonsynonymous mutations too (Hurst 2006; Forman et al. 2008; Guo et al. 2008; Liu et al. 2015).

The great majority of the above additional levels of information have been identified via hypothesis led approaches (e.g., if ESEs impose selective constraints, we should see ESE-associated synonymous sites conserved at exon ends). An alternative approach is to explore unusual codon or amino acid patterns as strong signals might act as excellent guides to features that are *a priori* important for the operation of cells. Here we highlight one such feature: in bacteria, there is a common bias at CDS fourth sites (i.e., immediately after the initiating codon) for amino acids whose codons start with adenine (*A*). The prevalence of *A*-starting second codons and positive influence on expression has previously been described (Looman et al. 1987; Stenstrom et al. 2001; Zalucki et al. 2007; Zamora-Romo et al. 2007), although these studies were only conducted in *Escherichia coli*. A large-scale multi-genome analysis by Tang et al. (2010) identified a preference for *A* in the first position and *C* in the second position of the second codon, but provided no context as to why the fourth site *A* bias may occur.

We begin by establishing how common bacterial fourth site *A* use is, asking whether it is simply explained by genome GC content influencing codon usage. We establish that the trend remains highly significant after such control in the great majority of bacterial genomes. In some cases, the bias is extraordinarily extreme (over 60% fourth site *A* usage in some genomes). We provide evidence that the fourth site is unusual, even compared with closer nucleotide neighbors. Consistent with strong selection on highly expressed genes, *A* usage is elevated in the most highly expressed genes (although the effect is not dramatic).

Having established that fourth site *A* enrichment is a common and potentially nontrivial feature, we propose and test a number of alternative hypotheses. We start by dismissing some possibilities and then consider three viable models: selection at the protein level requires an *A*-starting codon; RNA level selection minimizes 5′ mRNA domain secondary structures; or that fourth site *A* acts as an immediate trap for +1 frameshifted ribosomes (ATGA becomes TGA on a +1 frameshift). We find that RNA structural selection contributes some of the bias (enrichment is still observed in genomes that don't use TGA as a stop, but only to the level of enrichment seen downstream), however the frameshift correction model makes for a parsimonious explanation. To the best of our knowledge, this frameshift hypothesis is novel and extends the current understanding of the role of out of frame stop codons, providing the first evidence for site-specific selection of stop codons out of frame. This preference for *A* at the fourth site may, in addition, have become canalized and so feature as part of the start codon recognition mechanism. It is also possible that usage of TGA as a stop codon may also have been related to the evolution of NTG as a start codon.

Results

Fourth Site *A* Enrichment Is Common, Sometimes Extreme and Exceptional

Controls for Nucleotide Content Confirm a Common and Sometimes Extreme Enrichment of A at CDS Fourth Sites

Analysis of bacterial genomes CDSs indicates that in most genomes there is enrichment of fourth site *A* content (fig. 1). The most extreme is *Polaribacter* sp. in which 63.26% of CDSs have *A* at the fourth site. To control for genomic GC effects, we performed a ratio test (see Materials and Methods) comparing the nucleotide usage in the first position of the second codon with nucleotide usage at the first position for all codons in genome. Ratios equal to 1 signify *A*-starting second codons are used proportionately to *A*-starting codons within the genome. We find a remarkable 640/646 genomes (99.07%) have an A_4 ratio significantly >1 ($P < 0.01$, Pearson's cumulative test statistic [χ^2], Bonferroni correction). In comparison, 31/646 (4.80%), 3/646 (0.46%), and 55/646 (8.51%) and genomes have C_4, G_4, and T_4 ratios >1, respectively, confirming fourth site enrichment is specific to *A* and not attributable to GC biases. This exceptionalism of the fourth site is further illustrated by the striking reduction in fourth site GC variation (supplementary fig. S1, Supplementary Material online).

Fourth Site A Is Conserved

Genomes with high "silent" GC content (GC3) tend to more readily employ the amino acids with GC rich nonsynonymous sites (Warnecke et al. 2010). This shift in amino acid content we term GC "pressure." If the usage of *A* at fourth sites is functionally relevant we would expect its usage to be more resilient to GC pressure than for *A*-starting codons within the genome. Comparing genomic GC3 with both the proportion of *A*-starting second codons and all *A*-starting codons (fig. 2), we observe that the regression coefficient for all *A*-starting codons (−0.245) is significantly more negative than for *A*-starting second codons (−0.160) ($P = 7.056 \times 10^{-19}$, $Z = 8.874$, two-tailed *Z*-test of equivalency) and thus *A* at the fourth site is more resilient to genomic GC pressures.

Further evidence of functionality arises from analysis of the conservation of fourth site *A* between *E. coli* and *Shigella flexneri*. *E. coli* and *Shigella* spp. are closely related (Pupo et al. 2000; Zuo et al. 2013), demonstrating high nucleotide similarity between species (Goris et al. 2007). *Shigella* spp. undergo accelerated gene loss when compared with *E. coli*, in part explained by weakened purifying selection associated with reduced effective population size (N_e) (Hershberg et al. 2007; Balbi et al. 2009). Thus, if there is selection at the fourth site, by focusing on *E. coli* residues we can ask whether fourth site *A* is particularly resilient to substitution to an alternative nucleotide under weaker purifying selection by comparing with a lower N_e comparator for which purifying selection, as a result of reduced N_e, will be less effective in purging deleterious substitutions. If the fourth site is under particularly strong selection, we expect substitutions at the fourth site to be reduced when compared with other sites. We find the proportion of CDSs differing from *A* at the fourth site in *S. flexneri* is lower than for other nucleotides (fig. 3). This result assumes the *E. coli* state to be more reflective of the ancestral state, particularly as the low N_e genome is expected to have a higher rate of change. Although other first codon positions demonstrate a relative reduction away from an *A*-genotype when compared with other nucleotides, loss of *A* in the fourth position is significantly reduced compared with

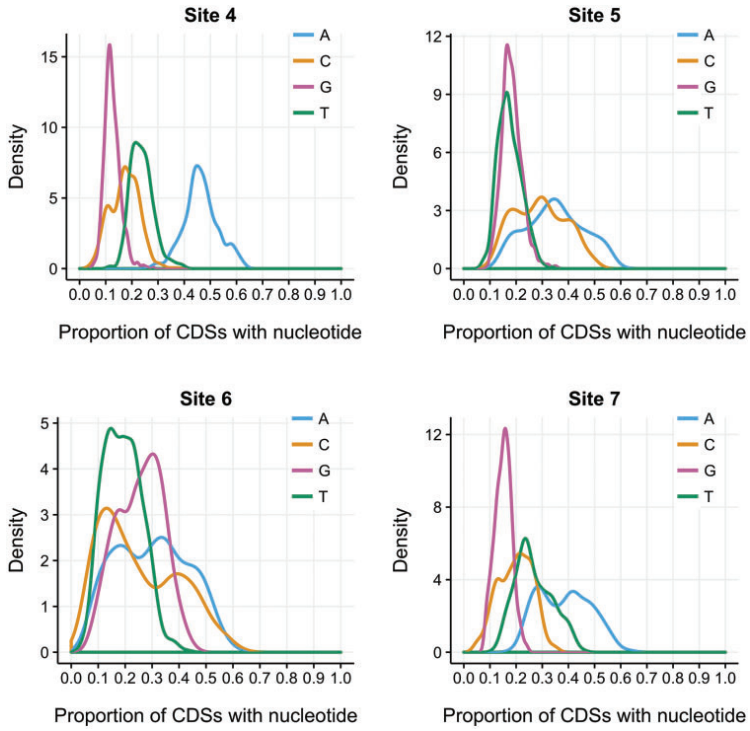


Fig. 1. Kernel density plots showing the proportion of coding sequences with each nucleotide (*A*, *C*, *G*, *T*) at coding sequence sites 4, 5, 6, and 7 (site 1 is defined as the first nucleotide of the start codon). Site 4 demonstrates a clear preference for *A* which is not observed at the other sites.

downstream positions ($P < 0.001$, one-sample *T*-test). This lack of change specific to the fourth site *A* genotype is indicative of purifying selection at the fourth site.

More Highly Expressed Genes Have Higher Fourth Site A Content

Selectively relevant features are often more pronounced in highly expressed genes (Urrutia and Hurst 2003; Doherty and McInerney 2013). To assay expression level, we consider the Codon Adaptation Index as a surrogate. For genomes in which suitable annotations were available, we compared the mean CAI for genes with and without fourth site *A* (N.B. this paired test controls for residual effects such as intergenome GC variation). We find a significantly higher CAI for genes with fourth site *A* ($P = 1.042 \times 10^{-12}$, $N = 232$, paired Wilcoxon rank-sum test), although the mean CAI value for CDSs with fourth site *A* (0.586 $\pm$ 0.088, $N = 232$) is only slightly greater (0.582 $\pm$ 0.088, $N = 232$) than for those without. Performing the test in the opposite direction, we find a significant increase ($P = 0.034$, Wilcoxon rank-sum test) in the proportion of CDSs with fourth site *A* in the highly expressed genes (0.457 $\pm$ 0.07, $N = 232$) compared with those less expressed (0.454 $\pm$ 0.082, $N = 232$).

The above result is most pronounced in high GC genomes. Genomes with extreme GC compositions demonstrate a reduced range of mean CAI values (supplementary fig. S2, Supplementary Material online) (Botzman and Margalit (2011) with codon usage in many CDSs similar to that for the ribosomal proteins. Repeating the same analyses for just 30 genomes with 20% $\leq$ GC3 $\geq$ 90% (supplementary fig. S2, red, Supplementary Material online) (reducing the mean CAI range to 0.576–0.743) we find mean CAI values for CDSs with fourth site *A* significantly higher ($P = 1.486 \times 10^{-6}$, paired *T*-test, $N = 30$) but again the difference in mean CAI in CDSs using *A* (mean CAI = 0.661 $\pm$ 0.034, $N = 30$) and non-*A* (mean CAI = 0.650 $\pm$ 0.036, $N = 30$) is small. For the GC-rich genomes, we find a significant difference in mean CAI ($P = 4.451 \times 10^{-10}$, paired *T*-test, $N = 18$) for CDSs using fourth site *A* (mean CAI = 0.581 $\pm$ 0.089, $N = 18$) when compared with those that do not *A* (mean CAI = 0.581 $\pm$ 0.089, $N = 18$). However, for AT-rich genomes mean CAI values are not significantly different between those using fourth site *A* and those not ($P = 0.243$, paired *T*-test, $N = 12$). These results suggest that fourth site *A* is more commonly utilized in highly expressed genes, albeit to a small degree, and even maintained under extreme GC restrictions. However, when conditions are inherently conducive to

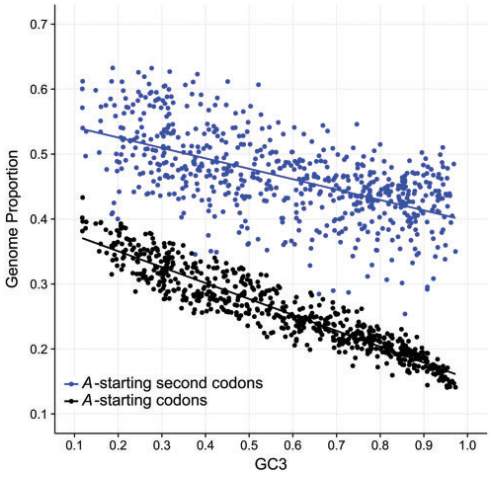


FIG. 2. The proportion of coding sequences with fourth site *A* is maintained above the proportion of *A*-starting codons as GC content increases. The regression coefficient for all *A*-starting codons is significantly greater than for *A*-starting second codons ($P = 7.056 \times 10^{-19}$, $Z = 8.874$, two-tailed *Z*-test of equivalency), suggesting enrichment of *A* at the fourth site becomes stronger with increasing GC content.

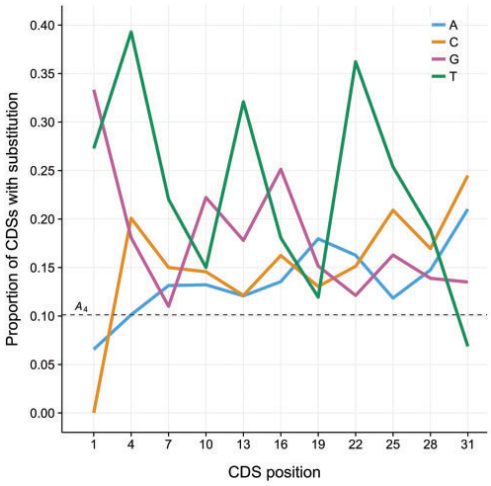


FIG. 3. The proportion of *Shigella flexneri* orthologs with a substitution of each nucleotide at the first position of codons from *Escherichia coli*. The proportion of sequences with a substitution from *A* at site 4 is displayed with the dotted line. Position 1 of the first codon demonstrates minimal variation away from an *A*-genotype confirming the preference for an ATG start codons. Substitutions from an *A*-genotype are reduced across the sites when compared with other nucleotides. The proportion of coding sequences with a change from *A* in codon 2 is significantly lower than neighboring codons ($P < 0.001$, one-sample *T*-test), suggesting fourth site *A* is under strong selection.

incorporating an *A*-starting second codon, we expect *A*-starting second codons to be used regardless of alternative selection pressures and therefore any enrichment signal is harder to detect.

Three Models to Explain Selection for Fourth Site *A* Content

Our results thus far all support the exceptionalism of the fourth site. Why might this be? The 5′ CDS is known to have distinct selection pressures to those acting on the remainder of the CDS. Although 5′ ends are enriched with nonoptimal codons (Tuller, Carmi et al. 2010; Pechmann and Frydman 2013; Tuller and Zur 2015), Bentele et al. (2013) have demonstrated that in bacteria selection favors codons that reduce mRNA folding around the translation start, regardless of whether these codons are frequent or rare. Notably, when a nonoptimal codon is GC-rich, they find preferences for optimal AT-rich codons. Thus, the trend is not explained by selection for nonoptimality (also concluded by Eyre-Walker and Bulmer 1993) but AT-content and therefore we do not consider this a selection pressure. An alternative explanation could be the presence of overlapping genes: a CDS employing the TGA stop codon overlapping a downstream CDS by four nucleotides will result in an *A* nucleotide in the fourth position of the subsequent CDS. However, after removing 165,357/2,173,531 (7.61%) CDSs with these four site overlaps, 635/646 (98.30%) genomes achieve an A_4 ratio >1 ($P < 0.01$, Pearson's cumulative test statistic (χ^2), Bonferroni correction) and therefore overlaps cannot account for the fourth site enrichment. Are there alternative explanations? We propose three possible models, which we proceed to test.

The Amino Acid Preference Model

Certain amino acids (lysine, serine) have been shown to be favored immediately following the start codon in both prokaryotes and eukaryotes (Shemesh et al. 2010) and evidence suggests that these amino acids may provide important functional roles (Stenstrom et al. 2001). Furthermore, Tats et al. (2006) and Bivona et al. (2010) note particular amino acids (alanine, cysteine, proline, serine, threonine, and lysine) may be used more frequently in the second position in highly expressed genes. These observations may be attributed to involvement of the second amino acid in posttranslational modifications. N-terminal methionine excision (NME) only occurs when the second amino acid is glycine, alanine, serine, threonine, cysteine, proline, or valine—amino acids with small side chains (Liao et al. 2004; Frottin et al. 2006; Ouidir et al. 2015). The second amino acid is implicated in the N-end rule pathway (overview in Tasaki et al. 2012), targeting proteins for degradation (Bachmair et al. 1986; Tobias et al. 1991) with the main determinants the amino acids not involved in NME (Varshavsky 2011). Signaling proteins requiring the inclusion of specific concentrations of hydrophobic amino acids (Ng et al. 1996) may also contribute to amino acid bias. A variety of protein-level selection pressures may therefore be acting upon the second amino acid.

If enrichment reflects protein-level selection on the second amino acid, we expect no difference in the use of *A*/non-*A* starting six fold degenerate amino acids as it is simply the amino acid, not the underlying nucleotide, that is important. We also expect other non-*A* starting amino acids to be favored given post-translational modification requirements.

The RNA Stability Model

Reducing secondary RNA structures in 5′ mRNA domains enhances the ability of the mRNA to interact efficiently with ribosomes and promotes translation efficiency (de Smit and van Duin 1990; Tuller, Waldman et al. 2010; Scharff et al. 2011). There indeed exists a relationship between 5′ mRNA folding strength and protein expression levels in prokaryotes and eukaryotes (Kudla et al. 2009; Li, Zheng, Ryvkin et al. 2012; Li, Zheng, Vandivier et al. 2012; Bentele et al. 2013; Goodman et al. 2013; Shah et al. 2013; Vandivier et al. 2013). Minimising the presence of these secondary structures, for example hairpin loops, by adopting destabilizing AT-rich 5′ domains (Qing et al. 2003; Kudla et al. 2009; Gu et al. 2010; Bentele et al. 2013; Goodman et al. 2013) could therefore promote more efficient translation by facilitating mRNA-ribosome interactions. Several studies have experimentally identified second codon AT preference promoting faster translation initiation (Zalucki et al. 2007) and correlating positively with expression levels (Stenstrom et al. 2001).

If reducing RNA stability can explain the fourth site *A* enrichment, we would expect enrichment at the fourth site to not be unique, but representative of neighboring codons in the 5′ mRNA binding domain. For instance, we would expect no significant difference between the fourth, seventh and tenth sites or between synonymous sites in these codons. Furthermore, if there is uniquely selection for increased AT-content to destabilize the RNA, we also expect to see a localized *T* enrichment.

The Frameshift Correction Model

Consider a CDS that starts NTGA, with *A* at the fourth site. Following a + 1 frameshift, this sequence becomes the TGA stop codon, immediately terminating or realigning translation and preventing the ribosome continuing on a + 1 reading frame (overview in fig. 4). We define this as the frameshift correction model, providing a novel and site-specific case of out of frame stop codons more generally.

This model presumes a + 1 frameshift is deleterious. Whilst viruses (Su et al. 2005; Melian et al. 2014), prokaryotes (Tsuchihashi and Kornberg 1990; Gupta et al. 2013) and eukaryotes (Wills et al. 2006; Belew et al. 2014) (reviewed in Caliskan et al. 2015) do employ frameshifting to encode multiple proteins from one mRNA strand (e.g., the *gag-pol* gene; Jacks et al. 1988), many ribosomal frameshifts are errors. Ribosomes leaving the correct reading frame and synthesizing proteins that were never "intended" are likely to incur cellular costs (Warnecke et al. 2010). For example, reduced ribosomal capability can be rate limiting for growth (Shachrai et al. 2010), whilst important cellular resources (tRNAs, amino acids) are misinvested. Furthermore, incorrectly folded mistranslated proteins may have an adverse effect on cellular interactions or form toxic aggregates (Tank and True 2009). The possible evolutionary advantage of capturing these frameshifts is conjectured to be reflected by an overrepresentation of out of frame stop codons, termed the "ambush hypothesis" (Seligmann and Pollock 2004; Singh and Pardasani 2009; Tse et al. 2010), although the frequency with which codons that form out of frame stops are used is largely predictable from the underlying GC pressure (Morgens et al. 2013). Alternatively, selection to reduce costs in genomes where frameshifting is most deleterious (notably GC rich ones) can explain the richer tRNA repertoire found in such genomes (Warnecke et al. 2010).

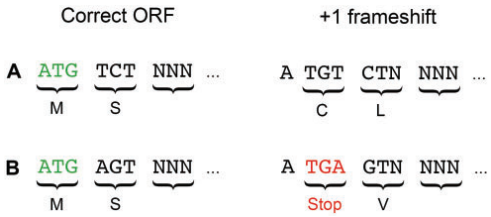


FIG. 4. A schematic representation of the frameshift correction model. Both CDSs encode methionine followed by serine and have identical GC content. However, following a +1 frameshift sequence A encodes a cysteine followed by a leucine, whereas translation of sequence B is immediately terminated by the presence of an out of frame TGA stop codon.

This +1 frameshift correction mechanism requires a NTG start codon. Prokaryotes are known to use a variety of non-ATG start codons with varying efficiencies (O'Donnell and Janssen 2001; Panicker et al. 2015), however 99.84% of CDSs within genomes in this study use a NTG start codon (supplementary table S1, Supplementary Material online), with ATG, GTG, and TTG the most highly represented (80.97%, 13.02%, and 5.72%, respectively). If this frameshift correction model can help to explain observed fourth site *A* enrichment, we can expect weaker enrichment in genomes that do not use TGA as a stop codon. Furthermore, the distance to the next +1 stop codon may be greater as initial frameshifts are captured immediately.

Testing the Models

The Amino Acid Preference Model Cannot Explain A-Starting Amino Acid Biases in the Second Peptide Position

A-Starting Codons Are Preferred Even If There Are Synonymous Alternatives. The structure of the genetic code provisions us with a natural test. Six-fold degenerates serine, leucine, and arginine are encoded by synonymous codons in two codon blocks, in which the first position nucleotide varies. *A*-starting codons for serine (S_A) and arginine (R_A) account for one third of the total codons available. Thus, if there is an amino acid level selection we expect to see mostly *T*-starting serine (S_T) and *C*-starting arginine (R_C).

Serine is especially informative. Assuming selection is primarily for the amino acid content of serine, we expect to see no difference between enrichment of both coding blocks as both maintain AT content destabilizing the 5′ mRNA domain. Whilst both S_A and S_T are more frequent in the second position than expected given genome amino acid usage ($P < 0.001$, Pearson's cumulative test statistic [χ^2]), the mean deviation within genomes from the expected number of CDSs utilizing serine as the second amino acid is greater for *A*-starting (mean observed—expected = 170.186) than *T*-starting serine (mean observed—expected = 70.774). In an unbiased genome, we would expect, all else being equal, the ratio of S_A:S_T to be 1:2. For all amino acids in the genome, we find the mean S_A:mean S_T ratio equal to 1:1.762 ($N = 646$), however for the second amino acid this ratio is 1:0.821, again indicating a strong *A*-starting second amino acid bias. Using genome serine use as our null, we find a significant increase of *A*-starting serine at the second site ($P < 0.001$, Pearson's cumulative test statistic [χ^2]). Furthermore, *A*-starting serine enrichment ratios (mean ratio = 3.429 $\pm$ 1.839, $N = 646$) are significantly greater ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test) than for *T*-starting serine (mean ratio = 1.535 $\pm$ 0.526, $N = 646$). It is apparent that there is a distinct overrepresentation of *A*-starting serine in the second site, indicating selection specific to the *A*-nucleotide.

A comparable analysis for *A*/*C*-starting arginine amino acids is slightly less discriminatory as *C*-starting arginine does not maintain the AT-content. Given genome amino acid usage, we find *A*-starting arginine overrepresented in the second position ($P < 0.001$, Pearson's cumulative test statistic [χ^2]; mean observed – expected = 49.107) with *C*-starting arginine underrepresented ($P < 0.001$, Pearson's cumulative test statistic [χ^2]; mean observed – expected = −26.319). A ratio of 1:4.390 for genome mean R_A:mean R_C ($N = 646$) use demonstrates greater dependence on *C*-starting arginine within CDSs, however a second amino acid ratio of 1:1.565 highlights the greater dependence on *A*-starting arginine at the second site. With genome arginine use as the null, we find a significant increase of *A*-starting arginine at the second site ($P < 0.001$, Pearson's cumulative test statistic [χ^2]). *A*-starting arginine (mean ratio = 3.492 $\pm$ 2.338, $N = 646$) enrichment ratios are significantly greater ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test) than for *C*-starting arginine (mean ratio = 0.892 $\pm$ 0.384, $N = 646$).

Evidently, *A*-starting synonyms of both serine and arginine are favored at the second position indicating selection is stronger for the *A* nucleotide in the first codon position and that selection is not at, or strongest at, the protein level.

No Individual Amino Acids Are Uniquely Preferred in the Second Peptide Position. We also consider whether enrichment reflects selection for specific *A*-starting amino acids in the second position, which could be expected were we witnessing selection at the peptide level. Conversely, if selection were at the nucleotide level we expect multiple amino acids with *A*-starting codons to be over-represented so long as they facilitate posttranslational modifications.

To determine second position amino acid preferences, we calculated average of difference (AOD) scores (see Tang et al. 2010). AOD scores distinguish whether there is a preference and enrichment of particular amino acids in the second position when compared with the whole transcriptome. In a similar manner to Tang et al. (2010), genomes were categorized into three equal groupings of low GC content (GC $\leq$ 44.19%), medium GC content (44.19% $<$ GC $\leq$ 60.91%) and high GC content (60.91% $<$ GC) to limit genomic GC effects. Each amino acid encoded for by *A*-starting codons is preferred at the second position regardless of genome GC content, except for methionine and isoleucine (fig. 5). Avoidance of methionine–methionine cannot be attributed to general avoidance of methionine pairs as they are found more frequently than expected given genome methionine usage ($P < 0.001$, Pearson's cumulative test statistic [χ^2]). However, as methionine in the second position doesn't facilitate NME, the avoidance may be related to the cleaving mechanism. Conversely, genome methionine-isoleucine pairs are less frequent than expected ($P < 0.001$, Pearson's cumulative test statistic [χ^2]) and therefore a general avoidance of methionine-isoleucine pairs may provide some explanation for second site avoidance.

Bonissone et al. (2013) propose that the primary role of NME is to expose serine and alanine rather than other NME substrates, possibly explaining why *T*-starting serine is the only non *A*-starting amino acid universally preferred across GC groupings. Regarding posttranslational modifications this makes sense—for CDSs with non-*A* starting second amino acids we still expect to see an amino acid capable of participating in NME. As we previously describe, both serine blocks are preferred, although *A*-starting serine amino acids are favored. The ability to facilitate NME may explain weak proline and alanine preferences and the preference for threonine and serine(*T*) in GC-rich genomes where *A*-starting codon usage is limited.

If selection is primarily for amino acid functionality, non-*A* starting amino acids involved in modifications should be preferred. This is not the case. Primary N-end rule pathway residues (leucine, phenylalanine, tyrosine, and tryptophan) recognized directly by the bacterial N-recognin ClpS (Dougan et al. 2012) are avoided. For secondary residues (methionine, lysine, and arginine) signaling for the attachment of a primary residue by leucyl/phenylalanyl-tRNA-protein transferase (LFTR) (Dougan et al. 2012), methionine is avoided with only *A*-starting amino acids preferred (avoidance of *C*-starting arginine). Conversely, if selection is at the protein level *A*-starting amino acids not involved in cleavage should be avoided. This is also not seen; *A*-starting asparagine is preferred but does not feature in either posttranslational modification pathway. More generally, the use of *A*-starting amino acids not involved in either pathway (lysine, asparagine, arginine) further suggests selection is operating on underlying nucleotide content.

5′ RNA Structure Requirements Cannot Fully Account for Fourth Site A Enrichment

The amino acid analysis suggests that selection is not for amino acids themselves but for *A*-starting codons (provided

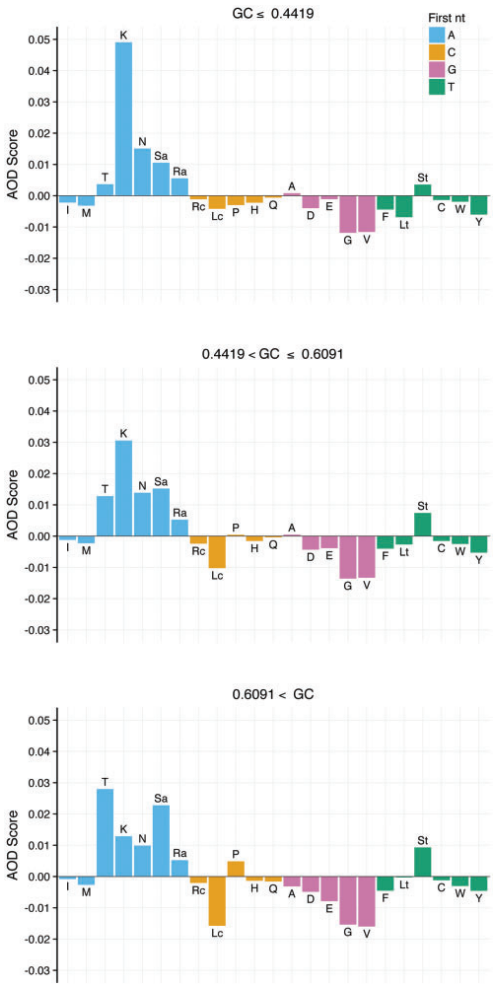


FIG. 5. Average of difference (AOD) scores for each amino acid, demonstrating enrichment or avoidance of each amino acid in the second peptide position when compared with amino acid use within the transcriptome. Genomes are grouped by GC content into three equal sizes grouping in order to minimize GC biases on amino acid choice (lysine for example, encoded by AAA and AAG, is expected to be used more frequently in GC-poor genomes). Amino acids encoded by two coding blocks are defined using the first nucleotide in the codon, for example, *A*-starting serine is denoted Sa. A preference for *A*-starting amino acids except methionine and isoleucine, regardless of genome GC content, is observed.

protein function is not overly compromised). If the selective constraint is to reduce 5′ mRNA stability, we also expect a degree of *T* enrichment within this domain. This prediction comes with the caveat that *G:U* noncanonical pairing is possible and could act to increase RNA stability (Varani and McClain 2000). T_4 ratios are significantly reduced compared with A_4 ratios for each genome ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test. Indeed the mean T_4 ratio is 0.796 ± 0.156 ($N = 646$) whereas the mean A_4 ratio is 1.873 ± 0.375 ($N = 646$), indicating that the effect is relatively *A* specific.

If selection is acting to increase *A* content, we expect little difference between *A* enrichment of the second codon and contiguous codons at both synonymous and nonsynonymous sites. GC variability at synonymous sites is more extreme than at other positions (Muto and Osawa 1987), allowing the possibility of regulation of local GC content independently of amino acid requirements (Babbitt et al. 2014). We therefore predict that if there is selection for *A*-rich codons in the 5′ domain, GC content at synonymous sites should be more independent of genome GC content than codons downstream. Results indicate this is the case (supplementary fig. S3, Supplementary Material online).

This resilience to GC pressure in the 5′ mRNA domain is suggestive of alternative selection pressures acting to determine synonymous site composition. If selection is being driven by RNA stability requirements, we might expect to observe selection on *A* content at all synonymous sites immediately 3′ of the start codon, but with little difference to synonymous sites of immediate codon neighbors. The mean A_6 ratio (1.954 ± 0.802, $N = 646$) confirms *A*-enrichment. Comparisons between A_6 ratios with A_9 and A_{12} ratios (in codons 4 and 5) show weakly significant *A* content variation at these synonymous sites ($P = 0.041$, KruskalWallis rank-sum test), however pairwise comparisons between *A* ratios indicate the second codon is not significantly different in terms of synonymous *A* enrichment (A_6–A_9: $P = 0.973$, A_6–A_{12}: $P = 0.057$, A_9–A_{12}: $P = 0.096$, pairwise Tukey–Kramer tests). Extending the analysis to the fifth codon, we find synonymous site *A* enrichment significantly decreases ($P < 0.01$, Kruskal–Wallis rank-sum test; A_6–A_{15}: $P = 1.2 \times 10^{-8}$, A_9–A_{15}: $P = 4.2 \times 10^{-8}$, A_{12}–A_{15}: $P = 0.001$, pairwise Tukey–Kramer tests), consistent with stronger selection toward 5′ ends. Enrichment is therefore considered comparable for codons two, three, and four.

But is there a unique enrichment specific to the fourth site? If selection on the fourth site is solely for RNA stability, we expect similar *A*-ratios between the nonsynonymous sites of these neighboring codons, as with synonymous sites. In contrast, we find that A_4 is elevated (fig. 6). There are significant differences between the *A*-ratios at the nonsynonymous sites (sites 4, 7, and 10) ($P < 2.2 \times 10^{-16}$, log-transformed *A*-ratios, Kruskal–Wallis rank-sum test), with pairwise comparisons suggesting enrichment at each site is significantly different ($P < 2.2 \times 10^{-16}$, pairwise Tukey–Kramer tests). We find the mean A_4 enrichment (1.873 ± 0.375, $N = 646$) greater than A_7 (1.488 ± 0.129, $N = 646$) and A_{10} (1.344 ± 0.105, $N = 646$).

These results highlight that despite AT requirements in the initial 5′ mRNA domain, the fourth site exhibits significant enrichment not observed at other nonsynonymous sites, a trend not seen for synonymous sites. We therefore cannot attribute the increased fourth site *A* content solely to RNA stability selection.

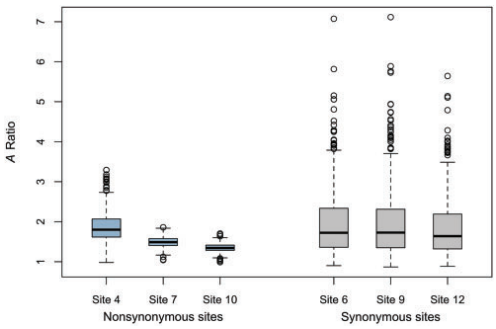


Fig. 6. Comparisons between *A* enrichment ratios for synonymous and nonsynonymous sites in codons 2–4. Enrichment ratios compare the use of *A* at each site with at comparable positions for all codons in the transcriptome (i.e., site 4 is compared with the first positions of all codons, site 5 is compared with the second positions of all codons and site 6 is compared with the third positions of all codons). Unlike synonymous sites in neighboring codons that display similar *A* enrichment ratio distributions, we observer greater variation in *A* enrichment ratios for the fourth site in comparison with the more tightly controlled ratios for sites 7 and 10. Enrichment ratios at the fourth site are significantly increased when compared with sites 7 and 10.

The Frameshift Correction Model Is a Parsimonious Explanation

The Frameshift Correction Model Predicts Weaker Enrichment at the Fourth Site in Genomes Not Using the TGA Stop Codon. The use of a NTG start codon dictates that under the frameshift model, the stop codon must be the TGA stop codon. If the frameshift model can best explain the enrichment observed, we would expect enrichment at synonymous sites in genomes not using TGA to only occur at levels similar to those in codons 3 and 4 due to 5′ RNA stability constraints.

Five of the 651 genomes within this study (*S. mirum*, *M. gallisepticum*, *M. florum*, *U. parvum*, and the Synthetic construct designed and chemically synthesized from *M. genitalium*; Gibson et al. 2008) use this alternative genetic code (NCBI translation table 4). A_4 ratios demonstrate an enrichment of *A* (1.277, 1.443, 1.548, 1.362, and 1.099, respectively), but, importantly, are significantly lower than the A_4 ratios for genomes using the standard genetic code ($P < 0.01$, Wilcoxon rank-sum test). After removing the Synthetic construct from the analysis, the difference remains significant ($P = 0.004$, Wilcoxon rank-sum test). Furthermore, the A_4, A_7, and A_{10} ratios for these genomes exhibit no significant difference between them ($P = 0.368$, Kruskal–Wallis rank-sum test). A_4 ratios are also not significantly different to the A_7 ratios of the third codon ($P = 0.053$, Welch two sample *T*-test) or A_{10} ratios for the fourth codon ($P = 0.835$, Welch two sample *T*-test) in genomes using the standard genetic code.

Might the lower A_4 ratio of genomes not using TGA reflect their high AT content more generally? In order to control for GC content, we performed a loess regression between total genomic GC content and A_4 enrichment ratios and compared the residuals for the two different translation tables. In this case, we find no significant difference between the enrichment ratios ($P = 0.234$, Kruskal–Wallis rank sum test). We note however, that the mean residual for the translation table 4 genomes (−0.103) is lower than for the genomes using the standard genetic code (−0.001) although not significant. This is however limited by the small sample size for table 4 genomes (5 genomes). If we include all table 4 genomes from the original data set ($N = 94$), although we introduce some phylogenetic nonindependence, we find the difference highly significant ($P < 0.001$, Kruskal–Wallis rank sum test) (supplementary fig. S4, Supplementary Material online). The mean residual for table 4 genomes is again negative and lower (−0.070) than for those using the standard genetic code (0.006). Supplementary figure S4*A*, Supplementary Material online, suggests that table 4 genomes may fall into two categories: those that have greatly reduced enrichment and those that are similar to genomes using the standard genetic code. This may result from phylogenetic nonindependence introduced when increasing the data set with the majority of genomes being *Mycoplasmas* (75/94; 79.79%). However Supplementary figure S4*C*, Supplementary Material online, suggests *Mycoplasma* residuals are varied. As these genomes are *AT*-rich, is it highly likely these genomes would utilize *A*-starting second codons regardless of fourth site selection, therefore the fact that there is reduced use in 57/94 (60.64%) genomes is suggestive of a difference in these table 4 genomes. Thus, these observations accord with a model in which the absence of TGA as a stop codon relaxes selection for especially high A_4 content. The remaining *A* excess seen can be accounted for in terms of selection for decreased 5′ mRNA stability (as also observed for A_7 and A_{10}). Assuming high AT content reflects weaker selection against a GC to AT mutation bias, the above results also suggest that the lower A_4 ratios in table 4 genomes cannot be owing to weakened purifying selection (assuming AT content is a proxy for N_e).

The Distance to the Next 1 Frameshift Stop Codon Is Greater for Genes with Fourth Site A. The excess of *A* at site four is consistent with preventing the ribosome initiating on the wrong reading frame. If the ribosome begins translation on an incorrect reading frame and is abruptly terminated, there is less demand for another local +1 stop codon (assuming selection for ambush codons). We therefore expect that the distance to the next +1 stop codon in genes with fourth site *A* is greater than those without. As the three standard stop codons are AT-rich (TAA, TAG, and TGA), we find a strong positive correlation between GC content and the mean nucleotide distance to the next +1 stop codon ($\rho = 0.966$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation) (supplementary fig. S5, Supplementary Material online). We therefore make within-genome comparisons as we can expect GC content to equally influence distances in CDSs with and without fourth site *A*.

The mean distance to a +1 stop codon is significantly greater in genes with fourth site A ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test) but not for genomes not using the

standard genetic code ($P = 0.461$, paired T-test). The presence of an immediate frameshift correction mechanism therefore appears to influence location of further downstream out of frame stop codons. The mean of mean genome distances shifts from 68.583 ± 37.091 ($N = 646$) nucleotides for genes without fourth site A to 72.533 ± 42.376 ($N = 646$) nucleotides in the presence of fourth site A with distances varying greatly between genomes. We observe increased distances to the second +1 stop codon from a mean of 141.334 ± 73.537 ($N = 646$) nucleotides without fourth site A to 144.718 ± 78.656 ($N = 646$) nucleotides ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test) and the third +1 stop codon from 212.226 ± 105.601 ($N = 646$) nucleotides without fourth site A to 215.238 ± 110.525 ($N = 646$) nucleotides ($P = 1.571 \times 10^{-13}$, paired Wilcoxon rank-sum test). In effect, the incorporation of an immediate +1 stop codon appears to subtly shift the sequence of frameshift capture codons downstream. Although these distances are highly variable (the effects of GC content varying between genomes), by comparing samples from within each genome we limit the effects of this variability. The preservation of A_4 under increased GC pressure (fig. 2) is consistent with stronger selection in GC rich genomes for A_4 preservation given the greater distance to the next +1 stop is likely to incur a greater cost.

Discussion

A_4 Content as Another Residue for Error Correction?

We have identified a series of variables that go some way to explaining the enrichment of A at fourth sites. For CDSs with upstream SD sequences, we find reduced fourth site A use (supplementary result 1, Supplementary Material online), consistent with the notion that SD sequences reduce the error rate at translation initiation when compared with genes lacking ribosome recruitment and initiation signals (Di Giacco et al. 2008). The presence of leader genes synthesizing nonfunctional peptides also go some way to explaining why sequences may lack fourth site A (supplementary result 2, Supplementary Material online). A multivariate model using genome A-starting codon use, 5′ A enrichment, leader gene use and the translation table explains over 50% of the variation in genome fourth site A use (supplementary result 3, Supplementary Material online). Given the validity of the frameshift model, we note that such a model might go some way to explain why start codons are in fact of the form NTG. We speculate that in early evolution there may have been coevolution of stop codon usage (we assume TGA to be ancestral) and choice of NTG codons as initiators prior to further dual-coding signals evolving in order to provide more stringent initiation pathways. If so, this provides, to the best of our knowledge, the first explanation as to why start codons are typically NTG and methionine.

The validity of the frameshift model is especially noteworthy given many dual coding signals relate to the control of errors (reviewed in Drummond and Wilke [2009] and Warnecke and Hurst [2011]). For example, splice control by ESEs may be considered as a control of missplicing errors (Dewey et al. 2006; Caceres and Hurst 2013; Wu and Hurst 2015) as ESEs are most abundant near longer introns where splicing error is most common. Selection to avoid amino acid misincorporation (Archetti 2004, 2006; Drummond et al. 2005; Stoletzki and Eyre-Walker 2007; Gilchrist et al. 2009) or codons in close mutational proximity to stop codons where nonsense mediated decay (NMD) cannot detect transcriptional errors (Cusack et al. 2011) may constrain codon choice. The presence of stop codons within introns appears to be NMD-mediated mechanism to catch splice errors (He et al. 1993; Jaillon et al. 2008; Farlow et al. 2010; Mekouar et al. 2010). This suggests a general theme coupling dual coding with error mitigation.

Is A_4 Enrichment Involved in Translation Initiation?

The notion that CDSs might incorporate +1 stop codons favored by selection is not new. Indeed, it has been proposed that the genetic code evolved such that it has the ability to encode frameshift traps (Itzkovitz and Alon 2007). The ambush hypothesis (Seligmann and Pollock 2004) proposes that there is an excess of out of frame stops and that coding sequences frequently use and are under selection for codons that have the potential to form out of stop codons (Seligmann and Pollock 2004; Singh and Pardasani 2009; Tse et al. 2010). However, the biases toward codons contributing to out of frame stops seems largely predictable from the underlying GC pressure (Morgens et al. 2013) with the ambush hypothesis not strictly observed at the gene level (Bertrand et al. 2015). The observation that the usage of A at the fourth site is significantly increased in genomes employing the TGA stop is perhaps the first evidence that selection does favor, at least at one specific site, out of frame stop codons.

Why might the fourth site be unusual and warrant a frameshift trap? We suggest that this might relate to the process of translation initiation itself. The results are consistent with a frameshift correction model, however the dynamics in which the ribosome may find itself incorrectly position on the reading frame, and the context in which an out of frame stop codon can regulate these errors, is somewhat less clear. We consider three models to this effect. First, the +1 stop codon may abort translation immediately if the ribosome slips following initiation, preventing the synthesis of a faulty protein and allowing ultrarapid recycling of ribosomes which are often rate-limiting (Shah et al. 2013; Subramaniam et al. 2014) (translation termination). Alternatively, the stop codon might provide a regulatory signal to increase the fidelity of the ribosome locating the correct initiation site (frameshift "stop and adjust"). It is reasonable to suppose that a slightly misaligned ribosome could read TGA as stop, blocking translation, realigning the ribosome on the correct start site whilst still in the presence of initiation factors. Finally, the +1 TGA may prevent read-through following the translation of an upstream gene (read-through termination), although there may well be many alternative sites for an out of frame stop to determine the fate of frameshifted translation.

We find evidence against the last of these models (see supplementary result 4, Supplementary Material online).

Regarding the "stop and adjust" model, this may be configured more generally in a context of start site recognition mechanisms. This model would concur with our observations that fourth site *A* content is associated with an absence of SD sequences or leader genes, both of which are implicated in start codon recognition. Yamamoto et al. (2016) propose that bacterial 70 S ribosomes have the ability to scan the mRNA and the presence of a SD sequence provides an important signal for selection of the correct start codon by allowing the fMet-tRNA to fix the ribosome at the canonical start codon. In its absence, the ribosome is not fixed and can continue to scan the mRNA. Our results in supplementary results 1 and 2, Supplementary Material online, are consistent with fourth site regulation of initiation by assistance in identifying and positioning the ribosome correctly at the start codon when lacking SD sequences and are suggestive of a direct involvement of the fourth site in the dynamics of translation initiation and start codon selection.

The identity of the start codon has also been shown to determine translation efficiency (O'Donnell and Janssen 2001; Osterman et al. 2013; Panicker et al. 2015; Hecht et al. 2017). We proposed two hypotheses that may implicate the start codon with fourth site *A* usage, either contributing to mRNA-ribosome stability for the more efficient start codons, or preventing the ribosomes from dissociating from weaker start codons. We find the *A* enrichment at the fourth site strongest for GTG, followed by ATG and TGT (supplementary result 5, Supplementary Material online) suggesting that the weakest binding initiator has weakest enrichment. Both Panicker et al. (2015) and Osterman et al. (2013) report GTG is the more efficient initiator. The increased enrichment at the more efficient start codons again implicates the fourth site in increasing initiation efficiency, although the evidence is not definitive. Interestingly, stop codons in 5′ leading regions allow termination of translation events that initiate before the ribosome reaches the correct start codon, increasing protein synthesis efficiency (Seligmann 2007). It is possible that the fourth site acts as a final checkpoint against these events, allowing recalibration or reinitiation of the ribosome at the correct initiation site. Such events may occur as increases in the number of alternative start codons in the 5′ region has a measurable increase on protein activity (Seligmann 2007). The evolution of 5′ stop codons to complement the use of these upstream start codons can provide stringent regulation of the ribosome initiation from the correct initiation site, where fourth site *A* can provide site-specific definition of the correct site.

Our results implicate involvement fourth site *A* in translation initiation and are consistent with in ensuring correct start codon selection. Assuming TGA to be an ancestral stop codon, the reduced enrichment for genomes not using the TGA stop suggest this control is functionally related to the presence of the stop codon. Upon losing the TGA stop, selection to maintain this enrichment was reduced and enrichment weakened to levels required for RNA stability.

A_4 Enrichment Observed in Archaea but Not in Eukaryotes Is Suggestive of Interactions Specific to the Prokaryotic Ribosome

One curiosity concerning fourth site usage is that different patterns are observed in nuclear eukaryotic genes. We find that A_4 enrichment ratios are significantly enriched >1 among archaea genomes (73/77, 94.81%), however we find no evidence for fourth site enrichment specific to *A* within eukaryotes (supplementary result 6, Supplementary Material online). As methionine removal is largely the same in the two taxa, a peptide-based argument seems unable to explain our observations. Furthermore, many human and plant genes tend instead to have GC rich terminal ends (Niimura et al. 2003). One notable distinction between the two is the ribosome. If frameshifting or start site recognition mechanisms differ between the 16S rRNA and 18S rRNA then we might expect differences between the taxa, even though TGA is a stop in almost all taxa. Notably the fourth site *A* enrichment observed in archaea, in which initiation resembles that of bacteria and utilizes 16S rRNA, provides a suggestion that the fourth site is a dual coding mechanism functionally linked with the prokaryotic ribosome and initiation mechanics. Given that leaderless mRNAs can be translated between domains (Grill et al. 2000), current leaderless mRNAs may have evolved from ancestral mRNA in which mRNA recognition and initiation the common ancestor occurred via a ribosome-initiation tRNA complex (Moll et al. 2002).

The strength of *A* bias in both bacteria and archaea, but lacking from eukaryotes, suggests the increased initiation complexity in eukaryotes (Asano 2014) may have allowed relaxed selection on ancestral fourth site *A*, given there are stringent alternative mechanisms for locating the correct start codon. The recruitment of ribosomes to eukaryotic mRNA and subsequent start codon identification requires a combination of eukaryotic initiation factors (eIFs) (Jackson et al. 2010; Shatsky et al. 2014) and further binding proteins, when only three initiation factors are found in bacteria (Laursen et al. 2005). Some bacterial leaderless genes do not require the presence of ribosomal proteins S1 or S2 (Moll et al. 2002), which are required for the 30 S ribosome pathway, or even the presence of initiation factors (Udagawa et al. 2004). Interactions between initiation factors forming multifactor complexes (MFC) provide stringent ATG recognition (reviewed in Asano 2014). eIF1A, a universally conserved eukaryotic homolog of bacterial eIF1 has evolved both N- and C-terminal domains stimulating recruitment of methionyl initiator tRNA to ATG but preventing and discriminating against non-ATG initiation (Pestova and Kolupaeva 2002; Fekete et al. 2005; Nanda et al. 2009; Saini et al. 2010). In addition, selection for nucleotides in the Kozak sequence (Kozak 1986, 1997), which acts to increase the efficiency of eukaryotic translation initiation, may be stronger than that on the fourth site *A* that would provide a similar regulation signal. Interestingly, *A* is the second most prevalent nucleotide at site 4 in Kozak sequences for eight eukaryotic organisms (Grzegorski et al. 2014) which may reflect ancestral selection on the fourth site for *A* that has now weakened

due to selection for nucleotides in the Kozak sequence, but still greater than for other nucleotides. The fidelity afforded to eukaryotic start codon recognition through the combination of initiation factors and initiation signals may explain the differences in enrichment between the domains at the fourth site.

Unresolved Issues

Although *A* enrichment is significantly greater at the fourth site compared with seventh and tenth sites of neighboring codons, both synonymous and nonsynonymous sites in the 5′ domain demonstrate an *A* enrichment. What is unclear about any RNA stability model is why *A*, and not *T*, is preferred. Localized *T* enrichment should provide a similar destabilizing effect as that of *A*, but *T* is consistently underrepresented in comparison with *A* in the first three codons. One possibility is the preference for *A* over *T* might reflect avoidance of *G*:*U* noncanonical base pairs that allow weak base pairing (Varani and McClain 2000) and could introduce unwanted mRNA stability. Results from archaea (supplementary result 6, Supplementary Material online) suggest that selection for *A*/*T* content in the 5′ domain reducing RNA stability is not limited to bacteria, but is infrequent in eukaryotes. Why are eukaryotes different in 5′ stability requirements?

Eukaryote analyses also raise further unresolved issues. Although *A* enrichment cannot be accounted for solely in terms of selection on the peptide in bacteria, the preference for particular non-*A* starting amino acids (alanine, proline, and *T*-starting serine) that facilitate methionine cleavage, and the avoidance of A starting methionine and isoleucine that do not, indicate a selection pressure for amino acids promoting cleavage. However, preferences for *A*-starting amino acids that promote cleavage (threonine, *A*-starting serine) are heightened. With evidence for methionine aminopeptidase activity and second amino acid specificity in eukaryotes (Giglione et al. 2000; Chen et al. 2002; Xiao et al. 2010), if selection was primarily for facilitative amino acids we should also observe an *A* enrichment in eukaryotes, yet this is not apparent. We do not know why this is.

The regulation of translation involves interactions with RNA binding proteins (RBPs) that influence ribosome binding and translation initiation (Babitzke et al. 2009; Van Assche et al. 2015). These interactions directly modulate ribosome binding, alter the mRNA secondary structures or act as a chaperone for the interactions of other RNA effectors. The most likely hypotheses implicating the fourth site in ribosome blocking interactions is one in which the fourth site acts as part of a binding site to which the RBPs bind, blocking initiation, or one in in which the fourth site is enriched to avoid these interactions. For example, the global regulator CsrA binds optimally to the sequence 5′-RUACARGGAUGU-3′ (Dubey et al. 2005; Schubert et al. 2007). The *B. subtilis trp* RNA binding attenuation protein (TRAP) binds with the *ycbK* putative efflux protein at NAG motifs across the initiation region, one of which may be GAG from sites 3 to 6, directly blocking the 30 S ribosome binding (Yakhnin et al. 2006). In a similar manner, the bacteriophage T4 regA binds near the start codons and interactions with the fourth site when binding to the to the consensus sequence 5′-AAAAUUGUUAUGUAA-3′ (Winter et al. 1987; Brown et al. 1997). Enrichment of fourth site *A* may reflect selection for avoidance of this interaction. For CsrA, the fourth site is the outermost nucleotide in the consensus sequence and we expect binding of this site to be less important and under weaker selection than binding with the 5′ UTR (Dubey et al. 2003; Edwards et al. 2011) and GGA core motif (Schubert et al. 2007). Binding of both TRAP and regA are likely to be organism specific. Whilst we cannot definitively discount selection against interactions with RBPs, it is unlikely to explain the near-universal enrichment we observe and are not investigated further within the scope of this work.

Future Prospects: Experimental Tests

Our observations provide an avenue for experimental testing. Adopting approaches similar to Napolitano et al. (2016) who mutated *A*-starting arginine codons to the CGT synonym would be especially valuable. Their preliminary data supports the exceptionalism of the fourth site. Notably 12 of 13 recalcitrant mutations, including 1 of 2 at the second codon, were in mRNA terminal domains highlighting the importance not only of the terminal domains, but the second codon in particular. Further targeted efforts to resolve the mechanistic basis for this would be valuable. A comparative analysis in both genomes that do and don't employ TGA as a stop would be especially valuable.

Materials and Methods

General

R version 3.2.3 (R Core Team 2015) was used for data plotting and statistical analyses. All further scripting was conducted using custom scripts in Python 2.7.10 and Python 3.6.1 (https://www.python.org/) with the Biopython 1.66 package (Cock et al. 2009) and Tcl (http://www.tcl.tk/). Scripts can be found at https://github.com/la466/fourth_site.git. For statistical analyses, N denotes the number of genomes used and means are given with one standard deviation.

Genome Downloads

Genome sequences of 3,731 bacterial genomes were downloaded from the European Molecular Biology Laboratory (EBML) database (http://www.ebi.ac.uk/Tools/dbfetch/emblfetch? db=embl, last accessed 12th January 2016). Genomes were filtered to include one genome per genus to control for phylogenetic nonindependence (additional genomes of that genus were discounted) larger than 500,000 base pairs leaving 651 genomes. Of these, 646 used translation table 11 and 5 translation table 4. CDS from 205 archaea genomes were downloaded from EMBL (accessed 27th October 2016) and subject to filtering leaving sequences from 77 genomes. Eukaryotic CDSs were downloaded from the Ensembl database (Yates et al. 2016) (ftp://ftp.ensembl.org/pub/release-86/fasta/, last accessed 31st October 2016). The analysis was based on CDSs from the following assemblies (Ensembl release 86 unless stated): *H. sapiens* (GRCh38.p7), *S. cerevisiae*

(R64-1-1), *D. melanogaster* (BDGP6), *M. musculus* (GRCm38.p4), *M. mulatta* (Mmul_8.0.1), *O. cuniculus* (OryCun2.0), *B. taurus* (UMD3.1), *G. gallus* (Gallus_gallus-5.0), *C. elegans* (WBcel235), and *A. thaliana* (TAIR10, release 33). 186 protist genomes were downloaded from the Ensembl database (Kersey et al. 2016) (ftp://ftp.ensemblgenomes.org/pub/protists/release-36, last accessed 22nd June 2017).

CDS Filtering

Every CDS within a genome was filtered, limiting the analysis to genes with a multiple of three nucleotides, containing only canonical *A*, *C*, *T*, or *G* nucleotides, without internal stop codons and those with a stop codon defined by the relevant translation table, either translation table 11 (TAA, TAG, and TGA) or translation table 4 (TAA and TAG) where TGA instead encodes tryptophan. For CDSs passing these filtering criteria, start codon frequencies were calculated. As the frameshift model assumes a NTG start codon, for subsequent analyses only CDSs starting with a NTG start codon (N = any nucleotide) were considered. In practice, non-NTG start codons are too rare for meaningful analysis. For eukaryotes, only ATG starts were allowed.

Calculation of Enrichment Ratios

To account for nucleotide bias within the genome, *A* enrichment ratios were calculated for each genome using

$$A_n = \frac{f_A(n)}{F_A(x)}, \quad (1)$$

where A_n = *A* ratio at position n, $f_A(n)$ = proportion of CDSs with *A* at site n and $F_A(x)$ = proportion of total codons with *A* in position x, where x corresponds to the intracodon position of n (i.e., if $n = 4$, $x = 1$, so we are considering all first codon sites in all CDSs in a genome). n can take any value from 1 to the length of the longest gene, although we consider events exclusively at 5′ ends. The same protocol was followed to calculate other nucleotide enrichment ratios and amino acid enrichment ratios.

Nucleotide Conservation

The variation in nucleotide content in each codon provides a representation of possible exceptionalism and conservation of particular positions. Methods for exploring GC content variation were as in Tang et al. (2010). For each codon position in codons 2–30, the proportion of each nucleotide usage was calculated across all CDSs in each genome. For each genome, the GC proportion for each position was then calculated across all CDSs. Finally, the variance in GC content at each position between genomes provided an overall GC variance.

Nucleotide Variability between Related Species

A local BLAST database was generated from filtered *E. coli* O157 CDSs using BLAST v2.4.0 (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). CDSs from *S. flexneri* were queried against the local database. If there was more than one match, the ortholog with the lowest expected value (*E*) and percentage match was chosen.

For each orthologous CDS pair, the nucleotide in the first position in each of the first 11 codons of the *E. coli* sequence was noted and losses from this nucleotide in *S. flexneri* orthologs counted. The proportion of sites changing from each nucleotide in each codon was calculated from the total counts. Comparisons between these species do not assume any evolutionary relationship but simply compares ortholog differences. These variations are conservative as orthologs with the most conserved sequences are chosen. We employ *E. coli* as the focal species and *S. flexneri* as the indicator of the effects of weakened purifying selection, as the strength of selection due to effective population size is considered to be smaller (Hershberg et al. 2007). Thus, we can ask whether a fourth site *A* in *E. coli* is more resilient to change. If so, this would indicate stronger purifying selection on the fourth site.

Codon Adaptation Index Analysis

Bacterial codon use is often highly nonrandom. Translational selection biases codons toward those rapidly translated tRNAs and with high availability (Ketteler 2012). Highly expressed genes, for which translational errors may prove more costly, typically use a restricted set of preferred codons corresponding to the tRNA repertoire (Rocha 2004) with codon bias strongest in these genes (Higgs and Ran 2008). The Codon Adaptation Index (CAI) (Sharp and Li 1987) is one method of quantifying codon bias. High expression correlates with a high CAI value in several organisms including *E. coli* (dos Reis et al. 2003), and therefore the CAI value is used as a proxy measure for gene expression.

For each genome, a reference set of CDSs for which codon usage was expected to be high was selected to represent the highly expressed genes to include 20 ribosomal genes from *rplA/1* to *rplF/6*, *rplI/9* to *rplU/21*, and *rpsB/2* to *rpsU/21*. Only genomes with annotations for 20 of these genes were considered. CAI indices for each gene in this reference set were calculated using CodonW v1.4.4 (https://sourceforge.net/projects/codonw/) using the "-coa_cu -coa_num 100%" parameters to include all reference CDSs. CAI values for the remaining genes within the genome were calculated using the "-all_indices" parameter, including the fop_file, cai_file, and cbi_file. For *E. coli* O157, CAI values were also calculated using the default indices provided by CodonW and correlated with those calculated from our reference set ($\rho = 0.987$, $P < 0.01$, Spearman's rank correlation) to ensure the reference set accurately represented the highly expressed genes.

Identification of Shine–Dalgarno Sequences

Potential Shine–Dalgarno (SD) sequences were identified using methods described in Starmer et al. (2006). For each genome, the 16 S rRNA genes were located and the 3′ tail isolated from the gene sequence. Tails were scanned for the 5′-GAT-3′ motif located closest to the 3′ end of the rRNA tail. If multiple tails were present, the most frequent was selected. Only tails between 8 and 15 nucleotides were considered.

For each CDS within the genome, the change in free energy ΔG° was calculated using the free_scan script from the

free2bind v1.0.1 package (https://sourceforge.net/projects/free2bind/) (Starmer et al. 2006). ΔG° describes the change in free energy required to bring the mRNA strand together with the identified 16S rRNA tail; ΔG° scores less than zero describe a likely interaction. For each CDSs, a 60-nucleotide window centered on the start codon, with *A* of the ATG representing nucleotide 30, was extracted and ΔG° was calculated by aligning the 16S rRNA tail at each position in this window. The position with minimal ΔG° was considered the optimal binding site.

A CDS was considered to have a SD sequence providing the optimal binding site had $\Delta G^\circ \leq -3.4535$ kcal/mol, derived from the average of free_scan calculations for core motifs 5′-GGAG-3′ (−3.60793 kcal/mol), 5′-GAGG-3′ (−3.60793 kcal/mol) and 5′-AGGA-3′ (−3.144505 kcal/mol) (Ma et al. 2002). Strong binding was defined as $\Delta G^\circ \leq -8.4$ kcal/mol obtained from binding of the sequence 5′-GGAGGT-3′. Relative gene distances were calculated as the distance of the 5′ *A* in the rRNA sequence flanking the core SD motif relative to the first nucleotide in the start codon, defined as 0. Distances less than one indicate a SD sequence upstream of the start codon.

Average of Difference Calculations

Preferences or avoidances of each amino acid in the second position was calculated using the average of difference (AOD) score (Tang et al. 2010). AOD scores calculate the difference between the frequencies of an amino acid in the second position compared with the average frequencies compared with all positions in the CDS, using the formula

$$AOD_x = \frac{\sum_n (f(x) - F_x)}{n}, \quad (2)$$

where AOD_x = average of difference score for amino acid *x*, $f(x)$ = frequency of amino acid *x* in the second peptide position, F_x = average frequency of amino acid *x* across all amino acids and n = number of CDSs. Genomes were further categorized equally into low ($GC \leq 44.19\%$), medium ($44.19\% < GC \leq 60.91\%$) and high ($GC < 60.91\%$) GC to account for underlying biases.

Distances to Out-of-Frame Stop Codons

For each CDS, removing the first nucleotide from the sequence provided the +1 frameshift sequence. For each codon from codon 2 within the shifted sequence was queried for a suitable stop codon. The position of the first nucleotide of the stop codon in the sequence was defined as the distance to the next stop codon. The same protocol was applied for second and third stop codons.

Identification of Leader Genes

Leader genes were identified as open reading frames (ORFs) 5′ to the structural CDS using similar methods to Lyubetsky et al. (2014) and Korolev et al. (2016). A CDS was considered providing it was longer than 200 nucleotides, shorter than 10,000 nucleotides and had met previous filtering criteria. For each qualifying CDS, the upstream intergenic region was extracted if >100 nucleotides and <1,400 nucleotides.

Within the intergenic region, all potential ORFs were identified providing they had a regular start codon, were a multiple of three nucleotides, without internal stop codons, had a stop codon defined by the relevant translation table and were longer than six codons. If more than one ORF was identified, the longest ORF was chosen. The algorithm was trained on the *E. coli* O157 genome to identify leader genes as found by Korolev et al. (2016) and subsequently applied to all genomes.

Multivariate Analysis

A multivariate analysis was conducted using 134 genomes with all available data points. These included: the proportion of CDSs with fourth site *A*, *A* content at sites 6, 7, 9, 10, and 12, the proportion of CDSs with a leader gene the proportion of *A*-starting codons and the genome translation table. Further analysis was conducted on all genomes ($N = 651$) and at the gene level ($N = 2164911$).

Supplementary Material

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

Acknowledgments

This work was supported by the European Research Council (Advanced grant ERC-2014-ADG 669207 to L.D.H.) and the Medical Research Council (grant number MR/L007215/1 to L.D.H.).

References

Agoglia RM, Fraser HB. 2016. Disentangling sources of selection on exonic transcriptional enhancers. *Mol Biol Evol.* 33(2):585–590.
Al-Shahib A, Breitling R, Gilbert DR. 2007. Predicting protein function by machine learning on amino acid sequences: a critical evaluation. *BMC Genomics* 8(1):1–10.
Archetti M. 2006. Genetic robustness and selection at the protein level for synonymous codons. *J Evol Biol.* 19(2):353–365.
Archetti M. 2004. Selection on codon usage for error minimization at the protein level. *J Mol Evol.* 59(3):400–415.
Asano K. 2014. Why is start codon selection so precise in eukaryotes?. *Translation* 2(1):e28387.
Babbitt GA, Alawad MA, Schulze KV, Hudson AO. 2014. Synonymous codon bias and functional constraint on GC3-related DNA backbone dynamics in the prokaryotic nucleoid. *Nucleic Acids Res.* 42:10915–10926.
Babitzke P, Baker CS, Romeo T. 2009. Regulation of translation initiation by RNA binding proteins. *Annu Rev Microbiol* 63:27–44.
Bachmair A, Finley D, Varshavsky A. 1986. In vivo half-life of a protein is a function of its amino-terminal residue. *Science* 234(4773):179–186.
Balbi KJ, Rocha EP, Feil EJ. 2009. The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. *Mol Biol Evol.* 26(2):345–355.
Behura SK, Severson DW. 2011. Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in *Aedes aegypti* and *Anopheles gambiae*. *Insect Mol Biol.* 20(2):177–187.
Belew AT, Meskauskas A, Musalgaonkar S, Advani VM, Sulima SO, Kasprzak WK, Shapiro BA, Dinman JD. 2014. Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. *Nature* 512(7514):265–269.
Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol* 9:675.

Bertrand RL, Abdel-Hameed M, Sorensen JL. 2015. Limitations of the 'ambush hypothesis' at the single-gene scale: what codon biases are to blame?. *Mol Genet Genomics* 290(2):493–504.
Bivona L, Zou Z, Stutzman N, Sun PD. 2010. Influence of the second amino acid on recombinant protein expression. *Protein Expr Purif.* 74(2):248–256.
Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.* 25(3):106–110.
Bonissone S, Gupta N, Romine M, Bradshaw RA, Pevzner PA. 2013. N-terminal protein processing: a comparative proteogenomic analysis. *Mol Cell Proteomics* 12(1):14–28.
Botzman M, Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol.* 12(10):R109.
Brown D, Brown J, Kang C, Gold L, Allen P. 1997. Single-stranded RNA recognition by the bacteriophage T4 translational repressor, regA. *J Biol Chem.* 272(23):14969–14974.
Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, Rodnina MV, Komar AA. 2016. Synonymous codons direct cotranslational folding toward different protein conformations. *Mol Cell* 61(3):341–351.
Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol.* 14(12):R143.
Caliskan N, Peske F, Rodnina MV. 2015. Changed in translation: mRNA recoding by -1 programmed ribosomal frameshifting. *Trends Biochem Sci.* 40(5):265–274.
Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol.* 62(1):89–98.
Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.* 3(4):285–298.
Chamary J-V, Hurst LD. 2005a. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else?. *Trends Genet.* 21(5):256–259.
Chamary JV, Hurst LD. 2005b. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.
Chen S, Vetro JA, Chang YH. 2002. The specificity in vivo of two distinct methionine aminopeptidases in *Saccharomyces cerevisiae*. *Arch Biochem Biophys.* 398(1):87–93.
Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.
Cohanim AB, Haran TE. 2009. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res.* 37(19):6466–6476.
Cusack BP, Arndt PF, Duret L, Crollius HR, Zhang J, 2011. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet.* 7(10):e1002276.
de Smit MH, van Duin J. 1990. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A.* 87(19):7668–7672.
Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.
Di Giacco V, Márquez V, Qin Y, Pech M, Triana-Alonso FJ, Wilson DN, Nierhaus KH. 2008. Shine–Dalgarno interaction prevents incorporation of noncognate amino acids at the codon following the AUG. *Proc Natl Acad Sci U S A.* 105(31):10715–10720.
Diwan GD, Agashe D. 2016. The frequency of internal Shine–Dalgarno: like motifs in prokaryotes. *Genome Biol Evol.* 8(6):1722–1733.
Doherty A, McInerney JO. 2013. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Mol Biol Evol.* 30(10):2263–2267.
dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* 31(23):6976–6985.
Dougan DA, Micevski D, Truscott KN. 2012. The N-end rule pathway: from recognition by N-recognins, to destruction by AAA + proteases. *Biochim Biophys Acta* 1823(1):83–91.
Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 102(40):14338–14343.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 10(10):715–724.
Dubey AK, Baker CS, Romeo T, Babitzke P. 2005. RNA sequence and secondary structure participate in high-affinity CsrA-RNA interaction. *RNA* 11(10):1579–1587.
Dubey AK, Baker CS, Suzuki K, Jones AD, Pandit P, Romeo T, Babitzke P. 2003. CsrA regulates translation of the *Escherichia coli* carbon starvation gene, cstA, by blocking ribosome access to the cstA transcript. *J Bacteriol.* 185(15):4450–4460.
Edwards AN, Patterson-Fortin LM, Vakulskas CA, Mercante JW, Potrykus K, Vinella D, Camacho MI, Fields JA, Thompson SA, Georgellis D, et al. 2011. Circuitry linking the Csr and stringent response global regulatory systems. *Mol Microbiol.* 80(6):1561–1580.
Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* 21(19):4599–4603.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2(9):E268.
Farlow A, Meduri E, Dolezal M, Hua L, Schlötterer C. 2010. Nonsense-mediated decay enables intron gain in Drosophila. *PLoS Genet.* 6(1):e1000819.
Fekete CA, Applefield DJ, Blakely SA, Shirokikh N, Pestova T, Lorsch JR, Hinnebusch AG. 2005. The eIF1A C-terminal domain promotes initiation complex assembly, scanning and AUG selection in vivo. *EMBO J.* 24(20):3588–3601.
Forman JJ, Legesse-Miller A, Coller HA. 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci U S A.* 105(39):14879–14884.
Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinnel T. 2006. The proteomics of N-terminal methionine cleavage. *Mol Cell Proteomics* 5(12):2336–2349.
Furlong RF, Yang Z. 2008. Diversifying and purifying selection in the peptide binding region of DRB in mammals. *J Mol Evol.* 66(4):384–394.
Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, et al. 2008. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319(5867):1215–1220.
Giglione C, Serero A, Pierre M, Boisson B, Meinnel T. 2000. Identification of eukaryotic peptide deformylases reveals universality of N-terminal protein processing mechanisms. *EMBO J.* 19(21):5916–5929.
Gilchrist MA, Shah P, Zaretzki R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics* 183(4):1493–1505.
Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342(6157):475–479.
Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57(Pt 1):81–91.
Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* 6(9):1197–1211.
Gray VE, Kumar S. 2011. Rampant purifying selection conserves positions with posttranslational modifications in human proteins. *Mol Biol Evol.* 28(5):1565–1568.
Grill S, Gualerzi CO, Londei P, Blasi U. 2000. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *EMBO J.* 19(15):4101–4110.

Grzegorski SJ, Chiari EF, Robbins A, Kish PE, Kahana A, Neuhauss SCF. 2014. Natural variability of kozak sequences correlates with function in a zebrafish model. *PLoS One* 9(9):e108475.
Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol.* 6(2):e1000664.
Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A.* 101(25):9205–9210.
Guo X, Gui Y, Wang Y, Zhu Q-H, Helliwell C, Fan L. 2008. Selection and mutation on microRNA target sequences during rice evolution. *BMC Genomics* 9:454.
Gupta P, Kannan K, Mankin AS, Vázquez-Laslop N. 2013. Regulation of gene expression by macrolide-induced ribosomal frameshifting. *Mol Cell* 52(5):629–642.
He F, Peltz SW, Donahue JL, Rosbash M, Jacobson A. 1993. Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1-mutant. *Proc Natl Acad Sci U S A.* 90(15):7034–7038.
Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. 2017. Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* 45(7):3615–3626.
Hershberg R, Tang H, Petrov DA. 2007. Reduced selection leads to accelerated gene loss in Shigella. *Genome Biol.* 8(8):R164.
Higgs PG, Ran W. 2008. Coevolution of Codon Usage and tRNA Genes Leads to Alternative Stable States of Biased Codon Usage. *Mol Biol Evol.* 25:2279–2291.
Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J Mol Evol.* 63(2):174–182.
Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* 17(4):405–412.
Jacks T, Power MD, Masiarz FR, Luciw PA, Barr PJ, Varmus HE. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* 331(6153):280–283.
Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev.* 11(2):113–127.
Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Ségurens B, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* 451(7176):359–362.
Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, et al. 2016. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* 44(D1):D574–D580.
Ketteler R. 2012. On programmed ribosomal frameshifting: the alternative proteomes. *Front Genet* 3:242.
Korolev SA, Zverkov OA, Seliverstov AV, Lyubetsky VA. 2016. Ribosome reinitiation at leader peptides increases translation of bacterial proteins. *Biol Direct* 11(1):20.
Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44(2):283–292.
Kozak M. 1997. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J.* 16(9):2482–2492.
Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324(5924):255–258.
Laursen BS, Sørensen HP, Mortensen KK, Sperling-Petersen HU. 2005. Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev* 69(1):101–123.
Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, Valladares O, Yang J, Bambina S, Sabin LR, et al. 2012. Global analysis of RNA secondary structure in two metazoans. *Cell Rep.* 1(1):69–82.
Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* 24(11):4346–4359.
Liao Y-D, Jeng J-C, Wang C-F, Wang S-C, Chang S-T. 2004. Removal of N-terminal methionine from recombinant proteins by engineered *E. coli* methionine aminopeptidase. *Protein Sci.* 13(7):1802–1810.
Liu G, Zhang R, Xu J, Wu C-I, Lu X. 2015. Functional conservation of both CDS- and 3′-UTR-located microRNA binding sites between species. *Mol Biol Evol.* 32(3):623–628.
Looman AC, Bodlaender J, Comstock LJ, Eaton D, Jhurani P, de Boer HA, van Knippenberg PH. 1987. Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in *Escherichia coli*. *EMBO J.* 6(8):2489–2492.
Lyubetsky VA, Korolev SA, Seliverstov AV, Zverkov OA, Rubanov LI. 2014. Gene expression regulation of the PF00480 or PF14340 domain proteins suggests their involvement in sulfur metabolism. *Comput Biol Chem.* 49:7–13.
Ma J, Campbell A, Karlin S. 2002. Correlations between Shine–Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol.* 184(20):5733–5745.
Ma LN, Cui P, Zhu J, Zhang ZH, Zhang Z. 2014. Translational selection in human: more pronounced in housekeeping genes. *Biol Direct* 9(1):17.
McFerrin LG, Stone EA. 2011. The non-random clustering of non-synonymous substitutions and its relationship to evolutionary rate. *BMC Genomics* 12(1):1–10.
Mekouar M, Blanc-Lenfle I, Ozanne C, Da Silva C, Cruaud C, Wincker P, Gaillardin C, Neuvéglise C. 2010. Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biol.* 11(6):R65.
Melian EB, Hall-Mendelin S, Du F, Owens N, Bosco-Lauth AM, Nagasaki T, Rudd S, Brault AC, Bowen RA, Hall RA, et al. 2014. Programmed ribosomal frameshift alters expression of west Nile virus genes and facilitates virus replication in birds and mosquitoes. *PLoS Pathog.* 10(11):e1004447.
Meyer IM, Miklós I. 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.* 33(19):6338–6348.
Moll I, Grill S, Gualerzi CO, Bläsi U. 2002. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol Microbiol.* 43(1):239–246.
Morgens DW, Chang CH, Cavalcanti ARO. 2013. Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in Prokaryotic Genomes. *BMC Genomics* 14(1):418.
Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A.* 84(1):166–169.
Nanda JS, Cheung YN, Takacs JE, Martin-Marcos P, Saini AK, Hinnebusch AG, Lorsch JR. 2009. eIF1 controls multiple steps in start codon recognition during eukaryotic translation initiation. *J Mol Biol.* 394(2):268–285.
Napolitano MG, Landon M, Gregg CJ, Lajoie MJ, Govindarajan L, Mosberg JA, Kuznetsov G, Goodman DB, Vargas-Rodriguez O, Isaacs FJ, et al. 2016. Emergent rules for codon choice elucidated by editing rare arginine codons in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 113(38):E5588–E5597.
Ng DTWB, Jeremy D, Walter P. 1996. Signal sequences specify the targeting route to the endoplasmic reticulum membrane. *J Cell Biol.* 134(2):269–278.
Niimura Y, Terabe M, Gojobori T, Miura K. 2003. Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Res.* 31(17):5195–5201.
O'Donnell SM, Janssen GR. 2001. The initiation codon affects ribosome binding and translational efficiency in *Escherichia coli* of cI mRNA with or without the 5′ untranslated leader. *J Bacteriol.* 183:1277–1283.
Osterman IA, Evfratov SA, Sergiev PV, Dontsova OA. 2013. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res.* 41(1):474–486.
Ouidir T, Jarnier F, Cosette P, Jouenne T, Hardouin J. 2015. Characterization of N-terminal protein modifications in *Pseudomonas aeruginosa* PA14. *J Proteomics* 114:214–225.

Panicker IS, 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.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23(2):301–309.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol.* 24(8):1600–1603.
Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of co-translational folding. *Nat Struct Mol Biol.* 20:237–243.
Pestova TV, Kolupaeva VG. 2002. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev.* 16(22):2906–2922.
Prendergast JG, Semple CA. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res.* 21(11):1777–1787.
Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A.* 97(19):10567–10572.
Qing G, Xia B, Inouye M. 2003. Enhancement of translation initiation by A/T-rich sequences downstream of the initiation codon in *Escherichia coli*. *J Mol Microbiol Biotechnol.* 6(3-4):133–144.
Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* 14(11):2279–2286.
Saini AK, Nanda JS, Lorsch JR, Hinnebusch AG. 2010. Regulatory elements in eIF1A control the fidelity of start codon selection by modulating tRNA(i)(Met) binding to the ribosome. *Genes Dev.* 24(1):97–110.
Sankararaman S, Kolaczkowski B, Sjölander K. 2009. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Res.* 37(Web Server):W390–W395.
Savisaar R, Hurst LD. 2017. Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. *Mol Biol Evol. 34(5):1110–1126*.
Scharff LB, Childs L, Walther D, Bock R, Casadesús J, 2011. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet.* 7(6):e1002155.
Schubert M, Lapouge K, Duss O, Oberstrass FC, Jelesarov I, Haas D, Allain FHT. 2007. Molecular basis of messenger RNA recognition by the specific bacterial repressing clamp RsmA/CsrA. *Nat Struct Mol Biol.* 14(9):807–813.
Seligmann H. 2007. Cost minimization of ribosomal frameshifts. *J Theor Biol.* 249(1):162–167.
Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23(10):701.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* 34(8):2428–2437.
Shachrai I, Zaslaver A, Alon U, Dekel E. 2010. Cost of unneeded proteins in *E. coli* is reduced after several generations in exponential growth. *Mol Cell* 38(5):758–767.
Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. *Cell* 153(7):1589–1601.
Sharp PM, Li WH. 1987. The codon Adaptation Index: a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3):1281.
Shatsky IN, Dmitriev SE, Andreev DE, Terenin IM. 2014. Transcriptome-wide studies uncover the diversity of modes of mRNA recruitment to eukaryotic ribosomes. *Crit Rev BiochemMol Biol.* 49(2):164–177.
Shemesh R, Novik A, Cohen Y. 2010. Follow the leader: preference for specific amino acids directly following the initial methionine in proteins of different organisms. *Genomics Proteomics Bioinformatics* 8(3):180–189.
Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: evidences for phylogenetic trends. *Comput Biol Chem* 33(3):239–244.
Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* 41(17):8220–8236.
Starmer J, Stomp A, Vouk M, Bitzer D. 2006. Predicting Shine–Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol.* 2(5):e57.
Stenstrom CM, Jin H, Major LL, Tate WP, Isaksson LA. 2001. Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* 263(1-2):273–284.
Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 342(6164):1367–1372.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 24(2):374–381.
Su M-C, Chang C-T, Chu C-H, Tsai C-H, Chang K-Y. 2005. An atypical RNA pseudoknot stimulator and an upstream attenuation signal for −1 ribosomal frameshifting of SARS coronavirus. *Nucleic Acids Res.* 33(13):4265–4275.
Subramaniam AR, Zid BM, O'Shea EK. 2014. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* 159(5):1200–1211.
Tang S-L, Chang BCH, Halgamuge SK. 2010. Gene functionality's influence on the second codon: a large-scale survey of second codon composition in three domains. *Genomics* 96(2):92–101.
Tank EM, True HL. 2009. Disease-associated mutant ubiquitin causes proteasomal impairment and enhances the toxicity of protein aggregates. *PLoS Genet.* 5(2):e1000382.
Tasaki T, Sriram SM, Park KS, Kwon YT. 2012. The N-end rule pathway. *Annu Rev Biochem.* 81:261–289.
Tats A, Remm M, Tenson T. 2006. Highly expressed proteins have an increased frequency of alanine in the second amino acid position. *BMC Genomics* 7:1–13.
Team RC. 2015. R: A Language and Environment for Statistical Computing. Version 4.3.2. Vienna, Austria: R Foundation for Statistical Computing.
Tobias J, Shrader T, Rocap G, Varshavsky A. 1991. The N-end rule in bacteria. *Science* 254(5036):1374–1377.
Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y. 2010. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics* 11(1):1–13.
Tsuchihashi Z, Kornberg A. 1990. Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. *Proc Natl Acad Sci U S A.* 87(7):2516–2520.
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.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A.* 107(8):3645–3650.
Tuller T, Zur H. 2015. Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Res.* 43(1):13–28.
Udagawa T, Shimizu Y, Ueda T. 2004. Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in Eubacteria. *J Biol Chem.* 279(10):8539–8546.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* 13(10):2260–2264.
Van Assche E, Van Puyvelde S, Vanderleyden J, Steenackers HP. 2015. RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Front Microbiol.* 6:141.
Vandivier LE, Li F, Zheng Q, Willmann MR, Chen Y, Gregory BD. 2013. Arabidopsis mRNA secondary structure correlates with protein function and domains. *Plant Signal Behav.* 8(6):e24301.
Varani G, McClain WH. 2000. The G·U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep.* 1:18–23.

Varshavsky A. 2011. The N-end rule pathway and regulation by proteolysis. *Protein Sci.* 20(8):1298–1345.
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, Huang Y, Przytycka TM, Hurst LD. 2010. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. *Genome Biol Evol.* 2:636–645.
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.
Willie E, Majewski J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet* 20(11):534–538.
Wills NM, Moore B, Hammer A, Gesteland RF, Atkins JF. 2006. A functional −1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. *J Biol Chem.* 281(11):7082–7088.
Winter RB, Morrissey L, Gauss P, Gold L, Hsu T, Karam J. 1987. Bacteriophage T4 regA protein binds to mRNAs and prevents translation initiation. *Proc Natl Acad Sci U S A.* 84(22):7822–7826.
Wu X, 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.
Xiao Q, Zhang F, Nacev BA, Liu JO, Pei D. 2010. Protein N-terminal processing: substrate specificity of *Escherichia coli* and human methionine aminopeptidases. *Biochemistry* 49(26):5588–5599.
Xing K, He X. 2015. Reassessing the "Duon" hypothesis of protein evolution. *Mol Biol Evol.* 32(4):1056–1062.
Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci U S A.* 102(38):13526–13531.
Yakhnin H, Yakhnin AV, Babitzke P. 2006. The trp RNA-binding attenuation protein (TRAP) of *Bacillus subtilis* regulates translation initiation of ycbK, a gene encoding a putative efflux protein, by blocking ribosome binding. *Mol Microbiol.* 61(5):1252–1266.
Yamamoto H, Wittek D, Gupta R, Qin B, Ueda T, Krause R, Yamamoto K, Albrecht R, Pech M, Nierhaus KH. 2016. 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria. *Proc Natl Acad Sci U S A.* 113(9):E1180–E1189.
Yang C, Hockenberry AJ, Jewett MC, Amaral LA. 2016. Depletion of Shine–Dalgarno sequences within bacterial coding regions is expression dependent. *G3 (Bethesda)* 6(11):3467–3474.
Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res.* 44(D1):D710–D716.
Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y. 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol Cell* 59(5):744–754.
Zalucki YM, Power PM, Jennings MP. 2007. Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. *Nucleic Acids Res.* 35(17):5748–5754.
Zamora-Romo E, Cruz-Vera LR, Vivanco-Domínguez S, Magos-Castro MA, Guarneros G. 2007. Efficient expression of gene variants that harbour AGA codons next to the initiation codon. *Nucleic Acids Res.* 35(17):5966–5974.
Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol.* 16(3):274–280.
Zhou Z, Fu XD. 2013. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* 122(3):191–207.
Zuo G, Xu Z, Hao B. 2013. Shigella strains are not clones of *Escherichia coli* but sister species in the genus *Escherichia*. *Genomics Proteomics Bioinformatics* 11(1):61–65.

Supplement to Chapter 5

Supplementary Results

The Supplementary Results presented below can also be found accompanying the published paper. These have been reformatted for this thesis. Further Supplementary Figures and Supplementary Tables can be found accompanying the published paper online and on the attached CD.

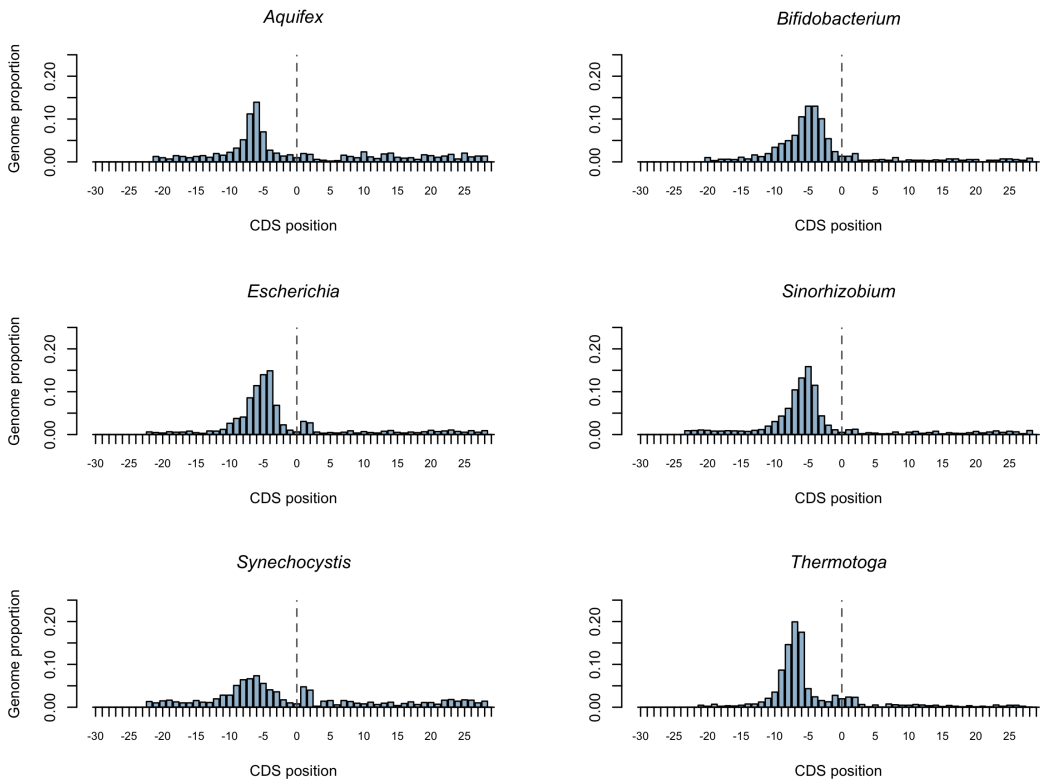
Supplementary Result 1

Sequences lacking upstream Shine-Dalgarno sequences have significantly greater fourth site *A* content

The differences in the translation initiation mechanisms between bacteria and eukaryotes suggest distinct pathways have evolved in the translation initiation process. Bacteria, eukaryotes and archaea, whose features resemble those found both in bacteria and eukaryotes, have the ability to translate genes both with and without additional initiation leaders signals in the 5' untranslated region (Tolstrup et al. 2000; Moll et al. 2002; Benelli et al. 2003; Ring et al. 2007; Akulich et al. 2016). Leaderless mRNAs are universally translatable (Grill et al. 2000) between bacteria, archaea and eukaryotes suggesting a common conserved mechanism of initiating leaderless genes. Leaderless initiation is likely the ancestral mechanism (Londei 2005; Zheng et al. 2011) and in bacteria occurs via strong, more stable interactions with intact 70S ribosomes (Moll et al. 2002; O'Donnell and Janssen 2002; Moll et al. 2004; Zuo et al. 2013) rather than the 30S subunit and is not dependant on ribosomal proteins or initiation factors (Moll et al. 2004; Udagawa et al. 2004). Thus, if leaderless genes reflect the ancestral state, why have leader signals evolved?

It is known that features of prokaryotic mRNA 5' untranslated regions (UTR) contribute to the ability and efficiency of translation (Teilhet et al. 1998; Hayashi et al. 2017). The conserved Shine-Dalgarno (SD) sequence, 5'-GGAGGT-3' sequence is complementary to the 16S rRNA antiSD sequence (Shine and Dalgarno 1974) and typically located 5-10 nucleotides upstream of the start codon (Chen et al. 1994) is found in bacteria and archaea but not eukaryotes. Full or partial complementarity of these sequences facilitate binding of the 16S rRNA to the mRNA, correctly positioning the 30S rRNA subunit at the correct start codon (Nakagawa et al. 2010). The SD sequence is found preferentially in highly expressed genes (Ma et al. 2002) with mutations in either the SD motif (Velazquez et al. 1991) or anti-SD motif (Jacob et al. 1987) reduce protein synthesis levels in *E. coli*, suggesting SD binding provides a precise and critical translation initiation signal. Furthermore in the absence of SD sequences, following initiation the first decoding step at the ribosomal A-site is highly error-prone resulting in significant incorporation of noncognate amino acids (Di Giacco et al. 2008). When SD sequences are present, there is no evidence of this misincorporation, suggesting SD sequences play a key role in translation initiation accuracy. In the 70S ribosomal scanning model proposed by Yamamoto et al. (2016), the absence of a SD sequence significantly weakened initiation. They conclude that the SD sequence provides a strong landing signal allowing the fMet-tRNA to fix the 70S ribosome at the cognate AUG. In the absence of a SD sequence the ribosome can continue to scan the mRNA. It is therefore feasible that leaders have evolved to increase initiation accuracy and that other initiation errors may be more frequent in genes lacking an SD. Could the fourth site be acting as an error control mechanism for genes without SD sequences that are likely to be more error prone in selection the correct start codon? If so, we expect a greater use of *A* in those without a SD sequence.

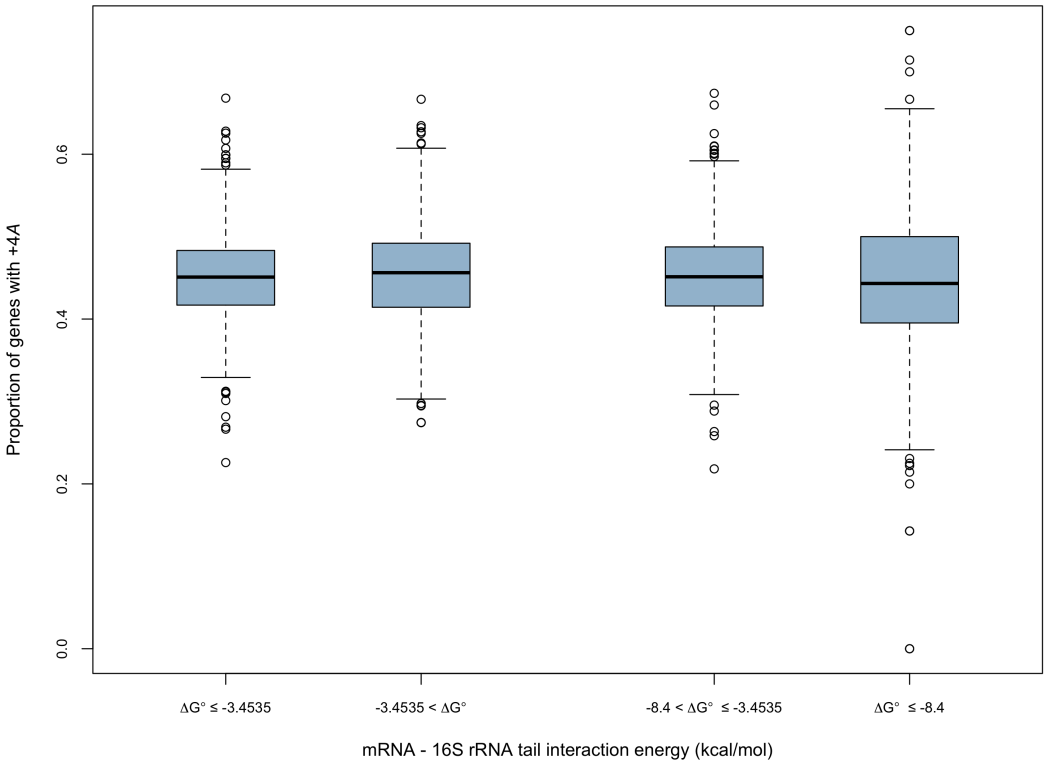
Potential SD sequences were calculated for 399 genomes with suitable 16S rRNA tails. Peaks in the proportion of SD sequences upstream of the start codon (Supplementary Result 1 Figure 1) are consistent with previous work locating SD sequences (Starmer et al. 2006). The proportion of genome CDSs with SD sequences varies considerably (95.57% in *G. kaustophilus* to 4.87% in *A. pleuropneumoniae*) and we find no correlation between GC3 content and the proportion of genes with a SD sequence ($P = 0.155$, Spearman's rank correlation).



Supplementary Result 1 Figure 1: The location of the strongest binding (ΔG°) between the mRNA and 16S rRNA tail identifies that Shine-Dalgarno (SD) sequences are located 5' of the start codon. Coding sequence position 0 is defined as the first nucleotide of the NTG start codon.

We find the distributions of the proportions of *A* content are extremely similar between CDSs with and without an SD sequence (Supplementary Result 1 Figure 2). The proportion of CDS's with fourth site *A* significantly differs between CDSs with and without a SD (P = 0.002, paired Wilcoxon rank-sum test), with the fourth site *A* proportion marginally greater in genes lacking an SD sequence (mean proportion of CDS with fourth site *A*: with SD: 0.451 ± 0.061; without SD: 0.455 ± 0.065, N = 399). Consistent with fourth site *A* being associated with a lack of SD we find a significant increase in a genome's proportion of genes with fourth site *A* for genes with a weak SD-antiSD interaction compared with strong SD-antiSD interactions (P = 0.013, paired Wilcoxon rank-sum test). As the distance of the SD sequence from the start codon is important (Chen et al. 1994), we may expect this distance to affect *A* content however we find no difference genome fourth site *A* usage between CDSs with a SD

sequence close to the start codon (defined as nearer the start codon than the mean SD distance) to those with a SD sequence further away ($P = 0.638$, paired Wilcoxon rank-sum test).



Supplementary Result 1 Figure 2: Distributions of the proportions of genes with +4*A* in the presence of Shine-Dalgarno (SD) sequences (ΔG° ≤ -3.4535 kcal/mol) are similar to those with no SD sequence (ΔG° > -3.4535 kcal/mol). The median proportion of genes with +4*A* of non-SD genes is slightly greater than SD-led genes. The proportion for genes with a strong SD sequence and high complementarity between the 5' mRNA UTR and anti-SD sequence, is more variable (ΔG° ≤ -8.4 kcal/mol) than those with a weak SD (-8.4 < ΔG° ≤ -3.4535 kcal/mol) and has a lower median proportion. These results support the model in which *A* at the fourth site is facilitating translation initiation accuracy in the absence of SD sequences.

Supplementary Result 2

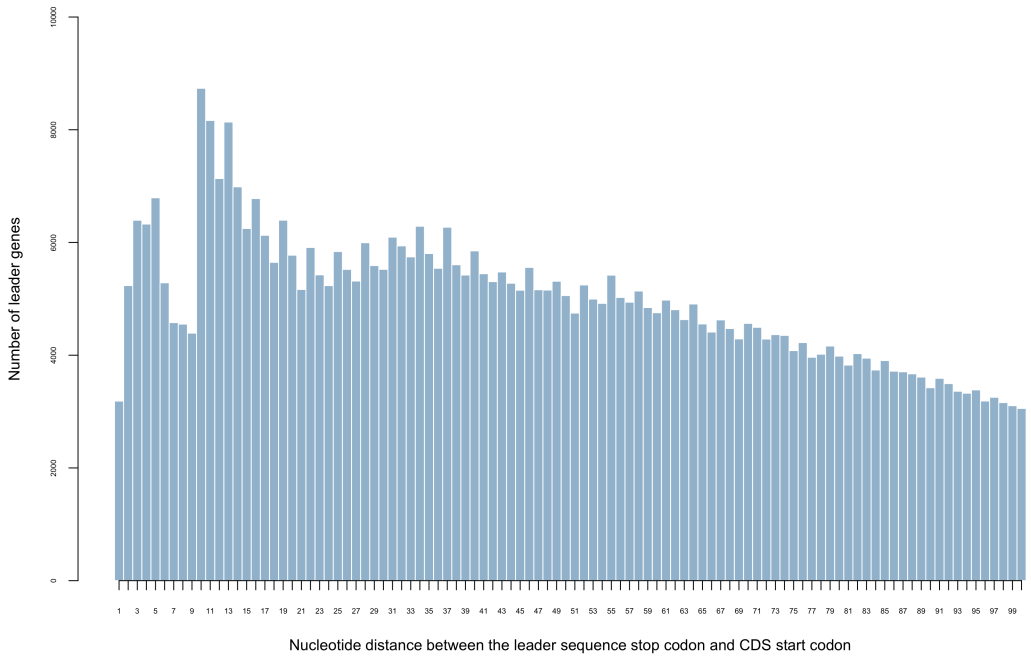
The presence of a leader gene reduces the fourth site *A* prevalence

If fourth site *A* enrichment assists in reducing initiation errors and increases 5' RNA stability, as the data seem to suggest, why then don't all genes use fourth site *A*? Naturally part of the explanation must be mutation-selection equilibrium, which will predict a dynamic equilibrium between mutations removing *A* and selection favouring *A*. The lack of a SD sequence also weakly predicts increased *A* usage. Is there an alternative explanation? Here we demonstrate that the presence of leader genes, as opposed to genes with an addition leader signal in Supplementary Result 1, appears to explain some instances of non-fourth site *A*.

Bacteria use a variety of premature termination signals to control gene expression. Genes regulated in this way contain termination signals located in non-protein coding leader genes 5' of the structural mRNA, with up to 10% of operons regulated by a transcription attenuation mechanism (Henkin and Yanofsky 2002) with attenuation signals varying between species and the expressed structural gene. Of particular interest in this study are 5' leader peptides translated prior to CDS translation to situate the ribosome within the vicinity of the CDS (Naville and Gautheret 2010). The influence a leader gene may have on a structural CDS fourth site *A* content is however unknown. High resolution 70S ribosome imaging in the elongation phase indicates approximately 30 nucleotides of the mRNA transcript are encompassed by the ribosome from positions -18 to +12 relative to the current translation site (Demeshkina et al. 2010). If the distance between the leader gene stop codon and the start codon of a structural CDS is short, is it possible the ribosome simultaneously accommodates both the leader gene and structural CDS, guiding the ribosome to the translation initiation site and facilitating the re-initiation of translation (Korolev et al. 2016). Leader genes could be described as a 'signpost', helping the ribosome track to the correct start codon and reducing initiation errors.

Potential leader genes were identified as described in the Methods. The specific function of the leader gene was not considered, merely the presence upstream of the CDS. The proportion of CDSs with potential leader genes varies across genomes, from

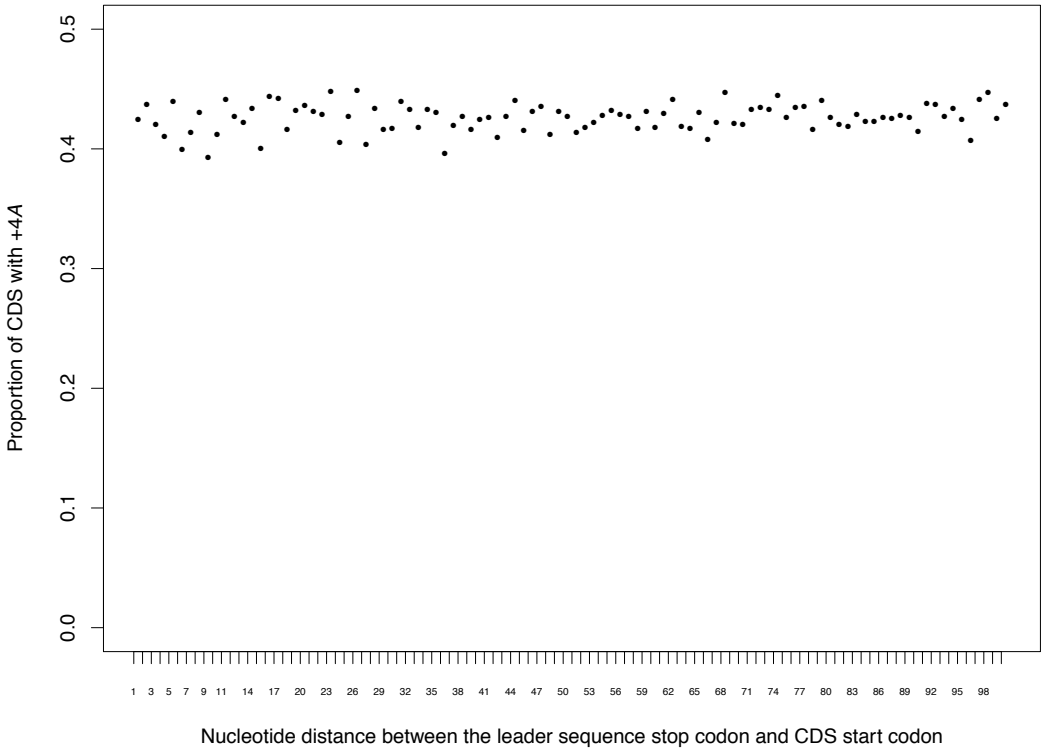
8.08% in *H. thermophilus* to 76.48% in *T. erythraeum* and is significantly but weakly correlated with GC content (ρ = -0.089, P = 0.023, Spearman's rank correlation). We find a peak in the frequency of leader genes 10-13 nucleotides upstream from the CDS (Supplementary Result 2 Figure 1) and a reduced peak 3-5 nucleotides upstream from the CDS. Leader genes at these peaks could accommodate the downstream CDS within the ribosome during translation. GC-rich genomes typically have longer CDSs (Xia et al. 2003) and so these short distances could be attributed to genomes having poor GC content. However, we find a significant negative correlation between GC content and mean distance from leader gene to the CDS (ρ = -0.422, $P < 2.2 \times 10^{-16}$, Spearman rank correlation).



Supplementary Result 2 Figure 1: The number of leader genes at each nucleotide distance to the downstream coding sequence start codon for all genomes. Two peaks of distances are observed at 3-5 and 10-13 nucleotides from the coding sequences.

Does the presence of a leader gene influence fourth site *A* content? Comparing the proportion of fourth site *A* in genes with a leader gene and those without in each genome, we find a significant reduction for those with a preceding leader gene ($P < 2.2 \times 10^{-16}$, paired Wilcoxon rank-sum test; mean *A* proportion for CDSs with a leader

= 0.436 ± 0.080 (N = 646), mean *A* proportion for CDSs without leader = 0.480 ± 0.062 (N = 646)). However, we find little variation in the proportion of genes with fourth site *A* as the distance from the CDS increases (Supplementary Result 2 Figure 2), with *A* content not correlated with the nucleotide distance of the leader gene from the CDS (ρ = 0.113, P = 0.263, Spearman's rank correlation). It would seem the presence of a leader gene does influence the *A* content of the fourth site, but is unaffected by the distance of the leader gene from the CDS.



Supplementary Result 2 Figure 2: The proportion of coding sequences with fourth site A in relation to the nucleotide distance of the leader gene from the coding sequence. There is no clear evidence that the distance of the coding sequence from the leader gene has an influence the incorporation of A at the fourth site.

Supplementary Result 3

Multivariate analysis

We have considered each model separately and discussed the implications for fourth site *A* content. May this trend be a combination of selection pressures that determine the ultimate composition of the fourth site? We performed a multivariate analysis predicting the proportions of CDSs with fourth site *A*. As we have measures for two weak predictors (SD and CAI) for a small subset of all genomes, we eliminate these variables. Models using these predictors do not significantly fit the data. Our resulting model (N = 651) explains 54.73% of the variation (adjusted R-squared: 0.5473) with overall significant fit (F-statistic: 197.5 on 4 and 646 df, $P < 2.2 \times 10^{-16}$), with 5' *A* richness ($P < 2 \times 10^{-16}$), proportion of leader genes ($P = 5.45 \times 10^{-11}$) and proportion of *A*-starting codons ($P = 5.80 \times 10^{-8}$) significant predictors. The genome translation table, determining whether a TGA stop is used, is a nearly significant predictor ($P = 0.053$).

We also consider a gene level model in which use of *A* at the fourth site is a binary variable. Under this model, the genome use of *A*-starting codons ($P < 2.2 \times 10^{-16}$), local 5' *A* richness ($P < 6.35 \times 10^{-9}$) and whether nor not the gene has a leader ($P < 2.2 \times 10^{-16}$) are significant predictors. We also find a significant interaction term between the leader gene and 5' *A* richness ($P < 2.2 \times 10^{-16}$). Again, the influence of the translation table is nearly significant ($P = 0.055$).

Supplementary Result 4

CDS fourth site *A* acting to prevent a ribosomal start codon readthrough

Mapping has shown that 60-70% of genes in prokaryotes are transcribed as part of an operon (Sorek and Cossart 2010). If the ribosome were to continue to scan the mRNA downstream of a gene, translating multiple sequences from an operon as part of polycistronic transcript, could the presence of an immediate +1 stop codon prevent readthrough of the start codon? The idea of ribosome scanning is not new (Adhin and van Duin 1990; Osterman et al. 2013). Furthermore, correct initiation is predominantly (Haimov et al. 2015) accomplished via a ribosomal scanning mechanism in eukaryotes (Kozak 1978; Agarwal and Bafna 1998; Hinnebusch 2014). Under a bacterial scanning model, Yamamoto et al. (2016) suggest the 70S ribosome does not dissociate following previous CDS translation termination but continues the surrounding sequence for a SD sequence.

If the fourth site prevents readthrough as the ribosome translocates the mRNA between CDSs, we would expect greater *A*-content in the CDSs with an upstream protein-coding CDS on the same strand. Do we find the number of CDSs with +4*A* and an upstream CDS on the same strand greater than expected by chance, given the total CDSs with +4*A* and an upstream CDS? Excluding overlapping genes (ensuring inter-CDS regions allowing scanning, we observe no significant effect of the strand of the upstream CDS on fourth site *A* content ($P \approx 1$, Pearson's cumulative test statistic (χ^2)). This analysis however accounts for genes located on different operons or at distances in which ribosome scanning is unlikely to occur (mean distance between CDS = 1872.64 nucleotides). Restricting the inter-CDS region to 10, 20, 30, 40, 50, 100, 150 or 200 nucleotides did not influence A content ($P \approx 1$, Pearson's cumulative test statistic (χ^2)). Fourth site *A* content is therefore unlikely to be under selection to provide a translocating ribosome assistance in locating the start codon.

Supplementary Result 5

CDSs with less efficient initiation codon TTG demonstrate weakest fourth site *A* enrichment

CDSs with different start codons are translated with different efficiencies (O'Donnell and Janssen 2001; Osterman et al. 2013; Panicker et al. 2015; Hecht et al. 2017). In vitro ribosome binding strength, as estimated from toeprint assays in *E. coli* assays, revealed 30S subunits bound most efficiently to leadered mRNA containing an ATG, followed by GTG and TTG (O'Donnell and Janssen 2001). We hypothesise two ways in which the start codon identity may determine fourth site *A* usage. First, the fourth site may be used more frequently for weaker binding start codons to prevent the ribosome dissociating with the correct initiation site prematurely. Alternatively, the fourth site *A* may be contributing towards the additional strength of binding for ATG start codons by providing an additional interaction between the ribosome and mRNA.

We find CDSs starting GTG (mean A_4 ratio = 2.607 ± 0.688, N = 646) and ATG (mean A_4 ratio = 1.887 ± 0.367, N = 646) demonstrate greater enrichment than TTG (mean A_4 ratio = 1.274 ± 0.319, N = 646), suggesting the weaker start codons are not compensated for with greater *A* content. This is suggestive that fourth site *A* is not assisting the particularly weak start codons. Panicker et al. (2015) and Osterman et al. (2013) report that in some cases, GTG is a more efficient initiator of translation. The role of fourth site *A* may be reflected in this increased initiation efficiency, although the evidence is not definitive and the ribosome may attempt to use an alternative start codon. The reduced *A* content in TTG might reflect lower expression and hence lower associated error cost due to an initiation error, as opposed to increasing the efficiency and accuracy of ATG and GTG. The mean CAI varies significantly dependant on the start codon ($P < 0.001$, Kruskal-Wallis rank sum test). Further, CDSs starting TTG have significantly lower CAI than those starting ATG ($P < 0.001$, pairwise Tukey-Kramer test) but not GTG ($P = 0.371$, pairwise Tukey-Kramer test). Thus, the reduced expression of TTG, in particular when compared with GTG, is unlikely to explain the differences in *A* use.

Supplementary Result 6

Fourth site *A* functionality is specific to prokaryotes

Drawing comparisons with other species may provide further insights into the fourth site functionality. For example, is enrichment specific to bacteria or prokaryotes more generally? Do we observe fourth site *A* enrichment in eukaryotes?

Supplementary Result 6.1 - Fourth site enrichment in archaea is comparable to bacteria

Archaea are an interesting domain to investigate evolutionary links between prokaryotes and eukaryotes. Features of archaeal translation initiation resemble those found both in bacteria and eukaryotes. Initiation factors, for example, have close homologues with eukaryote initiation factors (Kyrpides and Woese 1998). Conversely, mRNA structure and mRNA-ribosome recognition via SD interactions with 16S rRNA anitSD motifs resembles initiation consistent with eubacteria (Condo et al. 1999; Tolstrup et al. 2000; Slupska et al. 2001; Sartorius-Neef and Pfeifer 2004). Archaeal genomes also possess a major proportion of CDSs that lack 5′ UTR sequences entirely (Condo et al. 1999; Chang et al. 2006). The ability to translate leaderless mRNA's, absent of SD sequences, and those with a SD sequence suggest that two distinct translational mechanisms exist in archaea (Tolstrup et al. 2000; Benelli et al. 2003; Ring et al. 2007). The unique archaeal initiation dynamics can therefore provide insights into fourth site functionality. If the fourth site is important in translation initiation, in particular with ribosome-mRNA interactions, we predict an *A* enrichment similar to that observed in eubacteria.

Replicating previous analyses, we observe significant enrichment of *A* at the fourth site in 73/77 genomes (94.81%) ($P < 0.01$, Pearson's cumulative test statistic (χ^2), Bonferroni correction). Enrichment in the 5' domain is suggestive of selection for determining RNA stability; synonymous sites each exhibit enrichment (mean A_6 = 1.533 ± 0.684, mean A_9 = 1.425 ± 0.458, mean A_{12} = 1.513 ± 0.540, N = 77) yet are not significantly different ($P = 0.587$, Kruskal-Wallis rank-sum test; A_6 - A_9: $P = 0.780$,

A_6 - A_{12}: P = 0.940, A_9 - A_{12}: P = 0.570, pairwise Tukey-Kramer tests). However, as with eubacteria, nonsynonymous sites do exhibit localised *A* enrichment (mean A_4 = 1.566 ± 0.405, mean A_7 = 1.187 ± 0.088, mean A_{10} = 1.181 ± 0.106, N = 77) with the fourth site is significantly enriched beyond neighbouring codons ($P < 2.2 \times 10^{-16}$, Kruskal-Wallis rank-sum test; $A_4 – A_7$: $P = 1.10 \times 10^{-13}$, $A_4 – A_{10}$: $P = 3.40 \times 10^{-14}$, A_7 - A_{10}: P = 0.980, pairwise Tukey-Kramer tests). These results suggest archaea are under similar selection pressures at the fourth site.

Supplementary Result 6.2 - Weak *A* enrichment in *S. cerevisiae* may reduce RNA stability but is not observed in the second codon

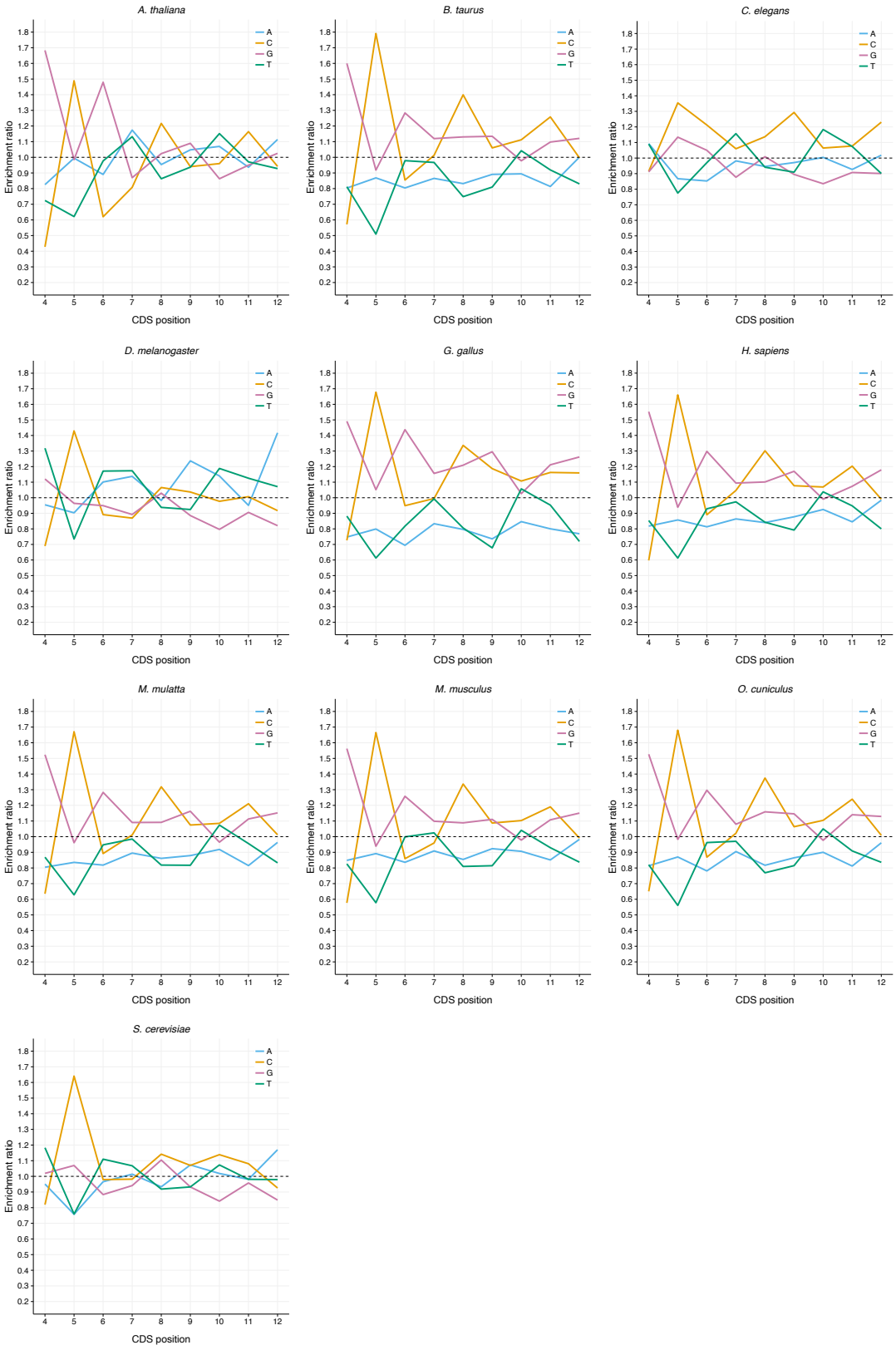
Does this enrichment extend to eukaryotes? Of interest is *S. cerevisiae*, in which 5' RNA stability is also known to effect expression (Shah et al. 2013). We therefore also expect an *A* enrichment in the 5' domain for *S. cerevisiae* CDSs. Both nonsynonymous A_7 (1.014) and A_{10} (1.018) ratios and synonymous A_9 (1.073) and A_{12} (1.170) ratios provide evidence of weak selection, yet we do not observe an *A* enrichment in the second codon (A_4 = 0.951, A_6 = 0.966). Interestingly, we find a weak *T* enrichment (T_4 = 1.183, T_6 = 1.110), which may provide the RNA destabilising effect. Notably, we find no evidence of selection specific to fourth site *A*.

Supplementary Result 6.3 - Eukaryotic species exhibit no fourth site enrichment specific to *A*

Is there any evidence of selection consistent with RNA stability or fourth site enrichment in other eukaryotes? We find variable enrichment profiles for codons 2-4 of various eukaryotes (Supplementary Result 6 Figure 1, Supplementary Result Figure 2). *C. elegans, D. melanogaster* and *A. thaliana* each exhibit *T* enrichment at the seventh and tenth sites, whilst *D. melanogaster* and *A. thaliana* exhibit an *A/T* preference in both first and synonymous sites of codons 3 and 4. A reduction in mRNA folding increasing the accessibility of the RNA in the CDS termini in each of these species has previously been documented (Li et al. 2012a; Li et al. 2012b; Vandivier et al. 2013) which these results seemingly confirm. A strong *G/C* bias at each position in the 5' domain of other eukaryotes would suggest that RNA stability selection is not

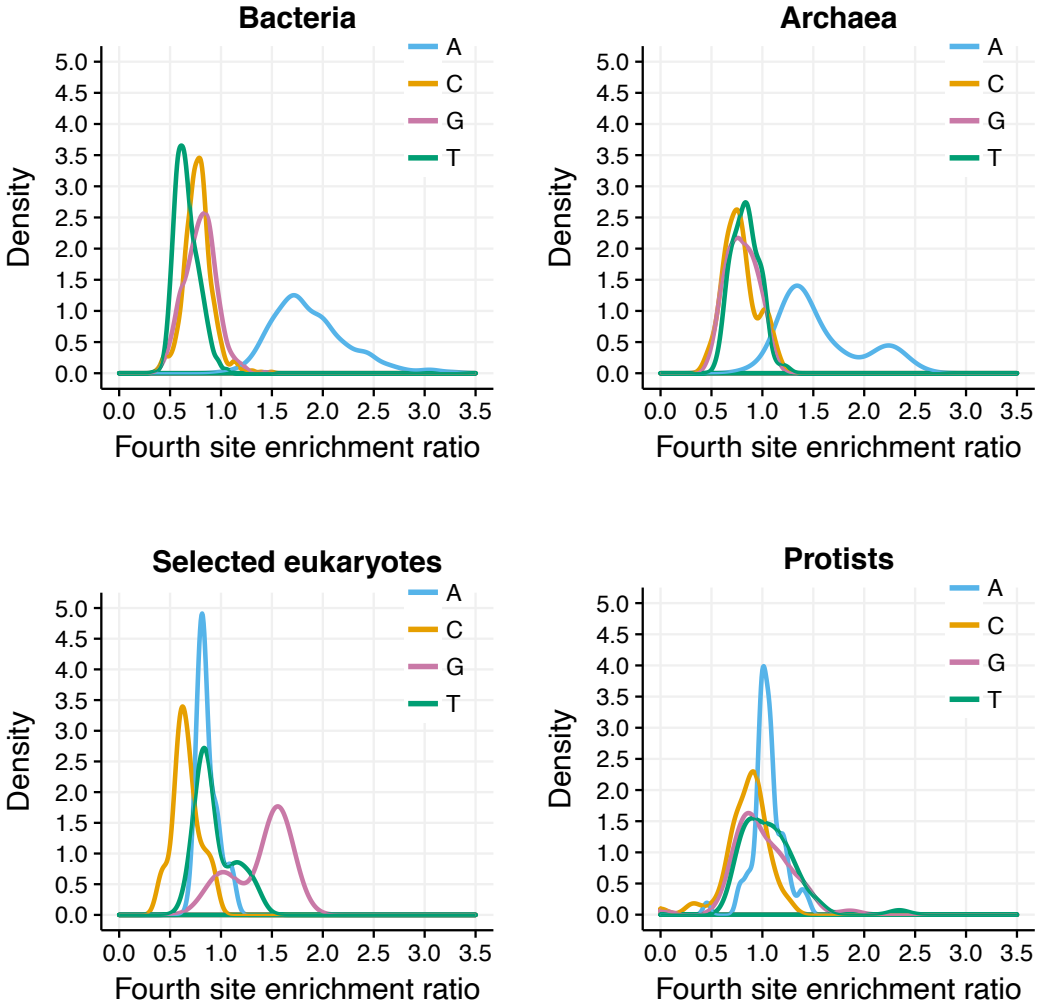
universal. In each eukaryotic species except *C. elegans*, we observe an enrichment of *G* in the fourth site. The Kozak sequence, for which fourth site *G* is an important nucleotide of the canonical GCCRCCAUGG (Kozak 1986, 1997) motif in eukaryotic ribosome binding may explain this enrichment. We also observe *T* enrichment at the fourth site in *C. elegans* and *D. melanogaster* that has previously been described in invertebrates and fungi (Nakagawa et al. 2008). Whilst *C. elegans* exhibits an enrichment of *A* at the fourth site, this is almost identical to *T* enrichment ($A_4 = 1.093$, $T_4 = 1.090$).

This change in enrichment profiles may however reflect the weakened purifying selection in eukaryotes not being able to maintain fourth site *A*. We consider the enrichment in protist genomes, for which effective population sizes are larger and therefore likely to be under stronger purifying selection. As with the previous selected eukaryotes, we find no evidence for specific fourth site *A* enrichment (Supplementary Result 6 Figure 2). The *Paramecium* genome, considered to have a large effective population size (Snoke et al. 2006), has an A_4 ratio of 1.204. However, 637/646 (98.61%) of bacterial and 68/77 (88.31%) of archaea genomes have a greater A_4 ratio greater than this value. Fourth site ratios for both bacteria ($P < 2.2 \times 10^{-16}$, Wilcoxon rank sum test) and archaea ($P = 1.786 \times 10^{-11}$, Wilcoxon rank sum test) are significantly greater in both cases. The maximum enrichment ratio for the protists is 1.449 for *P. tricornutum*, lower than 586/646 (90.71%) bacterial genomes. Thus, the reduced enrichment is consistent across the eukaryotic domain and unlikely to be due to weakened purifying selection not being able to maintain this enrichment.



Supplementary Result 6 Figure 1: Enrichment ratios in eukaryotes provide no evidence of selection for increased *A* content specific to coding sequence fourth sites. *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *A. thaliana* demonstrate bases towards

A/T in both synonymous and nonsynonymous sites. A general *G/C* bias is observed in vertebrate CDS, with the fourth site under strong selection for *G* content.



Supplementary Result 6 Figure 2: Enrichment ratios in the selected eukaryotes and protists demonstrate no bias specific to *A* at the fourth site that is observed for both bacteria and archaea. Eukaryotes demonstrate a clear enrichment of *G* at the fourth site, likely to reflect selection for nucleotides within the Kozak sequence.

References

Adhin MR, van Duin J. 1990. Scanning model for translational reinitiation in eubacteria. *J. Mol. Biol.* 213:811-818.

Agarwal P, Bafna V. 1998. The ribosome scanning model for translation initiation: implications for gene prediction and full-length cDNA detection. *Proc Int Conf Intell Syst Mol Biol* 6:2-7.

Akulich KA, Andreev DE, Terenin IM, Smirnova VV, Anisimova AS, Makeeva DS, Arkhipova VI, Stolboushkina EA, Garber MB, Prokofjeva MM, et al. 2016. Four translation initiation pathways employed by the leaderless mRNA in eukaryotes. *Sci Rep* 6:37905.

Benelli D, Maone E, Londei P. 2003. Two different mechanisms for ribosome/mRNA interaction in archaeal translation initiation. *Mol. Microbiol.* 50:635-643.

Chang B, Halgamuge S, Tang SL. 2006. Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene* 373:90-99.

Chen H, Bjerknes M, Kumar R, Jay E. 1994. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res.* 22:4953-4957.

Condo I, Ciammaruconi A, Benelli D, Ruggero D, Londei P. 1999. Cis-acting signals controlling translational initiation in the thermophilic archaeon Sulfolobus solfataricus. *Mol. Microbiol.* 34:377-384.

Demeshkina N, Jenner L, Yusupova G, Yusupov M. 2010. Interactions of the ribosome with mRNA and tRNA. *Curr. Opin. Struct. Biol.* 20:325-332.

Di Giacco V, Marquez V, Qin Y, Pech M, Triana-Alonso FJ, Wilson DN, Nierhaus KH. 2008. Shine-Dalgarno interaction prevents incorporation of noncognate amino acids at the codon following the AUG. *Proc Natl Acad Sci U S A* 105:10715-10720.

Grill S, Gualerzi CO, Londei P, Blasi U. 2000. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *EMBO J.* 19:4101-4110.

Haimov O, Sinvani H, Dikstein R. 2015. Cap-dependent, scanning-free translation initiation mechanisms. *Biochim Biophys Acta* 1849:1313-1318.

Hayashi R, Sugita C, Sugita M. 2017. The 5' untranslated region of the rbp1 mRNA is required for translation of its mRNA under low temperatures in the cyanobacterium Synechococcus elongatus. *Arch. Microbiol.* 199:37-44.

Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. 2017. Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* 45:3615-3626.

Henkin TM, Yanofsky C. 2002. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays* 24:700-707.

Hinnebusch AG. 2014. The scanning mechanism of eukaryotic translation initiation. *Annu Rev Biochem* 83:779-812.

Jacob WF, Santer M, Dahlberg AE. 1987. A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. *Proc Natl Acad Sci U S A* 84:4757-4761.

Korolev SA, Zverkov OA, Seliverstov AV, Lyubetsky VA. 2016. Ribosome reinitiation at leader peptides increases translation of bacterial proteins. *Biol Direct* 11:20.
Kozak M. 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* 15:1109-1123.
Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44:283-292.
Kozak M. 1997. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J.* 16:2482-2492.
Kyrpides NC, Woese CR. 1998. Universally conserved translation initiation factors. *Proc Natl Acad Sci U S A* 95:224-228.
Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, Valladares O, Yang J, Bambina S, Sabin LR, et al. 2012a. Global analysis of RNA secondary structure in two metazoans. *Cell Rep* 1:69-82.
Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012b. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* 24:4346-4359.
Londei P. 2005. Evolution of translational initiation: new insights from the archaea. *FEMS Microbiol. Rev.* 29:185-200.
Ma J, Campbell A, Karlin S. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.* 184:5733-5745.
Moll I, Grill S, Gualerzi CO, Blasi U. 2002. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol. Microbiol.* 43:239-246.
Moll I, Hirokawa G, Kiel MC, Kaji A, Blasi U. 2004. Translation initiation with 70S ribosomes: an alternative pathway for leaderless mRNAs. *Nucleic Acids Res.* 32:3354-3363.
Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K. 2008. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.* 36:861-871.
Nakagawa S, Niimura Y, Miura K, Gojobori T. 2010. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc Natl Acad Sci U S A* 107:6382-6387.
Naville M, Gautheret D. 2010. Transcription attenuation in bacteria: theme and variations. *Brief Funct Genomics* 9:178-189.
O'Donnell SM, Janssen GR. 2001. The initiation codon affects ribosome binding and translational efficiency in Escherichia coli of cI mRNA with or without the 5' untranslated leader. *J. Bacteriol.* 183:1277-1283.
O'Donnell SM, Janssen GR. 2002. Leaderless mRNAs bind 70S ribosomes more strongly than 30S ribosomal subunits in Escherichia coli. *J. Bacteriol.* 184:6730-6733.
Osterman IA, Evfratov SA, Sergiev PV, Dontsova OA. 2013. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res.* 41:474-486.
Panicker IS, 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:e0127911.
Ring G, Londei P, Eichler J. 2007. Protein biogenesis in Archaea: addressing translation initiation using an in vitro protein synthesis system for Haloferax volcanii. *FEMS Microbiol. Lett.* 270:34-41.

Sartorius-Neef S, Pfeifer F. 2004. In vivo studies on putative Shine-Dalgarno sequences of the halophilic archaeon Halobacterium salinarum. *Mol. Microbiol.* 51:579-588.

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. *Cell* 153:1589-1601.

Shine J, Dalgarno L. 1974. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* 71:1342-1346.

Slupska MM, King AG, Fitz-Gibbon S, Besemer J, Borodovsky M, Miller JH. 2001. Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum. *J. Mol. Biol.* 309:347-360.

Snoke MS, Berendonk TU, Barth D, Lynch M. 2006. Large global effective population sizes in Paramecium. *Mol. Biol. Evol.* 23:2474-2479.

Sorek R, Cossart P. 2010. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.* 11:9-16.

Starmer J, Stomp A, Vouk M, Bitzer D. 2006. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol* 2:e57.

Teilhet M, Rashid MB, Hawk A, Al-Qahtani A, Mensa-Wilmot K. 1998. Effect of short 5' UTRs on protein synthesis in two biological kingdoms. *Gene* 222:91-97.

Tolstrup N, Sensen CW, Garrett RA, Clausen IG. 2000. Two different and highly organized mechanisms of translation initiation in the archaeon Sulfolobus solfataricus. *Extremophiles* 4:175-179.

Udagawa T, Shimizu Y, Ueda T. 2004. Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in eubacteria. *J. Biol. Chem.* 279:8539-8546.

Vandivier L, Li F, Zheng Q, Willmann M, Chen Y, Gregory B. 2013. Arabidopsis mRNA secondary structure correlates with protein function and domains. *Plant Signal Behav* 8:e24301.

Velazquez L, Camarena L, Reyes JL, Bastarrachea F. 1991. Mutations affecting the Shine-Dalgarno sequences of the untranslated region of the Escherichia coli gltBDF operon. *J. Bacteriol.* 173:3261-3264.

Xia X, Xie Z, Li WH. 2003. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. *J. Mol. Evol.* 56:362-370.

Yamamoto H, Wittek D, Gupta R, Qin B, Ueda T, Krause R, Yamamoto K, Albrecht R, Pech M, Nierhaus KH. 2016. 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria. *Proc Natl Acad Sci U S A* 113:E1180-1189.

Zheng X, Hu G-Q, She Z-S, Zhu H. 2011. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* 12:361.

Zuo G, Xu Z, Hao B. 2013. Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia. *Genomics Proteomics Bioinformatics* 11:61-65.

Chapter 6

Discussion

That gene expression is a complex multi-step process means errors can and do happen at each step of the expression pathway. Consequently, although selection on the gene sequence encoding the correct amino acids is a strong and necessary constraint, it is becoming increasingly acknowledged that a significant proportion of the selective constraint is devoted to the control and mitigation of errors (Drummond and Wilke 2009; Warnecke and Hurst 2011). Often selection for error control overlaps CDS. Can we identify genomic patterns that are involved in the mitigation of errors? Are these patterns themselves a cause of the errors?

In this thesis, I have considered several approaches to these questions using stop codons as an exemplar. Stop codons are ideal sequence motifs to understand error-related because their canonical function is well-understood, and so any recurring presence or absence of stop codons not at the end of the CDS could be suggestive of function. As work demonstrating splicing is a key genome constraint (Parmley et al. 2007; Savisaar and Hurst 2018), I first show that the regulatory splicing signals themselves are constrained by protein-coding requirements. In itself, this is a logical insight, but one that has not previously been considered. However, the implications of such a constraint are wider-reaching. First, that noncoding sequences are depleted in stop codons is *a priori* unexpected. However, as lincRNA sequences also are thought to be processed and employ the same regulatory ESEs to ensure accurate splicing as protein-coding transcripts (reviewed in Will and Luhrmann 2011; De Conti et al. 2013; Krchnakova et al. 2019), the depletion, specific to the portions of genes thought to match ESE, makes sense. This work therefore highlights a novel pattern of sequence evolution – that the pattern observed in a gene may not relate directly to the functioning of that gene.

Although itself an interesting result, this depletion of stop codons in ESEs due to the protein-coding constraints has a second consequence. In Chapter 3, I consider how nonsense mutations can be associated with exon skipping, hypothesising that ESE motifs are vulnerable to stop codon-creating mutations (in any frame). While mutations disrupting splicing is nothing new (Baralle and Baralle 2005; Parmley et al. 2006; Hurst and Batada 2017; Anna and Monika 2018), the effects of mutations creating nonsense codons (PTCs) are less well understood. We find a non-negligible proportion of nonsense mutations do appear to exert their effects via splicing.

Importantly, both the computational and experimental work is consistent with exon skipping being the result of the disruption of splice motifs and not a "scanning" feedback mechanism. That such errors occur but would not typically be considered splice-disrupting variants argues for experimental validation of the effects of nonsense mutations to be sure of their true effects.

Several other questions arose during this work. First, are the exons that are skipped exceptional in any way? I find no evidence to support this notion. If NAS was an adaptive mechanism to save transcripts from NMD, these exons should be more frequently of length three and thus having minimised downstream reading frame effects, or be shorter exons than expected, but this is not the case. The genes disrupted by the PTCs do tend to be more tissue-specific, suggesting that disruption of these, on the whole, has less of a phenotypic consequence. This leads to the second question - how do we interpret the level of genome-wide NAS? One could argue that 6% is relatively low, although this is far from a negligible proportion. However, this estimate was calculated from "healthy" individuals and is therefore a conservative baseline frequency. I therefore made use of the publicly available ClinVar dataset cataloguing pathogenic mutations, finding that when disease-related PTCs occur they are distributed as expected if many of them are disrupting splicing. Unlike the 1000 Genomes PTCs, I find ≈33% of pathogenic nonsense mutations may affect splicing and that they disproportionately hit ESE motifs. Thus, that stop codons are found infrequently in ESEs is not only of evolutionary importance but is also likely clinically and therapeutically relevant.

The regularity with which ESEs appear throughout the genome (at least in humans) and their strong impact on sequence evolution (Parmley and Hurst 2007; Caceres and Hurst 2013; Savisaar and Hurst 2017b) means it is unlikely that out of frame stop codons themselves can function as common error-proofing signals in protein-coding sequences that require splicing. I therefore focused my attention on genomes where the effects of splicing and ESE-related constraints are not applicable. Much has been made of OSCs and the ambush hypothesis in the literature (Seligmann and Pollock 2004; Singh and Pardasani 2009; Morgens et al. 2013; Bertrand et al. 2015) although no work has demonstrated conclusive evidence supporting their utility, particularly in

an evolutionary context. However, although each of these studies has its limitations, my primary concern was why the simulation models (Morgens et al. 2013) did not preserve the protein-coding sequence and whether this could explain the results. That codon usage bias is most pronounced in highly expressed genes (Ikemura 1981) and codon usage is predictive of rates of gene expression (Sharp and Li 1987; Sharp et al. 2005; Brandis and Hughes 2016) suggests any selection for OSCs is likely to occur at synonymous sites. In my models, I find results remain inconclusive. However, this result only considers the cost of an error after occurring. When I consider the rate at which errors might occur, we find the data is then consistent with two strategies: first, prevent frameshift errors occurring or second, if they occur, select for OSCs to catch them.

Yet, several questions remain. Why do the bacterial genomes favour TAA and TGA as OSCs and not TAG? Nucleotide biases no doubt contribute, but TAG and TGA are identical in nucleotide content and so TAG should be favoured as it is less error-prone (Meng et al. 1995; Korkmaz et al. 2014; Wei et al. 2016). One model, in which TAG is selectively less favourable (although the reason why is not understood) may explain this trend (Povolotskaya et al. 2012). Why, in addition, do we find that OSC excesses are predominantly in the +1 reading frame? I conjecture that this involves translational dynamics – that a +1 frameshift is more likely to occur than a +2 frameshift that requires more substantial remodelling of the mRNA in the ribosome, or a -1 frameshift which is in the opposite direction to a translating ribosome, however this is purely hypothetical. I therefore conclude that although this dual-strategy model can shed light on the OSC selection and solving the problem of frameshifts, there remains unsolved questions to be addressed.

The work in Chapter 4 suggests that the ability to identify specific error-proofing mechanisms more generally may be better served by identifying processes that would most benefit from increased fidelity. Chapter 5 considers the process of translation initiation. Several factors suggest initiation may be error-prone. First, protein activity (as a proxy of gene expression) in *E. coli* increases relative to the number of stop codons in any frame in the 5' UTR (Seligmann 2007), suggesting that ribosomes employ stop codons to initiate at the correct codon. Second, the ability of ribosomes to correctly bind the mRNA is affected by RNA secondary structures (Kudla et al.

2009; Gu et al. 2010). Thus, being able to first bind and then locate the CDS start codon is difficult. I therefore hypothesised that, similar to Seligmann (2007), stop codons may be located 3' to the start codon. The work in Chapter 5 suggests a strong A enrichment (in 99% of bacterial genomes) immediately following the start codon is most parsimoniously explained in terms of the creation of a +1 stop codon. If so this would reinforce the notion that translation initiation is error-prone and that selection acts to reinforce the process.

Understanding the costs of errors

A fundamental assumption of the work in this thesis is that when errors occur, they have negative net fitness costs whether this is to the cell or organism as a whole. Consistent with this, many of the examples cited in Chapter 1 describing dual coding mechanisms under selection, including those involving stop codons, are motivated by hypotheses that assume that selection acts to minimise any costs of errors. However, the work in Chapter 4 highlighted that the cumulative "cost" of an error needs to be considered - that is the frequency of the error, the direct negative effects and secondary effects of compensatory selection.

It is therefore interesting to consider the broader evolutionary context of errors and error-proofing. At first sight, the optimal strategy for any genome would be to prevent the error from occurring altogether, for example by increasing ESE density towards exon ends to reduce the chance of a splicing error when there might be competing signals (Wu and Hurst 2015) or by selecting against potentially disruptive motifs (Li et al. 2012; Diwan and Agashe 2016; Yang et al. 2016). However, if an error prone site is under particularly strong purifying selection, the increased accuracy brought about by such selection might itself have negative and detrimental consequences. A widely cited case is the selection for increased translation accuracy where ribosome accuracy can be increased (reducing the probability of misincorporation) (Ruusala et al. 1984) but compromises translational speed (Wohlgemuth et al. 2010; Wohlgemuth et al. 2011; Ieong et al. 2016). Equally, not selecting for a most optimal sequence for translation may apply more generally to other processes, including splicing (Warnecke and Hurst 2007). Indeed, Melamud and Moult (2009) suggest that splicing is intrinsically noisy and that most alternative splice forms are toxic. However, the

system's equilibrium state is one where enough of the required product is formed but selection cannot act to reduce the unwanted splicing to zero as it is too costly. Thus, what first might seem like the optimal strategy may be less optimal when considering other factors.

A different mode of selection may explain why stop codon containing ESE motifs persist if they do not have the most optimal sequence for inclusion in protein-coding sequence. As suggested, this could be explained by their prevalence and functioning in noncoding sequence that defines a greater proportion of the spliced genome. However, they are per motif more frequent in protein-coding sequence. Thus, although these motifs contain stop codons, they may bind SR proteins that provide specific and as yet unknown splice functionality and the inability to be included in any frame compromised for binding utility. Even if a stop codon containing motif is included one nucleotide further from the splice junction than it would be if in frame, it may still be able to recruit the SR proteins and thus spliceosome close enough to the splice junction. Whether such a subtle change in the distance of an ESE to the splice junction has a significant effect is as yet unknown. Equally, could the pattern of OSCs be influenced by selection to promote/avoid RBP binding in bacteria (much like Savisaar and Hurst (2017a))? For example, one could imagine a scenario where OSCs are avoided in the CDS because the motifs RBPs bind are also depleted similarly to ESEs, although as ribosomal RBPs are thought to make up a larger proportion of bacterial RBPs this is less likely (Holmqvist and Vogel 2018).

The questions addressed in this work also assume that errors incur a net negative fitness cost. However, while not commonplace, errors may instead have beneficial consequences. For example, the effects of nonsense mutations in the *DMD* gene resulting in loss of functional protein (Aartsma-Rus et al. 2016) and in Duchenne muscular dystrophy (DMD) are somewhat alleviated in Becker muscular dystrophy (BMD) (Shiga et al. 1997; Carsana et al. 2005; Helderman-van den Enden et al. 2010; Flanigan et al. 2011; Anthony et al. 2014; Bello et al. 2016; Moore et al. 2017) due to a different error, exon skipping, although we find evidence that this is not a widespread phenomenon. Equally, the OSCs we suggest are under selection in Chapter 3 and Chapter 4 may instead be selected against for genes that undergo programmed ribosomal frameshifting (Farabaugh 1996; Dinman 2006, 2012; Ketteler 2012). If

translational “errors” occur at rates that are not detrimental to the cell and allow for the synthesis of functional protein products, such errors may instead be tolerated and selected for (e.g. selection for sequence motifs and RNA secondary structures inducing frameshifting in HIV Gag-Pol (Namy et al. 2006)). Gene duplications and rearrangements, changes in gene expression and regulation or the translation of noncoding genes are also sources of errors that could lead to genetic novelty and determine evolutionary fate of *de novo* genes (McLysaght and Guerzoni 2015; Schlotterer 2015; McLysaght and Hurst 2016; Stewart and Rogers 2019). Thus, it is important to consider that the absence of selection against particular errors may not indicate that errors do not occur, but rather that the cost of an error may be somewhat alleviated by other unintended positive effects or increased costs associated with preventing the error altogether.

Chapter 4 also highlights a wider methodological concern with the work carried out in Chapter 2 and Chapter 3, namely the “averaging” of selection effects across genes and genomes. For example, when searching for OSC selection, I have looked across all positions in all genes. Not only does this assume that all genes are subject to such selection, but that all positions in a gene are under selection for such a signal. The analyses are therefore sensitive to localised nucleotide compositional biases that could either underestimate the strength of any selection (e.g. few sites have very strong selection for an OSC, but is masked by many sites with no OSC resulting from other selective pressures) or increase the incidence of false-positive hits (e.g. many OSCs occur more frequently than expected, but for reasons other than frameshift errors). This applies more generally to the ESE analyses, where it is assumed that all motifs are functional which need not be the case (as, for example, some motifs will have splice-independent roles or be avoided altogether (Savisaar and Hurst 2016, 2017a)). Although I have attempted to control for such issues by performing nucleotide-/dinucleotide-matched controls or using intra-gene comparisons, I acknowledge the limitations with such methods. The work in Chapter 5 looking at one specific location overcomes this issue. Encouragingly, the results of the applied work in Chapter 3 suggest the more exploratory analyses I have performed are well informed and can be experimentally validated. Furthermore, a recent study suggests OSCs are more frequently downstream of frameshift-prone codons in a variety of genes (Seligmann

2019). Thus, searching for site- and context-specific selection such as that in Chapter 5, but averaged across many genes, does have utility.

Outlook and future perspective

The majority of the analyses in this thesis have been performed computationally, with the exception of the verification of the PTC-associated exon skipping top hit of *ACP1*. Although this one example provides evidence that the computational methods can be robust, it is evident that much of the work could be improved with the support of experimental methods to provide a molecular basis behind the results. For example, the work in Chapter 2 could be supported by experimental evidence using minigene constructs in which the ESE motif is varied between a stop codon containing motif and one not containing a stop (e.g. GATGGA and GATGAA) and either using RNA-seq to determine exon inclusion or using CLIP-seq (Stork and Zheng 2016) to determine the binding efficiencies. If the stop codon motifs are not less efficient at binding SR proteins or encouraging exon inclusion, this would support the hypothesis that the depletion is simply a result of being located in CDS. Work in Chapter 3 could have been improved by also including a similar variant into *ACP1* to show that the effect is not necessarily specific to the PTC, but more generally a disruption of the splice motif. Experimental work that could show a discrepancy between the frameshift rate in AT-rich and GC-rich bacteria would be of utility, for example by engineering genes to express green fluorescent protein (GFP) (Chalfie et al. 1994) if frameshifted.

Perhaps the most interesting analysis that I would like to have performed is one comparing the rates of PTC-associated exon skipping in healthy/non-healthy tissues. It is well documented that the regulation of accurate splicing plays a critical role in pathogenicity particularly in cancers (Srebrow and Kornblihtt 2006; David and Manley 2010), with selection acting against particularly disruptive synonymous mutations (Hurst and Batada 2017). Thus, by comparing tissues in this way, a PTC mutation in a diseased tissue can be compared not only with rates of exon skipping without the PTC in the diseased tissue but the comparable rates of exon skipping in the healthy tissue. I have tried to employ several datasets to do this including data from the Personal Genome Project UK (PGP-UK) (Consortium 2018; Chervova et al. 2019) and the Texas Cancer Research Biobank (Becnel et al. 2016), however, the relative

lack of information makes any analysis redundant. Using data from database such as Genotype-Tissue Expression (GTEx) (Consortium 2013) or the Cancer Genomes Atlas (TCGA) (reviewed in Tomczak et al. (2015)) would be of great utility, however access to these datasets is restricted and by no means guaranteed after application, nor would the results be replicable for other researchers. Thus, the best estimate of disease-associated NAS that I can provide to date is that provided.

Beyond the academic, it is also important to acknowledge the broader context of the work in this thesis. The main aim of the research group is to improve the understanding of transgenes and improve therapeutics. Thus, being able to translate how genomes have evolved to prevent errors and how selection operates on synonymous sites can inform the diagnosis of disease and improve the design of the transgenes themselves. Taking the work in this thesis example, if you are designing a transgene that requires splicing, knowing that a stop codon containing ESE is prone to stop codon creating mutations is valuable. Equally, the knowledge that some stop codon mutations are splice disrupting, rather than acting as early canonical stop codons, can inform treatments regimes that involve PTC-skipping therapies (Keeling et al. 2014; Dabrowski et al. 2018). Equally, knowledge of whether the expression of one gene affects its neighbours (Ghanbarian and Hurst 2015) and how genes insulate themselves from the effects of expression of other genes can help inform us as to where to target transgene insertion. Despite the ongoing work, many error-proofing mechanisms are likely to be unknown and not elucidated. For example, how many phosphorylation or dephosphorylation events happen off-target or at the wrong time? How are such events prevented? How often are RNAs or proteins incorrectly located and what error traps might there be in such circumstances? How do genes evolve to compensate/depend on duplicates of the original gene (Diss et al. 2017)? Have specific motifs evolved to ensure accurate alternative splicing? If there are tissue-specific splice enhancers (Badr et al. 2016), are there motifs that are required more generally? A broader knowledge of such mechanisms would no doubt of great utility as we move towards an era of personalised and precision medicine.

References

Aartsma-Rus A, Ginjaar IB, Bushby K. 2016. The importance of genetic diagnosis for Duchenne muscular dystrophy. *J. Med. Genet.* 53:145-151.

Anna A, Monika G. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* 59:253-268.

Anthony K, Arechavala-Gomeza V, Ricotti V, Torelli S, Feng L, Janghra N, Tasca G, Guglieri M, Barresi R, Armaroli A, et al. 2014. Biochemical characterization of patients with in-frame or out-of-frame DMD deletions pertinent to exon 44 or 45 skipping. *JAMA neurology* 71:32-40.

Badr E, ElHefnawi M, Heath LS. 2016. Computational Identification of Tissue-Specific Splicing Regulatory Elements in Human Genes from RNA-Seq Data. *PloS one* 11:e0166978-e0166978.

Baralle D, Baralle M. 2005. Splicing in action: assessing disease causing sequence changes. *J. Med. Genet.* 42:737-748.

Becnel LB, Pereira S, Drummond JA, Gingras MC, Covington KR, Kovar CL, Doddapaneni HV, Hu J, Muzny D, McGuire AL, et al. 2016. An open access pilot freely sharing cancer genomic data from participants in Texas. *Sci Data* 3:160010.

Bello L, Campadello P, Barp A, Fanin M, Semplicini C, Soraru G, Caumo L, Calore C, Angelini C, Pegoraro E. 2016. Functional changes in Becker muscular dystrophy: implications for clinical trials in dystrophinopathies. *Sci Rep* 6:32439.

Bertrand RL, Abdel-Hameed M, Sorensen JL. 2015. Limitations of the 'ambush hypothesis' at the single-gene scale: what codon biases are to blame? *Mol. Genet. Genomics* 290:493-504.

Brandis G, Hughes D. 2016. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLoS Genet.* 12:e1005926.

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Carsana A, Frisso G, Tremolaterra MR, Lanzillo R, Vitale DF, Santoro L, Salvatore F. 2005. Analysis of dystrophin gene deletions indicates that the hinge III region of the protein correlates with disease severity. *Ann Hum Genet* 69:253-259.

Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. 1994. Green fluorescent protein as a marker for gene expression. *Science* 263:802-805.

Chervova O, Conde L, Guerra-Assuncao JA, Moghul I, Webster AP, Berner A, Larose Cadieux E, Tian Y, Voloshin V, Jesus TF, et al. 2019. The Personal Genome Project-UK, an open access resource of human multi-omics data. *Sci Data* 6:257.

Consortium GT. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45:580-585.

Consortium P-U. 2018. Personal Genome Project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. *BMC Med Genomics* 11:108.

Dabrowski M, Bukowy-Bieryllo Z, Zietkiewicz E. 2018. Advances in therapeutic use of a drug-stimulated translational readthrough of premature termination codons. *Mol Med* 24:25.

David CJ, Manley JL. 2010. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.* 24:2343-2364.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* 4:49-60.
Dinman JD. 2006. Programmed Ribosomal Frameshifting Goes Beyond Viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift. *Microbe (Washington, D.C.)* 1:521-527.
Dinman JD. 2012. Mechanisms and implications of programmed translational frameshifting. *Wiley Interdiscip Rev RNA* 3:661-673.
Diss G, Gagnon-Arsenault I, Dion-Coté A-M, Vignaud H, Ascencio DI, Berger CM, Landry CR. 2017. Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science* 355:630.
Diwan GD, Agashe D. 2016. The Frequency of Internal Shine-Dalgarno-like Motifs in Prokaryotes. *Genome Biol Evol* 8:1722-1733.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* 10:715-724.
Farabaugh PJ. 1996. Programmed translational frameshifting. *Annu. Rev. Genet.* 30:507-528.
Flanigan KM, Dunn DM, von Niederhausern A, Soltanzadeh P, Howard MT, Sampson JB, Swoboda KJ, Bromberg MB, Mendell JR, Taylor LE, et al. 2011. Nonsense mutation-associated Becker muscular dystrophy: interplay between exon definition and splicing regulatory elements within the DMD gene. *Hum. Mutat.* 32:299-308.
Ghanbarian AT, Hurst LD. 2015. Neighboring Genes Show Correlated Evolution in Gene Expression. *Mol. Biol. Evol.* 32:1748-1766.
Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* 6:e1000664.
Helderman-van den Enden AT, Straathof CS, Aartsma-Rus A, den Dunnen JT, Verbist BM, Bakker E, Verschuuren JJ, Ginjaar HB. 2010. Becker muscular dystrophy patients with deletions around exon 51; a promising outlook for exon skipping therapy in Duchenne patients. *Neuromuscular disorders : NMD* 20:251-254.
Holmqvist E, Vogel J. 2018. RNA-binding proteins in bacteria. *Nat. Rev. Microbiol.* 16:601-615.
Hurst LD, Batada NN. 2017. Depletion of somatic mutations in splicing-associated sequences in cancer genomes. *Genome Biol* 18:213.
Ieong KW, Uzun U, Selmer M, Ehrenberg M. 2016. Two proofreading steps amplify the accuracy of genetic code translation. *Proc Natl Acad Sci U S A* 113:13744-13749.
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.* 151:389-409.
Keeling KM, Xue X, Gunn G, Bedwell DM. 2014. Therapeutics based on stop codon readthrough. *Annu Rev Genomics Hum Genet* 15:371-394.
Ketteler R. 2012. On programmed ribosomal frameshifting: the alternative proteomes. *Front Genet* 3:242.
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.* 289:30334-30342.

Krchnakova Z, Thakur PK, Krausova M, Bieberstein N, Haberman N, Muller-McNicoll M, Stanek D. 2019. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Res.* 47:911-928.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255-258.

Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484:538-541.

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* 370:20140332.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17:567-578.

Melamud E, Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res.* 37:4873-4886.

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.* 211:40-48.

Moore RS, Tirupathi S, Herron B, Sands A, Morrison PJ. 2017. Dystrophin Exon 29 Nonsense Mutations Cause a Variably Mild Phenotype. *Ulster Med J* 86:185-188.

Morgens DW, Chang CH, Cavalcanti AR. 2013. Ambushing the Ambush Hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. *BMC Genomics* 14:418.

Namy O, Moran SJ, Stuart DI, Gilbert RJ, Brierley I. 2006. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* 441:244-247.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* 23:301-309.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* 24:1600-1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.

Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biol Direct* 7:30.

Ruusala T, Andersson D, Ehrenberg M, Kurland CG. 1984. Hyper-accurate ribosomes inhibit growth. *EMBO J.* 3:2575-2580.

Savisaar R, Hurst LD. 2016. Purifying Selection on Exonic Splice Enhancers in Intronless Genes. *Mol. Biol. Evol.* 33:1396-1418.

Savisaar R, Hurst LD. 2017a. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Mol. Biol. Evol.* 34:1110-1126.

Savisaar R, Hurst LD. 2017b. Estimating the prevalence of functional exonic splice regulatory information. *Hum. Genet.* 136:1059-1078.

Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* 28:1442-1454.

Schlotterer C. 2015. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet.* 31:215-219.

Seligmann H. 2007. Cost minimization of ribosomal frameshifts. *J. Theor. Biol.* 249:162-167.
Seligmann H. 2019. Localized Context-Dependent Effects of the "Ambush" Hypothesis: More Off-Frame Stop Codons Downstream of Shifty Codons. *DNA Cell Biol.* 38:786-795.
Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* 23:701-705.
Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33:1141-1153.
Sharp PM, Li WH. 1987. The Rate of Synonymous Substitution in Enterobacterial Genes Is Inversely Related to Codon Usage Bias. *Mol. Biol. Evol.* 4:222-230.
Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M. 1997. Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. *J. Clin. Invest.* 100:2204-2210.
Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: Evidences for phylogenetic trends. *Comput. Biol. Chem.* 33:239-244.
Srebrow A, Kornblihtt AR. 2006. The connection between splicing and cancer. *J. Cell Sci.* 119:2635-2641.
Stewart NB, Rogers RL. 2019. Chromosomal rearrangements as a source of new gene formation in Drosophila yakuba. *PLoS Genet.* 15:e1008314.
Stork C, Zheng S. 2016. Genome-Wide Profiling of RNA-Protein Interactions Using CLIP-Seq. *Methods Mol Biol* 1421:137-151.
Tomczak K, Czerwinska P, Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 19:A68-77.
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.* 24:2755-2762.
Warnecke T, Hurst LD. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat. Rev. Genet.* 12:875-881.
Wei Y, Wang J, Xia X. 2016. Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. *Mol. Biol. Evol.* 33:2357-2367.
Will CL, Luhrmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 3:a003707.
Wohlgemuth I, Pohl C, Mittelstaet J, Konevega AL, Rodnina MV. 2011. Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philos Trans R Soc Lond B Biol Sci* 366:2979-2986.
Wohlgemuth I, Pohl C, Rodnina MV. 2010. Optimization of speed and accuracy of decoding in translation. *EMBO J.* 29:3701-3709.
Wu X, 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.* 32:1847-1861.
Yang C, Hockenberry AJ, Jewett MC, Amaral LAN. 2016. Depletion of Shine-Dalgarno Sequences Within Bacterial Coding Regions Is Expression Dependent. *G3 (Bethesda, Md.)* 6:3467-3474.