Selection on synonymous sites: the unwanted transcript hypothesis

Sofia Radrizzani[1,2], Grzegorz Kudla[3], Zsuzsanna Izsvák[4] and Laurence D. Hurst[1†]

[1]Milner Centre for Evolution, Department of Life Sciences, University of Bath, Bath, UK.
[2]Milner Therapeutics Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK.
[3]MRC Human Genetics Unit, Institute for Genetics and Cancer, The University of Edinburgh, Edinburgh, UK.
[4]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Society, Berlin, Germany.
[†]e-mail: bssldh@bath.ac.uk

Abstract

Although translational selection to favour codons that match the most abundant tRNAs is not readily observed in humans, there is nonetheless selection in humans on synonymous mutations. We hypothesise that much of this synonymous-site selection can be explained in terms of protection against unwanted RNAs — spurious transcripts, mis-spliced forms or RNAs derived from transposable elements or viruses. We propose that selection on synonymous sites not only acts to reduce the rate of creation of unwanted transcripts (e.g. via selection on exonic splice enhancers and cryptic splice sites), but, additionally, that high-GC content (but low CpG content), along with intron presence and position, is both particular to functional native mRNA and employed to recognize transcripts as native. Evidence supports this hypothesis with transcription, nuclear export, liquid phase condensation and RNA degradation all recently being shown to promote GC rich transcripts and suppress AU/CpG rich ones. With traps set against AU/CpG rich transcripts, codon usage of native genes in turn evolves to avoid such suppression. That parallel filters against AU/CpG rich transcripts also affect the endosomal import of RNAs further supports the unwanted transcript hypothesis of synonymous site selection and explains the similar design rules that have enabled the successful use of transgenes and RNA vaccines.

[H1] Introduction

As they do not alter the amino acid that is specified, synonymous mutations **[G]** were originally assumed to be neutrally evolving **[G]**[1]. However, further analysis led to the well-accepted notion that selection favours synonymous mutations forming codons that partner with the most abundant iso-acceptor tRNAs in that organism[2-4]. This is thought to reflect selection operating on translational velocity[5], possibly owing to shorter ribosomal dwell times of ribosomes on the mRNA **[Au:OK?]** (although other studies dispute this[6]), or on translational accuracy[7,8]. Such

translational selection **[G]** is more pronounced in fast growing species[9] and in those with large effective population sizes **[G]** (N_e), these being subject to more efficient purifying selection to remove deleterious alleles[10] (**Box 1**).

Within species, translational selection is more pronounced in highly expressed genes, and is thus typically inferred from stronger codon usage bias **[G]** in such genes and from positive correlations between codon usage and tRNA copy numbers[2]. In humans, as expected of a species with small N_e (ref.[11]) (**Box 1**), these hallmarks of translational selection are not readily observed[2,12]. It was thus supposed in early discussions [2] that selection does not operate on human synonymous mutations nor cause genetic disease.

However, this supposition has gradually changed (as reviewed by refs[13,14]). Although there is still no evidence for commonplace translational selection in humans[10], the supposition that codon usage is determined solely by translational selection is now recognised to be wrong. For example, increasing GC content at synonymous sites of transgenes in mammals leads to increased protein levels but not through increased rates of translation[15-17]. Estimates of the proportion of human synonymous sites that are under selection vary from a few percent[18,19] to between 10% and 20%[20-22]. The largest database reports thousands of synonymous mutations 'associated' with ~1,500 diseases[23]. There is now evidence for selection on synonymous codon usage throughout the gene expression pathway, from transcription onwards, as we discuss here.

Rather than simply review this evidence for selection on synonymous mutations, we ask why such selection might be so apparently pervasive. More particularly, whereas the translational selection model has dominated discussion of selection on synonymous codon usage for the past half century, is there an alternative framework that can explain this selection in humans? Here, we present the unwanted transcript hypothesis.

We argue that with low N_e and the resulting high proportion of 'junk' DNA in the human genome (**Box 1**), the main problem for gene expression in humans is the presence of unwanted RNA transcripts. Unwanted transcripts are defined as those for which their removal is advantageous, even if the process of their creation is itself beneficial (for example, via transcriptionally mediated modification of chromatin[24]). We propose that much of the selection on synonymous sites curtails the creation of unwanted transcripts (through effects on transcription and splicing) or traps them to prevent translation (**Fig. 1**). These mechanisms of selection use generic patterns, often nucleotide skews, present in non-functional or foreign RNA sequences that discriminate them from functional native sequences. Understanding this

form of selection, and the resulting trends in synonymous codon usage, is important for diagnostics and for the design of transgenes.

[H1] The unwanted transcript hypothesis

We start by laying out what we consider to be the problem of unwanted transcripts, arguing that they will commonly be costly and thus provide the context for their suppression. We then discuss possible solutions that might be broadly classified as either reduction in the rate of their creation or the setting of traps. Both processes are affected by codon identity and thus can be impacted by synonymous mutations.

[H2] The problem

There are multiple sources of unwanted transcripts. Spurious transcripts **[G]** are likely to be common in humans as eukaryotic transcription factor (TF) binding sites are small, degenerate and thus common[25]. Of the 15.6 million TF binding sites in the human genome, fewer than 13% have sequence conservation[26]. Thus, novel sequences could be a rich source of transcripts. Two recent preprints report that naïve (i.e. not native) sequences introduced into yeast (with mostly open chromatin) result in large amounts of spurious transcripts, many of which are rare[27,28]. Analysis of random sequences of 120 nucleotides integrated into yeast reveals that half have promoter activity, and only 1–5% of intergenic transcripts in yeast cannot be attributed to similar spurious transcriptional activity[29].

Some spurious transcription may be unavoidable. Most mRNA promoters are inherently bidirectional, with transcription upstream of transcription start sites commonly generating non-functional transcripts[30] known as promoter upstream transcripts (PROMPTs) that are degraded, for example by the nuclear RNA exosome complex **[G]**. Other types of potentially non-functional transcripts include long noncoding RNAs (lncRNAs), enhancer RNAs and natural antisense transcripts. Although some have documented functions, many (especially the rare transcripts[31,32]) are poorly conserved in sequence[31,33] and presence[34], and it seems likely that many of them are functionless[35]. Indeed, many are destroyed so soon after creation that they are defined as cryptic transcripts[36], visible only after the degradation machinery is removed. Similarly, lncRNA transcription commonly terminates prematurely, co-transcriptional splicing is less efficient than for protein-coding transcripts and lncRNA transcripts tend to be unstable with expression levels 10-fold lower than for mRNAs (reviewed in ref.[35]). Functional studies seem to reinforce the conclusion that few lncRNAs have evident function[37,35]. lncRNA transcripts that are rare and rapidly degraded are less likely to have any RNA sequence-dependent functionality[38].

Much transcription is either parasitic or associated with remnants of parasitic sequences (such as viruses and transposable elements **[G]**) as their successful genomic colonization requires them to have TF binding sites[39]. Even though only about 1% of transposable elements in the human genome can still transpose, they are likely to be a rich source of unwanted transcripts. Indeed, primate-specific TF binding sites tend to be unconserved, found within transposable elements and not identified in genome-wide association studies and hence with no evident phenotypic effects [26]. More than 85% of primate-specific TF binding sites, and more than 20% of all TF binding sites, are derived from transposable elements[26]. Although some transposable elements have become domesticated to serve novel functions in the host cell (for example[40,41]), many transposable element-derived transcripts are likely to be spurious. Thus, humans may have particularly high levels of spurious transcription, as more than half of human DNA is derived from transposable elements[11] (**Box 1**). These processes likely explain why nearly all the DNA of the human genome is transcribed[42] and why 83% of lncRNAs have exonised transposable elements[43].

Immature functional transcripts can also pose a problem in humans as splicing is error prone[44,45]. This can be owing to mutation or to the intrinsic nature of the process, with human transcripts often having weak splice sites[46]. Experimental reports of the proportion of exonic point mutations that disrupt splicing vary from 20% to nearly 100% (reviewed in ref.[47]). Just as small N_e is associated with weak translational selection, a recent preprint reports that it is associated with increased transcriptome diversity[45]. Alternative splicing is increased when N_e is small and, importantly, this is largely owing to low-abundance splice forms, many of which are frameshifted and not associated with selection to preserve the splice site[45]. The increase in splice forms when N_e is small is thus more likely owing to weaker selection on splicing[45] (**Box 1**) rather than reflecting functional splice forms associated with increasing organismal complexity.

[H2] Is expression of unwanted transcripts costly?

A key assumption of the unwanted transcript hypothesis is that such transcripts are costly and hence that there is selection against them. Transcription and translation of unwanted RNAs are expected to be costly because both, especially translation[48], are energetically demanding[48]. In addition, just as the translation of functionally irrelevant reporter genes reduces cellular fitness[49,50] by engaging ribosomes that could otherwise process required transcripts, so too we expect the translation of unwanted RNAs to be deleterious. Depending on the nature of the unwanted transcripts, there may be additional costs associated with their expression, such as direct toxicity[51], interference with RNA metabolism or eliciting of an

immune response[52] (e.g. via inappropriate triggering of viral sensing RIG-1-like receptors leading to secretion of type 1 interferons[53]).

One line of evidence that unwanted transcription can be costly comes from the association between mutations in pathways responsible for the removal of such transcripts (see also below) and human genetic diseases[53-55]. For example, mutations in *RBM7*, part of the nuclear exosome targeting (NEXT) complex that directs RNA to the nuclear exosome complex for degradation, result in spinal motor neuropathy[56]. Mutations in cyclin-dependent kinase 13 (CDK13), which modulate RNA stability, are associated with melanoma[57]. This reflects the role of CDK13 in phosphorylation of ZC3H14, which in turn is required for nuclear RNA degradation. Failure to activate nuclear RNA surveillance owing to CDK13 mutation results in aberrant prematurely terminated transcripts (ptRNAs) that are translated and, in turn, promote melanoma progression[57]. Mutations in genes encoding nuclear RNA surveillance components are seen in many malignancies[57]; a curiosity in this case is that the production and translation of ptRNAs increases cellular growth rates.

For transpositionally competent transposable elements, additional costs will be associated with damage to host DNA through insertion and double-stand breaks that are mutagenic, can cause break-induced genome instability and are toxic[52]. Furthermore, transposable elements can interfere with host mRNAs through multiple mechanisms and the accumulation of transposable element transcripts can trigger immune responses (reviewed in ref.[52]). Unsurprisingly, there is antagonistic coevolution between transposable elements and mechanisms to suppress them (for example, see refs[58,59]), which is itself indicative of costs associated with their unwanted activity. Transposable elements are silenced by multiple mechanisms, notably DNA and histone methylation. Particular to mammals are nuclear paraspeckles, ribonucleoprotein bodies that function as traps for double-stranded RNA (typical of viruses) and transcripts with inverted repeat Alu elements **[G]** in the 3′ untranslated region (UTR), thereby preventing their nuclear export[60].

The importance of these mechanisms to suppress transposable elements is suggested by pathologies that are associated with the impairment of a specific repression mechanism (for example, see refs[61-65]) and by the increased transcription of transposable elements in autoimmune diseases, cancer and neurodegenerative pathologies[61,66-69]. Typically, these pathological conditions worsen with age or manifest with aging (such as Alzheimer disease, age-related macular degeneration and tauopathies), and aging itself is associated with increased transposable element transcription and activity (for example of the SIRT6–LINE1 **[G]** (L1) silencer[70]).

However, it is often hard to be certain that some of the apparent costs of unwanted transcripts are direct costs rather than an incidental correlate. For example, transposable element upregulation may be an irrelevant consequence of weaker chromatin control or there may be unrelated pleiotropic effects of mutations in genes associated with RNA metabolism and surveillance, including a dearth of functional transcripts that are regulated by the same quality control mechanisms. In some cases, the evidence for causality of the unwanted transcripts is robust. For example, introduction of either mutant CDK13 or phosphorylation-modifying mutations of ZC3H14 promote tumour progression and both are associated with the generation of unusual transcripts[57].

In the case of splicing disruption, the cause–effect relationship can be clearer as there can be mutations not only in the RNA-binding proteins (such as serine–arginine-rich (SR) proteins) that coordinate splicing (which are long recognised as being associated with disease[71]) but also in the target transcripts. For example, mutations in exonic splice enhancers **[G]** (ESEs) in *BRCA1* cause mis-splicing and disease[72]. Recent evidence from analysis of the site frequency spectrum suggests that there is strong purifying selection in humans against synonymous mutations that disrupt ESEs[20].

Several other points of evidence would be hard to understand were there not a cost to unwanted transcripts. For example, the increase in apparently functionless splice forms in species with small N_e is most parsimoniously explained by inefficient selection against weakly deleterious splice forms[45]. It is also notable that the excess transcripts seen in species with small N_e are predominantly rare transcripts[45]. This can be explained by a balance between selection against mis-splicing and the costs associated with mis-spliced transcripts, which we assume to be lower for rarer transcripts. The fact that noise in gene expression is higher for the rarer transcripts[45] is also consistent with a failure of selection to operate on weakly deleterious, rare mis-spliced transcripts. Similarly, it is notable that in yeast, nearly all bases of naive sequence are transcribed, generating multiple rare transcripts, whereas native sequence seems to have much ‘cleaner’ transcription, with rare transcripts largely suppressed and functional transcripts at high abundance[27,28]. This would be the expected outcome of selection against costly spurious transcription.

[H2] The solutions

Assuming that unwanted transcripts are on average costly, there are two modes by which organisms can address this problem. One is to curtail their creation, which would involve either suppression of transcription or improved accuracy of transcript processing, notably splicing.

Gene body methylation in humans is, for example, necessary to reduce rates of spurious intragene transcriptional initiation from cryptic promoters[73]. The other is to have a quality control system to filter transcripts. Quality control could involve either active degradation (for example, by RNAses[74]) or physical isolation (for example, preventing nuclear export).

Humans have multiple such quality control filters throughout the gene expression pathway (**Fig. 1**). For example, ptRNAs and upstream antisense RNAs are targeted for nuclear degradation by the polyA exosome targeting (PAXT) complex[75,76], containing MTR4 and ZFC3H1. MTR4 is also involved in the NEXT complex, which targets early and unprocessed RNAs for destruction[77]. The physical separation of translation (in the cytoplasm) and transcription (in the nucleus) underpins much translational suppression; for example, retention in the nucleus of ADAR-edited pairs of inverted repeat Alu elements[78]. Even if a transcript can escape the nucleus, there are further traps (such as cytoplasmic processing bodies **[G]**) and mechanisms of translational[79] or post-translational suppression, including nonsense-mediated decay **[G]** (NMD), no-go mRNA decay **[G]**, codon optimality-mediated RNA decay **[G]** and non-stop RNA decay **[G]**. Cytoplasmic stress granules are considered a triage centre[80] for differentiating wanted from unwanted transcripts. Multiple systems selectively target non-native transcripts for destruction (such as DICER[81], zinc finger antiviral protein **[G]** (ZAP)[82] and RNases[74]). Even after translation has occurred, there are additional error checks such as degradation of proteins with a hydrophobic carboxyl terminus[83] as these are more likely to be products of spurious transcription.

N^6-methyladenosine (m^{6}A) deposition on RNA has a complex involvement in the quality control process. It is associated with retrotransposon suppression[84], accumulation in stress granules of potentially foreign, long-exon transcripts[85] and nuclear retention and decay when on introns of transcripts with unprocessed 5′ splice sites[86]. Conversely, small exons typical of native genes are protected from m^{6}A deposition by the exon-junction complex **[G]**[87]. Although this implicates m^{6}A deposition in RNA suppression, its effects are sensitive to the 'reader' proteins with which it is associated (reviewed in ref.[88]). It can, for example, also enable nuclear export of endogenous transcripts with few or large exons[17] and release from transcriptional suppression of a domesticated endogenous retrovirus[89].

Our hypothesis is that much of the selection on synonymous sites reflects selection for reduced production or improved quality control of unwanted transcripts. The core problem of any such system is how to differentiate wanted from unwanted transcripts. This differentiation must depend on the recognition of 'fingerprints' that would not be expected to occur in a random sequence or in a newly arrived sequence (for example, a new retrovirus). Similar logic

is recognized in the context of innate immunity in which conserved pathogen-associated molecular patterns are recognised by pattern-recognition receptors such as Toll-like receptors (TLRs) to trigger an immune response[90]. In this language, we propose that humans have a similar set of pattern-recognition receptors for unwanted transcripts. At the level of retroviruses, innate immunity and selection against unwanted transcripts combine, with the HUSH complex involved in transcriptional silencing being considered a pattern-recognition receptor (**Box 2**).

As introns are especially abundant in mammals, and vertebrates generally[91], their presence (and positioning) is indicative of a native transcript. Conversely, longer exons are rare in humans (80% of human exons are less than 200bp[92]) and potentially indicative of foreign transcripts. Intron presence is indeed known to inform quality control. For example, the human NMD system uses location relative to an exon-junction complex to infer prematurity[93], and introns have long been recognised as important for human transgene expression[94]. The most 'suspicious' transcripts are therefore transcripts with long exons and intronless transcripts. Also, humans cannot rely on intron-mediated NMD for quality control in intronless transcripts[95]. However, as some intronless genes are derived from retrotransposition, and thereby inherit the exon splice motifs of the parental multi-exon gene, SR protein binding to such motifs has splicing-independent activity[96] that could be used for quality control (for example, nuclear export related).

Importantly in the context of selection on codon usage, we suggest that nucleotide content provides a further useful guide to wanted versus unwanted transcripts. As the rate of GC to AT mutation is higher than the converse (known as a mutation bias)[97], a sequence that is selectively neutral will evolve low-GC content[97]. Human AT-rich gene deserts are close to mutational equilibrium **[G]**[98] and consequently the dinucleotide content of the genome as a whole is skewed towards A and T (**Fig. 2**). High-GC content can thus be an indication of functional constraint, because proteins will commonly need amino acids whose underlying codons are G/C-rich, there will be a dearth of A/T-rich stop codons, and A/T-rich codons tend to specify amino acids that are more metabolically expensive to manufacture (i.e. in terms of ATP consumption) and are thus counter-selected[99]. In addition to mutation bias, A enrichment of single-exon transcripts reflects the reliance of retroviruses on structurally poor A-rich RNA necessary for cDNA generation by reverse transcription[100]. We thus suggest that GC richness at non-synonymous sites naturally marks native transcripts as opposed to spurious transcripts from AT-rich intergenic sequences, and that synonymous site usage has evolved to reinforce the high-GC content. In particular, third sites of a codon are the freest to evolve without consequence, so are a preferable solution for indicating that a protein coding gene is native.

As would be assumed on this basis, human coding sequences are most enriched, compared with genomic mean dinucleotide content, for G/C dinucleotides and most lacking in A/T dinucleotides (**Fig. 2**).

The idea that high-GC content is evidence of functionality is complicated by the fact that GC-biased gene conversion **[G]** (gBGC) acts in the opposite direction to the mutation bias[101]. Thus, sequence that is GC-rich, especially at third codon sites, could be subject to gBGC rather than selection[101]. However, this observation makes sense if humans use GC content as a signal of nativity: one reason that gBGC is so biased in humans — but not yeast[102] — could be that it preserves beneficial high-GC content in the face of weak selection in mammals to do the same. As gBGC in humans acts predominantly on gene-rich domains[101], native genes acquire high-GC content at third codon sites (GC3) whereas spurious transcripts from the larger AT-rich part of the genome do not[101]. Additionally, with paralogs often in clustered arrays inter-locus recombination and associated gBGC can boost their GC3 content, as for example witnessed with histone genes[103]. It is unclear whether the high-GC3 content of human exons is entirely owing to gBGC as most exons have GC3 content higher than that of the flanking introns[104]. Although this can be explained in some part as being owing to AT-rich transposable elements being more abundant in introns than in coding sequence[105], a difference between GC3 content in exons and introns remains after transposable elements are masked[105].

In our proposal then, the evolution of suppression mechanisms against A/U rich transcripts, favouring G/C rich ones, both forces selection on synonymous mutations and acts on the evolution of the direction of gBGC's bias. This bias along with selection, given the GC preference, then act to force synonymous site GC content above that expected at mutation-selection equilibrium, this being highly AT biased. More generally, we expect that all quality control mechanisms will co-evolve with native genes synonymous site nucleotide content.

The exception to high-GC content indicating native sequence **[Au:OK?]** is CpG, as methylated CpG is hypermutable. Indeed, although CpG is more common in coding sequences than in the genome as a whole (**Fig. 2**), CpG is the only C/G dinucleotide below the frequency that would be expected given mononucleotide occurrence, and it has low frequency in coding sequences, with only TpA being less frequent (**Fig. 2**). This same hypermutability forces increased levels of TpG and CpA, which causes, for statistical reasons, under-representation of both CpG and TpA[106]. At the transcript level, high-CpG or high-UpA content is therefore indicative of foreign transcripts. As viral sequences also need to encode amino acids, their sequences could have high-G/C content but also high CpG (although some viruses, such as SARS-CoV-2, have such extreme mutation bias favouring U that they are GC poor despite

selection operating against the mutation bias[107]). For similar reasons, CpG should be used as a suppressor of transcription (for example of transposable elements) mediated by methylation.

In sum, we expect that GC-rich and CpG-poor (including at synonymous sites) transcripts will pass pre-translation quality control tests, whereas high CpG, high U/T and high-A content, including high UpA, will lead to activity that prevents eventual translation. Similarly, intron position and presence should be used as a guide to nativity. Intronless transcripts and those with especially large exons should thus be regarded with suspicion. GC-rich multi- and small-exon genes are top candidates for nativity, whereas AT-rich single- or long-exon transcripts are most likely to be spurious. Functional viral sequences are likely to be caught by small-exon, multi-exon, low-CpG, high-GC3 traps. In addition, we expect selection against synonymous mutations to occur if these generate spurious splice forms of native transcripts.

[H1] What can the hypothesis explain?

Here, we note that the unwanted transcript hypothesis can explain many previously unconnected observations, with nucleotide content being a guide to transcript nativity such that high-GC content promotes gene expression, whereas high-CpG and high-AT content suppress expression.

[H2] Suppression of transcription

In contrast to the prediction of the translational selection model, codon usage bias does affect transcription rates. The first such evidence was derived from *Neurospora crassa*[108], in which codon usage bias predicted both protein levels and intra-nuclear RNA levels. Similar results have been replicated in numerous species, including humans[109], with various mechanisms involved including histone 3 lysine 9 trimethylation (H3K9me3) and transcription factor binding[110]. Most pertinent, however, seems to be the effects of 5′ GC content on RNA polymerase II (Pol II) processivity. After transcription initiation, processive elongation by Pol II is necessary for completion of transcription rather than premature transcript termination that generates micro 5′ transcripts. In mammals, two features promote Pol II processivity: GC-rich 5′ codons[109,111] and a prospective splice site early in the transcript[111]. We presume that removing micro transcripts is preferable to the consequences of potentially extending transcription of non-functional sequences.

Conversely, AT-rich genes are bound by the zinc-finger protein SALL4, which prevents premature expression of such genes[112]. Intronless transcripts are considered to be most ‘suspicious’ and it is notable that humans have a specialist system (the HUSH complex) to mediate silencing of long intronless transcripts (or those with unusually long exons) that are

especially A-rich[100] (**Box 2**). Less than 0.4% of native exons, in about 3.5% of genes, have high enough A content and are long enough (>1500bp) to be affected by HUSH (**Fig. 3**). About a quarter of the affected genes encode zinc finger proteins, which suggests that HUSH and zinc finger proteins might be alternative suppression systems, with the latter inactivated by the former.

As expected, given that CpG is rare in native human coding sequences, the default state in the non-coding genome is transcriptional inactivation via CpG methylation[113,114], which in part is an anti-transposable element strategy.

[H2] Prevention of mis-splicing

The best studied mode of selection on synonymous sites in humans is to enable accurate splicing. Humans have strong selection on synonymous mutations to preserve ESEs[20,115,116] near exon ends and such ESE disruption is associated with disease[117]. Similarly, selection on synonymous mutations reflects selection against ESEs in unwanted (core) exonic locations[118]. Unusual codon usage trends towards the ends of human exons can then be explained by selection for synonymous sites that promote ESE incorporation[19,119,120]. If codon usage selection at exon ends needs to be balanced between accurate splicing and faster translation (in species with translational selection), then accurate splicing dominates[121]. The avoidance of cryptic splice sites **[G]** in proximity to exon junctions is similarly advantageous[122] and synonymous mutations that generate them are deleterious as they generate unwanted transcripts[123].

[H2] Quality control

Broadly, in supposing that high-GC3 content enables quality control filters to be passed, this model can explain why high-GC content at synonymous sites enables the expression of intronless genes[15,16]. It also resolves the enigma[124] of why non-recombining (hence not subject to gBGC) intronless *Sry* has an unusually high synonymous GC3 content. Additionally, it can explain why native intronless retrogenes **[G]** have higher GC3 content than their parental (intron-containing) genes[16]. It is unknown to what extent a net flux of retrogenes[125] from the AT-rich X chromosome to the GC-richer autosomes might explain the same trend.

More generally, for a gene to be expressed in humans it needs to have an intron near the 5′ end or be intronless but GC-rich[16]. Inspired by the finding that retrogenes have high-GC3 content, one study constructed many intronless variants encoding green fluorescent protein that differed in GC3 content and found that the higher the GC3 content, the higher the level of the protein expressed[16]. Both intronic and intronless human genes tend to be GC-rich at their

5′ ends[126]. This contrasts with bacteria, in which the 5′ ends of sequences are AT rich, which leads to low RNA stability and thus easier translation initiation[49,127]. Although low 5′ RNA stability is a common feature across much of the evolutionary tree, it does not force high AT content in mammals[128].

The increased protein expression of GC-rich transcripts described in GFP transgenes[16] likely involved both quality control measures and transcriptional effects. The most obvious location for any form of 'passport' quality control of mRNA is nuclear export. Indeed, this study identified that, at least in part, the GC-mediated expression effect is owing to increased relative cytoplasmic presence of GC3-rich transcripts. This suggests a role for either nuclear export, consistent with prior evidence[129], or some other intra-nuclear filter in restricting cytoplasmic access of GC-poor transcripts.

Further studies subsequently identified the nuclear export pathway as being core to such observations[17]. Of the two canonical nuclear RNA export pathways, NXF1 was required for the export of single-exon transcripts, transcripts with long exons and A/U-rich, multi-exon transcripts[17], whereas the TREX complex facilitated export of spliced and G/C-rich transcripts. It is, we suggest, no coincidence that the three transcript characteristics associated with NXF1-mediated nuclear export are the three that we propose to be indicative of potentially unwanted transcripts. Similar mechanisms seem to apply to m^6A : on native transcripts it promotes NXF1-mediated export of few-exon or large-exon transcripts[17] but it is absent from small-exon mRNAs (commonly exported via TREX) owing to the exon-junction complex[87]. Intriguingly, splice factor binding (in other words, SR protein binding) is conserved in intron-bearing and intronless transcripts alike, as revealed by synonymous site conservation in ESEs[96]. Synonymous sites in ESE motifs are preserved in intronless genes as binding of SR proteins to such motifs enables nucleocytoplasmic export via NFX1 (refs[130,131]). That the HUSH complex identifies A-rich transcripts is of particular note as nuclear export controls also restrict transcripts with A-rich 5′ ends[129] which suggests that multiple quality control filters have similar effects.

Nuclear export control seems to be a key part of the multi-layered quality control mechanisms and, more generally, retention and triage of unwanted transcripts may be relevant strategies. Cytoplasmic processing bodies contain mostly AU-rich mRNAs, and as a result, AU-rich and GC-rich mRNAs are subject to distinct translational control and decay pathways[132]. AU-rich transcripts are also targeted in the cytoplasm by AU binding proteins that target transcripts to the exosome complex[133]. Similarly, stress granules inhibit viruses and transposable elements, functioning as a triage[80] for retention of AU-rich transcripts[134,135]. Stress granules also have a

preference for transcripts that are rare[135], have poor translatability[135] and have long m^{6}A-tagged exons[85], which further suggests that their function is part of a quality control mechanism.

Dinucleotide content (or longer *k*-mers) is also central to RNA degradation in the manner predicted by the unwanted transcript hypothesis. CpG-rich transcripts tend to be directly and preferentially bound by ZAP to promote their decay[136]. Similarly, UpA dinucleotides, which are rare in native transcripts, are targets of ribonuclease L (RNase L)[137]. Likewise, it makes sense that the NEXT complex[77] targets unprocessed U-rich RNAs[138], particularly PROMPTs and intron-terminating sequences[139].

Although translational selection is relatively unimportant in mammals, it may also be part of a quality control mechanism. Codon optimality modulates ribosome velocity and hence protein folding in yeast[140]. Co-translational folding in mammals may also be modulated by ribosome velocity, with phenotypic effects owing to synonymous mutations[141]. Thus, even if a transcript passes prior quality control filters and makes it to the ribosome, the resulting protein may end up misfolded and hence recycled. Furthermore, whereas it was commonly assumed that codon optimality effects are mediated by ribosomal velocity or misincorporation rates, in yeast, codon optimality is monitored by Dhh1p and Ccr4-Not, leading to degradation if the translation process is too slow[142-144]. This may also be a quality control mechanism in that any transcript being processed too slowly is considered to be 'suspicious' and hence translation is curtailed. Codon composition is thought to affect RNA stability in mammals, possibly by similar routes[145,146].

Can we be confident that these quality control filters are not simply means to regulate genes? One line of evidence comes from genes that can co-evolve with these filters. If such filters exist to counter unwanted transcripts, co-evolving genes would be under selection to counter-adapt to these constraints. Consistent with this, CpG avoidance is commonly observed as an adaptation to human hosts of SARS-CoV-2[147] and other viruses[148]. As would be expected if nuclear export is a key quality control filter, nuclear RNA viruses have higher GC content than cytoplasmic RNA viruses[126].

A further line of evidence comes from examining the filters that restrict the import of nucleic acids into the cell via endosomes. Such filters mirror those for the rejection of non-native transcripts. Nuclear paraspeckles and cytoplasmic DICER and TLR3 capture long double-stranded RNAs, nuclear export and cytoplasmic transcription both select against AT-rich transcripts, as does cytoplasmic RIG-I (activated by polyU/UC), and ZAP and endosomal

TLR9 both filter out CpG-rich RNAs. These parallels we suggest are there because both sets of processes are there to filter out non-native sequences, such as viruses

[H1] What can the hypothesis not explain?

There are features with which the unwanted transcript hypothesis does not seem to be compatible. Although we predict that high-GC and low-CpG content function as signs of native genes and, consistent with this, human promoters tend to have high-GC content[149], they can also have more CpG than GpC in the form of CpG islands **[G]**. Human cells recognise these CpG-rich sequences as native and do not methylate them. Consequently, CpG methylation can be used as a suppressor of transcription (e.g. of transposable elements[150], or of intragene cryptic promoters[73]) and a high density of unmethylated CpG can be employed as an activator of transcription (for example, as a binding site for the transcriptional activator CFP1 (ref.[151])). Unmethylated CpG is not hyper-mutagenic, so its preservation does not contradict our hypothesis. However, it does seem contradictory that mammals have evolved systems to suppress transcripts with high CpG content (for example, via ZAP) but have also evolved CpG islands as indicators of transcript nativity. Why do mammals not methylate all occurrences of CpG? Likewise, other effects of CpG on transcription are not as expected. For example, in some genic contexts, CpG depletion in coding sequences results in reduced protein expression owing to nucleosome positioning-mediated transcriptional effects[152].

Whereas the unwanted transcript hypothesis can explain why active and successful human transposable elements (Alu elements, short interspersed nuclear elements (SINEs) and variable number of tandem repeat (VNTR) elements) tend to be GC rich, it is not clear how intronless GC-poor L1, the only autonomously transposing transposable element, evades all quality control filters. Indeed, even with the large number of host mechanisms that repress L1 at transcriptional, post-transcriptional and post-translational levels, *de novo* L1 insertions still take place within somatic tissues[153]. Part of the explanation is that open reading frame 1 (ORF1) of L1 is about 1kb less than the minimum length targeted by the HUSH complex[100]. The longer A-rich ORF2 of L1 is usually suppressed, but as transcription is needed for this suppression[100] some leakage is perhaps inevitable.

Further explanations for the apparent exceptions to the predictions of the unwanted transcript hypothesis might be provided, but whether these are truly ‘special case’ scenarios or instead reflect a weakness of our model is unclear. One could argue that no quality control filter can be absolute and that leaky expression is always expected to some extent. Indeed, we suggest that the problem of why no filter should be perfect is a theoretically interesting one. Beyond the notion that, as with mutation rates, perfect suppression of errors may well be too costly, a

perfect filter against ancestrally intronless transcripts would, for example, have prevented the expression of histones and G-protein-coupled receptors.

Also, our model does not allow for the possibility that in early human embryogenesis, the various filters against transposable elements and unwanted transcripts are not yet fully in place[52,154]. Although transposable elements are tightly suppressed in the zygote, they are derepressed during early embryonic development and gametogenesis, when the genome is globally demethylated to allow for epigenetic reprogramming[52]. Recent evidence suggests that active retrotransposition at this stage is kept in check by the removal of highly affected cells rather than by transcript-level processes[154].

[H1] What do we not try to explain?

The unwanted transcript hypothesis has little to say about selection on synonymous sites in some contexts. Indeed, we do not suppose that all selection on synonymous sites in mammals is associated with control of unwanted transcripts. For example, human coding sequences are depleted of ultra-stable G-quadruplexes **[G]** with selection against synonymous mutations that would generate them[155]. This is likely a result of selection for unhindered transcript processing[155], although an association with the inclusion of transposable elements in stress granules (reviewed in ref.[156]) might suggest that it also has a quality control function.

Selection on synonymous sites for or against micro RNA (miRNA) pairing sites[157-159] may be a further case in point. Notwithstanding the possibility that miRNA was originally an anti-transposable element mechanism (and is used to control L1 (ref.[160])), it now functions also for the control of native gene expression. Importantly, miRNA-mediated regulation is relatively specific in the transcripts affected (a feature of miRNA–RNA pairing) and thus relatively ill-suited as a general-purpose quality control filter. Indeed, it is striking that many of the human quality control filters — HUSH being an exemplar (**Box 2**) — use general 'fingerprints' rather than specific motifs for transcript recognition. The one example that uses specific motifs, the preservation of ESEs, is an example of selection to prevent the creation of spurious forms not the recognition of such transcripts.

[H1] Testing the hypothesis

How might we test the unwanted transcript hypothesis? Here we start by considering necessary but not sufficient tests. These include possible tests of the assumptions of the hypothesis, notably that such transcripts are costly, and transgene experiments that seek to further define the nucleotide biases that lead to successful expression (and why). We then consider tests of the effects of synonymous mutations. Here we consider both mechanism

554 derived predictions of the synonymous sites that will be under selection and predictions of
555 what happens when synonymous sites are mutated. Finally, we address comparative genomic
556 predictions.
557 ***[H2] Testing the cost assumption***
558 Although there is robust evidence that unwanted transcripts pose fitness costs (see above),
559 experimental evolution provides a further means to test this assumption. Naive sequence
560 introduced into yeast produces a large number of presumably functionless transcripts[27,28]. If
561 our model is right, we would expect that if we select over time for increasing cell fitness coupled
562 with retention of the naive sequence, there will be a reduction in the rate of production of
563 spurious transcripts, particularly those that were originally translated. In addition, given that
564 mutations in the quality control surveillance systems are associated with disease in humans
565 (see above), we expect that comparable mutations in cell lines that disable some quality
566 control filters should be both costly (reduced cell fitness) and lead to compensatory evolution.
567 Predicting what such compensation might involve in mechanistic terms is difficult but a
568 reduction in unwanted protein is a clearer prediction. These predictions come with the caveat
569 that we do not expect that accumulation of the translated products of ptRNAs should lead to
570 increased cell proliferation, although this is seen in humans leading to cancer[57].
571
572 ***[H2] Testing the effects of synonymous mutations***
573 The above tests would not address whether synonymous codon usage affects fitness in a
574 predictable manner. In this respect, transgene codon randomization experiments to modulate
575 ESE usage and SR protein binding, as well as GC content, would be valuable. We predict that
576 codon usage that promotes binding in intronless transcripts of those SR proteins associated
577 with nuclear export will promote higher levels of protein expression. We also predict that
578 modifications of synonymous sites in mammals will mediate their effects in a manner that is
579 predicated on the mutation bias. Changing either A for T (or vice versa), or G for C (or vice
580 versa), thus having no effect on GC3 content, should have relatively little effect on expression
581 levels of a short protein from an intronless transcript (<1.5kb), allowing for alterations in UpA
582 and CpG content. By contrast, changing AT to GC, which increases GC3 content, should
583 increase protein expression, whereas the reverse should typically decrease protein levels. We
584 predict that such changes would be mediated by increased filtering (of AT-rich transcripts) or
585 decreased filtering (of GC-rich transcripts) by quality control mechanisms (rather than being
586 exclusively ribosome mediated), with control of nuclear export being the main such
587 mechanism. Furthermore, for intronless transcripts >1.5kb and so within the scope of HUSH-
588 mediated silencing, the A-to-T or T-to-A effects will be different than for transcripts <1.5kb as
589 the HUSH complex specifically targets A-rich transcripts (**Box 2**).
590

Arginine codon usage is potentially especially informative. Our hypothesis is that there should be preference for G or C not just at third codon sites but more generally in native coding sequence. Arginine is specified by AGA, AGG or CGN. We predict that in human cells in which the anti-CpG systems such as ZAP are removed, transgenes with CGC or CGG codons should express more protein than those bearing AGA (all else being equal). AGG-, CGA- and CGT-bearing transgenes should have comparable and intermediate expression levels. Similar logic applies to leucine (specified by UUA, UUC or CUN), with predicted protein-boosting effects of the use of CUC and CUG codons compared with UUA. Serine usage may be less informative as both the two-fold degenerate (AG[UC]) and four-fold degenerate (UCN) blocks are GC/AT-matched at the first two sites, GC variation therefore being exclusive to codon third sites. The evidence also suggests a primacy to 5′ codon usage[16,111], such that these effects should be more pronounced when modification occurs in the first 20 codons of a transcript. Examination of 5′-to-3′ trends of arginine and leucine codon usage in native genes may also be instructive but comes with the caveat that CGN and UUA may attract anti-CpG or anti-UpA quality control measures, respectively.

[H2] Analysis of selection on synonymous mutations

Results from such transgene experiments will be necessary but not sufficient to test the unwanted transcript hypothesis as they do not address the fitness consequences of synonymous mutations in native genes. For predictions about native genes, there are two alternative but related means to investigate the relationship between selection on synonymous mutations and mechanisms of action. First, we can examine the effects of mutations at synonymous sites that are presumed to be under selection[18]. Second, we can predict the differential fitness effects of synonymous mutations given a particular mode of action. Fitness effects can be inferred by consideration of between-species conservation or by reference to within-species single nucleotide polymorphism (SNP) frequencies, as selection causes deleterious SNPs to have low frequencies. From the frequency spectrum of SNPs one can, in addition, derive a distribution of fitness effects[161,20], although care must be taken to exclude artefacts owing to demography. Both conservation-based[115] and SNP frequency-based[19,20,116,162] analyses report that synonymous sites in ESEs are under purifying selection. Both statistical approaches can be augmented by considering association with genetic diseases, either in well-worked single mutation analyses (**Supplementary Table 1**) or through genome-wide association studies. Many synonymous mutations associated with disease are known to cause deleterious novel splice forms (**Supplementary Table 1**).

Unfortunately, as there seem to be a large number of quality control filters that affect a range of processes from transcription to translation via nuclear export, there is no comparable,

single-mechanism prediction for the unwanted transcript hypothesis as simple as that for translational selection (that codons enriched in highly expressed genes match the most abundant tRNAs). Rather, the unwanted transcript hypothesis proposes that multiple factors affect the activity of synonymous mutations, broadly predicting that mutations that make an RNA look more like foreign RNA should be deleterious, as should those that generate spurious transcripts. Equally, the model predicts which foreign sequences should evade quality control filters, such as GC-rich transposable elements.

Some effects of synonymous mutations may result from the simple gain or loss of key nucleotides. For example, we predict that 5′ mutations in single-exon genes converting GC to AT would be deleterious, whereas the opposite would be advantageous. Generation of CpG and TpA should similarly be deleterious. There should also be a fitness difference on average between GC-to-AT mutations in single-exon genes and in multi-exon genes.

Effects of synonymous mutations may also manifest through changes to RNA stability and structure. In rodents, synonymous sites that bear fixed substitutions have different effects on *in silico*-predicted RNA stability than those that have no substitution[163]. We can extend this concept to more specifically predict that selection on synonymous mutations will sometimes be mediated by selection to avoid long-stem RNA structures, as these trigger various sensing mechanisms for viruses and transposable elements. It is of note that a synonymous mutation in the *COMT* gene that is associated with reduced protein titre markedly alters predicted RNA folding, giving rise to stable and long-stem structures[164].

In addition, although CpG and UpA rates are low in human genes, we expect that selection will function to physically 'hide' these dinucleotides in native transcripts, probably mediated by RNA secondary structures. Thus, we expect that there will mutations at synonymous conserved sites that 'unhide' these dinucleotides by altering such structures. This process may account for the observation that coding sequences subject to extensive selection on synonymous mutations are enriched for TpA[165]. If this reflects selection on synonymous sites because they modulate RNA structures[165], a lack of UpA exposure is predicted to be disrupted by mutation at the conserved synonymous sites. Differences in native RNA structures near AGA and AGG versus CGN arginine codons would be predicted, the latter being more likely to occur in stem structures than loop structures, assuming that anti-CpG mechanisms target loop-exposed CpG.

Other aspects of the profile of conserved synonymous sites are predicted by the unwanted transcript hypothesis, in line with the mechanisms that confer foreign versus native status to

transcripts. For example, with the same logic as outlined in the ESE-modifying transgene experiment, we also expect (perhaps counter-intuitively) selection in functional retrogenes to favour the retention of ESEs by conserving synonymous sites but only if these permit binding of nuclear export-promoting SR proteins (derived from the parental gene). Similarly, if m^6A can offer protection to what might otherwise look like a foreign transcript (for example, a single-exon transcript), synonymous mutations that disrupt m^6A deposition in native single-exon or few-exon transcripts near GC-rich motifs[17] will lead to reduced protein titre owing to disrupted NFX1-mediated nuclear export.

[H2] Comparative genomic predictions

In principle, the unwanted transcript hypothesis can make predictions about the need for quality control across taxa. Species with 'lithe' genomes (small intergenic distance, high ratio of coding sequence to genome size, few and small introns, few transposable elements) such as *Saccharomyces cerevisiae* should have fewer problems than humans with unwanted transcripts and so should need fewer quality control mechanisms or to invest less in them. However, the obvious problem is how to assay the number of or investment in quality control filters. It might be notable that the nuclear paraspeckles, which trap Alu elements and double-stranded RNA, are specific to mammals[60]. Also, higher ESE densities in species with larger and more common introns are indicative of greater difficulties in suppressing noisy splicing when N_e is low[166].

An alternative mode of testing is to examine the nature of the quality control mechanisms. For example, whereas ZAP attacks CpG and RNAse L attacks UpA, the two rarest dinucleotides in human coding sequence, we predict that in mammals there should be no common degradation system that targets the most common dinucleotides (GpG, CpC and GpC) prior to translation. It is intriguing that over evolutionary time, the preference of multiple RNase A family members has shifted from UpG (in fish) to UpA (in mammals)[167] consistent with mammalian pressure to target RNA rich in the dinucleotides least common in human CDS. The RNase preferences for A/U need not be the same in a species in which the mutation bias is AT to GC[97], for which high-GC content could occur owing to neutrality. Unfortunately, we are unaware of any species with both a small N_e and a mutation bias in the opposite direction to that seen in humans, which would be an ideal testing ground for the unwanted transcript hypothesis. However, we suggest that dependency of Pol II-mediated elongation on 5′ introns and high 5′ GC content will not be universal, that the split functions of the nuclear export channels might be taxonomically restricted to species with a high proportion of junk DNA in their genomes, and that any quality control functionality of m^6A would be particular to such genomes.

[H1] Discussion

Although rapidly growing microorganisms show selection on codons affecting the translation process (classical translational selection), in this Perspective article we put forward an alternative form of selection based on selection against unwanted transcripts that we suggest is more relevant for error-prone genomes with large amounts of junk DNA such as the human genome. We suggest that it is no coincidence that all of the quality control filters that we have observed function against transcripts rich in A, U, UpA and CpG, the nucleotides that are under-used in native coding sequence in humans.

An objection to this hypothesis could be that we have mistaken components of gene regulatory systems as components of filters for unwanted transcripts. We suggest that to suppose that any such system must be classified as either a mode of regulation or a filter is a false dichotomy. In evolutionary terms, it would be remarkable if a system that evolved to, for example, suppress a given transposable element, would not also sometimes be repurposed for spurious transcript control and for gene expression regulation (or vice versa). For example, NMD is used to suppress transposable elements, remove erroneous splice forms and regulate amino acid biosynthesis[168]. Paraspeckles sequester Alu elements and are used to control circadian gene expression[169]. Similarly, gene regulation occurs via modification of the length of the 3′ UTR, by removing or adding the Alu element[170]. Stress granules both sequester key cell division transcripts for future release and inhibit viruses and transposable elements[80]. In the context of translational selection, that optimal codons are scanned and non-optimal transcripts are degraded is often considered as evidence of a regulatory process controlling protein levels of native transcripts, but it could equally be a means to suppress transcripts without host codon usage patterns[79,171]. Evidence for the regulatory control of some transcripts by using what might otherwise be considered quality control filters doesn't then necessarily support the above objection. We emphasise the importance of not assuming that everything in a cell has an adaptive regulatory function.

Not only do we not presume that all selection on synonymous sites is mediated by filters or prevention systems, we also do not presume that all prevention and filtering is mediated by synonymous codon choice. Indeed, we expect there to be a wide range of systems to promote the expression of wanted transcripts and suppress others. For example, mRNA cap structures function as effective guides to wanted transcripts. For this reason, some viruses (such as influenza A virus) have evolved cap-snatching from host RNAs[172]. The polyA tail of mRNAs is similarly used for quality control[173]. What we do suggest is that to understand codon usage trends, it can be helpful to think in terms of keeping unwanted transcripts in check, rather than

739 focusing only on translational selection. Indeed, whereas codon optimization of synthetic
740 RNAs premised on the concept of tRNA-codon availability, is routinely carried out, the
741 assumptions of the method are not upheld in humans[174]. It is instead likely that any effects of
742 codon optimization are mediated by the described quality control filters[16].
743
744 Aside from the complex role of m^6A in quality control, we highlight two major unknowns
745 concerning our model. First, is the extent to which the model has validity across taxa. We have
746 highlighted relevance of the hypothesis to humans as we are interested in human diseases
747 and possible cures, and because humans are an exemplar of a species that we expect to have
748 problems with unwanted transcripts and for which translational selection has mostly not been
749 observed. This is because humans have low N_e and hence inefficient selection, which can
750 explain our large genome with a low ratio of coding sequence to genome size (**Box 1**), a
751 genome in which we expect poorer control of processes such as splicing and transcription.
752 Despite the fact that many unwanted transcripts, such as PROMPTs, may be universal,
753 generally, we expect that as N_e increases so the weak selection associated with translational
754 selection should become more important in determining codon usage and the unwanted
755 transcript problem should diminish, as selection over the control of gene expression and
756 splicing should increase. Evidence that rare spurious transcripts are less common when N_e is
757 large[45] is consistent with this, as is, for example, evidence that non-optimal stop codon usage
758 is predicted by small N_e[175] and that purifying selection acts on non-coding RNA in Drosophila
759 but not so obviously in humans[33].
760
761 One means to establish the phylogenetic extent of selection for synonymous sites to be
762 recognized as native may be through analysis of coding sequence 5′ ends. In many taxa,
763 coding sequence 5′ ends are under selection to generate unstable RNAs so as to mediate
764 more efficient translation initiation[128], thus explaining high 5′ AT content[176], including at third
765 codon sites. In mammals, this selection for low RNA stability is not evident[128]. We suggest that
766 this is because the counteracting selection for high GC content so as to appear as a native
767 transcript is dominant. If this logic is correct, cross-species analysis of 5′ GC3 content may
768 help to determine the taxonomic relevance of selection based on wanted versus unwanted
769 transcripts.
770
771 The second major unknown is how this model relates to the mechanisms by which
772 synonymous mutations might induce genetic diseases. Selection on most synonymous
773 mutations is expected to be weak; indeed, whereas altering the GC3 content of a human
774 transgene has major expression effects[15,16], the impact of changing just one synonymous site
775 must, one average, be more subtle. By contrast, individual splice-disrupting mutations have

776 the potential to have major fitness consequences. Estimates suggest that ~10–30% of all
777 disease mutations mediate their effects through disruption of splice motifs such as ESEs (see
778 refs[177-179] and references therein). Alternative mechanisms whereby a single synonymous
779 mutation has a large enough effect to cause disease are unclear; although examples can be
780 provided, their commonality is not known (**Supplementary Table 1**; **Fig. 4**). Although splicing
781 seems to be the dominant mode of disease-associated action, ascertainment bias is likely to
782 be profound.
783
784 The unwanted transcript hypothesis has implications for gene therapy, recombinant protein
785 production and the problem of how to make an intronless transcript. An obvious solution given
786 the above is to render at least the 5′ end codons GC3 rich[180]. CpG avoidance (or enrichment
787 to make attenuated live vaccines[181]) is a common strategy in multiple contexts[114,182]. It is
788 notable that similar themes apply to RNA vaccines. Indeed, one breakthrough in their
789 development was recognition that RNA base modifications suppress TLR activity[183]. The use
790 of pseudouridine in mRNA vaccines seems to enable RNA to be translated but simultaneously
791 evades the U-sensing filters[183]. Parallels can be made to m^6A deposition enabling the nuclear
792 export[17] of what might otherwise appear to be foreign transcripts (those with few long exons
793 or single exons). It is similarly notable that reduced CpG content and high-GC content both
794 enable nuclear transgene activity[16] and the efficacy of externally delivered mRNA[184]. That
795 TLRs and the control of nuclear transcripts have evolved similar rules to detect that which is
796 foreign explains why RNA vaccines and transgenes benefit from the same sort of modifications
797 at synonymous sites and why, more generally, their design principles are similar.
798
799 **References**

800 1 King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788-798, (1969).
801 2 Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. DNA sequence
802 evolution: the sounds of silence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **349**, 241-
803 247, (1995).
804 3 Ikemura, T. Correlation between the abundance of *Escherichia coli* transfer
805 RNAs and the occurrence of the respective codons in its protein genes: a
806 proposal for a synonymous codon choice that is optimal for the E. coli
807 translational system. *J. Mol. Biol.* **151**, 389-409, (1981).
808 4 Ikemura, T. Codon usage and tRNA content in unicellular and multicellular
809 organisms. *Mol. Biol. Evol.* **2**, 13-34, (1985).
810 5 Sharp, P. M. & Li, W.-H. The codon adaptation index-a measure of directional
811 synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*
812 **15**, 1281-1295, (1987).

6 Qian, W., Yang, J. R., Pearson, N. M., Maclean, C. & Zhang, J. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* **8**, e1002603, (2012).
7 Akashi, H. Synonymous codon usage in *Drosophila melanogaster:* Natural selection and translational accuracy. *Genetics* **136**, 927-935, (1994).
8 Stoletzki, N. & Eyre-Walker, A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol. Biol. Evol.* **24**, 374-381, (2007).
9 Sharp, P. M., Emery, L. R. & Zeng, K. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 1203-1212, (2010).
10 dos Reis, M. & Wernisch, L. Estimating translational selection in eukaryotic genomes. *Mol. Biol. Evol.* **26**, 451-461, (2009).
11 Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401-1404, (2003).
12 Duret, L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640-649, (2002).
13 Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E. & Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30**, 308-321, (2014).
14 Bali, V. & Bebok, Z. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *Int. J. Biochem. Cell Biol.* **64**, 58-74, (2015).
15 Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* **4**, e180, (2006).
16 Mordstein, C. *et al.* Codon usage and splicing jointly influence mRNA localization. *Cell Syst.* **10**, 351-362 e358, (2020).
17 Zuckerman, B., Ron, M., Mikl, M., Segal, E. & Ulitsky, I. Gene architecture and sequence composition underpin selective dependency of nuclear export of long RNAs on NXF1 and the TREX Complex. *Mol. Cell* **79**, 251-267 e256, (2020).
18 Lin, M. F. *et al.* Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res.* **21**, 1916-1928, (2011).
19 Caceres, E. F. & Hurst, L. D. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* **14**, R143, (2013).
20 Savisaar, R. & Hurst, L. D. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* **28**, 1442-1454, (2018).
21 Keightley, P. D. & Halligan, D. L. Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics* **188**, 931-940, (2011).
22 Eory, L., Halligan, D. L. & Keightley, P. D. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol. Biol. Evol.* **27**, 177-192, (2010).
23 Wen, P., Xiao, P. & Xia, J. dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics* **32**, 1914-1916, (2016).

24 Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96-118, (2021).
25 Hurst, L. D. Evolutionary genomics and the reach of selection. *J. Biol.* **8**, 12, (2009).
26 Andrews, G. *et al.* Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930, (2023).
27 Luthra, I. *et al.* Biochemical activity is the default DNA state in eukaryotes. Preprint at bioRxiv https://www.biorxiv.org/node/2906314.abstract (2022).
28 Camellato, B., Brosh, R., Maurano, M. T. & Boeke, J. D. Genomic analysis of a synthetic reversed sequence reveals default chromatin states in yeast and mammalian cells. Preprint at bioRxiv https://www.biorxiv.org/content/10.1101/2023.06.20.545713v2 (2022).
29 Xu, H., Li, C., Xu, C. & Zhang, J. Chance promoter activities illuminate the origins of eukaryotic intergenic transcriptions. *Nat Commun* **14**, 1826, (2023).
30 Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**, 7179-7193, (2011).
31 Schuler, A., Ghanbarian, A. T. & Hurst, L. D. 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, (2014).
32 Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A. & Koonin, E. V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol* **3**, 1390-1404, (2011).
33 Haerty, W. & Ponting, C. P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* **14**, R49, (2013).
34 Johnsson, P., Lipovich, L., Grandér, D. & Morris, K. V. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta* **1840**, 1063-1071, (2014).
35 Ponting, C. P. & Haerty, W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. *Annu Rev Genomics Hum Genet* **23**, 153-172, (2022).
36 Wyers, F. *et al.* Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**, 725-737, (2005).
37 Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, aah7111, (2017).
38 Schlackow, M. *et al.* Distinctive patterns of transcription and RNA processing for Human lincRNAs. *Mol. Cell* **65**, 25-38, (2017).
39 Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71-86, (2017).
40 Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405-409, (2014).

41 Raskó, T. *et al.* A novel gene controls a new structure: PiggyBac Transposable Element-Derived 1,unique to mammals, controls mammal-specific neuronal paraspeckles. *Molecular Biology and Evolution* **39**, (2022).

42 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, (2012).

43 Carlevaro-Fita, J. *et al.* Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res.* **29**, 208-222, (2019).

44 Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* **6**, e1001236, (2010).

45 Bénitière, F., Necsulea, A. & Duret, L. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. Preprint at bioRxiv https://www.biorxiv.org/content/10.1101/2022.12.09.519597v5 (2023).

46 Irimia, M. *et al.* Complex selection on 5' splice sites in intron-rich organisms. *Genome Res.* **19**, 2021-2027, (2009).

47 Savisaar, R. & Hurst, L. D. Estimating the prevalence of functional exonic splice regulatory information. *Hum. Genet.* **136**, 1059-1078, (2017).

48 Wagner, A. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**, 1365-1374, (2005).

49 Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255-258, (2009).

50 Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat. Biotechnol.* **36**, 1005-1015, (2018).

51 Mittal, P., Brindle, J., Stephen, J., Plotkin, J. B. & Kudla, G. Codon usage influences fitness through RNA toxicity. *Proc Natl Acad Sci U S A* **115**, 8639-8644, (2018).

52 Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 199, (2018).

53 Lionetti, M. *et al.* A compendium of DIS3 mutations and associated transcriptional signatures in plasma cell dyscrasias. *Oncotarget* **6**, (2015).

54 Fasken, M. B. *et al.* The RNA exosome and human disease. *Methods Mol Biol* **2062**, 3-33, (2020).

55 Morton, D. J. *et al.* The RNA exosome and RNA exosome-linked disease. *RNA* **24**, 127-142, (2018).

56 Giunta, M. *et al.* Altered RNA metabolism due to a homozygous RBM7 mutation in a patient with spinal motor neuropathy. *Hum. Mol. Genet.* **25**, 2985-2996, (2016).

57 Insco, M. L. *et al.* Oncogenic CDK13 mutations impede nuclear RNA surveillance. *Science* **380**, eabn7625, (2023).

58 Luo, S. *et al.* The evolutionary arms race between transposable elements and piRNAs in Drosophila melanogaster. *BMC Evol. Biol.* **20**, 14, (2020).

59 Bertozzi, T. M., Elmer, J. L., Macfarlan, T. S. & Ferguson-Smith, A. C. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proc Natl Acad Sci U S A* **117**, 31290-31300, (2020).
60 Fox, A. H. & Lamond, A. I. Paraspeckles. *Cold Spring Harb. Perspect. Biol.* **2**, a000687, (2010).
61 Kaneko, H. *et al.* DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. *Nature* **471**, 325-330, (2011).
62 Muotri, A. R. *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443-446, (2010).
63 Tsivion-Visbord, H. *et al.* Increased RNA editing in maternal immune activation model of neurodevelopmental disease. *Nat Commun* **11**, 5236, (2020).
64 Ansell, B. R. E. *et al.* A survey of RNA editing at single-cell resolution links interneurons to schizophrenia and autism. *RNA* **27**, 1482-1496, (2021).
65 Li, P. *et al.* Aicardi-Goutieres syndrome protein TREX1 suppresses L1 and maintains genome integrity through exonuclease-independent ORF1p depletion. *Nucleic Acids Res.* **45**, 4619-4631, (2017).
66 Stearrett, N. *et al.* Expression of Human endogenous retroviruses in systemic lupus erythematosus: multiomic integration with gene expression. *Front. Immunol.* **12**, 661437, (2021).
67 Dembny, P. *et al.* Human endogenous retrovirus HERV-K(HML-2) RNA causes neurodegeneration through Toll-like receptors. *JCI Insight* **5**, (2020).
68 Ramirez, P. *et al.* Pathogenic tau accelerates aging-associated activation of transposable elements in the mouse central nervous system. *Prog. Neurobiol.* **208**, 102181, (2022).
69 Grundy, E. E., Diab, N. & Chiappinelli, K. B. Transposable element regulation and expression in cancer. *FEBS J.* **289**, 1160-1179, (2022).
70 Van Meter, M. *et al.* SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat Commun* **5**, 5011, (2014).
71 Hastings, M. L. & Krainer, A. R. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* **13**, 302-309, (2001).
72 Liu, H. X., Cartegni, L., Zhang, M. Q. & Krainer, A. R. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat. Genet.* **27**, 55-58, (2001).
73 Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72-77, (2017).
74 Ilinskaya, O. N. & Mahmud, R. S. Ribonucleases as antiviral agents. *Mol. Biol.* **48**, 615-623, (2014).
75 Meola, N. *et al.* Identification of a nuclear exosome decay pathway for processed transcripts. *Mol. Cell* **64**, 520-533, (2016).
76 Ogami, K. *et al.* An Mtr4/ZFC3H1 complex facilitates turnover of unstable nuclear RNAs to prevent their cytoplasmic transport and global translational repression. *Genes Dev* **31**, 1257-1271, (2017).

77 Lubas, M. *et al.* Interaction profiling identifies the human nuclear exosome targeting complex. *Mol. Cell* **43**, 624-637, (2011).
78 Chen, L. L., DeCerbo, J. N. & Carmichael, G. G. Alu element-mediated gene silencing. *EMBO J.* **27**, 1694-1705, (2008).
79 Monaghan, L., Longman, D. & Cáceres, J. F. Translation-coupled mRNA quality control mechanisms. *EMBO J.* **42**, e114378, (2023).
80 Anderson, P. & Kedersha, N. Stress granules: the Tao of RNA triage. *Trends Biochem. Sci.* **33**, 141-150, (2008).
81 Ding, S. W. & Voinnet, O. Antiviral immunity directed by small RNAs. *Cell* **130**, 413-426, (2007).
82 Gao, G. X., Guo, X. M. & Goff, S. P. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. *Science* **297**, 1703-1706, (2002).
83 Kesner, J. S. *et al.* Noncoding translation mitigation. *Nature* **617**, 395-402, (2023).
84 Liu, J. *et al.* The RNA m6A reader YTHDC1 silences retrotransposons and guards ES cell identity. *Nature* **591**, 322-326, (2021).
85 Ries, R. J., Pickering, B. F., Poh, H. X., Namkoong, S. & Jaffrey, S. R. m6A governs length-dependent enrichment of mRNAs in stress granules. *Nat. Struct. Mol. Biol.*, (2023).
86 Lee, E. S. *et al.* N-6-methyladenosine (m6A) promotes the nuclear retention of mRNAs with intact 5′ splice site motifs. Preprint at bioRxiv https://www.biorxiv.org/content/10.1101/2023.06.20.545713v2 (2023).
87 He, P. C. *et al.* Exon architecture controls mRNA m(6)A suppression and gene expression. *Science* **379**, 677-682, (2023).
88 Delaunay, S., Helm, M. & Frye, M. RNA modifications in physiology and disease: towards clinical applications. *Nat. Rev. Genet.*, doi.org/10.1038/s41576-41023-00645-41572, (2023).
89 Sun, T. *et al.* Crosstalk between RNA m6A and DNA methylation regulates transposable element chromatin activation and cell fate in human pluripotent stem cells. *Nat. Genet.* **55**, 1324-1335, (2023).
90 Janeway, C. A., Jr. Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54 Pt 1**, 1-13, (1989).
91 Logsdon, J. M. The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* **8**, 637-648, (1998).
92 Sakharkar, M. K., Chow, V. T. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol.* **4**, 387-393, (2004).
93 Zhang, J., Sun, X. L., Qian, Y. M., LaDuca, J. P. & Maquat, L. E. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol* **18**, 5272-5283, (1998).
94 Le Hir, H., Nott, A. & Moore, M. J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* **28**, 215-220, (2003).

95 Brocke, K. S., Neu-Yilik, G., Gehring, N. H., Hentze, M. W. & Kulozik, A. E. The human intronless melanocortin 4-receptor gene is NMD insensitive. *Hum. Mol. Genet.* **11**, 331-335, (2002).
96 Savisaar, R. & Hurst, L. D. Purifying selection on Exonic Splice Enhancers in intronless genes. *Mol. Biol. Evol.* **33**, 1396-1418, (2016).
97 Long, H. *et al.* Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol* **2**, 237-240, (2018).
98 Ho, A. T. & Hurst, L. D. Unusual mammalian usage of TGA stop codons reveals that sequence conservation need not imply purifying selection. *PLoS Biol.* **20**, e3001588, (2022).
99 Charneski, C. A., Honti, F., Bryant, J. M., Hurst, L. D. & Feil, E. J. Atypical at skew in Firmicute genomes results from selection and not from mutation. *PLoS Genet.* **7**, e1002283, (2011).
100 Seczynska, M., Bloor, S., Cuesta, S. M. & Lehner, P. J. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature* **601**, 440-445, (2022).
101 Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285-311, (2009).
102 Liu, H. *et al.* Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat Ecol Evol* **2**, 164-173, (2018).
103 Galtier, N. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* **19**, 65-68, (2003).
104 D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. & Bernardi, G. Correlations between the compositional properties of human genes, codon usage, and amino-acid-composition of proteins. *J. Mol. Evol.* **32**, 504-510, (1991).
105 Duret, L. & Hurst, L. D. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.* **18**, 757-762, (2001).
106 Duret, L. & Galtier, N. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.* **17**, 1620-1625, (2000).
107 Morales, A. C. *et al.* Causes and consequences of purifying selection on SARS-CoV-2. *Genome Biol Evol* **13**, 17, (2021).
108 Zhou, Z. *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A* **113**, E6117-E6125, (2016).
109 Newman, Z. R., Young, J. M., Ingolia, N. T. & Barton, G. M. Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proc Natl Acad Sci U S A* **113**, E1362-1371, (2016).
110 Zhao, F. *et al.* Genome-wide role of codon usage on transcription and identification of potential regulators. *Proc Natl Acad Sci U S A* **118**, (2021).
111 Vlaming, H., Mimoso, C. A., Field, A. R., Martin, B. J. E. & Adelman, K. Screening thousands of transcribed coding and non-coding regions reveals sequence

determinants of RNA polymerase II elongation potential. *Nat. Struct. Mol. Biol.* **29**, 613-620, (2022).
112 Pantier, R. *et al.* SALL4 controls cell fate in response to DNA base composition. *Mol. Cell* **81**, 845-858, (2021).
113 Hisano, M., Ohta, H., Nishimune, Y. & Nozaki, M. Methylation of CpG dinucleotides in the open reading frame of a testicular germ cell-specific intronless gene, Tact1/Actl7b, represses its expression in somatic cells. *Nucleic Acids Res.* **31**, 4797-4804, (2003).
114 Hodges, B. L., Taylor, K. M., Joseph, M. F., Bourgeois, S. A. & Scheule, R. K. Long-term transgene expression from plasmid DNA gene therapy vectors is negatively affected by CpG dinucleotides. *Mol. Ther.* **10**, 269-278, (2004).
115 Parmley, J. L., Chamary, J. V. & Hurst, L. D. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* **23**, 301-309, (2006).
116 Carlini, D. B. & Genut, J. E. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* **62**, 89-98, (2006).
117 Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12**, 683-691, (2011).
118 Savisaar, R. & Hurst, L. D. Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. *Mol. Biol. Evol.* **34**, 1110-1126, (2017).
119 Parmley, J. L. & Hurst, L. D. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* **24**, 1600-1603, (2007).
120 Willie, E. & Majewski, J. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**, 534-538, (2004).
121 Warnecke, T. & Hurst, L. D. 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, (2007).
122 Eskesen, S. T., Eskesen, F. N. & Ruvinsky, A. Natural selection affects frequencies of AG and GT dinucleotides at the 5 ' and 3 ' ends of exons. *Genetics* **167**, 543-550, (2004).
123 Livingstone, M. *et al.* Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Hum. Mutat.* **38**, 1336-1347, (2017).
124 Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nat. Rev. Genet.* **2**, 549-555, (2001).
125 Potrzebowski, L. *et al.* Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* **6**, e80, (2008).
126 Mordstein, C. *et al.* Transcription, mRNA Export, and Immune Evasion Shape the Codon Usage of Viruses. *Genome Biol Evol* **13**, (2021).
127 Goodman, D. B., Church, G. M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475-479, (2013).

128 Gu, W., Zhou, T. & Wilke, C. O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* **6**, e1000664, (2010).
129 Palazzo, A. F. *et al.* The signal sequence coding region promotes nuclear export of mRNA. *PLoS Biol.* **5**, e322, (2007).
130 Huang, Y., Gattoni, R., Stevenin, J. & Steitz, J. A. SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Mol. Cell* **11**, 837-843, (2003).
131 Huang, Y. & Steitz, J. A. Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol. Cell* **7**, 899-905, (2001).
132 Courel, M. *et al.* GC content shapes mRNA storage and decay in human cells. *Elife* **8**, e49708, (2019).
133 Chen, C.-Y. *et al.* AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell* **107**, 451-464, (2001).
134 Namkoong, S., Ho, A., Woo, Y. M., Kwak, H. & Lee, J. H. Systematic characterization of stress-induced RNA granulation. *Mol. Cell* **70**, 175-187 e178, (2018).
135 Khong, A. *et al.* The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. *Mol. Cell* **68**, 808-820.e805, (2017).
136 Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124-127, (2017).
137 Duan, J. & Antezana, M. A. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *J. Mol. Evol.* **57**, 694-701, (2003).
138 Hrossova, D. *et al.* RBM7 subunit of the NEXT complex binds U-rich sequences and targets 3'-end extended forms of snRNAs. *Nucleic Acids Res.* **43**, 4236-4248, (2015).
139 Lubas, M. *et al.* The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Rep.* **10**, 178-192, (2015).
140 Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237-243, (2013).
141 Kimchi-Sarfaty, C. *et al.* A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525-528, (2007).
142 Radhakrishnan, A. *et al.* The DEAD-Box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* **167**, 122-132 e129, (2016).
143 Radhakrishnan, A. & Green, R. Connections underlying translation and mRNA stability. *J. Mol. Biol.* **428**, 3558-3564, (2016).
144 Buschauer, R. *et al.* The Ccr4-Not complex monitors the translating ribosome for codon optimality. *Science* **368**, (2020).
145 Medina-Munoz, S. G. *et al.* Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. *Genome Biol* **22**, 14, (2021).

146 Shu, H. *et al.* FMRP links optimal codons to mRNA stability in neurons. *Proc Natl Acad Sci U S A* **117**, 30400-30411, (2020).
147 Kumar, A. *et al.* The slowing rate of CpG depletion in SARS-CoV-2 genomes is consistent with adaptations to the human host. *Mol. Biol. Evol.* **39**, (2022).
148 Ficarelli, M. *et al.* CpG dinucleotides inhibit HIV-1 replication through Zinc Finger Antiviral Protein (ZAP)-dependent and -independent mechanisms. *J. Virol.* **94**, e01337-01319, (2020).
149 Hurst, L. D. *et al.* A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol* **15**, 413, (2014).
150 Bestor, T. H. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **326**, 179-187, (1990).
151 Voo, K. S., Carlone, D. L., Jacobsen, B. M., Flodin, A. & Skalnik, D. G. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol* **20**, 2108-2121, (2000).
152 Bauer, A. P. *et al.* The impact of intragenic CpG content on gene expression. *Nucleic Acids Res.* **38**, 3891-3908, (2010).
153 Singer, T., McConnell, M. J., Marchetto, M. C., Coufal, N. G. & Gage, F. H. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends Neurosci.* **33**, 345-354, (2010).
154 Singh, M. *et al.* A new human embryonic cell type associated with activity of young transposable elements allows definition of the inner cell mass. *PLoS Biol.* **21**, e3002162, (2023).
155 Mirihana Arachchilage, G., Hetti Arachchilage, M., Venkataraman, A., Piontkivska, H. & Basu, S. Stable G-quadruplex enabling sequences are selected against by the context-dependent codon bias. *Gene* **696**, 149-161, (2019).
156 Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat. Rev. Mol. Cell Biol.* **21**, 459-474, (2020).
157 Wang, Y., Qiu, C. & Cui, Q. A large-scale analysis of the relationship of synonymous SNPs changing microRNA regulation withfunctionality and disease. *Int. J. Mol. Sci.* **16**, 23545-23555, (2015).
158 Brest, P. *et al.* A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat. Genet.* **43**, 242-245, (2011).
159 Gartner, J. J. *et al.* Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A* **110**, 13481-13486, (2013).
160 Hamdorf, M. *et al.* miR-128 represses L1 retrotransposition by binding directly to L1 RNA. *Nat. Struct. Mol. Biol.* **22**, 824-831, (2015).

161 Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610-618, (2007).
162 Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**, E268, (2004).
163 Chamary, J. V. & Hurst, L. D. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* **6**, R75, (2005).
164 Nackley, A. G. *et al.* Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* **314**, 1930-1933, (2006).
165 Schattner, P. & Diekhans, M. Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Res.* **34**, 1700-1710, (2006).
166 Wu, X. M. & Hurst, L. D. 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, (2015).
167 Prats-Ejarque, G., Lu, L., Salazar, V. A., Moussaoui, M. & Boix, E. Evolutionary trends in RNA base selectivity within the RNase A superfamily. *Front. Pharmacol.* **10**, 1170, (2019).
168 Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F. & Dietz, H. C. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat. Genet.* **36**, 1073-1078, (2004).
169 Torres, M. *et al.* Paraspeckles as rhythmic nuclear mRNA anchorages responsible for circadian gene expression. *Nucleus (Calcutta)* **8**, 249-254, (2017).
170 Prasanth, K. V. *et al.* Regulating gene expression through RNA nuclear retention. *Cell* **123**, 249-263, (2005).
171 Lucks, J. B., Nelson, D. R., Kudla, G. R. & Plotkin, J. B. Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol* **4**, e1000001, (2008).
172 De Vlugt, C., Sikora, D. & Pelchat, M. Insight into influenza: A virus cap-snatching. *Viruses* **10**, (2018).
173 Jalkanen, A. L., Coleman, S. J. & Wilusz, J. Determinants and implications of mRNA poly(A) tail size--does this protein make my tail look big? *Semin Cell Dev Biol* **34**, 24-32, (2014).
174 Mauro, V. P. Codon optimization in the production of recombinant biotherapeutics: potential risks and considerations. *Biodrugs* **32**, 69-81, (2018).
175 Ho, A. T. & Hurst, L. D. Effective population size predicts local rates but not local mitigation of read-through errors. *Mol. Biol. Evol.* **38**, 244-262, (2021).
176 Allert, M., Cox, J. C. & Hellinga, H. W. Multifactorial determinants of protein expression in prokaryotic open reading frames. *J. Mol. Biol.* **402**, 905-918, (2010).
177 Wu, X. & Hurst, L. D. Determinants of the usage of splice-associated cis-motifs predict the distribution of human pathogenic SNPs. *Mol. Biol. Evol.* **33**, 518-529, (2016).

178 Abrahams, L. *et al.* Evidence in disease and non-disease contexts that nonsense mutations cause altered splicing via motif disruption. *Nucleic Acids Res.* **49**, 9665-9685, (2021).
179 Soemedi, R. *et al.* Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* **49**, 848-+, (2017).
180 Mühlhausen, S. & Hurst, L. D. Transgene-design: a web application for the design of mammalian transgenes. *Bioinformatics*, btac139, (2022).
181 Sharp, C. P. *et al.* CpG dinucleotide enrichment in the influenza A virus genome as a live attenuated vaccine development strategy. *PLoS Pathog* **19**, e1011357, (2023).
182 Yew, N. S. *et al.* CpG-depleted plasmid DNA vectors with enhanced safety and long-term gene expression in vivo. *Mol. Ther.* **5**, 731-738, (2002).
183 Kariko, K., Buckstein, M., Ni, H. & Weissman, D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* **23**, 165-175, (2005).
184 Vaidyanathan, S. *et al.* Uridine Depletion and Chemical Modification Increase Cas9 mRNA Activity and Reduce Immunogenicity without HPLC Purification. *Mol Ther Nucleic Acids* **12**, 530-542, (2018).
185 Ohta, T. The nearly neutral theory of molecular evolution. *Ann Rev Ecol System* **23**, 263-286, (1992).
186 Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943, (2023).
187 Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704-714, (2016).
188 Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A* **109**, 18488-18492, (2012).
189 Seczynska, M. & Lehner, P. J. The sound of silence: mechanisms and implications of HUSH complex function. *Trends Genet.* **39**, 251-267, (2023).

Acknowledgements

The authors thank the Humboldt Foundation for the award of the Humboldt Prize to LDH to enable him to spend time in Germany. SR is funded by the Evolution Education Trust. **[Au: do you wish to include any Acknowledgements?]**

Author contributions

LDH and SR researched data for the article. All authors contributed substantially to discussion of the content. LDH wrote the article. All authors reviewed and/or edited the manuscript before submission. ZI contributed to transposable element issues specifically. **[Au: please provide relevant information]**

Competing interests

The authors declare no competing interests. **[Au:OK?]**

Peer review information

Nature Reviews Genetics thanks the anonymous reviewers for their contribution to the peer review of this work.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s415XX-XXX-XXXX-X

Figure 1 | **Processes acting on transcription, processing and filtering of transcripts with emphasis on roles of codon content or intronic signals.** On the left is the regular gene expression pathway from transcription to translation, highlighting the features that allow transcripts to be recognised as native and correctly expressed as proteins. Arrows coming off towards the right are the features for which transcripts are recognised as "unwanted" and filtered from being translated. In bold are those steps for which codon content or intronic presence is known to be relevant. Relevant abbreviations: RNA Pol II (RNA Polymerase II), CFP1 (CxxC Zinc Finger Protein 1), HUSH (Human Silencing Hub), SALL4 (Spalt-Like) zinc finger protein, ESE (Exonic Splicing Enhancer), SR (Splicing Regulatory) protein, PROMPTs (Promoter Upstream Transcripts), lncRNA (long non-coding RNA), NATs (Natural Antisense Transcripts), dsRNA (double-stranded RNA), IRAlus (Inverte Repeat Alu) elements, ptRNA (prematurely terminated RNA), uaRNA (upstream antisense RNA), PAXT (PolyA Exosome Targeting) complex, RNase (ribonuclease), NEXT (Nuclear Exosome Targeting) complex, TREX (Transcription Exon) complex, NXF1/Tap (Nuclear RNA Export Factor 1), P bodies (Processing bodies), ZAP (Zinc-finger Antiviral Protein), NMD (Nonsense Mediated Decay), Dicer (endoribonuclease Dicer), Dhh1p (DEAD/H-box protein) Ccr4-Not (Carbon Catabolite Repression-Negative On TATA-less) transcription complex.

Figure 2 | **Dinucleotide frequencies of human coding sequences, single-exon coding sequences and genomic DNA, and as expected from mononucleotide frequencies. a,** The data are ordered according to the difference between dinucleotide frequency in the genome and that in coding sequence. Note that the four dinucleotides bearing only A and/or T are all rare in coding sequence, whereas the four dinucleotides bearing only G and/or C are the most enriched in coding sequence compared with genomic DNA. CpG is unique amongst these four in that its frequency in coding sequence is less than would be expected from its mononucleotide content, most probably owing to the mutagenicity of CpG. **b,** The absolute difference in dinucleotide content between coding sequence and genomic DNA. Bars are colour coded by G and/or C content, with darker colour indicating higher GC content. Data derived by the authors from human genome release GRCh38.

Figure 3 | **Native human exons predicted to be targeted by the HUSH complex.** Exons longer than ~1500bp and with an A content greater than ~27% are targeted for silencing by the HUSH complex (from Extended Data Fig. 3c of ref.[100]). This population is shown as the top right quadrant and accounts for less than 0.4% of native exons and about 3.5% of native human genes. We highlight a few zinc finger (ZNF) genes and a proto-cadherin. Single-exon coding sequence genes are shown in yellow. Note the rarity of human exons likely to be

subject to HUSH. This is consistent with both HUSH largely targeting foreign genes and forcing selection on the anatomy of native genes. Note too the commonality of ZNF genes in the group that are likely to human silenced by HUSH suggesting that HUSH and ZNF comprise alternative transposable element suppression systems, the one regulating the other. Data derived by the authors from human genome release GRCh38.

Figure 4 | **The mode of action of disease-associated synonymous mutations.** Synonymous mutations can interfere with post-transcriptional processing by disrupting constitutive splice sites or activating cryptic splice sites, therefore altering SR protein binding and the production of correctly spliced mature mRNAs. Synonymous mutations can also influence mRNA secondary structure and stability and miRNA-mediated gene regulation, in turn altering downstream protein expression. During translation, synonymous mutations can alter ribosome processivity, influencing translational kinetics and thus co-translational protein folding. Production of aberrant or truncated proteins can result in non-functional proteins or, in most cases, degradation. Examples of pathogenic synonymous mutations are indicated (in bold) next to their proposed mechanism of pathogenicity. See **Supplementary Table 1** for full list and references. Relevant abbreviations: RNA Pol II (RNA Polymerase II), ESE (Exonic Splicing Enhancer), NMD (Nonsense Mediated Decay).

Box 1 | **The nearly-neutral theory and the importance of effective population size**

What is the fate of mutations that have little or no effect on fitness? One of the major advances in theoretical population genetics has been the realization that mutations with little effect on fitness have a different fate to those with no effect. The fate of mutations that have no effect is explained by the neutral (or strictly-neutral) theory, whereas the fate of mutations with little effect is explained by the nearly-neutral theory[185]. Both theories assume that random changes in allele frequency (known as drift) are an influential process.

According to the strictly-neutral theory, the probability that one of two neutral alleles will drift to fixation is its current frequency. Thus, when a new mutation arrives in a diploid organism, it will be at a frequency $1/2N$, where N is the population size, and its probability of fixation is $1/2N$. The rate at which such mutations arrive is $2N \times mu$, where mu is the per generation mutation rate. Thus, the net rate of evolution is $2N \times mu/2N = mu$. Importantly, the rate of strictly-neutral evolution is independent of the population size.

In the nearly-neutral model, mutations reduce fitness by a very small degree, s. The key postulate of the nearly-neutral model is that as the population size increases (or, rather, the effective population size, N_e), selection is a more effective force, meaning that the chance that

a deleterious mutation will be eliminated increases with increasing N_e. This is because in a small population, a mutation with small effect requires fewer chance increases in frequency to go from being rare to common, whereas in a large population, selection has longer (in terms of the number of generations) to remove the deleterious mutation from the population before it reaches fixation by chance.

A key question concerns the relationship between s, N_e and the rate of fixation. The nearly-neutral theory divides deleterious mutations into one of three classes. Some mutations are of such small effect that their rate of fixation is so close to *mu* that they can't be distinguished from strictly-neutral mutations. These 'effectively neutral' mutations occur when $s<<1/2\ N_e$. A further class of mutations are so deleterious that they would never reach fixation; these 'strongly deleterious' mutations occur when $s>>1/2\ N_e$. The third class are those intermediary mutations that can reach fixation but do so at a slower rate than the mutation rate. These 'weakly deleterious' mutations occur when $s \cong 1/2\ N_e$.

A key result of the nearly-neutral model is that, if we assume a mutation has a small deleterious effect on fitness, its chance of fixation increases when N_e is small as it is more likely to be within the 'effectively neutral' or 'weakly deleterious' mutation classes. For example, insertion of a small segment of DNA with little phenotypic consequence could well be 'strongly deleterious' in a species with a large N_e and hence removed from the population, but 'effectively neutral' or 'weakly deleterious' in a species with a small N_e and hence not always removed, despite s being the same in all cases. In this way, the nearly-neutral model explains[11] why the human genome contains so much unconstrained 'junk' DNA (only about 10% of human DNA is constrained[186]), why we have large introns that are hard to splice accurately and why much DNA is prone to spurious binding of transcription factors. In each case we evoke the idea that a harmful mutation happens, such as insertion of a transposable element, a mutation decreasing splicing fidelity or one providing a spurious TF binding site etc.. The fate of such weak effect mutations will be dependent on the efficiency of selection being able to remove the mutation from the population, efficiency that is lower when N_e is small. Junk DNA is DNA that is 'weakly deleterious' and thus cannot always be prevented from reaching fixation by drift. The model evokes a balance between drift and selection to keep genomes free of 'strongly deleterious' mutations.This drift–selection balance can, in addition, explain why the mutation rate is higher when N_e is low[187,188].

Box 2 | **The HUSH complex**

The human silencing hub (HUSH) is a transcriptional silencing system, constitutively expressed in most cell types, that, amongst other activities, safeguards against retroelement

invasion and controls the flow of genetic information from RNA to DNA in mammalian genomes. HUSH represses retroviruses and retroelements (non-self) while largely avoiding the inappropriate repression of intron-containing host genes (self) (for review, see ref.[189]).

HUSH is a complex of M-phase phosphoprotein 8 (MPP8), transcription activation suppressor (TASOR) and periphilin. In the canonical pathway, periphilin binds RNA and then guides HUSH to its target loci. At the targets, HUSH promotes deposition of histone 3 lysine 9 trimethylation (H3K9me3), as well as microrchidia CW-type zinc finger 2 (MORC2) and the nuclear exosome targeting (NEXT) complex. Although, canonically, HUSH is recruited via RNA (and hence requires some expression[100]), it can also be recruited to DNA by DNA-binding protein nuclear protein 220 (NP220).

HUSH targets tend to be long exons[100] (>1.5kb) of either multi-exon genes or long intronless genes, neither of which are common for human transcripts. In addition, it targets mRNAs with high adenine (A) content. This is thought to enable the targeting of retroviruses and retrotransposons as retroposition is associated with high A content. Through its ability to distinguish self from non-self without the need for prior exposure (**Fig. 3**), HUSH may be considered a component of the innate immune system and intronless cDNA may be considered a pathogen-associated molecular pattern.

Given its mode of pattern recognition, HUSH suppresses retroelements from outside the cell (retroviruses such as HIV) as well as from within the cell (retrotransposons). HUSH deficiency phenotypes are likely owing to the release from repression of LINE1 transposons. HUSH thus guards against the activation of interferon and other immune genes that would otherwise be stimulated by LINE1 and associated double-stranded RNA (sensed by cytoplasmic RIG-I-like receptors).

Although HUSH is a system that has evolved to target mobile elements — and lentiviruses have co-evolved mechanisms to degrade HUSH by Vpr and Vpx — this suppressor system can be co-opted to control host genes. Native gene targets of HUSH include KRAB-zinc finger proteins (encoded by a single exon of 4.5kb) and some protocadherins (encoded by a single exon >2.5kb). The HUSH-mediated suppression of protocadherins is important for neuronal development.

HUSH can also repress recombinant adeno-associated virus (AAV), which is commonly used as a delivery vector in gene therapy. As HUSH recognizes unusually long exons and intronless transcripts, it presents an impediment to gene therapy for larger genes. That introns enable

escape from HUSH-dependent silencing through splicing-independent routes (unspliced mutant introns also enable avoidance of HUSH) suggests that it may be possible to design transgenes that evade this silencing.

Glossary

GC-biased gene conversion

(gBGC). Biased gene conversion describes the recombination of short stretches of genetic material from a donor sequence to an acceptor sequence that is biased as to which is the donor strand and which is the acceptor strand. In gBGC, AT:GC mismatches in recombinant sections are resolved with a preference towards G or C.

Codon usage bias

A pattern in which synonymous codons occur at frequencies different from some null expectation (often the null expectation is of equal frequency).

CpG islands

Long (>200bp) stretches of DNA rich in CpG dinucleotides that commonly occur in the vicinity of mammalian promoters. They are commonly unmethylated if the associated gene is expressed.

Cryptic splice sites

Splice sites within exons away from the canonical splice site that can result in alternative splicing.

Effective population size

(N_e). The number of individuals that an idealized neutrally evolving population would require for it to have properties (such as genetic diversity) equivalent to the real population. N_e is often smaller than the census population size (N). In simple cases, it approximates to the number of breeding individuals.

Exonic splice enhancers

(ESEs). Short (6–8bp) RNA motifs that enable more accurate splicing, particularly in the vicinity of weak splice sites. ESEs often function as RNA binding sites for serine–arginine-rich (SR) proteins.

Exon-junction complex

Protein complex that forms on a pre-mRNA molecule at the junction between two exons joined by RNA splicing.

Exosome complex

A multi-protein intracellular protein complex that degrades many types of RNA. In eukaryotes it is present in the cytoplasm, the nucleus and the nucleolus. In humans the cytoplasmic form is associated with the DIS3-like exonuclease 1 (DIS3L), while the nuclear complex employs DIS3. The nuclear form is termed the nuclear exosome complex, that which is targeted by NEXT complex. The (RNA) exosome complex is not to be confused with exosomes, extracellular vesicles generated by cells.

G-quadruplexes

Stable secondary structures formed in guanine-rich DNA and RNA sequences.

Inverted repeat Alu elements

Single-stranded Alu elements that are followed downstream by their reverse complement. This characteristic often allows inverted repeats to fold on themselves and form double-stranded structures.

LINE1

(L1). LINE elements, of which L1 is an example, are the most abundant class of transposable elements in the human genome, formed by autonomous retrotransposition. **[Au:OK?]**

Mutational equilibrium

The frequency (for example, of nucleotides) in a population at which, if only mutation bias and neutral evolution affect the frequency, the frequency does not change.

Neutral evolution

The process by which allele frequency is determined by chance events alone operating on alleles of the same fitness.

Nonsense-mediated decay

(NMD). Process by which mRNA molecules containing premature stop codons are degraded.

No-go mRNA decay

(NGD). Process by which RNA with stacks of stalled ribosomes are degraded.

Codon optimality-mediated RNA decay

(COMD). Process by which mRNA with abundant non-optimal codons (those matching rare tRNAs) are subject to decay, probably mediated by slow ribosomal progression.

Non-stop RNA decay

(NSD). Process by which mRNA without a proper stop codon are identified and targeted for decay. In eukaryotes, NSD discharges ribosomes stalled at the 3' end, directing the mRNA to the exosome complex.

Processing bodies

Ribonucleoprotein bodies found in the cytoplasm that retain mRNA molecules and contain proteins required for mRNA decay. A similar role is carried out by cytoplasmic stress granules.

Retrogenes

Processed copies of genes formed from reverse transcription of mature mRNA molecules of a parental gene (hence without introns).

Spurious transcripts

Transcripts generated by functionally irrelevant cellular events (such as TF binding to random sequence).

Synonymous mutations

Single base-pair mutations in a protein-coding exon that change the codon to a different but synonymous one and hence do not a priori change the amino acid sequence of the encoded protein.

Translational selection

Selection that favours synonymous mutations owing to their effects on translation, typically assumed to be mediated by faster or more accurate translation. Commonly evidenced by a codon usage bias that favours synonymous codons matching more abundant iso-acceptor tRNAs.

Transposable elements

DNA sequences that can move within genomes and replicate without depending on gene replication of the host cell.

Zinc finger antiviral protein

(ZAP). Protein involved in preventing retroviral infection by binding of CpG-rich sequences and recruitment of the RNA degradation machinery.

Table of Contents

Multiple mechanisms have evolved to prevent or trap deleterious unwanted transcripts. The unwanted transcript hypothesis proposes that selection at synonymous sites favours mutations that prevent the generation of transcriptional rubbish or that make native transcripts look “wanted” by being GC-rich.