# Read between the Lines: Diversity of Nontranslational Selection Pressures on Local Codon Usage

Martijn Callens, Léa Pradier, Michael Finnegan, Caroline Rose, and Stéphanie Bedhomme*

Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, Ecole Pratique des Hautes Etudes, Institut de Recherche pour le Développement, Montpellier, France

*Corresponding author: E-mail: stephanie.bedhomme@cefe.cnrs.fr.

**Accepted:** 28 April 2021

## Abstract

Protein coding genes can contain specific motifs within their nucleotide sequence that function as a signal for various biological pathways. The presence of such sequence motifs within a gene can have beneficial or detrimental effects on the phenotype and fitness of an organism, and this can lead to the enrichment or avoidance of this sequence motif. The degeneracy of the genetic code allows for the existence of alternative synonymous sequences that exclude or include these motifs, while keeping the encoded amino acid sequence intact. This implies that locally, there can be a selective pressure for preferentially using a codon over its synonymous alternative in order to avoid or enrich a specific sequence motif. This selective pressure could—in addition to mutation, drift and selection for translation efficiency and accuracy—contribute to shape the codon usage bias. In this review, we discuss patterns of avoidance of (or enrichment for) the various biological signals contained in specific nucleotide sequence motifs: transcription and translation initiation and termination signals, mRNA maturation signals, and antiviral immune system targets. Experimental data on the phenotypic or fitness effects of synonymous mutations in these sequence motifs confirm that they can be targets of local selection pressures on codon usage. We also formulate the hypothesis that transposable elements could have a similar impact on codon usage through their preferred integration sequences. Overall, selection on codon usage appears to be a combination of a global selection pressure imposed by the translation machinery, and a patchwork of local selection pressures related to biological signals contained in specific sequence motifs.

**Key words:** codon usage, synonymous mutations, gene expression regulation, sequence targeting antiviral immune systems, transposable elements.

## Significance statement

The frequency of use of synonymous codons varies between species and is known to be under selection for translation speed and accuracy. In this review, we argue that an additional local selection pressure on codon usage is generated by sequence motifs conveying different biological signals such as transcription and translation initiation, mRNA maturation, antiviral immune system targets or preferred transposable elements insertion sequences. Alternative synonymous sequences can be favored or disfavored because they contain these motif sequences. We review experimental and bioinformatic evidence for these local selection pressures.

## Introduction

The redundancy of the genetic code is a consequence of the existence of synonymous codons, which differ by their nucleotide triplets but code for the same amino acid. The different codons within a synonymous codon family are not used at equal frequencies; this codon usage bias (CUB) can vary between species and between genes within a species (Grantham et al. 1981; Ikemura 1985). CUB is shaped by mutation, selection, and drift (Bulmer 1991; Hershberg and Petrov 2008; Plotkin and Kudla 2011; Shah and Gilchrist 2011). Selection on CUB is generally assumed to be driven by its effects on translation efficiency (Tuller, Waldman, et al.

2010) and accuracy (Kurland 1992; Stoletzki and Eyre-Walker 2007), mediated by the coevolution of translation machinery and CUB: an association between the frequency of use of a codon and the availability of the corresponding decoding tRNA has been established for various genomes (e.g., Duret 2000; Rocha 2004). Codon usage has been shown to modulate the rate and efficiency of translation, with examples ranging from decreases in viral capsid protein production leading to virus attenuation (Coleman et al. 2008) to 58% translation elongation rate increases in human cell lines (Yan et al. 2016).

Selection on codon usage does not always act in the direction of higher translation efficiency, and this direction can vary across the genome and within genes. For example, in many prokaryotic and eukaryotic species the first 30–50 bp of genes often present an accumulation of codons which are at low frequency in the rest of the genome. This has been associated with a localized slow translation, preventing ribosomal collisions downstream (Tuller, Carmi, et al. 2010). In bacteria, it has been established that the corresponding part of the mRNA presents a reduced folding energy compared to the rest of the mRNA, which is assumed to favor translation initiation. An analysis of over 400 bacteria genomes confirmed that codons overrepresented at the beginning of the genes are those that reduce mRNA folding around the translation start, regardless of whether these codons are frequent or rare (Bentele et al. 2013).

Ribosome profiling and other technical advances have led to an in-depth understanding of the complex relationship between codon usage, translation efficiency regulation, and proteome composition. They enabled, for example, descriptions of the effect of codon usage on mRNA secondary structure (Katz 2003) and accessibility to ribosomes (Kudla et al. 2009) as well as the measure of the rate of ribosomal drop-off at low-frequency codons producing truncated proteins (Yang et al. 2019). The kinetic coupling of translational speed and protein folding has been described in detail (Pechmann and Frydman 2013; Yu et al. 2015; Chaney et al. 2017; Zhao et al. 2017). Finally, the modulatory role of codon usage in mRNA decay and stability has been documented in bacteria (Boël et al. 2016), single celled eukaryotic yeast (Radhakrishnan et al. 2016), and between different tissues in humans (Burow et al. 2018). In particular, in human cells, codon usage is a key determinant of the routing of mRNA towards P-bodies which are cytoplasmic organelles involved in mRNA storage and decay (Courel et al. 2019). These phenomena have been reviewed by Brule and Grayhack (2017) and are not the focus of the present review.

The existence of alternative synonymous sequences suggests that protein coding genes could potentially contain or exclude sequence motifs with biologically meaningful signals in addition to simply coding for an amino acid sequence. These biological signals can take the form of motifs in the actual nucleotide sequence, or in the biophysical properties of this sequence (secondary structure, hairpins, stiffness, etc.). The presence of these "other codes" is particularly recognized for biological signals involved in gene expression (e.g., Bergman and Tuller 2020), and it has been suggested that the genetic code is better suited for encoding this additional information than the vast majority of the potential alternative genetic codes (Itzkovitz and Alon 2007). We argue here that the potential for genes to contain information beyond the code of the amino acid sequence implies that specific nucleotide sequences can be favored or disfavored, because of the biological signal they carry. This can result in selection on local codon usage for reasons other than its consequences on translation accuracy and efficiency. In this review, we compile the different biological signals that can be contained in nucleotide sequences. We further discuss patterns of avoidance or enrichment of these sequence motifs and, when available, we present experimental evidence of the phenotypic effects of synonymous mutations in relation to these biological signals. Figure 1 provides a summary of the elements discussed in this review.

## Sequence motifs involved in gene expression regulation

### Promoter, Near-Promoter and Antisense Promoter Sequences

Promoters in bacteria are characterized by two consensus sequences, TATAAT and TTGACA, respectively located 10 and 35 bp upstream of the transcriptional start site (Browning and and Busby 2004). Active promoter sequences are not necessarily an exact consensus sequence but usually contain only three or four of the six nucleotides (Kinney et al. 2010). Promoter sequences, or sequences within a short mutational distance from a promoter sequence, are likely to occur within DNA sequences because they are short and moderately conserved. Indeed, 10% of 100 bp random sequences exhibit promoter activity in *Escherichia coli*, and within 250 generations 60% of random sequences evolved functional promoter activity due to a single mutation (Yona et al. 2018). The potential of a given sequence to evolve a functional promoter can be beneficial in terms of plasticity and evolvability of the transcription network. It can even be beneficial when occurring in a coding sequence: for example, in bacteria, synonymous mutations at the end of the coding sequence of a gene have been shown to be beneficial because they create a promoter from which the next gene in the operon is transcribed, and this overexpression is advantageous in specific environmental conditions (Ando et al. 2014; Kershner et al. 2016). However, the appearance of a new promoter within a coding sequence can also lead to an overproduction of RNA transcripts, sequestration of RNA polymerase, and an overall reduction in gene expression (Lamberte et al. 2017). Hahn (2003) found that coding sequences across

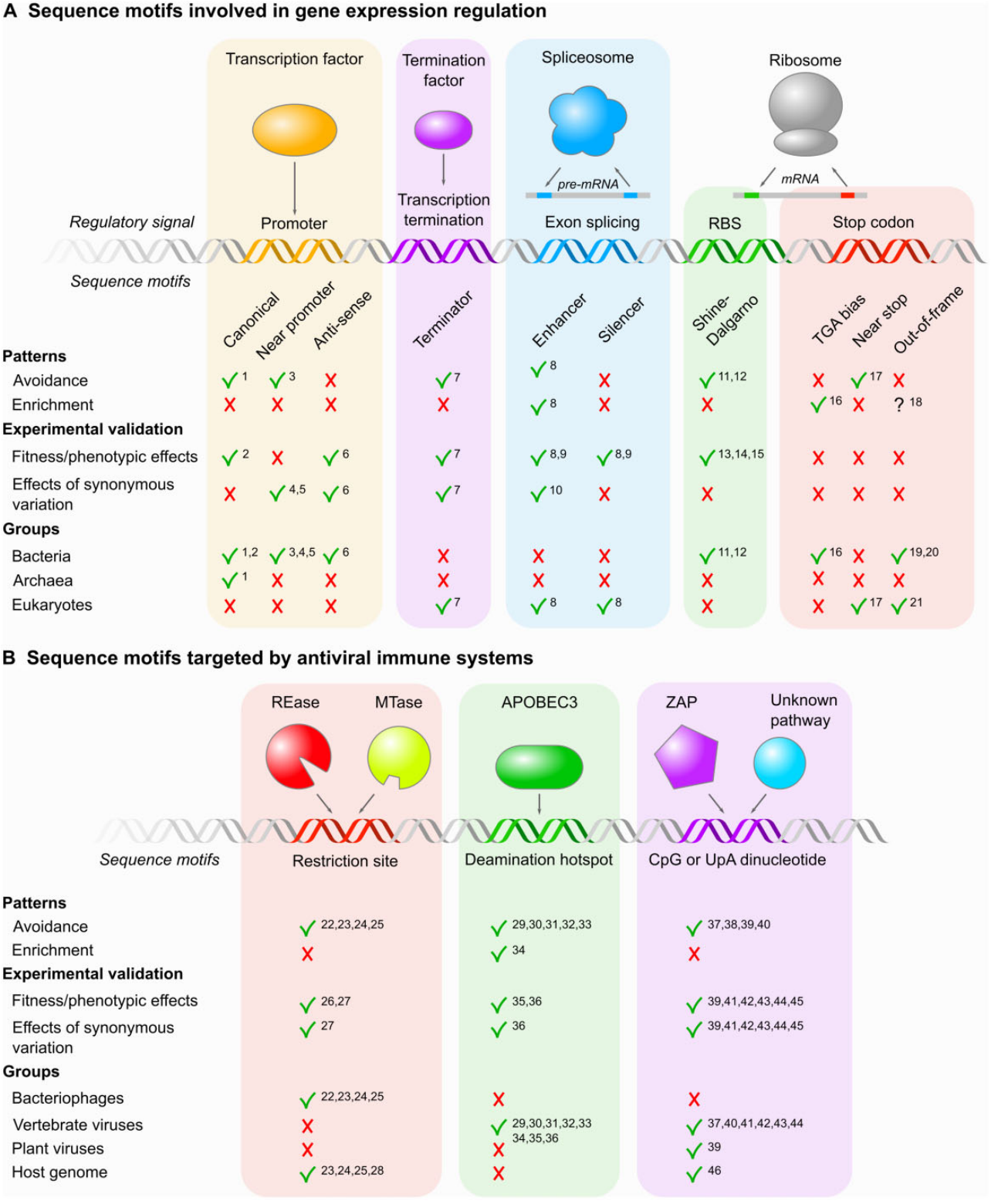


FIG. 1.—(A) Observed avoidance or enrichment of sequence motifs involved in gene expression regulation and potential phenotypic effects. Different processes depend on particular sequence motifs in the DNA or mRNA for their regulation (colored boxes from left to right: transcription initiation, transcription termination, gene splicing, translation initiation, translation termination). Green checks indicate if there is evidence in the literature for avoidance or enrichment of particular sequence motifs, if the presence or absence of these sequence motifs has observable phenotypic effects and if these phenotypic effects can be modified through synonymous variation. An "?" indicates this issue is debated. The bottom rows indicate in which domains of life these

Eubacteria and Archaea are under selection to avoid canonical promoter sequences, and Yona et al. (2018) computationally showed that the *E. coli* coding genome is depleted in sequences close to promoter sequences. Furthermore, this avoidance pattern is even stronger for essential genes, for which perturbation is extremely costly. This suggests that specific intragenic combinations of codons corresponding to promoter or near-promoter sequences are generally disadvantageous but can also be beneficial in specific genomic and environmental situations.

Intragenic promoters are, however, present on the antisense strand in a diversity of bacterial species (Cohen et al. 2016). Transcription from antisense promoters produces RNA fragments that are strictly complementary to the mRNAs produced from the sense strand and can hybridize with them. Antisense transcripts often lead to some repression of translation because the presence of RNA duplexes along mRNA can inhibit translation and target mRNA for degradation (Brantl 2007; Brophy and Voigt 2016). It is unclear when and to what degree the presence of these antisense promoters is spurious or favored by selection because of their role in translational regulation (Gophna 2018). Urtecho et al. (2020) showed experimentally that *E. coli* genes containing antisense promoter sequences had lower transcript levels. This study also revealed that the portions of the sense strand complementary to the antisense promoters contain many codons present at low frequency in the rest of the genome. These sequences thus seem to be constrained both by their role in amino acid coding and as antisense promoters with a regulatory function. In this context, synonymous mutations could have a phenotypic impact by affecting the functionality of antisense promoters and consequently the transcript levels of the genes containing them.

## Ribosome Binding Sequences

Translation of mRNA is initiated by the binding of a ribosome to the ribosomal binding site (RBS). Across all bacterial species, the consensus RBS consists of a 6–7 bp motif found 5–10 bp upstream of the start codon and complementary to the 3′ tail of the 16S ribosomal RNA (Shine and Dalgarno 1974). RBSs are relatively short and sequences that are one or two mutations away from the consensus Shine–Dalgarno sequence can be a functional RBS (Omotajo et al. 2015). Intragenic RBSs may promote spurious internal translation initiation leading to the production of frame-shifted or truncated protein (Whitaker et al. 2015), which is expected to have negative fitness effects (Drummond and Wilke 2009). Intragenic RBSs are also known to increase the rate of ribosomal frameshifting during translation elongation. In some cases, this has been shown to be "programmed frameshifting" allowing the production of two different functional proteins from the same coding sequence (Devaraj and Fredrick 2010; Chen et al. 2014). However, cases of spurious ribosomal frameshifting during translation elongation are likely to have negative consequences. In various bacterial species, internal RBSs have also been shown to induce translational pauses by directly binding to the ribosome and thereby reducing the local translation elongation rate (Li et al. 2012; Schrader et al. 2014), leading to a reduction in the quantity of protein produced (Osterman et al. 2020). This slow local translation can have a positive effect on fitness by allowing correct protein folding or down-regulating protein translation (Fluman et al. 2014; Frumkin et al. 2017), or a negative effect if this down-regulation is maladaptive. Like promoter sequences, RBSs also have a high probability of occurring by chance in coding sequences, given their small size. It is difficult to predict whether these motifs will be favored or disfavored by selection because of the diversity of mechanistic and fitness consequences intragenic RBSs can have. The vast majority of prokaryotic protein coding sequences are depleted of internal RBSs (Itzkovitz et al. 2010; Diwan and Agashe 2016). Using a comparative approach, Hockenberry et al. (2018) showed that strong intragenic RBSs detected in *E. coli* present a low level of conservation across *Enterobacteriales* and that sequences downstream of internal RBSs are strongly depleted of ATG start codons. Both observations suggest a negative effect of the presence of these sequences. The general

observations have been made. References: [1]Hahn (2003), [2]Lamberte et al. (2017), [3]Yona et al. (2018), [4]Ando et al. (2014), [5]Kershner et al. (2016), [6]Urtecho et al. (2020), [7]Zhou et al. (2018), [8]Savisaar and Hurst (2017), [9]Sterne-Weiler et al. (2011), [10]Mueller et al. (2015), [11]Itzkovitz et al. (2010), [12]Diwan and Agashe (2016), [13]Schrader et al. (2014), [14]Li et al. (2012), [15]Osterman et al. (2020), [16]Eyre-Walker (1996), [17]Johnson et al. (2011), [18]Morgens et al. (2013), [19]Tse et al. (2010), [20]Abrahams and Hurst (2018), and [21]Bertrand et al. (2015). (B) Observed avoidance or enrichment of sequence motifs targeted by antiviral immune systems and potential phenotypic effects. Different types of antiviral immune systems are considered (colored boxes from left to right: bacterial R-M systems [Rease-MTase]; mammalian apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 [APOBEC3] mediated innate immunity; eukaryotic antiviral pathways targeting CpG or UpA dinucleotides of which the zinc-finger antiviral protein [ZAP] is known to act in vertebrates but for plants the molecular pathways are yet to be elucidated). Green checks indicate if there is evidence in the literature for avoidance or enrichment of particular sequence motifs, if the presence or absence of these sequence motifs has observable phenotypic effects and if these phenotypic effects can be modified through synonymous variation. The bottom rows indicate in which host groups observations have been made in their infecting viruses or in the host genome itself. References : [22]Sharp (1986), [23]Karlin et al. (1992), [24]Rocha (2001), [25]Rusinov et al. (2018), [26]Pleška et al. (2016), [27]Pleška and Guet (2017), [28]Gelfand and Koonin (1997), [29]Warren et al. (2015), [30]Poulain et al. (2020), [31]Martinez et al. (2019), [32]Chen and MacCarthy (2017), [33]Verhalen et al. (2016), [34]Monajemi et al. (2014), [35]Armitage et al. (2012), [36]Sato et al. (2014), [37]Chen et al. (2014), [38]Simmonds et al. (2013), [39]Ibrahim et al. (2019), [40]Xia (2020), [41]Burns et al. (2009), [42]Gaunt et al. (2016), [43]Takata et al. (2017), [44]Fros et al. (2017), [45]Trus et al. (2020), and [46]Burge et al. (1992).

pattern emerging from these data is a pattern of selection against intragenic RBSs although they may be favored by local selection when their regulatory effect on protein elongation is beneficial. Regardless of the direction of selection on intragenic RBSs, these selective pressures have the potential to impact local codon usage (Li et al. 2012).

### Overlapping and Near-Overlapping Genes

Overlapping genes are widespread in bacterial genomes because of their high gene density: a study analyzing 699 bacterial species revealed more than 90% have at least one overlapping gene pair (OGP), while some genomes harbor up to 3,000 OGPs (Ahnert et al. 2008). Additionally, a high proportion of codirectional gene pairs are "near-OGPs" with less than 40 bps between the two genes (Pallejà et al. 2009). As a consequence, the upstream gene sequence provides both the code for its own amino acid sequence and the promoter and RBS of the downstream gene. For OGPs, the 3′ end of the upstream gene also codes for the amino acid sequence of the downstream gene (Huvet and Stumpf 2014). The double role of these regions constrains the codon usage and partially explains why CUB on the end of bacterial genes is often different from the rest of the genome (Eyre-Walker 1996).

### Stop, Near-Stop and Out-of-Frame Stop Codons

Stop-codon usage is under similar global selection pressure as other codons. In particular, a correlation has been established between stop codon use and availability of the corresponding release factor (Korkmaz et al. 2014). Stop-codon usage is additionally under specific selection pressure in many upstream genes of OGPs in prokaryotes; which often share 1 or 4 bp with the downstream gene, resulting in the overlap of the upstream gene stop codon with the downstream gene ATG start codon. This overlap restricts the choice for stop codons and favors the use of TGA (Eyre-Walker 1996).

Some amino-acid coding codons, called *near-stop codons*, have only one nucleotide difference from stop codons. Near-stop codons can lead to processivity errors when mutations or transcription/translation errors occur (Freistroffer et al. 2000). As processivity errors lead to the production of truncated proteins, they are costly, particularly if they occur late in translation. Selection is predicted to disfavor near-stop codons within coding regions, with a gradual increase in selection pressure along the coding sequence. To our knowledge, only one study has attempted to test this prediction (Johnson et al. 2011), which found evidence for the predicted pattern in coding regions of yeast and humans. Additionally, this selection pressure against near-stops seems to be released in the 30–50 codons upstream of the stop codon. However, certain amino-acids are coded only by near-stop codons, while other amino-acids can be coded by both near-stop and nonnear-stop codons. This result should therefore be regarded with some caution because no correction was made for amino-acid usage. If the hypothesis were verified across species, this would indicate that avoidance of near-stop codons partially shapes the CUB for the four amino-acids coded both by near-stop and non near-stop codons (Leucine, Serine, Arginine, and Glycine).

Finally, the ambush hypothesis proposes that selection might favor out-of-frame stop codons in coding regions, allowing translation to be rapidly aborted when ribosomal frame-shifts occur, thereby reducing the cost of producing a long nonfunctional polypeptide (Seligmann and Pollock 2004). Various studies (Singh and Pardasani 2009; Tse et al. 2010; Bertrand et al. 2015; Abrahams and Hurst 2018) have tried to test the ambush hypothesis, but disagree on the interpretation of the analysis performed and no general conclusion has been reached for now. Indeed, a vast majority of the studies detected an enrichment of out-of-frame stop codons in coding sequences but this enrichment is not significantly more pronounced than the enrichment in other out of frame codons (Morgens et al. 2013). If out-of-frame stop codons are indeed enriched in coding regions, this will have an impact on the specific in-frame codons used.

### Transcription Termination Sequences

Transcription termination signals may play an important role in shaping CUB in eukaryotes. Endonucleolytic cleavage of nascent eukaryotic mRNAs is followed by synthesis of the polyadenosine (poly(A)) tail at specific *cis*-acting polyadenylation sites. These sites, called poly(A) signals, are generally highly conserved AU-rich motifs, mutations in which lead to defects in mRNA processing (Tian and Manley 2017). Using the eukaryotic model organism *Neurospora crassa*, Zhou et al. (2018) demonstrated experimentally that rare codons led to premature transcription termination by creating putative poly(A) sequences. This is because there is a strong preference for C/G nucleotides at the wobble positions of *N. crassa* codons, so genes with rare codons contain higher A/U frequencies and are more likely to lead to the formation of poly(A) signal motifs. Zhou et al (2018) also showed, using a bioinformatics approach, a similar consequence of rare codon usage in mice. The authors suggest that preferences in codon usage may have coevolved with transcription termination machinery to avoid costly premature termination of transcription in GC-rich eukaryotes.

### Exon Silencing and Exon Enhancer Sequences

In eukaryotic gene expression, transcription is followed by splicing—a process through which nonprotein coding introns are removed from the pre-mRNA, and protein-coding exons are joined to produce a mature mRNA. Splicing is catalyzed by a large RNA–protein complex that recognizes specific sequence motifs in the pre-mRNA, both within introns and

exons (Abramowicz and Gos 2018). Exons contain Exonic Splice Enhancers (ESE) and Exonic Splice Silencers (ESS), which enhance integration into the mature mRNA or silence it, respectively. Disruption of ESE sites can cause the skipping of exons, leading to the production of dysfunctional proteins. Conversely, the creation of new ESS sites can lead to a similar outcome, by skipping previously included exons. Many ESE sites are involved in interactions with RNA-binding proteins (RBPs) and a selective pressure to conserve or avoid RBP motifs has been shown in primates and rodents (Savisaar and Hurst 2017). Interestingly, the strength of selection to conserve ESEs has been linked to effective population size. Wu and Hurst (2015) showed, in a study across 30 different species that mean intron size predicts ESE density, with mean intron size negatively correlating with effective population size. This argument also holds within species, with higher ESE density at genes with larger and more numerous introns.

Perturbation of exon encoded regulatory information has been associated with numerous human pathologies, including cystic fibrosis, Lynch syndrome, breast cancer, muscular dystrophy and haemophilia B (Sterne-Weiler et al. 2011; Savisaar and Hurst 2017). A comparative study (Fairbrother et al. 2004) showed that exon ends, where ESE are located, contain fewer single nucleotide polymorphisms than the central region of exons, and linked this pattern to the highly conserved splicing regulatory information encoded at exon extremities. Additionally, an experimental approach determined that 23% of synonymous mutations across exon 7 of the human *SMN1* gene decrease exon integration into mRNA (Mueller et al. 2015). This suggests that for some genes, splicing signals are encoded over the whole length of the exon. Thus, avoidance and maintenance of splice signals and other nonsplicing-associated RBP motifs could influence codon usage over extensive portions of the coding genome.

## Sequence Motifs Targeted by Antiviral Immune Systems

Viral reproduction depends on their host's cellular machinery because viruses release their genetic material directly into the cytoplasm of host cells where replication, transcription, and translation occur. The genetic material of viruses is thus a direct target for intracellular antiviral immune systems that recognize foreign nucleic acids based on specific sequence motifs, subsequently degrade the viral genetic material, and thus impede viral replication. In response, viruses have evolved sophisticated mechanisms to evade host immune responses such as DNA modification, the production of proteins that inhibit the action of certain restriction systems, the use of unusual bases in their genetic material and virus-encoded methylation (Tock and Dryden 2005; Harris and Dudley 2015). However, to evade immune systems that rely on the recognition of specific sequence motifs, the simplest strategy is to avoid these sequence motifs in their genetic material. Viruses have been shown to effectively evade host immune responses through synonymous mutations that remove target sequence motifs from their genome—while keeping the integrity of their coding sequences (Pleška and Guet 2017; Takata et al. 2017). This mechanism appears to be widespread, and the following sections provide an overview of the avoidance of sequence motifs in viral genomes that can be recognized by different antiviral immune systems.

### Recognition Sites for Restriction–Modification Systems

Bacterial restriction–modification (R-M) systems target recognition sites on double stranded DNA molecules that are generally composed of a 4–8 bp palindromic sequence. R-M systems consist of two enzymes: a restriction endonuclease (REase) and a methyltransferase (MTase). The REase cleaves the DNA at the recognition site, creating a double strand break. During bacterial DNA replication, the MTase methylates cytosine and adenine bases at the same recognition site, protecting it from cleavage by the REase. Through the combined action of the MTase and the REase, R-M systems can discriminate between host and foreign DNA containing recognition sites, and consequently cleave only the foreign DNA (Tock and Dryden 2005).

The biological consequences of recognition sites have been widely studied in bacteriophages, because they are the primary target of REases. The increasing availability of phage genomes from the 1980s onward has allowed testing for the avoidance of recognition sites that could be cleaved by the REases of their hosts (e.g., Sharp 1986; Karlin et al. 1992; Blaisdell et al. 1996; Rocha 2001; Rusinov et al. 2018). Indeed, in many phages, there seems to be selection for eliminating recognition sites that could be targeted by their host, resulting in a significant avoidance of these motifs (Sharp 1986). However, this strategy of avoiding host immune defences does not seem to be universal among phages, and three general factors have been identified that influence the occurrence of recognition site avoidance. First, recognition site avoidance is strongly dependent on the genetic material of the phage: dsDNA and ssDNA phages often avoid recognition site motifs, while RNA phages do not (Rocha 2001; Rusinov et al. 2018). This pattern is expected, as RNA phages are not targeted by REases, which only act on double stranded DNA. Although ssDNA phages are also resistant to restriction during their infective stage, they go through a double stranded stage during replication within the host, providing a window for REase attack and thus for selection to act against recognition site motifs. Second, the occurrence of restriction site avoidance depends on the type of R-M system: avoidance is often observed for recognition sites targeted by orthodox Type II R-M systems, but usually not for recognition sites of Type I and Type III R-M systems (Sharp 1986; Rusinov et al. 2018). There

are several explanations for this observation. In Type II systems, the REase and the MTase are independent enzymes with separate DNA recognition domains, while Type I and Type III systems function as hetero-oligomeric complexes with a single sequence recognition domain (Tock and Dryden 2005). Sharing of recognition domains between R and M factors makes it easier to change the specificity of Type I and Type III systems than that of Type II systems. This instigates a phage-bacteria arms-race with rapid changes in the specificity of host defence, rendering recognition site avoidance a less efficient strategy for long-term avoidance of host immune defence using Type I or Type III R-M systems (Rusinov et al. 2018). Several phages are also known to produce universal antirestriction proteins that can inhibit the action of Type I or Type III R-M systems, and are thus protected against restriction even when recognition sites are present in their genome (e.g., SAMase in phage T3, Karlin et al. 1992). Due to the high diversity in Type II R-M systems, such a universal defence could be more difficult to establish (Rusinov et al. 2018). Type I and Type III systems also often require two recognition sites to be present on opposing strands, so avoidance can additionally be achieved by removing a recognition site from only one strand (Tock and Dryden 2005). Third, bacteriophage lifestyle also seems to be a determining factor for the strength of selection against recognition sites, with lytic phages showing a higher degree of recognition site avoidance than temperate phages (Sharp 1986; Karlin et al. 1992; Rocha 2001; Rusinov et al. 2018), probably because temperate phages integrate into the genome of the host where their DNA will be methylated and thereby escape restriction.

Pleška and Guet (2017) provided direct experimental support for the phenotypic effect of synonymous mutation through recognition site changes in bacteriophage $\lambda$ cI857, a conditionally lytic phage of *E. coli*. This phage contains five EcoRI restriction sites, into which synonymous mutations were introduced. They observed that all individual synonymous point mutations increased the likelihood of phage escape, although at a variable rate. The combination of five synonymous mutations, one in each restriction site, provided full escape from restriction by EcoRI. These experimental data represent direct evidence for strong phenotypic effects of synonymous mutations located in a restriction site.

Although the genomes of bacteria encoding R-M systems are assumed to be protected from self-restriction through methylation of recognition sites, several studies have found that many bacterial genomes also show significant recognition site avoidance (Karlin et al. 1992; Gelfand and Koonin 1997; Rocha 2001; Rusinov et al. 2018). This indicates that there is a substantial selective pressure on bacterial genomes to avoid recognition sites and prevent self-restriction. For example, the EcoRI recognition site is avoided in the *E. coli* genome (Gelfand and Koonin 1997). Pleška et al. (2016) experimentally demonstrated that the genomic DNA of *E. coli* is frequently cleaved by EcoRI, and this might be caused by differences in expression levels of the REase and MTase. By comparing the probability of escaping restriction and levels of selfrestriction by two restriction enzymes, Pleška et al. (2016) suggested a trade-off between the efficiency of defence against phages and selfrestriction, which can be mitigated by restriction site avoidance in the host genome.

### APOBEC3 Hotspots

APOBEC3 (apolipoprotein B mRNA-editing enzyme, catalytic subunit 3 or A3) enzymes belong to a family of mutagenic cytidine deaminases that transform cytidine to uracil in DNA or RNA. A3s participate in mammalian innate immunity against retrotransposons, exogenous viruses and endogenous viruses, in which they induce mutations that restrict their replication (Harris and Dudley 2015). A3s have a specific preferred deamination context, called a deamination "hotspot." For example, the 5′TC motif is a hotspot for A3B, while 5′CCC is a hotspot for A3G. Preferred motifs of a particular APOBEC can be changed through a small number of amino-acid changes in the hotspot recognition loop (Kohli et al. 2009), and the expanded A3 gene repertoire in mammals is assumed to be the result of gene duplication and diversification of preferred motifs in response to selective pressures from various viral infections (Münk et al. 2012).

The antiviral action of A3s has been found to exert a mutational and selective pressure on many viral genomes. Recent studies indicated an elevated C to U mutation rate in SARS-CoV2, which can be attributed to the action of A3 (Di Giorgio et al. 2020; Ratcliff and Simmonds 2021; Rice et al. 2021). Viral genomes also often exhibit a depletion of A3 hotspots (Warren et al. 2015; Chen and MacCarthy 2017; Martinez et al. 2019; Poulain et al. 2020). Such a depletion has been recorded in as many as 22% of all human viruses, and is most striking for 5′TC motifs that occupy the second and third position in a codon, where a deamination of the third codon position is always synonymous (Poulain et al. 2020). Furthermore, a high genomic GC content also provides protection against A3s because it tends to minimize the presence of hotspots (Chen and MacCarthy 2017). However, a complete avoidance of A3 hotspots is generally difficult to obtain, because it often requires multiple nonsynonymous mutations that would be detrimental to the virus (Martinez et al. 2019).

Depletion of A3 hotspots is only apparent in certain viral families, with members of the papillomaviruses, polyomaviruses, coronaviruses, and autonomous parvoviruses showing the strongest depletion (Verhalen et al. 2016; Warren et al. 2015; Poulain et al. 2020). This pattern could be caused by a higher A3 pressure on these viral families, either because they infect cell types with higher A3 expression levels, because they induce A3 expression in their host, or because they lack

proteins that inhibit A3 activity (Warren et al. 2015; Verhalen et al. 2016). HIV, for example, is highly susceptible to A3G, but can effectively avoid deamination by the production of the vif protein that neutralizes A3G, reducing the need for A3G motif avoidance (Harris and Dudley 2015).

Although the action of A3-induced hypermutation is expected to have predominantly inactivating effects on HIV-1 (Armitage et al. 2012), some studies found evidence that during early infection HIV-1 can sometimes benefit from A3-induced hypermutation (Wood et al. 2009; Monajemi et al. 2014; Sato et al. 2014). This benefit is caused by accelerated evolution and diversification of positions targeted by the adaptive immune system, allowing for a quick evasion from the initial immune response. There are indications for positive selection on several codon sites within A3 hotspots of the envelope gene of HIV-1 that diversify during the early stages of infection (Wood et al. 2009). Sato et al. (2014) furthermore experimentally showed that in HIV-1 vif mutants, the action of A3D/F can promote in vivo viral diversification leading to a conversion of coreceptor usage. It has been hypothesized that this could explain an observed enrichment of A3 hotspots in cytotoxic T-cell epitope encoding portions of the HIV genome (Monajemi et al. 2014), but it remains unclear how selection for deaminated hotspots during early infection is counteracted by selection for unmodified hotspots during viral transmission.

### CpG and UpA Dinucleotides

Frequencies of CpG and UpA dinucleotides are often significantly depleted in both vertebrate and plant RNA viruses (Cheng et al. 2013; Simmonds et al. 2013; Ibrahim et al. 2019; Xia 2020). This depletion can be partially caused by the viral genome mirroring the nucleotide composition of the host mRNA, which avoids CpG and UpA for reasons other than interactions with antiviral immune systems (Beutler et al. 1989). However, experimental evidence suggests that plant- and vertebrate RNA viruses are additionally subjected to a selective pressure for CpG and UpA avoidance imposed by the host's antiviral immunity. Artificially increasing CpG and UpA dinucleotides, through synonymous mutations in protein coding genes or mutations in noncoding regions, was shown to strongly decrease replication in a large variety of viruses such as poliovirus (Burns et al. 2009), Influenza A (Gaunt et al. 2016), HIV-1 (Takata et al. 2017), the human enteric echovirus 7 (Fros et al. 2017), the potato virus Y (Ibrahim et al. 2019), and Zika virus (Trus et al. 2020). Fros et al. (2017) furthermore inferred that this effect was not caused by a lower translation efficiency due to changes in codon usage, thus suggesting the action of an intrinsic defence pathway present in the host cells acting on CpG and UpA dinucleotides. Takata et al. (2017) partially confirmed this by showing that the zinc-finger antiviral protein (ZAP) is involved in inhibiting virion production through targeting CpG dinucleotides in the RNA of HIV-1. Based on these findings, Xia (2020) proposed that the extreme CpG deficiency in SARS-CoV-2 could contribute to its high virulence in humans by allowing it to successfully avoid ZAP-mediated antiviral immunity. The immune pathways targeting CpG and UpA dinucleotides of plant viruses have not been elucidated, but analogous processes to those in vertebrates might also operate in plants (Ibrahim et al. 2019).

## Conclusions and Perspectives

We have reviewed a number of biological mechanisms that are likely to exert selection pressure on local codon usage for reasons other than selection for translation accuracy and efficiency. In the light of these different elements, selection on codon usage appears to be a combination of a global selection pressure imposed by the translation machinery, and a patchwork of local selection pressures linked to the enrichment or avoidance of specific nucleotide sequences that contain biological signals. However, contrary to the translational selection, the local, nontranslational selection pressures do not apply to all genomes, as some are specific to viruses or to prokaryotes (see fig. 1 for an overview). It is also important to realize that some sequence patterns could be subject to multiple selection pressures. For example, a palindromic sequence could be under selection both because it is the preferred insertion site for certain Transposable Elements (TEs) (see Box 1) and also because it is a restriction site. Specific selection pressures can therefore not be simply deduced by finding that a specific pattern is avoided or enriched in a genome, or a part of the genome. Knowledge of the evolutionary history of the species is generally necessary to make inferences about selective pressures (e.g., associations with specific TEs, specific restriction enzymes encoded and levels of selfrestriction). Additionally, for most mechanisms reviewed (except R-M motifs and CpG/UpA motifs), there are reports of both avoidance and enrichment of the same motif or of positive and negative effects on fitness of the addition or removal of these motifs. In these cases, the direction of selection is determined by factors that range from environmental conditions to surrounding sequences. Testing for avoidance or enrichment at a scale at which both might occur can lead to negative results or to errors in the estimation of the strength of the selection pressure. Finally, for all motifs, avoidance or enrichment patterns can be obtained through both synonymous and nonsynonymous mutations, but synonymous mutations are generally expected to have lower direct fitness effects and for this reason represent a priori a preferred way of avoiding or enriching specific patterns. Yet, when an avoidance or enrichment is observed, it cannot be excluded that nonsynonymous mutations contributed to this pattern.

From a methodological point of view, the detection of over- or under-representation of a particular sequence motif in a genome is often not a trivial task, and is an important issue in computational biology. This detection requires an appropriate model of the genome that assumes the absence of a selective pressure on the sequence motif to which observed frequencies can be compared. A wide range of methods have been developed for this task, including simple estimations using the product of nucleotide or k-mer frequencies and approaches using Markov models (see e.g., Rusinov et al. 2018 for a comparison of methods). Given these methodological difficulties, several authors have noted that some observations of sequence motif avoidance or enrichment are inconclusive and can be artifacts of an erroneous methodology (Sharp 1986; Morgens et al. 2013). It is also a well-known problem that the inference of selection on codon usage by comparative sequence analysis can be confounded by mutational bias, as both processes can produce similar motif enrichment/avoidance and codon usage patterns (Laurin-Lemay et al. 2018). Mutation biases can affect codon usage on both a genome-wide and a local scale (Duret 2002). Disentangling the effects of selection and mutational bias on codon usage is thus not an easy task, and is still a subject of much debate (Galtier et al. 2018; Laurin-Lemay et al. 2018). Along the same lines, inference of selection on codon usage can be erroneous because factors such as amino acid usage bias or gene expression are not considered. For example, it was assumed that translational inefficient codons are selected at the 5′ end of bacterial signal peptides because they can facilitate protein secretion (Power et al. 2004). However, Cope et al. (2018) refuted this hypothesis by showing that the 5′ end of bacterial signal peptides show no differences in CUB compared to cytoplasmic proteins after correcting for amino acid usage and gene expression. In the studies cited in the present review, selection is often inferred based on deviations from genome-wide nucleotide or k-mer frequencies. However, these generally do not account for context-dependent mutational biases or amino acid usage (although see e.g., Wood et al. 2009 accounting for mutational hotspots). The usage of more elaborate models accounting for multiple confounding factors could thus nuance the assumption of selection when observing avoidance or enrichment of a particular sequence motif. Ideally, the fitness effects of synonymous mutations are empirically determined to provide unequivocal evidence for selective pressures on these synonymous positions (Pleška and Guet 2017).

Patterns of avoidance or enrichment in specific motifs or codons are thus not necessarily the product of selection. Conversely, the existence of selection for or against a motif does not necessarily result in the enrichment or avoidance of this motif because it depends on the selection coefficients and the effective population size. For translational selection, selection coefficients on synonymous mutations are generally assumed to be weak (Sharp and Li 1986) and translational selection is only expected to shape codon usage when the effective population size is large enough so that selection can overcome drift, as stated by the nearly neutral theory (Ohta and Gillespie 1996). Consequently, translational selection is assumed to shape the codon usage of species with large effective population sizes, such as many microorganisms and some invertebrate animals, but not (or less) in species with a small effective population size such as larger mammals (Galtier et al. 2018). For nontranslational selection on codon usage, selection coefficients are generally unknown, but they probably vary widely between selective pressures and synonymous sites (e.g., selection against near-stop codons might be weak while selection on avoiding sequence motifs targeted by antiviral immune systems might be stronger). To estimate the potential impact of nontranslational selective pressures on the codon usage of a particular species, both the selection coefficient acting on synonymous variation and the effective population size of the species will need to be considered. However, sometimes extrapolation might not be so straightforward as selection coefficients on synonymous variation might be indirectly affected by the effective population size (Wu and Hurst 2015). Future studies investigating the importance of nontranslational selective pressures for shaping codon usage in a wide variety of organisms will be of particular interest to address this issue.

Selection on codon usage thus appears as a complex phenomenon composed of a mix of global and local pressures. The local pressures are both diverse and specific to certain genome groups, the level of evidence of their existence also varies and it is very likely that some "other codes" of DNA have yet to be uncovered. For example, all the elements for selection against or for the presence of preferred target sequences for TEs are present (see Box 1), but to our knowledge, these patterns and the potential effects on selection and evolution of local codon usage have not yet been investigated. To get a complete and accurate picture of the patchwork of local selective pressures on codon usage and its evolution, more work is required to rigorously identify their molecular signature, to experimentally measure the fitness effects of synonymous mutations in the identified patterns, and to test new hypotheses.

## Text box 1.

### Is Local Codon Usage Influenced by Transposable Elements?

Transposable elements (TEs) are DNA sequences that have the ability to change their position (i.e., to transpose) within or between genomes. TEs are widely spread across all eukaryotic and prokaryotic genomes, and their effects on genome structure and organism fitness are manifold (see Bourque et al. 2018 for a review): 1) TEs increase genome size by accumulating in genomes (Naville et al. 2019). 2) They create new recombination sites and thereby induce chromosome rearrangements (Lönnig and Saedler 2002). 3) They enhance the expression of genes, for example, by introducing new *cis*-regulatory elements in their neighborhood (Salces-Ortiz et al. 2020). (iv) They are a source of novel mutations: either by disrupting the expression of the genes they integrate into, or by introducing new genes (Jangam et al. 2017). Thus, the phenotypic changes induced by TEs range from adaptive (Salces-Ortiz et al. 2020) to lethal (Tsugeki et al. 1996). The sign and amplitude of the fitness effect depends mainly on the TE content and on its insertion site.

Many TE families show strong preferences for their insertion sites (Levin and Moran 2011), but some have dispersed integration patterns, and exhibit low or no preference, for example, ~500,000 copies of the L1 retroelement can be found throughout the human genome. For TEs showing an integration site preference, a precise nucleotide pattern is often required, for example the conserved 60 bp attnTn7 sequence required for the integration of Tn7 in bacterial chromosomes (Kuduvalli 2001; Parks and Peters 2007). The preferred integration site can also be a shorter, less conserved palindromic sequence, as for example the 6 bp motif where Tn10 preferentially inserts (Halling and Kleckner 1982). Other TE families show preferences for certain parts of the genome: some integrate in gene-rich regions but avoid coding regions, for example, *Drosophila* P element often integrates 500 bp upwards of transcription start sites (Bellen et al. 2011) and others integrate specifically in heterochromatin and other weakly expressed regions, for example, in *Saccharomyces cerevisiae*, 90% of Ty5 integration events occur in heterochromatin at telomeres (Zou and Voytas 1997). In many cases, the likelihood of transposition to a site mostly depends on DNA mechanical properties: namely DNA deformability, curvature, and melting (see Arinkin et al. 2019 for a review). Unwinding and bending of DNA allows precise cleavage of the target site, and renders integration irreversible (Morris et al. 2016; Ru et al. 2018). DNA melting allows the conjugative transposons to easily recombine with many insertion sites regardless of homology (Rubio-Cosials et al. 2018). Even when recognition by the transposase requires a few precise invariant base pairs (e.g., several DDE transposases require invariant T/A nucleotides in the sequence in order to integrate), DNA helix flexibility may be necessary to allow recognition and integration through base-flipping and formation of a base-specific contact zone with the transposase (Morris et al. 2016). Structural properties of DNA directly depend on sequence composition. GC content decreases thermostability and bendability but increases DNA curvature (Vinogradov 2003). The deformability of TE integration sites is suggested to be linked to their palindromicity, to their enrichment in T/A pairs (Arinkin et al. 2019) and in pyrimidine-purine base steps (Maskell et al. 2015; Morris et al. 2016).

The codon usage of transposable elements and the evolutionary forces shaping it have been investigated and debated (Lerat et al. 2002; Jia and Xue 2009; Southworth et al. 2019). It is also well established that the observed distribution of TEs in genomes is the result of both TE integration preferences and selection against the integration of TEs at certain loci (Sultana et al. 2017). However, to our knowledge, selection pressure on DNA motifs preferred for TE insertion, the resulting avoidance or enrichment and the potential impact on local codon usage has not been studied. However, by combining knowledge on TE insertion fitness effects and on the nature of preferred insertion sites, predictions can be derived. Local codon usage is likely to be a determinant of the local abundance of TE integration sites, either because synonymous versions of local sequences differ in their content of sequence-specific integration sites or palindromes, or because nucleotide sequence determines DNA mechanical properties (Olson et al. 1998) which favor or disfavor TE integration. Synonymous polymorphisms that increase the likelihood of TE integration will be less fit and purged from the population. This would give rise to a local codon usage preference that reduces the number of insertion motifs in coding regions. This evolutionary scenario should be most prevalent when fitness is highly correlated with gene expression, that is, in organisms with few redundant genes and/or a fast life cycle, and this selection for avoidance of integration sites should also be stronger for essential genes.

TE insertions can also have positive fitness effects, as adaptation to novel environments can be achieved by loss-of-function mutations, particularly in bacteria (reviewed in Hottes et al. 2013). In fluctuating environments, it might be advantageous to have the capacity to remobilize previously lost gene functions. In this context, we could imagine that gene expression could switch between "off" and "on" states through the integration/excision of nonreplicative TEs

(e.g., via cut-and-paste transposition mechanism). Local codon usage preference could thus be under selection to increase the likelihood of transposon integration in these genes. Both predictions for enrichment and avoidance of TE integration sites can be tested by comparing the frequency of TE integration sites in different gene categories. Predictions for enrichment can additionally be tested by analyzing whole genome sequencing data from experimental evolution studies involving stressful conditionsfluctuating over an extended period.

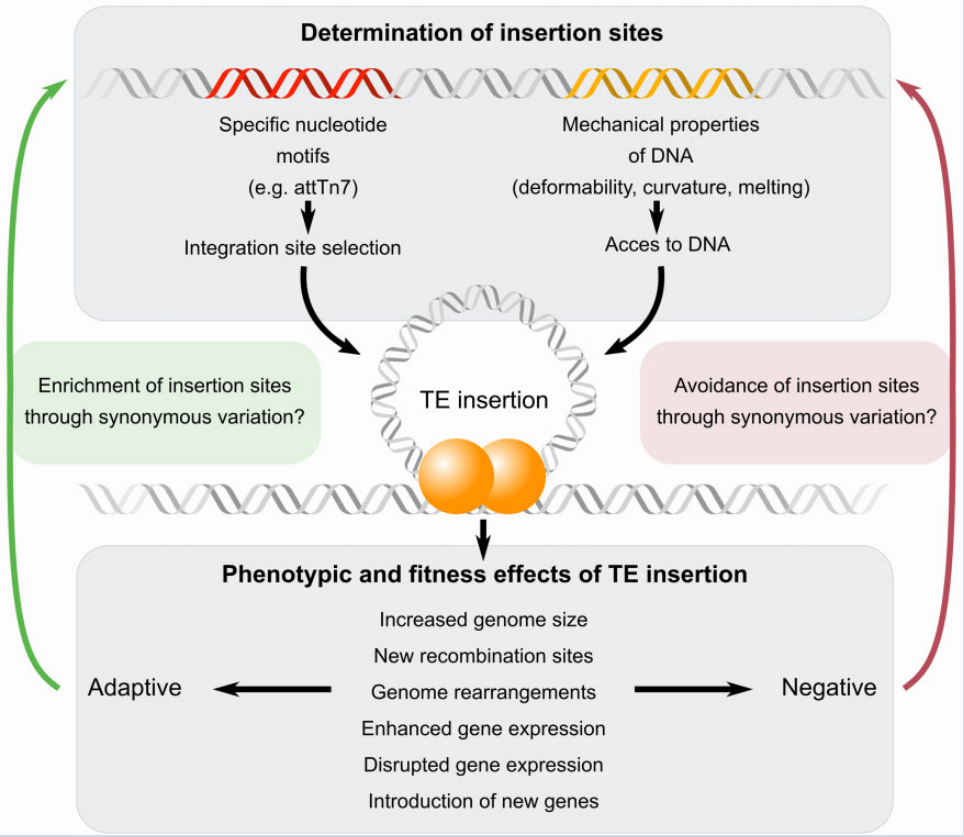


FIG. —How could transposable elements exert local selection pressures on codon usage?

## Acknowledgments

This work was supported by an ERC grant (HGTCODONUSE grant number 682819) to S.B.

## Data Availability

This paper does not include new data.

## Literature Cited

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

Abramowicz A, Gos M. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 59:253–268.

Ahnert SE, Fink TMA, Zinovyev A. 2008. How much non-coding DNA do eukaryotes require? J Theor Biol. 252:587–592.

Ando H, Miyoshi-Akiyama T, Watanabe S, Kirikae T. 2014. A silent mutation in *mabA* confers isoniazid resistance on *Mycobacterium tuberculosis*: *mabA* mutation confers INH resistance on *Mtb*. Mol Microbiol. 91:538–547.

Arinkin V, Smyshlyaev G, Barabas O. 2019. Jump ahead with a twist: DNA acrobatics drive transposition forward. Curr Opin Struct Biol. 59:168–177.

Armitage AE, et al. 2012. APOBEC3G-induced hypermutation of human immunodeficiency virus type-1 is typically a discrete "all or nothing" phenomenon. PLoS Genet. 8:e1002550.

Bellen HJ, et al. 2011. The *Drosophila* gene disruption project: progress using transposons with distinctive site specificities. Genetics 188:731–743.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 9:675.

Bergman S, Tuller T. 2020. Widespread non-modular overlapping codes in the coding regions. Phys Biol. 17:031002.
Bertrand RL, Abdel-Hameed M, Sorensen JL. 2015. Limitations of the 'ambush hypothesis' at the single-gene scale: what codon biases are to blame? Mol Genet Genomics. 290(2):493–504.
Beutler E, Gelbart T, Han JH, Koziol JA, Beutler B. 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc Natl Acad Sci U S A. 86:192–196.
Blaisdell BE, Campbell AM, Karlin S. 1996. Similarities and dissimilarities of phage genomes. Proc Natl Acad Sci U S A. 93:5854–5859.
Boël G, et al. 2016. Codon influence on protein expression in *E. coli* correlates with mRNA levels. Nature 529:358–363.
Bourque G, et al. 2018. Ten things you should know about transposable elements. Genome Biol. 19(1):199.
Brantl S. 2007. Regulatory mechanisms employed by cis-encoded antisense RNAs. Curr Opin Microbiol. 10:102–109.
Brophy JAN, Voigt CA. 2016. Antisense transcription as a tool to tune gene expression. Mol Syst Biol. 12:854.
Browning DF, Busby SJW. 2004. The regulation of bacterial transcription initiation. Nat Rev Microbiol. 2:57–65.
Brule CE, Grayhack EJ. 2017. Synonymous codons: choose wisely for expression. Trends Genet. 33:283–297.
Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.
Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A. 89(4):1358–1362.
Burns CC, et al. 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol. 83:9957–9969.
Burow DA, et al. 2018. Attenuated codon optimality contributes to neural-specific mRNA decay in Drosophila. Cell Rep. 24:1704–1712.
Chaney JL, et al. 2017. Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput Biol. 13:e1005531.
Chen J, et al. 2014. Dynamic pathways of −1 translational frameshifting. Nature 512:328–332.
Chen J, MacCarthy T. 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Comput. Biol. 13:e1005471.
Cheng X, et al. 2013. CpG usage in RNA viruses: data and hypotheses. PLoS ONE 8:e74109.
Cohen O, et al. 2016. Comparative transcriptomics across the prokaryotic tree of life. Nucleic Acids Res. 44:W46–53.
Coleman JR, et al. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320:1784–1787.
Cope AL, Hettich RL, Gilchrist MA. 2018. Quantifying codon usage in signal peptides: gene expression and amino acid usage explain apparent selection for inefficient codons. Biochim Biophys Acta BBA - Biomembr. 1860:2479–2485.
Courel M, et al. 2019. GC content shapes mRNA storage and decay in human cells. eLife 8:e49708.
Devaraj A, Fredrick K. 2010. Short spacing between the Shine-Dalgarno sequence and P codon destabilizes codon-anticodon pairing in the P site to promote +1 programmed frameshifting: ribosomal frameshifting. Mol Microbiol. 78:1500–1509.
Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 6:eabb5813.
Diwan GD, Agashe D. 2016. The frequency of internal Shine–Dalgarno-like motifs in prokaryotes. Genome Biol Evol. 8:1722–1733.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715–724.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.
Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16(7):287–289.
Eyre-Walker A. 1996. The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. J Mol Evol. 42:73–78.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2:e268.
Fluman N, Navon S, Bibi E, Pilpel Y. 2014. mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. eLife 3:e03440.
Freistroffer DV, Kwiatkowski M, Buckingham RH, Ehrenberg M. 2000. The accuracy of codon recognition by polypeptide release factors. Proc Natl Acad Sci U S A. 97:2046–2051.
Fros JJ, et al. 2017. CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. eLife 6:e29112.
Frumkin I, et al. 2017. Gene architectures that minimize cost of gene expression. Mol Cell. 65:142–153.
Galtier N, et al. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol. 35:1092–1103.
Gaunt E, et al. 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. eLife 5:e12735.
Gelfand MS, Koonin EV. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucleic Acids Res. 25:2430–2439.
Gophna U. 2018. The unbearable ease of expression—how avoidance of spurious transcription can shape G+C content in bacterial genomes. FEMS Microbiol Lett. 365:fny26. doi: 10.1093/femsle/fny267.
Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9(1):213–213.
Hahn MW. 2003. The effects of selection against spurious transcription factor binding sites. Mol Biol Evol. 20:901–906.
Halling SM, Kleckner N. 1982. A symmetrical six-base-pair target site sequence determines Tn10 insertion specificity. Cell 28:155–163.
Harris RS, Dudley JP. 2015. APOBECs and virus restriction. Virology 479-480:131–145
Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.
Hockenberry AJ, Jewett MC, Amaral LAN, Wilke CO. 2018. Within-gene Shine–Dalgarno sequences are not selected for function. Mol Biol Evol. 35:2487–2498.
Hottes AK, et al. 2013. Bacterial adaptation through loss of function. PLoS Genet. 9:e1003617.
Huvet M, Stumpf MP. 2014. Overlapping genes: a window on gene evolvability. BMC Genomics. 15:721.
Ibrahim A, et al. 2019. A functional investigation of the suppression of CpG and UpA dinucleotide frequencies in plant RNA virus genomes. Sci Rep. 9:18359.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.
Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 17:405–412.
Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. Genome Res. 20:1582–1589.

Jangam D, Feschotte C, Betrán E. 2017. Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet. 33:817–831.
Jia J, Xue Q. 2009. Codon usage biases of transposable elements and host nuclear genes in *Arabidopsis thaliana* and *Oryza sativa*. Genomics Proteomics Bioinformatics. 7(4):175–184.
Johnson LJ, et al. 2011. Stops making sense: translational trade-offs and stop codon reassignment. BMC Evol Biol. 11:227.
Karlin S, Burge C, Campbell AM. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. 20:1363–1370.
Katz L. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 13:2042–2051.
Kershner JP, et al. 2016. A synonymous mutation upstream of the gene encoding a weak-link enzyme causes an ultrasensitive response in growth rate. J Bacteriol. 198:2853–2863.
Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A. 107:9158–9163.
Kohli RM, et al. 2009. A portable hot spot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. J Biol Chem. 284:22898–22904.
Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem. 289:30334–30342.
Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255–258.
Kuduvalli PN. 2001. Target DNA structure plays a critical role in Tn7 transposition. EMBO J. 20:924–932.
Kurland CG. 1992. Translational accuracy and the fitness of bacteria. Annu Rev Genet. 26:29–50.
Lamberte LE, et al. 2017. Horizontally acquired AT-rich genes in *Escherichia coli* cause toxicity by sequestering RNA polymerase. Nat Microbiol. 2:16249.
Laurin-Lemay S, Philippe H, Rodrigue N. 2018. Multiple factors confounding phylogenetic detection of selection on codon usage. Mol. Biol. Evol. 35:1463–1472.
Lerat E, Capy P, Biémont C. 2002. Codon usage by transposable elements and their host genes in five species. J Mol Evol. 54:625–637.
Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 12:615–627.
Li G-W, Oh E, Weissman JS. 2012. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484:538–541.
Lönnig W-E, Saedler H. 2002. Chromosome rearrangements and transposable elements. Annu Rev Genet. 36:389–410.
Martinez T, Shapiro M, Bhaduri-McIntosh S, MacCarthy T. 2019. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. Virus Evol. 5:vey040. doi: 10.1093/ve/vey040.
Maskell DP, et al. 2015. Structural basis for retroviral integration into nucleosomes. Nature 523:366–369.
Monajemi M, et al. 2014. Positioning of APOBEC3G/F mutational hotspots in the human immunodeficiency virus genome favors reduced recognition by CD8+ T cells. PLoS One 9:e93428.
Morgens DW, Chang CH, Cavalcanti AR. 2013. Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. BMC Genomics. 14:418.
Morris ER, Grey H, McKenzie G, Jones AC, Richardson JM. 2016. A bend, flip and trap mechanism for transposon integration. eLife 5:e15537.
Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. 2015. The silent sway of splicing by synonymous substitutions. J Biol Chem. 290:27700–27711.
Münk C, Willemsen A, Bravo IG. 2012. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol. 12:71.
Naville M, et al. 2019. Massive changes of genome size driven by expansions of non-autonomous transposable elements. Curr Biol. 29:1161–1168.e6.
Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theor Popul Biol. 49:128–142.
Olson WK, Gorin AA, Lu X-J, Hock LM, Zhurkin VB. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A. 95:11163–11168.
Omotajo D, Tate T, Cho H, Choudhary M. 2015. Distribution and diversity of ribosome binding sites in prokaryotic genomes. BMC Genomics. 16:604.
Osterman IA, et al. 2020. Translation at first sight: the influence of leading codons. Nucleic Acids Res. 48:6931–6942.
Pallejà A, García-Vallvé S, Romeu A. 2009. Adaptation of the short intergenic spacers between co-directional genes to the Shine-Dalgarno motif among prokaryote genomes. BMC Genomics. 10:537.
Parks AR, Peters JE. 2007. Transposon Tn7 is widespread in diverse bacteria and forms genomic islands. J. Bacteriol. 189:2170–2173.
Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 20:237–243.
Pleška M, et al. 2016. Bacterial autoimmunity due to a restriction-modification system. Curr Biol 26:404–409.
Pleška M, Guet CC. 2017. Effects of mutations in phage restriction sites during escape from restriction–modification. Biol Lett. 13(12):20170646.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.
Poulain F, Lejeune N, Willemart K, Gillet NA. 2020. Footprint of the host restriction factors APOBEC3 on the genome of human viruses. PLoS Pathog. 16:e1008718
Power PM, Jones RA, Beacham IR, Bucholtz C, Jennings MP. 2004. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of *Escherichia coli*. Biochem Biophys Res Commun. 322:1038–1044.
Radhakrishnan A, et al. 2016. The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. Cell 67:122–132.e9.
Ratcliff J, Simmonds P. 2021. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. Virology 556:62–72.
Rice AM, et al. 2021. Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol. 38:67–83.
Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14:2279–2286.
Rocha EPC. 2001. Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. Genome Res. 11:946–958.
Ru H, et al. 2018. DNA melting initiates the RAG catalytic pathway. Nat Struct Mol Biol. 25:732–742.
Rubio-Cosials A, et al. 2018. Transposase-DNA complex structures reveal mechanisms for conjugative transposition of antibiotic resistance. Cell 173:208–220.e20.
Rusinov IS, Ershova AS, Karyagina AS, Spirin SA, Alexeevski AV. 2018. Avoidance of recognition sites of restriction-modification systems is

a widespread but not universal anti-restriction strategy of prokaryotic viruses. BMC Genomics. 19:885.
Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. 2020. Transposable elements contribute to the genomic response to insecticides in *Drosophila melanogaster*. Philos. Trans. R. Soc. B Biol. Sci. 375:20190341.
Sato K, et al. 2014. APOBEC3D and APOBEC3F potently promote HIV-1 diversification and evolution in humanized mouse model. PLoS Pathog. 10:e1004453.
Savisaar R, Hurst LD. 2017. Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. Mol Biol Evol. 34:1110–1126.
Schrader JM, et al. 2014. The coding and noncoding architecture of the *Caulobacter crescentus* genome. PLoS Genet. 10:e1004463.
Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol. 23:701–705.
Shah P, Gilchrist MA. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc Natl Acad Sci U S A. 108:10231–10236.
Sharp P. 1986. Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes. Mol Biol Evol. 3:75–83.
Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 24:28–38.
Shine J, Dalgarno L. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A. 71:1342–1346.
Simmonds P, Xia W, Baillie J, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics. 14:610.
Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: evidences for phylogenetic trends. Comput Biol Chem. 33:239–244.
Southworth J, Grace CA, Marron AO, Fatima N, Carr M. 2019. A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage. Mob DNA. 10:44.
Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res. 21:1563–1571.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol. 24:374–381.
Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 18:292–308.
Takata MA, et al. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124–127.
Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol. 18:18–30.
Tock MR, Dryden DT. 2005. The biology of restriction and anti-restriction. Curr Opin Microbiol. 8:466–472.
Trus I, et al. 2020. CpG-recoding in zika virus genome causes host-age-dependent attenuation of infection with protection against lethal heterologous challenge in mice. Front Immunol. 10:3077.
Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y. 2010. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics. 11:491.
Tsugeki R, Kochieva EZ, Fedoroff NV. 1996. A transposon insertion in the Arabidopsis SSR16 gene causes an embryo-defective lethal mutation. Plant J. 10:479–489.
Tuller T, Carmi A, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 107:3645–3650.
Urtecho G, et al. 2020. Genome-wide functional characterization of *Escherichia coli* promoters and regulatory elements responsible for their function. bioRxiv. doi: 10.1101/2020.01.04.894907.
Verhalen B, Starrett GJ, Harris RS, Jiang M. 2016. Functional upregulation of the DNA cytosine deaminase APOBEC3B by polyomaviruses. J Virol. 90:6379–6386.
Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. Nucleic Acids Res. 31:1838–1844.
Warren CJ, Van Doorslaer K, Pandey A, Espinosa JM, Pyeon D. 2015. Role of the host restriction factor APOBEC3 on papillomavirus evolution. Virus Evol. 1(1):vev015.
Whitaker WR, Lee H, Arkin AP, Dueber JE. 2015. Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. ACS Synth Biol. 4:249–257.
Wood N, et al. 2009. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. 5:e1000414.
Wu X, Hurst LD. 2015. Why selection might be stronger when populations are small: intron size and density predict within and between-species usage of exonic splice associated *cis*-motifs. Mol Biol Evol. 32:1847–1861.
Xia X. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Mol Biol Evol. 37:2699–2705
Yan X, Hoek TA, Vale RD, Tanenbaum ME. 2016. Dynamics of translation of single mRNA molecules in vivo. Cell 165:976–989.
Yang Q, et al. 2019. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. Nucleic Acids Res. 47:9243–9258.
Yona AH, Alm EJ, Gore J. 2018. Random sequences rapidly evolve into de novo promoters. Nat Commun. 9:1530.
Yu C-H, et al. 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell. 59:744–754.
Zhao F, Yu C-H, Liu Y. 2017. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res. 45:8484–8492.
Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. 2018. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. eLife 7:e33569. doi: 10.7554/eLife.33569.
Zou S, Voytas DF. 1997. Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5. Proc Natl Acad Sci U S A. 94:7412–7416.

**Associate editor:** Paul Sharp