



HAL
open science

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

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

► **To cite this version:**

Martijn Callens, Léa Pradier, Michael Finnegan, Caroline Rose, Stéphanie Bedhomme. Read between the Lines: Diversity of Nontranslational Selection Pressures on Local Codon Usage. *Genome Biology and Evolution*, 2021, 13 (9), 10.1093/gbe/evab097 . hal-03421370

HAL Id: hal-03421370

<https://hal.science/hal-03421370>

Submitted on 19 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 *Review for the GBE special section on “Phenotypic presentation of codon usage*
2 *preferences”*

3

4 **Title:** Read between the lines: Diversity of non-translational selection pressures on local
5 codon usage.

6

7 **Authors:** Martijn Callens¹, Léa Pradier¹, Michael Finnegan¹, Caroline Rose¹, Stéphanie
8 Bedhomme^{1,*}.

9 ¹ Centre d’Ecologie Fonctionnelle et Evolutive, CNRS, Université de Montpellier, Université
10 Paul Valéry Montpellier 3, Ecole Pratique des Hautes Etudes, Institut de Recherche pour le
11 Développement, 34000 Montpellier, France

12 *Author for Correspondence: Stéphanie Bedhomme, Centre d’Ecologie Fonctionnelle et
13 Evolutive, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, Ecole
14 Pratique des Hautes Etudes, Institut de Recherche pour le Développement, 34000
15 Montpellier, France

16 Telephone number: 00 33 4 67 61 32 11

17 Fax number: 00 33 4 67 61 33 36

18 Email address: stephanie.bedhomme@cefe.cnrs.fr

19 Abstract

20 Protein coding genes can contain specific motifs within their nucleotide sequence that
21 function as a signal for various biological pathways. The presence of such sequence motifs
22 within a gene can have beneficial or detrimental effects on the phenotype and fitness of an
23 organism, and this can lead to the enrichment or avoidance of this sequence motif. The
24 degeneracy of the genetic code allows for the existence of alternative synonymous
25 sequences that exclude or include these motifs, while keeping the encoded amino acid
26 sequence intact. This implies that locally, there can be a selective pressure for preferentially
27 using a codon over its synonymous alternative in order to avoid or enrich a specific sequence
28 motif. This selective pressure could –in addition to mutation, drift and selection for translation
29 efficiency and accuracy– contribute to shape the codon usage bias.

30 In this review, we discuss patterns of avoidance of (or enrichment for) the various biological
31 signals contained in specific nucleotide sequence motifs: transcription and translation
32 initiation and termination signals, mRNA maturation signals, and antiviral immune system
33 targets. Experimental data on the phenotypic or fitness effects of synonymous mutations in
34 these sequence motifs confirm that they can be targets of local selection pressures on codon
35 usage. We also formulate the hypothesis that transposable elements could have a similar
36 impact on codon usage through their preferred integration sequences.

37 Overall, selection on codon usage appears to be a combination of a global selection
38 pressure imposed by the translation machinery, and a patchwork of local selection pressures
39 related to biological signals contained in specific sequence motifs.

40

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

43 Significance statement

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

52 Introduction

53 The redundancy of the genetic code is a consequence of the existence of
54 synonymous codons, which differ by their nucleotide triplets but code for the same amino
55 acid. The different codons within a synonymous codon family are not used at equal
56 frequencies; this codon usage bias (CUB) can vary between species and between genes
57 within a species (Grantham et al. 1981; Ikemura 1985). CUB is shaped by mutation,
58 selection and drift (Bulmer 1991; Hershberg & Petrov 2008; Plotkin & Kudla 2011; Shah &
59 Gilchrist 2011). Selection on CUB is generally assumed to be driven by its effects on
60 translation efficiency (Tuller, Waldman, et al. 2010) and accuracy (Kurland 1992; Stoletzki &
61 Eyre-Walker 2007), mediated by the co-evolution of translation machinery and CUB: an
62 association between the frequency of use of a codon and the availability of the
63 corresponding decoding tRNA has been established for various genomes (e.g. Duret 2000;
64 Rocha 2004). Codon usage has been shown to modulate the rate and efficiency of
65 translation, with examples ranging from decreases in viral capsid protein production leading
66 to virus attenuation (Coleman et al. 2008) to 58% translation elongation rate increases in
67 human cell lines (Yan et al. 2016).

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

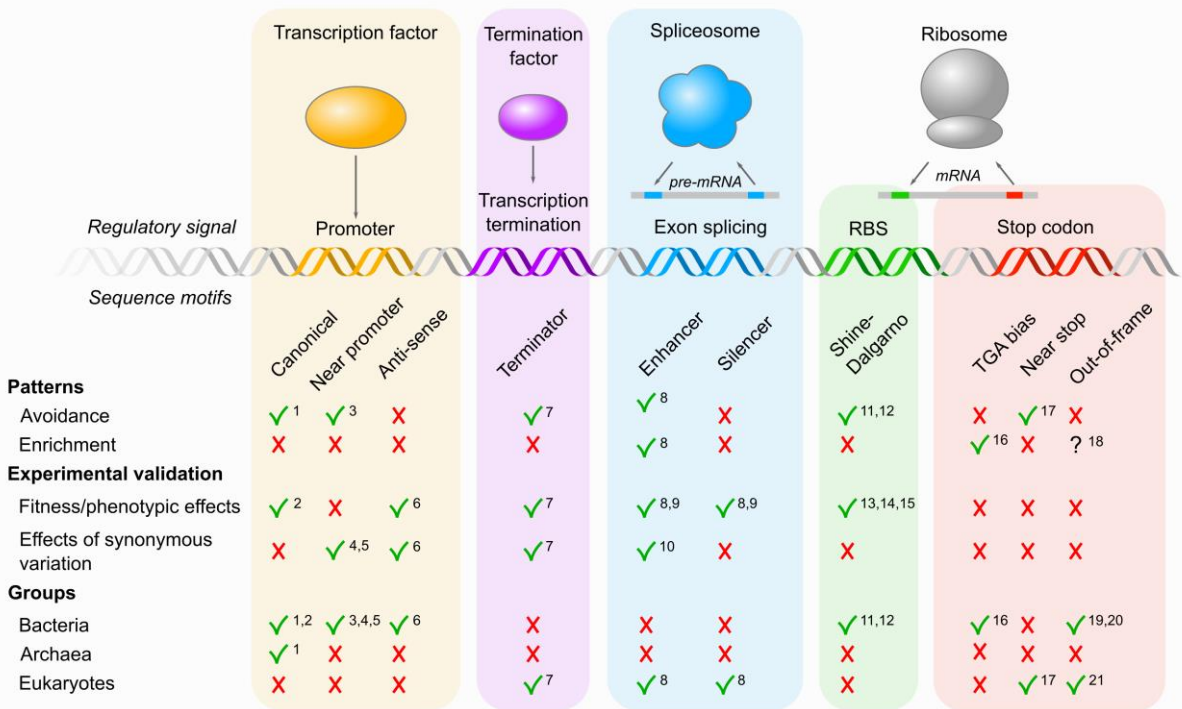
79 Ribosome profiling and other technical advances have led to an in-depth understanding of
80 the complex relationship between codon usage, translation efficiency regulation and
81 proteome composition. They enabled, for example, descriptions of the effect of codon usage
82 on mRNA secondary structure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

93

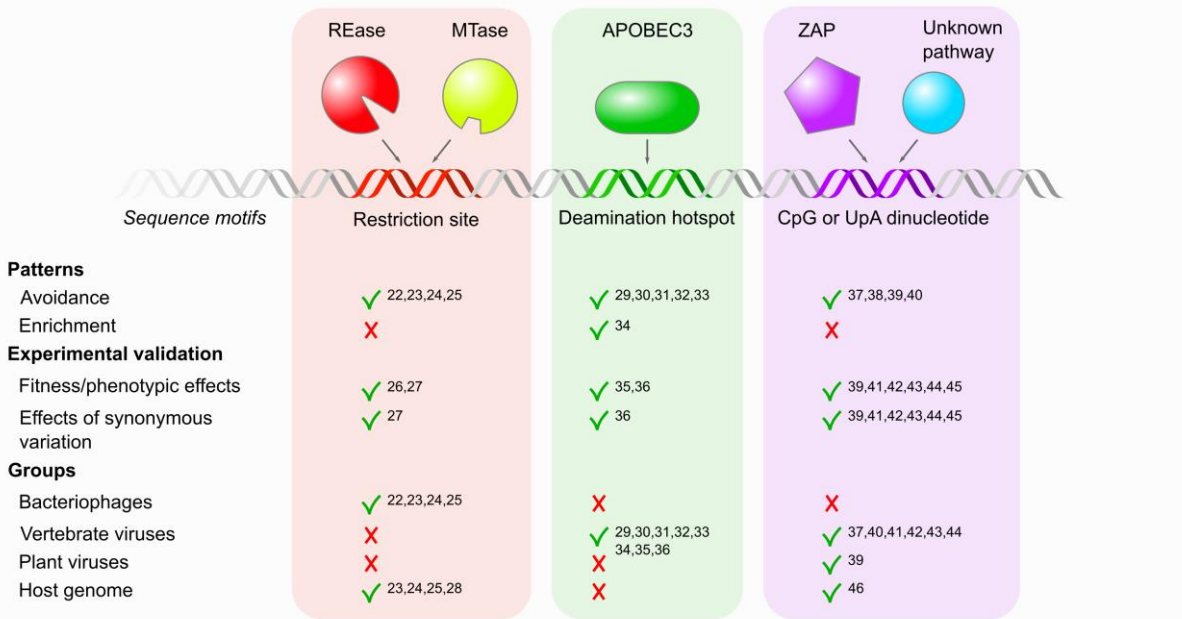
94 The existence of alternative synonymous sequences suggests that protein coding
95 genes could potentially contain or exclude sequence motifs with biologically meaningful
96 signals in addition to simply coding for an amino acid sequence. These biological signals can
97 take the form of motifs in the actual nucleotide sequence, or in the biophysical properties of
98 this sequence (secondary structure, hairpins, stiffness, etc.). The presence of these “other
99 codes” is particularly recognised for biological signals involved in gene expression (e.g.
100 Bergman & Tuller 2020), and it has been suggested that the genetic code is better suited for
101 encoding this additional information than the vast majority of the potential alternative genetic
102 codes (Itzkovitz & Alon 2007). We argue here that the potential for genes to contain
103 information beyond the code of the amino acid sequence implies that specific nucleotide
104 sequences can be favoured or disfavoured, because of the biological signal they carry. This
105 can result in selection on local codon usage for reasons other than its consequences on

106 translation accuracy and efficiency. In this review, we compile the different biological signals
107 that can be contained in nucleotide sequences. We further discuss patterns of avoidance
108 or enrichment of these sequence motifs and, when available, we present experimental
109 evidence of the phenotypic effects of synonymous mutations in relation to these biological
110 signals. Figure 1 provides a summary of the elements discussed in this review.

A. Sequence motifs involved in gene expression regulation



B. Sequence motifs targeted by antiviral immune systems



111 **Figure 1: A. Observed avoidance or enrichment of sequence motifs involved in gene**
 112 **expression regulation and potential phenotypic effects.** Different processes depend on
 113 particular sequence motifs in the DNA or mRNA for their regulation (colored boxes from left
 114 to right: transcription initiation, transcription termination, gene splicing, translation initiation,
 115 translation termination). Green checks indicate if there is evidence in the literature for
 116 avoidance or enrichment of particular sequence motifs, if the presence or absence of these

117 sequence motifs has observable phenotypic effects and if these phenotypic effects can be
118 modified through synonymous variation. An “?” indicates this issue is debated. The bottom
119 rows indicate in which domains of life these observations have been made. References:
120 ¹Hahn et al. (2003), ²Lambert et al. (2017),³Yona et al. (2018),⁴Ando et al. (2013),⁵Kershner
121 et al. (2016),⁶Urtecho et al. (2020),⁷Zhou et al. (2018),⁸Savisaar & Hurst (2017),⁹Sterne-
122 Weiler et al. (2011),¹⁰Mueller et al. (2015),¹¹Itzkovitz et al. (2010),¹²Diwan & Agashe
123 (2016),¹³Schrader et al. (2014),¹⁴Li et al. (2012),¹⁵Osterman et al. (2020),¹⁶Eyre-Walker
124 (1996),¹⁷Johnson et al. (2011),¹⁸Morgens et al. (2013),¹⁹Tse et al. (2010),²⁰Abrahams & Hurst
125 (2018),²¹Bertrand et al. (2015). **B. Observed avoidance or enrichment of sequence**
126 **motifs targeted by antiviral immune systems and potential phenotypic effects.** Different
127 types of of antiviral immune systems are considered (colored boxes from left to right:
128 bacterial restriction-modification systems (Rease-MTase); mammalian apolipoprotein B
129 mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3) mediated innate immunity;
130 eukaryotic antiviral pathways targeting CpG or UpA dinucleotides of which the zinc-finger
131 antiviral protein (ZAP) is known to act in vertebrates but for plants the molecular pathways
132 are yet to be elucidated). Green checks indicate if there is evidence in the literature for
133 avoidance or enrichment of particular sequence motifs, if the presence or absence of these
134 sequence motifs has observable phenotypic effects and if these phenotypic effects can be
135 modified through synonymous variation. The bottom rows indicate in which host groups
136 observations have been made in their infecting viruses or in the host genome itself.
137 References:²²Sharp et al. (1986),²³Karlin et al. (1992),²⁴Rocha et al. (2001),²⁵Rusinov et al.
138 (2018),²⁶Pleška et al. (2016),²⁷Pleška & Guet (2017),²⁸Gelfand & Koonin (1997),²⁹Warren et
139 al. (2015),³⁰Poulain et al. (2020),³¹Martinez et al. (2019),³²Chen & MacCarthy
140 (2017),³³Verhalen et al. (2016),³⁴Monajemi et al. (2014),³⁵Armitage et al. (2012),³⁶Sato et al.
141 (2014),³⁷Chen et al. (2013),³⁸Simmonds et al. (2013),³⁹Ibrahim et al. (2019),⁴⁰Xia
142 (2020),⁴¹Burns et al. (2009),⁴²Gaunt et al. (2016),⁴³Takata et al. (2017),⁴⁴Fros et al.
143 (2017),⁴⁵Trus et al. (2020),⁴⁶Burge et al. (1992).

144 Sequence motifs involved in gene expression regulation.

145 *Promoter, near-promoter and anti-sense promoter sequences.*

146 Promoters in bacteria are characterized by two consensus sequences, TATAAT and
147 TTGACA, respectively located 10 and 35 base pairs upstream of the transcriptional start site
148 (Browning & Busby 2004). Active promoter sequences are not necessarily an exact
149 consensus sequence, but usually contain only three or four of the six nucleotides (Kinney et
150 al. 2010). Promoter sequences, or sequences within a short mutational distance from a
151 promoter sequence, are likely to occur within DNA sequences because they are short and
152 moderately conserved. Indeed, 10% of 100bp random sequences exhibit promoter activity in
153 *Escherichia coli*, and within 250 generations 60% of random sequences evolved functional
154 promoter activity due to a single mutation (Yona et al. 2018). The potential of a given
155 sequence to evolve a functional promoter can be beneficial in terms of plasticity and
156 evolvability of the transcription network. It can even be beneficial when occurring in a coding
157 sequence: for example, in bacteria, synonymous mutations at the end of the coding
158 sequence of a gene have been shown to be beneficial because they create a promoter from
159 which the next gene in the operon is transcribed and this over-expression is advantageous in
160 specific environmental conditions (Ando et al. 2014; Kershner et al. 2016). However, the
161 appearance of a new promoter within a coding sequence can also lead to an overproduction
162 of RNA transcripts, sequestration of RNA polymerase and an overall reduction in gene
163 expression (Lamberte et al. 2017). Hahn et al. (2003) found that coding sequences across
164 Eubacteria and Archaea are under selection to avoid canonical promoter sequences, and
165 Yona et al. (2018) computationally showed that the *E. coli* coding genome is depleted in
166 sequences close to promoter sequences. Furthermore, this avoidance pattern is even
167 stronger for essential genes, for which perturbation is extremely costly. This suggests that
168 specific intragenic combinations of codons corresponding to promoter or near-promoter
169 sequences are generally disadvantageous but can also be beneficial in specific genomic and
170 environmental situations.

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

186 *Ribosome binding sequences*

187 Translation of mRNA is initiated by the binding of a ribosome to the
188 ribosomal binding site (RBS). Across all bacterial species, the consensus RBS consists of a
189 6-7bp motif found 5-10 bp upstream of the start codon and complementary to the 3' tail of the
190 16S ribosomal RNA (Shine & Dalgarno 1974). RBSs are relatively short and sequences that
191 are one or two mutations away from the consensus Shine-Dalgarno sequence can be a
192 functional RBS (Omotajo et al. 2015). Intragenic RBSs may promote spurious internal
193 translation initiation leading to the production of frame-shifted or truncated protein (Whitaker
194 et al. 2015), which is expected to have negative fitness effects (Drummond & Wilke 2009).
195 Intragenic RBSs are also known to increase the rate of ribosomal frame-shifting during
196 translation elongation. In some cases, this has been shown to be “programmed
197 frameshifting” allowing the production of two different functional proteins from the same

198 coding sequence (Chen et al. 2014; Devaraj & Fredrick 2010). However, cases of spurious
199 ribosomal frame-shifting during translation elongation are likely to have negative
200 consequences. In various bacterial species, internal RBSs have also been shown to induce
201 translational pauses by directly binding to the ribosome and thereby reducing the local
202 translation elongation rate (Schrader et al. 2014; Li et al. 2012), leading to a reduction in the
203 quantity of protein produced (Osterman et al. 2020). This slow local translation can have a
204 positive effect on fitness by allowing correct protein folding or down-regulating protein
205 translation (Fluman et al. 2014; Frumkin et al. 2017), or a negative effect if this down-
206 regulation is maladaptive. Like promoter sequences, RBSs also have a high probability of
207 occurring by chance in coding sequences, given their small size. It is difficult to predict
208 whether these motifs will be favoured or disfavoured by selection because of the diversity of
209 mechanistic and fitness consequences intragenic RBSs can have. The vast majority of
210 prokaryotic protein coding sequences are depleted of internal RBSs (Itzkovitz et al. 2010;
211 Diwan & Agashe 2016). Using a comparative approach, Hockenberry et al. (2018) showed
212 that strong intragenic RBSs detected in *E. coli* present a low level of conservation across
213 *Enterobacteriales* and that sequences downstream of internal RBSs are strongly depleted of
214 ATG start codons. Both observations suggest a negative effect of the presence of these
215 sequences. The general pattern emerging from these data is a pattern of selection against
216 intragenic RBSs although they may be favoured by local selection when their regulatory
217 effect on protein elongation is beneficial. Regardless of the direction of selection on
218 intragenic RBSs, these selective pressures have the potential to impact local codon usage
219 (Li et al. 2012).

220 *Overlapping and near-overlapping genes.*

221 Overlapping genes are widespread in bacterial genomes because of their high gene
222 density: a study analysing 699 bacterial species revealed more than 90% have at least one
223 overlapping gene pair (OGP), while some genomes harbour up to 3000 OGPs (Ahnert et al.
224 2008). Additionally, a high proportion of codirectional gene pairs are “near-OGPs” with less

225 than 40bps between the two genes (Pallejà et al. 2009). As a consequence, the upstream
226 gene sequence provides both the code for its own amino acid sequence and the promoter
227 and RBS of the downstream gene. For OGPs, the 3' end of the upstream gene also codes
228 for the amino acid sequence of the downstream gene (Huvet & Stumpf 2014). The double
229 role of these regions constrains the codon usage and partially explains why CUB on the end
230 of bacterial genes is often different from the rest of the genome (Eyre-Walker 1996).

231 *Stop, near-stop and out-of-frame stop codons*

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

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

252 avoidance of near-stop codons partially shapes the CUB for the four amino-acids coded both
253 by near-stop and non near-stop codons (Leucine, Serine, Arginine and Glycine).

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

265

266 *Transcription termination sequences*

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

278 preferences in codon usage may have co-evolved with transcription termination machinery to
279 avoid costly premature termination of transcription in GC-rich eukaryotes.

280

281 *Exon Silencing and Exon Enhancer Sequences*

282 In eukaryotic gene expression, transcription is followed by splicing - a process
283 through which non-protein coding introns are removed from the pre-mRNA, and protein-
284 coding exons are joined to produce a mature mRNA. Splicing is catalysed by a large RNA-
285 protein complex that recognizes specific sequence motifs in the pre-mRNA, both within
286 introns and exons (Abramowicz & Gos 2018). Exons contain Exonic Splice Enhancers (ESE)
287 and Exonic Splice Silencers (ESS), which enhance integration into the mature mRNA or
288 silence it, respectively. Disruption of ESE sites can cause the skipping of exons, leading to
289 the production of dysfunctional proteins. Conversely, the creation of new ESS sites can lead
290 to a similar outcome, by skipping previously included exons. Many ESE sites are involved in
291 interactions with RNA-binding proteins (RBPs) and a selective pressure to conserve or avoid
292 RBP motifs has been shown in primates and rodents (Savisaar & Hurst 2017). Interestingly,
293 the strength of selection to conserve ESEs has been linked to effective population size. Wu
294 et. Hurst (2015) showed, in a study across 30 different species, that mean intron size
295 predicts ESE density, with mean intron size negatively correlating with effective population
296 size. This argument also holds within species, with higher ESE density at genes with larger
297 and more numerous introns.

298 Perturbation of exon encoded regulatory information has been associated with
299 numerous human pathologies, including cystic fibrosis, Lynch syndrome, breast cancer,
300 muscular dystrophy and haemophilia B (Sterne-Weiler et al. 2011; Savisaar & Hurst 2017). A
301 comparative study (Fairbrother et al. 2004) showed that exon ends, where ESE are located,
302 contain fewer single nucleotide polymorphisms than the central region of exons, and linked
303 this pattern to the highly conserved splicing regulatory information encoded at exon

304 extremities. Additionally, an experimental approach determined that 23% of synonymous
305 mutations across exon 7 of the human *SMN1* gene decrease exon integration into mRNA
306 (Mueller et al. 2015). This suggests that for some genes, splicing signals are encoded over
307 the whole length of the exon. Thus, avoidance and maintenance of splice signals and other
308 non-splicing associated RBP motifs could influence codon usage over extensive portions of
309 the coding genome.

310

311 Sequence motifs targeted by antiviral immune systems

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

328 *Recognition sites for restriction-modification systems*

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

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

355 sites targeted by orthodox Type II R-M systems, but usually not for recognition sites of Type I
356 and Type III R-M systems (Sharp 1986; Rusinov et al. 2018). There are several explanations
357 for this observation. In Type II systems, the REase and the MTase are independent enzymes
358 with separate DNA recognition domains, while Type I and Type III systems function as
359 hetero-oligomeric complexes with a single sequence recognition domain (Tock & Dryden
360 2005). Sharing of recognition domains between R and M factors makes it easier to change
361 the specificity of Type I and Type III systems than that of Type II systems. This instigates a
362 phage-bacteria arms-race with rapid changes in the specificity of host defence, rendering
363 recognition site avoidance a less efficient strategy for long-term avoidance of host immune
364 defence using Type I or Type III R-M systems (Rusinov et al. 2018). Several phages are also
365 known to produce universal anti-restriction proteins that can inhibit the action of Type I or
366 Type III R-M systems, and are thus protected against restriction even when recognition sites
367 are present in their genome (e.g. SAMase in phage T3, Karlin et al. 1992). Due to the high
368 diversity in Type II R-M systems, such a universal defence could be more difficult to establish
369 (Rusinov et al. 2018). Type I and Type III systems also often require two recognition sites to
370 be present on opposing strands, so avoidance can additionally be achieved by removing a
371 recognition site from only one strand (Tock & Dryden 2005). Third, bacteriophage lifestyle
372 also seems to be a determining factor for the strength of selection against recognition sites,
373 with lytic phages showing a higher degree of recognition site avoidance than temperate
374 phages (Sharp 1986; Karlin et al. 1992; Rocha 2001; Rusinov et al. 2018), probably because
375 temperate phages integrate into the genome of the host where their DNA will be methylated
376 and thereby escape restriction.

377 Pleška & Guet (2017) provided direct experimental support for the phenotypic effect
378 of synonymous mutation through recognition site changes in bacteriophage λ cl857, a
379 conditionally lytic phage of *E. coli*. This phage contains five EcoRI restriction sites, into which
380 synonymous mutations were introduced. They observed that all individual synonymous point
381 mutations increased the likelihood of phage escape, although at a variable rate. The

382 combination of five synonymous mutations, one in each restriction site, provided full escape
383 from restriction by EcoRI. These experimental data represent direct evidence for strong
384 phenotypic effects of synonymous mutations located in a restriction site.

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

397 *APOBEC3 hotspots*

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

409 The antiviral action of A3s has been found to exert a mutational and selective
410 pressure on many viral genomes. Recent studies indicated an elevated C to U mutation rate
411 in SARS-CoV2, which can be attributed to the action of A3 (Di Giorgio et al. 2020; Rice et al.
412 2021; Ratcliff & Simmonds 2021). Viral genomes also often exhibit a depletion of A3 hotspots
413 (Warren et al. 2015; Poulain et al. 2020; Martinez et al. 2019; Chen & MacCarthy 2017).
414 Such a depletion has been recorded in as many as 22% of all human viruses, and is most
415 striking for 5'TC motifs that occupy the second and third position in a codon, where a
416 deamination of the third codon position is always synonymous (Poulain et al. 2020).
417 Furthermore, a high genomic GC content also provides protection against A3s because it
418 tends to minimize the presence of hotspots (Chen & MacCarthy 2017). However, a complete
419 avoidance of A3 hotspots is generally difficult to obtain, because it often requires multiple
420 non-synonymous mutations that would be detrimental to the virus (Martinez et al. 2019).

421 Depletion of A3 hotspots is only apparent in certain viral families, with members of the
422 papillomaviruses, polyomaviruses, coronaviruses and autonomous parvoviruses showing the
423 strongest depletion (Verhalen et al. 2016; Warren et al. 2015; Poulain et al. 2020). This
424 pattern could be caused by a higher A3 pressure on these viral families, either because they
425 infect cell types with higher A3 expression levels, because they induce A3 expression in their
426 host, or because they lack proteins that inhibit A3 activity (Warren et al. 2015; Verhalen et al.
427 2016). HIV, for example, is highly susceptible to A3G, but can effectively avoid deamination
428 by the production of the vif protein that neutralizes A3G, reducing the need for A3G motif
429 avoidance (Harris & Dudley 2015).

430 Although the action of A3-induced hypermutation is expected to have predominantly
431 inactivating effects on HIV-1 (Armitage et al. 2012), some studies found evidence that during
432 early infection HIV-1 can sometimes benefit from A3-induced hypermutation (Wood et al.
433 2009; Sato et al. 2014; Monajemi et al. 2014). This benefit is caused by accelerated
434 evolution and diversification of positions targeted by the adaptive immune system, allowing
435 for a quick evasion from the initial immune response. There are indications for positive

436 selection on several codon sites within A3 hotspots of the envelope gene of HIV-1 that
437 diversify during the early stages of infection (Wood et al. 2009). Sato et al. (2014)
438 furthermore experimentally showed that in HIV-1 vif mutants, the action of A3D/F can
439 promote *in vivo* viral diversification leading to a conversion of co-receptor usage. It has been
440 hypothesized that this could explain an observed enrichment of A3 hotspots in cytotoxic T-
441 cell epitope encoding portions of the HIV genome (Monajemi et al. 2014), but it remains
442 unclear how selection for deaminated hotspots during early infection is counteracted by
443 selection for unmodified hotspots during viral transmission.

444

445

446 *CpG and UpA dinucleotides*

447 Frequencies of CpG and UpA dinucleotides are often significantly depleted in both
448 vertebrate and plant RNA viruses (Cheng et al. 2013; Simmonds et al. 2013; Ibrahim et al.
449 2019; Xia 2020). This depletion can be partially caused by the viral genome mirroring the
450 nucleotide composition of the host mRNA, which avoids CpG and UpA for reasons other than
451 interactions with antiviral immune systems (Beutler et al. 1989). However, experimental
452 evidence suggests that plant- and vertebrate RNA viruses are additionally subjected to a
453 selective pressure for CpG and UpA avoidance imposed by the host's antiviral immunity.
454 Artificially increasing CpG and UpA dinucleotides, through synonymous mutations in protein
455 coding genes or mutations in non-coding regions, was shown to strongly decrease
456 replication in a large variety of viruses such as poliovirus (Burns et al. 2009), Influenza A
457 (Gaunt et al. 2016), HIV-1 (Takata et al. 2017), the human enteric echovirus 7 (Fros et al.
458 2017), the potato virus Y (Ibrahim et al. 2019), and Zika virus (Trus et al. 2020). Fros et al.
459 (2017) furthermore inferred that this effect was not caused by a lower translation efficiency
460 due to changes in codon usage, thus suggesting the action of an intrinsic defence pathway
461 present in the host cells acting on CpG and UpA dinucleotides. Takata et al. (2017) partially

462 confirmed this by showing that the zinc-finger antiviral protein (ZAP) is involved in inhibiting
463 virion production through targeting CpG dinucleotides in the RNA of HIV-1. Based on these
464 findings, Xia (2020) proposed that the extreme CpG deficiency in SARS-CoV-2 could
465 contribute to its high virulence in humans by allowing it to successfully avoid ZAP-mediated
466 antiviral immunity. The immune pathways targeting CpG and UpA dinucleotides of plant
467 viruses have not been elucidated, but analogous processes to those in vertebrates might
468 also operate in plants (Ibrahim et al. 2019).

469

470 Conclusions and perspectives

471 We have reviewed a number of biological mechanisms that are likely to exert
472 selection pressure on local codon usage for reasons other than selection for translation
473 accuracy and efficiency. In the light of these different elements, selection on codon usage
474 appears to be a combination of a global selection pressure imposed by the translation
475 machinery, and a patchwork of local selection pressures linked to the enrichment or
476 avoidance of specific nucleotide sequences that contain biological signals. However, contrary
477 to the translational selection, the local, non-translational selection pressures do not apply to
478 all genomes, as some are specific to viruses or to prokaryotes (see figure 1 for an overview).
479 It is also important to realise that some sequence patterns could be subject to multiple
480 selection pressures. For example, a palindromic sequence could be under selection both
481 because it is the preferred insertion site for certain Transposable Elements (TEs) (see Box 1)
482 and also because it is a restriction site. Specific selection pressures can therefore not be
483 simply deduced by finding that a specific pattern is avoided or enriched in a genome, or a
484 part of the genome. Knowledge of the evolutionary history of the species is generally
485 necessary to make inferences about selective pressures (e.g. associations with specific TEs,
486 specific restriction enzymes encoded and levels of self-restriction). Additionally, for most
487 mechanisms reviewed (except R-M motifs and CpG/UpA motifs), there are reports of both

488 avoidance and enrichment of the same motif or of positive and negative effects on fitness of
489 the addition or removal of these motifs. In these cases, the direction of selection is
490 determined by factors that range from environmental conditions to surrounding sequences.
491 Testing for avoidance or enrichment at a scale at which both might occur can lead to
492 negative results or to errors in the estimation of the strength of the selection pressure.
493 Finally, for all motifs, avoidance or enrichment patterns can be obtained through both
494 synonymous and non-synonymous mutations, but synonymous mutations are generally
495 expected to have lower direct fitness effects and for this reason represent *a priori* a preferred
496 way of avoiding or enriching specific patterns. Yet, when an avoidance or enrichment is
497 observed, it cannot be excluded that non-synonymous mutations contributed to this pattern.

498 From a methodological point of view, the detection of over- or under-
499 representation of a particular sequence motif in a genome is often not a trivial task, and is an
500 important issue in computational biology. This detection requires an appropriate model of the
501 genome that assumes the absence of a selective pressure on the sequence motif to which
502 observed frequencies can be compared. A wide range of methods have been developed for
503 this task, including simple estimations using the product of nucleotide or k-mer frequencies
504 and approaches using Markov models (see e.g. Rusinov et al. 2018 for a comparison of
505 methods). Given these methodological difficulties, several authors have noted that some
506 observations of sequence motif avoidance or enrichment are inconclusive and can be
507 artefacts of an erroneous methodology (Sharp 1986; Morgens et al. 2013). It is also a well-
508 known problem that the inference of selection on codon usage by comparative sequence
509 analysis can be confounded by mutational bias, as both processes can produce similar motif
510 enrichment/avoidance and codon usage patterns (Laurin-Lemay et al. 2018). Mutation biases
511 can affect codon usage on both a genome-wide and a local scale (Duret 2002).
512 Disentangling the effects of selection and mutational bias on codon usage is thus not an
513 easy task, and is still a subject of much debate (Galtier et al. 2018; Laurin-Lemay et al.
514 2018). Along the same lines, inference of selection on codon usage can be erroneous

515 because factors such as amino acid usage bias or gene expression are not considered. For
516 example, it was assumed that translational inefficient codons are selected at the 5' end of
517 bacterial signal peptides because they can facilitate protein secretion (Power et al. 2004).
518 However, (Cope et al. 2018) refuted this hypothesis by showing that the 5' end of bacterial
519 signal peptides show no differences in CUB compared to cytoplasmic proteins after
520 correcting for amino acid usage and gene expression. In the studies cited in the present
521 review, selection is often inferred based on deviations from genome-wide nucleotide or k-mer
522 frequencies. However, these generally do not account for context-dependent mutational
523 biases or amino acid usage (although see e.g. Wood et al. 2009 accounting for mutational
524 hotspots). The usage of more elaborate models accounting for multiple confounding factors
525 could thus nuance the assumption of selection when observing avoidance or enrichment of a
526 particular sequence motif. Ideally, the fitness effects of synonymous mutations are
527 empirically determined to provide unequivocal evidence for selective pressures on these
528 synonymous positions (Pleška & Guet 2017).

529 Patterns of avoidance or enrichment in specific motifs or codons are thus not
530 necessarily the product of selection. Conversely, the existence of selection for or against a
531 motif does not necessarily result in the enrichment or avoidance of this motif because it
532 depends on the selection coefficients and the effective population size. For translational
533 selection, selection coefficients on synonymous mutations are generally assumed to be weak
534 (Sharp & Li 1986) and translational selection is only expected to shape codon usage when
535 the effective population size is large enough so that selection can overcome drift, as stated
536 by the nearly neutral theory (Ohta & Gillespie 1996). Consequently, translational selection is
537 assumed to shape the codon usage of species with large effective population sizes, such as
538 many microorganisms and some invertebrate animals, but not (or less) in species with a
539 small effective population size such as larger mammals (Galtier et al. 2018). For non-
540 translational selection on codon usage, selection coefficients are generally unknown, but
541 they probably vary widely between selective pressures and synonymous sites (e.g. selection

542 against near-stop codons might be weak while selection on avoiding sequence motifs
543 targeted by antiviral immune systems might be stronger). To estimate the potential impact of
544 non-translational selective pressures on the codon usage of a particular species, both the
545 selection coefficient acting on synonymous variation and the effective population size of the
546 species will need to be considered. However, sometimes extrapolation might not be so
547 straightforward as selection coefficients on synonymous variation might be indirectly affected
548 by the effective population size (Wu & Hurst 2015). Future studies investigating the
549 importance of non-translational selective pressures for shaping codon usage in a wide
550 variety of organisms will be of particular interest to address this issue.

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

562

563 Text Box 1

564 Is local codon usage influenced by transposable elements?

565 Transposable elements (TEs) are DNA sequences that have the ability to change
566 their position (i.e. to transpose) within or between genomes. TEs are widely spread across all

567 eukaryotic and prokaryotic genomes, and their effects on genome structure and organism
568 fitness are manifold (see Bourque et al. 2018 for a review): (i) TEs increase genome size by
569 accumulating in genomes (Naville et al. 2019). (ii) They create new recombination sites and
570 thereby induce chromosome rearrangements (Lönnig & Saedler 2002). (iii) They enhance
571 the expression of genes, e.g. by introducing new cis-regulatory elements in their
572 neighbourhood (Salces-Ortiz et al. 2020). (iv) They are a source of novel mutations: either by
573 disrupting the expression of the genes they integrate into, or by introducing new genes
574 (Jangam et al. 2017). Thus, the phenotypic changes induced by TEs range from adaptive
575 (Salces-Ortiz et al. 2020) to lethal (Tsugeki et al. 1996). The sign and amplitude of the fitness
576 effect depends mainly on the TE content and on its insertion site.

577 Many TE families show strong preferences for their insertion sites (Levin & Moran
578 2011), but some have dispersed integration patterns, and exhibit low or no preference, e.g.
579 ~500,000 copies of the L1 retroelement can be found throughout the human genome. For
580 TEs showing an integration site preference, a precise nucleotide pattern is often required, for
581 example the conserved 60bp attnTn7 sequence required for the integration of Tn7 in
582 bacterial chromosomes (Kuduvalli 2001; Parks & Peters 2007). The preferred integration site
583 can also be a shorter, less conserved palindromic sequence, as for example the 6bp motif
584 where Tn10 preferentially inserts (Halling & Kleckner 1982). Other TE families show
585 preferences for certain parts of the genome: some integrate in gene-rich regions but avoid
586 coding regions, e.g. *Drosophila* P element often integrates 500bp upwards of transcription
587 start sites (Bellen et al. 2011) and others integrate specifically in heterochromatin and other
588 weakly expressed regions, e.g. in *Saccharomyces cerevisiae*, 90% of Ty5 integration events
589 occur in heterochromatin at telomeres (Zou & Voytas 1997). In many cases, the likelihood of
590 transposition to a site mostly depends on DNA mechanical properties: namely DNA
591 deformability, curvature, and melting (see Arinkin et al. 2019 for a review). Unwinding and
592 bending of DNA allows precise cleavage of the target site, and renders integration
593 irreversible (Morris et al. 2016; Ru et al. 2018). DNA melting allows the conjugative

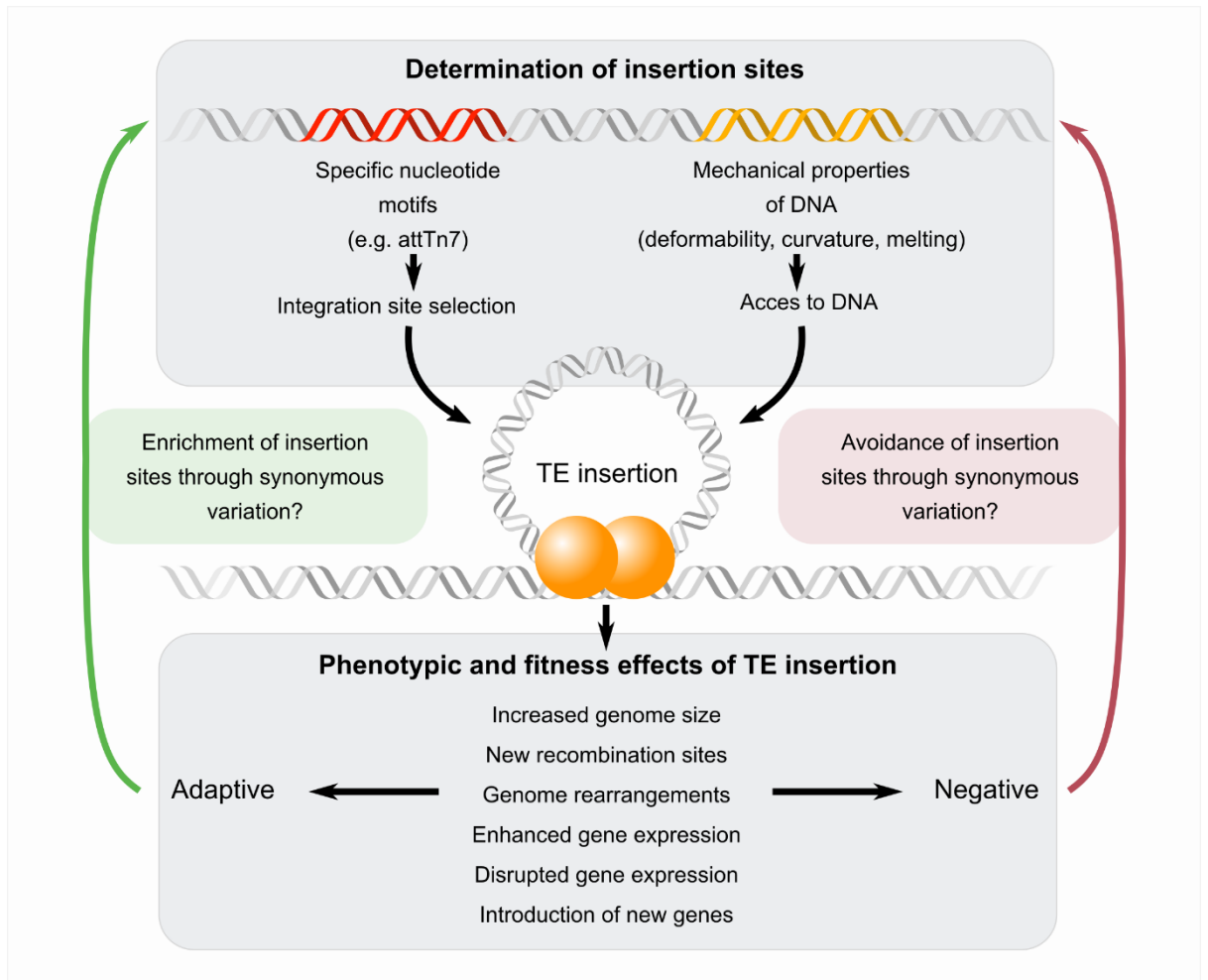
594 transposons to easily recombine with many insertion sites regardless of homology (Rubio-
595 Cosials et al. 2018). Even when recognition by the transposase requires a few precise
596 invariant base pairs (e.g. several DDE transposases require invariant T/A nucleotides in the
597 sequence in order to integrate), DNA helix flexibility may be necessary to allow recognition
598 and integration through base-flipping and formation of a base-specific contact zone with the
599 transposase (Morris et al. 2016). Structural properties of DNA directly depend on sequence
600 composition. GC content decreases thermostability and bendability but increases DNA
601 curvature (Vinogradov 2003). The deformability of TE integration sites is suggested to be
602 linked to their palindromicity, to their enrichment in T/A pairs (Arinkin et al. 2019) and in
603 pyrimidine-purine base steps (Maskell et al. 2015; Morris et al. 2016).

604 The codon usage of transposable elements and the evolutionary forces shaping it
605 have been investigated and debated (Southworth et al. 2019; Jia & Xue 2009; Lerat et al.
606 2002). It is also well established that the observed distribution of TEs in genomes is the
607 result of both TE integration preferences and selection against the integration of TEs at
608 certain loci (Sultana et al. 2017). However, to our knowledge, selection pressure on DNA
609 motifs preferred for TE insertion, the resulting avoidance or enrichment and the potential
610 impact on local codon usage has not been studied. However, by combining knowledge on TE
611 insertion fitness effects and on the nature of preferred insertion sites, predictions can be
612 derived. Local codon usage is likely to be a determinant of the local abundance of TE
613 integration sites, either because synonymous versions of local sequences differ in their
614 content of sequence-specific integration sites or palindromes, or because nucleotide
615 sequence determines DNA mechanical properties (Olson et al. 1998) which favour or
616 disfavour TE integration. Synonymous polymorphisms that increase the likelihood of TE
617 integration will be less fit and purged from the population. This would give rise to a local
618 codon usage preference that reduces the number of insertion motifs in coding regions. This
619 evolutionary scenario should be most prevalent when fitness is highly correlated with gene

620 expression, i.e. in organisms with few redundant genes and/or a fast life cycle, and this
621 selection for avoidance of integration sites should also be stronger for essential genes.

622 TE insertions can also have positive fitness effects, as adaptation to novel
623 environments can be achieved by loss-of-function mutations, particularly in bacteria
624 (reviewed in Hottes et al. 2013). In fluctuating environments, it might be advantageous to
625 have the capacity to remobilize previously lost gene functions. In this context, we could
626 imagine that gene expression could switch between 'off' and 'on' states through the
627 integration/excision of nonreplicative TEs (e.g. via cut-and-paste transposition mechanism).
628 Local codon usage preference could thus be under selection to increase the likelihood of
629 transposon integration in these genes. Both predictions for enrichment and avoidance of TE
630 integration sites can be tested by comparing the frequency of TE integration sites in different
631 gene categories. Predictions for enrichment can additionally be tested by analysing whole
632 genome sequencing data from experimental evolution studies involving stressful conditions
633 fluctuating over an extended period.

634



635

636 Figure Box 1. How could transposable elements exert local selection pressures on codon
 637 usage?

638

639 Literature cited

- 640 Abrahams L, Hurst LD. 2018. Refining the Ambush Hypothesis: Evidence That GC- and AT-
641 Rich Bacteria Employ Different Frameshift Defence Strategies. *Genome Biol. Evol.* 10:1153–
642 1173. doi: 10.1093/gbe/evy075.
- 643 Abramowicz A, Gos M. 2018. Splicing mutations in human genetic disorders: examples,
644 detection, and confirmation. *J. Appl. Genet.* 59:253–268. doi: 10.1007/s13353-018-0444-7.
- 645 Ahnert SE, Fink TMA, Zinovyev A. 2008. How much non-coding DNA do eukaryotes require?
646 *J. Theor. Biol.* 252:587–592. doi: 10.1016/j.jtbi.2008.02.005.
- 647 Ando H, Miyoshi-Akiyama T, Watanabe S, Kirikae T. 2014. A silent mutation in *mabA* confers
648 isoniazid resistance on *Mycobacterium tuberculosis*: *mabA* mutation confers INH resistance
649 on *Mtb*. *Mol. Microbiol.* 91:538–547. doi: 10.1111/mmi.12476.
- 650 Arinkin V, Smyshlyaev G, Barabas O. 2019. Jump ahead with a twist: DNA acrobatics drive
651 transposition forward. *Curr. Opin. Struct. Biol.* 59:168–177. doi: 10.1016/j.sbi.2019.08.006.
- 652 Armitage AE et al. 2012. APOBEC3G-Induced Hypermutation of Human Immunodeficiency
653 Virus Type-1 Is Typically a Discrete “All or Nothing” Phenomenon Worobey, M, editor. *PLoS*
654 *Genet.* 8:e1002550. doi: 10.1371/journal.pgen.1002550.
- 655 Bellen HJ et al. 2011. The *Drosophila* Gene Disruption Project: Progress Using Transposons
656 With Distinctive Site Specificities. *Genetics.* 188:731–743. doi: 10.1534/genetics.111.126995.
- 657 Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. 2013. Efficient translation initiation
658 dictates codon usage at gene start. *Mol. Syst. Biol.* 9:675. doi: 10.1038/msb.2013.32.
- 659 Bergman S, Tuller T. 2020. Widespread non-modular overlapping codes in the coding
660 regions. *Phys. Biol.* 17:031002. doi: 10.1088/1478-3975/ab7083.
- 661 Bertrand RL, Abdel-Hameed M, Sorensen JL. 2015. Limitations of the ‘ambush hypothesis’
662 at the single-gene scale: what codon biases are to blame? *Mol. Genet. Genomics MGG.*
663 290:493–504. doi: 10.1007/s00438-014-0937-y.
- 664 Beutler E, Gelbart T, Han JH, Koziol JA, Beutler B. 1989. Evolution of the genome and the
665 genetic code: selection at the dinucleotide level by methylation and polyribonucleotide
666 cleavage. *Proc. Natl. Acad. Sci. U. S. A.* 86:192–196. doi: 10.1073/pnas.86.1.192.
- 667 Blaisdell BE, Campbell AM, Karlin S. 1996. Similarities and dissimilarities of phage genomes.
668 *Proc. Natl. Acad. Sci.* 93:5854–5859. doi: 10.1073/pnas.93.12.5854.
- 669 Boël G et al. 2016. Codon influence on protein expression in *E. coli* correlates with mRNA
670 levels. *Nature.* 529:358–363. doi: 10.1038/nature16509.
- 671 Bourque G et al. 2018. Ten things you should know about transposable elements. *Genome*
672 *Biol.* 19:199. doi: 10.1186/s13059-018-1577-z.
- 673 Brantl S. 2007. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr.*
674 *Opin. Microbiol.* 10:102–109. doi: 10.1016/j.mib.2007.03.012.
- 675 Brophy JAN, Voigt CA. 2016. Antisense transcription as a tool to tune gene expression. *Mol.*
676 *Syst. Biol.* 12:854. doi: 10.15252/msb.20156540.

- 677 Browning DF, Busby SJW. 2004. The regulation of bacterial transcription initiation. *Nat. Rev.*
678 *Microbiol.* 2:57–65. doi: 10.1038/nrmicro787.
- 679 Brule CE, Grayhack EJ. 2017. Synonymous Codons: Choose Wisely for Expression. *Trends*
680 *Genet. TIG.* 33:283–297. doi: 10.1016/j.tig.2017.02.001.
- 681 Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics.*
682 129:897–907.
- 683 Burns CC et al. 2009. Genetic Inactivation of Poliovirus Infectivity by Increasing the
684 Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region
685 Codons. *J. Virol.* 83:9957–9969. doi: 10.1128/JVI.00508-09.
- 686 Burow DA et al. 2018. Attenuated Codon Optimality Contributes to Neural-Specific mRNA
687 Decay in *Drosophila*. *Cell Rep.* 24:1704–1712. doi: 10.1016/j.celrep.2018.07.039.
- 688 Chaney JL et al. 2017. Widespread position-specific conservation of synonymous rare
689 codons within coding sequences Wilke, CO, editor. *PLOS Comput. Biol.* 13:e1005531. doi:
690 10.1371/journal.pcbi.1005531.
- 691 Chen J et al. 2014. Dynamic pathways of –1 translational frameshifting. *Nature.* 512:328–
692 332. doi: 10.1038/nature13428.
- 693 Chen J, MacCarthy T. 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine
694 deaminases have differential effects when mutating retrotransposon and virus sequences
695 compared to host genes Matsen, FA, editor. *PLOS Comput. Biol.* 13:e1005471. doi:
696 10.1371/journal.pcbi.1005471.
- 697 Cheng X et al. 2013. CpG Usage in RNA Viruses: Data and Hypotheses Burk, RD, editor.
698 *PLoS ONE.* 8:e74109. doi: 10.1371/journal.pone.0074109.
- 699 Cohen O et al. 2016. Comparative transcriptomics across the prokaryotic tree of life. *Nucleic*
700 *Acids Res.* 44:W46-53. doi: 10.1093/nar/gkw394.
- 701 Coleman JR et al. 2008. Virus attenuation by genome-scale changes in codon pair bias.
702 *Science.* 320:1784–1787. doi: 10.1126/science.1155761.
- 703 Cope AL, Hettich RL, Gilchrist MA. 2018. Quantifying codon usage in signal peptides: Gene
704 expression and amino acid usage explain apparent selection for inefficient codons. *Biochim.*
705 *Biophys. Acta BBA - Biomembr.* 1860:2479–2485. doi: 10.1016/j.bbamem.2018.09.010.
- 706 Courel M et al. 2019. GC content shapes mRNA storage and decay in human cells. *eLife.*
707 8:e49708. doi: 10.7554/eLife.49708.
- 708 Devaraj A, Fredrick K. 2010. Short spacing between the Shine-Dalgarno sequence and P
709 codon destabilizes codon-anticodon pairing in the P site to promote +1 programmed
710 frameshifting: Ribosomal frameshifting. *Mol. Microbiol.* 78:1500–1509. doi: 10.1111/j.1365-
711 2958.2010.07421.x.
- 712 Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-
713 dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* 6:eabb5813. doi:
714 10.1126/sciadv.abb5813.
- 715 Diwan GD, Agashe D. 2016. The Frequency of Internal Shine–Dalgarno-like Motifs in
716 Prokaryotes. *Genome Biol. Evol.* 8:1722–1733. doi: 10.1093/gbe/evw107.

- 717 Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein
718 synthesis. *Nat. Rev. Genet.* 10:715–724. doi: 10.1038/nrg2662.
- 719 Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*
720 12:640–649. doi: 10.1016/S0959-437X(02)00353-2.
- 721 Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-
722 adapted for optimal translation of highly expressed genes. *Trends Genet.* 16:287–289. doi:
723 10.1016/S0168-9525(00)02041-2.
- 724 Eyre-Walker A. 1996. The close proximity of *Escherichia coli* genes: Consequences for stop
725 codon and synonymous codon use. *J. Mol. Evol.* 42:73–78. doi: 10.1007/BF02198830.
- 726 Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single Nucleotide Polymorphism-
727 Based Validation of Exonic Splicing Enhancers Sean Eddy, editor. *PLoS Biol.* 2:e268. doi:
728 10.1371/journal.pbio.0020268.
- 729 Fluman N, Navon S, Bibi E, Pilpel Y. 2014. mRNA-programmed translation pauses in the
730 targeting of *E. coli* membrane proteins. *eLife.* 3:e03440. doi: 10.7554/eLife.03440.
- 731 Freistroffer DV, Kwiatkowski M, Buckingham RH, Ehrenberg M. 2000. The accuracy of codon
732 recognition by polypeptide release factors. *Proc. Natl. Acad. Sci.* 97:2046–2051. doi:
733 10.1073/pnas.030541097.
- 734 Fros JJ et al. 2017. CpG and UpA dinucleotides in both coding and non-coding regions of
735 echovirus 7 inhibit replication initiation post-entry. *eLife.* 6:e29112. doi: 10.7554/eLife.29112.
- 736 Frumkin I et al. 2017. Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell.*
737 65:142–153. doi: 10.1016/j.molcel.2016.11.007.
- 738 Galtier N et al. 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural
739 Selection, Effective Population Size, and GC-Biased Gene Conversion. *Mol. Biol. Evol.*
740 35:1092–1103. doi: 10.1093/molbev/msy015.
- 741 Gaunt E et al. 2016. Elevation of CpG frequencies in influenza A genome attenuates
742 pathogenicity but enhances host response to infection. *eLife.* 5:e12735. doi:
743 10.7554/eLife.12735.
- 744 Gelfand MS, Koonin EV. 1997. Avoidance of palindromic words in bacterial and archaeal
745 genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25:2430–2439.
746 doi: 10.1093/nar/25.12.2430.
- 747 Gophna U. 2018. The unbearable ease of expression—how avoidance of spurious
748 transcription can shape G+C content in bacterial genomes. *FEMS Microbiol. Lett.* 365. doi:
749 10.1093/femsle/fny267.
- 750 Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a
751 genome strategy modulated for gene expressivity. *Nucleic Acids Res.* 9:213–213. doi:
752 10.1093/nar/9.1.213-b.
- 753 Hahn MW. 2003. The Effects of Selection Against Spurious Transcription Factor Binding
754 Sites. *Mol. Biol. Evol.* 20:901–906. doi: 10.1093/molbev/msg096.
- 755 Halling SM, Kleckner N. 1982. A symmetrical six-base-pair target site sequence determines
756 Tn10 insertion specificity. *Cell.* 28:155–163. doi: 10.1016/0092-8674(82)90385-3.

- 757 Harris RS, Dudley JP. 2015. APOBECs and virus restriction. *Virology*. 479–480:131–145.
758 doi: 10.1016/j.virol.2015.03.012.
- 759 Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu. Rev. Genet.* 42:287–299.
760 doi: 10.1146/annurev.genet.42.110807.091442.
- 761 Hockenberry AJ, Jewett MC, Amaral LAN, Wilke CO. 2018. Within-Gene Shine–Dalgarno
762 Sequences Are Not Selected for Function Agashe, D, editor. *Mol. Biol. Evol.* 35:2487–2498.
763 doi: 10.1093/molbev/msy150.
- 764 Hottes AK et al. 2013. Bacterial Adaptation through Loss of Function Matic, I, editor. *PLoS*
765 *Genet.* 9:e1003617. doi: 10.1371/journal.pgen.1003617.
- 766 Huvet M, Stumpf MP. 2014. Overlapping genes: a window on gene evolvability. *BMC*
767 *Genomics.* 15:721. doi: 10.1186/1471-2164-15-721.
- 768 Ibrahim A et al. 2019. A functional investigation of the suppression of CpG and UpA
769 dinucleotide frequencies in plant RNA virus genomes. *Sci. Rep.* 9:18359. doi:
770 10.1038/s41598-019-54853-0.
- 771 Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms.
772 *Mol. Biol. Evol.* doi: 10.1093/oxfordjournals.molbev.a040335.
- 773 Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional
774 information within protein-coding sequences. *Genome Res.* 17:405–412. doi:
775 10.1101/gr.5987307.
- 776 Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences.
777 *Genome Res.* 20:1582–1589. doi: 10.1101/gr.105072.110.
- 778 Jangam D, Feschotte C, Betrán E. 2017. Transposable Element Domestication As an
779 Adaptation to Evolutionary Conflicts. *Trends Genet.* 33:817–831. doi:
780 10.1016/j.tig.2017.07.011.
- 781 Jia J, Xue Q. 2009. Codon Usage Biases of Transposable Elements and Host Nuclear
782 Genes in *Arabidopsis thaliana* and *Oryza sativa*. *Genomics Proteomics Bioinformatics.*
783 7:175–184. doi: 10.1016/S1672-0229(08)60047-9.
- 784 Johnson LJ et al. 2011. Stops making sense: translational trade-offs and stop codon
785 reassignment. *BMC Evol. Biol.* 11:227. doi: 10.1186/1471-2148-11-227.
- 786 Karlin S, Burge C, Campbell AM. 1992. Statistical analyses of counts and distributions of
787 restriction sites in DNA sequences. *Nucleic Acids Res.* 20:1363–1370. doi:
788 10.1093/nar/20.6.1363.
- 789 Katz L. 2003. Widespread Selection for Local RNA Secondary Structure in Coding Regions
790 of Bacterial Genes. *Genome Res.* 13:2042–2051. doi: 10.1101/gr.1257503.
- 791 Kershner JP et al. 2016. A Synonymous Mutation Upstream of the Gene Encoding a Weak-
792 Link Enzyme Causes an Ultrasensitive Response in Growth Rate Metcalf, WW, editor. *J.*
793 *Bacteriol.* 198:2853–2863. doi: 10.1128/JB.00262-16.
- 794 Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize
795 the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.*
796 *S. A.* 107:9158–9163. doi: 10.1073/pnas.1004290107.

- 797 Kohli RM et al. 2009. A Portable Hot Spot Recognition Loop Transfers Sequence
798 Preferences from APOBEC Family Members to Activation-induced Cytidine Deaminase. *J.*
799 *Biol. Chem.* 284:22898–22904. doi: 10.1074/jbc.M109.025536.
- 800 Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon
801 usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.*
802 289:30334–30342. doi: 10.1074/jbc.M114.606632.
- 803 Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-Sequence Determinants of Gene
804 Expression in *Escherichia coli*. *Science*. 324:255–258. doi: 10.1126/science.1170160.
- 805 Kuduvali PN. 2001. Target DNA structure plays a critical role in Tn7 transposition. *EMBO J.*
806 20:924–932. doi: 10.1093/emboj/20.4.924.
- 807 Kurland CG. 1992. Translational Accuracy and the Fitness of Bacteria. *Annu. Rev. Genet.*
808 26:29–50. doi: 10.1146/annurev.ge.26.120192.000333.
- 809 Lamberte LE et al. 2017. Horizontally acquired AT-rich genes in *Escherichia coli* cause
810 toxicity by sequestering RNA polymerase. *Nat. Microbiol.* 2:16249. doi:
811 10.1038/nmicrobiol.2016.249.
- 812 Laurin-Lemay S, Philippe H, Rodrigue N. 2018. Multiple Factors Confounding Phylogenetic
813 Detection of Selection on Codon Usage Pupko, T, editor. *Mol. Biol. Evol.* 35:1463–1472. doi:
814 10.1093/molbev/msy047.
- 815 Lerat E, Capy P, Biéumont C. 2002. Codon Usage by Transposable Elements and Their Host
816 Genes in Five Species. *J. Mol. Evol.* 54:625–637. doi: 10.1007/s00239-001-0059-0.
- 817 Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their
818 hosts. *Nat. Rev. Genet.* 12:615–627. doi: 10.1038/nrg3030.
- 819 Li G-W, Oh E, Weissman JS. 2012. The anti-Shine–Dalgarno sequence drives translational
820 pausing and codon choice in bacteria. *Nature*. 484:538–541. doi: 10.1038/nature10965.
- 821 Lönnig W-E, Saedler H. 2002. Chromosome Rearrangements and Transposable Elements.
822 *Annu. Rev. Genet.* 36:389–410. doi: 10.1146/annurev.genet.36.040202.092802.
- 823 Martinez T, Shapiro M, Bhaduri-McIntosh S, MacCarthy T. 2019. Evolutionary effects of the
824 AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. *Virus Evol.* 5.
825 doi: 10.1093/ve/vey040.
- 826 Maskell DP et al. 2015. Structural basis for retroviral integration into nucleosomes. *Nature*.
827 523:366–369. doi: 10.1038/nature14495.
- 828 Monajemi M et al. 2014. Positioning of APOBEC3G/F Mutational Hotspots in the Human
829 Immunodeficiency Virus Genome Favors Reduced Recognition by CD8+ T Cells Sandberg,
830 JK, editor. *PLoS ONE*. 9:e93428. doi: 10.1371/journal.pone.0093428.
- 831 Morgens DW, Chang CH, Cavalcanti AR. 2013. Ambushing the ambush hypothesis:
832 predicting and evaluating off-frame codon frequencies in Prokaryotic Genomes. *BMC*
833 *Genomics*. 14:418. doi: 10.1186/1471-2164-14-418.
- 834 Morris ER, Grey H, McKenzie G, Jones AC, Richardson JM. 2016. A bend, flip and trap
835 mechanism for transposon integration. *eLife*. 5:e15537. doi: 10.7554/eLife.15537.

- 836 Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. 2015. The Silent Sway of
837 Splicing by Synonymous Substitutions. *J. Biol. Chem.* 290:27700–27711. doi:
838 10.1074/jbc.M115.684035.
- 839 Münk C, Willemsen A, Bravo IG. 2012. An ancient history of gene duplications, fusions and
840 losses in the evolution of APOBEC3 mutators in mammals. *BMC Evol. Biol.* 12:71. doi:
841 10.1186/1471-2148-12-71.
- 842 Naville M et al. 2019. Massive Changes of Genome Size Driven by Expansions of Non-
843 autonomous Transposable Elements. *Curr. Biol.* 29:1161-1168.e6. doi:
844 10.1016/j.cub.2019.01.080.
- 845 Ohta T, Gillespie JH. 1996. Development of Neutral and Nearly Neutral Theories. *Theor.*
846 *Popul. Biol.* 49:128–142. doi: 10.1006/tpbi.1996.0007.
- 847 Olson WK, Gorin AA, Lu X-J, Hock LM, Zhurkin VB. 1998. DNA sequence-dependent
848 deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci.* 95:11163–
849 11168. doi: 10.1073/pnas.95.19.11163.
- 850 Omotajo D, Tate T, Cho H, Choudhary M. 2015. Distribution and diversity of ribosome
851 binding sites in prokaryotic genomes. *BMC Genomics.* 16:604. doi: 10.1186/s12864-015-
852 1808-6.
- 853 Osterman IA et al. 2020. Translation at first sight: the influence of leading codons. *Nucleic*
854 *Acids Res.* 48:6931–6942. doi: 10.1093/nar/gkaa430.
- 855 Pallejà A, García-Vallvé S, Romeu A. 2009. Adaptation of the short intergenic spacers
856 between co-directional genes to the Shine-Dalgarno motif among prokaryote genomes. *BMC*
857 *Genomics.* 10:537. doi: 10.1186/1471-2164-10-537.
- 858 Parks AR, Peters JE. 2007. Transposon Tn7 Is Widespread in Diverse Bacteria and Forms
859 Genomic Islands. *J. Bacteriol.* 189:2170–2173. doi: 10.1128/JB.01536-06.
- 860 Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden
861 signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20:237–243. doi:
862 10.1038/nsmb.2466.
- 863 Pleška M et al. 2016. Bacterial Autoimmunity Due to a Restriction-Modification System. *Curr.*
864 *Biol.* 26:404–409. doi: 10.1016/j.cub.2015.12.041.
- 865 Pleška M, Guet CC. 2017. Effects of mutations in phage restriction sites during escape from
866 restriction–modification. *Biol. Lett.* 13:20170646. doi: 10.1098/rsbl.2017.0646.
- 867 Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of
868 codon bias. *Nat. Rev. Genet.* 12:32–42. doi: 10.1038/nrg2899.
- 869 Poulain F, Lejeune N, Willemart K, Gillet NA. 2020. Footprint of the host restriction factors
870 APOBEC3 on the genome of human viruses Lambert, PF, editor. *PLOS Pathog.*
871 16:e1008718. doi: 10.1371/journal.ppat.1008718.
- 872 Power PM, Jones RA, Beacham IR, Bucholtz C, Jennings MP. 2004. Whole genome
873 analysis reveals a high incidence of non-optimal codons in secretory signal sequences of
874 *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 322:1038–1044. doi:
875 10.1016/j.bbrc.2004.08.022.

- 876 Radhakrishnan A et al. 2016. The DEAD-Box Protein Dhh1p Couples mRNA Decay and
877 Translation by Monitoring Codon Optimality. *Cell*. 167:122-132.e9. doi:
878 10.1016/j.cell.2016.08.053.
- 879 Ratcliff J, Simmonds P. 2021. Potential APOBEC-mediated RNA editing of the genomes of
880 SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology*.
881 556:62–72. doi: 10.1016/j.virol.2020.12.018.
- 882 Rice AM et al. 2021. Evidence for Strong Mutation Bias toward, and Selection against, U
883 Content in SARS-CoV-2: Implications for Vaccine Design Townsend, J, editor. *Mol. Biol.*
884 *Evol.* 38:67–83. doi: 10.1093/molbev/msaa188.
- 885 Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization,
886 and efficient decoding for translation optimization. *Genome Res.* 14:2279–2286. doi:
887 10.1101/gr.2896904.
- 888 Rocha EPC. 2001. Evolutionary Role of Restriction/Modification Systems as Revealed by
889 Comparative Genome Analysis. *Genome Res.* 11:946–958. doi: 10.1101/gr.GR-1531RR.
- 890 Ru H et al. 2018. DNA melting initiates the RAG catalytic pathway. *Nat. Struct. Mol. Biol.*
891 25:732–742. doi: 10.1038/s41594-018-0098-5.
- 892 Rubio-Cosials A et al. 2018. Transposase-DNA Complex Structures Reveal Mechanisms for
893 Conjunctive Transposition of Antibiotic Resistance. *Cell*. 173:208-220.e20. doi:
894 10.1016/j.cell.2018.02.032.
- 895 Rusinov IS, Ershova AS, Karyagina AS, Spirin SA, Alexeevski AV. 2018. Avoidance of
896 recognition sites of restriction-modification systems is a widespread but not universal anti-
897 restriction strategy of prokaryotic viruses. *BMC Genomics*. 19:885. doi: 10.1186/s12864-018-
898 5324-3.
- 899 Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. 2020. Transposable
900 elements contribute to the genomic response to insecticides in *Drosophila melanogaster*.
901 *Philos. Trans. R. Soc. B Biol. Sci.* 375:20190341. doi: 10.1098/rstb.2019.0341.
- 902 Sato K et al. 2014. APOBEC3D and APOBEC3F Potently Promote HIV-1 Diversification and
903 Evolution in Humanized Mouse Model Ross, SR, editor. *PLoS Pathog.* 10:e1004453. doi:
904 10.1371/journal.ppat.1004453.
- 905 Savaisaar R, Hurst LD. 2017. Both maintenance and avoidance of RNA-binding protein
906 interactions constrain coding sequence evolution. *Mol. Biol. Evol.* msx061. doi:
907 10.1093/molbev/msx061.
- 908 Schrader JM et al. 2014. The coding and noncoding architecture of the *Caulobacter*
909 *crescentus* genome. *PLoS Genet.* 10:e1004463. doi: 10.1371/journal.pgen.1004463.
- 910 Seligmann H, Pollock DD. 2004. The Ambush Hypothesis: Hidden Stop Codons Prevent Off-
911 Frame Gene Reading. *DNA Cell Biol.* 23:701–705. doi: 10.1089/dna.2004.23.701.
- 912 Shah P, Gilchrist MA. 2011. Explaining complex codon usage patterns with selection for
913 translational efficiency, mutation bias, and genetic drift. *Proc. Natl. Acad. Sci. U. S. A.*
914 108:10231–10236. doi: 10.1073/pnas.1016719108.
- 915 Sharp P. 1986. Molecular evolution of bacteriophages: evidence of selection against the
916 recognition sites of host restriction enzymes. *Mol. Biol. Evol.* doi:
917 10.1093/oxfordjournals.molbev.a040377.

- 918 Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in
919 unicellular organisms. *J. Mol. Evol.* 24:28–38. doi: 10.1007/BF02099948.
- 920 Shine J, Dalgarno L. 1974. The 3'-Terminal Sequence of *Escherichia coli* 16S Ribosomal
921 RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Proc. Natl. Acad.*
922 *Sci.* 71:1342–1346. doi: 10.1073/pnas.71.4.1342.
- 923 Simmonds P, Xia W, Baillie J, McKinnon K. 2013. Modelling mutational and selection
924 pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in
925 cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics.* 14:610. doi:
926 10.1186/1471-2164-14-610.
- 927 Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: Evidences for phylogenetic
928 trends. *Comput. Biol. Chem.* 33:239–244. doi: 10.1016/j.combiolchem.2009.04.002.
- 929 Southworth J, Grace CA, Marron AO, Fatima N, Carr M. 2019. A genomic survey of
930 transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on
931 codon usage. *Mob. DNA.* 10:44. doi: 10.1186/s13100-019-0189-9.
- 932 Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a
933 common mechanism of human inherited disease. *Genome Res.* 21:1563–1571. doi:
934 10.1101/gr.118638.110.
- 935 Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for
936 translational accuracy. *Mol. Biol. Evol.* 24:374–381. doi: 10.1093/molbev/msl166.
- 937 Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by
938 retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* 18:292–308. doi:
939 10.1038/nrg.2017.7.
- 940 Takata MA et al. 2017. CG dinucleotide suppression enables antiviral defence targeting non-
941 self RNA. *Nature.* 550:124–127. doi: 10.1038/nature24039.
- 942 Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol.*
943 *Cell Biol.* 18:18–30. doi: 10.1038/nrm.2016.116.
- 944 Tock MR, Dryden DT. 2005. The biology of restriction and anti-restriction. *Curr. Opin.*
945 *Microbiol.* 8:466–472. doi: 10.1016/j.mib.2005.06.003.
- 946 Trus I et al. 2020. CpG-Recoding in Zika Virus Genome Causes Host-Age-Dependent
947 Attenuation of Infection With Protection Against Lethal Heterologous Challenge in Mice.
948 *Front. Immunol.* 10:3077. doi: 10.3389/fimmu.2019.03077.
- 949 Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y. 2010. Natural selection retains
950 overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC*
951 *Genomics.* 11:491. doi: 10.1186/1471-2164-11-491.
- 952 Tsugeki R, Kochieva EZ, Fedoroff NV. 1996. A transposon insertion in the *Arabidopsis*
953 *SSR16* gene causes an embryo-defective lethal mutation. *Plant J.* 10:479–489. doi:
954 10.1046/j.1365-313X.1996.10030479.x.
- 955 Tuller T, Carmi A, et al. 2010. An Evolutionarily Conserved Mechanism for Controlling the
956 Efficiency of Protein Translation. *Cell.* 141:344–354. doi: 10.1016/j.cell.2010.03.031.

- 957 Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by
958 both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* 107:3645–3650. doi:
959 10.1073/pnas.0909910107.
- 960 Urtecho G et al. 2020. Genome-wide Functional Characterization of *Escherichia coli*
961 Promoters and Regulatory Elements Responsible for their Function. doi:
962 10.1101/2020.01.04.894907.
- 963 Verhalen B, Starrett GJ, Harris RS, Jiang M. 2016. Functional Upregulation of the DNA
964 Cytosine Deaminase APOBEC3B by Polyomaviruses Ross, SR, editor. *J. Virol.* 90:6379–
965 6386. doi: 10.1128/JVI.00771-16.
- 966 Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res.*
967 31:1838–1844. doi: 10.1093/nar/gkg296.
- 968 Warren CJ, Van Doorslaer K, Pandey A, Espinosa JM, Pyeon D. 2015. Role of the host
969 restriction factor APOBEC3 on papillomavirus evolution. *Virus Evol.* 1:vev015. doi:
970 10.1093/ve/vev015.
- 971 Whitaker WR, Lee H, Arkin AP, Dueber JE. 2015. Avoidance of Truncated Proteins from
972 Unintended Ribosome Binding Sites within Heterologous Protein Coding Sequences. *ACS*
973 *Synth. Biol.* 4:249–257. doi: 10.1021/sb500003x.
- 974 Wood N et al. 2009. HIV evolution in early infection: selection pressures, patterns of insertion
975 and deletion, and the impact of APOBEC. *PLoS Pathog.* 5:e1000414. doi:
976 10.1371/journal.ppat.1000414.
- 977 Wu X, Hurst LD. 2015. Why Selection Might Be Stronger When Populations Are Small: Intron
978 Size and Density Predict within and between-Species Usage of Exonic Splice Associated *cis*-
979 Motifs. *Mol. Biol. Evol.* 32:1847–1861. doi: 10.1093/molbev/msv069.
- 980 Xia X. 2020. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host
981 Antiviral Defense Kumar, S, editor. *Mol. Biol. Evol.* 37:2699–2705. doi:
982 10.1093/molbev/msaa094.
- 983 Yan X, Hoek TA, Vale RD, Tanenbaum ME. 2016. Dynamics of Translation of Single mRNA
984 Molecules In Vivo. *Cell.* 165:976–989. doi: 10.1016/j.cell.2016.04.034.
- 985 Yang Q et al. 2019. eRF1 mediates codon usage effects on mRNA translation efficiency
986 through premature termination at rare codons. *Nucleic Acids Res.* 47:9243–9258. doi:
987 10.1093/nar/gkz710.
- 988 Yona AH, Alm EJ, Gore J. 2018. Random sequences rapidly evolve into de novo promoters.
989 *Nat. Commun.* 9:1530. doi: 10.1038/s41467-018-04026-w.
- 990 Yu C-H et al. 2015. Codon Usage Influences the Local Rate of Translation Elongation to
991 Regulate Co-translational Protein Folding. *Mol. Cell.* 59:744–754. doi:
992 10.1016/j.molcel.2015.07.018.
- 993 Zhao F, Yu C-H, Liu Y. 2017. Codon usage regulates protein structure and function by
994 affecting translation elongation speed in *Drosophila* cells. *Nucleic Acids Res.* 45:8484–8492.
995 doi: 10.1093/nar/gkx501.
- 996 Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. 2018. Codon usage biases co-evolve with
997 transcription termination machinery to suppress premature cleavage and polyadenylation.
998 *eLife.* 7. doi: 10.7554/eLife.33569.

999 Zou S, Voytas DF. 1997. Silent chromatin determines target preference of the
1000 *Saccharomyces* retrotransposon Ty5. *Proc. Natl. Acad. Sci.* 94:7412–7416. doi:
1001 10.1073/pnas.94.14.7412.

1002