



HAL
open science

Potential role of the X circular code in the regulation of gene expression

Julie D. Thompson, Raymond Ripp, Claudine Mayer, Olivier Poch, Christian J. Michel

► **To cite this version:**

Julie D. Thompson, Raymond Ripp, Claudine Mayer, Olivier Poch, Christian J. Michel. Potential role of the X circular code in the regulation of gene expression. *BioSystems*, 2021, 203, pp.104368. 10.1016/j.biosystems.2021.104368 . hal-03236059

HAL Id: hal-03236059

<https://u-paris.hal.science/hal-03236059>

Submitted on 13 Oct 2022

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.



**Potential role of the X circular code in the regulation of
gene expression**

Journal:	<i>BioEssays</i>
Manuscript ID	Draft
Wiley - Manuscript type:	Hypothesis
Date Submitted by the Author:	n/a
Complete List of Authors:	Thompson, Julie; ICube, Computer Science Ripp, Raymond; ICube, Computer Science Mayer, Claudine; ICube, Computer Science; Institut Pasteur, Unité de Microbiologie Structurale; Université Paris Diderot, Sorbonne Paris Cité Poch, Olivier; ICube, Computer Science Michel, Christian; ICube, Computer Science
Keywords:	gene expression, codon optimization, genetic code, circular code
Scientific Area:	

SCHOLARONE™
Manuscripts

Potential role of the *X* circular code in the regulation of gene expression

Julie D. Thompson^{1,*}, Raymond Ripp¹, Claudine Mayer^{1,2,3}, Olivier Poch¹, Christian J. Michel^{1,*}

¹ Department of Computer Science, ICube, CNRS, University of Strasbourg, Strasbourg, France

² Unité de Microbiologie Structurale, Institut Pasteur, CNRS, 75724 Paris Cedex 15, France

³ Université Paris Diderot, Sorbonne Paris Cité, 75724 Paris Cedex 15, France

Email: thompson@unistra.fr, raymond.ripp@unistra.fr, mayer@pasteur.fr, olivier.poch@unistra.fr,
c.michel@unistra.fr,

* To whom correspondence should be addressed; Email: thompson@unistra.fr

Keywords: gene expression; codon optimization; genetic code; circular code

Abstract

It is proposed that the *X* circular code, a set of 20 trinucleotides (codons) that has been identified in the protein-coding genes of most organisms from bacteria to eukaryotes, represents a new genetic signal that contributes to the regulation of gene expression. Circular codes have the important mathematical property of being error-correcting codes, and it has been hypothesized that the *X* circular code represents an ancestor of the standard genetic code, used in primordial systems to simultaneously decode amino acids and synchronize the reading frame. Here, using previously published experimental data, we demonstrate important correlations between the presence of *X* motifs in genes and more efficient gene expression. Our contributions suggest that *X* motifs continue to play a functional role in extant genomes and represent a new genetic signal, contributing to the maintenance of the correct reading frame and the optimization and regulation of gene expression.

1 Introduction

Codes are ubiquitous in genomes: for example, the genetic code, the nucleosome positioning code, the histone code, the splicing code, mRNA degradation code, or the 'codon usage' code [1-6], to name but a few. The standard genetic code [1] is probably the most well-known genome code, and represents one of the greatest discoveries of the 20th century. All known life on Earth uses the (quasi-)same triplet genetic code to control the translation of genes into functional proteins. The fact that there are 64 possible nucleotide triplet combinations but only 20 amino acids to encode, means that the genetic code is redundant and most amino acids are encoded by more than one codon. This redundancy allows for the encoding of supplementary information in addition to the amino acid sequence [7-9], and significant efforts have been applied recently to understand the multiple layers of information or 'codes within the code' [10] that can be exploited to increase the versatility of genome decoding.

Here, we focus on an important class of genome codes, called the circular codes, first introduced by Arquès and Michel [11] and reviewed in [12,13]. In coding theory, a circular code is also known as an error-correcting code or a self-synchronizing code, since no external synchronization is required for reading frame identification. In other words, circular codes have the ability to detect and maintain the correct reading frame. For example, comma-free codes are a particularly efficient subclass of circular codes,

1 where the reading frame is detected by a single codon. The genetic code was originally proposed to be
2 a comma-free code in order to explain how a sequence of codons could code for 20 amino acids, and at
3 the same time how the correct reading frame could be retrieved and maintained [14]. However, it was
4 later proved that the modern genetic code could not be a comma-free code [15], when it was discovered
5 that *TTT*, a codon that cannot belong to a comma-free code, codes for phenylalanine. Other circular codes
6 are less restrictive than comma-free codes, as a frameshift of 1 or 2 nucleotides in a sequence entirely
7 consisting of codons from a circular code will not be detected immediately but after the reading of a
8 certain number of nucleotides.
9

10 By excluding the four periodic codons {*AAA,CCC,GGG,TTT*} and by assigning each codon to a preferential
11 frame (i.e. each codon is assigned to the frame in which it occurs most frequently), the so-called *X*
12 circular code (Fig. 1a,b) was identified in the reading frame of protein coding genes from eukaryotes
13 and prokaryotes [11,16]. Other circular codes, and notably variations of the common *X* circular code, are
14 hypothesized to exist in different organisms [16-18].

15 The *X* circular code has important mathematical properties, in particular it is self-complementary [11],
16 meaning that if a codon belongs to *X* then its complementary trinucleotide also belongs to *X*. Like the
17 comma-free codes, the *X* circular code also has the property of synchronizability. It has been shown that,
18 in any sequence generated by the *X* circular code, at most 13 consecutive nucleotides are enough to
19 always retrieve the reading frame [11]. In other words, any sequence 'motif' containing 4 consecutive *X*
20 codons is sufficient to determine the correct reading frame (Fig. 1c) [20]. More formal definitions of the
21 mathematical properties of the *X* circular code are available [12-13,21].

22 The hypothesis of the *X* circular code in genes is supported by evidence from several statistical analyses
23 of modern genomes. We previously showed in a large-scale study of 138 eukaryotic genomes that *X*
24 motifs (defined as series of at least 4 codons from the *X* circular code) are found preferentially in protein-
25 coding genes compared to non-coding regions with a ratio of ~8 times more *X* motifs located in genes
26 [22]. More detailed studies of the complete gene sets of yeast and mammal genomes confirmed the strong
27 enrichment of *X* motifs in genes and further demonstrated a statistically significant enrichment in the
28 reading frame compared to frames 1 and 2 [23-24]. In addition, it was shown that most of the mRNA
29 sequences from these organisms (e.g. 98% of experimentally verified genes in *S. cerevisiae*) contain *X*

1 motifs. Intriguingly, conserved *X* motifs have also been found in many tRNA genes [25], as well as in
2 important functional regions of the 16S/18S ribosomal RNA (rRNA) from bacteria, archaea and
3 eukaryotes [26-27], which suggest their involvement in universal gene translation mechanisms. More
4 recently, a circular code periodicity 0 modulo 3 was identified in the 16S rRNA, covering the region that
5 corresponds to the primordial proto-ribosome decoding center and containing numerous sites that
6 interact with the tRNA and messenger RNA (mRNA) during translation [20].
7

8 Based on these observations, it has been proposed that the *X* circular code represents an ancestor of the
9 modern genetic code that was used to code for a smaller number of amino acids and simultaneously
10 identify and maintain the reading frame [27]. Intriguingly, the theoretical minimal RNA rings, short RNAs
11 designed to code for all coding signals without coding redundancy among frames, are also biased for
12 codons from the *X* circular code [28]. These RNA rings attempt to mimic primitive tRNAs and potentially
13 reflect ancient translation machineries [29-30].
14

15 The question remains of whether the *X* motifs observed in modern genes are simply a vestige of an
16 ancient code that might have existed in the early stages of cellular life, or whether they still play a role
17 in the complex translation systems of extant organisms.
18

19 In this work, we define a simple density parameter representing the coverage of *X* motifs in genes.
20 Unexpectedly, this parameter identifies several relations between the *X* circular code and translation
21 efficiency and/or kinetics. Our observations provide evidence supporting the idea that motifs from the
22 *X* circular code represent a new genetic signal, participating in the maintenance of the correct reading
23 frame and the optimization and regulation of gene expression.
24
25
26
27
28
29
30
31
32
33

34 **2. The *X* circular code and 'optimal' codons/dicodons for translation efficiency**

35 **2.1 *X* codons correlate with optimal codons**

36 We compared the 20 codons that belong to the *X* circular code with the 'codon optimality code' resulting
37 from various statistical and experimental studies in metazoan [31], as well as in *S. cerevisiae* [32]. In these
38 studies, the codon stabilization coefficient (CSC), defined as the Pearson correlation coefficient between
39 the occurrence of each codon and the half-life of each mRNA, was used as a robust and conserved
40 measure of how individual codons contribute to shape mRNA stability and translation efficiency. Thus,
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 codons found more frequently in genes with longer mRNA half-lives have higher CSC values. The 61
3
4 coding codons can then be ranked according to their CSC scores in different organisms. Fig. 2 shows the
5
6 mean ranking of optimal codons, according to the CSC score, from four different experiments (in *S.*
7
8 *cerevisiae*, zebrafish, *Xenopus* and *Drosophila*), where the highest ranking codon is the most optimal one.
9
10 The *X* codons are ranked significantly higher than non-*X* codons (i.e. the 41 coding codons which do not
11
12 belong to the circular code *X*), according to a Mann-Whitney signed rank test (z -score = 4.3, p -value <
13
14 0.00001). In other words, optimal codons for mRNA stability and elongation rate are significantly
15
16 enriched in *X* codons.
17
18
19
20

21 **2.2 *X* codons correlate with the dicodons associated with increased expression**

22
23 In recent years, emerging evidence has shown that translational rates may be encoded by dicodons
24
25 rather than single codons [33-35]. For example, a large-scale screen in *S. cerevisiae* [33] assessed the degree
26
27 to which codon context modulates eukaryotic translation elongation rates beyond effects seen at the
28
29 individual codon level. The authors screened yeast cell populations housing libraries containing random
30
31 sets of triplet codons within an ORF encoding superfolder Green Fluorescent Protein (GFP). They found
32
33 that 17 dicodons were strongly associated with reduced GFP expression, i.e. associated with a
34
35 substantial reduction of the translation elongation rate. This set included the known inhibitory dicodon
36
37 *CGA-CGA* and was enriched for codons decoded by wobble interactions. Of these 17 dicodons associated
38
39 with slower translation elongation rates, none are composed of 2 *X* codons (Supplementary Table 1).
40
41

42
43 A subsequent statistical analysis of coding sequences of nine organisms [34] identified dicodons with
44
45 significant different frequency usage for coding either lowly or highly abundant proteins. The working
46
47 hypothesis was that sequences encoding abundant proteins should be optimized, in the sense of
48
49 translation efficiency. 16 dicodons were identified with a preference for low abundance proteins, while
50
51 40 dicodons presented a preference for high abundance proteins. None of the 16 dicodons associated
52
53 with low abundance proteins are composed of 2 *X* codons (Supplementary Table S1). In contrast, 27 of
54
55 the 40 dicodons associated with high abundance proteins correspond to 2 *X* codons, and only 3 dicodons
56
57 do not contain any *X* codons (Supplementary Table S1).
58
59
60

1
2 These recent studies support the idea that codons in coding sequences are likely arranged in an
3 organized way, and that the local sequence context contributes to the effects of codon usage bias on gene
4 regulation. Strikingly, our observations support the hypothesis that codon context may be linked in
5 some way to the *X* circular code. In the next section, we describe more detailed analyses that test this
6 hypothesis further.
7
8
9
10
11
12
13
14

15 **3. *X* motifs are enriched in the minimal gene set**

16
17 Based on the increasing evidence of the importance of codon context [35-39], we hypothesized that if the
18 *X* circular code plays a role in gene regulation, then we might expect to see a non-random use, or
19 'clusters', of *X* codons along the length of the gene. In previous work [23-24], we defined an *X* motif as a
20 series of consecutive *X* codons (of length at least 4 codons in order to always retrieve the reading frame,
21 see Supplementary Materials) in a gene sequence and searched for such *X* motifs in the reading frames
22 of different genes. This approach allowed us to demonstrate that the reading frames of genes in yeasts
23 and in mammals are significantly enriched in such *X* motifs.
24
25
26
27
28
29
30
31

32 To test the hypothesis that the *X* motifs represent a more universal signature, we analyzed a set of 81
33 genes that were previously defined as a 'minimal gene set' [40]. At that time, the 'minimal gene set' genes
34 were found to be conserved in all species. We used the *Mycoplasma genitalium* genes provided in the
35 original study, as well as 15,822 orthologous sequences (5503 eukaryotes, 9205 bacteria and 1114
36 archaea) from the OrthoInspector 3.0 database [41], and identified all *X* motifs in the reading frame with
37 a minimum length of 4 codons. Fig. 3 shows the density (defined as the number of *X* motifs per kilobase,
38 see Supplementary Materials) of the *X* motifs in the mRNA sequences. To evaluate the significance of the
39 enrichment, as in previous work [23-24], we used a randomization model in which we generated N=100
40 random codes that preserved most of the properties to the *X* code, except the circularity. We then
41 identified all random motifs from the 100 random codes and calculated mean values for the 100 codes.
42 The density of *X* motifs found in the minimal gene set sequences belonging to the three domains of life,
43 is significantly higher than the density of random motifs according to a one-sided Student's *t*-test ($p <$
44 10^{-100}) for each set of sequences from archaea, bacteria and eukaryota. This study demonstrates that *X*
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 motifs are significantly enriched in the minimal gene set, and seem to be a universal feature of gene
2 sequences in all three domains of life.
3
4
5
6
7

8 **4. X motifs are enriched in codon-optimized genes**

9
10 If *X* motifs modify the codon usage in favor of optimal codons for translational efficiency, then we would
11 expect that increasing the number of *X* motifs in a gene would increase the expression level. In an
12 indirect way, we have shown that this is indeed the case. We previously showed that synthetic genes,
13 which were re-designed for optimized protein expression, generally have more *X* motifs [24].
14 Supplementary Fig. 1A shows an example of the protein L1h from human papillomavirus (HPV-16),
15 optimized for expression in mammalian cell lines and leading to significantly increased expression [42].
16 Here, the wild type gene contains only 3 *X* motifs, while the optimized gene construct has a total of 21 *X*
17 motifs. It is important to note that classical codon optimization strategies do not always increase protein
18 expression levels. Supplementary Fig. 1B shows another example involving the L1s protein from human
19 papillomavirus (HPV-11) optimized for expression in the potato *Solanum tuberosum* [43]. In this case,
20 only a low level of L1 expression was observed for the codon-optimized gene. In this example, we did
21 not observe a significant difference between the number of *X* motifs in the wild type and optimized
22 sequences (5 *X* motifs in the wild type gene compared to 4 in the optimized construct).
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37

38 Codon replacement strategies have also been applied to the design of attenuated viruses, although in
39 this case frequent codons are replaced with rare ones. Using quantitative proteomics and RNA
40 sequencing, the molecular basis of attenuation in a strain of bacteriophage T7 (*Escherichia coli* K-12)
41 was investigated [44]. The authors engineered the *E. coli* major capsid protein gene (gene 10A) to carry
42 different proportions of suboptimal, rare codons. Transcriptional effects of the recoding were not
43 observed, but proteomic observations revealed that translation was halved for the completely recoded
44 major capsid gene, with subsequent effects on virus fitness (measured as doublings/hour). We obtained
45 the sequences with 10%, 20%, 30% and 50% recoding from [45] and identified the density (defined in
46 Supplementary Materials) of *X* motifs in each construct. Fig. 4 clearly shows the correlation between the
47 fitness obtained for each recoded sequence and the density of *X* motifs observed. The authors suggested
48
49
50
51
52
53
54
55
56
57
58
59
60

1 that recoding of gene 10A reduced capsid protein abundance probably by ribosome stalling rather than
2 ribosome fall-off.
3

4
5
6 In general, codon optimization is a successful strategy for improving protein expression in heterologous
7 systems. However, simply replacing all rare codons by frequent codons can have negative effects *in vivo*
8 [46]. Rare codons have the potential to slow down the translation elongation rate, due to the relatively
9 long dwell time of the ribosome while searching for rare tRNAs. Several studies have suggested that
10 gene-wide codon bias in favor of slowly translated codons serves as a regulatory means to obtain low
11 expression levels of protein when desired, for example, in the case of regulatory genes, or where excess
12 of the protein may be detrimental or lethal to the cell. An example, in *Neurospora crassa*, demonstrated
13 that codon optimization of the central clock protein FRQ actually abolished circadian rhythms [47].
14 Different optimized constructs of the wild type gene *frq* were used in the study, where either the N-
15 terminal end (codons 1-164) or the middle region (codons 185-530) was optimized. All optimized
16 constructs gave higher levels of FRQ protein, this led to a different structural conformation. The density
17 of *X* motifs (defined in Supplementary Materials) identified in the different wild type and optimized
18 constructs is shown in Supplementary Table 22. As in the previous examples, the optimized constructs
19 contain significantly more *X* motifs (for instance, density of 10.2 in the N-terminal end of the fully
20 optimized construct compared to 4.1 in the wild type). This example shows how non-optimal codon
21 usage, and the associated reduction in the number of *X* motifs, can be used *in vivo* to regulate protein
22 expression and to achieve optimal protein structure and function.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 In nature, the translation efficiency of a gene may vary at different conditions, cell types and tissues [48-
43 51]. Thus, it has been proposed that the codon optimization should take into account other factors in
44 addition to replacing rare codons by frequent ones, a process termed 'codon harmonization' [52-54]. Taken
45 together, the examples described above suggest that it may be important for such harmonization
46 strategies to consider the effect of codon replacement on the insertion or deletion of *X* motifs.
47
48
49
50
51
52
53
54

55 **5. *X* motifs correlate with translation efficiency and mRNA stability**

56
57
58
59
60

1
2 In this section, we examine the recent evidence resulting from high-throughput technologies such as
3
4 ribosome profiling, and demonstrate that the presence of *X* motifs in genes can be used as a predictor of
5
6 gene expression level.
7

8 We have shown previously that the reading frames of genes in *S. cerevisiae* are significantly enriched in
9
10 *X* motifs [16]. Since then, ribosomal profiling has enabled a more detailed study of translation efficiency
11
12 for a large set of 5450 genes from this organism [55]. A central assumption of ribosome profiling is that
13
14 indirect measurement of the kinetics of translation *via* ribosome footprint occupancy on transcripts is
15
16 directly reflective of true protein synthesis. The authors thus performed a simulation of translation
17
18 based on the totally asymmetric simple exclusion process (TASEP) model, using experimental
19
20 measurements of the number of ribosomes on each transcript as well as RNA copy numbers to calibrate
21
22 the parameters. They used the simulation to estimate the average translation rate of each gene, using
23
24 experimental measurements of ribosome occupancy. Again, we identified the *X* motifs in the complete
25
26 set of 5450 genes and calculated the density of *X* motifs (defined in Supplementary Materials) in three
27
28 subsets of the genes having different estimated translation rates (Fig. 5). We observed that genes with
29
30 higher translation rates had significantly more *X* motifs than those with lower translation rates. The
31
32 density of *X* motifs is higher for the sequences with medium translation rates than for those with low
33
34 translation rates (one-sided Student's t-test $p < 10^{-10}$) and for the sequences with high translation rates
35
36 than for those with medium translation rates (one-sided Student's t-test $p < 10^{-14}$). This result
37
38 demonstrates the link between the total time needed for ribosome transition on a mRNA and density of
39
40
41
42
43 *X* motifs along the length of the sequence.

44 To investigate whether *X* motifs might play a role in modulating ribosome speed in specific regions in
45
46 mRNA, we considered single protein studies, where local translation elongation rate has been studied
47
48 in detail. The first example concerns the study of a gene in *S. cerevisiae*, to investigate the link between
49
50 translational elongation and mRNA decay [56]. In this study, various HIS3 protein constructs (length of
51
52 699 nucleotides) were designed with increasing codon optimality (measured by the CSC index) from
53
54 0% to 100%. We identified *X* motifs in the different constructs as before and compared them to the
55
56 experimentally measured mRNA half-life. As the authors point out, the mRNA half-life is largely
57
58 determined by the codon-dependent rate of translational elongation, since mRNAs whose translation
59
60

1 elongation rate is slowed by inclusion of non-optimal codons are specifically degraded. The density of *X*
2 motifs ranges from 0 in the 0% optimized construct to more than 7 in the 100% optimized sequence
3
4 (Fig. 6a). The results suggest that the introduction of individual *X* motifs in specific regions can be used
5
6 to increase the mRNA half-life.
7
8

9
10 The second example concerns a *Drosophila* cell-free translation system that was used to directly
11 compare the rate of mRNA translation elongation for different luciferase constructs with synonymous
12 substitutions [57]. The OPT construct was designed with the most preferred codons in all positions except
13 for the first 10 codons, while the dOPT construct had the least preferred codons in all positions. The N-
14 OPT, M-OPT and C-OPT constructs were created by replacing the N-terminal part (codons 11–223),
15 middle part (codons 224–423) and C-terminal part (codons 424–550) of the dOPT sequence with the
16 corresponding optimized sequence, respectively. For each construct, the authors measured the time
17 when the luminescence signal was first detected after start of translation. The time of first appearance
18 (TFA) should thus reflect the speed of translation process. Higher TFA values were observed for each
19 construct in the order dOPT < C-OPT < M-OPT < N-OPT < OPT, correlating well with an increasing
20 density of *X* motifs (Fig. 6b). These results suggest that the introduction of *X* motifs in different regions
21 of the gene significantly increased the rate of translation elongation, probably by speeding up ribosome
22 movement on the mRNA.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37

38 We have highlighted the potential effects of *X* motifs on translation elongation speed, protein folding
39 and function. The examples selected include studies in very different organisms, including viruses, fungi
40 and insects with different codon usage bias (codon usage tables for these organisms are provided in
41 Supplementary Table S3), but in all the examples a strong correlation is observed between ‘optimal’
42 codons and *X* codons. Taken together, the results support the idea that the use of *X* motifs is a conserved
43 mechanism from viruses to animals that may participate in the modulation or regulation of the
44 translation elongation rate along the mRNA.
45
46
47
48
49
50
51
52
53
54

55 **6. Conclusions and Prospects**

56
57 In this work, we have combined two very distinct research domains: gene translation through the
58 genetic code and the theory of circular codes which allows two processes simultaneously: reading frame
59
60

1 retrieval and amino acid coding. Our hypothesis is that at least two codes operate in genes: the standard
2 genetic code, experimentally proved to be functional, and the *X* circular code that has been shown to be
3 statistically enriched in genes. For the first time here, we shed light on a number of biological
4 experimental results by using the definition of a very simple parameter to analyze the density of *X* motifs
5 in genes, i.e. motifs from the circular code *X*.
6
7
8
9
10
11

12 We would first like to make some comments about the mathematical structure of these two codes. The
13 standard genetic code consists of 60 codons coding for 19 amino acids, the start codon *ATG* that codes
14 for the amino acid *Met* and establishes the reading frame, and three non-coding stop codons
15 {*TAA, TAG, TGA*}. The genetic code has a weak mathematical structure: a surjective coding map for the 60
16 codons and an incomplete self-complementary property for the 60 codons (e.g. the complementary
17 codon of *TTA* coding the amino acid *Leu* is the stop codon *TAA*) implying a non self-complementary
18 property for the four start/stop codons. The circular code *X* consists of 20 codons coding for 12 amino
19 acids and has a strong mathematical structure: circularity for retrieving the reading frame, a surjective
20 coding map, a complete self-complementary property for the 20 codons, a C^3 property, etc. (reviewed in
21 [12-13]).
22
23
24
25
26
27
28
29
30
31
32
33

34 We propose that the theory of circular codes can be used to shed light on many of the observed
35 phenomena related to optimal codons/dicodons and the effects of codon optimization on different
36 factors of gene expression, from transcriptional regulation to translation initiation, retrieval of the open
37 reading frame, translation elongation velocities, and protein folding. We showed that optimal codons at
38 the species and gene levels correlate well with the 20 codons that define the *X* circular code. Importantly,
39 the optimal codons identified in diverse species [31] that increase translation elongation rates and mRNA
40 stability are significantly enriched in *X* codons. We then studied a number of published experiments that
41 used recent technologies to perform more detailed investigations of codon usage along the length of a
42 gene, which suggest that codon context and local clusters of optimal or non-optimal codons may
43 represent important regulatory signals for translation bursts and pauses [6,58]. In all these experiments,
44 increased translation efficiency correlates with the number of *X* motifs present in the gene sequences.
45 These observations raise the question: do *X* motifs somehow orchestrate elongation rate? Since it is
46 known that translational elongation rate is intimately connected to mRNA stability, it is also tempting
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 to suggest that *X* motifs are linked to the universal code of codon-mediated mRNA decay proposed by
3
4 Chen and Coller [59].

5
6 The theory of the *X* circular code will have practical implications for improving the prediction of gene
7
8 expression levels based on the gene sequence. Most of the current codon usage measures are dependent
9
10 on the studied organism and the chosen expression system. In contrast, the presence of *X* motifs
11
12 represents a universal signature that is significantly correlated with increased expression. Our previous
13
14 work has already shown that *X* motifs can predict functional versus dubious genes in yeast [23] and can
15
16 be used for rational gene design [24].

17
18 Translation of mRNA by the ribosome is a universal mechanism, and the most parsimonious explanation
19
20 for the observed correlation between the presence of *X* motifs and increased translation elongation rates
21
22 is that *X* motifs are somehow recognized by the ribosome. It is known that codon usage has effects on
23
24 the major steps of translation elongation, including codon-anticodon decoding and peptide bond
25
26 formation [58], as well as translocation which can be slowed down by mRNA secondary structure
27
28 elements, such as pseudoknots and stem-loops [60]. Our hypothesis that *X* motifs in mRNA are
29
30 recognized by the ribosome is further supported by recent ribosome profiling experiments in
31
32 *Neurospora crassa*, which suggest that codon optimization increases the rate of ribosome movement on
33
34 mRNA [61], and by the observation that translation elongation and mRNA stability are coupled through
35
36 the ribosomal A-site [62]. Interestingly, our previous work has identified *X* motifs in the anticodon region
37
38 of multiple tRNA genes, as well as in important functional regions of the ribosomal rRNA including the
39
40 decoding centre [25-27].

41
42 How could motifs from the *X* circular code work? If the decoding unit at the ribosome is the anticodon
43
44 then the comma-free codes would immediately return to the reading phase while the general circular
45
46 codes would have a delay associated with reading at most four codons (exactly 13 nucleotides). If the
47
48 decoding unit at the ribosome is the anticodon with adjacent nucleotides then the general circular codes
49
50 could also immediately return to the reading phase. Does the self-complementary property of the *X*
51
52 circular code contribute to coordination between *X* motifs in mRNA and *X* motifs in tRNA and/or rRNA?
53
54 So far we have mainly discussed the effects of codon choices on the throughput of translation, but
55
56 changes in the translation elongation process can clearly affect translation fidelity and accuracy,
57
58
59
60

1 reviewed in ^[63]. For example, clustering of rare codons could deplete cognate tRNAs, increasing the
2 probability of a near- or non-cognate tRNA occupying the decoding site, and this probability could be
3
4 reflected in the frequency of miss-incorporation. In this case, it has been shown that the standard genetic
5
6 code minimizes the impact of the mutations on the translated protein ^[56]. Clustering of identical rare
7
8 codons also increases the probability of a frameshift during translation. Ribosome stalling at *Lys* codons
9
10 triggers ribosome sliding on successive *AAA* codons. When ribosomes resume translation, they may shift
11
12 in an incorrect reading frame. The ribosomes translating in the -1 or +1 frame usually quickly encounter
13
14 out-of-frame stop codons that result in termination. Again, it has been suggested that the genetic code
15
16 might be in some way optimized for frameshift mutations ^[64]. Given the inherent error correcting
17
18 properties of circular codes, it is possible that the *X* circular code may play a role in the synchronization
19
20 of the correct reading frame.
21
22

23
24 In the future, we hope to show that the simple parameter defined in this work to estimate the coverage
25
26 of *X* motifs in genes is a useful factor that should be taken into account in codon optimization strategies
27
28 or other experimental approaches involving gene expression. We also plan to investigate more complex
29
30 parameters linked to *X* motifs, such as localized density patterns within specific regions of the genes.
31
32
33
34
35

36 **Acknowledgements**

37
38 This work was supported by Institute funds from the French Centre National de la Recherche
39
40 Scientifique and the University of Strasbourg. The authors would like to thank the BISTRO and BICS
41
42 Bioinformatics Platforms for their assistance. This work was supported by French Infrastructure
43
44 Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013, and the ANR Grants Elixir-Excelsior:
45
46 GA-676559 and RAinRARE: ANR-18-RAR3-0006-02. The authors declare that there are no conflicts of
47
48 interest.
49
50
51
52
53
54
55
56
57
58
59
60

References

1. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. *Nature*. **1961**, *192*, 1227–1232.
2. Eslami-Mossallam B, Schram RD, Tompitak M, van Noort J, Schiessel H. Multiplexing genetic and nucleosome positioning codes: A computational approach. *PLoS One*. **2016**, *11*, e0156905.
3. Prakash K, Fournier D. Evidence for the implication of the histone code in building the genome structure. *Biosystems*. **2018**, *164*, 49-59.
4. Baralle M, Baralle FE. The splicing code. *Biosystems*. **2018**, *164*, 39-48.
5. Cakiroglu SA, Zaugg JB, Luscombe NM. Backmasking in the yeast genome: encoding overlapping information for protein-coding and RNA degradation. *Nucleic Acids Research*. 2016, *44*, 8065-8072.
6. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol Cell*. **2015**, *59*, 744-754.
7. Weatheritt RJ, Babu MM. Evolution. The hidden codes that shape protein evolution. *Science*. **2013**, *342*, 1325-1326.
8. Maraia RJ, Iben JR. Different types of secondary information in the genetic code. *RNA*. **2014**, *20*, 977-984.
9. Bergman S, Tuller T. Widespread non-modular overlapping codes in the coding regions. *Phys Biol*. **2020**, *1088*, 1478-3975/ab7083.
10. Babbitt GA, Coppola EE, Mortensen JS, Ekeren PX, Viola C, Goldblatt D, Hudson AO. Triplet-based codon organization optimizes the impact of synonymous mutation on nucleic acid molecular dynamics. *J Mol Evol*. **2018**, *86*, 91-102.
11. Arquès DG, Michel CJ. A complementary circular code in the protein coding genes. *J Theor Biol*. **1996**, *182*, 45-58.
12. Michel CJ. A 2006 review of circular codes in genes. *Computer and Mathematics with Applications* **2008**, *55*, 984-988.
13. Fimmel E, Strüngmann L. Mathematical fundamentals for the noise immunity of the genetic code. *Biosystems*. **2018**, *164*, 86-198.
14. Crick FH, Griffith JS, Orgel LE. Codes without commas. *Proc Natl Acad Sci U SA*. **1957**, *43*, 416-421.
15. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U SA*. **1961**, *47*, 1588-1602.
16. Michel CJ. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. *Life*. **2017**, *7*, 20.
17. Frey G, Michel CJ. Circular codes in archaeal genomes. *J Theor Biol* **2003**, *223*, 413-431.
18. Frey G, Michel CJ. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. *Comp Biol Chem* **2006**, *30*, 87-101.
19. Grosjean H, Westhof E. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res*. **2016**, *44*, 8020-8040.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
20. Michel CJ, Thompson JD. Identification of a circular code periodicity in the bacterial ribosome, origin of codon periodicity in genes? *RNA Biol.* **2020**, 17, 571-583.
21. Fimmel E, Michel CJ, Pirot F, Sereni J-S, Strüngmann L. Mixed circular codes. *Math Biosci* **2019**, 317, 108231.
22. El Soufi K, Michel CJ. Circular code motifs in genomes of eukaryotes. *J Theor Biol.* **2016**, 408, 198-212.
23. Michel CJ, Ngoune VN, Poch O, Ripp R, Thompson JD. Enrichment of circular code motifs in the genes of the yeast *Saccharomyces cerevisiae*. *Life.* **2017**, 7.
24. Dila G, Michel CJ, Poch O, Ripp R, Thompson JD. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. *Biosystems.* **2019**, 175, 57-74.
25. Michel CJ. Circular code motifs in transfer RNAs. *Comp Biol Chem.* **2013**, 45, 17-29.
26. Michel CJ. Circular code motifs in transfer and 16S ribosomal RNAs, a possible translation code in genes. *Comp Biol Chem.* **2012**, 37, 24-37.
27. Dila G, Ripp R, Mayer C, Poch O, Michel CJ, Thompson JD. Circular code motifs in the ribosome, a missing link in the evolution of translation? *RNA* **2019**, 25, 1714-1730.
28. Demongeot J, Seligmann H. Spontaneous evolution of circular codes in theoretical minimal RNA rings. *Gene* **2019**, 705, 95-102.
29. Demongeot J, Moreira A. A possible circular RNA at the origin of life. *J Theor Biol.* **2007**, 249, 314-324.
30. Demongeot J, Seligmann H. The uroboros theory of life's origin, 22-nucleotide theoretical minimal RNA rings reflect evolution of genetic code and tRNA-rRNA translation machineries. *Acta Biotheor.* **2019**, 67, 273-297.
31. Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, Yao J, Khokha MK, Giraldez AJ. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.* **2016**, 35, 2087-2103.
32. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Mol Cell Biol.* **2018**, 19, 20-30.
33. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell.* **2016**, 166, 679-690.
34. Diambra LA. Differential bicodon usage in lowly and highly abundant proteins. *PeerJ.* **2017**, 5, e3081.
35. Guo FB, Ye YN, Zhao HL, Lin D, Wei W. Universal pattern and diverse strengths of successive synonymous codon bias in three domains of life, particularly among prokaryotic genomes. *DNA Res.* **2012**, 19, 477-485.
36. Clarke TF, Clark PL. Rare codons cluster. *PLoS One.* **2008**, 3, e3412.
37. Brar GA. Beyond the triplet code, Context cues transform translation. *Cell.* **2016**, 167, 1681-1692.

- 1
2 38. Chevance FFV, Hughes KT. Case for the genetic code as a triplet of triplets. *Proc Natl Acad Sci U SA*.
3 **2017**, 114, 4745-4750.
- 4
5 39. Sharma AK, Sormanni P, Ahmed N, Ciryam P, Friedrich UA, Kramer G, O'Brien EP. A chemical kinetic
6 basis for measuring translation initiation and elongation rates from ribosome profiling data. *PLoS*
7 *Comp Biol*. **2019**, 15, e1007070.
- 8
9 40. Koonin EV. How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics*
10 *Hum Genet*. **2000**, 1, 99-116.
- 11
12 41. Nevers Y, Kress A, Defosset A, Ripp R, Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector
13 3.0: open portal for comparative genomics. *Nucleic Acids Res*. **2019**, 47, D411-D418.
- 14
15 42. Leder C, Kleinschmidt JA, Wiethe C, Müller M. Enhancement of capsid gene expression: Preparing
16 the human papillomavirus type 16 major structural gene L1 for DNA vaccination purposes. *J Virol*.
17 **2001**, 75, 9201-9209.
- 18
19 43. Warzecha H, Mason HS, Lane C, Tryggvesson A, Rybicki E, Williamson AL, Clements JD, Rose RC.
20 Oral immunogenicity of human papillomavirus-like particles expressed in potato. *J Virol*. **2003**, 77,
21 8702-8711.
- 22
23 44. Jack BR, Boutz DR, Paff ML, Smith BL, Bull JJ, Wilke CO. Reduced protein expression in a virus
24 attenuated by codon deoptimization. *G3 (Bethesda)*. 2017, 7, 2957-2968.
- 25
26 45. Bull JJ, Molineux IJ, Wilke CO. Slow fitness recovery in a codon-modified viral genome. *Mol Biol Evol*.
27 **2012**, 29,2997-3004.
- 28
29 46. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. *Mol Syst Biol*. **2011**, 7, 481.
- 30
31 47. Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, Sachs MS, Liu Y. Non-optimal codon usage affects
32 expression, structure and function of FRQ clock protein *Nature*. **2013**, 495, 111-115.
- 33
34 48. Charneski CA, Hurst LD. Positively charged residues are the major determinants of ribosomal
35 velocity. *PLoS Biol*. **2013**, 11, e1001508.
- 36
37 49. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. Measurement of average decoding rates
38 of the 61 sense codons in vivo. *Elife*. **2014**, 3.
- 39
40 50. Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. Improved ribosome-
41 footprint and mRNA measurements provide insights into dynamics and regulation of yeast
42 translation. *Cell Rep*. **2016**, 14, 1787-1799.
- 43
44 51. Wu CC, Zinshteyn B, Wehner KA, Green R. High-resolution ribosome profiling defines discrete
45 ribosome elongation states and translational regulation during cellular stress. *Mol Cell*. **2019**, 73,
46 959-970.e5.
- 47
48 52. Brule CE, Grayhack EJ. Synonymous codons: Choose wisely for expression. *Trends in Genetics*. **2017**,
49 33, 283-297.
- 50
51 53. Villada JC, Brustolini OJB, Batista da Silveira W. Integrated analysis of individual codon contribution
52 to protein biosynthesis reveals a new approach to improving the basis of rational gene design. *DNA*
53 *Res*. **2017**, 24, 419-434.
- 54
55
56
57
58
59
60

- 1
2 54. Mignon C, Mariano N, Stadthagen G, Lugari A, Lagoutte P, Donnat S, Chenavas S, Perot C, Sodoyer R,
3 Werle B. Codon harmonization - going beyond the speed limit for protein expression. *FEBS Letts.*
4 **2018**, 592, 1554-1564.
5
6 55. Diamant A, Feldman A, Schochet E, Kupiec M, Arava Y, Tuller T. The extent of ribosome queuing in
7 budding yeast. *PLoS Comp Biol.* **2018**, 14, e1005951.
8
9 56. Boël G, Letso R, Neely H, Price WN, Wong KH, Su M, Luff J, Valecha M, Everett JK, Acton TB, Xiao R,
10 Montelione GT, Aalberts DP, Hunt JF. Codon influence on protein expression in *E. coli* correlates
11 with mRNA levels. *Nature.* **2016**, 529, 358-363.
12
13 57. Błażej P, Wnętrzak M, Mackiewicz D, Mackiewicz P. Optimization of the standard genetic code
14 according to three codon positions using an evolutionary algorithm. *PLoS one.* **2018**, 13, e0201715.
15
16 58. Rodnina MV. The ribosome in action: Tuning of translational efficiency and protein folding. *Protein*
17 *Science.* **2016**, 25, 1390-1406.
18
19 59. Chen YH, Collier J. A universal code for mRNA stability? *Trends in Genetics.* **2016**, 32, 687-688.
20
21 60. Wu B, Zhang H, Sun R, Peng S, Cooperman BS, Goldman YE, Chen C. Translocation kinetics and
22 structural dynamics of ribosomes are modulated by the conformational plasticity of downstream
23 pseudoknots. *Nucleic Acids Res.* **2018**, 46, 9736-9748.
24
25 61. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, and Liu Y. Codon usage is an important
26 determinant of gene expression levels largely through its effects on transcription *Proc Natl Acad Sci*
27 *USA.* **2016**, 113, E6117-E6125.
28
29 62. Hanson G, Alhusaini N, Morris N, Sweet T, Collier J. Translation elongation and mRNA stability are
30 coupled through the ribosomal A-site. *RNA.* 2018, 24, 1377-1389.
31
32 63. Liu Y, Sharp J.S, Do DH, Kahn RA, Schwalbe H, Buhr F, Prestegard JH. Mistakes in translation:
33 Reflections on mechanism. *PLoS One.* **2017**, 12, e0180566.
34
35 64. Geyer R, Madany Mamlouk A. On the efficiency of the genetic code after frameshift mutations. *PeerJ.*
36 **2018**, 6, e4825.
37
38
39
40
41
42
43

Figure Legends

44
45
46
47 Fig 1. a) Definition of the *X* circular code. b) Circular representation of the genetic code, adapted from
48 [19], with the 20 codons of the *X* circular code shown on the circumference. The numbers after the
49 nucleotides indicate their position in the codon. *X* codons that are complementary to each other are
50 highlighted in the same color. c) Retrieval of the reading frame in a *X* motif constructed with the *X*
51 circular code. Codons belonging to the *X* circular code are indicated in blue, while non-*X* codons are
52 shown in red. Among the three possible frames, only the reading frame 0 contains codons of the *X*
53 circular code exclusively.
54
55
56
57
58
59
60

1
2
3
4 Fig 2. Optimal codons for translation elongation rate and mRNA stability in different eukaryotic species
5
6 (*S. cerevisiae*, zebrafish, *Xenopus* and *Drosophila*). Codons are ordered according to their mean ranking
7
8 obtained in four different experiments. Codons belonging to the X code are identified by a blue star.
9

10
11
12 Fig 3. Density of X motifs in the mRNA sequences of the 'minimal gene set'. The distributions of the
13
14 number of X motifs identified in the sequences from the three domains of life are indicated by boxplots
15
16 representing the mean number with a ± 0.99 confidence interval. The distributions of the number of R
17
18 random motifs (see Supplementary Materials) identified in the same sequences are shown for statistical
19
20 evaluation. There is a very strong statistical significance as confirmed by a one-sided Student's *t*-test
21
22 with a *p*-value $p < 10^{-100}$ for each set of sequences from archaea, bacteria and eukaryota.
23
24
25

26
27 Fig 4. Histogram of the density of X motifs in the recoded version of the gene 10A from *Escherichia coli*
28
29 K-12, compared to the wild type sequence. The orange plot indicates the viral fitness values
30
31 corresponding to each construct.
32
33

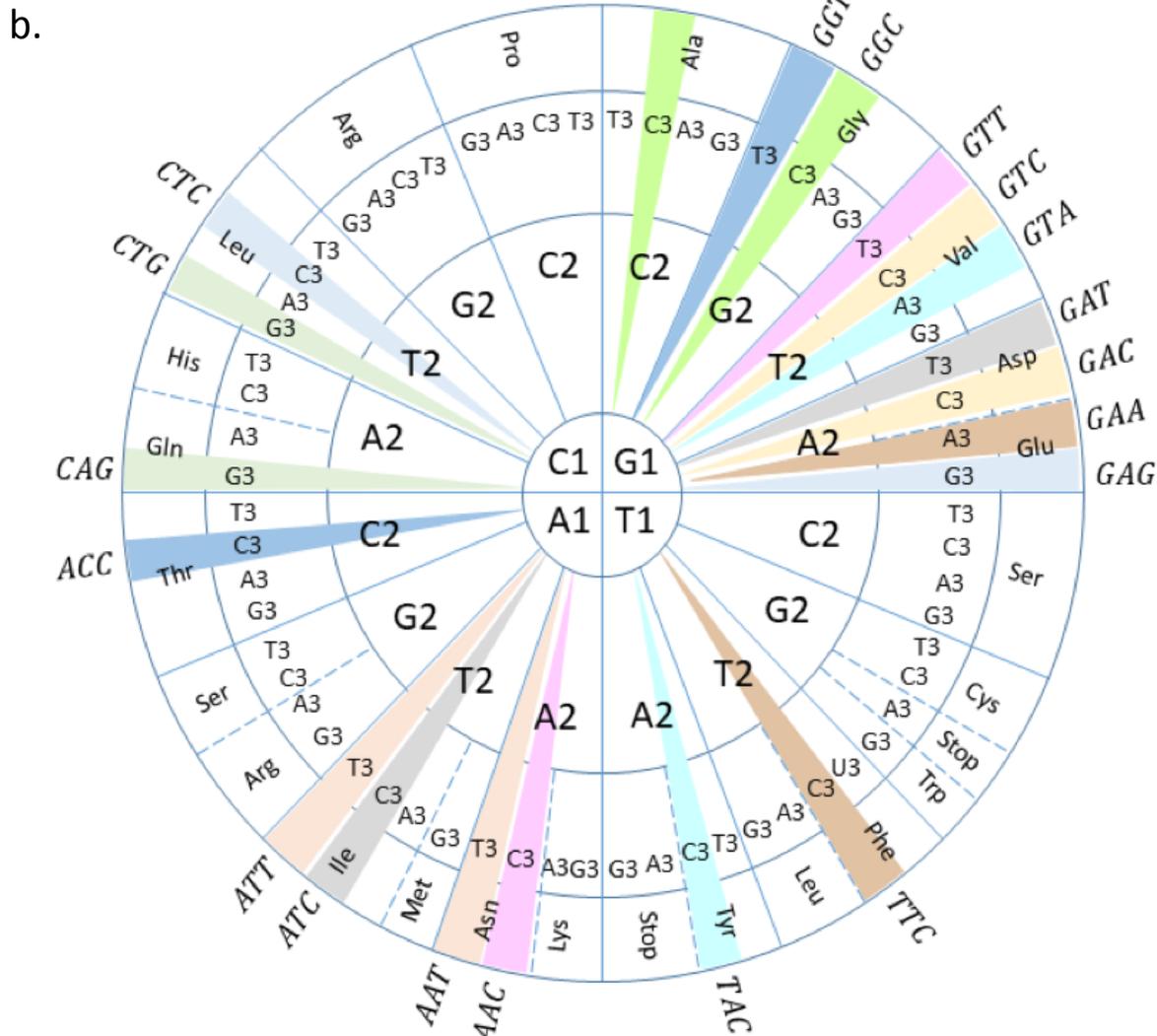
34
35 Fig 5. Density of X motifs for *S. cerevisiae* genes: 1323 genes with low translation rates (estimated
36
37 translation rate < 0.03), 1378 genes with medium translation rates (estimated translation rate 0.05-
38
39 0.09) and 1324 genes with high translation rates (estimated translation rate > 1.1). The distributions of
40
41 the number of X motifs identified in the genes are indicated by boxplots representing the mean number
42
43 with a ± 0.99 confidence interval. The statistical significance is confirmed by two one-sided Student's *t*-
44
45 tests with: $p < 10^{-10}$ between the sequences with medium translation rates and those with low
46
47 translation rates; and $p < 10^{-14}$ between the sequences with high translation rates and those with medium
48
49 translation rates.
50
51
52

53
54 Fig 6. a) Histogram of the density of X motifs for different constructs corresponding to the *S. cerevisiae*
55
56 HIS3 gene with 0-100% optimized codons. The orange plot indicates the mRNA half-life values
57
58 corresponding to each construct. b) Density of X motifs in the different constructs corresponding to the
59
60

1
2 *Drosophila* luciferase gene. Sequence regions shown in blue are codon optimized, and in red are the wild
3
4 type sequence. The numbers above the sequences indicate the codon positions of the optimized regions.
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

- a. X circular code contains 20 codons:
 {AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}
- X codes for 12 amino acids:
 {Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val}



- c.
- f_0 ATG ... **G G T A A T T A C G A G** ... TAA
- f_1 ATG ... G **G T A A T T A C G** A G ... TAA
- f_2 ATG ... G G **T A A T T A C G A** G ... TAA

Figure 1

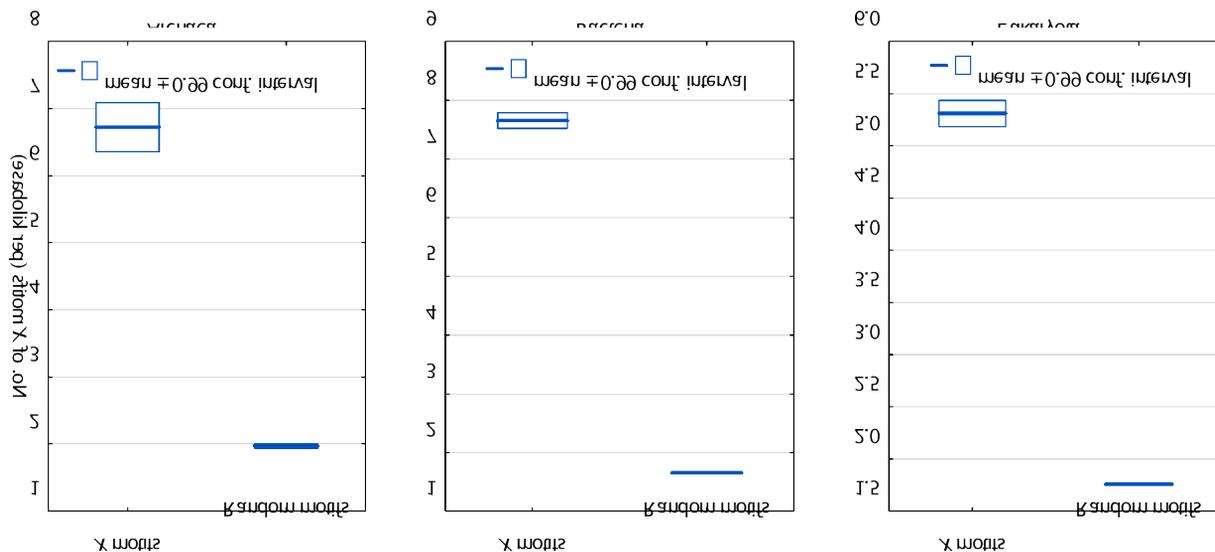


Figure 3

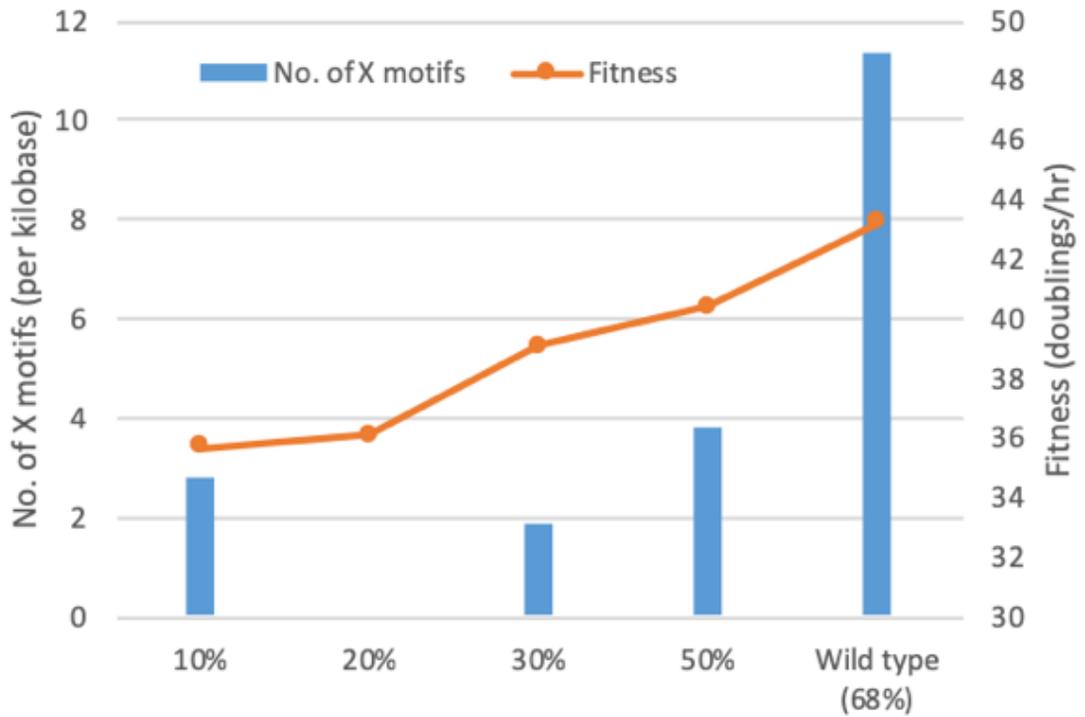


Figure 4

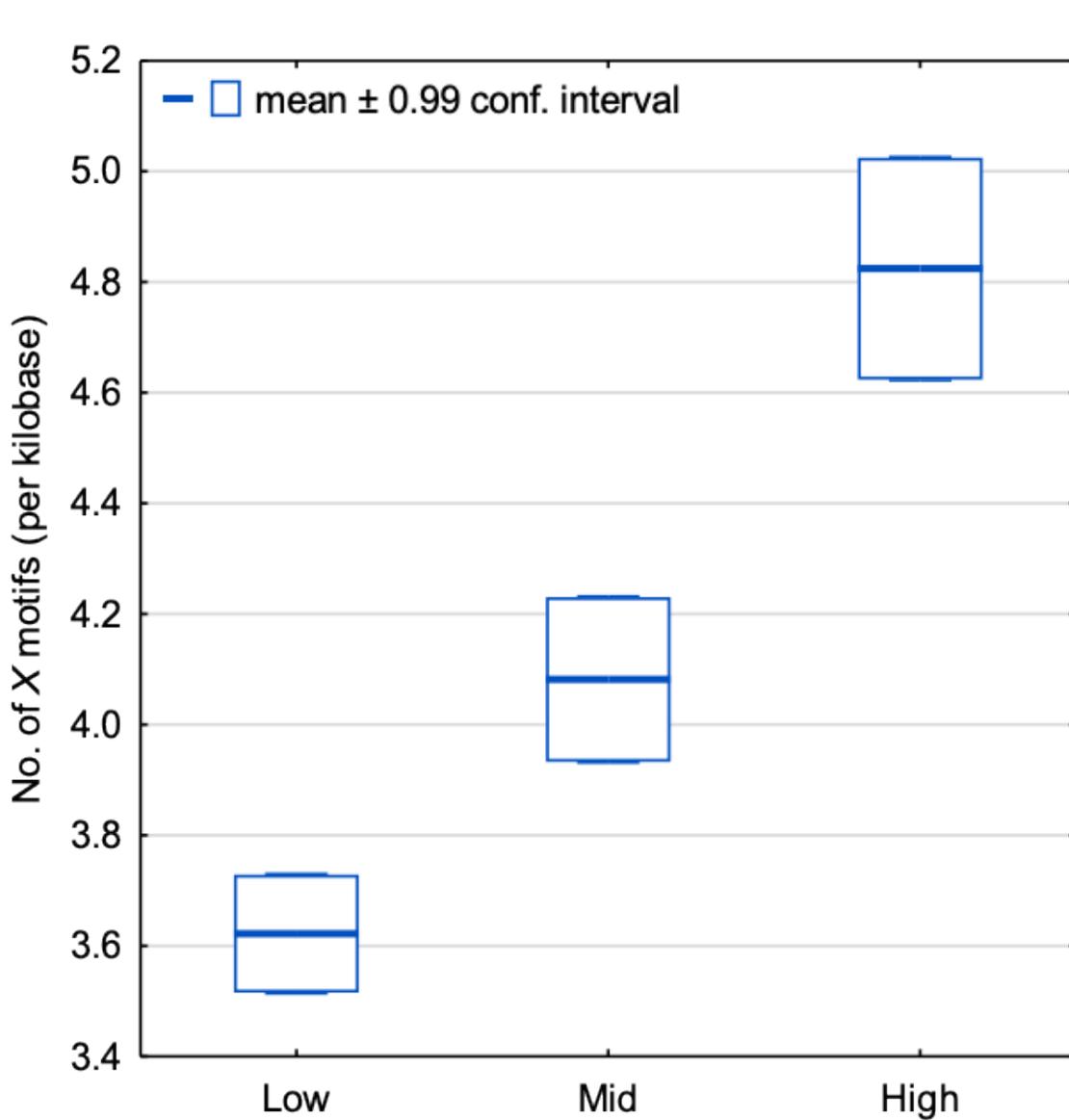


Figure 5

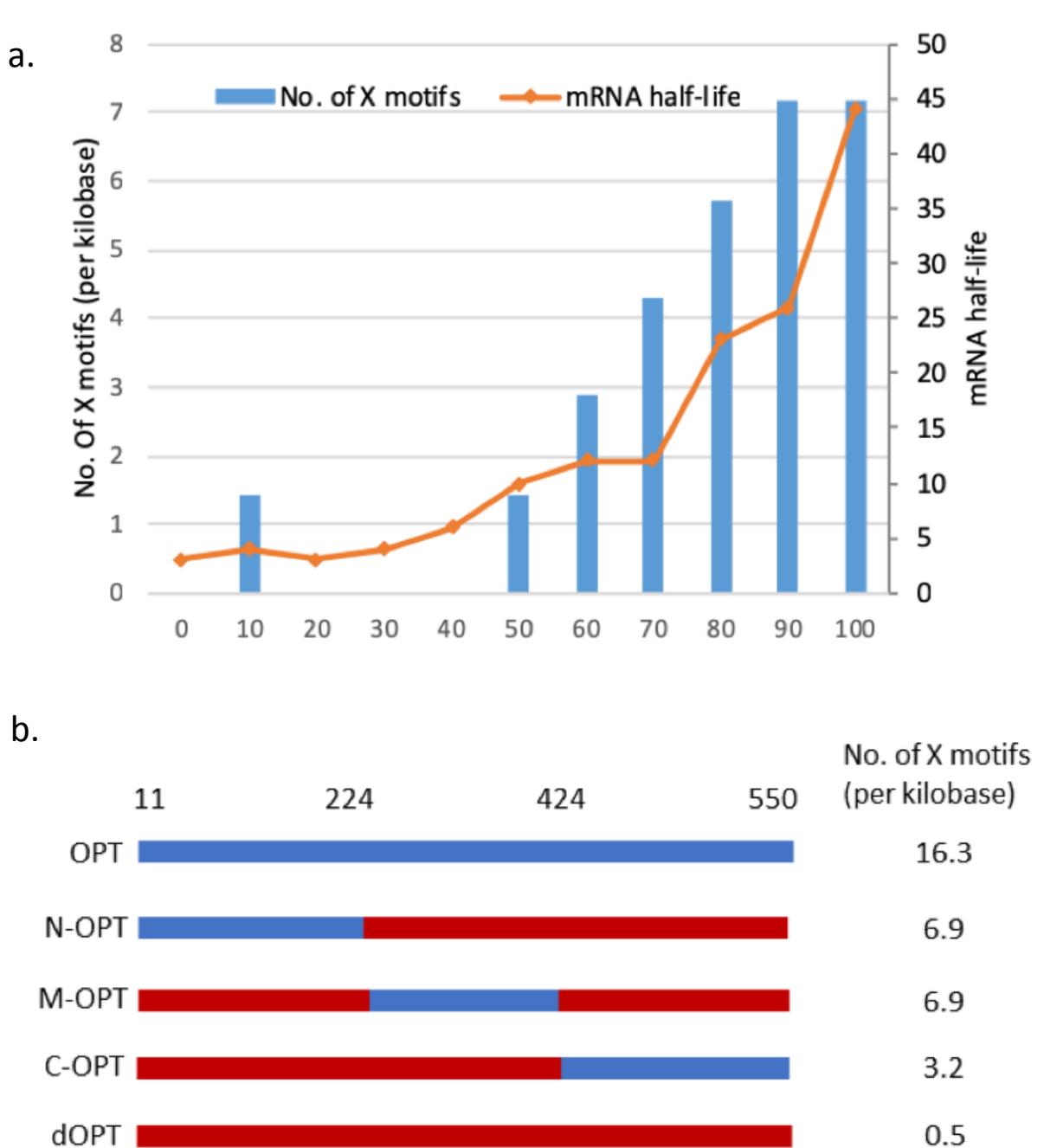


Figure 6