the plant journal

S E B SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **112,** 919–945 doi: 10.1111/tpj.15970

Mechanisms governing codon usage bias and the implications for protein expression in the chloroplast of *Chlamydomonas reinhardtii*

Maxime Fages-Lartaud[1,*], **Kristoffer Hundvin**[1] and **Martin Frank Hohmann-Marriott**[2]

[1] *Department of Biotechnology, Norwegian University of Science and Technology, Trondheim N-7491, Norway, and*
[2] *United Scientists CORE (Limited), 2 Tewsley St, Dunedin 9016, New Zealand*

Received 8 March 2022; revised 29 August 2022; accepted 1 September 2022; published online 7 September 2022.

*For correspondence (e-mail fages.maxime@outlook.fr).

SUMMARY

Chloroplasts possess a considerably reduced genome that is decoded via an almost minimal set of tRNAs. These features make an excellent platform for gaining insights into fundamental mechanisms that govern protein expression. Here, we present a comprehensive and revised perspective of the mechanisms that drive codon selection in the chloroplast of *Chlamydomonas reinhardtii* and the functional consequences for protein expression. In order to extract this information, we applied several codon usage descriptors to genes with different expression levels. We show that highly expressed genes strongly favor translationally optimal codons, while genes with lower functional importance are rather affected by directional mutational bias. We demonstrate that codon optimality can be deduced from codon–anticodon pairing affinity and, for a small number of amino acids (leucine, arginine, serine, and isoleucine), tRNA concentrations. Finally, we review, analyze, and expand on the impact of codon usage on protein yield, secondary structures of mRNA, translation initiation and termination, and amino acid composition of proteins, as well as cotranslational protein folding. The comprehensive analysis of codon choice provides crucial insights into heterologous gene expression in the chloroplast of *C. reinhardtii*, which may also be applicable to other chloroplast-containing organisms and bacteria.

Keywords: chloroplast, genetic code, codons, codon usage bias, tRNA anticodon, protein expression.

INTRODUCTION

Codons are the fundamental link between genes and proteins. This position makes codons the target of a complex array of evolutionary forces. By analyzing the interactions of these forces, we may uncover fundamental connections and learn how to apply them in biotechnological applications.

Genetic information is universally encoded in DNA and RNA through sequences of nucleotides (A, T/U, C, and G). The genetic information encoding proteins is organized into nucleotide triplets called codons, offering $4^3 = 64$ possible combinations for encryption. To translate DNA into proteins, 61 triplets code are used for the 20 canonical amino acids and three are used to terminate translation (UAA, UAG, and UGA). The excess of possible nucleotide triplet combinations versus the number of codable amino acids leads to redundancy of the genetic code, where one amino acid is encoded by several codons (see Figure 1a). The segregation of the genetic code into codon families or 'boxes' is a consequence of nucleotide base pairing rules. Codon decryption is achieved through specific pairing with the anticodon of a tRNA that carries a specific amino acid. The codon bases at positions 1, 2, and 3 pair with positions N_{36}, N_{35}, and N_{34} of the anticodon loop, respectively (anticodon positions 3, 2, and 1, respectively) (see Figure 1b). Codon–anticodon recognition follows Watson–Crick pairing rules (A:U, U:A, G:C, C:G) for the first and second positions of the codon with bases N_{36} and N_{35} of the anticodon. In contrast, the interaction between the third codon position and the first anticodon base (N_{34}) is less specific and follows an extended set of combinations expressed in the 'wobble rules' (Agris, 2004; 1991; Crick, 1966). In addition, a plethora of nucleotide modifications in the anticodon loop, especially in the wobble position N_{34} and anticodon adjacent N_{37}, modulate codon discrimination (Agris, 2008; Osawa et al., 1992) by increasing or decreasing codon–anticodon pairing specificity. Therefore, with the exception of methionine and tryptophan, amino acids are often associated with several near-cognate codons in duet, triplet, or quartet boxes, or even possess two decoding boxes (e.g., serine, leucine, and arginine) (see Figure 1a).

The precision of the genetic code relies on the supply of correctly charged aminoacyl-tRNAs. Amino acids are ligated onto their corresponding tRNAs by specific aminoacyl-tRNA synthetases. This specificity constitutes a crucial step in preserving the fidelity of the genetic code. Mature aminoacyl-tRNAs enter the ribosomal A-site and pair with anticodons by complementarity rules. Accurate codon–anticodon pairing allows movement into the ribosome P-site. Here, a peptidyl bond is created with the previous amino acid of the nascent polypeptide chain (see Figure 1b). The tRNA exits the ribosome and the process is iterated to create mature proteins.

Historically, synonymous codons were considered equivalent because they resulted in 'silent' mutations with no consequences on the protein sequence. However, although the genetic code is nearly universal, it became evident that its utilization varies across species depending on the deciphering optimization strategy imposed by evolutionary forces (Grosjean et al., 2010; Grosjean & Westhof, 2016). In consequence, codon usage profiles differ widely across species. Furthermore, codon usage differs within the same species and correlates with gene expression in unicellular organisms (Bennetzen & Hall, 1982; Gouy & Gautier, 1982; Lloyd & Sharp, 1992; Sharp & Cowe, 1991; Sharp & Li, 1986). The species preference for certain redundant codons is known as codon usage bias. Prokaryotes and eukaryotes developed increasingly complex regulatory mechanisms of their genetic codes that are reflected in codon usage bias. Codon usage bias is a consequence of a multi-layered complexity combining genetic code deciphering properties with elaborated regulatory mechanisms that span the entire protein expression process. The deconvolution of these interdependencies is not trivial and renders the implications of codon usage for protein expression difficult to understand in their entirety.

Microorganisms with a simple genetic makeup can be useful targets for gaining insights in the complexity of the genetic code. *Mycoplasma*, Archaea, and organelles possess simple versions of the genetic code (Grosjean et al., 2010; Grosjean & Westhof, 2016; Osawa et al., 1990, 1992) and are targets for exploring the mechanisms underlying codon usage bias. For example, *Mycoplasma capricolum* and mitochondria use UGA as a tryptophan codon instead of a stop codon (Grosjean & Westhof, 2016); mitochondria leave the arginine duet-box AGA/G unassigned (Grosjean & Westhof, 2016); while Archaea exhibit a reassignment of the UAG stop codon to the non-canonical amino acid pyrrolysine (Ambrogelly et al., 2007). These deviations from the 'universal' code can be viewed as relics of an ancestral genetic code that is more than a billion years old (Jukes, 1973) or as the result of an evolutionary sparing strategy imposed by small tRNA sets (Grosjean et al., 2010; Grosjean & Westhof, 2016). The ancestral genetic code proposed by Jukes in 1973 (Jukes, 1973) coded for 10 amino acids before expanding to the contemporary code due to increasing encrypting capacities introduced by tRNA modifications. Organelles, such as mitochondria and chloroplasts, present a minimal tRNA set with a reduced complexity in tRNA modifications. This

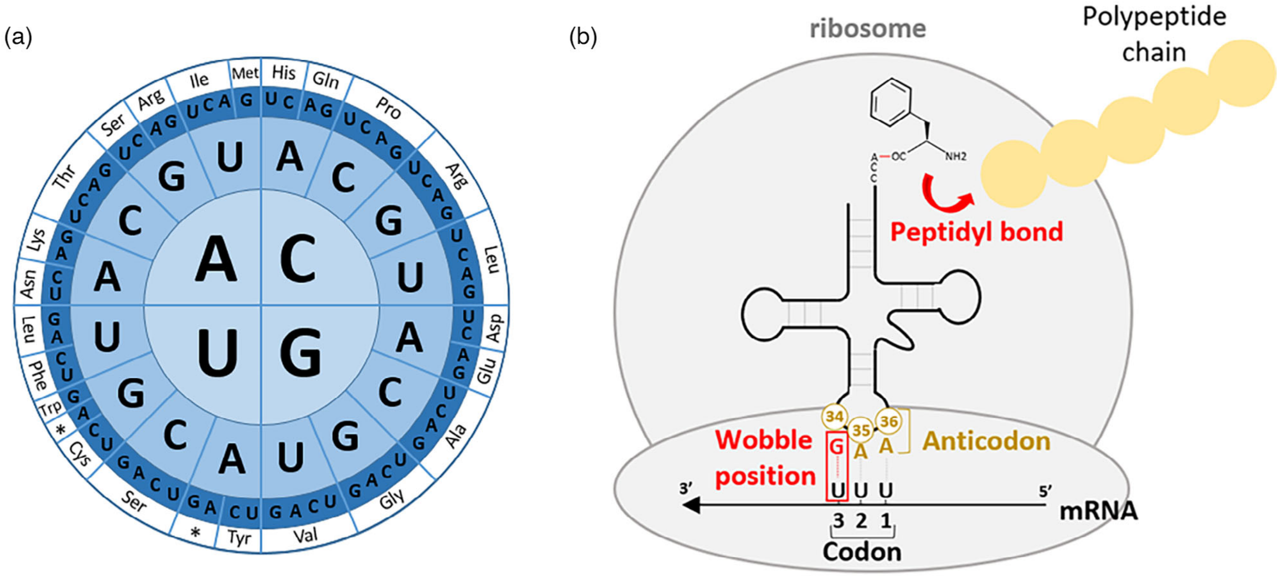

Figure 1. Overview of the genetic code and tRNA–mRNA interactions during translation. (a) Graphic representation of the genetic code. Codons are read from the letter in the center outwards. Corresponding amino acids are indicated by the three-letter code and an asterisk for stop codons. (b) Pairing of a codon with the anticodon of an aminoacyl-tRNA in the ribosome P-site before peptide elongation.

apparent simplicity makes the deciphering of their genetic code highly dependent on physical properties of nucleotides. Thus, organelles like chloroplasts are entities of choice to study the first layers of complexity of the genetic code and its evolution.

In this study, we focus on the chloroplast of the model organism *Chlamydomonas reinhardtii*, although this work may be applicable to other chloroplast-containing organisms due to a common evolution (Suzuki & Morton, 2016). Plant and microalgal plastids represent attractive platforms for biotechnological applications, including the production of energy, therapeutics, animal food, and high-value nutritional and biochemical coproducts for the industry (Almaraz-Delgado et al., 2014; Bock, 2015; Cardi et al., 2010; Doron et al., 2016; Dyo & Purton, 2018; Rosales-Mendoza et al., 2012; Scaife et al., 2015; Scranton et al., 2015; Specht et al., 2010). Biotechnological applications and fundamental research can take advantage of a sophisticated set of genetic tools that have been developed for the chloroplast (Bock, 2015; Doron et al., 2016; Scaife et al., 2015; Wang et al., 2009). Expressing genes is an essential outcome for most biotechnological work in chloroplasts. It is therefore important to understand each aspect of gene expression to realize the biotechnological potential of the chloroplast. The utilization of codon optimization tools in chloroplasts with methodologies established in bacteria (Weiner et al., 2020) ignored some key aspects of codon usage bias and its impact on gene expression that we present in this study.

The evolutionary history of the chloroplast is reflected by the organization of its protein expression machinery. Chloroplasts originate from endocytosis, i.e., the engulfment of an ancient member of the cyanobacterial clade (Douglas & Turner, 1991; Gray, 1989; Martin & Kowallik, 1999). The ensuing symbiotic relationship gave rise to metabolite exchange, deletion of dispensable genes or transfer of essential genes from the plastome to the nucleus, and the subsequent development of a protein import machinery that reroutes nuclear gene products to the chloroplast (Bock, 2015; Scharff & Bock, 2014). Consequently, the polyploid plastome was considerably reduced so its size slightly exceeds a couple hundred kilobases and contains on average 120 genes (Bock, 2015; Dyo & Purton, 2018; Gallaher et al., 2018). The majority of plastid genes are involved in photosynthesis and in the chloroplast's personal transcription/translation apparatus and ATP synthesis (Gallaher et al., 2018; Maul et al., 2002). This small prokaryotic-like genome is maintained probably due to the necessity of protein coexpression with either cofactors or nuclear-encoded counterparts (Stern et al., 2010).

Transcription in the plastid is mediated by two types of RNA polymerases, the nuclear-encoded T7 phage-type polymerases (NEPs) and the eubacterial plastid-encoded polymerases (PEPs) (Hess & Börner, 1999; Shiina et al., 2005). Chloroplastic gene expression is regulated by various nuclear-encoded sigma factors, which activate translation of proteins involved in abiotic stress responses, light and redox signals, and development, depending on promoter type (Barkan, 2011; Kanamaru & Tanaka, 2004). After transcription, mRNAs are processed. Polycistronic transcripts are cleaved into smaller fragments and stabilized by sequence-specific tetra-, penta-, or octatricopeptide repeat (TPR, PPR, and OPR) proteins (Barkan, 2011; Del Campo, 2009; Jalal et al., 2015; Raynaud et al., 2007; Schmitz-Linneweber & Small, 2008; Shikanai & Fujii, 2013; Stern et al., 2010). The chloroplast translation machinery, composed of prokaryotic orthologs, proceeds to translation of mRNA into proteins.

Plastids encode their own almost minimal set of tRNAs, which is often close to the minimal 25-tRNA set required to decipher the genetic code (Alkatib et al., 2012), and there is no evidence of tRNA import from the nucleus (Duchene et al., 2005; Marechal-Drouard et al., 1993). In the chloroplast of *C. reinhardtii*, regulation of protein expression occurs mainly at the translation level (Veronica et al., 2004) by *cis*-elements such as the Shine–Dalgarno (SD) sequence (Scharff et al., 2017; Shine & Dalgarno, 1974; Weiner et al., 2019), mRNA secondary structure (Mauger et al., 2013; Scharff et al., 2011), codon usage (Nakamura & Sugiura, 2007; Pfitzinger et al., 1987), noncoding RNA (Anand & Pandi, 2021; Dietrich et al., 2015), nascent peptide elements (Zoschke & Bock, 2018), and *trans*-elements like the abovementioned PPR proteins.

In this paper, we investigate the effects of codon usage along the entire protein expression process. We enrich our understanding of the functional aspects by including key subtleties to obtain a global and accurate view of codon usage regulation in chloroplasts. We provide an overview of molecular mechanisms that underlie codon usage bias and explore the implications on protein expression. Our work advances the interpretations of codon usage bias in the chloroplast beyond the boundaries of existing literature.

RESULTS AND DISCUSSION

Examination of the codon usage bias fingerprint of the chloroplast

In the chloroplast, codon usage is traditionally analyzed for the total coding sequences (CDSs) of the genome. However, previous studies showed that codon usage can vary widely between functional sets of genes within a single organism (Osawa et al., 1990, 1992). This complexity has often been overlooked or was not investigated in its entirety in the chloroplast. In this section, we will develop and apply codon usage descriptors to define and assess the codon usage bias of the chloroplast in more detail.

Codon usage bias correlates with gene expression

The presence of an intraspecies codon usage bias between genes with different expression levels was observed for a wide range of organisms (Bennetzen & Hall, 1982; Gouy & Gautier, 1982; Lloyd & Sharp, 1992; Sharp & Cowe, 1991; Sharp & Li, 1986). This bias can be estimated by the Codon Adaptation Index (CAI) (Sharp & Li, 1987), in which the codon usage of each gene is compared to the supposed codon optimality of a reference set composed of highly expressed genes (see Methods). Codon usage bias profiles were correlated with tRNA gene copy number, or more accurately tRNA concentrations (Duret, 2000; Ikemura, 1982; 1981), and can be assessed with the tRNA Adaptation Index (tAI) (dos Reis et al., 2004). In the chloroplast, the tRNA gene copy number is either one or two, rendering this correlation inadequate (Data S3). The tRNA reads from RNA sequencing (RNAseq) data (Gallaher et al., 2018) do not accurately represent the mature tRNA population because of the difficulty in detecting such structurally complex RNAs and it ignores aminoacyl-tRNA maturation. Interestingly, quantification from 2D PAGE migration showed a high correlation between amino acid occurrence and tRNA concentrations (Pfitzinger et al., 1987). Although this analysis is valid for leucine, serine, and arginine, all possessing two distinct codon boxes each, and for the special case of isoleucine's three-codon box, it is not suitable to explain codon bias for other amino acids since only one tRNA is responsible for reading all their corresponding codons. Therefore, we used CAI to investigate the presence of a codon usage bias in the chloroplast associated with gene expression.

The CAI of each gene was calculated and plotted against its mRNA expression level (fragments per kilobase of exon per million mapped fragments [FPKM]). Transcript levels were used as a proxy for gene expression, as it was previously shown that transcript levels and protein content often correlate in their native cellular environment (Bennetzen & Hall, 1982; Gouy & Gautier, 1982; Ikemura, 1981; Pfitzinger et al., 1987), despite variations in translation efficiency. A strong correlation (Pearson coefficient = 0.78) was identified between codon adaptation and gene expression (Figure 2). This correlation demonstrates the strong preference for certain codons in highly expressed genes (Top 18 genes). Interestingly, functional groups of genes are distributed according to this correlation. Genes responsible for photosynthesis, such as genes encoding components of photosystems I and II, Rubisco (*rbcL*), cytochrome *b₆f*, and ATP synthase, display both high CAI and high expression. A certain correlation is expected by design, since CAI is based on a reference set that consists of highly expressed genes. However, the aforementioned genes also cluster together when a smaller reference set is used (Top 7 genes), demonstrating the weak effect of the choice of reference group. Only *psbI* is far outside the correlation; this may be explained by a systematic RNAseq under quantification due to particular mRNA instability or secondary structure, or high translation efficiency, protein stability, and protein turnover, or may be related to the function of *PsbI* (Wang et al., 2015). In contrast, genes responsible for chlorophyll synthesis and RNA polymerase (PEP) subunits are the two functional groups with low expression and a significantly different CAI (Low 8 group). The pool of ribosomal proteins is intermediate and appears just slightly biased.

We also performed an analysis of codon usage bias within previously identified operons (Shahar et al., 2019), in an attempt to identify a correlation between codon usage bias and operon expression (Data S1). Some operons show good CAI versus mRNA level correlation (e.g., *rpl2-rps19*, *rpl16-rps14*, *rps18-ycf3*, and *psbJ-atpl-psaJ-rps12*). However, other operons, despite showing similar CAI values, possessed different transcript levels for their respective genes (e.g., *psaC-petL*, *psbT-psbB*, *rps8-psaA_exon1*, *psaA_exon2-psbD*). This suggests the presence of gene-specific promoters within each operon, decoupling the quantitative transcript levels of each gene despite a basal readthrough for the complete operon.

Analysis of codon optimality in the chloroplast

The notion of codon optimality refers to translation efficiency and is different from total codon frequency at the genomic level. This distinction has often been disregarded and led to ambiguous interpretations of the genetic code optimality in the chloroplast (Alkatib et al., 2012; Nakamura & Sugiura, 2007). Since biased, highly expressed genes can constitute a relatively small group in comparison to the total number of CDSs of a cell, the codon optimality information they contain is diluted among the evolutionary constraints of the majority. Ideally, codon demand should consider protein levels quantitatively and include cellular dynamics such as protein stability and turnover, mature aminoacyl-tRNA concentrations, and tRNA modifications. For example, a cell regulates tRNA aminoacylation and tRNA modifications to express different genes under a range of conditions (Jayabaskaran et al., 1990). However, obtaining comprehensive data to analyze this is tremendously demanding. A common simplification is to ignore translation regulation processes and use transcriptomic data (mRNA and tRNA) or genomic data while including a gene expression component (such as CAI) (Bennetzen & Hall, 1982; Gouy & Gautier, 1982; Ikemura, 1981; Pfitzinger et al., 1987). In order to extract this information from the chloroplast, we analyzed the number of each codon, their frequency per 1000 codons, the Relative Synonymous Codon Usage (RSCU), and the relative adaptiveness of a codon (W_{ij}) (see Methods for calculations) (Data S2). This analysis was performed on total CDSs, as well as the high-

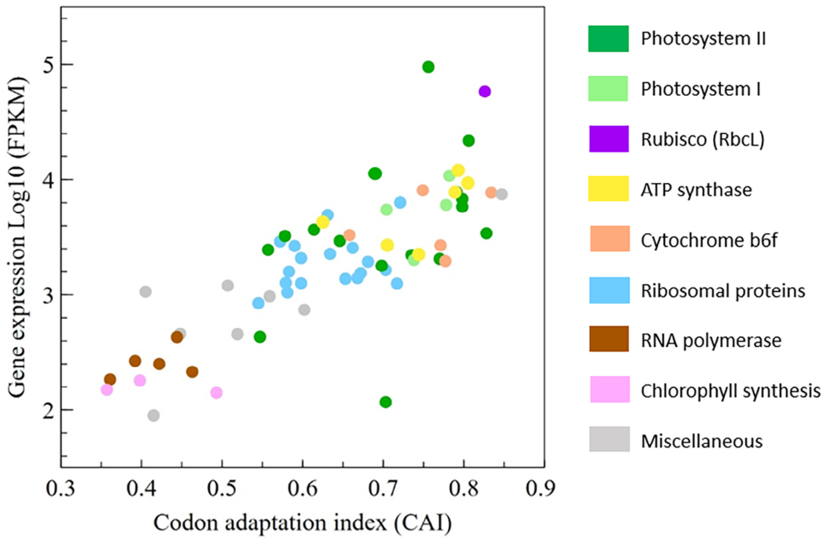

Figure 2. Correlation between gene expression (mRNA FPKM values) and the Codon Adaptation Index (CAI) for CDSs of the *C. reinhardtii* chloroplast (Pearson coefficient 0.78). Categories of functional genes are represented in different colors. Reference set = Top 18 (see Methods).

and low-expression groups (Top 18 and Low 8, respectively). Using the high-expression group, these calculations permit to identify 'optimal' and 'non-optimal' codons as occurring frequently and rarely, respectively. The fold differences between groups were also calculated (Data S2).

The results of the frequency per thousand calculations are presented in Figure 3 and permit to distinguish three categories of codons. The first group is composed of the NNU/C duet boxes (Asn, Asp, Cys, His, Phe, and Tyr). The low-expression group is richer in NNU codons, while high-expression genes show significant opposite enrichment with NNC codons (Figures 3a and 4). Only cysteine is an exception to this rule; in all expression groups UGU is preferred. Cysteine may be a special case due to its low occurrence in proteins (Figure 5), a particular tRNA architecture, or its involvement in protein tertiary structure through the formation of disulfide bridges. Isoleucine AUU/C and serine AGU/C can be included in this group (Figure 3d). Even though the isoleucine codon AUC appears more often in high-expression genes, AUU remains the favored codon for this amino acid across all expression groups. However, its third synonymous codon (AUA) is almost absent from the Top 18 group and relatively common in the Low 8 group, suggesting that another regulatory mechanism is driving codon choice for isoleucine. For the duet box of serine, the AGU codon is less used in high-expression genes, but it is compensated by using favored codons of the quartet box (UCA/U) rather than the AGC codon. The RSCU fold differences for the NNU/C duet boxes suggest that C-ending codons are favored in high-expression genes, while U-ending codons are overrepresented in the low-expression group (Figure 4).

A second group of codons can be identified by the possession of NNA/G duet boxes (Glu, Gln, Lys, Leu [UUA/G], and Arg [AGA/G]). In this group, the A-ending codon is always favored and there is a slight increase in G-ending codons in low-expression genes (Figures 3b and 4). The duet box of arginine (AGA/G) follows the same rule, although its usage is relatively low in comparison to the quartet box (CGN) and almost absent in the Top 18 expression group. The NNA/G duet boxes show less variation across expression groups, suggesting a lower impact on codon regulation compared to the NNU/C duet boxes.

Finally, the group of quartet boxes (Pro, Arg, Leu, Ala, Gly, Val, Ser, and Thr) shows a strong preference for A/U-ending codons regardless of gene expression levels. Nevertheless, G/C-ending codons are more frequent in the low-expression group (Figures 3c,d and 4). This G/C-ending codon bias could have a potential influence on codon regulation even though it might be moderate considering the low occurrence of these codons. Exceptions to these rules are glycine and arginine, which both strongly favor only their respective U-ending codons in the quartet box.

Using a detail-rich analysis our study reveals a codon optimality that deviates from studies that use bulk genome-wide codon usage to determine 'optimality'. Our study suggests that codon optimality is hidden in a relatively small set of highly expressed genes and becomes obscured by including unweighted genome-wide codon usage. The main overlooked codon optimality concerns the group of NNU/C duet boxes that display opposite preference between highly expressed genes and genome-wide codon usage.

In order to refine codon representation in transcripts and aminoacyl-tRNA demand during translation, we estimated the representation of each codon in mRNA by combining sequences with expression data (Data S1). Indeed, even though highly expressed genes are adapted to higher tRNA concentrations (Duret, 2000; Ikemura, 1982; 1981), the codon overrepresentation in highly expressed mRNAs will consume the respective tRNAs at a higher rate, and this is likely to even out the tRNA concentration effects. Since the chloroplast mainly contains one tRNA per codon box, this

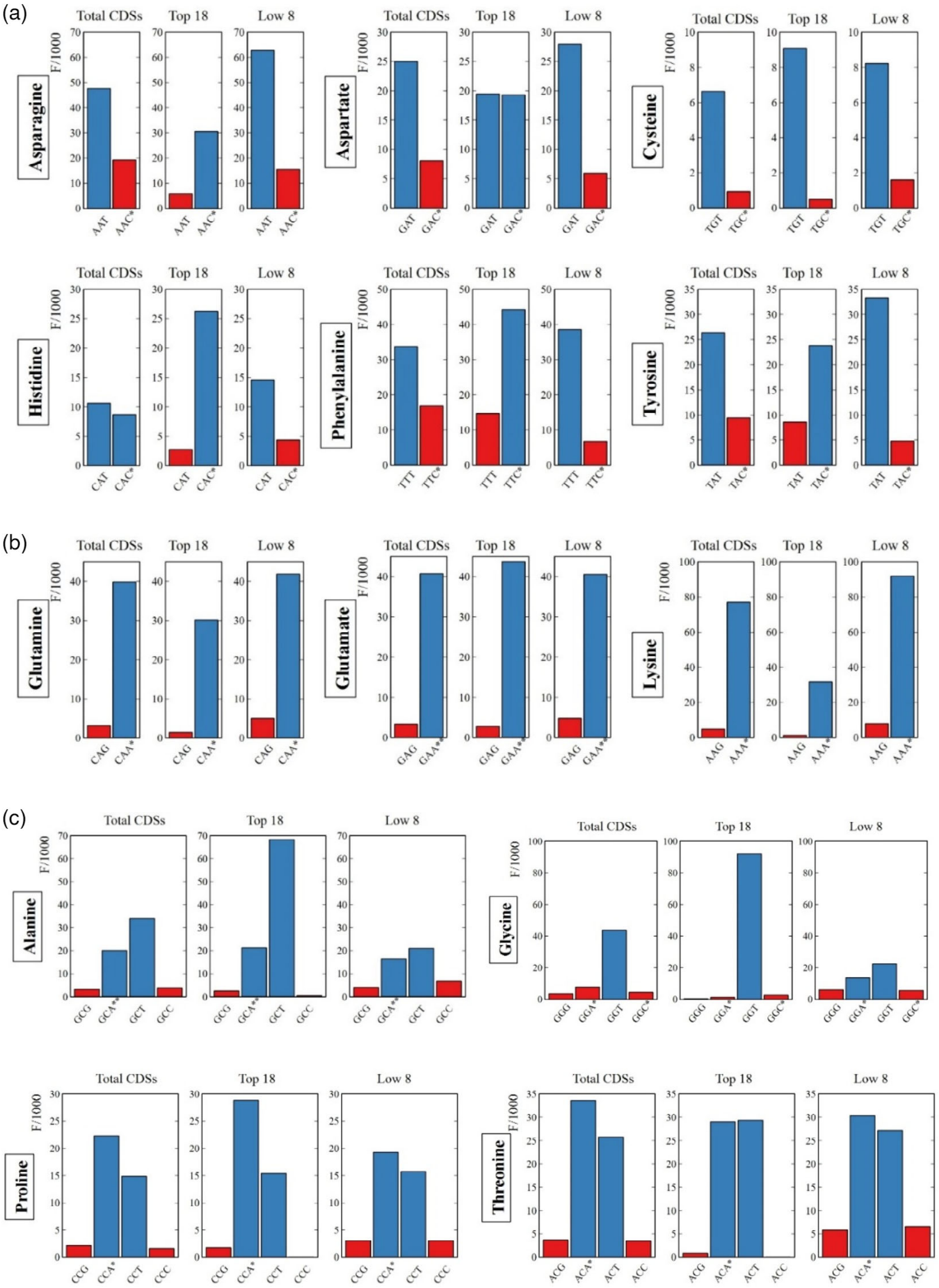

Figure 3. Codon usage frequency statistics. Histograms represent the frequency per thousand of each codon in DNA CDSs. The analysis was made for total CDSs and expression groups Top 18 and Low 8. Blue shows the preferential codon(s) compared to the less preferred codon in red for each group. Asterisks mark the presence and the copy number of corresponding tRNAs. (a) The NNU/C duet boxes show an enrichment in NNC codons in high-expression genes while NNU is preferred in low-expression genes. The total codon sequences hide the optimality of codons of highly expressed genes. (b) For the NNA/G duet boxes, NNA is always favored across all expression groups and NNG is repressed in high-expression genes. (c, d) For quartet boxes, NNA/U codons are always favored (glycine and arginine favor only their respective NNU codon), while NNG/C codons are also repressed in high-expression genes. For amino acids with several tRNAs, we can identify the favored tRNA isoacceptor: the quartet boxes of serine and arginine, the duet box of leucine and isoleucine. (e) Stop codons are almost exclusively TAA; three TAG codons and no TGA codons are present.

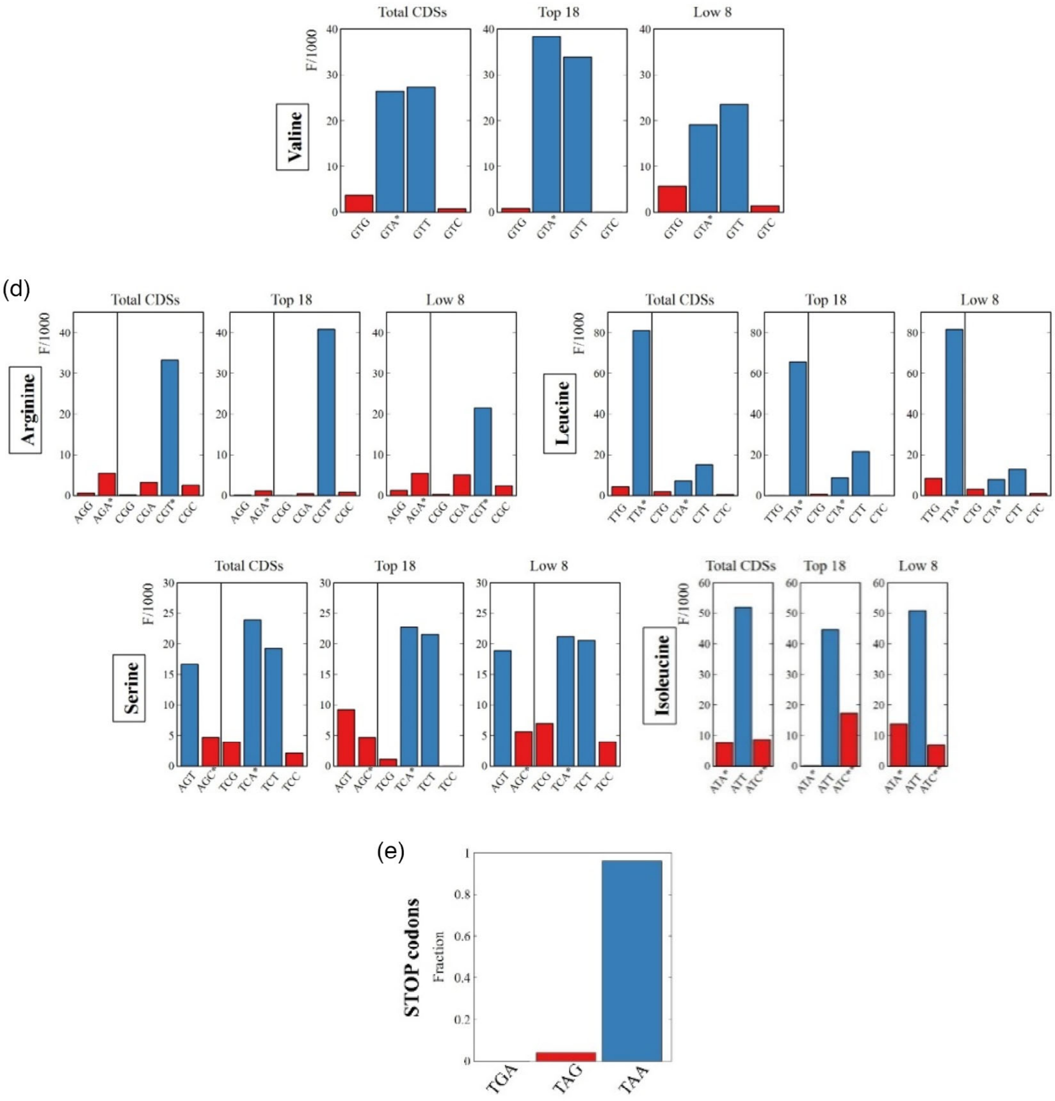

Figure 3. Continued

analysis should support the frequency per thousand and CAI results at the genomic level. As expected, codon demand reinforces the distinction into three codon groups found with genomic frequency per thousand (Data S2). This analysis also provides a finer-grained perception of the preponderance of the bias between expression groups (Data S1). The average bias between Top and Low expression groups for NNU/C duet boxes, NNA/G duet boxes, and quartet boxes is 20.9-, 3.1-, and 13.3-fold, respectively, displaying their effect on translation regulation. We constructed a list of optimal and non-optimal codons and a codon usage table of the Top expression group that can be used for codon optimization of heterologous genes (Data S1).

Overall, optimal codons represent 89.6% of codons in high-expression genes, 74.3% in total CDSs, and only 65.9% in the low-expression group (Figure 6). As demonstrated in this section, codons can be arranged into three groups that have different weights in codon bias. The natural question that ensues concerns the origin of this bias, as well as its function and its biological implications.

Mechanisms responsible for codon usage bias in the chloroplast

In this section, we will explore forces that shape the nucleotide and the codon composition of the plastome. First, we describe how the evolutionary mutational bias directed the nucleotide composition to reach the genomic signature of the chloroplast. Then, we investigate the forces that influenced the codon composition of CDSs to deviate from the genomic nucleotide signature.

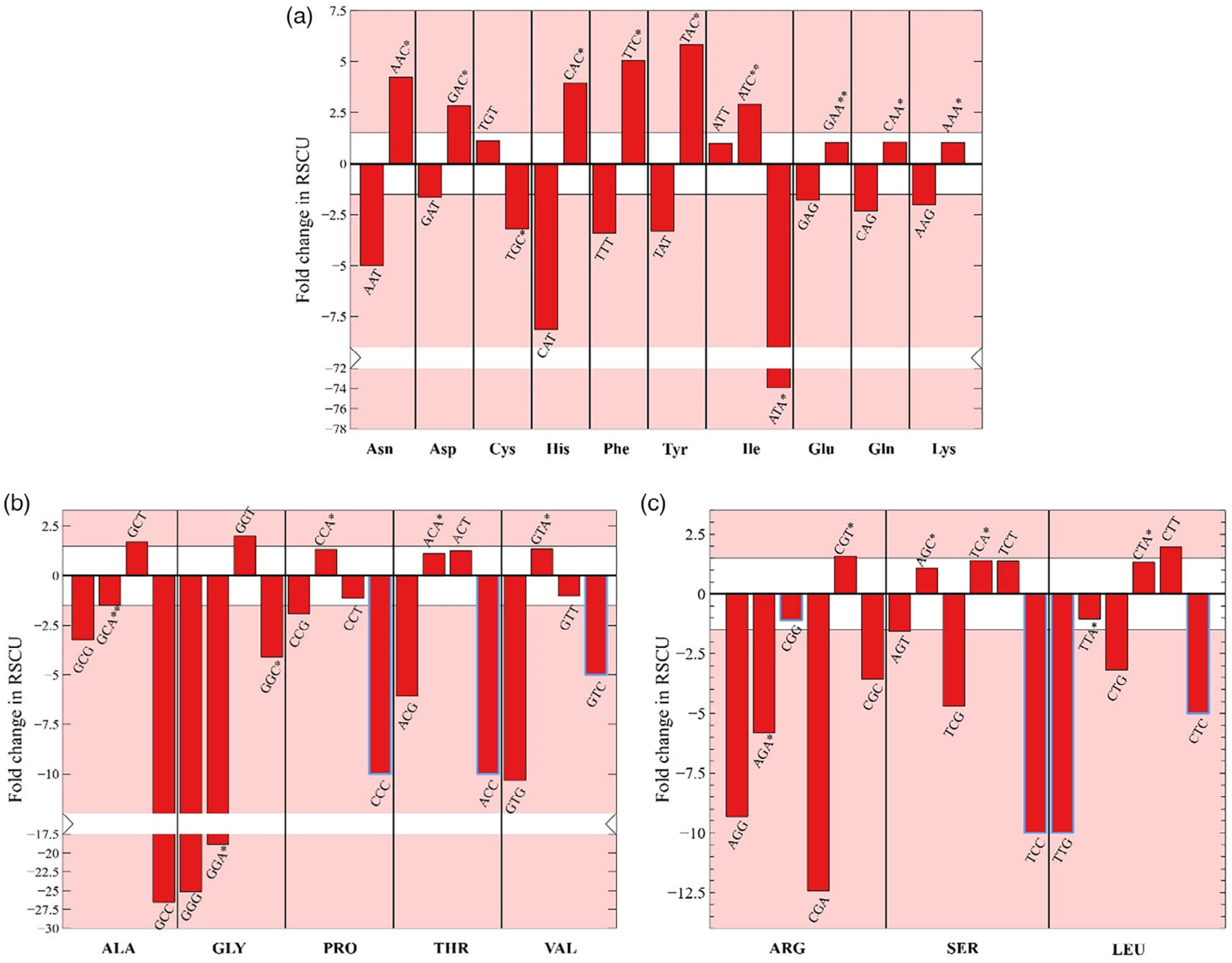

Figure 4. RSCU fold difference. Differences in RSCU for each codon between Top 18 and Low 8 expression groups for (a) duet boxes, (b) quartet boxes, and (c) six-codon boxes are displayed. Asterisks mark the presence of a cognate tRNA and its copy number. The red zones delineate >1.5-fold change in RSCU. Blue rectangles show artificially set fold change values because of the complete absence of the codon in the Top 18 group.

Based on studies in *Escherichia coli* and yeast, Ikemura proposed two evolutionary forces dictating codon selection. He formulated the following rules: (i) when two different tRNA isoacceptors are present for the same amino acid, the codon with the most abundant tRNA is used more frequently; (ii) when only one tRNA reads several codons through wobble base pairing, a higher affinity between codon and anticodon leads to a higher usage (Ikemura, 1982, 1981). We examined the preponderance of both rules for shaping the codon usage of the chloroplast. Finally, we examine potential mechanisms thought to influence the choice of codon juxtaposition.

Evolutionary mutational bias shapes nucleotide composition

The genomic nucleotide composition varies widely across bacterial species from 25 to 75% GC content (Osawa et al., 1990). These variations are the consequence of directional substitution mutation rates (toward A:T or C:G) and the mutational equilibrium represents the genomic signature of an organism (Sueoka, 1988; 1962). Mutation rates are intrinsically associated with the DNA replication and repair systems. On the one hand, replication may generate context-dependent mutations or operate more efficiently on specific sequences (Karlin et al., 1997). On the other hand, variations in composition of DNA repair enzymes, such as the *mut* gene family, can promote transversions from A:T to C:G or inversely (Bai & Lu, 2007; Cabrera et al., 1988; Denamur et al., 2000; Fowler & Schaaper, 1997; Nghiem et al., 1988). This resulting organism-specific GC content is a result of selective constraints exerted to eliminate deleterious mutants and optimize cell growth (Kimura, 1991; Osawa et al., 1992). In opposition to Darwinian evolution processes, Kimura formulated the neutral

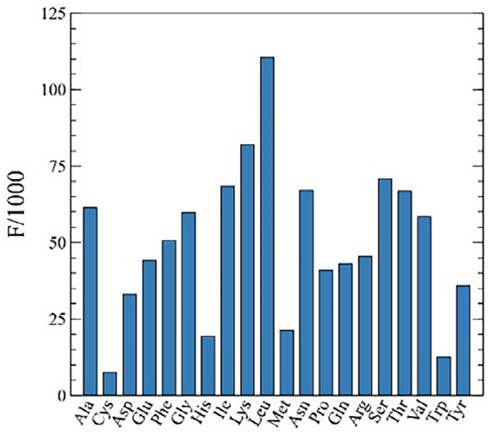

Figure 5. Genome-wide codon frequency. Frequency per thousand of amino acid usage in *C. reinhardtii*'s chloroplast CDSs.

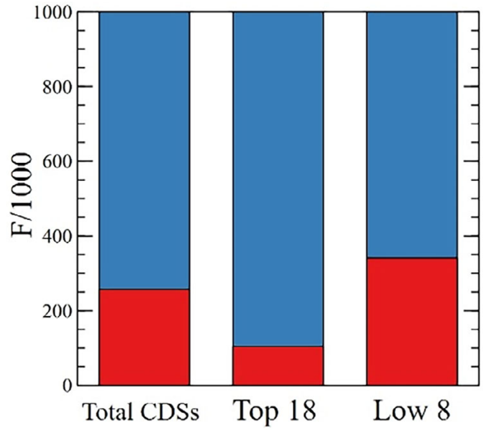

Figure 6. Rare and frequent codons. Frequency per thousand of all rare codons (red) and high-frequency codons (blue) for total CDSs, the Top 18 group, and the Low 8 group.

theory of evolution, stating that genetic regions with lower functional importance evolve through random nucleotide fixation (Kimura, 1991). The neutral sites in question are constituted of genetic spacer regions and protein and DNA polymorphisms (i.e., amino acid similarities [Grantham, 1974; Yampolsky & Stoltzfus, 2005] and third codon positions). This phenomenon is referred to as evolutionary directional mutational bias (Bulmer, 1990; Sharp & Li, 1986). An asymmetric, strand-specific mutational bias has also been described (Lobry & Sueoka, 2002).

It was demonstrated that the GC content of all three codon positions presents a positive linear correlation with genomic GC content (Muto & Osawa, 1987). Under the neutral theory of evolution assumption, the GC content in the third codon position (GC3) should match the expected value from the genomic correlation. The genomic GC content of the chloroplast is 34.5 and 31.9% for CDSs and 36.1% for non-CDS DNA. The expression groups Top 18 and Low 8 present GC contents of 39.0 and 30.9%, respectively, which, based on the neutral theory of evolution, should result in GC3 values close to 34 and 19%, respectively. The GC content was analyzed for all three codon positions across the expression groups (Data S2). The GC3 value of the low-expression group is 16.2 ± 2.5%, which is close to the expected value of 19%. Therefore, the codon usage of genes with lower functional importance follows the rules of the neutral theory of evolution and evolutionary mutational bias. However, the highly expressed gene set possesses a GC3 content of 20.6 ± 7.6%, which is much lower than the expected 34%.

Additionally, the GC3 content between expression groups deviates more drastically from the genomic correlation when considering the three groups of codons established in the previous section (Data S2). While the low-expression group generally follows the neutral theory, it is noteworthy that the NNA/G duet box is slightly underrepresented in NNG codons (11.5 ± 4.5%). This indicates an active selection to decrease NNG codons that could be unfavorable for protein expression. In the case of highly expressed genes, the third codon position analysis draws the same conclusion regarding codon overrepresentation among codon family boxes. In brief, NNA/G duet boxes and quartet boxes show very low GC3 percentages (2.5 ± 1.0 and 5.4 ± 4.5%, respectively) strongly deviating from mutational bias predictions, while NNU/C duet boxes show a strong enrichment in NNC codons with the exception of cysteine (58 ± 31.7 and 66.8 ± 23.5% with and without including cysteine, respectively). The high deviation from the genomic GC content correlation for highly expressed genes indicates that there is another mechanism involved in codon selection than solely evolutionary mutational bias.

tRNA concentrations influence codon usage

In this section, we will explore the codon families that follow Ikemura's first rule: when two different tRNA isoacceptors are present for the same amino acid, the codon with the most abundant tRNA is used more frequently (Ikemura, 1981; Ikemura, 1982). In the chloroplast, only a few amino acids possess several tRNAs (Leu, Arg, Ser, Ile, and Gly). The isoleucine three-codon box possesses a tRNA-G_{34}AU reading the two codons AUU and AUC; in addition, a second isoacceptor, tRNA-k_2C_{34}AU, reads its third codon (AUA). As mentioned in Section 1, codon usage bias of the AUU/C duet box does not exactly follow the typical behavior of NNU/C duet boxes. Indeed, while there is an enrichment in AUC codons in high-expression genes, the main codon for isoleucine remains AUU. Codon regulation seems to occur rather on the AUA codon, which is almost absent from the high-expression group but common in the

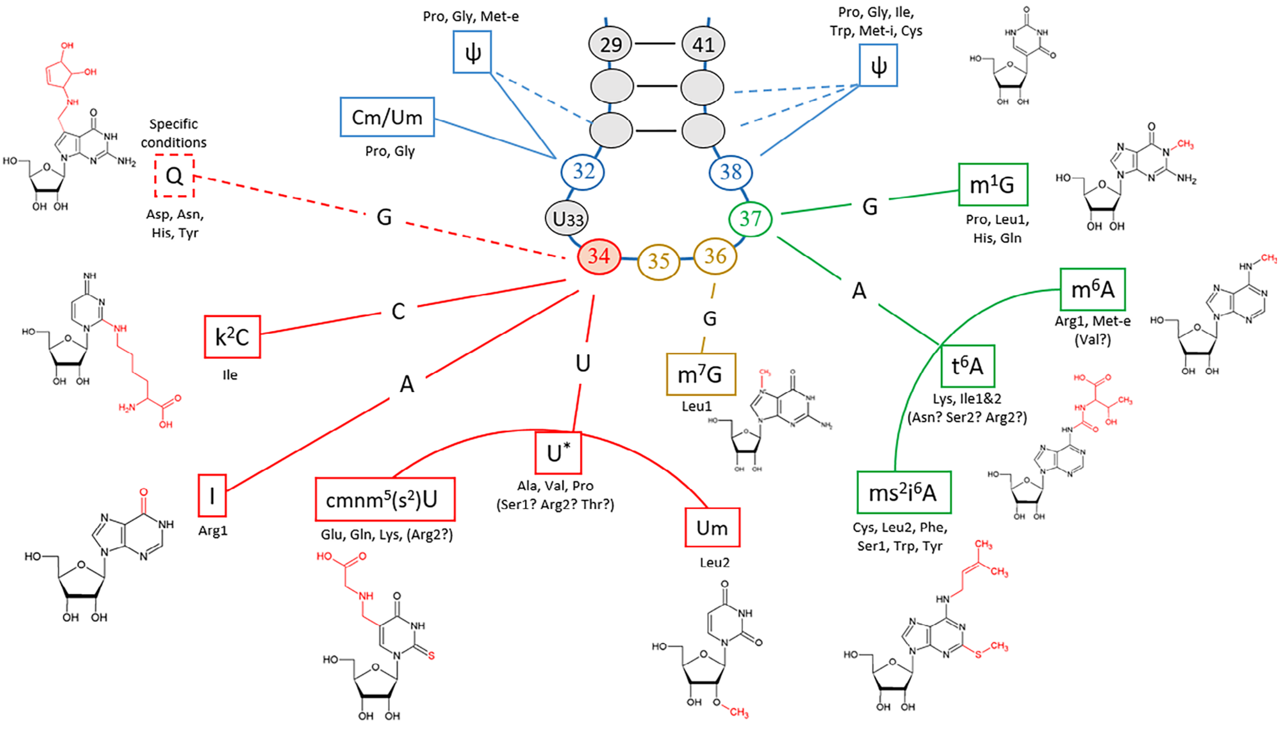

Figure 7. Principal modifications of the tRNA anticodon loop. The principal modifications of the tRNA anticodon loop are shown in connection to their position (Figure from Fages-Lartaud & Hohmann-Marriott, 2022, submitted). Each nucleotide position of the anticodon loop is associated with corresponding modifications depending on the type of original nucleotide. The modifications at position 37 (in green) maintain the decoding accuracy by avoiding interferences between codon boxes. The nature of anticodons, especially with the modifications of the wobble base N_{34} (in red), determines the codon–anticodon pairing affinity and defines codon optimality.

low-expression group (Figures 3d and 4). We hypothesize that tRNA-k_2C_{34}AU is rare in comparison to tRNA-G_{34}AU, thus making AUA the key regulatory codon for isoleucine. Through this mechanism, the effects of mutational bias, favoring A/T in the third codon position, do not have to be countered, while tRNA concentrations regulate isoleucine codon usage. For the four-codon box of glycine, the combination of tRNA-G_{34}CC and tRNA-U_{34}CC reading properties favors the U-ending codon but tRNA concentrations might not be a dominant factor of regulation. Moreover, tRNA-G_{34}CC has been shown to be dispensable in the chloroplast (Rogalski et al., 2008).

For each six-codon box (Ser, Arg, and Leu), one tRNA reads a duet box and the other one reads a quartet box. For serine, the AGU codon from the duet box is less used in high- than in low-expression genes, but it is compensated by the favored codons of the quartet box (UCA/U) rather than with the synonymous AGC codon of the duet box. Here, the usage of the duet box is lower but reasonable in high-expression genes, while the duet box is used significantly in the low-expression group. This suggests that the tRNA-U_{34}GA concentration is higher than that of the other isoacceptor. Additionally, the codon choice strategy for serine also lowers the GC3 content, in accordance with mutational bias pressure. In the case of arginine, the duet box AGA/G is the key of its codon regulation. It is rarely used in the high-expression group but slightly more common in the low-expression group. Thus, tRNA-I_{34}CG is preponderant compared to tRNA-U_{34}CU, and makes CGU the preferred codon. For the leucine six-codon box, across the duet and quartet box, A/U-ending codons are almost exclusively present in highly expressed genes, while C/G-ending codons are more frequently found in low-expression genes. The UUA codon is used most frequently, suggesting that the tRNA-Um_{34}AA concentration is higher than the tRNA-U_{34}AG concentration. However, it seems like C/G-ending codons play a yet to be defined role in codon regulation for leucine.

The first rule of Ikemura helps to explain tendencies of codon usage for some codons but only for a few amino acids. Therefore, we will investigate in the following section how Ikemura's second rule affects the codon usage of the chloroplast.

tRNAs pairing properties are responsible for codon optimality within each codon family

Every organism possesses specific characteristics, such as a mutational bias signature, a certain tRNA set, and particular tRNA expression levels. Depending on these

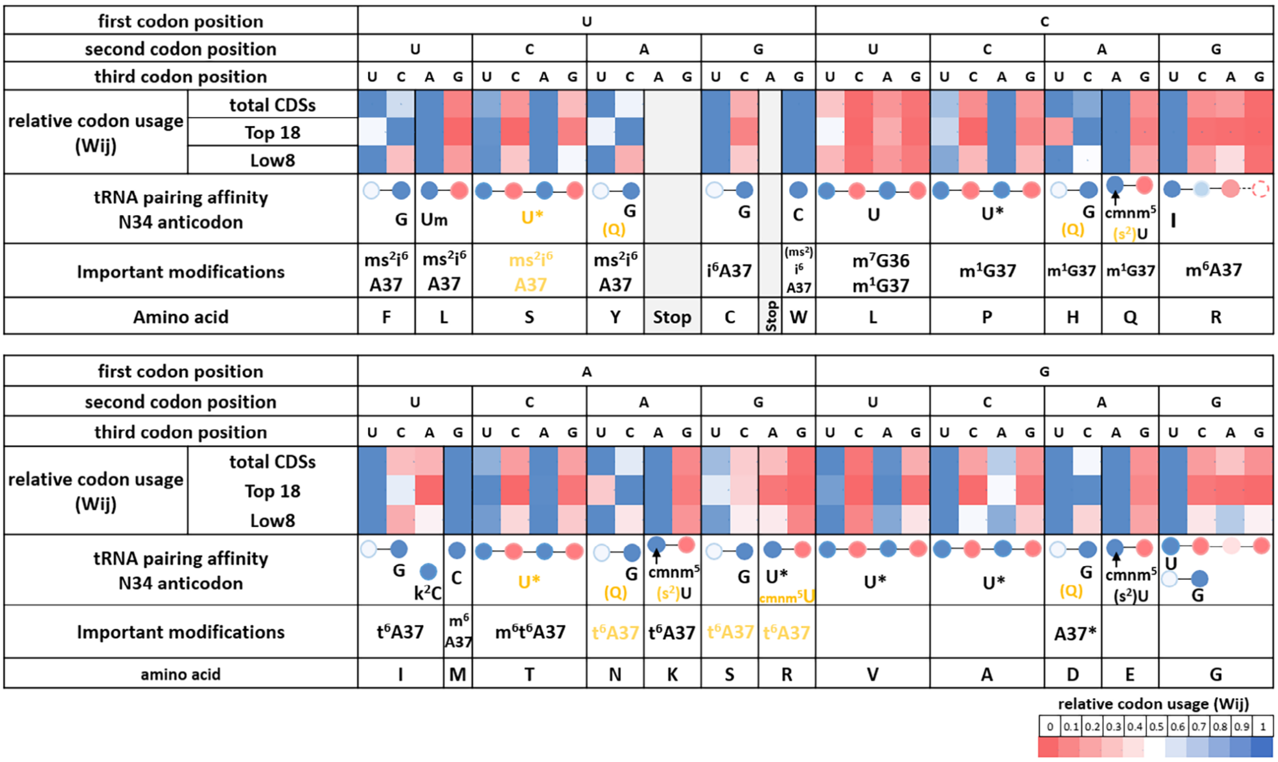

Figure 8. Correlation between tRNA properties and codon usage bias. Codon usage for total CDSs, and the high- and low-expression groups (Top 18 and Low 8, respectively). Codon usage is displayed with tRNA modifications and their pairing affinities to show the effect of codon–anticodon pairing on codon usage bias. The relative codon usage (Wij) is represented for each codon box associated with an amino acid. The relative usage of the different codons encoding the same amino acid is indicated. More precisely, the subtlety of codon usage bias in highly expressed genes compared to the total CDSs or the low-expression group (Wij is red for low toward blue for high) is displayed. The last row contains the tRNA modifications of anticodon position 34 and other important modifications. The codon–anticodon pairing efficiency is represented with the same color code as Wij (from Fages-Lartaud & Hohmann-Marriott, 2022). Base 34 of each anticodon is represented directly below the codon it recognizes by Watson–Crick pairing (whenever it is possible). Shown in black are modifications determined experimentally and shown in yellow are modifications postulated from bioinformatics analysis. This figure shows the correspondence between codon usage of highly expressed genes and codon–anticodon pairing efficiency. For amino acids with multiple tRNAs, it shows which isoacceptor is preponderant. Codon optimality is represented by the high-expression group and correlates with codon–anticodon pairing affinity or tRNA concentration.

characteristics, evolution drove organisms to adopt different decoding strategies. These strategies have been categorized into four groups by Grosjean et al. (Grosjean et al., 2010). The chloroplast falls into the third sparing strategy, which consists in a total depletion of tRNA harboring A_{34} and C_{34} in the anticodon. Additionally, the ability of U_{34} to read all 4-fold degenerate codons by superwobbling permits decryption of the genetic code with only 25 tRNAs (Alkatib et al. 2012; Rogalski et al., 2008).

Ikemura's second rule states that when only one tRNA reads several codons by wobbling, a higher affinity between codon and anticodon leads to a higher usage (Ikemura, 1981, Ikemura, 1982). Optimization of the binding energy between codons and anticodons is a selection criterion that drives evolution. During translation, matching Watson–Crick base pairing provides an advantage over wobble base pairing by decreasing codon deciphering rates (Grosjean & Westhof, 2016; Letzring et al., 2010, Stadler & Fire, 2011). Importantly, Watson–Crick geometry and tRNA modifications are fundamental structural features that influence the acceptance of the appropriate tRNA species by the ribosome during translation, affecting its kinetics beyond purely codon–anticodon recognition (Cochella, 2005; Gromadski et al., 2006; Ogle et al., 2001, 2002). Pairing affinity is highly dependent on the identity of the anticodon, the respective nucleotide modifications, and tRNA secondary structures. In order to verify the applicability of Ikemura's second rule to codon selection in the chloroplast, we collated bioinformatics analysis with available experimental data to build a comprehensive picture of the current knowledge concerning tRNA modifications in the chloroplast (Figure 7) (Fages-Lartaud & Hohmann-Marriott, 2022). The results of this study are used to infer codon–anticodon affinity and translation efficiency.

The chloroplastic tRNA set resembles the one of *M. capricolum* analyzed by Grosjean et al. in terms of tRNA modifications, the ability of U_{34} to read an entire quartet box, and a relatively similar GC content (Grosjean &

Westhof, 2016). Substituents such as Um, cmnm^5, and cmnm^{5}s^2 on U_{34} restrict anticodon pairing to NNA/G boxes with a strong preference for A-ending codons (Fages-Lartaud & Hohmann-Marriott, 2022; Grosjean et al., 2010; Kurata et al., 2008; Lim, 1994; Takai & Yokoyama, 2003). Such a large difference in codon–anticodon affinity explains the omnipresence of NNA codons for duet boxes regardless of the expression groups (Figure 8). Additionally, favoring NNA codons is in accordance with the mutational bias; hence, there are no opposing evolutionary forces.

For the quartet boxes, the unmodified U_{34} nucleotide efficiently pairs with U- and A-ending codons, while a G_3:U_{34} wobble base is unstable and C-ending codons are not read efficiently (Grosjean & Westhof, 2016). Anticodons containing an unknown modification of U_{34} were hypothesized to lead to higher translation efficiency of A- and U-ending codons within quartet boxes in the chloroplast of *C. reinhardtii* and in *M. capricolum*, in accordance with Ikemura's second rule (Fages-Lartaud & Hohmann-Marriott, 2022; Grosjean & Westhof, 2016). These codon–anticodon affinities, plus the mutational bias, explain the strong codon usage bias toward NNU/A codons in quartet boxes across all gene expression groups (Figure 8). Once again, NNG/C codons are particularly repressed in high-expression genes but not completely absent, which may be due to mutational bias or an underlying hidden functionality. The only exceptions to these rules are arginine and glycine quartet boxes. Indeed, the inosine of $tRNA_{Arg}$-I_{34}CG pairs preferentially with U, slightly less with C, barely with A, and almost never with G. Therefore, arginine almost exclusively uses the CGU codon (Figure 8). For glycine, it was shown that its tRNA often favors exclusively the GGU codons for reasons that are not completely understood. A possible reason for the preferred use of GGU may originate from the stacking properties of C_{32} in $tRNA_{Gly}$-U_{34}CC (Claesson et al., 1995; Lustig et al., 1993) or the combination with the second $tRNA_{Gly}$ also reading GGU.

Finally, the NNU/C duet boxes show the largest difference in usage between low- and high-expression groups (Figures 4 and 8). In these duet boxes, each tRNA contains a G_{34} in the anticodon that shows a pairing efficiency for NNC codons that is about three times higher than that to NNU codons (Chan et al., 2017; Grosjean & Westhof, 2016). Interestingly, highly expressed genes tend to overuse NNC codons despite the directional mutational bias to increase translation efficiency, while low-expression genes succumb to mutational bias and rather use NNU codons (Figure 8). The concordance of tRNA pairing energy and codon usage bias for the NNU/C duet boxes is hidden when looking at total CDSs, but is revealed when the focus is on highly expressed genes.

In conclusion, the tRNA set of the chloroplast is fairly well adapted to the nucleotide composition equilibrium caused by the mutational bias. The effect of tRNA concentrations affects the codon usage of only a few amino acids and correlates with the mutational bias, while codon–anticodon pairing affinity appears as the dominant factor dictating codon usage and consequent bias in the chloroplast. The interactions between the anticodon loop and the mRNA strand are not exactly restricted to the anticodon itself, especially concerning interactions with ribosomes; therefore, a codon context dependency emerges that may influence codon choice. We will now explore if such a contextual effect is present in the chloroplast.

The influence of codon context on codon choice

Protein function and features dictate the amino acid composition of protein primary sequences; hence, the use of the appropriate codon family at each position of coding mRNA primary sequences is imposed. As presented above, the choice among redundant codons for an amino acid is determined based on codon usage bias. In addition, different types of protein features (such as α-helices) present diverse composition enrichments in certain categories of amino acids (e.g., aliphatic amino acids) due to their functionality and substitutability (Grantham, 1974; Yampolsky & Stoltzfus, 2005). The non-randomness in the juxtaposition of amino acids constitutes the dipeptide bias (Ghadimi et al., 2018), i.e., two amino acids are associated with a different frequency than their coupled respective frequency of occurrence.

Besides codon usage and dipeptide biases, another mechanism related to codon context was hypothesized to influence codon choice in coding mRNAs. Across all three kingdoms of life, the existence of a species-specific codon pair bias was evidenced (Buchan et al., 2006; Moura et al., 2007, 2011; Tats et al., 2008). There are fundamental differences in the mechanisms for this codon bias between eukaryotes and prokaryotes. In eukaryotes, codon pair bias was shown to be a direct consequence of dinucleotide bias and DNA methylation at the junction of two codons (Kunec & Osterrieder, 2016; Moura et al., 2007). In prokaryotes, however, codon pair bias seems to arise from constraints imposed by the translational machinery (Boycheva et al., 2003; Moura et al., 2007, 2011). While there are no common preferred codon pairs across organisms (Moura et al., 2007), each bacteria seems to specifically overuse certain codon pairs in relation to their respective translation apparatus (Buchan et al., 2006; Moura et al., 2007; Moura et al., 2011).

In order to identify a potential influence of codon pair bias in the chloroplast, we sought to compare the actual occurrence of codon pairs in CDSs with their expected

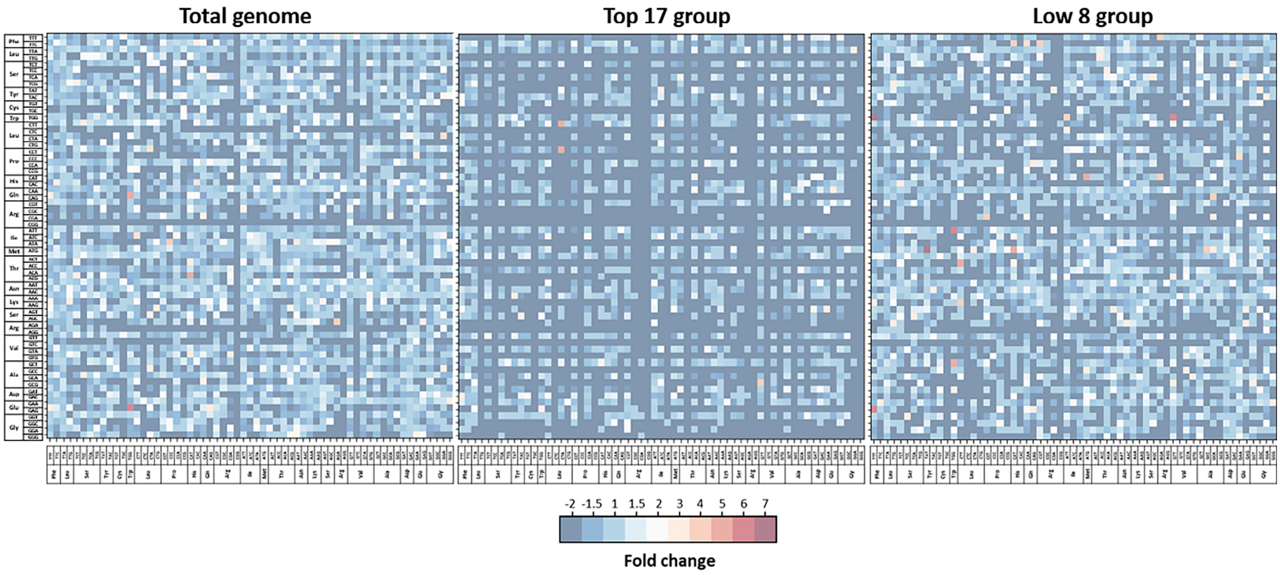

Figure 9. Codon pair bias of different expression groups. The colored matrix represents codon pair bias within each expression group. Blue indicates underrepresented codon pairs, white indicates the absence of significant bias, and red indicates significantly overrepresented codon pairs (>3-fold). Each codon with their encoded amino acids is represented on the *y*-axis (codon 1: ribosomal P-site) and *x*-axis (codon 2: ribosomal A-site). Since rare events were taken into consideration for the calculations, there are no artificially overexpressed codon pairs. Thus, the number of significantly overrepresented codon pairs is relatively low, and their identity often differs between groups. These matrices may indicate the presence of codon context dependencies influencing codon choice. Data are available in Data S4.

occurrence based on codon usage of individual codons. We first removed the effect of dipeptide bias to match over/underrepresented dipeptide associations (Buchan et al., 2006; Ghadimi et al., 2018) (see Methods and Data S4). After this step, we compared the number of codon pairs observed in CDSs with expectations calculated from a random codon distribution for the 3721 possible combinations (61 × 61 codons) (see Methods and Data S4). Caution must be applied with the calculation of the codon pair bias because the method relies on a probabilistic model using codon counts and codon frequencies. Therefore, this method is not suitable for the prediction of rare events, as is the case for pairs involving rare codons. As an illustration, the rare leucine codon CTC, which is present only 15 times in the genome, will show by default a low probability of pairing with any other codons (below one occurrence or very low frequency). This low expected occurrence/frequency will artificially inflate the codon pair bias as soon as a pair is present because of the division by an artificially low number. To avoid this issue, some studies excluded rare codons from the analysis (Buchan et al., 2006; Gutman & Hatfield, 1989); others did not take it into consideration, leading to artificial inflation of codon pair bias by including rare codons. Since rare codons constitute a significant fraction of sense codons in the chloroplast, we aimed to include them in the analysis. To circumvent distortions by low-frequency codons, while preserving rare codons in the analysis, we adopted a mixed approach. This approach consists of a conventional probabilistic model for common codons and a deterministic model for rare codons (see Methods). In brief, a Poisson probability law was used to assess the potential of each codon to make on average one pair or less with a 75% confidence threshold. The codons falling into that category were considered to be rare and therefore followed a deterministic model. For simplification, both models follow the same calculations for the number of expected pairs, but the value is rounded up to the nearest whole number to avoid artificial inflation of the codon pair bias. We present the differences between our analysis based on these two models and an analysis that does not consider the issue of artificial inflation in the calculations (Figure S1).

The analysis of codon pair bias indicates that there are far fewer overrepresented codon pairs than suggested by previous studies. Although there are still some overrepresented codon pairs in the chloroplast, it is difficult to draw conclusions from the candidates showing deviations from expectations. It is important to note that the relatively low total number of codons in the chloroplast may hide some existing traits of codon pair bias.

The results presented in Figure 9 show under/overused codon pairs in the genome and across expression groups. Since underused codon pairs are very common due to the deterministic model and the low usage of certain codons, we focused solely on overused pairs. Indeed, as mentioned above, bacteria tend to overuse certain codon pairs (Buchan et al., 2006); we have extracted several codon pairs from chloroplastic CDSs overrepresented more than 3-fold compared to expectations (Data S4).

A first observation is that most overrepresented codon pairs contain at least one translationally unfavorable codon, as exemplified by the following pairs: Arg-Arg (AGA-AGA), Ile-Pro (ATA-CCC), Ala-Val (GCG-GTT), Glu-Gln (GAG-CAG), and Leu-Thr (TTG-ACC). Codon pair bias can result from the juxtaposition of unfavorable codons acting in concert to slow down translation (Irwin et al., 1995). Indeed, the occurrence of a codon pair bias was correlated with translation efficiency and accuracy (Gamble et al., 2016; Kurland, 1987; Shpaer, 1986) and with the presence of protein tertiary structures (Widmann et al., 2008).

Although mechanisms involved have not been fully identified, the tRNA wobble base interaction with the third nucleotide of the P-site codon (cP_3) and at least the first, up to all three nucleotides of the A-site codon ($cA_{1,2,3}$), have been correlated with codon pair bias (Bossi & Roth, 1980; Buchan et al., 2006). In the chloroplast, codon context preferences were observed with the first nucleotide of the adjacent codon (cA_1) (Morton, 2003). In addition, it was hypothesized that tRNA–tRNA interactions in the ribosomal A- and P-sites may influence translation efficiency, thus shaping codon pair usage (Buchan et al., 2006; Smith & Yarus, 1989). In the case of the chloroplast, there seems to be no relationship between the types of aminoacyl-tRNA either in the P- or A-site and overused codon pairs. Since most codon families are decoded by only one tRNA species, it is unlikely that translationally detrimental tRNA–tRNA interactions were sustained throughout evolution. Although it could be the case for amino acids decoded by several tRNAs (Leu, Ser, Arg, Gly, and Ile), we found no evidence of preferences between their two respective tRNA isoacceptors for any codon family either in the A- or P-site. The evolutionary tendency toward choosing a preferred tRNA isoacceptor for these amino acids may have started from the lack of detrimental tRNA–tRNA interactions before concentrations were optimized. In the chloroplast, it is difficult to draw conclusions from this analysis, due to the very low occurrence of certain codons, which drowns out a potential effect of tRNA–tRNA interaction.

Across all kingdoms of life, evolution suppressed NNU_3-A_1NN and NNU_3-G_1NN motifs in codon pairs to avoid potential out-of-frame UAA and UGA stop codons (cP_3-$cA_{1,2}$) (Moura et al., 2007; Tats et al., 2008). In addition to stop codons, codon pair bias tends to avoid mononucleotide repeats that can lead to frameshifts and loss of protein function (Berg & Silva, 1997; Gu et al., 2010; Tats et al., 2008). In the chloroplast, we found no particular repression involving overlapping stop codons either in cP_3-$cA_{1,2}$ or $cP_{2,3}$-cA_1. Similarly, we found no active avoidance of tetra-, penta-, or hexanucleotide repeats (Data S4).

Dinucleotide bias at the junction of two adjacent codons is responsible for codon pair bias in eukaryotes, but it is not evident in bacteria (Kunec & Osterrieder, 2016; Moura et al., 2007). We analyzed the proportion of the 16 potential overlapping dinucleotides in all overrepresented pairs. We found a slightly higher proportion of dinucleotides corresponding to G or C in the third position of the P-site codon (cP_3); AT and GC contents of the cP_3 position are 41 and 59%, respectively (Data S4). However, the higher presence of GC in cP_3 is simply a consequence of the overuse of translationally unfavorable codons, usually containing G or C in the third codon position of chloroplastic CDSs.

A correlation between overuse of certain codon pairs and gene expression was initially hypothesized (Gutman & Hatfield, 1989; Yarus & Folley, 1985), but later questioned due to the low number of genes in the initial analysis. Correlation of codon pairing with gene expression was demonstrated to be only moderate (Boycheva et al., 2003; Buchan et al., 2006). We found that there are only a few overrepresented pairs in the Top expression group and a larger proportion in the Low expression group. Although some of these pairs are present across groups, such as the Ala-Val pair GCG-GTT, most overrepresented pairs differ among the three expression groups (Figure 9 and Data S4). Thus, there is no correlation between gene expression and the identity of particular codon pairs. However, there is an increase in the total number of overrepresented pairs with relatively low expression.

To conclude, codon pairing does not seem to be a major mechanism for shaping codon usage in mRNA CDSs that can be separated from codon usage bias. The effects of codon context might not be preponderant in the chloroplast due to its minimalistic nature or might be difficult to detect due to the very low occurrence of rare codons.

Functional implications of codon usage bias in the chloroplast context

In this final section, we show the impact of codon usage bias on cellular processes leading to protein expression. In addition to optimizing protein translation rates, codon usage bias and amino acid usage may also provide relative pauses for protein folding and affect mRNA secondary structures involved in ribosome recruitment. Furthermore, it has been hypothesized that clusters of rare codons decrease early translation rates due to space limitation of ribosomes along the mRNA strand.

Codon usage affects translation rates and protein yield

The main purpose of codon usage bias is to optimize translation yield and accuracy and conserve protein structure and functionality, while managing resources to optimize cell fitness. The average translation rate in prokaryotes is about 20 amino acids per second (Gouy &

Grantham, 1980) and can vary by more than one order of magnitude (Chevance et al., 2014; Sørensen et al., 1989; Sørensen & Pedersen, 1991). Ribosome profiling studies aim to identify factors influencing translation speed; however, limits in sensitivity allow to focus mainly on the most extreme translational pauses, for example those caused by internal SD sequences (Li et al., 2012). New analysis designed to quantify comparatively subtle differences in translation rate demonstrated that rare codons are translated more slowly than synonymous optimal codons (Gardin et al., 2014; Stadler & Fire, 2011). Indeed, all the factors contributing to codon usage bias were demonstrated to affect the translation rate. The additive effect of decreased translation speed originating from rare codons constitutes a key regulatory factor affecting low-expression genes. In contrast, translationally optimal codon usage increases protein production of highly expressed genes (as shown in Figure 2). Both factors responsible for codon usage bias, tRNA concentrations and codon–anticodon pairing affinity, influence translation rates and protein yield.

The biological implications of tRNA concentrations—A predominant factor that determines the efficiency of translation is the availability, turnover, and relative concentrations of tRNA isoacceptor species (Gouy & Grantham, 1980; Pedersen, 1984; Varenne et al., 1984). The entry events of rare tRNA isoacceptors into the ribosome A-site are less frequent, making the ribosome stall while waiting for the correct tRNA isoacceptor, thus slowing down translation. In the chloroplast, only a handful of amino acids are decoded by several tRNAs (Leu, Arg, Ser, Ile, and Gly). Since highly expressed genes are biased toward an increased translation yield, we correlated codon usage for these amino acids with their respective tRNA isoacceptor concentrations in Section 2.2 (Figure 3 and Figure 4). Briefly, leucine uses preferentially the UUA/G duet box for optimized translation; arginine and serine use rather their quartet boxes CGN and UCN, respectively; isoleucine uses the rare AUA codon to regulate translation of low-expression genes, while AUU is the translationally favored codon; glycine might use its two tRNA isoacceptors to favor the GGU codon, although here, tRNA concentrations are not the most significant parameter. The tRNA concentration effects can be assessed by examining the correlation between codon usage bias and tRNA properties presented in Figure 8. The reliance on codon–anticodon pairing energy inside a codon box is decreased because the second box offers alternative codons. For example, the serine codon AGC of the duet box would be expected to be used preferentially in highly expressed genes; instead, the optimal UCU/A codons from the quartet box are used (Figure 8). Therefore, codon usage bias of the serine duet box is not correlated with the pairing energy of $tRNA_{Ser}$-G_{34}CU. Interestingly, for these amino acids, tRNA concentration-mediated codon regulation concords with a directional mutational bias toward AT_3, so the pressure to counterbalance mutational bias toward C-ending codons is less severe (as presented in Section 2.1). It is interesting to note that regulation of aminoacyl-tRNA levels, charging, and modifications can cause changes in the set of genes preferentially expressed, for example due to the cell cycle, growth conditions, or circadian rhythms (Jayabaskaran et al., 1990; Matsuo et al., 2006; Nedialkova & Leidel, 2015).

The effects of codon–anticodon pairing affinity on translation rates—Another dominant factor affecting translation is codon–anticodon pairing affinity (Grosjean & Westhof, 2016; Letzring et al., 2010). Indeed, perfectly matching Watson–Crick base pairing between codon and anticodon decreases the time necessary to recruit and utilize the aminoacyl-tRNA in comparison to wobble base pairing at the third codon nucleotide (Grosjean & Westhof, 2016; Letzring et al., 2010; Stadler & Fire, 2011). Therefore, translation efficiency increases with codon–anticodon pairing stability and interaction with ribosome geometry (Grosjean & Westhof, 2016; Letzring et al., 2010).

We previously presented the codon–anticodon affinities for the three groups of codons in Section 2.3. Briefly, the NNA/G duet boxes strongly favor A-ending codons (Grosjean et al., 2010; Kurata et al., 2008; Lim, 1994; Takai & Yokoyama, 2003), which benefits translation rates, while overuse of NNG codons could be detrimental to protein expression yield. In quartet boxes, the modified or unmodified U_{34} of anticodons generally shows a strong preference for U- and A-ending codons (Grosjean & Westhof, 2016) (for exceptions see Section 2.3), which should be reflected in high translation rates for these codons. The NNU/C duet boxes present an affinity for NNC codons around three times higher than for NNU codons. Here, both codons can be distributed along a CDS to modulate translation yield without severe consequences.

In order to verify the effects of codon–anticodon pairing affinity on translation rates, we examined previous data that analyzed ribosomal proteogenic site occupancy by aminoacyl-tRNAs (Gawroński et al., 2018). Interaction of tRNA with mRNA in ribosomes can be evaluated by ribosome profiling. This method analyzes ribosomal-protected mRNA fragments to establish enzyme densities along coding mRNAs. From these data, pausing sites can be extracted based on the assumption that an increase in ribosomal density reflects a decrease in translation speed. In a previous study aiming to identify the major translational pauses in the chloroplast of *Arabidopsis thaliana*, Gawroński et al. did not find a correlation between codon usage and ribosomal pausing (Gawroński et al., 2018). However, this analysis was not flawless, because it used the genomic codon usage, which often causes erroneous conclusions regarding codon usage bias. We performed a new analysis of their data, taking into consideration that

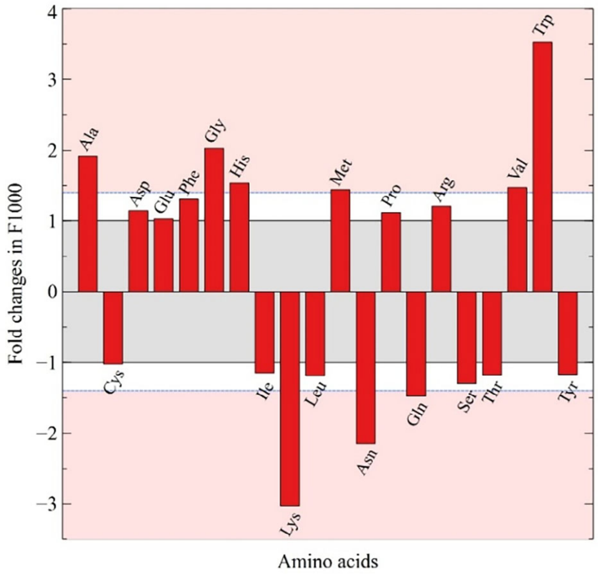

Figure 10. Fold changes in amino acid composition. The changes in amino acid composition of proteins (frequency per thousand [F/1000]) between the Top 18 and Low 8 expression groups are shown. Red areas represent fold change values considered significant (>1.4-fold).

translationally optimal codons are reflected in high-expression genes due to the arguments presented in previous sections. While the sensitivity of the method may not provide the required accuracy for precise quantification of codon translation speed, it is possible to extract tendencies. We analyzed the relative pausing score of allegedly favored over detrimental codons for all codon boxes when they are present in the P-site only, EP-sites, and EPA-sites. Overall, apparent translation speed correlates with tRNA pairing affinity and codon usage as presented in Figure 8 with the exception of threonine, glutamate, and aspartate. For NNA/G duet boxes, the G-ending codons present an average pausing score that is 49% higher than their A-ending counterparts (69% excluding glutamate). For NNC/U duet boxes, the G_{34}:U_3 wobble base pauses 68% longer than the G_{34}:C_3 Watson–Crick pair (80% excluding aspartate; 38 and 48% when considering the EP-sites, respectively). Quartet boxes show an on average 35% increased pausing score for unfavorable codons versus their optimal counterparts (47% excluding threonine, 38 and 50% in the EP-sites, respectively) (Data S5). These results are in accordance with the finding that codons deciphered through Watson–Crick codon–anticodon pairing are translated faster than their synonymous wobble-pairing codons (Wang et al., 2017). In addition, it was shown that optimal and frequent codons were decoded more quickly than rare codons and that AT-rich codons were translated faster than GC-rich codons (Gardin et al., 2014). We can also distinguish a longer decoding time for tRNA with hypothesized lower concentrations than the second isoacceptor (for Ile, Leu, and Arg), but not for serine since similar tRNA concentrations were postulated. These data point toward our model of codon optimality in chloroplasts. Nonetheless, it is important to note that *C. reinhardtii* and *A. thaliana* are relatively distant species and may present some evolutionary differences regarding this topic. Although these relative changes in translation kinetics are subtle, their additive effects constitute a major factor that affects translation yield. We mapped all codons considered as translationally 'slower' along all CDSs (Figure S2), and it is clear that there is a gradient from rarely present toward abundant with decreasing gene expression.

The amino acid composition influences translation rates

Similar to the availability of tRNA, regulation could also be exerted through a differential usage of amino acids in proteins from distinct expression groups; however, analysis of the amino acid bias did not show the presence of such a mechanism in the chloroplast. As described before, there is a directional nucleotide mutational bias resulting from an equilibrium in transversion kinetics (Sueoka, 1962; Sueoka, 1988). However, the effect of directional mutational bias is not restricted to neutral regions and strongly influences the amino acid composition of proteins (Gu et al., 1998; Knight et al., 2001; Lobry, 1997; Singer & Hickey, 2000; Sueoka, 1961). Evidently, the determining factor of amino acid composition results from selection based on their functionality (Lobry & Gautier, 1994). For example, membrane proteins are enriched in hydrophobic amino acids while cytoplasmic proteins contain more hydrophilic amino acids. The amino acid bias correlates with GC content and is exerted in near-neutral sites of proteins (Osawa et al., 1990, 1992) or through similar amino acid characteristics (Grantham, 1974; Yampolsky & Stoltzfus, 2005). Hence, higher GC content correlates with an increase in codons containing one or two G:C bases while higher AT content favors codons with two or three A:T bases. The difference in GC content between high- and low-expression genes (39.0 and 30.9%, respectively) is responsible for their bias in amino acid composition. The high-expression set is clearly enriched in Ala, Gly, Trp, His, Val, and Met (40% increase) and to a lower extent in Pro, Arg, Asp, Glu, and Phe (Figure 10). This is in contrast to the low-expression set, which is enriched mainly in Lys, Asn, and Gln (40% decrease) and to a lower extent in Tyr, Ile, Leu, Thr, and Ser (Figure 10). Last, it is interesting to note that the positively charged amino acid lysine is used 3-fold more in low-expression genes. The peptide exit tunnel of the ribosome is composed of negatively charged residues and it has been hypothesized that interaction with peptides enriched in positively charged amino acids slows down translation (Charneski & Hurst, 2013; Lu & Deutsch, 2008).

The impact of codon usage on protein folding

In addition to encoding the primary sequence of a protein, mRNA possesses additional functions. While optimal codon usage and amino acid composition improve protein yield, an overabundance of optimal codons increases the amount of insoluble or misfolded proteins, illustrating the necessity for rare codons (Angov, 2011; Crombie et al., 1992; Jacobson & Clark, 2016; Komar et al., 1999; Rosano & Ceccarelli, 2009; Siller et al., 2010; Spencer et al., 2012). Indeed, fine-tuned translation rates ensure protein quality and cotranslational folding efficiency by inducing ribosomal pausing around protein domain boundaries, leading to the proper folding pathway (Buhr et al., 2016; Kaiser et al., 2011; Purvis et al., 1987; Thanaraj & Argos, 1996; Tsai et al., 2008). In bacteria, a local decrease in translation speed affects cotranslational protein folding, cofactor binding, multi-domain assembly, chemical modifications, and membrane targeting (Gloge et al., 2014). In the chloroplast, in addition to accurate protein folding, a multitude of cofactors, such as pigments, quinones, hemes, and metal ions, are cotranslationally associated with nascent proteins (Nickelsen & Rengstl, 2013; Schöttler et al., 2011, 2015). Specific ribosomal pausing sites along coding mRNA, for *psbA*, *psaA*, *psaB*, and *psaC*, were suggested to facilitate the integration of stabilizing cofactors (e.g., chlorophyll) and promote the accurate folding and biogenesis of photosystem I and II multi-protein complexes (Gawroński et al., 2018; Kim et al., 1991, 1994; van Wijk et al., 1996). Another important factor for the chloroplast is the targeting of nascent proteins to the thylakoid membrane, which occurs cotranslationally for a significant fraction of plastidial proteins (Celedon & Cline, 2013; Friemann & Hachtel, 1988; Margulies et al., 1987; Zoschke & Barkan, 2015).

The various mechanisms responsible for codon usage that affect translation kinetics were correlated with enhanced cotranslational protein folding (Komar, 2009; Zhang & Ignatova, 2011). Altering these parameters through synonymous codon substitutions lead to slight differences but substantial consequences on protein structure, function, and cell fitness (Buhr et al., 2016; Tsai et al., 2008; Walsh et al., 2020; Zhang et al., 2009). The ribosomal exit tunnel can accommodate 30–40 amino acids of the nascent peptide chain and allows the formation of secondary structures such as α-helices (Bhushan et al., 2010; Gloge et al., 2014; Holtkamp et al., 2015). However, tertiary structures form when polypeptides exit the tunnel (Holtkamp et al., 2015; Lin et al., 2012; Lu & Deutsch, 2005); thus, there is a polypeptide spacer between the pause-causing event and the formation of the tertiary structure. The pausing events can be caused at the P-site (e.g., related to codons/tRNAs), upstream (e.g., SD sequence or positively charged amino acids), or downstream from the P-site (e.g., mRNA structure).

One of the major mechanisms causing strong ribosomal pauses is mRNA secondary structures (Chen et al., 2013; Qu et al., 2011). The duration of ribosomal pauses depends on the strength of these structures and reflects the ability of the ribosome to remove these obstacles (Chursov et al., 2013; Qu et al., 2011). In the chloroplast, it was shown that pauses due to mRNA secondary structures are the main factor influencing protein domain folding (Gawroński et al., 2018). Codon choice can indirectly affect protein expression by influencing mRNA secondary structures, hindering or improving transit of ribosomes during translation (Chursov et al., 2013; Qu et al., 2011). Synonymous codon substitutions can disrupt mRNA secondary structures; therefore, they can affect the proper folding of protein domains. This may explain why a few rare codons remain in specific positions of high-expression genes that can be conserved across species (Chursov et al., 2013; Pechmann & Frydman, 2013).

The second mechanism proposed to explain ribosomal pauses is the presence of an internal SD sequence that interacts with the anti-SD sequence of the ribosomal 16S RNA (Li et al., 2012). However, some of these claims have been disputed and assigned to technical artifacts (Mohammad et al., 2016). In chloroplast of *C. reinhardtii*, the anti-SD sequence is 3′-UCCUCC-5′, corresponding to a 5′-AGGAGG-3′ SD sequence. In plastidial CDSs, exactly matching penta- and hexanucleotide SD sequences rarely occur (a total of nine times in the Low 8 group). While tetranucleotides matching the SD sequence, such as AGGA, GGAG, or GAGG, are more common, they are rare in the high-expression group. To consider imperfect matches, Gawroński et al. calculated the hybridization energy of an 8-nucleotide window with the anti-SD sequence of the chloroplast. Their results indicate that the strongest anti-SD–mRNA interactions upstream of the pause sites correlate with the duration of the pausing event (Gawroński et al., 2018). Interactions with the anti-SD sequence require a certain nucleotide composition, especially rich in guanosine. The corresponding codons potentially interact with the anti-SD sequence and often represent translationally unfavorable codons, such as glycine GGA, glutamate GAG, and arginine AGG. As for mRNA secondary structures, these codons might be sustained in specific positions of coding mRNA in order to maintain anti-SD-related ribosomal pausing.

In addition to these two mechanisms, the presence of positively charged amino acids in the nascent polypeptide chain were shown to interact with the negatively charged ribosomal exit tunnel, thus causing a decrease in translation rate (Charneski & Hurst, 2013; Lu & Deutsch, 2008). In the chloroplast, the positively charged amino acids (lysine, arginine, and histidine) conferred ribosomal pauses, which were, however, weaker than those caused by the two previous mechanisms (Gawroński et al., 2018). We showed in the previous section that low-yield proteins are enriched in lysine by a factor of 3 compared to high-yield proteins. This distribution bears witness to the evolutionary pressure applied on the protein primary sequence that also shapes protein expression (Gu et al., 1998; Knight et al., 2001; Lobry, 1997; Singer & Hickey, 2000; Sueoka, 1961). The distribution of lysine residues along a protein's primary sequence induces small decreases of translational speed that may correlate with minor requirements of protein tertiary structures.

The three mechanisms influencing ribosomal pausing mentioned in the previous paragraphs can be directly affected by synonymous codon substitutions. Since recoding of synonymous codons was demonstrated to disrupt protein functionality and increase misfolding (Rosano & Ceccarelli, 2009; Widmann et al., 2008; Zhang et al., 2009), seemingly unfavorable codons might be sustained in specific positions to ensure protein quality. Across diverse eukaryotic, bacterial, and archaeal genomes, homologous CDSs show conserved rare codon clusters separating small protein structural motifs (Chaney et al., 2017; Clarke IV & Clark, 2008). The positive selection for these position-specific clusters, observed independently of gene expression, suggest a functional role in protein maturation (Chaney et al., 2017; Clarke IV & Clark, 2008). While individual rare codons decrease translation rates, it is not sufficient to induce ribosomal pausing that is sufficient for protein domain folding. These small pauses might be involved in less demanding structural properties (Jacobs & Shakhnovich, 2017). It is unclear if the cumulative effects of decreased translation rates from rare codon clusters can provide pauses that are long enough or if they act indirectly through the mechanisms presented previously. The occurrence of several consecutive rare codons might take advantage of low tRNA concentrations (Fedyunin et al., 2012; Parmley & Huynen, 2009), tRNA turnover and diffusion (Gouy & Grantham, 1980), and tRNA wobble properties to create an aggregate of reduction in translation rate (Zhang et al., 2009). On the contrary, some clusters contain repeats of the same rare codon to optimize translation, because the unusual proximity reduces the time necessary for tRNA turnover and diffusion (Cannarozzi et al., 2010). Nevertheless, these rare codon clusters were demonstrated to play a regulatory role in the folding of important domains (Chartier et al., 2012; Liu, 2020; Widmann et al., 2008; Zhang et al., 2009).

We mapped all the codons considered as translationally unfavorable on the CDSs of the chloroplast (Figure S2). First, we note that there is an increasing proportion of rare codons with decreasing gene expression, pointing toward the relationship between codon usage bias and protein yield. While some genes present a uniform distribution of their rare codons along the CDS, such as *rbcL*, other genes, like *atpA* and *psaB*, present small clusters of at least three unfavorable codons within a 30-codon window. This accumulation is even more visible on intermediate-expression genes such as *rps2* or *rps3*. For low-expression genes, local increases in the density of rare codons are still discernable but are concealed by the relatively high general presence of these codons. Interestingly, some enrichments of rare codon occur in the 5′- or 3′-termini of CDSs. This 5′- and 3′-terminal enrichment has been linked to membrane targeting or protein secretion and to translation termination, respectively (Clarke & Clark, 2010; Gerresheim et al., 2020). In the chloroplast, significant enrichment of the 5′-terminus with rare codons occurs for example in *psaA*, *psaB*, *psbB*, *rps3*, or *ycf1*, and this occurs at the 3′-terminus in *rps4* or *rps7*.

In addition to the aforementioned mechanisms, codon context was demonstrated to affect translation kinetics (Chevance et al., 2014) but only moderately correlated with gene expression (Boycheva et al., 2003; Buchan et al., 2006; Chevance et al., 2014). Thus, the presence of overrepresented, unfavorable codon pairs was hypothesized to locally decrease translation rates in relation with protein maturation (Seligmann & Warthi, 2017). As presented previously, for high-expression genes, there are only a few significantly overrepresented codon pairs and their total occurrence is relatively low. Moreover, since mechanisms responsible for codon pairing are not understood, it is difficult to estimate the contribution of codon context to protein maturation. Most overrepresented codon pairs possess at least one slow codon and are supposedly exploiting the first adjacent nucleotide (cA_1) to intensify their kinetic effects on the ribosome. Thus, codon context may participate in protein maturation to a similar extent as, or in coordination with, rare codon clusters.

The relationship between mRNA structural features and codon choice influences protein expression

Protein expression yield depends on efficient translation initiation. The recruitment of ribosomes at the 5′-untranslated region (UTR) of mRNA sequences is a crucial step for initiating translation. The recruitment process occurs either through interactions between the 5′-UTR SD sequence and the anti-SD sequence of the 16S rRNA or through non-canonical translation initiation mechanisms (Chang et al., 2006; Nakagawa et al., 2017). In the chloroplast, although SD-dependent recruitment certainly plays a role (Scharff et al., 2017), a large portion of genes use alternative mechanisms for ribosome recruitment (Fargo et al., 1998; Nakagawa et al., 2017; Scharff et al., 2011; Weiner et al., 2019). This type of translation initiation requires a low amount of mRNA secondary structure around the start of the gene. Indeed, efficient protein expression is a compromise between mRNA stability and ribosome entry site accessibility (Espah Borujeni et al., 2014, 2017; Espah Borujeni & Salis, 2016; Mignone et al., 2002). Plastidial SD-less mRNAs show an increase in free energy from −30 to +20 nucleotides around the start codon, which is indicative of an absence of secondary structure (Scharff et al., 2011). The decrease in mRNA stability promotes ribosome recruitment and often correlates with a lower local GC content on each side of the start codon (Li & Qu, 2013).

Once ribosomes are recruited, translation initiation rates are influenced by processing of the start codon (Esposito, 2003) and the nucleotide composition at the start of the gene, a region referred to as the gene 'ramp' (T. Tuller, Carmi, et al., 2010a; T. Tuller, Waldman, et al., 2010b). The ramp is composed of the first 30 to 50 codons of CDSs and was shown to be enriched in slow codons (Fredrick & Ibba, 2010; T. Tuller, Carmi, et al., 2010 a; T. Tuller, Waldman, et al., 2010b; Tuller & Zur, 2015; Villada et al., 2017). This enrichment in slow codons occurs across a wide range of organisms. These slow codons are not involved in protein cotranslational folding because the nascent polypeptide is still in the ribosomal exit tunnel. The presence of rare codon clusters at the 5′-termini of genes was hypothesized to limit early translation rates, so as to allow spacing between ribosomes, hence avoiding traffic jams and ribosome fall-offs (T. Tuller, Carmi, et al., 2010a; Zhang et al., 1994). However, it has been a point of debate whether the unusual codon usage in the ramp is solely a consequence of the selection for local mRNA secondary structures or these codons are present in order to slow down translation (Bentele et al., 2013; Goodman et al., 2013; Shah et al., 2013; T. Tuller, Carmi, et al., 2010a). Overall, the nucleotide composition of the ramp appears to be a compromise between its functional elements. The ramp starts with a decreased mRNA structure around the start codon, followed by a region enriched in slow codons and positively charged amino acids, which is followed by a stronger mRNA loop inside the ramp for stability (T. Tuller, Carmi, et al., 2010a; Tuller & Zur, 2015). In chloroplasts, it was demonstrated that not solely CAI dictates protein yield, suggesting that other parameters were equally important (Weiner et al., 2020).

To gain insight into the extent to which these different features are present in chloroplasts, we first quantified the actual enrichment of slow codons in gene ramps. We excluded from the analysis genes with a length of 150 nucleotides or below. We found that rare codons are 84% more frequent in the gene ramp compared to the rest of the mRNA sequence for the high-expression group and 10% more frequent for the low-expression group, but that there was no difference for intermediate genes (Data S6). Additionally, the occurrence of rare codons is highly gene-dependent; some genes, such as *psbA* and *rbcL*, do not show any differential enrichment in rare codons for the ramp, while it is significant for other genes, such as *psaA*, *psaB*, *psbB*, *rps3*, and *ycf1*, which may indicate a special characteristic of membrane proteins (Figure S2). There is a slight increase in positively charged amino acids in the ramp of 25, 12, and 33% for low-, intermediate-, and high-expression genes (Data S6). Furthermore, we analyzed the cumulative GC content of the first 30 codons as a proxy for secondary structure. There is an active pressure to decrease the GC content of the ramp up to the 10th codon for both high- and low-expression groups (Figure S3) and before the start codon (Scharff et al., 2011). However, this pressure does not seem to act similarly on the nucleotide codon positions between the two expression groups. Finally, we also tested if recoding the rare codons of the ramps of *psaA*, *psaB*, and *psbB* with their favored synonymous counterpart had any effect on the mRNA structure of these genes (−50 to +150 nt). We found that recoding with common codons increased the minimal free energy of the centroid structure by 11.3, 3.8 and 1.7 kcal/mol for *psaA*, *psaB*, and *psbB*, respectively, and disrupted some mRNA structural characteristics. These results indicate that some rare codons within the ramp are involved in disrupting the mRNA secondary structures, either around the start of these genes or in the downstream stability loop. However, slow codons and charged amino acids may also be involved in limiting early translation rates.

CONCLUSION

Our work advances our understanding of the complex interactions of codon usage with protein expression. First, we showed that codon usage is highly biased in correlation with the level of gene expression. Highly expressed genes have evolved to utilize a restricted set of codons that is deemed optimal. By comparing expression groups, we demonstrate that this codon optimality information is diluted by the genomic codon usage for proteins that require less optimal expression. This dilution effect was not identified by previous studies that investigated the relationship between codon usage and protein expression, thus leading to interpretations that lacked precision.

The optimization of highly expressed genes permitted to identify the favored codons, which obtain the highest protein yield. The directional mutational bias drives plastidial DNA composition toward an AT-rich equilibrium. Directional mutational bias shapes codon usage of functionally less important genes, while for highly expressed genes other mechanisms counteract the mutational pressure to optimize their translation. In summary, optimal codon usage from duet boxes favors NNC over NNU codons and NNA over NNG codons. Quartet boxes favor their NNU/A codons, except for arginine and glycine, which favor only their respective U-ending codons. While the usage of NNU/C duet boxes can be balanced to modulate expression, G- and C-ending codons of NNA/G duet boxes and quartet boxes are actively avoided in highly expressed genes because their wobble base properties

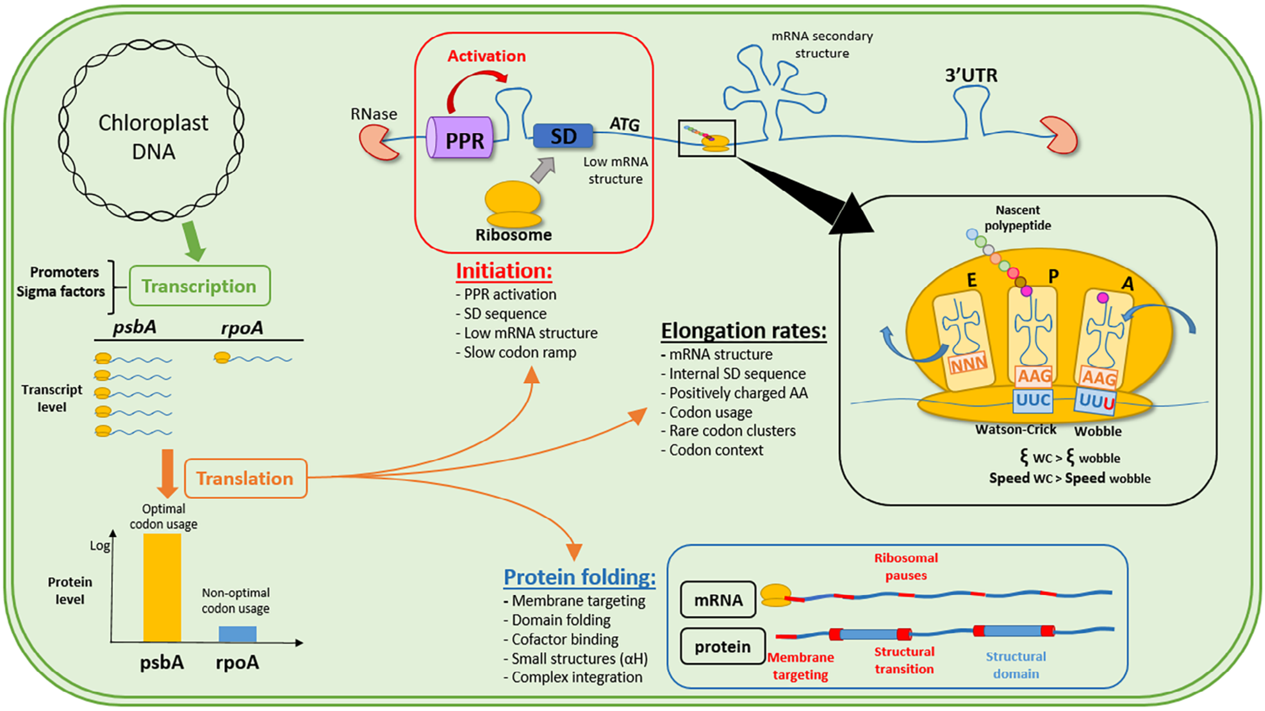

Figure 11. Summary of the main mechanisms affecting protein expression in chloroplasts. From plastidial DNA, genes are transcribed at different levels depending on promoter strength, sigma factors, and growth conditions. Translation yield is optimized through codon usage for high-expression genes (represented by *psbA*), while codons from low-expression genes are rather subject to mutational bias. At the transcript level, PPRs and mRNA structure stabilize the transcripts and enhance translation initiation (in addition to SD sequences). Enrichment in slow codons at the beginning of the gene (ramp) allows ribosome spacing to avoid traffic jams and fall-offs. During elongation, codon usage and codon context modulate translation rates, while mRNA structures, internal SD sequences, and positively charged amino acids create longer ribosomal pauses. These pauses are necessary for protein cotranslational folding, cofactor binding, integration of proteins into multi-protein complexes, or membrane targeting of proteins.

excessively affect translation rates. Additionally, tRNA concentrations affect codon optimality of leucine, serine, arginine, and isoleucine and determine which codon box is favored.

Codon optimality relates to ribosomal translation rates when encountering a given codon. Our analysis shows that tRNA characteristics such as anticodon loop modifications and codon–anticodon pairing affinity are the main factors determining translation rates. This is reflected in the codon usage optimality of highly expressed genes. Apart from protein yield, codon usage is involved directly or indirectly in protein cotranslational folding by influencing mRNA secondary structures, internal SD sequences, the codon context, and rare codon clusters. Synonymous substitutions in specific positions can persist through evolution, in order to maintain features causing ribosomal pauses. This is crucial for membrane targeting, protein domain folding, binding cofactors, such as chlorophyll or metal ions, or the integration of a protein to a multi-enzymatic complex, such as photosystems I and II. Codon composition in the gene ramp is also modulated by mRNA structures and rare codon clusters, which regulate translation initiation.

Overall, our results reconcile the codon usage of the chloroplast with eminent theories of bacterial codon usage (Figure 11). This includes codon usage bias associated with strong gene expression, the influence of mutational bias, and the codon–anticodon pairing affinity properties that modulate the deciphering ability of tRNAs and influence translation rates. Additionally, plastidial local codon usage is in line with the realization that mRNA contains more information than the primary amino acid sequence and is involved in translation initiation and the maintenance of protein features.

The comprehensive insights presented in this work will help codon optimization of heterologous genes for biotechnological applications in algae and plants. Additionally, the chloroplast appears as an entity of choice for advancing our understanding of the entanglement of genetic features with other cellular processes. The simplicity of the chloroplast provides unobstructed view at the first layers of the complexity of the genetic code, which cannot easily be gleaned from more complex organisms.

METHODS

Data

The *C. reinhardtii* chloroplast genome sequence CPv4 was retrieved from the supplementary information of Gallaher et al. (Gallaher et al., 2018) and annotated with the features of the previous version FJ423446.1 (Smith & Lee, 2009). Corrections in annotations were performed according to the work of Gallaher et al., especially for *rpoC1* and *rps2*, which are read as one single open reading frame (ORF). The mRNA quantification data of the chloroplast were retrieved from the same source (Gallaher et al., 2018). We used the mRNA sequencing data obtained under light conditions as an assumption for high/low gene expression, as transcriptomic data are much easier to obtain compared to quantitative proteomics and data are more accurate. Previous research establishing the principles for codon usage bias showed the correlation between codon usage, mRNA content, and protein content (Bennetzen & Hall, 1982; Gouy & Gautier, 1982, Ikemura, 1981; and for protein synthesis rates specifically for chloroplasts: Pfitzinger et al., 1987). These principles are the basis of codon optimization, where codon occurrence is correlated with tRNA content and protein synthesis rates. This previous research showed that transcripts could be used as proxy for the determination of codon usage bias under the assumption that translation can be associated to a proportional increase in the quantities of the encoded proteins. Two sets of genes were created based on high or low mRNA expression. The high-expression set was composed of 18 genes (*psbA*2*, *rbcL*, *psaA-B-C*, *psbB-C-D-F*, *atpA-B-H*, *ycf12*, *petB-D*, *tufA*, *rps12*), and the low-expression set contained 8 genes (*rpoA-B1-B2-C1-C2*, *chlB-L-N*) (referred to as Top 18 and Low 8, respectively) (Data S1). Non-native plastid genes or not characterized genes such as I-cre, Wendy transposons, or ORF 528 were excluded to avoid bias in codon usage analysis. At first, the reference set for the high-expression group was composed of the seven genes (Top 7) that presented an FPKM value of >10 000 in RNAseq analysis (Data S1). The reference set was then extended to 18 genes (Top 18), with FPKM values of >5000, to include a higher number of codons, equivalent to the low-expression group, in order to avoid biases. The two groups provide a very similar outcome in terms of correlation of codon usage bias (Data S2); hence, the largest group was kept for the analyses. The low-expression group was composed of genes with FPKM values of <500, avoiding genes that could provide artifacts, due to lack of characterization or arising from potential RNAseq defects (e.g., *psbN* and *psbI*). The low- and high-expression groups contain a similar number of total codons (8762 and 6151, respectively). Gene grouping did not affect much the correlation between CAI and expression, as the codon usage in the respective expression range is extremely close (Figure 2).

Bioinformatics analysis

The number of codon occurrences in the total CDSs and the different expression groups was calculated with online resources (www.bioinformatics.org/sms2/) and a python-based program we created (github.com/hundvin93/recodon-python). Frequencies of occurrence per 1000 codons were used as normalized values. Codon usage tables for total CDSs and high-expression gene sets were constructed from the codon usage database with the *countcodons* program (www.kazusa.or.jp/codon/countcodon.html) (Data S2).

Codon usage calculations

We analyzed the RSCU, which is a measure of codon frequency, assuming equal codon usage for one amino acid. From the RSCU values, we calculated the relative adaptiveness of a codon W_{ij}, which normalizes the codon frequency to the most used synonymous codon. This permits to identify the so-called 'optimal' and 'non-optimal' codons as frequent and rare, respectively. We used CAI (Sharp & Li, 1987) to assess codon usage bias. CAI uses a set of highly expressed genes hypothesized to possess the most optimal codons as a reference. The high-expression gene sets (Top 7 and Top 18) were used as reference set for codon optimality. CAI values were computed using the code of Fox and Erill (Fox & Erill, 2010) (erilllab.umbc.edu/research/software/201-2/). These calculations were performed on total CDSs and the different expression groups. Fold differences between groups were calculated for each codon (Data S2) as follows:

$$\mathrm{RSCU_{ij}} = \frac{Xi,j}{\frac{1}{ni}\sum_{j=1}^{ni} Xi,j},$$

where X_{ij} is the number of occurrences of the *j*th codon for the *i*th amino acid and n_{ij} is the number of alternative synonymous possibilities for the *i*th amino acid (one to six). Moreover,

$$\mathrm{W_{ij}} = \frac{RSCUi,j}{RSCUi,max} = \frac{Xi,j}{Xi,max},$$

where the values for RSCU and occurrence X_{ij} are compared for the *j*th codon and the most frequent codon (*max*) for the *i*th amino acid. CAI was calculated as follows:

$$\mathrm{CAI} = \left(\prod_{j=1}^{L} Wj\right)^{1/\mathrm{L}},$$

where W_j is the relative adaptiveness for the *j*th codon of a gene of codon length L.

Codon pair calculations

The codon_pair.py function of the previously mentioned python program was used to perform the calculations (github.com/hundvin93/recodon-python). The program extracts the observed occurrence (o_{ij}) of all 3721 possible codon pairs (61 × 61) within CDSs. It also calculates the expected number of occurrences of a codon pair (e_{ij})

similarly to previously described calculations (Buchan et al., 2006; Gutman & Hatfield, 1989):

$$e_{ij} = \frac{C_i\ x\ C_j}{N_p} = c_i \times F_j,$$

where c is the number of occurrences of the ith and jth codons and N_P is the number of codon pairs (number of codons $[N_{TOT}] - 1$). Using this relationship, for a fixed codon i, the probability of encountering a following codon j can be approximated to its frequency of occurrence ($F_j = c_j / N_{TOT}$).

The expected codon pair frequency (e_{ij}) was corrected for the dipeptide bias as previously described (Buchan et al., 2006; Gutman & Hatfield, 1989). This correction compensates the uneven distribution of amino acids along protein primary structures caused by their structural and functional properties. In brief, the over/underrepresentation of a dipeptide is estimated by dividing the actual occurrence of each amino acid pair (P_{kl}), for the kth and lth amino acids, by the expected dipeptide frequency (Q_{kl}) calculated from individual amino acid frequencies:

$$dipeptide\ bias\ (DiPB) = \frac{P_{kl}}{Q_{kl}}.$$

The expected frequency of occurrence for a codon pair (e_{ij}) was corrected by the dipeptide bias to render the following equation:

$$e_{ij,\ norm} = c_i \times F_j \times DiPB_{kl},$$

where c_i is the codon count for the ith codon, F_j is the frequency of occurrence of the jth codon, and $DiPB$ is the dipeptide bias for the kl dipeptide (encoded by the ij codon pair).

The codon pair bias was calculated by comparing the number of occurrences of the observed codon pair (o_{ij}) with its normalized expected frequency ($e_{ij,\ norm}$) for a given codon pair ij:

$$codon\ pair\ bias\ (ij) = \frac{(o_{ij}) - (e_{ij,norm})}{(e_{ij,norm})}.$$

However, this calculation creates an artificial overestimation of codon pair bias for rare codons. Indeed, the expected occurrence/frequency of pairs containing rare codons is very low; thus, the division artificially inflates the codon pair bias (Moura et al., 2005, 2007). Nonetheless, because rare codons exist, they will axiomatically pair with a j codon. Therefore, pairing events involving rare codons are semi-deterministic. Some studies chose to exclude rare codons to avoid this artifact; however, in the chloroplast, rare codons represent a significant fraction and should be considered in this analysis. To circumvent this artifact, we first chose to work with the number of occurrences of codon pairs instead of their frequencies to relate the values obtained with their biological meaning. Then, we classified common codons following a probabilistic model and rare codons following a deterministic model. We used a Poisson law to assess the probability of each codon i to form maximum one codon pair with each codon j. The probability P of the number of occurrences (X) of a pair to be zero or one follows the law

$$P(X \le k) = \sum\nolimits_{k=0}^{1} \frac{\lambda^k}{k!}.e^{-\lambda},$$

where k is the possible number of occurrences of the codon pair ij and λ is the predicted frequency of occurrence of the event. For each pair ij, we have $\lambda = c_i \times F_j$. The average probability for each codon i to form only zero or one pair with all other codons was calculated and codons above a threshold of $P(X \le 1) = 0.75$ were considered to follow a deterministic model. Because of the deterministic nature, the expected number of codon pair occurrences for rare codons cannot be below one. For simplicity, we combined the probabilistic and deterministic models for common and rare codons, respectively, under the same codon pair bias calculation, but we rounded up the expected pair values to the nearest whole number. This way, rare codons are not excluded from calculations and do not provoke an artificial bias, although it makes most pairs involving rare codons artificially underrepresented.

Calculation of free energy of RNA structures

The free energy of RNA structures was calculated using the web-based RNAfold implementation (http://rna.tbi.univie.ac.at/). In order to assess if codons considered as rare present in the ramp have a significant impact on mRNA structures, they were substituted with one of their common counterparts.

ACKNOWLEDGMENT

This study was funded through a PhD fellowship to MFL by the Faculty of Natural Sciences of the Norwegian University of Science and Technology.

AUTHOR CONTRIBUTIONS

MFL and MHM conceived the study and wrote the manuscript. MFL performed the literature review and data analysis. KH wrote the python scripts and analyzed codon usage features.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

The supplemental data are accessible online at https://onlinelibrary.wiley.com. Expression data for mRNA sequencing were retrieved from the online supplemental material of Gallaher et al., 2018. The python programs we created were deposited on GitHub (githubcom/hundvin93/recodon-python) and all other bioinformatic tools utilized in the study are listed in the Methods section.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Method comparison of codon pair bias calculations.

Figure S2. Gene maps of translationally unfavorable codons within expression groups.

Figure S3. Cumulative GC contents of the ramps of Top 18 and Low 8 groups.

Data S1. CAI analysis.

Data S2. Codon usage analysis within gene groups.

Data S3. tRNA copy number.

Data S4. Codon pair bias.

Data S5. Ribosome profiling.

Data S6. Codon usage within gene ramps.

REFERENCES

Agris, P.F. (1991) Wobble position modified nucleosides evolved to select transfer RNA codon recognition: A modified-wobble hypothesis. *Biochimie*, **73**, 1345–1349.

Agris, P.F. (2004) Decoding the genome: A modified view. *Nucleic Acids Research*, **32**, 223–238.

Agris, P.F. (2008) Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications. *EMBO Reports*, **9**, 629–635.

Alkatib, S., Scharff, L.B., Rogalski, M., Fleischmann, T.T., Matthes, A., Seeger, S. *et al.* (2012) The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS Genetics*, **8**(11), e1003076.

Almaraz-Delgado, A.L., Flores-Uribe, J., Pérez-España, V.H., Salgado-Manjarrez, E. & Badillo-Corona, J.A. (2014) Production of therapeutic proteins in the chloroplast of Chlamydomonas reinhardtii. *AMB Express*, **4**, 1–9.

Ambrogelly, A., Palioura, S. & Söll, D. (2007) Natural expansion of the genetic code. *Nature Chemical Biology*, **3**, 29–35.

Anand, A. & Pandi, G. (2021) Noncoding RNA: an insight into chloroplast and mitochondrial gene expressions. *Life*, **11**, 49.

Angov, E. (2011) Codon usage: Nature's roadmap to expression and folding of proteins. *Biotechnology Journal*, **6**, 650–659.

Bai, H. & Lu, A.-L. (2007) Physical and functional interactions between Escherichia coli MutY glycosylase and mismatch repair protein MutS. *Journal of Bacteriology*, **189**(3), 902–910.

Barkan, A. (2011) Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. *Plant Physiology*, **155**, 1520–1532.

Bennetzen, J.L. & Hall, B.D. (1982) Codon selection in yeast. *The Journal of Biological Chemistry*, **257**, 3026–3031.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. (2013) Efficient translation initiation dictates codon usage at gene start. *Molecular Systems Biology*, **9**, 1–10.

Berg, O.G. & Silva, P.J.N. (1997) Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Research*, **25**, 1397–1404.

Bhushan, S., Gartmann, M., Halic, M., Armache, J.P., Jarasch, A., Mielke, T. *et al.* (2010) α-Helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel. *Nature Structural & Molecular Biology*, **17**, 313–317.

Bock, R. (2015) Engineering plastid genomes: methods, tools, and applications in basic research and biotechnology. *Annual Review of Plant Biology*, **66**, 211–241.

Bossi, L. & Roth, J.R. (1980) The influence of codon context on genetic code translation. *Nature*, **286**, 123–127.

Boycheva, S., Chkodrov, G. & Ivanov, I. (2003) Codon pairs in the genome of Escherichia coli. *Bioinformatics*, **19**, 987–998.

Buchan, J.R., Aucott, L.S. & Stansfield, I. (2006) tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Research*, **34**, 1015–1027.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H. *et al.* (2016) Synonymous codons direct Cotranslational folding toward different protein conformations. *Molecular Cell*, **61**, 341–351.

Bulmer, M. (1990) The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Research*, **18**, 2869.

Cabrera, M., Nghiem, Y. & Miller, J.H. (1988) mutM, a second mutator locus in Escherichia coli that generates G.C----T.A transversions. *Journal of Bacteriology*, **170**(11), 5405–5407.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C. *et al.* (2010) A role for codon order in translation dynamics. *Cell*, **141**(2), 355–367.

Cardi, T., Lenzi, P. & Maliga, P. (2010) Chloroplasts as expression platforms for plant-produced vaccines. *Expert Review of Vaccines*, **9**, 893–911.

Celedon, J.M. & Cline, K. (2013) Intra-plastid protein trafficking: How plant cells adapted prokaryotic mechanisms to the eukaryotic condition. *Biochimica et Biophysica Acta - Molecular Cell Research*, **2**, 1833.

Chan, S., Ch'ng, J.H., Wahlgren, M. & Thutkawkorapin, J. (2017) Frequent GU wobble pairings reduce translation efficiency in plasmodium falciparum. *Scientific Reports*, **7**, 723.

Chaney, J.L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A.T., Ngo, K. *et al.* (2017) Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Computational Biology*, **13** (5), e1005531.

Chang, B., Halgamuge, S. & Tang, S.L. (2006) Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene*, **373**, 90–99.

Charneski, C.A. & Hurst, L.D. (2013) Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLOS Biology*, 11(3), e1001508.

Chartier, M., Gaudreault, F. & Najmanovich, R. (2012) Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics*, **28**, 1438–1445.

Chen, C., Zhang, H., Broitman, S.L., Reiche, M., Farrell, I., Cooperman, B.S. *et al.* (2013) Dynamics of translation by single ribosomes through mRNA secondary structures. *Nature Structural & Molecular Biology*, **20**(5), 582–588.

Chevance, F.F.V., Le Guyon, S. & Hughes, K.T. (2014) The effects of codon context on in vivo translation speed. *PLoS Genetics*, **10**(6), e1004392.

Chursov, A., Frishman, D. & Shneider, A. (2013) Conservation of mRNA secondary structures may filter out mutations in Escherichia coli evolution. *Nucleic Acids Research*, **41**, 7854–7860.

Claesson, C., Lustig, F., Boré, T., Simonsson, C., Barciszewska, M. & Lagerkvist, U. (1995) Glycine codon discrimination and the nucleotide in position 32 of the anticodon loop. *Journal of Molecular Biology*, **247**, 191–196.

Clarke, T.F. & Clark, P.L. (2010) Increased incidence of rare codon clusters at 5′ and 3′ gene termini: implications for function. *BMC Genomics*, **11**, 188.

Clarke, T.F., IV & Clark, P.L. (2008) Rare codons cluster. *PLoS One*, **3**(10), e3412.

Cochella, L. (2005) An active role for tRNA in decoding beyond codon:anticodon pairing. *Science*, **80-88**, 308.

Crick, F.H.C. (1966) Codon-anticodon pairing: the wobble hypothesis. *Journal of Molecular Biology*, **19**, 548–555.

Crombie, T., Swaffield, J.C. & Brown, A.J.P. (1992) Protein folding within the cell is influenced by controlled rates of polypeptide elongation. *Journal of Molecular Biology*, **228**(1), 7–12.

Del Campo, E.M. (2009) Post-transcriptional control of chloroplast gene expression. *Gene Regulation and Systems Biology*, **3**, 31–47.

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C. *et al.* (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*, **103**(5), 711–721.

Dietrich, A., Wallet, C., Iqbal, R.K., Gualberto, J.M. & Lotfi, F. (2015) Organellar non-coding RNAs: emerging regulation mechanisms. *Biochimie*, **117**, 48–62.

Doron, L., Segal, N. & Shapira, M. (2016) Transgene expression in microalgae - from tools to applications. *Frontiers in Plant Science*, **7**, 505.

Douglas, S.E. & Turner, S. (1991) Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. *Journal of Molecular Evolution*, **33**, 267–273.

Duchene, A.-M., Giritch, A., Hoffmann, B., Cognat, V., Lancelin, D., Peeters, N.M. *et al.* (2005) Dual targeting is the rule for organellar aminoacyl-tRNA synthetases in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, **102**, 16484–16489.

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

Dyo, Y.M. & Purton, S. (2018) The algal chloroplast as a synthetic biology platform for production of therapeutic proteins. *Microbiol*, **164**, 113–121.

Espah Borujeni, A. & Salis, H.M. (2016) Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. *Journal of the American Chemical Society*, **138**, 7016–7023.

Espah Borujeni, A., Channarasappa, A.S. & Salis, H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Research*, **42**, 2646–2659.

Espah Borujeni, A., Cetnar, D., Farasat, I., Smith, A., Lundgren, N. & Salis, H.M. (2017) Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Research*, **45**, 5437–5448.

Esposito, D. (2003) In vivo evidence for the prokaryotic model of extended codon-anticodon interaction in translation initiation. *EMBO Journal*, **22** (3), 651–656.

Fages-Lartaud, M. & Hohmann-Marriott, M.F. (2022) Overview of tRNA modifications in chloroplasts. *Microorganisms*, **10**, 226.

Fargo, D.C., Zhang, M., Gillham, N.W. & Boynton, J.E. (1998) Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. *Molecular Genetics and Genomics MGG*, **257**(3), 271–282.

Fedyunin, I., Lehnhardt, L., Böhmer, N., Kaufmann, P., Zhang, G. & Ignatova, Z. (2012) tRNA concentration fine tunes protein solubility. *FEBS Letters*, **586**(19), 3336–3340.

Fowler, R.G. & Schaaper, R.M. (1997) The role of the mutT gene of Escherichia coli in maintaining replication fidelity. *FEMS Microbiology Reviews*, **2**(1), 43–55.

Fox, J.M. & Erill, I. (2010) Relative codon adaptation: a generic codon bias index for prediction of gene expression. *DNA Research*, **17**(3), 185–196.

Fredrick, K. & Ibba, M. (2010) How the sequence of a gene can tune its translation. *Cell*, **141**(2), 227–229.

Friemann, A. & Hachtel, W. (1988) Chloroplast messenger RNAs of free and thylakoid-bound polysomes from Vicia faba L. *Planta*, **175**(1), 50–59.

Gallaher, S.D., Fitz-Gibbon, S.T., Strenkert, D., Purvine, S.O., Pellegrini, M. & Merchant, S.S. (2018) High-throughput sequencing of the chloroplast and mitochondrion of Chlamydomonas reinhardtii to generate improved de novo assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. *The Plant Journal*, **93**, 545–565.

Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S. & Grayhack, E.J. (2016) Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell*, **166**, 679–690.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S. & Futcher, B. (2014) Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, **3**, e03735.

Gawroński, P., Jensen, P.E., Karpinski, S., Leister, D. & Scharff, L.B. (2018) Plastid ribosome pausing is induced by multiple features and is linked to protein complex assembly. *Plant Physiology*, **176**, 2557–2569.

Gerresheim, G.K., Hess, C.S., Shalamova, L.A., Fricke, M., Marz, M., Andreev, D.E. *et al.* (2020) Ribosome pausing at inefficient codons at the end of the replicase coding region is important for hepatitis c virus genome replication. *International Journal of Molecular Sciences*, **21**, 1–22.

Ghadimi, M., Heshmati, E. & Khalifeh, K. (2018) Distribution of dipeptides in different protein structural classes: an effort to find new similarities. *European Biophysics Journal*, **47**(1), 31–38.

Gloge, F., Becker, A.H., Kramer, G. & Bukau, B. (2014) Co-translational mechanisms of protein maturation. *Current Opinion in Structural Biology*, **24**, 24–33.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013) Causes and effects of N-terminal Codon bias in bacterial genes. *Science*, **342**(6157), 475–479.

Gouy, M. & Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, **10**, 7055–7074.

Gouy, M. & Grantham, R. (1980) Polypeptide elongation and tRNA cycling in Escherichia coli: A dynamic approach. *FEBS Letters*, **115**(2), 151–155.

Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**(4154), 862–864.

Gray, M.W. (1989) The evolutionary origins of organelles. *Trends in Genetics*, **5**, 294–299.

Gromadski, K.B., Daviter, T. & Rodnina, M.V. (2006) A uniform response to mismatches in codon-anticodon complexes ensures ribosomal Fidelity. *Molecular Cell*, **21**, 369–377.

Grosjean, H. & Westhof, E. (2016) An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Research*, **44**, 8020–8040.

Grosjean, H., de Crécy-Lagard, V. & Marck, C. (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Letters*, **584**, 252–264.

Gu, X., Hewett-Emmett, D., and Li, W.-H. (1998) Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. In: Mutation and Evolution. Contemporary Issues in Genetics and Evolution, pp. 383–391.

Gu, T., Tan, S., Gou, X., Araki, H. & Tian, D. (2010) Avoidance of long mononucleotide repeats in codon pair usage. *Genetics*, **186**, 1077–1084.

Gutman, G.A. & Hatfield, G.W. (1989) Nonrandom utilization of codon pairs in Escherichia coli. *Genetics*, **86**, 3699–3703.

Hess, W.R. & Börner, T. (1999) Organellar RNA polymerases of higher plants. *International Review of Cytology*, **190**, 1–59.

Holtkamp, W., Kokic, G., Jager, M., Mittelstaet, J., Komar, A.A. & Rodnina, M.V. (2015) Cotranslational protein folding on the ribosome monitored in real time. *Science*, **350**(6264), 1104–1107.

Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, **151**, 389–409.

Ikemura, T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes differences in synonymous codon choice patterns of yeast and Escherichiu coli with reference to the abundance of Isoaccepting transfer RNAs. *Journal of Molecular Biology*, **158**(4), 573–587.

Irwin, B., Denis Heck, J. & Hatfield, G.W. (1995) Codon pair utilization biases influence translational elongation step times. *The Journal of Biological Chemistry*, **270**, 22801–22806.

Jacobs, W.M. & Shakhnovich, E.I. (2017) Evidence of evolutionary selection for cotranslational folding. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, 11434–11439.

Jacobson, G.N. & Clark, P.L. (2016) Quality over quantity: optimizing co-translational protein folding with non-'optimal' synonymous codons. *Current Opinion in Structural Biology*, **38**, 102–110.

Jalal, A., Schwarz, C., Schmitz-Linneweber, C., Vallon, O., Nickelsen, J. & Bohne, A.V. (2015) A small multifunctional pentatricopeptide repeat protein in the chloroplast of chlamydomonas reinhardtii. *Molecular Plant*, **8**, 412–426.

Jayabaskaran, C., Kuntz, M., Guillemaut, P. & Weil, J.-H. (1990) Variations in the levels of chloroplast tRNAs and aminoacyl-tRNA synthetases in senescing leaves of Phaseolus vulgaris. *Plant Physiology*, **92**, 136–140.

Jukes, T.H. (1973) Possibilities for the evolution of the genetic code from a preceding form. *Nature*, **246**, 22–26.

Kaiser, C.M., Goldman, D.H., Chodera, J.D., Tinoco, I., and Bustamante, C. (2011) The Ribosome Modulates Nascent Protein Folding. *Science* **334** (6063), 1723–1727.

Kanamaru, K. & Tanaka, K. (2004) Roles of chloroplast RNA polymerase sigma factors in chloroplast development and stress response in higher plants. *Bioscience, Biotechnology, and Biochemistry*, **68**, 2215–2223.

Karlin, S., Mrazek, J. & Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, **179**, 3899–3913.

Kim, J., Klein, P.G. & Mullet, J.E. (1991) Ribosomes pause at specific sites during synthesis of membrane-bound chloroplast reaction center protein D1. *The Journal of Biological Chemistry*, **266**, 14931–14938.

Kim, J., Klein, P.G & Mullet, J.E (1994) Synthesis and turnover of photosystem II reaction center protein D1. Ribosome pausing increase during chloroplast development. *The Journal of Biological Chemistry*, **269**, 17918–17923.

Kimura, M. (1991) The neutral theory of molecular evolution: a review of recent evidence. *The Japanese Journal of Genetics*, **66**, 367–386.

Knight, R.D., Freeland, S.J. & Landweber, L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, **2**(4), research0010.

Komar, A.A. (2009) A pause for thought along the co-translational folding pathway. *Trends in Biochemical Sciences*, **34**(1), 16–24.

Komar, A.A., Lesnik, T. & Reiss, C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Letters*, **462**(3), 387–391.

Kunec, D. & Osterrieder, N. (2016) Codon pair bias is a direct consequence of dinucleotide bias. *Cell Reports*, **14**, 55–67.

Kurata, S., Weixlbaumer, A., Ohtsuki, T., Shimazaki, T., Wada, T., Kirino, Y. ***et al.*** (2008) Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U·G wobble pairing during decoding. *The Journal of Biological Chemistry*, **283**, 18801–18811.

Kurland, C.G. (1987) Features strategies for efficiency and accuracy in gene expression. *Trends in Biochemical Sciences*, **12**, 126–128.

Letzring, D.P., Dean, K.M. & Grayhack, E.J. (2010) Control of translation efficiency in yeast by codon-anticodon interactions. *RNA*, **16**, 2516–2528.

Li, Q. & Qu, H.Q. (2013) Human coding synonymous single nucleotide polymorphisms at ramp regions of mRNA translation. *PLoS One*, **8**(3), e59706.

Li, G.W., Oh, E. & Weissman, J.S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538–541.

Lim, V.I. (1994) Analysis of action of wobble nucleoside modifications on codon-anticodon pairing within the ribosome. *Journal of Molecular Biology*, **240**(1), 8–18.

Lin, K.-F., Sun, C.-S., Huang, Y.-C., Chan, S.I., Koubek, J., Wu, T.-H. ***et al.*** (2012) Cotranslational protein folding within the ribosome tunnel influences trigger-factor recruitment. *Biophysical Journal*, **102**(12), 2818–2827.

Liu, Y. (2020) A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Communication and Signaling: CCS*, **18**, 145.

Lloyd, A.T. & Sharp, P.M. (1992) Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. *Nucleic Acids Research*, **20**, 5289–5295.

Lobry, J.R. (1997) Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**(1-2), 309–316.

Lobry, J.R. & Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acid Research*, **22**(15), 3174–3180.

Lobry, J.R. & Sueoka, N. (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biology*, **3**, research0058.1.

Lu, J. & Deutsch, C. (2005) Folding zones inside the ribosomal exit tunnel. *Nature Structural & Molecular Biology*, **12**, 1123–1129.

Lu, J. & Deutsch, C. (2008) Electrostatics in the ribosomal tunnel modulate chain elongation rates. *Journal of Molecular Biology*, **384**(1), 73–86.

Lustig, F., Bortn, T., Claesson, C., Simonsson, C., Barciszewskat, M. & Lagerkvist, U. (1993) The nucleotide in position 32 of the tRNA anticodon loop determines ability of anticodon UCC to discriminate among glycine codons. *Proceedings of the National Academy of Sciences USA*, **90**, 3343–3347.

Marechal-Drouard, L., Weil, J.H. & Dietrich, A. (1993) Transfer RNAs and transfer RNA genes in plants. *Review of Plant Physiology and Plant Molecular Biology*, **44**, 13–32.

Margulies, M.M., Tiffany, H.L. & Hattori, T. (1987) Photosystem I reaction center polypeptides of spinach are synthesized on thylakoid-bound ribosomes. *Archives of Biochemistry and Biophysics*, **254**(2), 454–461.

Martin, W. & Kowallik, K. (1999) Annotated english translation of mereschkowsky's 1905 paper 'Über natur und ursprung der chromatophoren impflanzenreiche'. *European Journal of Phycology*, **34**, 287–295.

Matsuo, T., Onai, K. & Minagawa, J. (2006) Real-Time Monitoring of Chloroplast Gene Expression by a Luciferase Reporter: Evidence for Nuclear Regulation of Chloroplast Circadian Period Size-dependent symbiont specificity in cnidarian-dinoflagellate symbiosis View project. *Arctic Molecular and Cell Biology*, **26**(3), 863–870.

Mauger, D.M., Siegfried, N.A. & Weeks, K.M. (2013) The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS Letters*, **587**, 1180–1188.

Maul, J.E., Lilly, J.W., Cui, L., Depamphilis, C.W., Miller, W., Harris, E.H. ***et al.*** (2002) The Chlamydomonas reinhardtii plastid chromosome: islands of genes in a sea of repeats. *Plant Cell*, **14**, 2659–2679.

Mignone, F., Gissi, C., Liuni, S. & Pesole, G. (2002) Untranslated regions of mRNAs. *Genome Biology*, **3**, 1–10.

Mohammad, F., Woolstenhulme, C.J., Green, R. & Buskirk, A.R. (2016) Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Reports*, **14**(4), 686–694.

Morton, B.R. (2003) The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *Journal of Molecular Evolution*, **56**, 616–629.

Moura, G., Pinheiro, M., Silva, R., Miranda, I., Afreixo, V., Dias, G. ***et al.*** (2005) Comparative context analysis of codon pairs on an ORFeome scale. *Moura al*, **6**(3), r28.

Moura, G., Pinheiro, M., Arrais, J., Gomes, A.C., Carreto, L., Freitas, A. ***et al.*** (2007) Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS One*, **2**(9), e847.

Moura, G.R., Pinheiro, M., Freitas, A., Oliveira, J.L., Frommlet, J.C., Carreto, L. ***et al.*** (2011) Species-specific codon context rules unveil non-neutrality effects of synonymous mutations. *PLoS One*, **10**(12), e0145593.

Muto, A. & Osawa, S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution (biased mutation pressure/codon usage/neutral theory). *Proceedings of the National Academy of Sciences of the United States of America USA*, **84**, 166–169.

Nakagawa, S., Niimura, Y. & Gojobori, T. (2017) Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine–Dalgarno sequence in prokaryotes. *Nucleic Acids Research*, **45**(7), 3922–3931.

Nakamura, M. & Sugiura, M. (2007) Translation efficiencies of synonymous codons are not always correlated with codon usage in tobacco chloroplasts. *The Plant Journal*, **49**, 128–134.

Nedialkova, D.D. & Leidel, S.A. (2015) Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell*, **161**(7), 1606–1618.

Nghiem, Y., Cabrera, M., Cupples, C.G. & Miller, J.H. (1988) The mutY gene: a mutator locus in Escherichia coli that generates G.C----T.A transversions. *Proceedings of the National Academy of Sciences*, **85**(8), 2709–2713.

Nickelsen, J. & Rengstl, B. (2013) Photosystem II assembly: from cyanobacteria to plants. *Annual Review of Plant Biology*, **64**, 609–635.

Ogle, J.M., Brodersen, D.E., Clemons, W.M., Tarry, M.J., Carter, A.P. & Ramakrishnan, V. (2001) Recognition of cognate transfer RNA by the 30 S ribosomal subunit. *Science*, **292**, 897–902.

Ogle, J.M., Murphy, F.V., Tarry, M.J. & Ramakrishnan, V. (2002) Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell*, **111**, 721–732.

Osawa, S., Muto, A., Ohama, T., Andachi, Y., Tanaka, R. & Yamao, F. (1990) Prokaryotic genetic code. *Expert Review*, **46**, 1097–1106.

Osawa, S., Jukes, T.H., Watanabe, K. & Muto', A. (1992) Recent evidence for evolution of the genetic code. *Microbiological Reviews*, **56**, 229–264.

Parmley, J.L. & Huynen, M.A. (2009) Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLoS Genetics*, **5**, e1000548.

Pechmann, S. & Frydman, J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology*, **20**, 237–243.

Pedersen, S. (1984) Escherichia coli ribosomes translate in vivo with variable rate. *EMBO Joumal*, **3**, 2895–2898.

Pfitzinger, H., Guillemaut, P., Weil, J.H. & Pillay, D.T.N. (1987) Adjustment of the tRNA population to the codon usage in chloroplasts. *Nucleic Acids Research*, **15**, 1377–1386.

Purvis, I.J., Bettany, A.J.E., Santiago, T.C., Coggins, J.R., Duncan, K., Eason, R. ***et al.*** (1987) The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. *Journal of Molecular Biology*, **193**, 413–417.

Qu, X., Wen, J.-D., Lancaster, L., Noller, H.F., Bustamante, C. & Tinoco, I. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature*, **475**, 118–121.

Raynaud, C., Loiselay, C., Wostrikoff, K., Kuras, R., Girard-Bascou, J., Wollman, F.-A. ***et al.*** (2007) Evidence for regulatory function of nucleus-encoded factors on mRNA stabilization and translation in the chloroplast. *Proceedings of the National Academy of Sciences*, **104**, 9093–9098.

dos Reis, M., Savva, R. & Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, **32**, 5036–5044.

Rogalski, M., Karcher, D. & Bock, R. (2008) Superwobbling facilitates translation with reduced tRNA sets. *Nature Structural & Molecular Biology*, **15**, 192–198.

Rosales-Mendoza, S., Paz-Maldonado, L.M.T. & Soria-Guerra, R.E. (2012) Chlamydomonas reinhardtii as a viable platform for the production of recombinant proteins: current status and perspectives. *Plant Cell Reports*, **31**, 479–494.

Rosano, G.L. & Ceccarelli, E.A. (2009) Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain. *Microbial Cell Factories*, **8**, 41.

Scaife, M.A., Nguyen, G.T.D.T., Rico, J., Lambert, D., Helliwell, K.E. & Smith, A.G. (2015) Establishing Chlamydomonas reinhardtii as an industrial biotechnology host. *The Plant Journal*, **82**, 532–546.

Scharff, L.B. & Bock, R. (2014) Synthetic biology in plastids. *The Plant Journal*, **78**, 783–798.

Scharff, L.B., Childs, L., Walther, D. & Bock, R. (2011) Local absence of secondary structure permits translation of mrnas that lack ribosome-binding sites. *PLoS Genetics*, **7**(6), e1002155.

Scharff, L.B., Ehrnthaler, M., Janowski, M., Childs, L.H., Hasse, C., Gremmels, J. *et al.* (2017) Shine-dalgarno sequences play an essential role in the translation of plastid mRNAs in tobacco. *Plant Cell*, **29**, 3085–3101.

Schmitz-Linneweber, C. & Small, I. (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science*, **13**, 663–670.

Schöttler, M.A., Albus, C.A. & Bock, R. (2011) Photosystem I: its biogenesis and function in higher plants. *Journal of Plant Physiology*, **168**, 1452–1461.

Schöttler, M.A., Tóth, S.Z., Boulouis, A. & Kahlau, S. (2015) Photosynthetic complex stoichiometry dynamics in higher plants: biogenesis, function, and turnover of ATP synthase and the cytochrome b6f complex. *Journal of Experimental Botany*, **66**(9), 2373–2400.

Scranton, M.A., Ostrand, J.T., Fields, F.J. & Mayfield, S.P. (2015) Chlamydomonas as a model for biofuels and bio-products production. *The Plant Journal*, **82**, 523–531.

Seligmann, H. & Warthi, G. (2017) Genetic code optimization for Cotranslational protein folding: codon directional asymmetry correlates with antiparallel Betasheets, tRNA synthetase classes. *Computational and Structural Biotechnology Journal*, **15**, 412–424.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J.B. (2013) Rate-limiting steps in yeast protein translation. *Cell*, **153**, 1589–1601.

Shahar, N., Weiner, I., Stotsky, L., Tuller, T. & Yacoby, I. (2019) Prediction and large-scale analysis of primary operons in plastids reveals unique genetic features in the evolution of chloroplasts. *Nucleic Acids Research*, **47**, 3344–3352.

Sharp, P. & Cowe, E. (1991) Synonymous codon usage in Saccharomyces cerevisiae. *Yeast*, **7**, 657–678.

Sharp, P.M. & Li, W.-H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution*, **24**, 28–38.

Sharp, P.M. & Li, W.-H. (1987) The codon adaptation index -a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**(3), 1281–1295.

Shiina, T., Tsunoyama, Y., Nakahira, Y. & Khan, M.S. (2005) Plastid RNA polymerases, promoters, and transcription regulators in higher plants. *International Review of Cytology*, **244**, 1–68.

Shikanai, T. & Fujii, S. (2013) Function of PPR proteins in plastid gene expression. *RNA Biology*, **10**, 1446–1456.

Shine, J. & Dalgarno, L. (1974) The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites (terminal labeling/stepwise degradation/protein synthesis/suppression). *Proceedings of the National Academy of Sciences*, **71**, 1342–1346.

Shpaer, E.G. (1986) Constraints on codon context in Escherichia coli genes their possible role in modulating the efficiency of translation. *Journal of Molecular Biology*, **188**, 555–564.

Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C. & Barral, J.M. (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *Journal of Molecular Biology*, **396**(5), 1310–1318.

Singer, G.A.C. & Hickey, D.A. (2000) Nucleotide bias causes a Genomewide bias in the amino acid composition of proteins. *Molecular Biology and Evolution*, **17**(11), 1581–1588.

Smith, D.R. & Lee, R.W. (2009) Nucleotide diversity of the Chlamydomonas reinhardtii plastid genome: addressing the mutational-hazard hypothesis. *BMC Evolutionary Biology*, (9), 120.

Smith, D. & Yarus, M. (1989) tRNA-tRNA interactions within cellular ribosomes (context effect/RNA structure/translation/suppressor). *Proceedings of the National Academy of Sciences*, **86**, 4397–4401.

Sørensen, M.A. & Pedersen, S. (1991) Absolute in vivo translation rates of individual codons in Escherichia coli the two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *Journal of Molecular Biology*, **222**, 265–280.

Sørensen, M.A., Kurland, C.G. & Pedersen, S. (1989) Codon usage determines translation rate in Escherichia coli. *Journal of Molecular Biology*, **207**, 365–377.

Specht, E., Miyake-Stoner, S. & Mayfield, S. (2010) Micro-algae come of age as a platform for recombinant protein production. *Biotechnology Letters*, **32** (10), 1373–1383.

Spencer, P.S., Siller, E., Anderson, J.F. & Barral, J.M. (2012) Silent substitutions predictably Alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology*, **422**(3), 326–335.

Stadler, M. & Fire, A. (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, **17**, 2063–2073.

Stern, D.B., Goldschmidt-Clermont, M. & Hanson, M.R. (2010) Chloroplast RNA metabolism. *Annual Review of Plant Biology*, **61**, 125–155.

Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proceedings of the National Academy of Sciences*, **47**(8), 1140–1149.

Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences*, **48**(4) 582–592.

Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution (guanine-plus-cytosine content/selective constraints/non-Darwinian evolution). *Proceedings of the National Academy of Science-sUSA*, **85**, 2653–2657.

Suzuki, H. & Morton, B.R. (2016) Codon adaptation of plastid genes. *PLoS One*, **11**, e0154306.

Takai, K. & Yokoyama, S. (2003) Roles of 5-substituents of tRNA wobble uridines in the recognition of purine-ending codons. *Nucleic Acids Research*, **31**, 6383–6391.

Tats, A., Tenson, T. & Remm, M. (2008) Preferred and avoided codon pairs in three domains of life. *BMC Genomics*, **9**, 463.

Thanaraj, T.A. & Argos, P. (1996) Ribosome-mediated translational pause and protein domain organization. *Protein Science*, **5**(8), 1594–1612.

Tsai, C.-J., Sauna, Z.E., Kimchi-Sarfaty, C., Ambudkar, S.V., Gottesman, M.M. & Nussinov, R. (2008) Synonymous mutations and ribosome stalling can Lead to altered folding pathways and distinct minima. *Journal of Molecular Biology*, **383**(2), 282–291.

Tuller, T. & Zur, H. (2015) Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Research*, **43**, 13–28.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J. *et al.* (2010a) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.

Tuller, T., Waldman, Y.Y., Kupiec, M. & Ruppin, E. (2010b) Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences*, **107**, 3645–3650.

Varenne, S., Buc, J., Lloubes, R. & Lazdunski, C. (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology*, **180**(3), 549–576.

Veronica, M., Beligni, V., Yamaguchi, K. & Mayfield, S.P. (2004) The translational apparatus of Chlamydomonas reinhardtii chloroplast. *Photosynthesis Research*, **82**, 315–325.

Villada, J.C., Brustolini, O.J.B. & da Silveira, W.B. (2017) Integrated analysis of individual codon contribution to protein biosynthesis reveals a new approach to improving the basis of rational gene design. *DNA Research*, **24**, 419–434.

Walsh, I.M., Bowman, M.A., Soto Santarriaga, F., Rodriguez, A. & Clark, P.L. (2020) Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences*, **117**, 3528–3534.

Wang, H.H., Yin, W.B. & Hu, Z.M. (2009) Advances in chloroplast engineering. *Journal of Genetics and Genomics*, **36**(7), 387–398.

Wang, F., Johnson, X., Cavaiuolo, M., Bohne, A.-V., Nickelsen, J. & Vallon, O. (2015) *Two Chlamydomonas OPR proteins stabilize chloroplast mRNAs encoding small subunits of photosystem II and cytochrome* b $_6$ f. *The Plant Journal*, **82**(5), 861–873.

Wang, H., McManus, J., and Kingsford, C. (2017) *Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast*. In: *Journal of Computational Biology*, pp. 486–500. Mary Ann Liebert Inc.

Weiner, I., Shahar, N., Marco, P., Yacoby, I. & Tuller, T. (2019) Solving the riddle of the evolution of Shine-Dalgarno based translation in chloroplasts. *Molecular Biology and Evolution*, **36**, 2854–2860.

Weiner, I., Feldman, Y., Shahar, N., Yacoby, I. & Tuller, T. (2020) CSO – A sequence optimization software for engineering chloroplast expression in Chlamydomonas reinhardtii. *Algal Research*, **46**, 101788.

Widmann, M., Clairo, M., Dippon, J. & Pleiss, J. (2008) Analysis of the distribution of functionally relevant rare codons. *BMC Genomics*, **9**, 207.

van Wijk, K.J., Andersson, B. & Aro, E.-M. (1996) Kinetic resolution of the incorporation of the D1 protein into photosystem II and localization of assembly intermediates in thylakoid membranes of spinach chloroplasts. *The Journal of Biological Chemistry*, **271**(16), 9627–9636.

Yampolsky, L.Y. & Stoltzfus, A. (2005) The exchangeability of amino acids in proteins. *Genetics*, **170**(4), 1459–1472.

Yarus, M. & Folley, L.S. (1985) Sense codons are found in specific contexts. *Journal of Molecular Biology*, **182**, 529–540.

Zhang, G. & Ignatova, Z. (2011) Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Current Opinion in Structural Biology*, **21**(1), 25–31.

Zhang, S., Goldman, E. & Zubay, G. (1994) Clustering of low usage codons and ribosome movement. *Journal of Theoretical Biology*, **170**(4), 339–354.

Zhang, G., Hubalewska, M. & Ignatova, Z. (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology*, **16**(3), 271–283.

Zoschke, R. & Barkan, A. (2015) Genome-wide analysis of thylakoid-bound ribosomes in maize reveals principles of cotranslational targeting to the thylakoid membrane. *Proceedings of the National Academy of Sciences*, **112**(13), e1678-87.

Zoschke, R. & Bock, R. (2018) Chloroplast translation: structural and functional organization, operational control and regulation. *Plant Cell*, **30**, 745–770.