International Journal of
Molecular Sciences

MDPI

Article

Analysis of Stop Codons within Prokaryotic Protein-Coding Genes Suggests Frequent Readthrough Events

Frida Belinky [1], Ishan Ganguly [1], Eugenia Poliakov [2], Vyacheslav Yurchenko [3,4,*] and Igor B. Rogozin [1,*]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; frida.belinky@nih.gov (F.B.); ganguly1708@gmail.com (I.G.)

[2] National Eye Institute, National Institutes of Health, Bethesda, MD 20892, USA; poliakove@nei.nih.gov

[3] Life Science Research Centre, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic

[4] Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov University, 119435 Moscow, Russia

[*] Correspondence: vyacheslav.yurchenko@osu.cz (V.Y.); rogozin@ncbi.nlm.nih.gov (I.B.R.)

Abstract: Nonsense mutations turn a coding (sense) codon into an in-frame stop codon that is assumed to result in a truncated protein product. Thus, nonsense substitutions are the hallmark of pseudogenes and are used to identify them. Here we show that in-frame stop codons within bacterial protein-coding genes are widespread. Their evolutionary conservation suggests that many of them are not pseudogenes, since they maintain dN/dS values (ratios of substitution rates at non-synonymous and synonymous sites) significantly lower than 1 (this is a signature of purifying selection in protein-coding regions). We also found that double substitutions in codons—where an intermediate step is a nonsense substitution—show a higher rate of evolution compared to null models, indicating that a stop codon was introduced and then changed back to sense via positive selection. This further supports the notion that nonsense substitutions in bacteria are relatively common and do not necessarily cause pseudogenization. In-frame stop codons may be an important mechanism of regulation: Such codons are likely to cause a substantial decrease of protein expression levels.

Keywords: in-fame stop codon; expression; short-term evolution; population polymorphism; negative selection

check for updates

Citation: Belinky, F.; Ganguly, I.; Poliakov, E.; Yurchenko, V.; Rogozin, I.B. Analysis of Stop Codons within Prokaryotic Protein-Coding Genes Suggests Frequent Readthrough Events. *Int. J. Mol. Sci.* **2021**, *22*, 1876. https://doi.org/10.3390/ijms22041876

Academic Editors: Anne M. Alvarez and Alfredo Ciccodicola

Received: 7 January 2021
Accepted: 9 February 2021
Published: 14 February 2021

1. Introduction

Most single nucleotide substitutions in protein-coding genes either change an encoded amino acid or are synonymous. These two types of substitutions are frequently used in measures of molecular evolution [1]. Another type of substitutions, nonsense mutations, is an understudied phenomenon. Nonsense mutations, by definition, turn a coding (sense) codon into a stop codon that is assumed to result in a truncated protein product. Thus, in-frame stop codons are the hallmark of pseudogenes and are used to identify them [2–5]. Nonsense mutations are potentially highly deleterious, and functional protein-coding genes are not expected to have them at all. Nonetheless, pseudogenes (frequently defined as protein-coding genes with in-frame stop codons) in pro- and eukaryotic genomes persist on the evolutionary timescale, implying that they are maintained by natural selection [6]. In addition, pseudogenes can be transcribed and translated [3,7].

The translation of pseudogenes is not a paradox that molecular biology cannot explain. For example, naturally isolated *Escherichia coli* strains display a wide range of ribosomal fidelity, suggesting that a high rate of translational errors may be favored under some conditions [8]. It was suggested that increased translational errors (including readthrough events) paradoxically provide benefits to microorganisms experiencing environmental stress [9–12]. For example, amino acid misincorporation in the β subunit of RNA polymerase increases resistance of mycobacteria to rifampicin [13,14], and translational errors increase bacterial tolerance to oxidative stress by activating their general stress response [15,16]. Interestingly, in such cases, only subpopulations of genetically identical cells survive severe stresses [14,15], suggesting that stress response activated by translational errors may be heterogeneous (noisy) at the level of individual cells. However, the origin and extent of such heterogeneity remain unknown. A recent study has suggested that fluctuations in the concentrations of translational components lead to UGA readthrough heterogeneity among single cells, which enhances phenotypic diversity of the genetically identical population and facilitates its adaptation to changing environments [17].

In addition, it is well documented that certain stop codons in species across all domains of life are reassigned to sense codons [18–20]. In such cases, it is obvious that a mutation toward a reassigned stop codon is not considered a nonsense mutation and, therefore, has a much-reduced effect on fitness. Interestingly, it has been suggested that there are defined evolutionary steps leading to stop reassignment, including a stop codon that becomes non-essential, as is the case for UAG in *E. coli* [21]. The non-essentiality of UAG does not mean that it does not function as a stop codon, but rather that genes that end with UAG are either less important or have downstream stop codons that can compensate for readthrough events. As a part of the evolutionary reassignment process, it is imperative to have a tRNA recognizing a stop as a sense codon [18,21]. Historically, such tRNAs have been identified as suppressor tRNAs, allowing for a readthrough of a stop codon [22]. It is also possible that regular tRNAs with lack of perfect codon–anticodon interactions would still have some affinity to stop codons and facilitate some level of readthrough events [23]. Recent advancements have provided a solid basis for the development of various experimental tools that are based on the incorporation of biologically occurring or chemically synthesized non-canonical amino acids into the recombinant proteins and even proteomes via reprogrammed protein translation [24]. This takes place in the frame of a greatly expanded genetic code with a variety of codons liberated from their current specificities [24]. An example of such expansions has been documented in some methanogenic archaeal species that synthesize proteins containing selenocysteine or pyrrolysine encoded by stop codons [25].

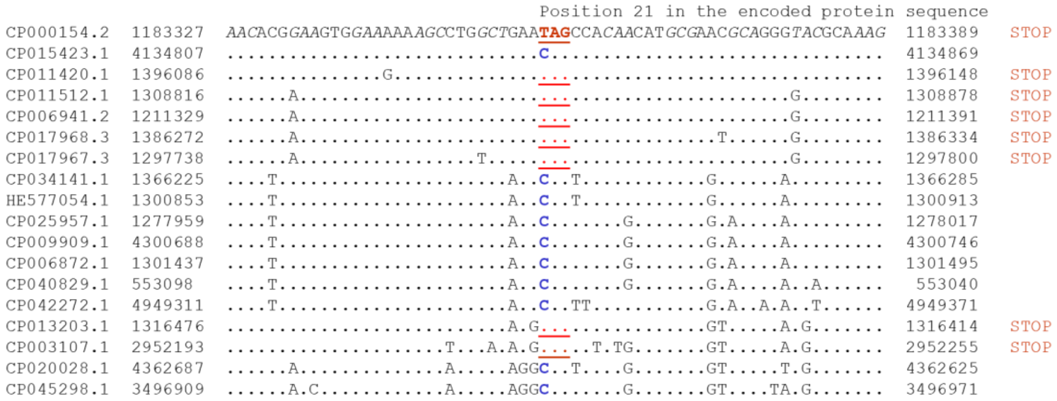
Here we show that stop codons within protein-coding genes are widespread in bacteria and are found in same positions of orthologous genes. This evidence of evolutionary conservation indicates that many of them are not pseudogenes, because they maintain dN/dS values downstream, to the stop codon significantly lower than 1. We also found that double substitutions, where an intermediate step is a nonsense mutation, show accelerated rates of evolution, as compared to null models, indicating that a stop codon was introduced and then changed back to a sense codon via positive selection. This further supports the notion that nonsense substitutions in bacteria are relatively common and do not necessarily cause pseudogenization.

2. Results

2.1. Large-Scale Study: Analysis of Complete Bacterial Genomes

The initial database searches revealed many potential stop codons in protein-coding genes (Supplementary Figure S1); however, we expect that a substantial fraction of them may have resulted from various artifacts (Supplementary Figure S2).

To minimize those artifacts, we focused on stop codons that are present within orthologous protein-coding genes shared by two or more bacterial species (exemplified in Figure 1, for a conserved 3′-phosphoadenosine 5′-phosphosulfate sulfotransferase protein family of *Paenibacillus* spp.; Supplementary Figure S3). In-frame TAG stop codons are in the same position of the alignment (orthologous stop codons) (Figure 1). We use this example to illustrate problems associated with the analysis of in-frame stop codons. There are eight instances of TAG stop codon and ten instances of CAG (Figures 1 and 2).



```
                                                    Position 21 in the encoded protein sequence
CP000154.2  1183327  AACACGGAAGTGGAAAAAAGCCTGGCTGAATAGCCACAACATGCGAACGCAGGGTACGCAAAG  1183389  STOP
CP015423.1  4134807  ..............................C................................  4134869
CP011420.1  1396086  ...............G...............................................  1396148  STOP
CP011512.1  1308816  ......A...............................................G........  1308878  STOP
CP006941.2  1211329  ......A...............................................G........  1211391  STOP
CP017968.3  1386272  ......A........................................T......G........  1386334  STOP
CP017967.3  1297738  ......A.................T.............................G........  1297800  STOP
CP034141.1  1366225  ....T......................A..C..T............G......A.........  1366285
HE577054.1  1300853  ....T......................A..C..T............G......A.........  1300913
CP025957.1  1277959  ....T......................A..C.......G.......G.A....A.........  1278017
CP009909.1  4300688  ....T......................A..C.......G.......G.A....A.........  4300746
CP006872.1  1301437  ....T......................A..C.......G.......G.A....A.........  1301495
CP040829.1  553098   ....T......................A..C.......G.......G.A....A..A......   553040
CP042272.1  4949311  ....T......................A..C..TT...........G.A..A.A..T......  4949371
CP013203.1  1316476  ...........................A.G................GT.....A.G.......  1316414  STOP
CP003107.1  2952193  .....................T...A.A.G.....T.TG.......GT.....A.G.......  2952255  STOP
CP020028.1  4362687  ......A..............A.....AGGC..T....G.......GT.....T.G.......  4362625
CP045298.1  3496909  ......A.C............A.....AGGC.......G.......GT....TA.G.......  3496971
```

Figure 1. Partial alignment of the genes encoding the conserved 3′-phosphoadenosine 5′-phosphosulfate sulfotransferase protein family of *Paenibacillus* spp. The encoded protein sequence is conserved in over 100 *Paenibacillus* spp. (we used BLASTP search in non-redundant protein NR database with default parameters, https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed on 11 February 2021)). Analyses of surrounding genes at NCBI Genomes website (https://www.ncbi.nlm.nih.gov/genome/ (accessed on 11 February 2021), three genes upstream and three genes downstream) did not reveal any obvious conserved gene neighborhood.

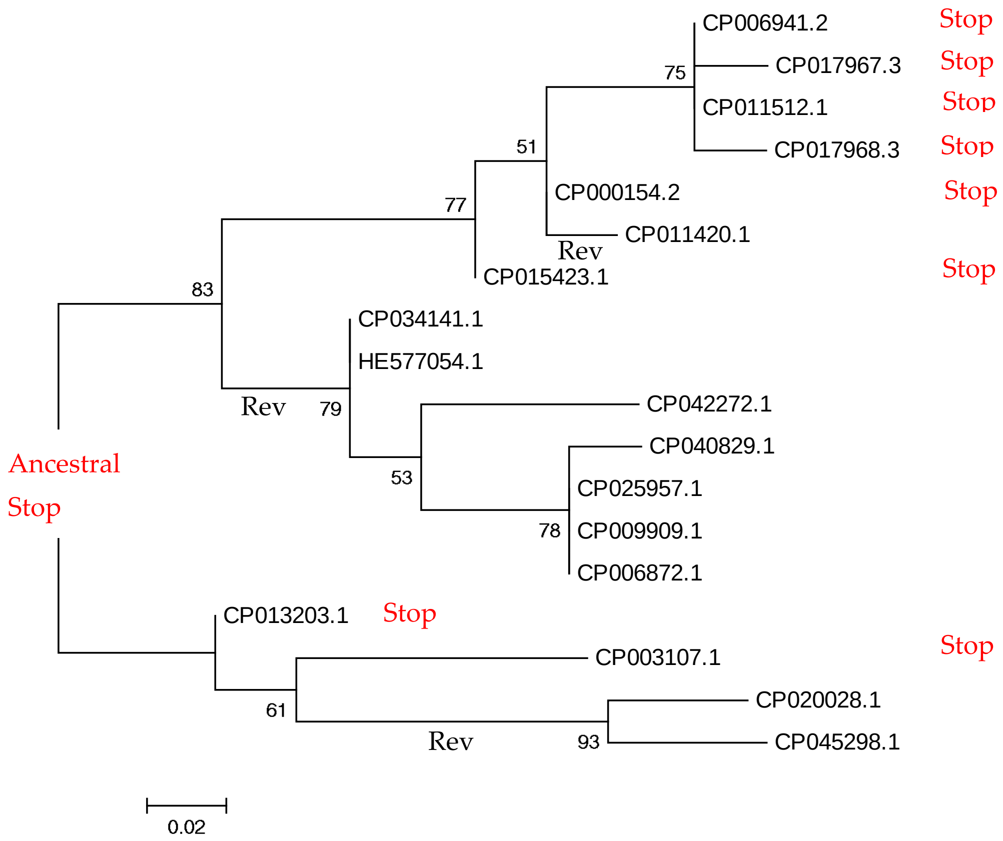


Figure 2. Molecular phylogenetic analysis by maximum-likelihood method. The tree with the highest log likelihood (−247.8) is shown. "Stop" indicated TAG codons, unmarked terminal branches contain "CAG" in the orthologous positions (Figure 1). "Rev" under corresponding branches indicates TAG > CAG substitutions. The bootstrap support values are shown next to the branches; the scale bar indicates the number of substitutions per site. The list of genomes is shown in Supplementary Table S1. The nearly perfect correspondence of the reconstructed (the current figure) tree and the species tree (Figure 2 in [26]) is a well-known property of the vertically inherited sequences [27] and suggests the absence of horizontal gene transfer events in the set of studied sequences (Figure 1).

To delineate evolutionary history of the in-frame stop codon, we mapped TAG/CAG on the maximum-likelihood phylogenetic tree, which was reconstructed using nucleotide sequences (Figure 2). The most parsimonious scenario involves three TAG>CAG reversals, assuming that stop codon was ancestral (if CAG was present in the last common ancestor, two CAG>TAG and two TAG>CAG changes would be needed). In both scenarios the impact of reversal is substantial. It should be noted that the analyzed region is likely to be under purifying selection, as expected for protein-coding genes (mean dN = 0.124, mean dS = 0.170, where dN is the number of nonsynonymous substitutions per nonsynonymous site and dS is the number of synonymous substitutions per synonymous site). The dN/dS value below 1 (0.73 in this case) is an indicator of negative selection.

We estimated the fraction of independent occurrences of orthologous in-frame stops by comparing the number of orthologous and "non-orthologous" (those located in different positions of the same protein family) stop codons (Figure 3). The average fraction of codons (in all protein-coding genes used in this study) that can produce a stop codon as the result of single substitution (stop-codon-prone) is 22%. We used 600 codons (200 amino acids) as a lower bound estimate of a protein-coding gene's length [28]. Moreover, 22% of 600 codons (132 codons) are stop-codon-prone. We observed a twofold excess of "non-orthologous" stop codons (Figure 3). Thus, the fraction of independent events among all "orthologous" stop codons is roughly 2/132 = 0.015. It should be noted that this is a conservative estimate. This indicates that the fraction of independent events (polyphyletic stop codons) is small and the vast majority of orthologous in-frame stop codon substitutions occurred just once.

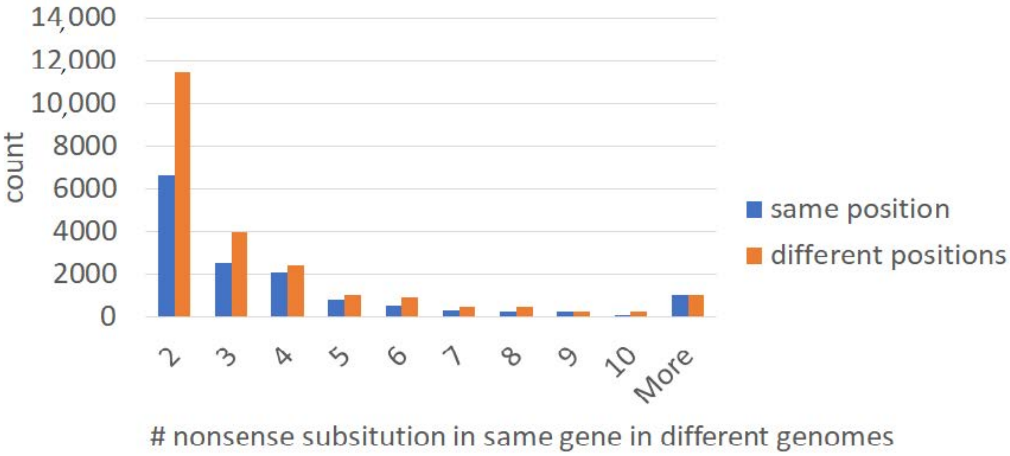


Figure 3. The number of orthologous and "non-orthologous" stop codons (located in different positions of the same protein family) in two or more genomes.

We decided to define polyphyletic and monophyletic stop codons, using a conservative "phyletic index" approach illustrated in the Supplementary Figure S4: If two stop codons were separated by branches that have codons other than stop in the same position of an alignment, we assumed that they evolved independently. The distribution of the phyletic indexes is shown in the Figure 4. A fraction of polyphyletic stop codons is small, confirming that independent origin of "orthologous" in-frame stop codons is, indeed, unlikely.

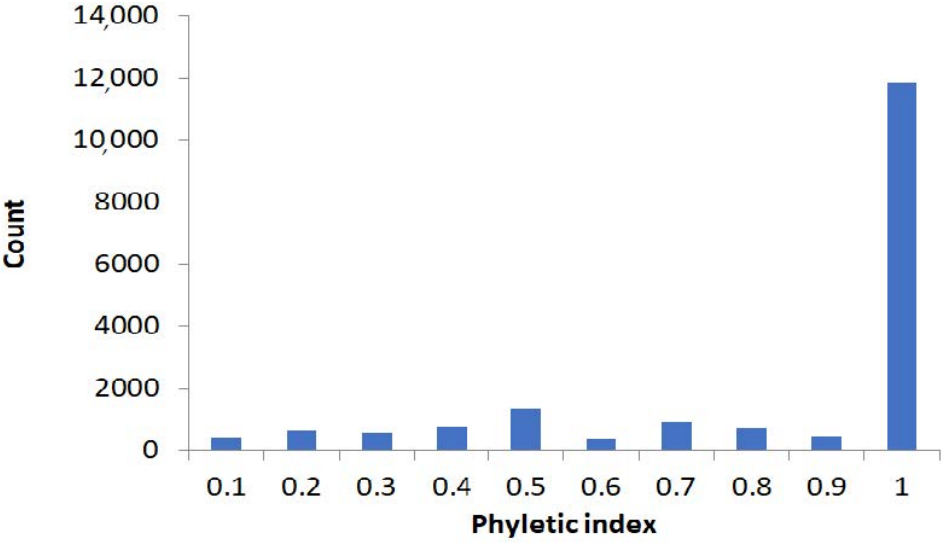


Figure 4. Distribution of the phyletic indexes of nonsense substitution on species trees.

Analysis of the number of mismatches around in-frame stop codons (window of 30 bases) suggested that the vast majority of pairwise comparisons have either a small number or no mismatches (71% pairwise alignments without mismatches, 15% with one mismatch, 7% with two mismatches, 3% with three mismatches, and 4% with more than three mismatches). This result strongly suggests that in-frame stop codons tend to persist for short evolutionary time at the scale of population polymorphism.

Protein-coding genes are expected to be under purifying selection (dN/dS < 1) (the ratio of the number of nonsynonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site, which can be used as an indicator of selective pressure acting on a protein-coding gene) [1,29,30]. If sequences with stop codons are pseudogenes or represent errors of annotation, no purifying selection is expected. The dN and dS values are shown in the Table 1. All three types of in-frame stop codons appeared to be under purifying selection ($p < 0.001$). Nevertheless, the dN/dS values are fairly high, suggesting that either a fraction of sequences are true pseudogenes or in-frame codons represent single nucleotide polymorphisms (SNPs): dN/dS values for negatively selected genes are typically closer to 1 when comparing intra-specific samples, as opposed to inter-specific samples [29].

Table 1. The dN and dS values for regions surrounding in-frame stop codons.

	dN/dS	**dN**	**dS**	***p* Value**
TAA	0.7444	0.0023	0.0031	<0.001
TAG	0.8153	0.0023	0.0029	<0.001
TGS	0.6602	0.0028	0.0043	<0.001

Next, we analyzed dN/dS values before and after stop codons in the first, second, and third terciles of studied genes with in-frame stop codons (Figure 5). The high dN/dS in regions before in-frame stop codons in the first tercile may reflect problems with 5′ end annotations (Supplementary Figure S2), whereas high dN/dS after in-frame stop codons in the third tercile may reflect variability of 3′ ends [30].

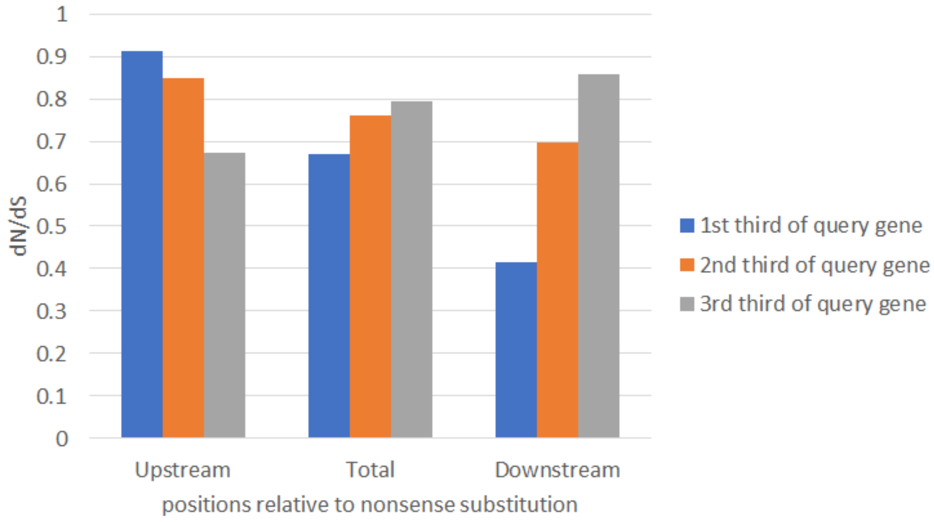


Figure 5. Graph of dN/dS values before and after stop codons in the first, second, and third terciles of studied genes with in-frame stop codons. Deviations from 1 were significant for all dN/dS values ($p < 0.001$).

Analysis of ATGC-COG functional categories did not reveal any major trends, except for an over-representation of the [X} "Mobilome: phages and transposons" (Supplementary Table S2) (9889 cases in total). It should be noted that the total number of cases of the next two most abundant categories ([G] "Carbohydrate transport and metabolism" [6126 cases] and [R] "General function prediction only" [5516 cases]) is substantially greater than the [X] functional category (12,713 vs. 9889) (Supplementary Table S2).

2.2. Small-Scale Study: Analysis of Stop Codons in Triplets of Species

We also performed a small-scale study in well-defined conditions that we used earlier, to study various types of substations [31–33]. It should be noted that signatures of positive selection have been found for double substitutions in stop codons in bacteria (UAG > UGA and UGA > UAG), which could be attributed to the deleterious, non-stop intermediate state, UGG [34]. We applied a similar model for analysis of in-frame stop codons; however, they are considered as intermediate steps in the current study.

Using triplets of genomes with reliable phylogenetic relationships, we calculated frequencies of double and single substitutions in codons, and in double synonymous controls (Supplementary Figure S5). An important distinction of this approach is in use of double synonymous substitutions that served as null models for the double substitutions in codons. This control is important because some replication/repair enzymes are known to produce excessive numbers of simultaneous double substitutions [35–37]. Therefore, we compared the frequencies of all codon double substitutions to all possible types of double synonymous substitutions that were captured in two null models (Supplementary Figure S6). The first null model (NM1, syn_31) included a synonymous substitution in the third position of a codon, followed by another synonymous substitution in the first position of the next codon. The second null model (NM2, syn_33) included non-adjacent synonymous substitutions in third codon positions of consecutive codons. We found that the double fraction (DF), i.e., the observed double substitution frequency divided by sum of the cumulative single substitution frequency and the double frequency (Supplementary Figure S5), was typically higher for the syn_31 model, compared to that of syn_33 model, suggesting a mutational bias toward double substitutions in adjacent positions (Figure 6). The DF is assumed to be proportional to the second-step substitution rate (Supplementary Figure S6). If the elevated DF of codon double substitutions results solely from a multi-nucleotide mutational bias, the comparison to the null model is expected to show no significant difference. Conversely, a significantly lower DF compared to that of the null model is indicative of purifying selection, whereas a significantly higher DF points to positive selection. For example, the double substitution CAG > TCG can be the result of changes two consecutive substitutions CAG > CCG > TCG. Another pathway CAG > TAG > TCG contains the stop codon TAG and was not included in calculations. There are 201 double-substitutions CAG > TCG and 1288 single-substitutions CAG > CCG. The numbers of corresponding synonymous substitutions (NM1, syn_31 model) are 39 and 1023; thus, the excess of double substitutions associated with in-frame stop codons is significant (Supplementary Table S3).

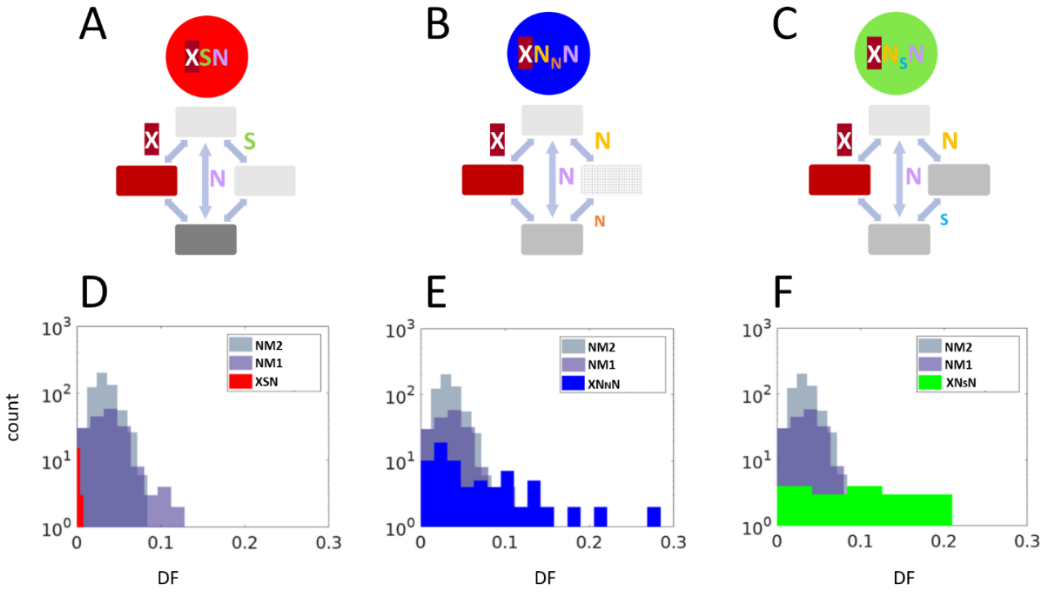


Figure 6. Types and distribution of double fraction (DF) relative to controls. (**A**) Double-substitutions XSN, in which a single substitution can be either synonymous or nonsense, while the double substitution is nonsynonymous, as compared to the ancestral state. (**B**) Double-substitution XN_NN, in which a single substitution can be either nonsynonymous or nonsense, while the double substitution is nonsynonymous to the ancestral state, and the substitution between the intermediate nonsynonymous to the final nonsynonymous is nonsynonymous, as well. (**C**) Double-substitution XN_SN, in which a single substitution can be either nonsynonymous or nonsense, while the double substitution is nonsynonymous to the ancestral state; however, the intermediate state is synonymous to the final state. (**D**) Distribution of DF values of XSN double substitution, compared to NM1 (syn_31) and NM2 (syn_33) control null models. (**E**) Distribution of DF values of XN_NN double substitution, compared to NM1 and NM2 control null models. (**F**) Distribution of DF values of XN_SN double substitution, compared to NM1 and NM2 control null models.

Representing all within-codon double substitutions as "ancestral-intermediate-final", we define the following three combinations, without taking into account the "stop codon intermediate" path (Figure 6A–C): (i) a double-substitution XSN, in which the first single substitution is synonymous, while the double substitution is nonsynonymous, compared to the ancestral state; (ii) a double substitution XN_NN, in which the first single substitution is nonsynonymous, while the double substitution is nonsynonymous to the ancestral state, and the substitution between the intermediate nonsynonymous to the final nonsynonymous is nonsynonymous, as well; (iii) a double substitution XN_SN, in which the first single substitution is nonsynonymous, while the double substitution is nonsynonymous to the ancestral state, however the intermediate state is synonymous to the final state.

The XSN changes are subject to purifying selection (Figure 6D and Supplementary Table S3), whereas many XN_NN and XN_SN changes are likely to have accelerated rates of evolution compared to neutral controls (Figure 6E,F). These results suggest that at least some double mutations have accelerated rates of evolution, due to the escape from in-frame stop-codon state. It should be noted that the analyzed triplets of species have the divergence rate of over 5%, and predicted mutations are unlikely to represent population polymorphisms (fixed mutations) [29]. The overall number of events is not high: For XNnN, we detected

12 positively selected versus 2 negatively selected cases; for XNsN, we detected three positively selected and no negatively selected cases (Supplementary Table S3). This is likely to be the result of small sample sizes.

3. Discussion

In general, the results of small- and large-scale analyses are consistent. According to the large-scale experiment, a vast majority of detected orthologous stop codons tend to exist for a short period of time (roughly corresponding to SNPs). Analyses of dN/dS (Table 1) are consistent with this time estimate. The accelerated evolution of some double mutations with an intermediate in-frame stop codon (Figure 6) reflects potential avoidance of in-frame stop codons for prolonged period of times. Indeed, the small-scale analysis was designed to study fixed mutations at the level of different species rather than SNPs [31–33]. Overall, our results suggest that in-frame stop codons and readthrough events (suppression of in-frame stop codons) are likely to be widespread biological phenomena in prokaryotes, and such stop codons could be advantageous for short periods of time.

We were able to detect numerous putative readthrough events (Figure 3). There are different mechanisms that can cause suppression and recording of stop codons that are detected as readthrough events at the genomic level. For example, in *Escherichia coli*, *Salmonella typhimurium*, and *Bacillus subtilis*, TGA is encountered less frequently than TAA and more frequently than TAG [38,39]. In *E. coli* and *S. typhimurium* TGA can be decoded at very low frequency; when this occurs, the amino acid inserted is tryptophan [40]. Thus, TGA is considered as a "leaky" termination codon [41]. In agreement with this, an unexpectedly high proportion of TGA nonsense mutations was obtained in a collection of chemically induced mutations in the *spoIIR* locus of *Bacillus subtilis* [42]. Six suppressors of the TGA mutations were isolated, and five of the suppressing mutations were mapped to the *prfB* gene encoding protein release factor 2. The five *prfB* mutations also resulted in suppression of the *catA86*-TGA mutation to 19–54% of the expression of *catA86*+, compared to the readthrough level of 6% in the *prfB*+ strain [42].

Other known mechanisms causing readthrough events at the genomic level are TAG to pyrrolysine translation via amber suppression [43], reassignment of TGA or TAG to selenocysteine [44], and TGA to selenocysteine and cysteine recoding [45]. It is not possible to estimate the overall impact of these or similar events [25]; however, it is likely that these events are functionally important and persist over long periods of evolutionary time. Nevertheless, results of the small-scale study of putative readthrough events suggest that many detected readthrough events exist for short periods of time; thus, they are likely to be deleterious at longer evolutionarily timescales. We assume that the vast majority of detected readthrough events do not have functional recording of in-frame stop codons.

Our results are consistent with previous studies of in-frame stop in eukaryotes. Some genes with in-frame stop codons in metazoan species exhibit evolutionary conservation of gene sequences, reduced nucleotide variability, excess synonymous over nonsynonymous nucleotide polymorphism, and other features that are expected in genes or DNA sequences that have functional roles [6]. The cytoplasmic inherited [PSI+] factor has long been known to reduce the efficiency of translation termination and, thereby, increase the readthrough of stop codons in the yeast *Saccharomyces cerevisiae* [46]. In addition, modest changes in the context surrounding the stop codon was shown to result in a substantial reduction in the efficiency of translation termination in eukaryotic organisms [47]. Thus, in-frame stop codons and associated readthrough events may represent a "short-term" mechanism of regulation, because lower levels of expression are expected, due to less efficient translation [48].

In general, gene dosage effect is likely to be an important factor in the evolution of gene families [49,50]. It was suggested that gene duplications that persist in an evolving lineage are beneficial from the time of their origin, due primarily to a protein dosage effect in response to variable environmental conditions [49]. However, a suppression of protein expression may be as important as an increase in protein expression. Thus, it is likely that a decrease in expression due to presence of in-frame stop codons is beneficial for functioning of prokaryotic cells (at least, for short periods of time). Importance of the gene downregulation in response to various factors has been shown for various pro- and eukaryotic species [51–55].

Future analyses of in-frame stop codons and readthrough events in pro- and eukaryotes can produce somewhat unexpected results. For example, recent studies suggested that efficiency of protein translation is likely to be associated with autism spectrum disorders (ASD) [56,57]. This observation connects environmental factors and genetic factors (SNPs and de novo mutations) because each can alter translation efficiency. Many SNPs and de novo mutations are positioned within the coding region of a gene, resulting in premature stop codons [58]. Thus, efficiency of readthrough events could have a functional effect on the protein translation and ASD phenotype [57].

Our results suggest frequent readthrough events in prokaryotes; however, many of them are likely to be deleterious at the scale of long-term evolution, and thus reversals (leading to full restoration of function) are advantageous, as suggested by the small-scale analyses (Figures 2 and 6). It is hard to estimate the frequency of reversal (Figure 2), although it may be high, as suggested by the phyletic index (Figure 4). Taking into account that the fraction of independent events among all "orthologous" stop codons is small (0.015), many orthologous in-frame stop codons with phyletic index less than 1 may reflect the substantial impact of reversals.

In-frame stop codons are most frequently found in the COG functional category [X] "Mobilome: phages and transposons" (Supplementary Table S2). Prophage regions of prokaryotic genomes have at least two evolutionary fates: either domestication or pseudogenization [59,60]. The observed excess of orthologous stop codons in the COG category [X] (Supplementary Table S2) may be explained by the fact that many of them are pseudogenes [59] and do not experience reversals (Figure 2). For example, phages with numerous recorded stop codons (e.g., CrAssphages, [61]) may create an excess of in-frame stop. However, they are much more likely to become pseudogenes rather than "domesticated genes", because of the presence of multiple in-frame stop codons in genes corresponding to recorded regions, e.g., "late" genes compared to phages with the standard genetic code. In general, the fraction of bacteria and phages with recorded stop codons is less than 1% (~0.044%) [19].

Frameshift mutations in protein-coding genes are caused by sequencing errors or programmed ribosomal frameshifting. Programmed ribosomal frameshift events are rare, but they are functionally important [62,63]. We avoided frameshift mutations by using the window (±30 nucleotides) surrounding in-frame stop codons. Even if frameshifts are present in our dataset, it will cause a false "positive selection" ($dN/dS > 1$), because the third position of affected codons (this position is known to be the most variable position) shifts to the first or second positions (these positions are known to be the most conserved positions) [30]. Thus, this can bias our estimates of dN/dS because dN values become artificially large and dS values become artificially small [64,65]. We observed the opposite trend ($dN/dS < 1$, Table 1 and Figure 5). Thus, frameshift mutations are not likely to affect our results and conclusions.

We removed *Mycoplasma* spp. genomes because this is the only known clade in the ATGC database [66] with the stop-codon recording. Our analysis of other potential stop-codon events did not reveal any deviations, except for expected *Mycoplasma* spp. (that we removed from our analyses) and a few *Rickettsia* spp. (Supplementary Figure S7). *Rickettsia* species have been known to contain numerous pseudogenes and are even used as model organisms for studies of pseudogene degradation [67]. Pseudogenes are expected to evolve according to the neutral mode of evolution ($dN/dS \sim 1$) [1]. Thus, pseudogenes cannot substantially bias our estimates of dN/dS in sequences surrounding in-frame stop codons (Table 1 and Figure 5).

There are at least four major sources of in-frame stop codons in our dataset: artifacts of annotations, sequencing errors, pseudogenes, and premature stop codons in functional genes (Supplementary Figure S2). All four types of sources cause problems for performing genome annotations and analyses. In this study, we used prokaryotic genomes with reliable annotations and a high-quality ATGC database, which is based on orthologous gene families [66]. All low-quality proteins were removed from this database. We also effectively removed effects of sequencing errors by using orthologous in-frame stop codons in two or more species (Supplementary Figure S1). These approaches are different from using (semi-)automated annotation pipelines, because all four sources of in-frame stop codons pose major challenges for these systems. Artifacts of previous annotations of closely related sequenced genomes may be, to some extent, resolved by using comparative genomics employed by some (semi-)automated annotation pipelines [68]. However, pseudogenes are an important source of in-frame stop codons in one or several species [67]. In addition to this problem, our analyses suggest that the readthrough mechanism is likely to function in many prokaryotes; thus, many putative "pseudogenes" with in-frame stop codons are functional genes instead. This poses a challenge for genome annotation pipelines, because in-frame stop codons without other signs of gene "degradation" (for example, multiple frameshifts or long deletions) appear to be poor markers of pseudogenization.

4. Materials and Methods

4.1. Identification of Nonsense Substitutions in Protein-Coding Genes

All *Mycoplasma* spp. genomes were removed from our analyses. All protein sequences from the ATGC database [66] were used as a query in TBLASTN (default parameters) searches against all ATGC genomes translated in six frames. When a stop codon was found aligned to an amino acid in the query and an accurate alignment was found 10 amino acids, upstream and downstream, the case was further considered. To reduce the possibility that a stop codon was the result of a sequencing error (Supplementary Figure S2), only orthologous cases with a stop codon in the same position in two or more independent genomes were considered. An example of orthologous an in-frame stop codon is shown in the Figure 2.

4.2. Phylogenetic Analysis and dN/dS Calculations

The maximum-likelihood phylogenetic tree was inferred, using the Tamura–Nei model in MEGA X [69]. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated by using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The analysis involved 18 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions with less than 75% site coverage were eliminated, resulting in a total of 57 positions in the final dataset. CodeML [70] was used to estimate the dN/dS in a 30 bases window, upstream and downstream from the stop. Protein-coding genes are expected to be under purifying selection (dN/dS < 1). If sequences with stop codons are pseudogenes or represent errors of annotation, no purifying selection is expected. Significance of deviations of dN/dS values from 1 was estimated by using the two-tail *t*-test.

4.3. Double Substitution with Stop Intermediate Classification

A double substitution with a stop intermediate is a codon substitution where one of the single substitutions of that double would result in a stop codon. We subdivide double substitutions with stop intermediates into 3 subclasses: (a) S-stop-N—where one intermediate is the stop while the other intermediate is synonymous (S), and the final codon is nonsynonymous (N) to the original codon; (b) N-stop-Nn—where one intermediate is the stop, while the other is N, and also the final codon is N to the original, and the also the final codon is nonsynonymous (n) to the intermediate sense codon; (c) N-stop-Ns—where one intermediate is the stop, while the other is N, and also the final codon is N to the original, and the also the final codon is synonymous (s) to the intermediate sense codon. A subclass 3 can be viewed as a subset of subclass 2, where no selection is expected to affect the second step between the intermediate sense codon and the final codon (because they are synonymous and, thus, identical to the original codon. Thus, the distribution of DFs in subclass 3 is expected to be no different than that double synonymous substitutions, if nonsense mutations are so deleterious that they are completely purged by purifying selection.

4.4. Estimation of Selection on Double Substitutions with Stop Intermediates

Frequencies of double and single substitutions were calculated from the changes between triplets of closely related genomes, as previously described [33]. The DF, calculated as the ratio between the double frequency and the single plus double frequencies, was used as an estimate of the selection pressure on the second step of the double substitutions. When the DF was not different than that of the double synonymous controls, no selection was inferred; when the DF was higher than that of the double synonymous controls, then positive selection was inferred; and when the DF was lower than that of the double synonymous controls, the negative/purifying selection was inferred [33]. Both the *t*-test and signed rank test (two-tail tests) were used to assess the differences of DF between each class and the controls.

4.5. Analysis of Stop Codons within Protein-Coding Genes

To assess a distinction between real pseudogenes (genes with no function at the protein level) and apparent pseudogenes (genes possessing in-frame stop codons that, at least to some degree, can be translated or skipped during translation, resulting in a functional protein), we calculated the dN/dS measure of selection.

Supplementary Materials: The following are available online, at https://www.mdpi.com/1422-0067/22/4/1876/s1. Figure S1: Pipeline steps and the number of detected stop codons. Figure S2: Major technical problems. Figure S3: PSI-BLAST (iteration#2, default parameters) output for putative 3′-phosphoadenosine 5′-phosphosulfate sulfotransferase proteins. Figure S4: Distribution of nonsense substitution on species tree: phyletic index. Figure S5: Conceptual scheme of double substitution analysis and the double fraction (DF) measure. Figure S6: Double synonymous substitutions in adjacent codons used as null models and calculation of DF. Figure S7: Fraction of nonsense substitution compared to the genome size. Table S1: List of *Paenibacillus* species that are shown in Figures 1 and 2. Table S2: Description of ATGC-COG functional categories where in-frame stop codons have been detected. Table S3: Statistics of double mutations and modes of selection.

Author Contributions: Conceptualization, F.B., I.B.R. and V.Y.; methodology, F.B. and I.B.R.; software, F.B., I.G. and I.B.R.; validation, F.B., E.P., I.B.R. and V.Y.; formal analysis, F.B. and I.B.R.; investigation, F.B., I.G., E.P., I.B.R. and V.Y.; resources, F.B. and I.B.R.; data curation, I.B.R., E.P., I.G. and F.B.; writing—original draft preparation, F.B., I.B.R., E.P. and V.Y.; writing—review and editing, F.B., I.G., I.B.R., E.P. and V.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (to F.B., I.G., and I.B.R.); the Intramural Research Program of the National Eye Institute, NIH (to E.P.); the European Regional Funds (CZ.02.1.01/16_019/0000759); and the Grant Agency of Czech Republic (18-15962S) (to V.Y.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We thank members of our laboratories for stimulating discussions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Koonin, E.V.; Rogozin, I.B. Getting positive about selection. *Genome Biol.* **2003**, *4*, 331. [CrossRef]
2. Andersson, J.O.; Andersson, S.G. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **1999**, *9*, 664–671. [CrossRef]
3. Goodhead, I.; Darby, A.C. Taking the pseudo out of pseudogenes. *Curr. Opin. Microbiol.* **2015**, *23*, 102–109. [CrossRef] [PubMed]
4. Holt, K.E.; Thomson, N.R.; Wain, J.; Langridge, G.C.; Hasan, R.; Bhutta, Z.A.; Quail, M.A.; Norbertczak, H.; Walker, D.; Simmonds, M.; et al. Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC Genom.* **2009**, *10*, 36. [CrossRef] [PubMed]
5. Lerat, E.; Ochman, H. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.* **2005**, *33*, 3125–3132. [CrossRef]
6. Balakirev, E.S.; Ayala, F.J. Pseudogenes: Are they "junk" or functional DNA? *Annu. Rev. Genet.* **2003**, *37*, 123–151. [CrossRef]
7. Schrimpe-Rutledge, A.C.; Jones, M.B.; Chauhan, S.; Purvine, S.O.; Sanford, J.A.; Monroe, M.E.; Brewer, H.M.; Payne, S.H.; Ansong, C.; Frank, B.C.; et al. Comparative omics-driven genome annotation refinement: Application across *Yersiniae*. *PLoS ONE* **2012**, *7*, e33903. [CrossRef]
8. Mikkola, R.; Kurland, C.G. Selection of laboratory wild-type phenotype from natural isolates of *Escherichia coli* in chemostats. *Mol. Biol. Evol.* **1992**, *9*, 394–402.
9. Bezerra, A.R.; Simoes, J.; Lee, W.; Rung, J.; Weil, T.; Gut, I.G.; Gut, M.; Bayes, M.; Rizzetto, L.; Cavalieri, D.; et al. Reversion of a fungal genetic code alteration links proteome instability with genomic and phenotypic diversification. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 11079–11084. [CrossRef]
10. Ling, J.; O'Donoghue, P.; Söll, D. Genetic code flexibility in microorganisms: Novel mechanisms and impact on physiology. *Nat. Rev. Microbiol.* **2015**, *13*, 707–721. [CrossRef] [PubMed]
11. Pan, T. Adaptive translation as a mechanism of stress response and adaptation. *Annu. Rev. Genet.* **2013**, *47*, 121–137. [CrossRef] [PubMed]
12. Ribas de Pouplana, L.; Santos, M.A.; Zhu, J.H.; Farabaugh, P.J.; Javid, B. Protein mistranslation: Friend or foe? *Trends Biochem. Sci.* **2014**, *39*, 355–362. [CrossRef] [PubMed]
13. Javid, B.; Sorrentino, F.; Toosky, M.; Zheng, W.; Pinkham, J.T.; Jain, N.; Pan, M.; Deighan, P.; Rubin, E.J. Mycobacterial mistranslation is necessary and sufficient for rifampicin phenotypic resistance. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 1132–1137. [CrossRef] [PubMed]
14. Su, H.W.; Zhu, J.H.; Li, H.; Cai, R.J.; Ealand, C.; Wang, X.; Chen, Y.X.; Kayani, M.U.; Zhu, T.F.; Moradigaravand, D.; et al. The essential mycobacterial amidotransferase GatCAB is a modulator of specific translational fidelity. *Nat. Microbiol.* **2016**, *1*, 16147. [CrossRef]
15. Fan, Y.; Wu, J.; Ung, M.H.; De Lay, N.; Cheng, C.; Ling, J. Protein mistranslation protects bacteria against oxidative stress. *Nucleic Acids Res.* **2015**, *43*, 1740–1748. [CrossRef]
16. Fredriksson, A.; Ballesteros, M.; Peterson, C.N.; Persson, O.; Silhavy, T.J.; Nystrom, T. Decline in ribosomal fidelity contributes to the accumulation and stabilization of the master stress response regulator sigmaS upon carbon starvation. *Genes Dev.* **2007**, *21*, 862–874. [CrossRef]
17. Fan, Y.; Evans, C.R.; Barber, K.W.; Banerjee, K.; Weiss, K.J.; Margolin, W.; Igoshin, O.A.; Rinehart, J.; Ling, J. Heterogeneity of stop codon readthrough in single bacterial cells and implications for population fitness. *Mol. Cell* **2017**, *67*, 826–836. [CrossRef]
18. Osawa, S.; Jukes, T.H. Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* **1989**, *28*, 271–278. [CrossRef]
19. Ivanova, N.N.; Schwientek, P.; Tripp, H.J.; Rinke, C.; Pati, A.; Huntemann, M.; Visel, A.; Woyke, T.; Kyrpides, N.C.; Rubin, E.M. Stop codon reassignments in the wild. *Science* **2014**, *344*, 909–913. [CrossRef]
20. Záhonová, K.; Kostygov, A.; Ševčíková, T.; Yurchenko, V.; Eliáš, M. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr. Biol.* **2016**, *26*, 2364–2369. [CrossRef]
21. Johnson, D.B.; Wang, C.; Xu, J.; Schultz, M.D.; Schmitz, R.J.; Ecker, J.R.; Wang, L. Release factor one is nonessential in *Escherichia coli*. *ACS Chem. Biol.* **2012**, *7*, 1337–1344. [CrossRef]
22. Li, L.; Linning, R.M.; Kondo, K.; Honda, B.M. Differential expression of individual suppressor tRNA(Trp) gene gene family members *in vitro* and *in vivo* in the nematode *Caenorhabditis elegans*. *Mol. Cell Biol.* **1998**, *18*, 703–709. [CrossRef]
23. Bienz, M.; Kubli, E. Wild-type tRNA$^{\mathrm{TyrG}}$ reads the TMV RNA stop codon, but Q base-modified tRNA$^{\mathrm{TyrQ}}$ does not. *Nature* **1981**, *294*, 188–190. [CrossRef]
24. Hoesl, M.G.; Budisa, N. Recent advances in genetic code engineering in *Escherichia coli*. *Curr. Opin. Biotechnol.* **2012**, *23*, 751–757. [CrossRef]
25. Rother, M.; Krzycki, J.A. Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea. *Archaea* **2010**, *2010*, 453642. [CrossRef]
26. Pasari, N.; Gupta, M.; Eqbal, D.; Yazdani, S.S. Genome analysis of *Paenibacillus polymyxa* A18 gives insights into the features associated with its adaptation to the termite gut environment. *Sci. Rep.* **2019**, *9*, 6091. [CrossRef] [PubMed]
27. Olendzenski, L.; Gogarten, J.P. Evolution of genes and organisms: The tree/web of life in light of horizontal gene transfer. *Ann. N. Y. Acad. Sci.* **2009**, *1178*, 137–145. [CrossRef] [PubMed]
28. Brocchieri, L.; Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **2005**, *33*, 3390–3400. [CrossRef]
29. Kryazhimskiy, S.; Plotkin, J.B. The population genetics of dN/dS. *PLoS Genet.* **2008**, *4*, e1000304. [CrossRef] [PubMed]

30. Rogozin, I.B.; Spiridonov, A.N.; Sorokin, A.V.; Wolf, Y.I.; Jordan, I.K.; Tatusov, R.L.; Koonin, E.V. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.* **2002**, *18*, 228–232. [CrossRef]
31. Rogozin, I.B.; Belinky, F.; Pavlenko, V.; Shabalina, S.A.; Kristensen, D.M.; Koonin, E.V. Evolutionary switches between two serine codon sets are driven by selection. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13109–13113. [CrossRef]
32. Belinky, F.; Rogozin, I.B.; Koonin, E.V. Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Sci. Rep.* **2017**, *7*, 12422. [CrossRef]
33. Belinky, F.; Sela, I.; Rogozin, I.B.; Koonin, E.V. Crossing fitness valleys via double substitutions within codons. *BMC Biol.* **2019**, *17*, 105. [CrossRef] [PubMed]
34. Belinky, F.; Babenko, V.N.; Rogozin, I.B.; Koonin, E.V. Purifying and positive selection in the evolution of stop codons. *Sci. Rep.* **2018**, *8*, 9260. [CrossRef]
35. Rogozin, I.B.; Pavlov, Y.I.; Bebenek, K.; Matsuda, T.; Kunkel, T.A. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat. Immunol.* **2001**, *2*, 530–536. [CrossRef]
36. Chan, K.; Gordenin, D.A. Clusters of multiple mutations: Incidence and molecular mechanisms. *Annu. Rev. Genet.* **2015**, *49*, 243–267. [CrossRef] [PubMed]
37. Chen, J.M.; Ferec, C.; Cooper, D.N. Complex multiple-nucleotide substitution mutations causing human inherited disease reveal novel insights into the action of translesion synthesis DNA polymerases. *Hum. Mutat.* **2015**, *36*, 1034–1038. [CrossRef] [PubMed]
38. Andersson, S.G.; Kurland, C.G. Codon preferences in free-living microorganisms. *Microbiol. Rev.* **1990**, *54*, 198–210. [CrossRef]
39. Eggertsson, G.; Soll, D. Transfer ribonucleic acid-mediated suppression of termination codons in *Escherichia coli*. *Microbiol. Rev.* **1988**, *52*, 354–374. [CrossRef]
40. Parker, J. Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.* **1989**, *53*, 273–298. [CrossRef]
41. Roth, J.R. UGA nonsense mutations in *Salmonella typhimurium*. *J. Bacteriol.* **1970**, *102*, 467–475. [CrossRef]
42. Karow, M.L.; Rogers, E.J.; Lovett, P.S.; Piggot, P.J. Suppression of TGA mutations in the *Bacillus subtilis spoIIR* gene by *prfB* mutations. *J. Bacteriol.* **1998**, *180*, 4166–4170. [CrossRef]
43. Wan, W.; Tharp, J.M.; Liu, W.R. Pyrrolysyl-tRNA synthetase: An ordinary enzyme but an outstanding genetic code expansion tool. *Biochim. Biophys. Acta* **2014**, *1844*, 1059–1070. [CrossRef] [PubMed]
44. Kotini, S.B.; Peske, F.; Rodnina, M.V. Partitioning between recoding and termination at a stop codon-selenocysteine insertion sequence. *Nucleic Acids Res.* **2015**, *43*, 6426–6438. [CrossRef] [PubMed]
45. Gonzalez-Flores, J.N.; Shetty, S.P.; Dubey, A.; Copeland, P.R. The molecular biology of selenocysteine. *Biomol. Concepts* **2013**, *4*, 349–365. [CrossRef]
46. Serio, T.R.; Lindquist, S.L. [PSI+]: An epigenetic modulator of translation termination efficiency. *Annu. Rev. Cell Dev. Biol.* **1999**, *15*, 661–703. [CrossRef] [PubMed]
47. Keeling, K.M.; Lanier, J.; Du, M.; Salas-Marco, J.; Gao, L.; Kaenjak-Angeletti, A.; Bedwell, D.M. Leaky termination at premature stop codons antagonizes nonsense-mediated mRNA decay in *S. cerevisiae*. *RNA* **2004**, *10*, 691–703. [CrossRef]
48. Kramarski, L.; Arbely, E. Translational read-through promotes aggregation and shapes stop codon identity. *Nucleic Acids Res.* **2020**, *48*, 3747–3760. [CrossRef]
49. Kondrashov, F.A.; Rogozin, I.B.; Wolf, Y.I.; Koonin, E.V. Selection in the evolution of gene duplications. *Genome Biol.* **2002**, *3*, RESEARCH0008. [CrossRef] [PubMed]
50. Rogozin, I.B. Complexity of gene expression evolution after duplication: Protein dosage rebalancing. *Genet. Res. Int.* **2014**, *2014*, 516508. [CrossRef]
51. Liu, Y.; Zhou, J.; Omelchenko, M.V.; Beliaev, A.S.; Venkateswaran, A.; Stair, J.; Wu, L.; Thompson, D.K.; Xu, D.; Rogozin, I.B.; et al. Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 4191–4196. [CrossRef]
52. Takahashi, S. Positive and negative regulators of the metallothionein gene (review). *Mol. Med. Rep.* **2015**, *12*, 795–799. [CrossRef] [PubMed]
53. Ojo, D.; Rodriguez, D.; Wei, F.; Bane, A.; Tang, D. Downregulation of *CYB5D2* is associated with breast cancer progression. *Sci. Rep.* **2019**, *9*, 6624. [CrossRef]
54. Havis, E.; Duprez, D. EGR1 transcription factor is a multifaceted regulator of matrix production in tendons and other connective tissues. *Int. J. Mol. Sci.* **2020**, *21*, 1664. [CrossRef] [PubMed]
55. Peredo, E.L.; Cardon, Z.G. Shared up-regulation and contrasting down-regulation of gene expression distinguish desiccation-tolerant from intolerant green algae. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 17438–17445. [CrossRef]
56. Rogozin, I.B.; Gertz, E.M.; Baranov, P.V.; Poliakov, E.; Schaffer, A.A. Genome-wide changes in protein translation efficiency are associated with autism. *Genome Biol. Evol.* **2018**, *10*, 1902–1919. [CrossRef] [PubMed]
57. Sokolowski, M.B. Functional testing of ASD-associated genes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26–28. [CrossRef]
58. Ji, X.; Kember, R.L.; Brown, C.D.; Bucan, M. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 15054–15059. [CrossRef]
59. Bobay, L.M.; Touchon, M.; Rocha, E.P. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 12127–12132. [CrossRef] [PubMed]
60. Czajkowski, R. May the phage be with you? Prophage-like elements in the genomes of soft rot Pectobacteriaceae: *Pectobacterium* spp. and *Dickeya* spp. *Front. Microbiol.* **2019**, *10*, 138. [CrossRef] [PubMed]

61. Li, Y.; Gordon, E.; Shean, R.C.; Idle, A.; Deng, X.; Greninger, A.L.; Delwart, E. CrAssphage and its bacterial host in cat feces. *Sci. Rep.* **2021**, *11*, 815. [CrossRef]
62. Baranov, P.V.; Gesteland, R.F.; Atkins, J.F. P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA* **2004**, *10*, 221–230. [CrossRef] [PubMed]
63. Lainé, S.; Thouard, A.; Komar, A.A.; Rossignol, J.M. Ribosome can resume the translation in both +1 or −1 frames after encountering an AGA cluster in *Escherichia coli*. *Gene* **2008**, *412*, 95–101. [CrossRef]
64. Kondrashov, A.S.; Rogozin, I.B. Context of deletions and insertions in human coding sequences. *Hum. Mutat.* **2004**, *23*, 177–185. [CrossRef] [PubMed]
65. Wei, X.; Zhang, J. A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol. Evol.* **2014**, *7*, 381–390. [CrossRef]
66. Kristensen, D.M.; Wolf, Y.I.; Koonin, E.V. ATGC database and ATGC-COGs: An updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation. *Nucleic Acids Res.* **2017**, *45*, D210–D218. [CrossRef] [PubMed]
67. Andersson, J.O.; Andersson, S.G. Pseudogenes, junk DNA, and the dynamics of *Rickettsia genomes*. *Mol. Biol. Evol.* **2001**, *18*, 829–839. [CrossRef]
68. Ejigu, G.F.; Jung, J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology* **2020**, *9*, 295. [CrossRef]
69. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
70. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef]