Why are frameshift homologs widespread within and across species?

Xiaolong Wang[*1], Quanjiang Dong*[2], Gang Chen*[1], Jianye Zhang*[1], Yongqiang Liu*[1], Jinqiao Zhao*[1], Haibo Peng*[1], Yalei Wang*[1], Yujia Cai*[1], Xuxiang Wang*[1], Chao Yang*[1]*

1. *College of Life Sciences, Ocean University of China, Qingdao, 266003, P. R. China*
2. *Qingdao Municipal Hospital, Qingdao, Shandong, 266003, P. R. China*

Abstract

Frameshifted coding genes presumably yield truncated and dysfunctional proteins. We report that frameshift homologs, including frameshift orthologs and frameshift paralogs, are actually widespread within and across species. We proposed that protein coding genes have a *ca*-0.5 quasi-constant shiftability: given any protein coding sequence, at least 50% of the amino acids remain conserved in a frameshifted protein sequence. In the natural genetic code, amino acid pairs assigned to frameshift codon substitutions are more conserved than those to random codon substitutions, and the frameshift tolerating ability of the natural genetic code ranks among the best 6% of all compatible genetic codes. Hence, the shiftability of protein coding genes was mainly predefined by the standard genetic code, while additional sequence-level shiftability was achieved through biased usages of codons and codon pairs. We concluded that during early evolution the genetic code was symmetrically optimized for tolerate frameshifts, so that protein coding genes were endowed an inherent ability to tolerate frameshifting in both forward and backward directions.

[1] To whom correspondence should be addressed: *Xiaolong Wang, Ph.D., Department of Biotechnology, Ocean University of China, No. 5 Yushan Road, Qingdao, 266003, Shandong, P. R. China*, Tel: *0086-139-6969-3150*, E-mail: *Xiaolong@ouc.edu.cn*.

1. Introduction

The genetic code was discovered in the early 1960s [1]. It consists of 64 triplet codons: 61 sense codons for the twenty amino acids and the remaining three nonsense codons for stop signals. The natural genetic code has a number of important properties: (1) The genetic code is universal for all organisms, with only a few variations found in some organelles or organisms, such as mitochondrion, archaea and yeast; (2) The triplet codons are redundant, degenerative and wobble (the third base tends to be interchangeable); (3) In an open reading frame, an insertion/deletion (InDel) causes a frameshift unless the size of the InDel is a multiple of three.

The natural genetic code was optimized for translational error minimization [2], which is extremely efficient at minimizing the effects of mutation or mistranslation errors [3], and optimization for kinetic energy conservation in polypeptide chains [4]. Moreover, it was presumed that the natural genetic code resists frameshift errors by increasing the probability that a stop signal is encountered upon frameshifts, because frameshifted codons for abundant amino acids overlap with stop codons [5].

Presumably, most frameshifted coding DNA sequences (CDSs) yield truncated, non-functional, potentially cytotoxic products, lead to waste of cell energy, resources and the activity of the biosynthetic machinery [6, 7]. Therefore, frameshift mutations were generally considered to be lost-of-function and of little importance for the evolution of novel proteins. However, it was found that frameshift mutations can be retained for millions of years and enable new gene functions to be acquired [8].

Moreover, frameshifted yet functional proteins and their coding genes have been frequently observed [9-13]. For example, in a frameshifted coding gene for yeast mitochondrial cytochrome c oxidase subunit II (COXII), the sequence is translated in an alternative frame by assuming that TGAs do not cause translation termination [13]. However, they have not been considered as a common phenomenon that shares a common underlying mechanism. Moreover, it was reported that frameshift mutations can be retained for millions of years and enable the acquisition of new gene functions [8], shed light into the role of frameshift mutation in molecular evolution.

A protein can be dysfunctioned even by changing a few residues, it is therefore a puzzle how the frameshift proteins kept their structures and functionalities while their sequence has been changed remarkably. Here we report that frameshifted protein homologs widespread within and across species, and this is because in early evolution the natural genetic code was symmetrically optimized for frameshift tolerating, and protein coding genes was endowed an inherent ability that can tolerate frameshifting in both forward and backward directions.

2. Materials and Methods

2.1 Protein and coding DNA sequences

All available protein sequences in all species (Release 2016_04 of 13-Apr-2016 of UniProtKB/TrEMBL, contains 63686057 sequence entries) were downloaded from the UniprotKB protein database. All available reference protein sequences and their coding DNA sequences (CDSs) in nine model organisms, including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Xenopus tropicalis*, *Mus musculus* and *Homo sapiens*, were retrieved from *UCSC, Ensembl* and/or *NCBI* Genome Databases. Ten thousand CDSs each containing 500 random sense codons were simulated by *Recodon* 1.6.0 using default settings [14]. The human/simian immunodeficiency virus (HIV/SIV) strains were derived from the seed alignment in Pfam (pf00516). The CDSs of their envelop glycoprotein (GP120) were retrieved from the HIV sequence database [15].

2.2 Blastp searching for frameshift homologs

A java program, *Frameshift-Translate*, was written and used to translate CDSs in the alternative reading frames, and the frameshift translations were used as queries to search against the UniprotKB protein database by local blastp, and the Blast hits were filtered with a stringent cutoff criterion (*E-value*≤1e-5, *identity*≥30%, and *alignment length*≥20 AAs).

Given a coding gene, its alternative reading frames often contain a certain number of off-frame stop codons. Therefore, frameshifted coding sequences are commonly translated into inconsecutive protein sequences interrupted by some stop signals (*).

In order to find frameshift homologs by blastp, it is better that the query sequences to be consecutive sequences devoid of stop signals. Therefore, in *Frameshift-Translate*, when the CDSs were translated into protein sequences in alternative reading frames, every internal nonsense codon was translated into an amino acid according to a set of *readthrough rules* (Table 1).

The *readthrough rules* were summarized from nonsense suppression tRNAs reported in *E. coli*. The suppressor tRNAs are expressed *in vivo* to correct nonsense mutations, including *amber suppressors* (*supD* [16], *supE* [17], *supF* [18]), *ochre suppressors* (*supG* [19]) and *opal suppressors* (*supU* [18], *su9* [20]). These suppressor tRNAs are taken as *readthrough rules*, because *translational readthrough* occurs upon activity of a suppressor tRNA with an anticodon matching a stop codon. The suppressor tRNAs frequently occur in the negative strand of a regular tRNA [21-23], they are usually undetected, but are expressed in specific conditions. It was found that these suppressor tRNAs off-frame peptides [24-27]. We assumed that suppressor tRNAs are used not only for the readthrough of the nonsense mutations, but also for nonsense codons emerging in the frameshifted coding sequences. This assumption does not require or imply that these readthrough rules must function in frameshifted coding genes, but only to obtain consecutive frameshift protein sequences without the interruption of stop signals.

2.3 Aligning and computing the similarity of the frameshifted protein sequences

A java program, *Frameshift-Align*, was written to translate CDSs in three reading frames, align the three translations and compute their similarities. Every CDS was translated into three protein sequences in its three reading frames in the same strand using the standard genetic code, while all internal nonsense codons were *readthrough* according to the above *readthrough rules* (Table 1). Each protein sequence and the two frameshifted protein sequences were aligned by ClustalW2 using default parameters. The pairwise similarity between a protein sequence and its frameshifted protein sequence is given by the percent of sites in which the matched amino acids are conserved (having a positive or zero amino acid substitution score in a scoring matrix, BLOSSUM62, PAM250 or GON250).

2.4 Computational analysis of frameshift codon substitutions

A protein sequence consisting of *n* amino acids is written as, $A_1 A_2 \ldots A_i A_{i+1} \ldots A_n$, where $A_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i = 1 \ldots n$; its coding DNA sequence consists of *n* triplet codons, which is written as,

$$\boldsymbol{B_1 B_2 B_3} \mid B_4 B_5 B_6 \mid \boldsymbol{B_7 B_8 B_9} \mid \ldots \mid \boldsymbol{B_{3i+1} B_{3i+2} B_{3i+3}} \mid B_{3i+4} B_{3i+5} B_{3i+6} \mid \ldots \mid B_{3n-2} B_{3n-1} B_{3n}$$

Where $B_k = \{A, G, U, C\}$, $k = 1 \ldots 3n$. Without loss of generality, let a frameshift be caused by deleting or inserting one or two bases in the start codon:

(1) *Delete one:* $\boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7} \mid \ldots \mid \boldsymbol{B_{3i+2} B_{3i+3}} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B_{3i+7}} \mid \ldots$

(2) *Delete two:* $\boldsymbol{B_3} B_4 B_5 \mid B_6 \boldsymbol{B_7 B_8} \mid \ldots \mid \boldsymbol{B_{3i+3}} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B_{3i+7} B_{3i+8}} \mid \ldots$

(3) *Insert one:* $B_0 B_1 B_2 \mid \boldsymbol{B_3} B_4 B_5 \mid B_6 \boldsymbol{B_7 B_8} \mid \ldots \mid \boldsymbol{B_{3i+3}} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B_{3i+7} B_{3i+8}} \mid \ldots$

(4) *Insert two:* $B_{-1} B_0 \boldsymbol{B_1} \mid \boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7} \mid \ldots \mid \boldsymbol{B_{3i+2} B_{3i+3}} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B_{3i+7}} \mid \ldots$

We can see that if a frameshift mutation occurred in the first codon, the second codon $B_4 B_5 B_6$ and its encoded amino acid A_2 has two and only two possible changes:

(1) *Forward frameshifting* (*FF*): $\boldsymbol{B_3} B_4 B_5$ ($\rightarrow A_{21}$)

(2) *Backward frameshifting* (*BF*): $B_5 B_6 \boldsymbol{B_7}$ ($\rightarrow A_{22}$)

So do the downstream codons. The results are two frameshifted protein sequences, which were denoted as *FF* and *BF*. In either case, in every codon all three bases are changed when compared base by base with the original codon. Traditionally, codon substitutions are classified into two types according to whether the encoded amino acid is changed or not: (1) *Synonymous substitution* (SS); (2) *Nonsynonymous substitution* (NSS). Based on the above analysis, we classified codon substitutions further into three subtypes: (1) *Random substitution*; (2) *Wobble substitution*; (3) *Frameshift substitution*.

The amino acid substitution score of a frameshift codon substitution is defined as frameshift substitution score (FSS). A java program, *Frameshift-CODON*, was written to compute the average substitution scores in different kinds of codon substitutions by using a scoring matrix (BLOSSUM62, PAM250 or GON250).

2.5 Computational analysis of alternative codon tables

A java program, *Frameshift-GC*, was written to produce “compatible” alternative codon tables according to the method used in reference [3], by changing amino acids assigned to sense codons randomly, while keeping all degenerative codons synonymous. One million alternative genetic codes were selected from all (20! = 2.43290201×10^{18}) “compatible” genetic codes. The sum and average FSSs for each genetic code were computed and sorted, and compared with that of the natural genetic code.

2.6 Analysis of codon pairs and their frameshift substitutions scores

For a given pair of amino acids, written as, $A_1 A_2$, where $A_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i = 1, 2$; its encoding codon pair is written as, $\boldsymbol{B_1 B_2 B_3} \mid B_4 B_5 B_6$, where $B_k = \{A, G, U, C\}$, $k = 1…6$. There are 400 different amino acid pairs and 4096 different codon pairs.

Without loss of generality, let a frameshift be caused by inserting or deleting one base in the first codon, the codon pair and its encoded amino acids has two and only two types of changes:

(1) *Forward frameshifting:* $B_0 \boldsymbol{B_1 B_2} \mid \boldsymbol{B_3} B_4 B_5$ $(\rightarrow A_{11}A_{21})$

(2) *Backward frameshifting:* $\boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7}$ $(\rightarrow A_{12}A_{22})$

A java program, *Frameshift-CODONPAIR*, was written to compute the average amino acid substitution scores for each codon pairs. The result of these calculations is a list of 4096 codon pairs with their corresponding FSSs.

2.7 Computational analysis of the usage of codon and codon pairs

The usage of codons and codon pairs was analyzed on the above dataset using the same method used in reference [28]. The program *CODPAIR* was rewritten in java as the original program is not available. For each sequence, it enumerates the total number of codons, and the number of occurrences for each codon and codon pair. The observed and expected frequencies were then calculated for each codon and codon pair. The result of these calculations is a list of 64 codons and 4096 codon pairs, each with an expected (E) and observed (O) number of occurrences, usage frequency, together with a value for $\chi_1^2 = (O - E)^2/E$. The codons and dicodons whose *O-value* is greater/smaller than their *E-value* were identified as *over-/under-represented*, their average FSSs and the total weighted average FSSs were computed and compared.

3. Results and Analysis

3.1 Frameshift homologs widespread within and across different species

Presumably, frameshift mutations disrupt the function of proteins, as every codon is changed, and often many nonsense codons emerge in a frameshifted CDS. However, we noticed that protein sequences encoded by frameshifted CDSs are actually highly similar to the wild-type protein sequences. For example, in different HIV/SIV strains, such as HIV1J3, SIVCZ and SIVGB, a number of whole or partial, forward or backward, frameshifting occurred in the envelop glycoprotein coding gene, *gp120* (Fig S1A), but their encoded protein sequences remain highly similar to each other (Fig S1B). In addition, these frameshifted GP120 are surely all functional in their host cells. Since HIV was originated from SIVCZ, and SIVCZ was from SIVGB [29-31], obviously, *gp120* underwent a series of evolutionary events, including insertion, deletion, frameshifting, substitution and/or recombination.

As we know, a frameshift mutation is caused by one or more InDels in a protein coding gene whose length is not a multiple of three. Consequently, the reading frame is altered, either fully or partially. In this study, a *frameshift homolog* is defined as a blastp hit using an artificially frameshifted protein sequence as a query. A frameshift homolog is not a frameshift pseudogene, which often contains a certain number of internal nonsense codons and is usually considered dysfunctional. A frameshift homolog, however, does not necessarily contain internal stop codons, and is usually a protein coding gene that encodes a functional protein.

By searching Uniprot protein database using blastp with artificially frameshifted protein sequences as queries, we found that frameshift homologs are actually widespread within a genome and across different species. These frameshift homologs were classified into two types:

(1) ***Frameshift orthologs****:* using a frameshifted protein *A* in a species as query, the blastp hits (frameshift homologs) *in another species*, say protein *a*, represents

functional frameshift coding genes in different species that evolved from a common ancestral gene via speciation and frameshifting (Fig 1A).

(2) ***Frameshift paralogs***: using a frameshifted protein *A* in a species as query, the blastp hits (frameshift homologs) *in the same species*, say protein *B*, represents functional frameshift coding genes in the same species that evolved from a common ancestral gene via duplication and frameshifting (Fig 1B).

As shown in Supplementary Dataset 1, large numbers of frameshift paralogs and orthologs were found exist in the genome of all species tested. For example, in *Homo sapiens*, using frameshifted protein sequences translated from the alternative reading frames of human reference CDSs (hg38, GRCh38) as queries, blastp detected 3974 frameshift paralogs in the human genome and 23224 frameshift homologs (including frameshift orthologs and paralogs) in all species. The blastp hits were filtered with rigorous cutoff criteria, therefore they were considered to be true frameshift homologs that evolved from a common ancestral gene via frameshifting rather than random similarities or artifacts. These frameshift homologs were mapped onto the human genome and displayed in the UCSC genome browser in two custom tracks, *frameshift homologs* and *frameshift paralogs* (Fig 1C), respectively. The supplementary dataset, source code of programs, and custom track files for the UCSC genome browser are available in a webpage on the website of our laboratory (http://www.dnapluspro.com/?page_id=392223).

3.2 Frameshift proteins are always highly similar to their wild-types

To test whether or not frameshifted protein sequences are always similar to their wild-types, their coding sequences were translated each into three protein sequences in the three different reading frames, the three translations were aligned by ClustalW, and their pairwise similarities were computed. For a given CDS, let $\delta_{ij} = \delta_{ji}$ ($i, j=1,2,3, i \neq j$) be the similarity between a pair of protein sequences encoded in reading frame i and j, the average pairwise similarity among the three protein sequences translated from the three different reading frames on the same strand is defined as *the shiftability of the protein coding gene* (δ),

$$\delta = \frac{1}{3}(\delta_{12} + \delta_{13} + \delta_{23})$$

By analyzing all available reference CDSs in nine major model organisms, We show that δ was centered approximately at 0.5 in all CDSs, in all species, as well as in the simulated CDSs (Table 2 and Supplementary Dataset 2). In other words, *in most coding genes*, the three protein sequences encoded in their three reading frames are always highly similar to each other, with an average similarity of ~50%. Therefore we proposed that *protein coding genes have ca-0.5 quasi-constant shiftability, i.e., in most protein coding genes, approximately 50% of the amino acids remain conserved in a completely frameshifted protein sequence*.

For a partial frameshifted coding sequence of length L, if a frameshift starts at L_s and ends at L_e, obviously, site conservation is inversely proportional to frameshifted sites, therefore the partial frameshifts are all highly similar to the wild-type. Hence it is guaranteed that in a frameshifted protein at least half of the sites are conserved when compared to the wide-type, forming the basis of frameshift tolerating. However, this does not imply that all frameshifted variants are functional, but at least some of them could maintain the function.

3.3 The genetic code was optimized for frameshift tolerating

In Table 2, the shiftability of the protein coding genes is similar in all species, and all genes, and the standard deviation is very small, suggesting that the shiftability is largely species- and sequence-independent. This implies that the shiftability is defined mainly by the genetic code rather than by DNA/protein sequences. This is also suggested by simulated protein coding sequences, whose shiftability is comparable with that of the real coding genes.

As described above in the method section, we computed the average amino acid substitution scores respectively for random, wobble and forward/backward frameshift codon substitutions. As shown in Table 3 and Supplementary Dataset 3, in all 4096 possible codon substitutions, most (192/230=83%) of the synonymous substitutions are wobble, and most (192/256=75%) wobble substitutions are synonymous, thus the average substitution score of the wobble substitutions is the highest. For frameshift codon substitutions, except for the 64 codons unchanged in frameshifting, only a small proportion (4.1%) of the changed codons are synonymous and the others (95.9%) are nonsynonymous. In addition, although only a small proportion (7.0%) of frameshift substitutions are synonymous (Table 4), a large proportion (35.9%) of them are positive (including SSs and positive NSSs), which is significantly higher than that of random substitutions (25.7%). In summary, in the natural genetic code, SSs are assigned mainly to wobble substitutions, while positive NSSs are assigned mainly to frameshift substitutions.

In addition, no matter which substitution scoring matrix (BLOSSUM62, PAM250 or GON250) was used for computation, the average FSSs are significantly higher than those of the random substitutions (t-test P << 0.01), suggesting that the amino acid substitutions assigned to the frameshift substitutions are more conservative than those to the random substitutions.

The scoring matrix is widely used to determine similarity and conservation in sequence alignment and blast searching, which forms the basis of most bioinformatics analysis. In any commonly used scoring matrix, either BLOSSUM62, PAM250 or GON250, most amino acid substitution scores are negative and the percent of positive scores is less than 30%. So random codon substitutions will has about 30% percent of positive scores. However, the percent of positive scores for frameshift substitution is about 50%. As shown in Table 3, for most coding sequence, a frameshifted protein will be always highly similar to the wild-type: ~35% similarity derived from the frameshift substitutions, plus ~25% similarity derived from the random substitutions, minus their intersection (~10%), explained the ~50% similarities observed among the wild-type and the corresponding frameshifted protein sequences (Table 2). Therefore, it is suggested that the shiftability of protein-coding genes was predefined mainly by the genetic code, and is largely independent on the proteins or coding sequences, clearly demonstrating that the genetic code has a feature of frameshift tolerating.

In order to further investigate optimization for frameshift tolerance of the natural genetic code, one million alternative genetic codes were randomly selected from all ($20! = 2.43290201 \times 10^{18}$) “compatible” genetic codes by changing the amino acids assigned to the sense codons randomly, while keeping all degenerative codons synonymous. By computing and sorting the average FSSs for these alternative genetic codes (Table 5), the FSSs of the natural genetic code ranks in the best 6.3% of all compatible genetic codes. Hence the genetic code was indeed optimized for tolerating frameshifts .

3.4 The genetic code is symmetric in frameshift tolerating

The genetic code shows the characteristics of symmetry in many aspects [32-34], and it evolved probably through progressive symmetry breaking [35-37]. Here in all CDSs both forward and backward frameshift proteins have comparable similarities with the wild-type (Table 2); In addition, in the natural genetic code both forward and backward frameshift substitutions have the same number of SSs/NSSs and frameshift substitution scores (Table 3). These data suggested that the genetic code is also symmetric in terms of shiftability and frameshift tolerating, so that a protein coding gene has an ability to tolerate frameshifting in both forward and backward directions at the same time (Fig 2). This could also explain why in the natural genetic code the codons are triplet but not tetrad: triplet codon could be kept symmetric for both forward and backward frameshifting easily, while for tetrad codons the situation will be more complicated in frameshifting.

3.5 The shiftability at sequence level

Although the shiftability of a coding sequence is predefined mainly by the genetic code, shiftability may also exist at the sequence level. Functionally important coding genes, such as housekeeping genes, which are more conserved, may also have greater shiftability when compared with other genes. At first, we thought that a biased usage of codons may contribute to the sequence-level shiftability. However, as shown in Table 6 and Supplementary Dataset 4, it is somewhat surprising that in *E. coli* and *C. elegans* the average FSSs weighted by their codon usages are even lower than for unweighted calculations (equal usage of codons). In the other species, although the weighted average FSSs are higher than for unweighted analyses, in all species the difference is never statistically significant ($P>0.05$), suggesting that the usage of codons has little or no direct impact on the shiftability, but it may influence the shiftability indirectly, *e.g.*, by shaping the pattern of codon pairs.

Given a pair of amino acids, $A_1 A_2$, if A_1 and A_2 have m_1 and m_2 degenerative codons, respectively, their encoding dicodon, $B_1 B_2 B_3 | B_4 B_5 B_6$, has $m_1 \times m_2$ possible combinations, called *degenerative codon pairs* (DCPs). It has been reported that codon pair usages are highly biased in various species, such as bacteria, human and animals [28, 38-43]. As shown in Table 7, and Supplementary Dataset 5, in all species tested, the average FSSs of the over-represented codon pairs are all positive, while those of the under-represented codon pairs are all negative; in addition, the weighted average FSSs of all codon pairs are positive, while that of the equal usage of codon pairs is negative, suggesting that in these genomes frameshift-tolerable DCPs are present more frequently than non-frameshift-tolerable DCPs. There have been many reports on the causes and consequences of the codon bias, such as gene expression level [44-49], mRNA structure [50-57], protein abundance [54, 58-60], and stability [61-63]. Based on the above analysis, it is suggested that the usages of codon pairs have an impact on the frameshift tolerability (shiftability) of the protein-coding genes. Therefore, sequence-level shiftability does exist, and was achieved through a biased usage of codons and codon pairs.

4. Discussion

4.1 The genetic code was optimized for frameshift tolerating

The natural genetic code results from selection during early evolution, as it seems optimized along several properties when compared with other possible genetic codes [64-75]. It was pointed out that the natural genetic code was optimized for translational error minimization, because amino acids whose codons differed by a single base in the first and third codon positions were similar with respect to polarity and hydropathy, and the differences between amino acids were specified by the second codon position is explained by selection to minimize the deleterious effects of translation errors during the early evolution of the genetic code [2]. In addition, it was reported that only one in every million alternative genetic codes is more efficient than the natural genetic code, which is extremely efficient at minimizing the effects of point mutation or translation errors [3]. It was demonstrated that the natural genetic code is nearly optimal for allowing additional information within coding sequences, such as out-of-frame hidden stop codons (HSCs) and secondary structure formation (self-hybridization) [5].

In the above, we showed that the code- and sequence-level shiftability of coding genes guaranteed at least half of the sites are kept conserved in a frameshifted protein when compared with the wild-type protein. This is the basis for frameshift tolerating, and explains why frameshift homologs were found widespread within and across species. In addition, the wild type is not necessarily the "*best*" form. In a frameshifted protein the other half of sites change into dissimilar amino acids, probably provides a fast and effective means of molecular evolution for improving or altering the structure and function of proteins.

4.2 The universality of the shiftability

Here we analyzed the shiftability of protein-coding genes only in some model organisms, thus it is interesting to ask whether or not the mechanism is preserved in other species. It has been reported that in some animal species frameshift mutations are tolerated by the translation systems in mitochondrial genes [76-78]. For example, a +1 frameshift insertion is tolerated in the *nad3* in some birds and reptiles [76]. Moreover, frameshifted overlapping genes have been found in mitochondria genes in fruit fly and turtles [79, 80]. It has been reported that in *E. coli* the levels of stop codon readthrough and frameshifting are both high and growth phase dependent [81]. Meanwhile, translational stop codon readthrough has been widely observed in many species [82-89]. Frameshift tolerating was explained by a *programmed translational frameshifting* mechanism [90-93]. However, the shiftability of protein coding genes might also contribute to the expression, functioning, repairing and evolution of the protein coding genes in many species.

5. Conclusion

The above analysis conclude that frameshift homologs are widespread within a genome and across species, because the natural genetic code was optimized symmetrically for frameshift tolerating. The codon- and sequence-level shiftability guarantees near-half conservation after a frameshifting event, endows protein coding genes an inherent ability to tolerate frameshifting in both forward and backward directions. The natural genetic code, which exists since the origin of life, seems optimized by competition with other variant codes during early evolution. The shiftability of the protein coding genes, as an ingenious "*underlying design*" of the natural genetic code, serves as an innate mechanism for cells to deal with frameshift mutations.

Author Contributions

Xiaolong Wang conceived the main ideas, designed the experiments, coded the programs, analyzed the data, prepared the figures, tables and wrote the paper; Quanjiang Dong proofread the paper and gave conceptual advices. Gang Chen and Jianye Zhang provided materials and supports. Yujia Cai analyzed the FSS data for alternative genetic codes. Yongqiang Liu, Jinqiao Zhao and Chao Yang analyzed some data; Xuxiang Wang, Haibo Peng and Yalei Wang performed some experiments. All authors discussed and suggested improvements.

Acknowledgements

This study was funded by *National Natural Science Foundation of China* through Grant 31571369. We acknowledge Dr. Hervé Seligmanna from Aix-Marseille University for his valuable comments and suggestions on an earlier preprint of this manuscript posted in the PeerJ PrePrints [94].

Figure Legends

Fig 1. Diagram of different frameshift homologs. (A) Frameshift orthologs; (B) Frameshift paralog; (C) Custom tracks for the frameshift homologs displayed in the UCSC genome browser;

Fig 2. The alignment of the coding DNA and the protein sequences of HIV GP120. (A) The alignment of coding DNA sequences of HIV GP120. (B) The alignment of protein sequences of HIV GP120, shows that the coding genes contain a number of frameshifting events, in other words, the coding gene is expressed in different reading frames in different virus strains.

Additional Information

We declare that the authors have no competing interests as defined by Nature Publishing Group, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Table 1. The natural suppressor tRNAs (*readthrough rules*) for nonsense mutations.

Site	tRNA	Wild type		Correction	
	(AA)	**Code**	**Anti-code**	**Code**	**Anti-code**
supD	Ser (S)	→ UCG	CGA←	→ UAG	CUA←
supE	Gln (Q)	→ CAG	CUG←	→ UAG	CUA←
supF	Tyr (Y)	→ UAC	GUA←	→ UAG	CUA←
supG	Lys (K)	→ AAA	UUU←	→ UAA	UUA←
supU	Trp (W)	→ UGG	CCA←	→ UGA	UCA←

Table 2. The similarities of natural and simulated proteins and their frameshift forms.

No.	***Species***	***Number of***	***Average Similarity***					
		CDSs	δ_{12}	δ_{13}	δ_{23}	δ	***MAX***	***MIN***
1	*H. sapiens*	71853	0.5217±0.0114	0.5044±0.0122	0.4825±0.0147	0.5028±0.0128	0.5948	0.4357
2	*M. musculus*	27208	0.5292±0.042	0.5058±0.0437	0.4869±0.0418	0.5073±0.0425	0.8523	0.1000*
3	*X. tropicalis*	7706	0.5190±0.0013	0.4987±0.0013	0.4855±0.0008	0.5010±0.0008	0.5962	0.4790
4	*D. rerio*	14151	0.5234±0.0007	0.5022±0.0008	0.4921±0.0005	0.5059±0.0004	0.5240	0.4784
5	*D. melanogaster*	23936	0.5162±0.0015	0.4921±0.001	0.4901±0.0013	0.4995±0.0008	0.6444	0.4667
6	*C. elegans*	29227	0.5306±0.0007	0.5035±0.0008	0.5002±0.001	0.5115±0.0006	0.6044	0.4864
7	*A. thaliana*	35378	0.5389±0.0508	0.5078±0.0481	0.5062±0.048	0.5176±0.0388	0.9540	0.2162*
8	*S. cerevisiae*	5889	0.5174±0.0011	0.4811±0.001	0.5072±0.0006	0.502±0.0007	0.5246	0.4577
9	*E.coli*	4140	0.5138±0.0019	0.4871±0.0046	0.481±0.0015	0.494±0.0012	0.7778	0.4074
10	*Simulated*	10000	0.5165±0.0282	0.4745±0.0272	0.4773±0.0263	0.4894±0.0013	0.6489	0.3539

* Very large and small similarity values were observed in a few very short or repetitive peptides.

Table 3. The amino acid substitution scores for different kind of codon substitutions.

Codon Substitution		*ALL (Random)*	*Frameshift*		*Wobble*
			FF	***BF***	
	All	4096	256	256	256
Type of Codon Substitution	*Unchanged (%)*	64 (1.6%)	4 (1.6%)	4 (1.6%)	64 (25%)
	Changed (%)	4032 (98.4%)	252 (98.4%)	252 (98.4%)	192 (75%)
	SS (%)	230 (5.6%)	18 (7.0%)	18 (7.0%)	192 (75%)
	NSS-Positive (%)	859 (20.1%)	76 (29.7%)	72 (28.1%)	40 (15.6%)
	NSS-Negative (%)	3007 (73.4%)	162 (63.3%)	166 (64.8%)	24 (9.4%)
Average Substitution Score	*BLOSSUM62*	-1.29	-0.61	-0.65	3.77
	PAM250	-4.26	-0.84	-0.84	3.68
	GON250	-10.81	-1.78	-1.78	35.60

SS/NSS: synonymous/nonsynonymous substitution; FF/BF: forward/backward frameshift codon substitution.

Table 4. The synonymous frameshift substitutions

Forward Frameshifting					*Backward Frameshifting*				
From			*To*		*From*			*To*	
1	AAA	K	AAA	K	1	AAA	K	AAA	K
2	AAA	K	AAG	K	2	AAG	K	AAA	K
3	GGG	G	GGA	G	3	GGA	G	GGG	G
4	GGG	G	GGG	G	4	GGG	G	GGG	G
5	GGG	G	GGC	G	5	GGC	G	GGG	G
6	GGG	G	GGT	G	6	GGT	G	GGG	G
7	CCC	P	CCA	P	7	CCA	P	CCC	P
8	CCC	P	CCG	P	8	CCG	P	CCC	P
9	CCC	P	CCC	P	9	CCC	P	CCC	P
10	CCC	P	CCT	P	10	CCT	P	CCC	P
11	CTT	L	TTA	L	11	TTA	L	CTT	L
12	CTT	L	TTG	L	12	TTG	L	CTT	L
13	TTT	F	TTC	F	13	TTC	F	TTT	F
14	TTT	F	TTT	F	14	TTT	F	TTT	F

Table 5. The frameshift substitution score of the natural and alternative genetic codes.

Number of alternative	*The natural genetic code*		*FSS of the alternative genetic codes*				
genetic codes Sampled	*FSS Score*	*Rank*	*MAX*	*MIN*	*Average A**	*Average B***	*Average*
1,000,000	-294	62007	-43	-814	-256.842	-438.930	-427.375

* Average A: the average FSS of the genetic codes ranks above (better than) the natural genetic code;

** Average B: the average FSS of the genetic codes ranks below (worse than) the natural genetic code;

Table 6. The usage of codons and their weighed average FSSs (Gon250)

NO	***Species*** ***(Codon Usage)***	***Weighted Average FSS***
1	*H. sapiens*	-9.82
2	*M. musculus*	-13.47
3	*X. tropicalis*	-12.75
4	*D. rerio*	-20.58
5	*D. melanogaster*	-19.43
6	*C. elegans*	-23.38
7	*A. thaliana*	-22.52
8	*S. cerevisiae*	-14.08
9	*E.coli*	-28.59
10	*Equal usage*	-22.27

1 Table 7. The usage of codon pairs and their weighed average FSSs (Gon250)

NO	***Species (Codon Usage)***	***Average FSS of over-represented Codon pairs***	***Average FSS of under-represented Codon pairs***	***Weighted Average FSS of All Codon pairs***
1	*H. sapiens*	41.30	-25.94	102.41
2	*M. musculus*	41.09	-26.09	98.55
3	*X. tropicalis*	42.20	-25.81	98.24
4	*D. rerio*	40.91	-26.17	87.38
5	*D. melanogaster*	39.77	-25.95	79.51
6	*C. elegans*	40.85	-26.18	81.48
7	*A. thaliana*	40.54	-26.09	90.64
8	*S. cerevisiae*	40.85	-26.18	99.21
9	*E.coli*	39.27	-30.75	77.03
10	*Equal Usage*	N/A	N/A	-28.50

2

3

References

1. Nirenberg, M.W. and J.H. Matthaei, *The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides.* Proc Natl Acad Sci U S A, 1961. **47**: p. 1588-602.

2. Haig, D. and L.D. Hurst, *A quantitative measure of error minimization in the genetic code.* J Mol Evol, 1991. **33**(5): p. 412-7.

3. Freeland, S.J. and L.D. Hurst, *The genetic code is one in a million.* Journal of Molecular Evolution, 1998. **47**(3): p. 238-248.

4. Guilloux, A. and J.L. Jestin, *The genetic code and its optimization for kinetic energy conservation in polypeptide chains.* Biosystems, 2012. **109**(2): p. 141-4.

5. Itzkovitz, S. and U. Alon, *The genetic code is nearly optimal for allowing additional information within protein-coding sequences.* Genome Research, 2007. **17**(4): p. 405-412.

6. Seligmann, H. and D.D. Pollock, *The ambush hypothesis: hidden stop codons prevent off-frame gene reading.* DNA Cell Biol, 2004. **23**(10): p. 701-5.

7. Tse, H., et al., *Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes.* BMC Genomics, 2010. **11**: p. 491.

8. Raes, J. and Y. Van de Peer, *Functional divergence of proteins through frameshift mutations.* Trends Genet, 2005. **21**(8): p. 428-31.

9. Xu, J., R.W. Hendrix, and R.L. Duda, *Conserved translational frameshift in dsDNA bacteriophage tail assembly genes.* Molecular Cell, 2004. **16**(1): p. 11-21.

10. Streisinger, G., et al., *Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday.* Cold Spring Harb Symp Quant Biol, 1966. **31**: p. 77-84.

11. Pai, H.V., et al., *A frameshift mutation and alternate splicing in human brain generate a functional form of the pseudogene cytochrome P4502D7 that demethylates codeine to morphine.* Journal of Biological Chemistry, 2004. **279**(26): p. 27383-27389.

12. Baykal, U., A.L. Moyne, and S. Tuzun, *A frameshift in the coding region of a novel tomato class I basic chitinase gene makes it a pseudogene with a functional wound-responsive promoter.* Gene, 2006. **376**(1): p. 37-46.

13. Fox, T.D., *Five TGA "stop" codons occur within the translated sequence of the yeast mitochondrial gene for cytochrome c oxidase subunit II.* Proc Natl Acad Sci U S A, 1979. **76**(12): p. 6534-8.

14. Arenas, M. and D. Posada, *Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography.* BMC Bioinformatics, 2007. **8**: p. 458.

15. Abecasis, A.B., A.M. Vandamme, and P. Lemey, *Sequence Alignment in HIV Computational Analysis*, in *HIV Sequence Compendium*, T. Thomas, et al., Editors. 2007, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory,: Los Alamos, NM. LA-UR 07-4826. p. 2-16.

16. Nagano, T., Y. Kikuchi, and Y. Kamio, *High expression of the second lysine decarboxylase gene, ldc, in Escherichia coli WC196 due to the recognition of the stop codon (TAG), at a position which corresponds to the 33th amino acid residue of sigma(38), as a*

serine residue by the amber suppressor, supD. Bioscience Biotechnology and Biochemistry, 2000. **64**(9): p. 2012-2017.

17. Kuriki, Y., *Temperature-Sensitive Amber Suppression of Ompf'-'Lacz Fused Gene-Expression in a Supe Mutant of Escherichia-Coli K12.* Fems Microbiology Letters, 1993. **107**(1): p. 71-76.

18. Johnston, H.M. and J.R. Roth, *UGA suppressor that maps within a cluster of ribosomal protein genes.* J Bacteriol, 1980. **144**(1): p. 300-5.

19. Prather, N.E., B.H. Mims, and E.J. Murgola, *supG and supL in Escherichia coli code for mutant lysine tRNAs+.* Nucleic Acids Res, 1983. **11**(23): p. 8283-6.

20. Chan, T.S. and A. Garen, *Amino acid substitutions resulting from suppression of nonsense mutations. V. Tryptophan insertion by the Su9 gene, a suppressor of the UGA nonsense triplet.* J Mol Biol, 1970. **49**(1): p. 231-4.

21. Seligmann, H., *Undetected antisense tRNAs in mitochondrial genomes?* Biol Direct, 2010. **5**: p. 39.

22. Seligmann, H., *Avoidance of antisense, antiterminator tRNA anticodons in vertebrate mitochondria.* Biosystems, 2010. **101**(1): p. 42-50.

23. Seligmann, H., *Pathogenic mutations in antisense mitochondrial tRNAs.* J Theor Biol, 2011. **269**(1): p. 287-96.

24. Seligmann, H., *Overlapping genetic codes for overlapping frameshifted genes in Testudines, and Lepidochelys olivacea as special case.* Computational Biology and Chemistry, 2012. **41**: p. 18-34.

25. Seligmann, H., *An overlapping genetic code for frameshifted overlapping genes in Drosophila mitochondria: Antisense antitermination tRNAs UAR insert serine.* Journal of Theoretical Biology, 2012. **298**: p. 51-76.

26. Seligmann, H., *Two genetic codes, one genome: Frameshifted primate mitochondrial genes code for additional proteins in presence of antisense antitermination tRNAs.* Biosystems, 2011. **105**(3): p. 271-285.

27. Faure, E., et al., *Probable presence of an ubiquitous cryptic mitochondrial gene on the antisense strand of the cytochrome oxidase I gene.* Biol Direct, 2011. **6**: p. 56.

28. Gutman, G.A. and G.W. Hatfield, *Nonrandom utilization of codon pairs in Escherichia coli.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(10): p. 3699-3703.

29. Holmes, E.C., *On the origin and evolution of the human immunodeficiency virus (HIV).* Biol Rev Camb Philos Soc, 2001. **76**(2): p. 239-54.

30. Rambaut, A., et al., *Human immunodeficiency virus. Phylogeny and the origin of HIV-1.* Nature, 2001. **410**(6832): p. 1047-8.

31. Paraskevis, D., et al., *Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using bayesian inference: implications for the origin of HIV-1.* Mol Biol Evol, 2003. **20**(12): p. 1986-96.

32. Findley, G.L., A.M. Findley, and S.P. McGlynn, *Symmetry characteristics of the genetic code.* Proc Natl Acad Sci U S A, 1982. **79**(22): p. 7061-5.

33. Frappat, L., P. Sorba, and A. Sciarrino, *Symmetry and codon usage correlations in the genetic code.* Physics Letters A, 1999. **259**(5): p. 339-348.

34. Koch, A.J. and J. Lehmann, *About a symmetry of the genetic code.* Journal of Theoretical Biology, 1997. **189**(2): p. 171-174.

35. Lenstra, R., *Evolution of the genetic code through progressive symmetry breaking.* J Theor Biol, 2014. **347**: p. 95-108.

36. Hornos, J.E.M., Y.M.M. Hornos, and M. Forger, *Symmetry and symmetry breaking: An algebraic approach to the genetic code.* International Journal of Modern Physics B, 1999. **13**(23): p. 2795-2885.

37. Antoneli, F. and M. Forger, *Symmetry breaking in the genetic code: Finite groups.* Mathematical and Computer Modelling, 2011. **53**(7-8): p. 1469-1488.

38. Das, G. and R.H.D. Lyngdoh, *Configuration of wobble base pairs having pyrimidines as anticodon wobble bases: significance for codon degeneracy.* Journal of Biomolecular Structure & Dynamics, 2014. **32**(9): p. 1500-1520.

39. Bizinoto, M.C., et al., *Codon pairs of the HIV-1 vif gene correlate with CD4+T cell count.* Bmc Infectious Diseases, 2013. **13**.

40. Wu, X.M., et al., *Computational identification of rare codons of Escherichia coli based on codon pairs preference.* Bmc Bioinformatics, 2010. **11**.

41. Tats, A., T. Tenson, and M. Remm, *Preferred and avoided codon pairs in three domains of life.* Bmc Genomics, 2008. **9**.

42. Boycheva, S., G. Chkodrov, and I. Ivanov, *Codon pairs in the genome of Escherichia coli.* Bioinformatics, 2003. **19**(8): p. 987-998.

43. Boycheva, S.S. and I.G. Ivanov, *Missing codon pairs in the genome of Escherichia coli.* Biotechnology & Biotechnological Equipment, 2002. **16**(1): p. 142-144.

44. Willie, E. and J. Majewski, *Evidence for codon bias selection at the pre-mRNA level in eukaryotes.* Trends Genet, 2004. **20**(11): p. 534-8.

45. Coghlan, A. and K.H. Wolfe, *Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae.* Yeast, 2000. **16**(12): p. 1131-45.

46. Goetz, R.M. and A. Fuglsang, *Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli.* Biochem Biophys Res Commun, 2005. **327**(1): p. 4-7.

47. Roymondal, U., S. Das, and S. Sahoo, *Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome.* DNA Res, 2009. **16**(1): p. 13-30.

48. Herbeck, J.T., D.P. Wall, and J.J. Wernegreen, *Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia.* Microbiology, 2003. **149**(Pt 9): p. 2585-96.

49. Li, H. and L. Luo, *The relation between codon usage, base correlation and gene expression level in Escherichia coli and yeast.* J Theor Biol, 1996. **181**(2): p. 111-24.

50. Shen, X., S. Chen, and G. Li, *Role for gene sequence, codon bias and mRNA folding energy in modulating structural symmetry of proteins.* Conf Proc IEEE Eng Med Biol Soc, 2013. **2013**: p. 596-9.

51. Pop, C., et al., *Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation.* Mol Syst Biol, 2014. **10**: p. 770.

52. Martinez-Perez, F., et al., *Influence of codon usage bias on FGLamide-allatostatin mRNA secondary structure.* Peptides, 2011. **32**(3): p. 509-17.

53. Carlini, D.B., Y. Chen, and W. Stephan, *The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr.* Genetics, 2001. **159**(2): p. 623-33.

54. Subramanian, A. and R.R. Sarkar, *Comparison of codon usage bias across Leishmania and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions.* Genomics, 2015. **106**(4): p. 232-41.

55. Griswold, K.E., et al., *Effects of codon usage versus putative 5'-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm.* Protein Expr Purif, 2003. **27**(1): p. 134-42.

56. Gambari, R., C. Nastruzzi, and R. Barbieri, *Codon usage and secondary structure of the rabbit alpha-globin mRNA: a hypothesis.* Biomed Biochim Acta, 1990. **49**(2-3): p. S88-93.

57. Zama, M., *Codon usage and secondary structure of mRNA.* Nucleic Acids Symp Ser, 1990(22): p. 93-4.

58. Klumpp, S., J. Dong, and T. Hwa, *On ribosome load, codon bias and protein abundance.* PLoS One, 2012. **7**(11): p. e48542.

59. Zhou, J.H., et al., *The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus.* Infect Genet Evol, 2013. **16**: p. 270-4.

60. McHardy, A.C., et al., *Comparing expression level-dependent features in codon usage with protein abundance: an analysis of 'predictive proteomics'.* Proteomics, 2004. **4**(1): p. 46-58.

61. Mukhopadhyay, P., S. Basak, and T.C. Ghosh, *Synonymous codon usage in different protein secondary structural classes of human genes: implication for increased non-randomness of GC3 rich genes towards protein stability.* J Biosci, 2007. **32**(5): p. 947-63.

62. Stenoien, H.K. and W. Stephan, *Global mRNA stability is not associated with levels of gene expression in Drosophila melanogaster but shows a negative correlation with codon bias.* J Mol Evol, 2005. **61**(3): p. 306-14.

63. Mishima, Y. and Y. Tomari, *Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish.* Mol Cell, 2016. **61**(6): p. 874-85.

64. Trifonov, E.N., *Evolution of the Genetic Code and the Earliest Proteins.* Origins of Life and Evolution of Biospheres, 2009. **39**(3-4): p. 184-184.

65. Koonin, E.V. and A.S. Novozhilov, *Origin and Evolution of the Genetic Code: The Universal Enigma.* Iubmb Life, 2009. **61**(2): p. 99-111.

66. Archetti, M. and M. Di Giulio, *The evolution of the genetic code took place in an anaerobic environment.* Journal of Theoretical Biology, 2007. **245**(1): p. 169-174.

67. Wiltschi, B. and N. Budisa, *Natural history and experimental evolution of the genetic code.* Applied Microbiology and Biotechnology, 2007. **74**(4): p. 739-753.

68. Travers, A., *The evolution of the genetic code revisited.* Origins of Life and Evolution of the Biosphere, 2006. **36**(5-6): p. 549-555.

69. Knight, R.D. and L.F. Landweber, *The early evolution of the genetic code.* Cell, 2000. **101**(6): p. 569-572.

70. Jimenez-Montano, M.A., *Protein evolution drives the evolution of the genetic code and vice versa.* Biosystems, 1999. **54**(1-2): p. 47-64.

71. Davis, B.K., *Evolution of the genetic code.* Progress in Biophysics & Molecular Biology, 1999. **72**(2): p. 157-243.

72. JimenezSanchez, A., *On the origin and evolution of the genetic code.* Journal of Molecular Evolution, 1995. **41**(6): p. 712-716.

73. Beland, P. and T.F.H. Allen, *The Origin and Evolution of the Genetic-Code.* Journal of Theoretical Biology, 1994. **170**(4): p. 359-365.

74. Baumann, U. and J. Oro, *3 Stages in the Evolution of the Genetic-Code.* Biosystems, 1993. **29**(2-3): p. 133-141.

75. Osawa, S., et al., *Recent-Evidence for Evolution of the Genetic-Code.* Microbiological Reviews, 1992. **56**(1): p. 229-264.

76. Russell, R.D. and A.T. Beckenbach, *Recoding of Translation in Turtle Mitochondrial Genomes: Programmed Frameshift Mutations and Evidence of a Modified Genetic Code.* Journal of Molecular Evolution, 2008. **67**(6): p. 682-695.

77. Masuda, I., M. Matsuzaki, and K. Kita, *Extensive frameshift at all AGG and CCC codons in the mitochondrial cytochrome c oxidase subunit 1 gene of Perkinsus marinus (Alveolata; Dinoflagellata).* Nucleic Acids Research, 2010. **38**(18): p. 6186-6194.

78. Haen, K.M., W. Pett, and D.V. Lavrov, *Eight new mtDNA sequences of glass sponges reveal an extensive usage of+1 frameshifting in mitochondrial translation.* Gene, 2014. **535**(2): p. 336-344.

79. Seligmann, H., *Overlapping genetic codes for overlapping frameshifted genes in Testudines, and Lepidochelys olivacea as special case.* Comput Biol Chem, 2012. **41**: p. 18-34.

80. Seligmann, H., *An overlapping genetic code for frameshifted overlapping genes in Drosophila mitochondria: antisense antitermination tRNAs UAR insert serine.* J Theor Biol, 2012. **298**: p. 51-76.

81. Wenthzel, A.M., M. Stancek, and L.A. Isaksson, *Growth phase dependent stop codon readthrough and shift of translation reading frame in Escherichia coli.* FEBS Lett, 1998. **421**(3): p. 237-42.

82. Namy, O., et al., *Identification of stop codon readthrough genes in Saccharomyces cerevisiae.* Nucleic Acids Research, 2003. **31**(9): p. 2289-2296.

83. Loughran, G., et al., *Evidence of efficient stop codon readthrough in four mammalian genes.* Nucleic Acids Research, 2014. **42**(14): p. 8928-8938.

84. Stiebler, A.C., et al., *Ribosomal Readthrough at a Short UGA Stop Codon Context Triggers Dual Localization of Metabolic Enzymes in Fungi and Animals.* Plos Genetics, 2014. **10**(10).

85. Jungreis, I., et al., *Evidence of abundant stop codon readthrough in Drosophila and other metazoa.* Genome Research, 2011. **21**(12): p. 2096-2113.

86. Howard, M.T., et al., *Readthrough of dystrophin stop codon mutations induced by aminoglycosides.* Annals of Neurology, 2004. **55**(3): p. 422-426.

87. Dunn, J.G., et al., *Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.* Elife, 2013. **2**.

88. Steneberg, P. and C. Samakovlis, *A novel stop codon readthrough mechanism produces functional Headcase protein in Drosophila trachea.* Embo Reports, 2001. **2**(7): p. 593-597.

89. Williams, I., et al., *Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae.* Nucleic Acids Research, 2004. **32**(22): p. 6605-6616.

90. Chen, J., et al., *Dynamic pathways of-1 translational frameshifting.* Nature, 2014. **512**(7514): p. 328-+.

91. Dinman, J.D., *Mechanisms and implications of programmed translational frameshifting.* Wiley Interdisciplinary Reviews-Rna, 2012. **3**(5): p. 661-673.

92. Smekalova, Z. and T. Ruml, *Programmed translational frameshifting - Translation of alternative products.* Chemicke Listy, 2006. **100**(12): p. 1068-1074.

93. Farabaugh, P.J., *Programmed translational frameshifting.* Microbiological Reviews, 1996. **60**(1): p. 103-&.

94. Wang, X.W., X.; Chen, G.; Zhang, J.; Liu, Y.; Yang C. , *The shiftability of protein coding genes: the genetic code was optimized for frameshift tolerating.* PeerJ PrePrints 2015. **3** p. e806v1.

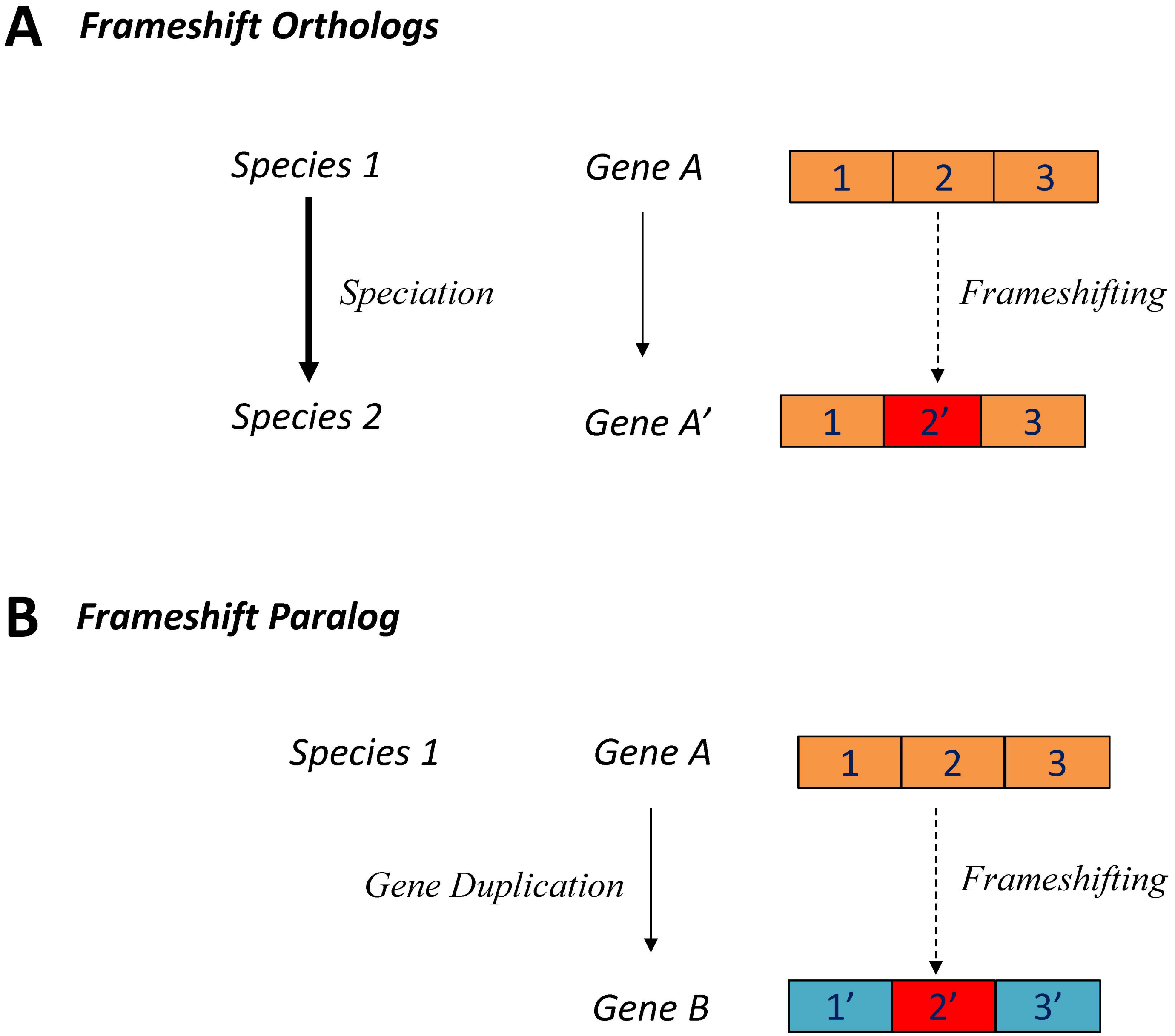
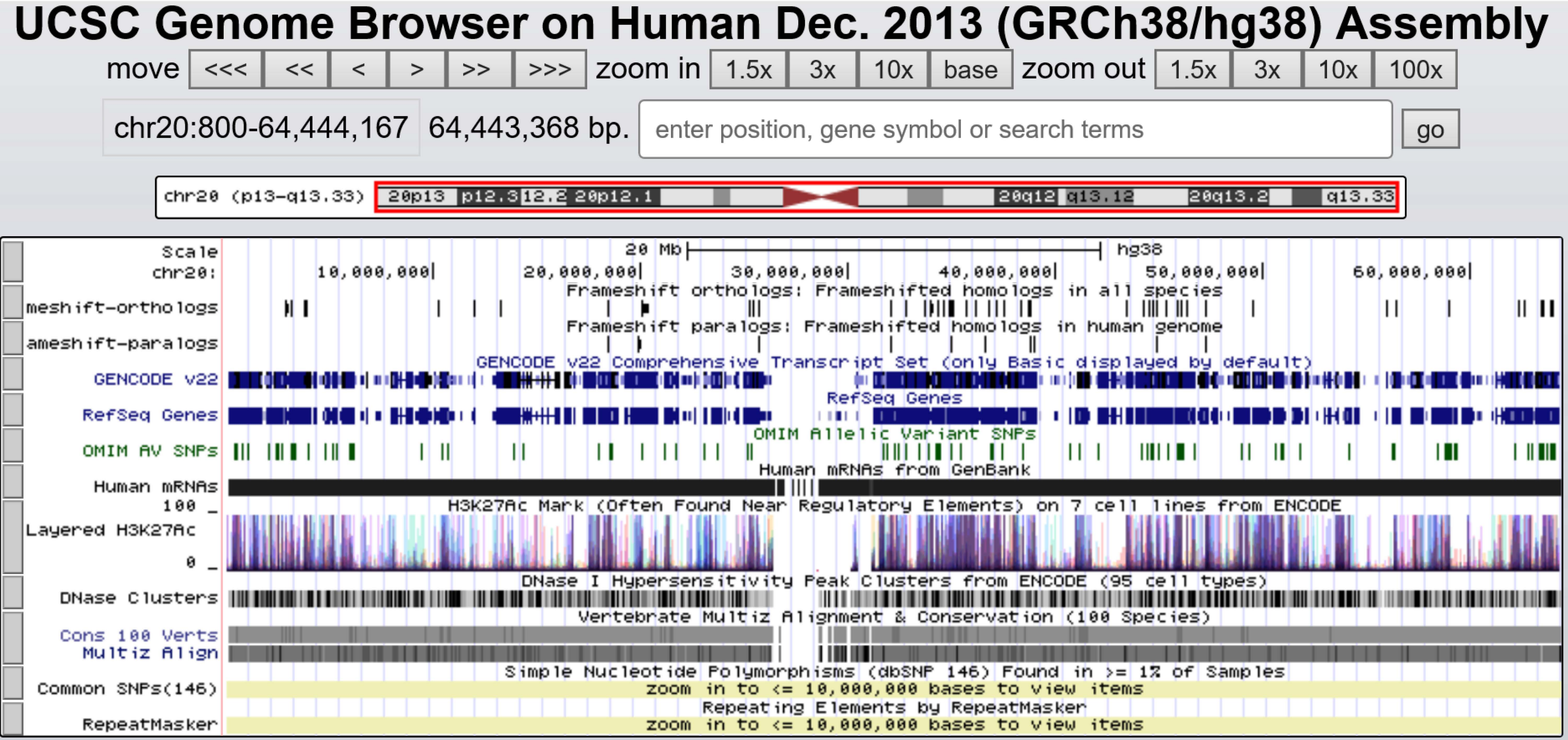

A
Frameshift Orthologs
Species 1
Gene A
1
2
3
Speciation
Frameshifting
Species 2
Gene A’
1
2’
3
B
Frameshift Paralog
Species 1
Gene A
1
2
3
Gene Duplication
Frameshifting
Gene B
1’
2’
3’

Fig 1

C



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr20:800-64,444,167 64,443,368 bp.

move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x 100x
enter position, gene symbol or search terms
go
chr20 (p13-q13.33)
20p13
p12.3
12.2
20p12.1
20q12
q13.12
20q13.2
q13.33
Scale
chr20:
20 Mb
hg38
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
meshift-orthologs
Frameshift orthologs: Frameshifted homologs in all species
ameshift-paralogs
Frameshift paralogs: Frameshifted homologs in human genome
GENCODE v22
GENCODE v22 Comprehensive Transcript Set (only Basic displayed by default)
RefSeq Genes
OMIM AV SNPs
OMIM Allelic Variant SNPs
Human mRNAs
Human mRNAs from GenBank
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
Cons 100 Verts
Multiz Align
Vertebrate Multiz Alignment & Conservation (100 Species)
Common SNPs(146)
Simple Nucleotide Polymorphisms (dbSNP 146) Found in >= 1% of Samples
zoom in to <= 10,000,000 bases to view items
RepeatMasker
Repeating Elements by RepeatMasker
zoom in to <= 10,000,000 bases to view items

Fig 1

```
                 *        20         *        40
HV1J3 : -----------ATGAGAGTGAAGGGGATCAGGAAGAA--TTA :    29
SIVCZ : -----------ATGAAAGTAATGGAGAAGAAGAAGAG--AGA :    29
SIVGB : ATGTCTACAGGAAACGTGTACCAGGAACTAATAAGAAGATAC :    42

               *        60         *        80
HV1J3 : TCAGCACTTGTGGAGATGGGGCACGATGCTCCTTGGGATATT :    71
SIVCZ : CTGGAACAGCTTATCCATAATTACAATCATAACAATCATTTT :    71
SIVGB : CTGGTAGTGGTGAAGAAGCTATACGAAGGTAAGTATGAAGTG :    84

             *       100         *       120
HV1J3 : GATGATCTGTAGTGCTGCAGAACAATTGTGGGTCACAGTC-- :   111
SIVCZ : GCTAACCCCATGTTTGACCTCTGAGTTATGGGTAACAGTA-- :   111
SIVGB : TCCAGGTCTTTTTCTTATACTATGTTTA-GCCTACTAGTAGG :   125

           *       140         *       160
HV1J3 : TATTATGGGGTACCTGTGTGGAAAGAAGCAGCCACCACTCTA :   153
SIVCZ : TATTATGGAGTACCTGTTTGGCATGATGCTGACCCGGTACTC :   153
SIVGB : TATTATAGGAAAACAATATGTGACAGT-CTTCTATGGAGTAC :   166

         *       180         *       200         *
HV1J3 : TTTTGTGCATCAGATGCTAAAGCATAT---------GATACA :   186
SIVCZ : TTTTGTGCCTCAGACGCTAAGGCACAT---------AGTACA :   186
SIVGB : CAGTATGGAA-GGAAGCTAAAACACATTTGATTTGTGCTACA :   207

               220         *       240         *
HV1J3 : GAGGTACATAATGTTTGGGCCACACATGCCTGTGTACCCACA :   228
SIVCZ : GAGGCTCATAATATTTGGGCCACACAGGCATGTGTACCTACA :   228
SIVGB : GATAATTCAAGTCTCTGGGTAACCACTAATTGCATACCTTCA :   249

             260         *       280         *
HV1J3 : GACCCCAACCCACAAGAAGTAGTATTGGAAAATGTGACAGAA :   270
SIVCZ : GATCCCAGTCCTCAGGAAGTATTTCTTCCAAATGTAATAGAA :   270
SIVGB : TTGCCAGATTATGATGAGGTAGAAATTCCTGATATAAAGGAA :   291

           300         *       320         *
HV1J3 : AAATTTAA------CATGTGGAAAAATAACATGGTAGAACAG :   306
SIVCZ : TCATTTAA------CATGTGGAAAAATAATATGGTGGACCAA :   306
SIVGB : AATTTTACAGGACTTATAAGGGAAAATCAGATAGTTTATCAA :   333
```

Fig 2 (A). Alignment of coding sequences of HIV/SIV GP120

```
                 *        20         *        40
HV1J3 : --------------MRVKGIRKNYQHLWRWGTMLLGILMICSA :  29
SIVCZ : --------------MKVMEKKKRDWNSLSIITIITIILLTPCL :  29
SIVGB : MSTGNVYQELIRRYLVVVKKLYEGKYEVSRSFSYTMFSLLVGI :  43
                      6 V  k k      s  t  t il6

              *        60         *        80
HV1J3 : AEQLWVTVYYGVPVWKEAATTLFCASDAKAYDTEVHNVWATHA :  72
SIVCZ : TSELWVTVYYGVPVWHDADPVLFCASDAKAHSTEAHNIWATQA :  72
SIVGB : IGKQYVTVFYGVPVWKEAKTHLICATDNSS-------LWVTTN :  79
           l5VTV5YGVPVWkeA t LfCA3Daka  te hn6WaT a

           *       100         *       120
HV1J3 : CVPTDPNPQEVVLENVTEKFN--MWKNNMVEQMHEDIISLWDQ : 113
SIVCZ : CVPTDPSPQEVFLPNVIESFN--MWKNNMVDQMHEDIISLWDQ : 113
SIVGB : CIPSLPDYDEVEIPDIKENFTGLIRENQIVYQAWHAMGSMLDT : 122
        C6P3dP pqEV 6p16 E Fn  6wkNn6V Qmhed6iS6wDq

        *       140         *       160         *
HV1J3 : SLKPCVKLTPLCVTLNCIDWGNDTSPNATNTTSSGGEKMEKGE : 156
SIVCZ : SLKPCVKLTPLCVTLQCSKANFSQAKNLTNQTSS-----PPLE : 151
SIVGB : ILKPCVKINPYCVKMQCQETENVSATTAKPITTPTTTSTVASS : 165
        sLKPCVK6tPlCVt6qC    n  a natn T3s        e

             180         *       200         *
HV1J3 : MKNCSFNITTSIRDKVQKEHALFY------KHDVVPINNSTKD : 193
SIVCZ : MKNCSFNVTTELRDKKKQVYSLFY------VEDVVNLG----- : 183
SIVGB : TEIYLDVDKNNTEEKVERNHVCRYNITGLCRDSKEEIVTNFRG : 208
        mkncsfn tt  rdKv   h lfY        dvv 6

          220         *       240         *       2
HV1J3 : NIKNDNSTRYRLISCNTSVITQACPKISFEPIPIHYCAPAGFA : 236
SIVCZ : ---NENNT-YRIINCNTTAITQACPKTSFEPIPIHYCAPAGFA : 222
SIVGB : DDVKCENNTCYMNHCNESVNTEDCQKG-LLIRCILGCVPPGYV : 250
           n nnt yr6i CNt3viT2aCpK sfepipIhyCaPaG5a

        60         *       280         *       300
HV1J3 : IIKCNDKKFNGTGPCTNVSTVQCTHGIKPVVSTQLLLNGSLAE : 279
SIVCZ : ILKCNDKDFSGKGKCTNVSTVHCTHGIKPVVTTQLLINGSLAE : 265
SIVGB : MLRYN-EKLNNNKLCSNISAVQCTQHLVATVSSFFGFNGTMHK : 292
        664cNdkkfng g C3N6StVqCThg6kpvV33qll NG36ae

                *       320         *       340
HV1J3 : EEVVIRSENFTDNAK-------TIIVQLKEPVVINCTRPSKTT : 315
SIVCZ : GNITVRVENKSKNTD-------VWIVQLVEAVSLNCHRPGNNT : 301
SIVGB : EGELIPIDDKYRGPEEFHQRKFVYKVPGKYGLKIECHRKGNRS : 335
        e   6r e1k  n         v iVqlke 6 6nChRpgn 3
```

Fig 2 (B). Alignment of protein sequences of HIV/SIV GP120