Pairs of compensatory frameshifting mutations contribute to evolution of protein-coding sequences in vertebrates and insects

Dmitry Biba[1], Galya Klink[2], Georgii Bazykin[1,2]

1 - Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Moscow Oblast, Russia, 121205

2 - Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevitch Institute), Moscow, Russia, 127051

Corresponding author e-mail: g.bazykin@skoltech.ru

Abstract

Insertions and deletions of lengths not divisible by 3 in protein-coding sequences cause frameshifts that usually induce premature stop codons and may carry a high fitness cost. However, this cost can be circumvented by a second compensatory indel restoring the reading frame. The role of such compensatory frameshifting mutations (CFMs) in evolution has not been studied systematically. Here, we use whole-genome alignments of protein coding genes of 100 vertebrate species, and of 122 insect species, studying the prevalence of CFMs in their divergence. After stringent filtering, we detect a total of 11 high-confidence genes carrying pairs of CFMs, including three human genes: *RAB36, ARHGAP6* and *NCR3LG1.* CFMs tended to occur in genes under relaxed negative selection, indicating that they are typically prevented at functionally important genes. In some instances, mutations closely predating or following the CFMs restored the biochemical similarity of the frameshifted segment to the ancestral sequence, possibly reducing or negating the fitness cost of a CFM. Typically, however, the resulting sequence bore no similarity to the ancestral one, indicating that the CFMs can uncover radically novel regions of sequence space. In total, CFMs represent a potentially important and previously overlooked source of novel variation in amino acid sequences.

Introduction

The origin of radically novel amino acid sequences remains somewhat of an enigma. By far the most common mode of evolution of protein-coding genomic segments is modification through single-nucleotide substitutions. While gradual accumulation of such substitutions can modify the characteristics, structure and, in some instances, function of the encoded protein, they only allow reaching a limited neighborhood of the existing sequence in the sequence space (1). This leaves the question of the possibility of origin of new segments of protein-coding sequences from scratch. Several mechanisms for such events are known. Some involve acquisition of new DNA segments within protein-coding regions, by repeat expansion (2) or non-repeat-associated insertions with lengths divisible by 3. Others involve redefinition of an existing DNA segment as protein-coding; these include emergence of exonic segments from intronic ones ((3)), sequences encoding N- or C-termini of proteins from 5'- or 3'-UTRs ((4), or even entire genes from non-coding DNA (5); or acquisition of an alternative reading frame for an existing protein-coding sequence (6).

Because a previously non-coding sequence is expected to carry an in-frame stop codon every ~20 codons, radically new protein-coding sequences are seldom long. Nevertheless, the complete novelty of the encoded sequence may result in reaching remote

domains of the sequence space, which may be then fine-tuned in the course of subsequent evolution (7).

Another possible mechanism of origin for a new amino acid-coding sequence is frameshifting insertions and deletions (hereafter, indels). Generally, they are considered to be an unlikely source of novelty (8) because they usually cause major disruptions in the amino acid sequence and its truncation through gain of premature stop-codons, and therefore should be extremely rare to fix due to fitness cost associated with them. Nevertheless, there are some examples of fixation of frameshifting indels and, though the protein function is often lost in these cases, sometimes it is retained (9). Moreover, in some cases such indels give rise to a totally new protein function (10, 11). Still, these mutations tend to fix only if they happen near the 5'- or 3'-end of a gene, which likely lowers their impact on protein structure (12, 13).

Another factor which may prevent a frameshifting indel from disrupting overly long segments of protein-coding sequence is a gain of a second compensatory indel in close proximity to the first one. If the total length of the two indels is divisible by 3, the reading frame will be restored and the only amino acids changed would be those between these indels. In experimental evolution of phages, indels are frequently compensated by frame-restoring indels - a finding which has been instrumental in the discovery of the structure of genetic code (14). However, the role of compensatory pairs of indels in natural evolution has not been studied systematically. Meanwhile, there are some reasons to believe that the changes to protein structure during frameshifts are not as substantial as replacement with a totally random sequence: some key physico-chemical properties of the encoded amino acid sequence tend to be preserved by frameshifts (15, 16). Finally, the intermediate, yet-to-be-compensated state of a protein may not be as deleterious as one could have expected, since ribosomes often bypass premature stop-codons, which are the most detrimental feature of such state (17). With all these considerations taken into account, we suggest that pairs of compensatory frameshifting mutations (hereafter, pairs of CFMs) may be of some importance in the evolution of novel amino acid sequences.

In this paper we systematically study CFMs that have occurred rapidly one after the other in the evolutionary history of vertebrate and insect genes. Using a phylogenetic reconstruction of ancestral states, we describe multiple instances of such events, describe the characteristics of the proteins and protein segments in which they occur, and predict their impact on the properties of the encoded protein sequence.

Materials & Methods

Sequence alignment

As a starting point, we used the 100-way vertebrate and 124-way insect (diptera) MULTIZ alignments obtained from the UCSC genome browser (18). These alignments are comprised of non-reference genomes being aligned to a reference (correspondingly, human or *Drosophila melanogaster).* This means that all the sites where the reference species has a deletion are absent, making these alignments unfit for our purposes. Therefore, we realigned the protein-coding sequences of genes from these species as follows. We

extracted the exonic sequences from genomes using the annotation from the corresponding MULTIZ alignments, concatenated these exons into genes, and aligned the resulting genes using mafft v7 (19) with the following parameters: --maxiterate 1000 --globalpair –preservecase. To allow for detection of frameshifts, we performed a nucleotide-level, rather than a codon-level, alignment. In the insects data set, we removed the D_pseudoobscura_1 and A_gambiae_1 genomes as they were redundant (the alignment already contained the droPse3 and anoGam3 genomes from the same species). Because the gene annotation was based on the genome of the reference species, we excluded the overhanging 5'- and 3'-ends as well as long (>50 nucleotides) internal DNA segments in non-reference species that had no orthologous sequence in the reference. High-quality alignments were obtained for 21208 out of 21521 genes of vertebrates and 27341 out of 30482 genes in insects (insect genes were comprised of all NCBI RefSeq Gene CDS regions including non protein-coding ones; non protein-coding genes were excluded at subsequent filtering steps). The phylogenetic trees for vertebrates and insects obtained from whole-genome alignments, with branch lengths scaled in estimated numbers of nucleotide substitutions per site, were obtained from the UCSC genome browser.

Inference of compensatory frameshifting mutations

We designed an algorithm for detection of CFMs using a multiple alignment and the corresponding phylogenetic tree.

In short, the algorithm uses a parsimony-based approach to detect insertions and deletions in the alignment. As the first step, it reconstructs past insertions and deletions that had occurred at specific phylogenetic branches, and are inherited by subsets of analyzed species. At the second step, to ensure usage of high quality and high confidence protein-coding sequences for CFM inference, the nucleotide alignments for each gene were filtered on a sequence-by-sequence basis. We excluded the species that did not meet all of the following criteria: had the number of nucleotides divisible by 3; carried a start codon at the sequence start; carried a stop codon at the sequence end; carried no in-frame stop codons. At the third step, the pairs of CFMs are identified as pairs of indels of length not divisible by 3 in the remaining sequences; at this step, we also excluded sequences with more than 2 such indels, assuming that our reconstruction of the pairs of CFMs in such sequences would be less trustworthy. We performed quality filtering after indel classification, assuming that the added information from the presumably lower-quality sequences allows for more robust classification of the evolutionary history of indels. Alignments for which fewer than four sequences were left after filtering were not considered further; these included all non-protein-coding genes.

For each detected CFM, the algorithm outputs its type (insertion or deletion), position, lengths, and the list of species it was detected in. A detailed description of the algorithm is provided in the Supplementary text 1. The algorithm is implemented as a Python script, available at https://github.com/Captain-Blackstone/Compensatory-frameshifts.

Post-filtering of compensatory frameshifting mutations

With all the CFMs detected, we then focused on the most reliable ones, narrowing down the considered sample as follows. First, while our algorithm detects pairs of CFMs independently of whether they occurred in the same on in different branches of the phylogenetic tree, in this work, we only consider those that happened on the same branch. Second, we only considered pairs of CFMs comprised of short indels, with combined lengths of no more than 4 nucleotides (i.e., two insertions of lengths 1 and 2; two deletions of lengths 1 and 2; an insertion and a deletion both of lengths 1; or an insertion and a deletion both of lengths 2). We made this decision reasoning that longer indels are more likely to be artefactual. Third, the remaining alignments were manually inspected, and only those with unambiguous CFMs were retained.

For the final analysis, we focused on those cases where the pair of CFMs was inherited by more than one species in our dataset. This way, we could have excluded some bona fide species-specific CFMs, but obtained additional support for the remaining ones, under the logic that coincident artefacts in different independently assembled genomes should be rare. We also excluded those alignments in which either the ancestral or the frameshifted sequence was not supported by running nBLAST (20) against the genome of the corresponding species in the NCBI nr database.

This left us with 10 high-confidence pairs of CFMs. For the vertebrate dataset, we then sought for additional support from other sequencing projects not included in the UCSC genome browser. For this, for each CFM-carrying gene, we used the ENSEMBL (21) to find orthologs in species that were more closely related to the CFM-carrying species than any other species in our alignment. This rescued one additional pair of CFMs for the final dataset, for a total of 11 pairs.

Statistical approaches

For each gene, we used PAML version 4.8 (22) to estimate the probabilities of nonsynonymous changes relative to synonymous ones (ω values) under the substitution model of Goldman and Yang (1994) described in Yang et al., 1998. We used the M0 model (model=0 parameter) to estimate a single ω value for the entire tree, and a model with two ω parameters (model=2 parameter) to fit distinct values of ω for the pre- and post-CFM branches. The remaining parameters were the same in both models: runmode=1, seqtype=1, NSsites=0.

For each gene, we estimated ω using as input only the subset of species that passed the filtering steps described in the “Inference of compensatory frameshifting mutations” section. As a result, the subset of considered species differed between genes.

To ask whether the CFM-carrying genes are characterized by increased or decreased ω values, we compiled a sample of non-CFM-carrying genes as a control. For this, we used the same filtering steps as above, namely, retained for each alignment only those sequences that had the number of nucleotides divisible by 3; carried a start codon at the sequence start; carried a stop codon at the sequence end; and carried no in-frame stop codon. To account for differences in numbers of considered species which could bias ω

estimates, for each CFM-carrying gene, we used as a control a subset of genes that were matched by the number of retained species.

To combine the signal across comparisons for individual genes, we used the Fisher's combined probability test under the null hypothesis of no effect in any of the genes.

Ancestral state reconstructions for nucleotide sites

Ancestral states for individual nucleotide sites and phylogenetic positions of single-nucleotide substitutions were reconstructed independently of the inference of indels (see above) with the help of the maximum likelihood method implemented in MEGA6 (23). For substitutions that had happened on the same phylogenetic branches as the CFM, which one came first could not be established unambiguously. In these cases, we assumed that all substitutions occurred simultaneously; that both CFMs occurred simultaneously; and considered two scenarios, assuming that the substitutions occurred either before or after the CFMs. For these branches, we identified the ancestral state as "A", the intermediate state in the mutation-first scenario as "A_{mut}", the intermediate state in the frameshift-first scenario as "A_{fr}", and the derived state as "E".

Similarity metrics

To study the effects of indels and single-nucleotide substitutions on the encoded amino acid sequence, we compared different states of these sequences using two similarity metrics. First, we measured the difference between the mean hydropathy indexes of the two sequences (24). This difference is close to 0 for very similar sequences and close to 9 for very distant ones. Second, we measured the mean Miyata distance (25) between the aligned amino acids in the pairwise alignment. This metric is also close to 0 for very similar sequences; for very distant sequences, it is close to 5.13, which is the maximum Miyata distance (namely, the distance between glycine and tryptophan). We considered gaps to have the maximum possible distance (5.13) to any amino acid.

Results

Detected CFMs

We have found 536 genes in vertebrates and 182 genes in insects that carried pairs of CFMs. The vast majority of these CFMs were only inherited by a single species each; multiple species inheriting the same CFM were observed only for 15 genes in vertebrates and for 12 genes in insects (Table 1). Because such cases are the most reliable, we focused on them. After post-filtering (see Methods), the final list included only 6 genes in vertebrates and 5 genes in insects (Figure 1, Table 2). Some of them were supported by transcriptomics and/or proteomics data (Table 2, Supplementary Table 2).

A previous work (12) has identified a pair of CFMs between human and dog in the FLJ43860 gene. However, we were unable to confirm their finding, as the corresponding amino acid-coding sequence is not detected by tblastn in the reference human genome, and

this gene is currently annotated as a pseudogene, according to the HGNC (26) gene symbol report.

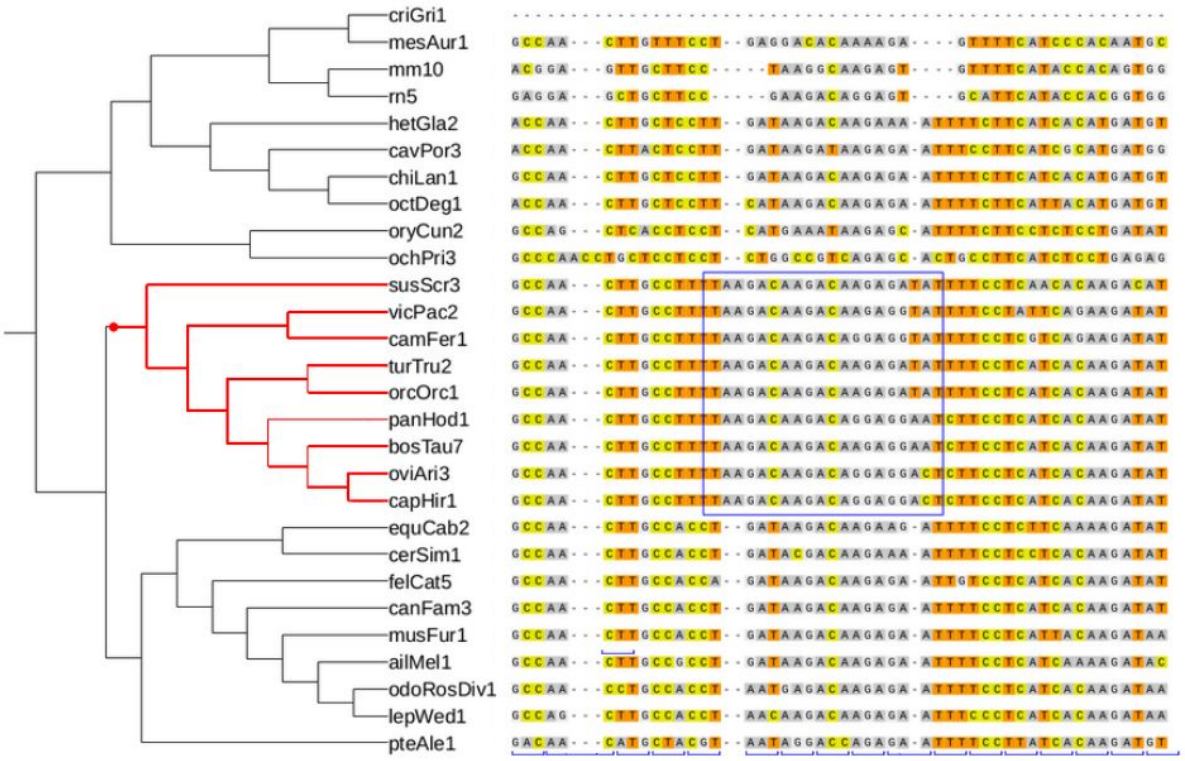

Figure 1. An example of a CFM: a fragment of alignment of the *INSL6* gene, one of the 11 identified genes carrying reliable CFMs. At the left, a cladogram of the corresponding species is shown, with the CFM-carrying clade marked in red, and a subset of the species devoid of the CFMs, in black (some of the non-CFM-carrying species are not shown). The reconstructed phylogenetic position of the CFM is marked with a red dot. The region with the shifted reading frame is marked by the blue rectangle. The original reading frame is indicated at the bottom of the alignment by blue square brackets.

	Vertebrates	Insects
Number of analysed protein-coding genes	21208	15283
Number of genes with CFMs of combined length <= 4	536	182
Number of genes with multiple species carrying CFMs	15	12
Number of genes with multiple species carrying CFMs that passed all the quality filters	5	5

Table 1. Number of genes surviving each of the filtering stages.

	Gene name	Number of species with CFMs*	Number of nucleotides between CFMs	Indel types (5', then 3')
Vertebrates	*RAB36†*	11	99	2-bp deletion, 1-bp deletion
	ARHGAP6†‡	9	7	1-bp deletion, 2-bp deletion
	INSL6+	9	14	2-bp insertion, 1-bp insertion
	NCR3LG1†‡	4	32	1-bp deletion, 1-bp insertion
	RASSF4	3	26	2-bp deletion, 1-bp deletion
	SPATA24¶	1	14	1-bp deletion, 1-bp insertion
Insects	*Diap1*	4	15	1-bp deletion, 2-bp deletion
	CG31530-RA	11	16	1-bp deletion, 2-bp deletion
	wds	2	10	2-bp deletion, 1-bp deletion
	Cyp6a23	4	19	2-bp deletion, 1-bp deletion
	Osi10	4	5	1-bp deletion, 2-bp deletion

Table 2. The 11 CFM-carrying genes retained after filtering. The number of species carrying the pair of CFMs (*) includes the species that were parts of the post-frameshift clades, but that were excluded at the filtering step. For *SPATA24,* a closely related CFM-carrying species (*Cavia aperea*) not present in the original alignment was found in ENSEMBL. See Supplementary Table 1 for complete lists of CFM-carrying species.
** derived state confirmed by mRNA transcripts
† both ancestral and derived state confirmed by mRNA transcripts
‡ derived state confirmed by protein sequence
¶ ancestral state confirmed by protein sequence

Frequency of CFMs is independent of position within a coding sequence

Uncompensated indels with lengths not in multiple of 3 occur more frequently near ends of coding sequence, probably because the deleterious effect of a frameshift is weaker

at these positions. By contrast, pairs of compensatory frameshifting mutations do not cause any disruption of the protein-coding potential of the sequence downstream of them. Therefore, we hypothesized that unlike uncompensated indels, CFMs should be uniformly distributed along the CDS.

To test this hypothesis, we obtained a distribution of frequencies of uncompensated indels in human polymorphism data from (27), and compared it to that of indels involved in pairs of CFMs (Figure 2). While uncompensated indels were biased towards gene ends, no such bias was observed for the CFMs, resulting in a significant difference between these two distributions (Kolmogorov-Smirnov test, p-value = 1.97e-17). No significant deviation from the uniform distribution along the gene length was found for the CFMs (p-value = 0.277).

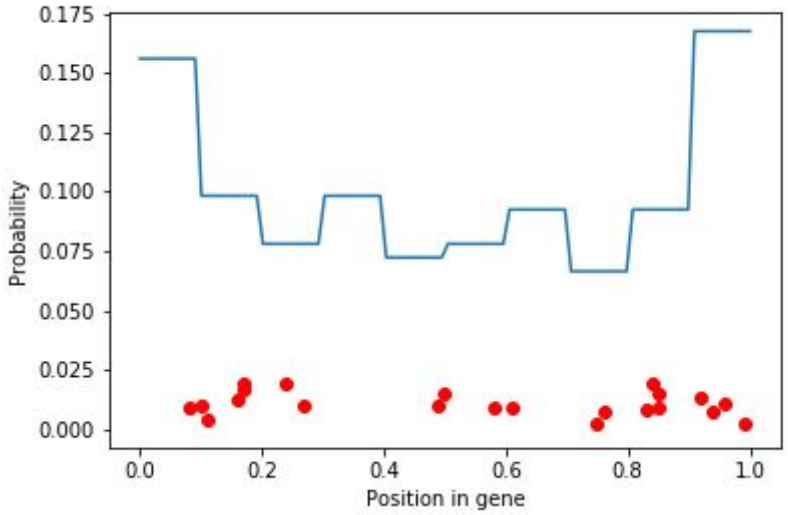

Figure 2. The distribution of uncompensated indels (blue line) and CFMs (red dots) along the gene coordinates. The horizontal axis indicates the position along the length of the coding sequence, with 0.0 corresponding to its start, and 1.0, to its end.

Genes with compensatory frameshifts are under relaxed negative selection

We hypothesized that compensatory frameshifts are more likely to fix in genes that are under relaxed negative selection. To test this, we estimated the ratio of nonsynonymous to synonymous substitution rates ω for the CFM-carrying genes, and compared it to the same value calculated for non-CFM-carrying genes (Figure 3).

The values of ω were above the 95th percentile of the distribution, indicating relaxed negative selection, for two of the 11 analyzed genes, *INSL6* and *Diap1*. They were also somewhat elevated for all genes combined (Fisher's combined probability test, p-value = 0.002, χ^2 = 46.3).

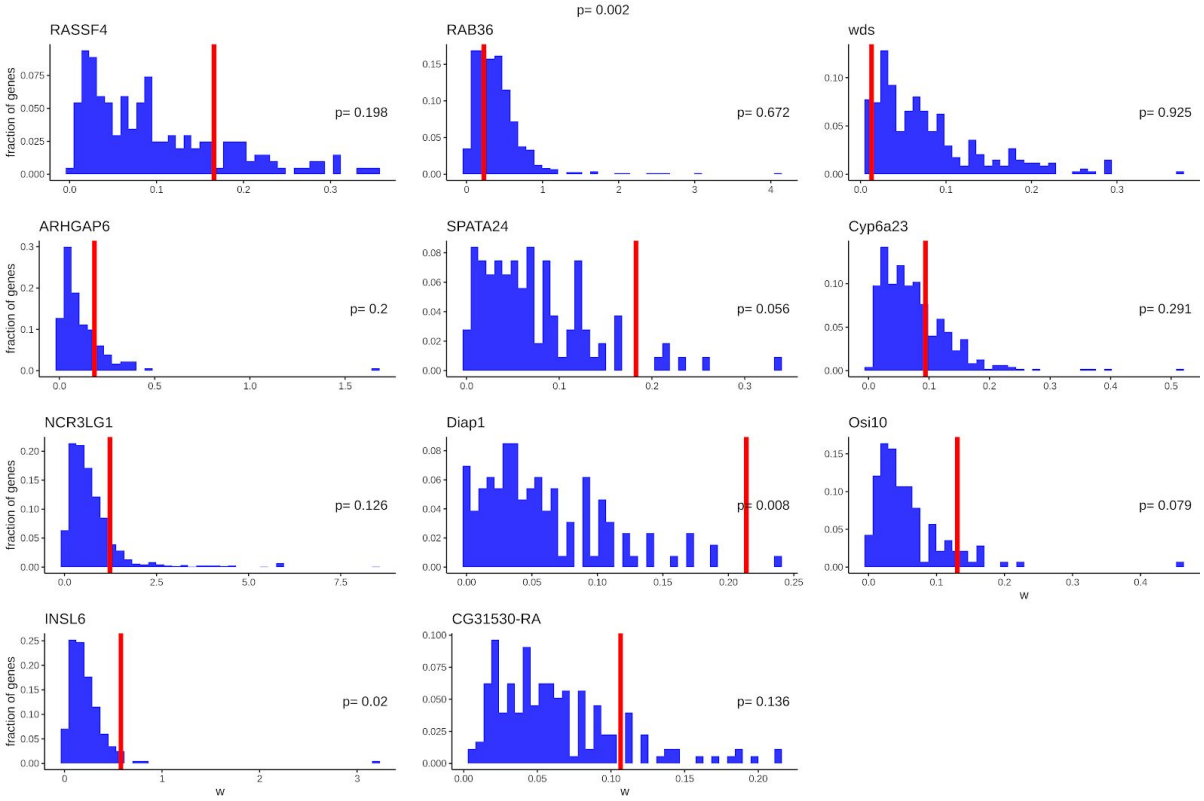

Figure 3. Negative selection at CFM-carrying genes. Each panel corresponds to a gene with a pair of compensatory frameshifts. The red line indicates the observed value of the tree-wide ω of the gene of interest. The blue histograms show the distribution of ω in non-CFM-carrying genes. The p-values are obtained as the proportion of the distribution to the right of the red line. Because ω were inferred using only species which passed all the quality filters, the numbers of species used to calculate ω differed among genes, which could affect ω estimates; therefore, for each panel, only those genes that had the same number of species were used for the reference distribution (see "Statistical approaches" section in Materials and Methods). The p-value of Fisher's combined probability test is shown at the top.

We also hypothesised that the mode of selection changes when the CFMs are fixed. To test this, for each gene, we fitted two ω values: one for the clade of the phylogenetic tree descendant from the CFM-carrying branch, and another for the remaining branches. To test whether this model describes our data better than the model with a single ω, we used the likelihood-ratio test. The two-ω model fitted the data significantly better than the null single-ω model for one of the genes (CG31530-RA in insects); for the remaining genes, no significant difference was found (Table 3).

Protein	ω for the one-parameter model	ω for the non-CFM-carrying branches in the two-parameter model	ω for the CFM-carrying clade in the two-parameter model	LRT statistics value	p-value
RASSF4	0.166	0.163	0.254	3.78	0.05

NCR3LG1	1.23	1.71	0.867	1.76	0.183
RAB36	0.229	0.24	0.211	0.22	0.636
SPATA24	0.183	0.181	0.284	0.734	0.391
INSL6	0.578	0.605	0.438	2.04	0.153
ARHGAP6	0.184	0.184	0.187	0.01	0.912
Diap1	0.214	0.213	0.235	0.207	0.649
CG31530-RA	0.107	0.101	0.155	11.013	0.001*
wds	0.013	0.013	0.015	0.128	0.721
Cyp6a23	0.094	0.094	2.129	0	1
Osi10	0.131	0.128	0.177	3.45	0.063

Table 3. Selection in CFR-carrying genes under different evolutionary models.
*p<0.05 after the Bonferroni correction (p-values themselves are presented without a correction).

Protein regions between the CFMs show no evidence for reduced conservation

To study the positions of CFMs within the protein-coding sequence, we hypothesized that they tend to fix in those gene segments that are generally less conserved. To test this hypothesis, we compared the amino acid-level conservation of the region between the CFMs to that of other CDS segments of the same gene, using a sliding window of the same length as the region between the CFMs to obtain the null distribution (Figure 4). We used Shannon's entropy as a measure of conservation of amino acid site. We see no obvious decrease in conservation of the region between CFMs (p-value of Fisher's combined probability test = 0.253, χ^2 = 25.9).

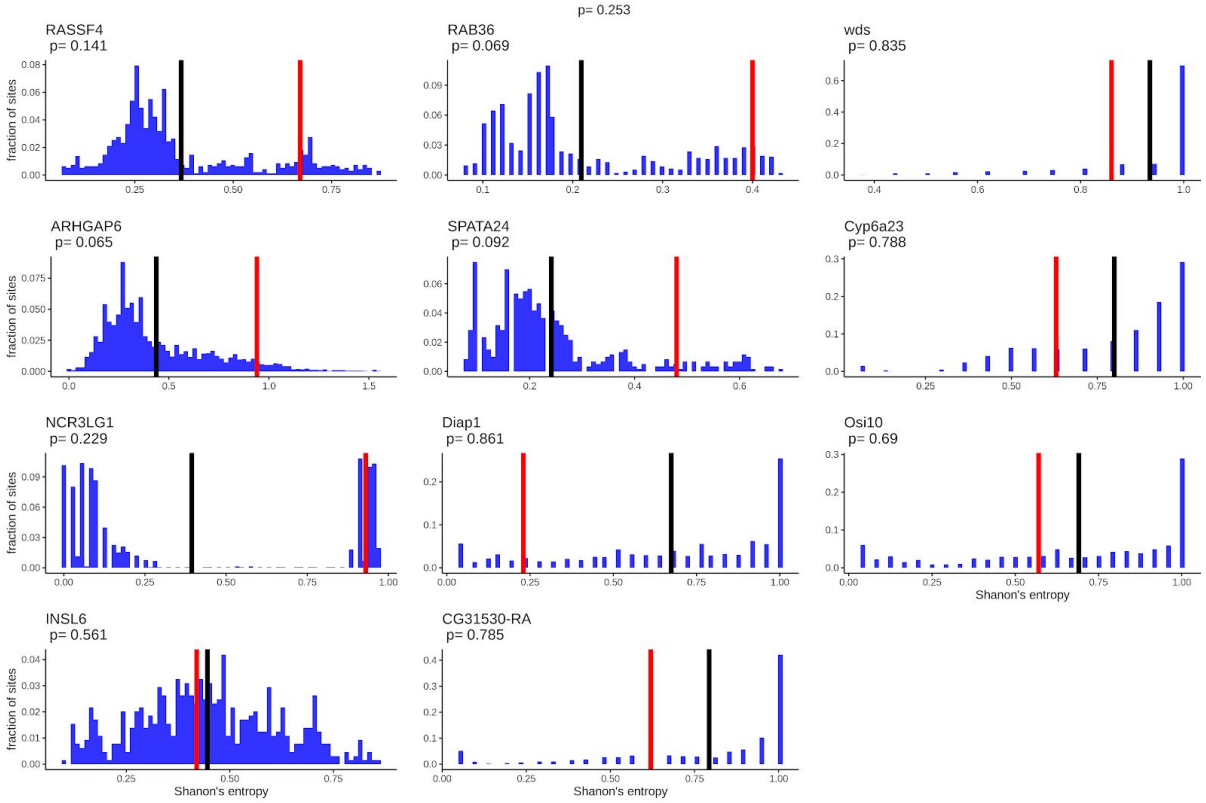

Figure 4. Conservation of the region between the CFMs. Each panel corresponds to a gene with a pair of compensatory frameshifts. The red line indicated the mean Shannon's entropy of the region between the CFMs. The blue histograms show the distribution of entropy at all positions of the same protein, calculated as the mean entropy in a sliding window of length equal to the length of the region between CFMs, with the mean across all sliding windows indicated with the black line. The p-value of Fisher's combined probability test is shown at the top.

The effect of CFMs on the amino acid sequence

Due to the structure of the genetic code and the characteristics of the structured amino acid sequences, frameshifting mutations generally preserve the physico-chemical properties of the encoded protein ((15, 16)). We asked if the pairs of CFMs additionally tend to occur at those positions where this preservation is more precise.

Specifically, we hypothesized that CFMs tend to fix at positions where they don't substantially affect the structure of the encoded protein according to two similarity metrics. To test this hypothesis, for each of the eleven CFM-carrying genes, we calculated the similarity between their amino acid sequences before and after the CFMs (see "Similarity metrics" subsection in Materials & Methods). We used ancestral state reconstructions (see "Ancestral states reconstruction" subsection in Materials & Methods) to infer the sequences in which the CFMs have fixed. We asked if a sequence resulting from a pair of CFMs was unexpectedly similar to its ancestral version prior to the CFMs. To calculate the expected distribution of the similarity values, we introduced *in silico* pairs of CFMs in different genes sampled randomly from the genome, preserving the indel types (insertion/deletion), lengths (1 or 2) and the number of nucleotides between them.

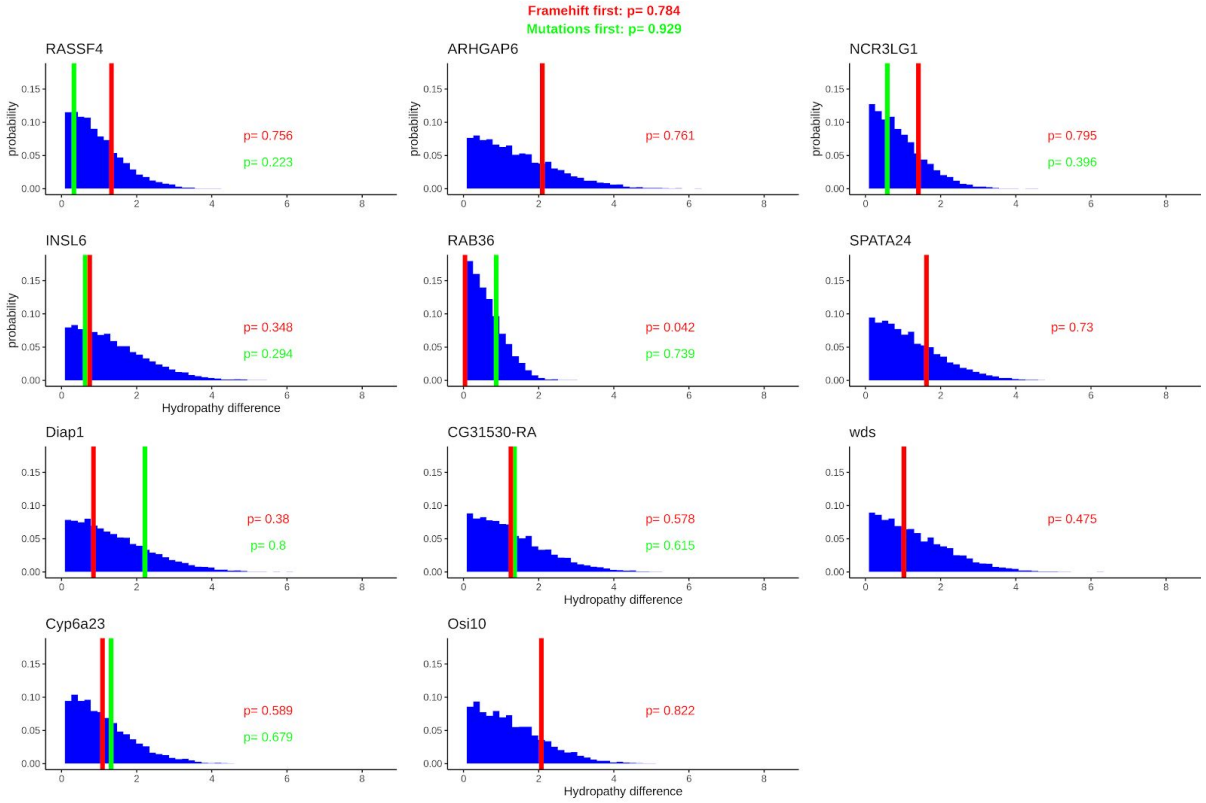

Figure 5. The effect of CFMs on hydropathy of the encoded amino acid sequence. The vertical lines represent the difference in the hydropathy index between the reconstructed ancestral sequence of the gene immediately prior to the CFMs, and the reconstructed (or observed) derived state of this sequence immediately after the CFMs. The red and green lines differ in treatment of single-nucleotide mutations that have happened on the same phylogenetic branch as the CFMs: such mutations are assumed to have occurred either after (red line) or before (green line) the CFM; if no such mutations were present, only the red line is shown. The blue bars represent the distribution of differences in hydropathy index between each of the 10,000 protein-coding sequences randomly drawn from the genome, and a version of this sequence with a pair of CFMs spaced identically to that in the considered gene. The numbers in each panel correspond to the p-values obtained as the percentile of the distribution, with low p-values corresponding to lower-than-expected differences in hydropathy. The p-values of the Fisher's combined probability test are shown at the top of the figure.

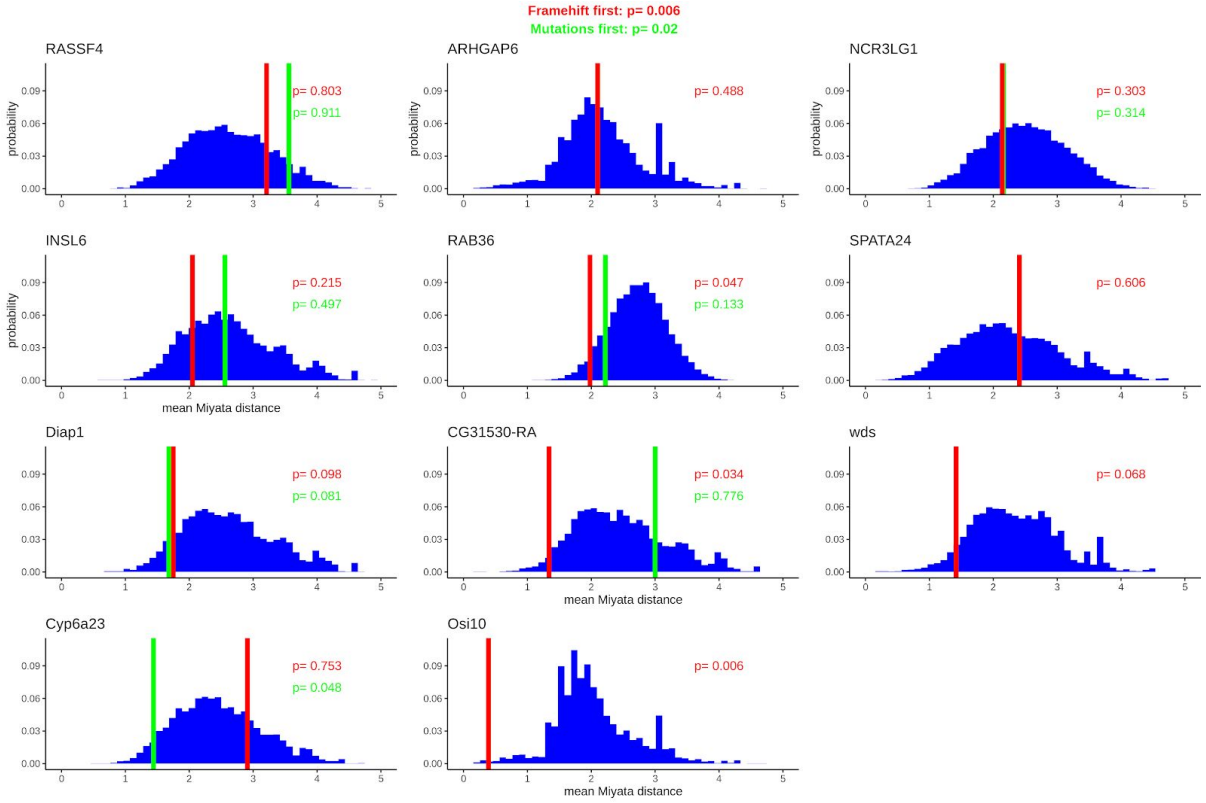

Figure 6. The effect of CFMs on the physico-chemical properties of the encoded amino acid sequence as measured by the mean Miyata distance. The notation is the same as in Figure 5.

We used two measures of sequence similarity between amino acid sequences: differences in hydropathy index, and the Miyata distance. The results are shown in Figures 5 and 6 respectively. For individual genes, the observed effect of the CFMs on the amino acid sequence matched the expectations, with none of the genes having significantly lower-than-expected differences between the ancestral and derived sequences after the Bonferroni correction. Nevertheless, overall the effect of the CFMs across all genes was lower than expected for the Miyata distance (Fisher's combined probability test, p-value = 0.006 (χ^2 = 42) and 0.02 (χ^2 = 37.6)). For the hydropathy difference, no difference was observed either in the single-gene or pooled comparison (p-value = 0.746 (χ^2 = 16.6) and 0.923 (χ^2 = 13.2).

Interaction between CFMs and single nucleotide substitutions

Conceivably, changes in physico-chemical properties of a protein segment caused by a pair of CFMs could be partially compensated by subsequent single-nucleotide substitutions in the region spanned by them, if such substitutions make the protein more similar to the pre-CFM state. Additionally, the CFMs could preferentially occur in regions where preceding amino acid substitutions made their effect less pronounced. To study this, we analyzed the

interactions between the CFMs and mutations that happened on the same phylogenetic branch in their effect on the characteristics of the encoded amino acid sequence.

In Figures 7 and 8, we measure the amino acid similarity between the reconstructed ancestral variant of the protein (A), the variant carrying the CFMs (A_{fr}), the variant carrying the single-nucleotide substitutions that occurred at the same branch as the CFMs (A_{mut}), and the variant carrying both the CFMs and the substitutions (E). The observed patterns of pairwise similarity between these variants differed between proteins.

Importantly, we were unable to infer the order in which CFMs and mutations occurred on the same branch. Therefore, we had to consider two possibilities: that the CFMs occurred prior to the mutations, and that the CFMs occurred after the mutations. We put forth two hypotheses with regard to the similarity patterns: (i) that the mutations occurred after the CFMs, and compensated for their effect; (ii) that the mutations occurred prior to the CFMs, and made them permissible, so that the effect of the CFM was lower on the mutated background.

Under (i), we expect the (A, E) distance (blue in Figs. 7-8) to be lower than the (A, A_{fr}) distance (red). This was the case for mutations in *RASSF4* and *NCR3LG1* by hydropathy measurements, in *Diap1,* by Miyata distance measurements, and in *Cyp6a23,* by both measurements.

Under (ii), we expect the (A_{mut}, E) distance (green) to be lower than the (A, A_{fr}) distance (red). This was the case for mutations in *RASSF4, INSL6* and *NCR3LG1* by hydropathy measurements, and in *Diap1* and *Cyp6a23* by Miyata distance measurements.

Nevertheless, we observe no general tendency for the substitutions to be compensatory or permissive.

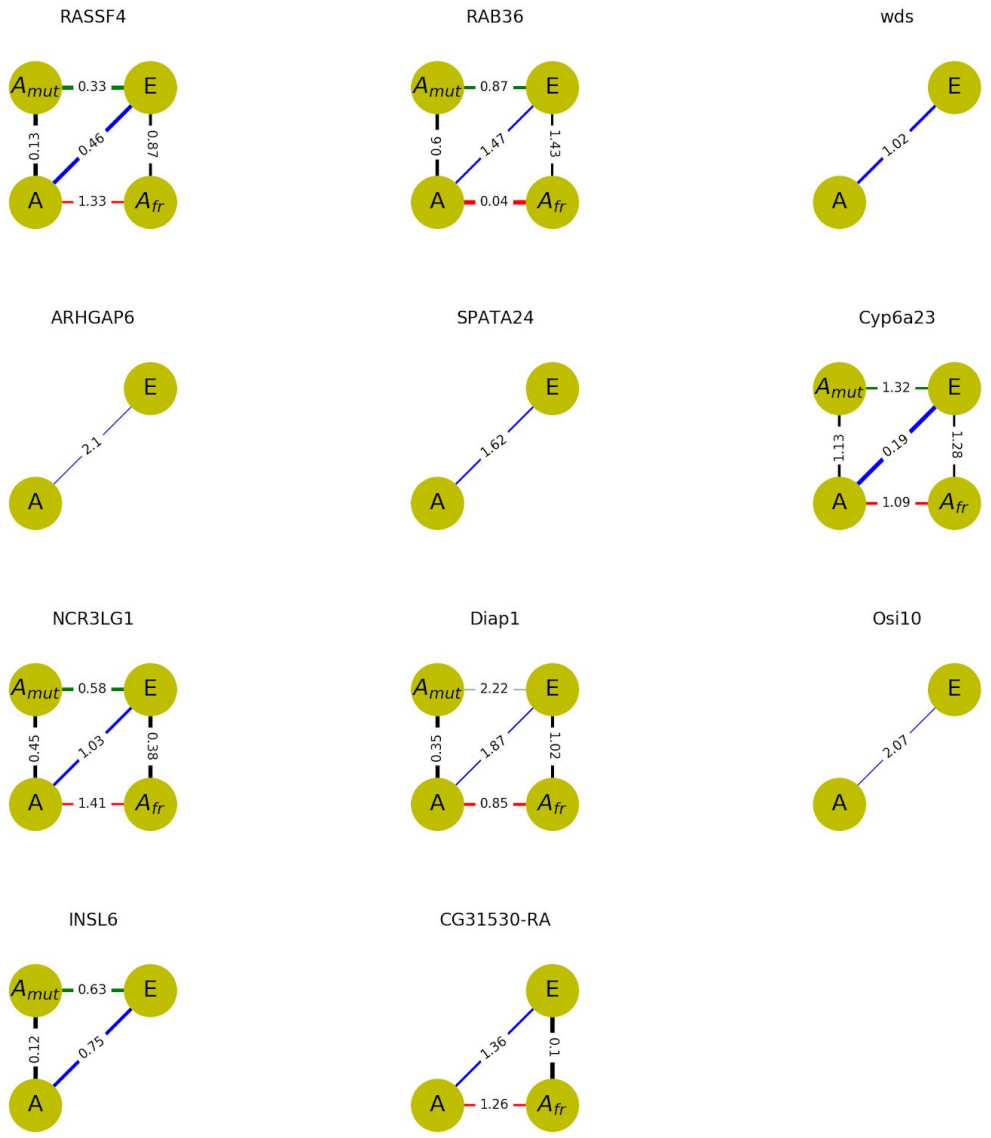

Figure 7. Effects of CFMs and amino acid substitutions occurring on the same phylogenetic branch as the CFMs on the physico-chemical properties of the encoded protein, according to the difference in hydropathy of the encoded amino acid sequence. The four circles in each panel represent the four sequences: without CFMs or mutations (ancestral state, A); with mutations but without CFMs (A_{mut}); with CFMs but without mutations (A_{fr}); with both CFMs and mutations (derived state, E). The numbers on the lines connecting these states represent the distances between corresponding sequences; smaller distances are depicted with bolder lines. Omitted intermediate states (A_{mut} or A_{fr}) correspond to cases when the mutations are synonymous in the context of the CFMs and nonsynonymous in the context of the ancestral state, or vice versa. Both intermediate states are omitted if no mutations happened on the considered branch.

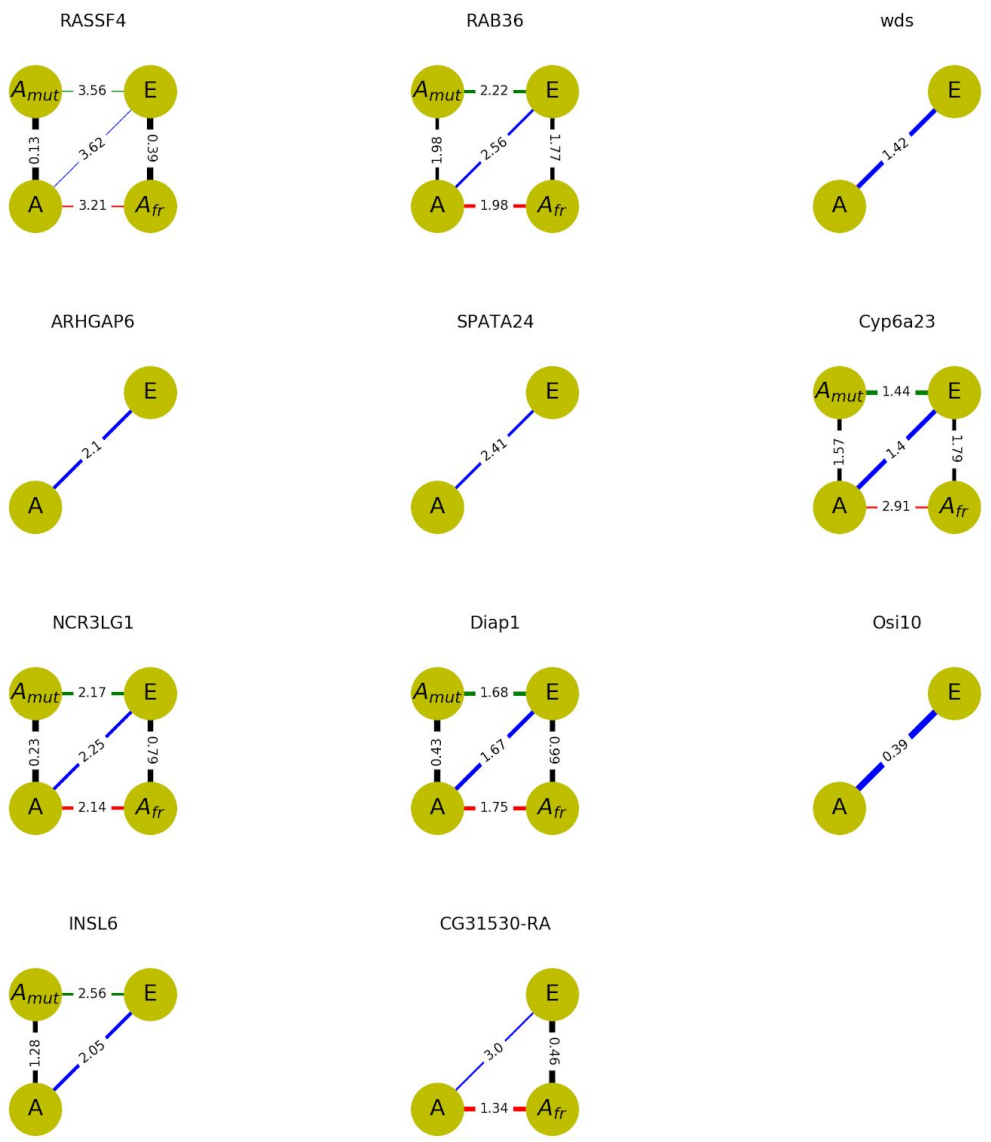

Figure 8. Effects of CFMs and amino acid substitutions occurring on the same phylogenetic branch as the CFMs on the physico-chemical properties of the encoded protein, according to the Miyata difference between the encoded amino acid sequences. The notation is the same as in Figure 7.

Next, we hypothesized that the deleterious effect of CFMs can be compensated by mutations occurring at subsequent branches in the phylogeny (Figures 9 and 10). If this hypothesis holds true, we expect the distance between A and terminal descendant branch (magenta in in Figs.9-10) to be smaller than the (A, E) distance (blue). This was the case in *ARHGAP6, RAB36, CG31530-RA* and *Osi10* by hydropathy measurements and in *RAB36, CG31530-RA* and *Cyp6a23* by Miyata distance measurements. Overall, however, we see no tendency for post-CFM mutations to be compensatory: while 13 of them decrease the distance to the ancestral sequence by hydropathy measurements (14 by Miyata distance), 7 increase it (4 by Miyata distance).

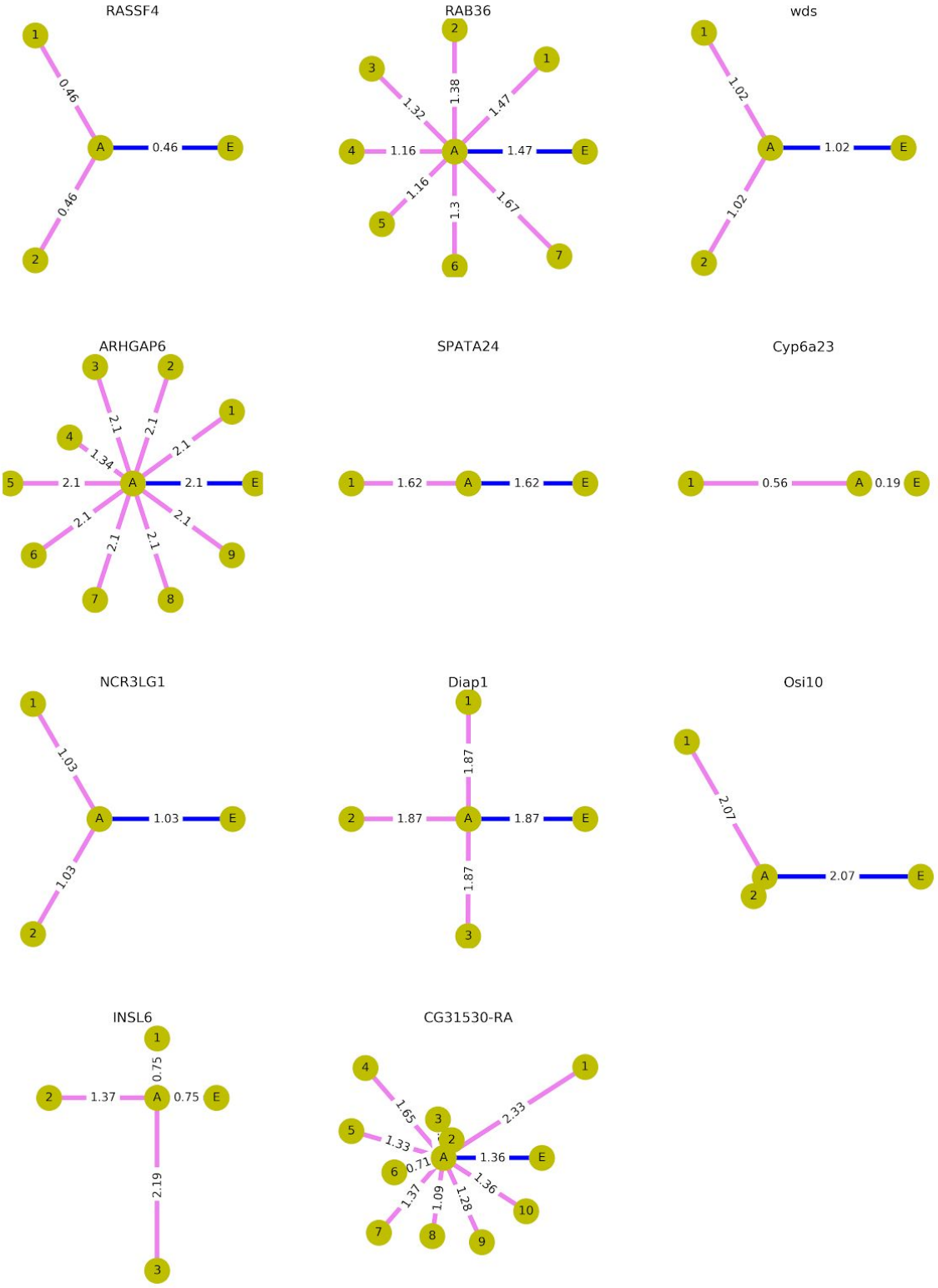

Figure 9. Effects of amino acid substitutions occurring on the phylogenetic branches descendant to the CFMs on the physico-chemical properties of the encoded protein, according to the difference in hydropathy of the encoded amino acid sequence. In each panel, the circles correspond to the different states of a sequence: without CFMs (A); with CFMs (E, always to the right of A); and with both CFMs and mutations on the descendant branches (terminal leaves states, denoted by numbers). The lengths of the lines and the numbers on top of them represent distances between the corresponding

states; many of the distances are identical, reflective of the substitutions common to the corresponding species.

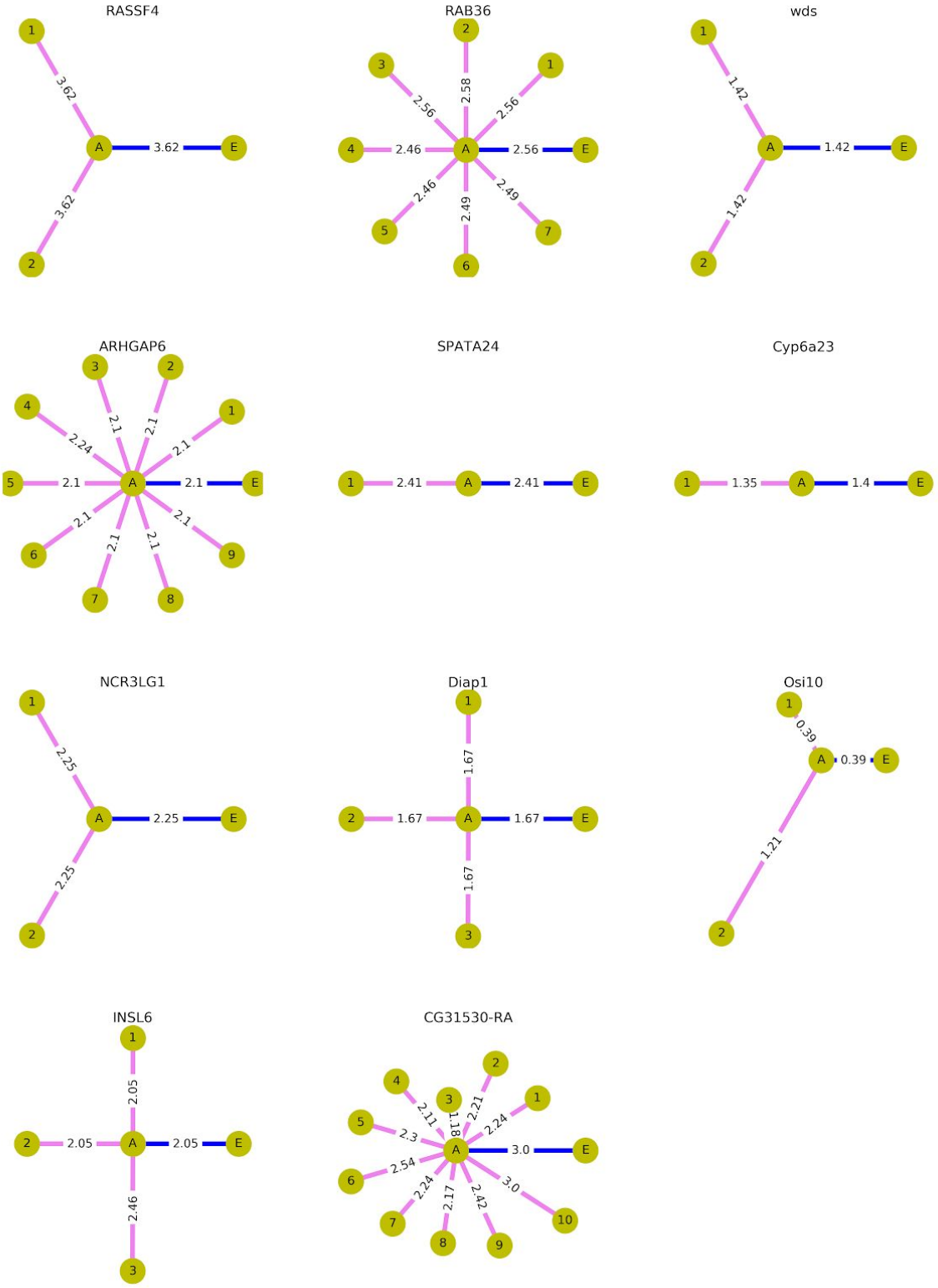

Figure 10. Effects of amino acid substitutions occurring on the phylogenetic branches descendant to the CFMs on the physico-chemical properties of the encoded protein, according to the Miyata difference between the encoded amino acid sequences. The notation is as in Figure 9.

Discussion

In a recent paper, Bartonek et al. ((15, 16)) indicated that frameshifts tend to preserve the physico-chemical properties of the encoded amino acid sequence, and asked: “can evidence be found that frameshifting has indeed played a relevant role during evolution of real proteins?”. Here, we address this question, and show that this is indeed the case. We describe 182 instances in the recent protein evolution of insects, and 536 cases in vertebrates, where a pair of CFMs has fixed in the course of evolution of the lineages that gave rise to extant species. By using stringent filtering criteria, we focus on the 6 insect and 5 vertebrate genes for which this evidence is the most robust, and study how they affect the encoded amino acid sequence. Importantly, we focus on pairs of CFMs that happened rapidly one after the other (so that they were observed on the same segment of the phylogenetic tree); thus, we ignore any possible instances of a longer time lag between the two events, including possible cases of resurrection of established pseudogenes (28).

Among our pairs of CFMs consisting of two same-type events (insertion-insertion or deletion-deletion), the deletion-deletion pairs are more frequent (Table 2), and this pattern is also observed in the larger unfiltered dataset. This is consistent with the fact that, at least in *Drosophila,* deletion mutations are more frequent than insertion mutations (mutational deletion bias). In functional regions, this bias is compensated by a higher probability of insertions to spread and be fixed by selection (fixational insertion bias) ((29)). As a result, among fixed in-frame exonic mutations, insertions and deletions are equally frequent (Fig. 2 in (29)). The fact that the deletion bias is retained for our fixed CFMs suggests that unlike in-frame indels, selection against out-of-frame insertions and deletions is similar, so that the excess of deletions among fixed mutations recapitulates the biases at origin of these mutations.

Expected frequency of CFMs

In a functional gene, a single CFM may cause a loss in fitness, which can then be compensated by the second CFM. There is extensive theoretical literature aiming to estimate the possibility and rates of compensatory evolution under various ranges of population genetics parameters. To our knowledge, the first such model was proposed by Kimura (30); using diffusion approximation, he showed that the rate of compensatory evolution at tightly linked loci is non-negligible over a wide range of $N_e s$, where N_e is the effective population size and s is the coefficient of selection disfavoring the intermediate variant. This approach was further developed by Stephan (31) who focused on RNA evolution and expanded the model assumptions to different selection coefficients against different intermediate mutants.

In subsequent work, two distinct mechanisms for such valley crossing were considered. The first one is sequential fixation of mutations: the first (single or multiple) mutations are deleterious and fixed by drift, and the last mutation is advantageous against

their background and is fixed by selection. Under the second mechanism, the intermediate deleterious variant is never fixed; instead, the haplotypes carrying it remain at low frequencies, either due to genetic drift (so-called stochastic tunneling (32)) or deterministically under mutation-selection balance, before they are "rescued" by subsequent adaptive mutations. The range of parameters leading to each regime was explored by Weissman (33) in asexual populations; for sexual populations, he has shown that low recombination rates may even facilitate valley-crossing (34). One of the results of this study was that, quite intuitively, tunneling is relevant for larger N_e, while sequential fixations are relevant for smaller N_e. Carter and Wagner (35) also obtained similar results, invoking differences in N_e to explain the differences in compensatory evolution rates of enhancers in vertebrates compared to *Drosophila*.

We adopt a simplistic model to estimate the number of CFMs we could expect to find in vertebrate and mammalian genome-scale datasets. We consider two tightly linked positions such that mutations are deleterious in each of them separately but are neutral in combination. The selection coefficient against both single mutants s is assumed to be the same, and to equal that against a loss of function in a non-essential gene. This selection is assumed to be strong, individual CFMs are unlikely to fix in a population under realistic N_e. Therefore, we only consider the rate of tunneling, assuming that the first CFM is maintained by mutation-selection balance before the second CFM occurs. These assumptions are close to those of Kimura's (30).

Let L be an average gene length; p_1, the rate of the first uncompensated mutation per nucleotide per generation; and p_2, the rate of the second, compensatory mutation per nucleotide per generation. The probability of a CFM at some position of this gene per generation m_1 then equals

$$m_1 = 1-(1-p_1)^L \ .$$

Similarly, the probability of a compensatory CFM at some position of this gene per generation m_2 equals

$$m_2 = 1-(1-p_2)^l \ ,$$

where l is the number of nucleotides surrounding the first CFM where the second CFM would be compensatory.

Under mutation-selection balance (36), the expected frequency of uncompensated frameshift in a population equals m_1/s. The compensated genotypes will then originate by mutation of the deleterious alleles at per generation rate of $m_1 m_2/s$.

If compensation is exact, so that the compensated genotype has the same fitness as the ancestral variant, the rate of fixation of such compensatory pairs equals the rate of the mutation giving rise to them: $P_{fix} = m_1 m_2/s$ (37). Over N generations, a total of $N m_1 m_2/s$ such pairs are expected to have occurred.

To estimate p_1 and p_2, we consider four types of deleterious CFMs: insertion of lengths 1 or 2 and deletions of lengths 1 or 2. For each of these mutations, we consider the sets of mutations that can compensate for it (for example, an insertion of length 1 can only be compensated by an insertion of length 2 or a deletion of length 1 but not by a deletion of length 2), and sum over the probabilities of these mutations. The resulting P_{fix} terms are the sums over these pairs of deleterious and compensatory mutations.

We use the estimate for the *de novo* mutation rate $= 10^{-8}$ for vertebrates from (38) and $= 4.9 * 10^{-9}$ for *Drosophila* from (39); the estimates for p_1 for each type of

frameshifting mutation from (29): 0.020, 0.010, 0.042 and 0.015 for insertion of length 1, insertion of length 2, deletion of length 1 and deletion of length 2, respectively; and the estimate for the selection coefficient against a loss of function mutation in non-essential genes $s = 0.0015$ in *Drosophila (40)* and $s = 0.005$ for vertebrates (estimated for human in (41)). We estimated N and L from our data. For each gene, N the mean length of a phylogenetic tree including only the species that have passed all filtering steps, without accounting for the terminal branches (since we only consider indels inherited by multiple species) and measured in the genome-wide number of synonymous substitutions. In our data, $N = 176\,536\,337$ for vertebrates, and $N = 403\,272\,889$ for insects. L is just the mean length of analysed genes, and equaled $L = 1897$ for vertebrates and $L = 1204$ for insects. We take l to equal 100 (the distance between all observed CFMs is less than that, Table 2).

Using these estimates, we obtain $Nm1m2/s = 0.002$ for vertebrates and 0.003 for insects. Given the number of protein-coding genes we analyzed (21208 for vertebrates and 15283 for insects), we expect to find 52±14 and 43±13 genes with CFMs in vertebrates and insects respectively (95 credible intervals are calculated using normal distribution approximation for binomial distribution with parameters *n*=21208, *p*=0.002 and *n*=15283, *p*=0.003).

The above values lie between the numbers of raw unfiltered indels (536 genes in vertebrates and 182 in insects) and filtered indels (6 and 5 respectively) that we obtain. This could indicate that our filtering steps are too stringent, perhaps excluding some of the *bona fide* CFMs. It could also result from violations of our assumptions. In particular, the above model assumes that the pair of fixed CFMs confers the same fitness as the wild type sequence; if it is selectively inferior, the expected numbers will be lower, and if it confers a selective advantage, these numbers will be higher.

CFMs and sequence conservation

Our results indicate that genes with fixed CFM pairs are generally weakly constrained, as suggested by their lower ω values. This is in line with the expected radical effect of CFMs on the encoded protein, and with the majority of CFMs being deleterious. Still, the deficit of CFM pairs at more constrained proteins could be due to selection against the intermediate single-CFM variant (which would have to be present at the population at some point unless both mutations were acquired simultaneously by the same individual), and/or the ultimate two-CFM variant.

The finding that CFMs do not preferentially fix in the least conserved regions of a protein favors the first explanation. Indeed, it suggests that the overall constraint of the protein imposes stronger limitations on the CFMs than the constraint of the region between the CFMs: CFMs may fix in any region if the protein is sufficiently unimportant. This would be the case if the limiting step was survival of some individuals with uncompensated mutation for some number of generations. By contrast, if the CFM positions were mainly determined by the effect of selection at the two-CFM stage, we would expect CFM pairs to preferentially happen at regions of low conservation, as such selection would discriminate between the protein versions with ancestral and derived amino acid sequences encoded by the region between the CFMs.

This does not mean that the pairs of CFMs are functionally unimportant. In one of the genes (CG31530-RA), we find that negative selection has relaxed at the same time as the CFMs were fixed, suggesting that the CFMs could have perturbed its function. This gene encodes a transmembrane protein involved in transmembrane transport ((42); https://www.uniprot.org/uniprot/D2NUL5). *Drosophila melanogaster* and 10 other species of flies have the version of this gene with indels.

Conclusions

We describe 718 candidate pairs of compensatory frameshifting mutations in the past evolution of vertebrate and insect species, including 11 high-confidence cases that passed all our stringent filtering steps. CFM pairs tend to fix in genes under relaxed negative selection, but not in particularly poorly conserved segments, suggesting that the rate of their fixation is mainly limited by selection against individual uncompensated CFMs. Overall, these results put forward CFMs as a previously undescribed source of novelty in protein evolution.

Data availability

All processed data (after realignment and clipping) is available at FigShare (https://figshare.com/articles/dataset/Re-aligned_data_from_UCSC_100-way_vertebrate_and_124-way_insect_exome_alignments_/13272974). The python script used to infer CFMs is available at the gitHub repository (https://github.com/Captain-Blackstone/Compensatory-frameshifts).

Acknowledgements

The authors are grateful to Alexey Kondrashov for discussions of the model for the expected frequency of compensatory frameshifting mutations; to Sergey Lysenkov for advice on statistics; and to members of Georgii Bazykin's and Alexey Kondrashov's labs for valuable discussions. Computations were performed using the Makarich HPC cluster of the Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University.

References

1. Povolotskaya,I.S. and Kondrashov,F.A. (2010) Sequence space and the ongoing expansion of the protein universe. *Nature*, **465**, 922–926.

2. Hancock,J.M. and Simon,M. (2005) Simple sequence repeats in proteins and their significance for network evolution. *Gene*, **345**, 113–118.

3. Artamonova,I.I. and Gelfand,M.S. (2007) Comparative genomics and evolution of alternative splicing: the pessimists' science. *Chem. Rev.*, **107**, 3407–3430.

4. Vakhrusheva,A.A., Kazanov,M.D., Mironov,A.A. and Bazykin,G.A. (2011) Evolution of

prokaryotic genes by shift of stop codons. *J. Mol. Evol.*, **72**, 138–146.

5. Neme,R. and Tautz,D. (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, **14**, 117.

6. Keese,P.K. and Gibbs,A. (1992) Origins of genes: “big bang” or continuous creation? *Proc Natl Acad Sci USA*, **89**, 9489–9493.

7. Carvunis,A.-R., Rolland,T., Wapinski,I., Calderwood,M.A., Yildirim,M.A., Simonis,N., Charloteaux,B., Hidalgo,C.A., Barbette,J., Santhanam,B., *et al.* (2012) Proto-genes and de novo gene birth. *Nature*, **487**, 370–374.

8. Ohno,S. (1970) Evolution by gene duplication Springer Berlin Heidelberg, Berlin, Heidelberg.

9. Hahn,Y. and Lee,B. (2005) Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics*, **21 Suppl 1**, i186-94.

10. Ohno,S. (1984) Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proc Natl Acad Sci USA*, **81**, 2421–2425.

11. Vandenbussche,M., Theissen,G., Van de Peer,Y. and Gerats,T. (2003) Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Res.*, **31**, 4401–4409.

12. Hu,J. and Ng,P.C. (2012) Predicting the effects of frameshifting indels. *Genome Biol.*, **13**, R9.

13. MacArthur,D.G., Balasubramanian,S., Frankish,A., Huang,N., Morris,J., Walter,K., Jostins,L., Habegger,L., Pickrell,J.K., Montgomery,S.B., *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.

14. Crick,F.H., Barnett,L., Brenner,S. and Watts-Tobin,R.J. (1961) General nature of the genetic code for proteins. *Nature*, **192**, 1227–1232.

15. Wang,X., Dong,Q., Chen,G., Zhang,J., Liu,Y., Zhao,J., Peng,H., Wang,Y., Cai,Y., Wang,X., *et al.* (2016) Why are frameshift homologs widespread within and across species? *BioRxiv*, 10.1101/067736.

16. Bartonek,L., Braun,D. and Zagrovic,B. (2020) Frameshifting preserves key physicochemical properties of proteins. *Proc Natl Acad Sci USA*, **117**, 5907–5912.

17. Rockah-Shmuel,L., Tóth-Petróczy,Á., Sela,A., Wurtzel,O., Sorek,R. and Tawfik,D.S. (2013) Correlated occurrence and bypass of frame-shifting insertion-deletions (InDels) to give functional proteins. *PLoS Genet.*, **9**, e1003882.

18. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M., *et al.* (2015) The UCSC

Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670-81.

19. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

20. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

21. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S., *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.

22. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

23. Tamura,K., Stecher,G., Peterson,D., Filipski,A. and Kumar,S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725–2729.

24. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

25. Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–236.

26. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO gene nomenclature committee (HGNC). *Hum. Genet.*, **109**, 678–680.

27. Ng,P.C., Levy,S., Huang,J., Stockwell,T.B., Walenz,B.P., Li,K., Axelrod,N., Busam,D.A., Strausberg,R.L. and Venter,J.C. (2008) Genetic variation in an individual human exome. *PLoS Genet.*, **4**, e1000160.

28. Esfeld,K., Berardi,A.E., Moser,M., Bossolini,E., Freitas,L. and Kuhlemeier,C. (2018) Pseudogenization and resurrection of a speciation gene. *Curr. Biol.*, **28**, 3776-3786.e7.

29. Leushkin,E.V., Bazykin,G.A. and Kondrashov,A.S. (2013) Strong mutational bias toward deletions in the Drosophila melanogaster genome is compensated by selection. *Genome Biol. Evol.*, **5**, 514–524.

30. Kimura,M. (1985) The role of compensatory neutral mutations in molecular evolution. *J. Genet.*, **64**, 7–19.

31. Stephan,W. (1996) The rate of compensatory evolution. *Genetics*, **144**, 419–426.

32. Iwasa,Y., Michor,F. and Nowak,M.A. (2004) Stochastic tunnels in evolutionary dynamics. *Genetics*, **166**, 1571–1579.

33. Weissman,D.B., Desai,M.M., Fisher,D.S. and Feldman,M.W. (2009) The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.*, **75**, 286–300.

34. Weissman,D.B., Feldman,M.W. and Fisher,D.S. (2010) The rate of fitness-valley

crossing in sexual populations. *Genetics*, **186**, 1389–1410.

35. Carter,A.J.R. and Wagner,G.P. (2002) Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. *Proc. Biol. Sci.*, **269**, 953–960.

36. Crow,J.F. and Kimura,M. (1970) An introduction to population genetics theory. *An introduction to population genetics theory.*

37. Kimura and M. (1983) The Neutral Theory of Molecular. *The Neutral Theory of Molecular Evolution*.

38. Kong,A., Frigge,M.L., Masson,G., Besenbacher,S., Sulem,P., Magnusson,G., Gudjonsson,S.A., Sigurdsson,A., Jonasdottir,A., Jonasdottir,A., *et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, **488**, 471–475.

39. Assaf,Z.J., Tilk,S., Park,J., Siegal,M.L. and Petrov,D.A. (2017) Deep sequencing of natural and experimental populations of Drosophila melanogaster reveals biases in the spectrum of new mutations. *Genome Res.*, **27**, 1988–2000.

40. Langley,C.H., Voelker,R.A., Brown,A.J., Ohnishi,S., Dickson,B. and Montgomery,E. (1981) Null allele frequencies at allozyme loci in natural populations of Drosophila melanogaster. *Genetics*, **99**, 151–156.

41. Cassa,C.A., Weghorn,D., Balick,D.J., Jordan,D.M., Nusinow,D., Samocha,K.E., O'Donnell-Luria,A., MacArthur,D.G., Daly,M.J., Beier,D.R., *et al.* (2017) Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.*, **49**, 806–810.

42. UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190-5.