SUPPLEMENTARY TABLE AND FIGURES LEGENDS

Supplementary Table 1. Summary of injections, indicating the concentration of mRNA or plasmid injected, genetic background of embryos injected and stage for protein extraction.

Supplementary Table 2. Properties of standard and codon modified versions of genes. Properties that are outside the threshold for 90% of highly expressed genes in zebrafish are highlighted.

Supplementary Figure 1. Codon usage in highly expressed genes.

(a) Highly expressed genes identified from independent data sources show significant overlap. The Himix set comprises all genes in the four sets.

(b) Relative synonymous codon usage for the Refseq (top row) and Himix (bottom row) sets. The most commonly used codon for each amino acid is highlighted, for the Refseq set (grey) and Himix set (yellow). The two amino acids encoded by six codons that are separated in this table are highlighted to aid comparison (Ser: blue, Arg: pink).

(c-f) RSCU for all 64 codons, in each of the highly expressed gene sets compared to the Refseq set. Red dotted line indicates equal usage. The UAA stop codon is colored green in each set.

(g-h) Comparison of RSCU in the Refseq set, to the relative abundance of tRNA decoders per amino acid, without adjusting for wobble decoding (g) and after adjusting for wobble decoding (h).

(i) Codon adaptation index using tRNA isoacceptor frequencies as a measure of the codon bias. TAI was calculated for each gene (excluding the single codon amino acids Met and Trp) and the box plot shows the distribution of TAIs for the Himix and the individual highly expressed gene sets. t-test * $p < 0.001$.

Supplementary Figure 2. Minor codon usage

(a) Distribution of codon usage (excluding stop codons, Met and Trp) for the Refseq gene set. Codon usage is the abundance of each codon, as a fraction of all codons. Red shaded bars were designated as minor codons: Ala-GCG, Arg-CGA, Arg-CGC, Arg-CGG, Arg-CGU, Gly-GGG, Ile-AUA, Leu-CUA, Leu-UUA, Pro-CCG, Ser-UCG, Thr-ACG, Val-GUA.

(b-e) Minor codon usage for all genes (Refseq) compared to each high expression gene set. Black shaded circles are Leu-UUA, Leu-CUA and Ile-AUA. Green shaded circles are Arg-CGU and Arg-CGC.

(f) Cumulative frequency histogram for rare codon use by genes in the Refseq (black) and Himix (blue) sets. Dotted line: less than 2.9% of codons are rare codons in 90% of highly expressed genes.

Supplementary Figure 3. Codon selection is influenced by dinucleotide frequency.

(a) Double stranded odds ratio for each dinucleotide in the Refseq set (black) and Himix set (blue). The double-stranded odds ratio ρ^*_{xy} (55) is the frequency of the dinucleotide XY (appearing on either strand), divided by the expected frequency of this dinucleotide based on the frequency of nucleotides X and Y. Both Refseq and Himix gene sets avoid the dinucleotides UA and CG.

(b) The frequency of GC (black) and CG (grey) dinucleotides in each frame of the mRNA for the Refseq (R) and Himix (H) sets.

(c) As for (b) but showing AU (black) and UA (grey) dinucleotides. CG and UA dinucleotides are depleted in all three frames.

(d) Sources of CG dinucleotide suppression. Depletion of CG dinucleotides in the 1-2 and 2-3 positions correlates with codon bias. CG*n* codons all encode Arg which can also use synonymous codons AGA and AGG. The AG*n* codons represent a total 47% RSCU for this codon and thus CG*n* is depleted. All instances where the CG dinucleotide is in the second and third position of the codon (*n*CG: Ser-UCG, Thr-ACG, Ala-GCG and Pro-CCG) are minor codons. Panel A shows that CG dinucleotides spanning two codons (*nn*C·G*nn*) are also avoided. None of the amino acids encoded by G*nn* have synonymous codons starting with a different nucleotide and so wherever amino acid usage could create a *nn*C·G*nn* sequence, the second codon can not avoid starting with a G. However many *nn*C codons have synonymous codons with a different third nucleotide and indeed for six *nn*C amino acids, the most used codon in both the Refseq and Himix gene sets is the *nn*C variant: Phe-UUC, Tyr-UAC, Ile-AUC, Asn-AAC, His-CAC, Asp-GAC. For these amino acids, we compared the frequency of the *nn*C codon when it came directly before a G*nn* codon (creating a CG dinucleotide) or when it came directly after a G*nn* amino acid (creating the sequence G*nn·nn*C). The *nn*C synonymous codon was reduced when it was immediately before a G*nn* codon (grey bars), in comparison to its usage when it came directly after the G*nn* codon (black bars) where values were similar to overall codon usage (Diamonds: the RSCU for each codon irrespective of its position in the sequence)

(e-f) Sources of UA dinucleotide suppression. Codon bias partly explains UA depletion in the 1-2 and 2-3 positions. UA*n* codons are tyrosine and stop codons. As UA*n* stop codons are less than 0.1% of all codon usage, this in part explains the relative infrequency of this dinucleotide. *n*UA are Leu-CUA, Val-GUA, Leu-UUA and Ile-AUA all of which are minor codons. UA dinucleotides created by the juxtaposition of two codons (*nn*T·A*nn*) are avoided. Five amino acids (Ile, Met, Thr, Asn and Lys) are encoded by A*nn* codons with no synonymous codons that can avoid an A in the first position. For Pro and Ala, the most frequently used codon in zebrafish has the pattern *nn*U but this codon was avoided when a Pro or Ala came directly before Ile, Met, Thr, Asn or Lys (grey bars), in comparison to usage directly after a A*nn* containing codon (black bars) that was similar to overall RSCU (diamonds). Similarly, A*nn* codons for Arg and Ser were avoided when they were immediately after a *nn*U codon (f, grey bars) compared to normal usage when appearing before a *nn*U codon (black bars).

(g-h) Cumulative frequency histograms showing %CG (g) and %TA (h) dinucleotides for the Refseq (black) and Himix (blue) sets. For 90% of highly expressed genes, CG dinucleotides were less than 8.2% of the total number of dinucleotides, and TA dinucleotides are less than 6.7% of the total.

Supplementary figure 4. Nucleotide sequence features enriched in highly expressed genes in mouse.

(a) Relative synonymous codon usage for the mouse Refseq (top row) and Himix (bottom row) sets. Highlighted frequencies are the most commonly used codon for each amino acid for the Refseq set (grey) and Himix set (yellow).

(b) RSCU for 62 codons (excluding Met and Trp) in the Himix set compared to the Refseq set, 25 of which show significant usage differences between sets (grey circles, X^2 test, $p < 0.001$). Red dotted line indicates equal usage.

(c) Optimal codon usage for the Refseq and Himix sets. t-test * $p < 0.001$.

(d) Distribution of codon usage (excluding stop codons, Met and Trp) for the Refseq set. Red shaded bars are 11 minor codons: Ala-GCG, Arg-CGA, Arg-CGC, Arg-CGU, Ile-AUA, Leu-CUA, Leu-UUA, Pro-CCG, Ser-UCG, Thr-ACG, Val-GUA.

(e) Minor codon usage for Refseq compared to Himix gene sets. Codons with significant difference in usage are indicated (filled circles, X^2 test, $p < 0.001$). We operationally designate the three codons used at a significantly reduced rate by highly expressed genes (Leu-CUA, Ile-AUA and Leu-UUA) as rare codons in mouse. Red dotted line indicates equal codon usage.

(f) Minor (closed circles) and rare (open circles) codon frequency for 30 bp windows along the coding sequence starting at the base indicated on the x-axis. 'Last' indicates the last 30 bp of the coding sequence. For minor codons across all points, $X^2 = 20.1$, $p = 0.04$, suggesting a trend for reduced use in the Himix set except in the region following the transcription start site. For rare codons, $X^2 = 181.4$, $p < 0.001$ and reduced use is seen throughout the coding sequence.

(g) Cumulative frequency histogram for minor (solid lines) and rare (dotted lines) codon use by genes in the Refseq (black) and Himix (blue) sets. Dotted line: less than 11.5 % of codons are minor in 90% and less than 2.9% of codons are rare in highly expressed genes.

(h) Double stranded odds ratio dinucleotides in the Refseq (black) and Himix sets (blue).

(i) Relative use frequency for stop tetranucleotides in the Refseq and Himix gene sets. * $p < 0.001$ (X^2 test).

(j) Cumulative frequency histograms for the free energy of the minimum energy secondary structure (dG) for Refseq (black) and Himix (blue) gene sets, calculated at 37°C. Dotted line indicates that for 90% of highly expressed genes, the dG was greater than -12.9 kcal/mol. Means are -9.1 ± 3.8 kcal/mol and -8.6 ± 3.3 kcal/mol for the Refseq and Himix sets (t-test, p=0.002).

(k) DNA logo representing the nucleotide usage frequency in the six bases before the initiator ATG for Refseq genes (top) and the Himix gene set (bottom). For both sets, the consensus, GCCACC, was also the most frequently used sequence, present in 1.03 and 3.28% of transcripts in the Refseq and Himix sets respectively.

Supplementary figure 5. Installation and operation guide for CodonZ.

Supplementary figure 6. Stop codon read-through and protein expression from different stop tetranucleotides.

(a) Normal exposure of Western blot for Cer (second AB coupled to IR800, green) and TagRFPT (second AB coupled to IR700, red) showing Cer and lack of expression of proteins of the predicted size (kDa) of TagRFPT and Cer-TagRFPT fusion proteins. Lanes are from constructs with the indicated intervening stop tetranucleotides.

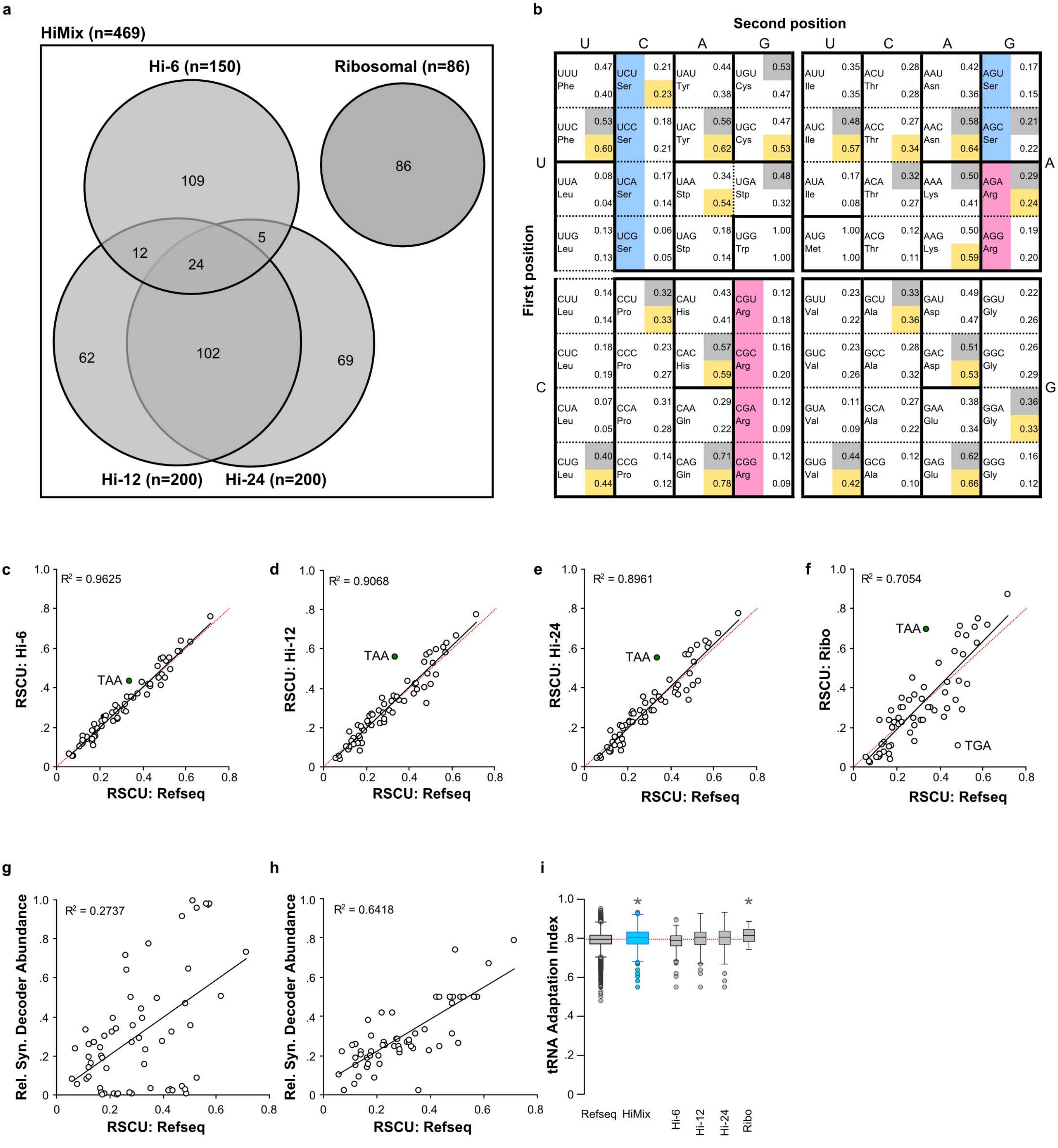
(b) Long exposures for the same blots showing a lack of specific bands that would indicate read-through transcription producing TagRFPT immunoreactivity.

(c-e) Images of 24 hpf embryos expressing Cer-stop-TagRFPT with the indicated stop tetranucleotides.

Supplementary figure 7. Effect of PRE element in 3' UTR on protein expression.

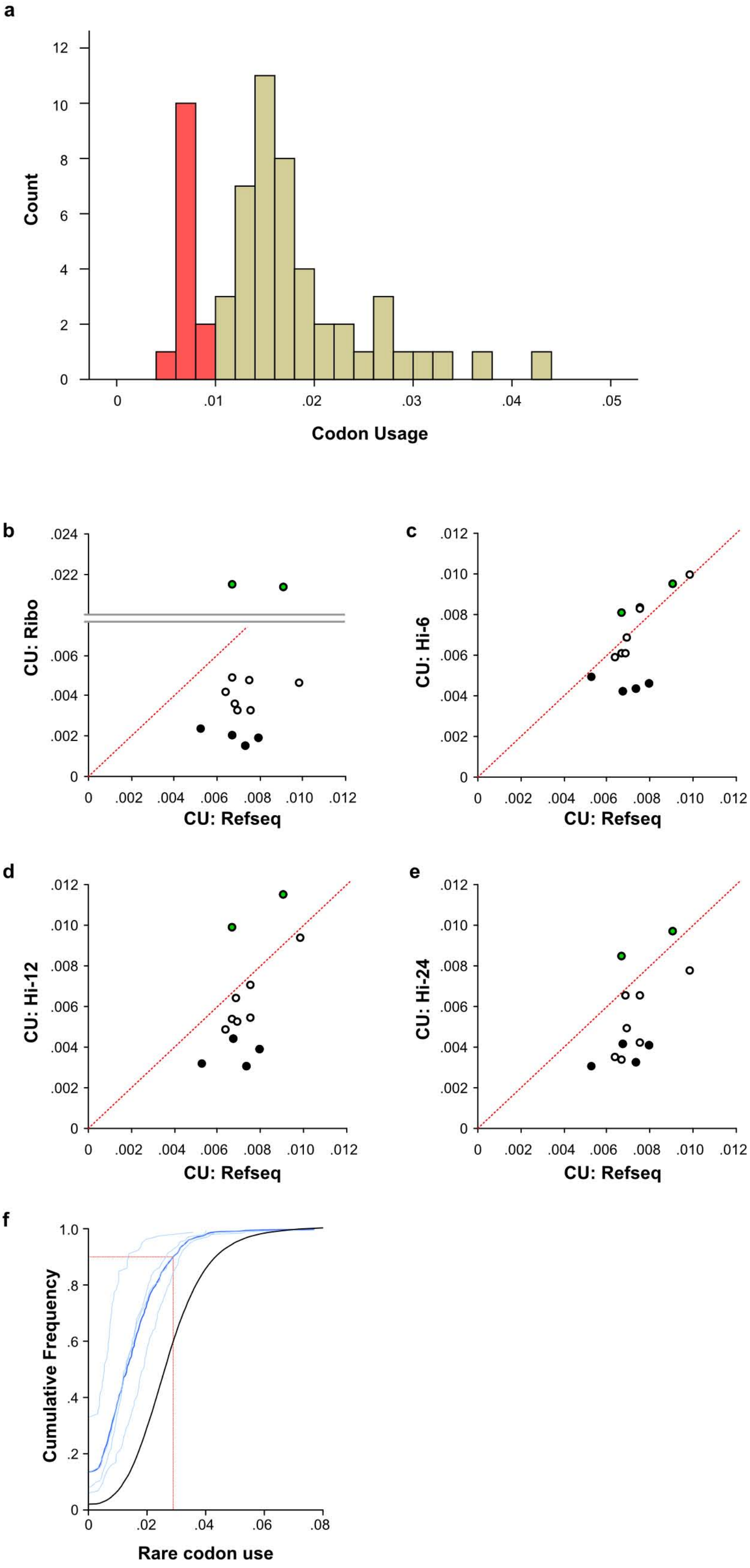
(a) TagRFPT expression in embryos injected with mRNA for TagRFPT with or without the woodchuck hepatitis virus PRE element (N=3).

(b) Effect of PRE element in transgenic embryos. Transgenic UAS:NTR-TagRFPT larvae with and without the PRE element in the 3'UTR were made by microinjection into transgenic line HuC:Gal4 in which Gal4 is expressed throughout the brain (92) (N=5).

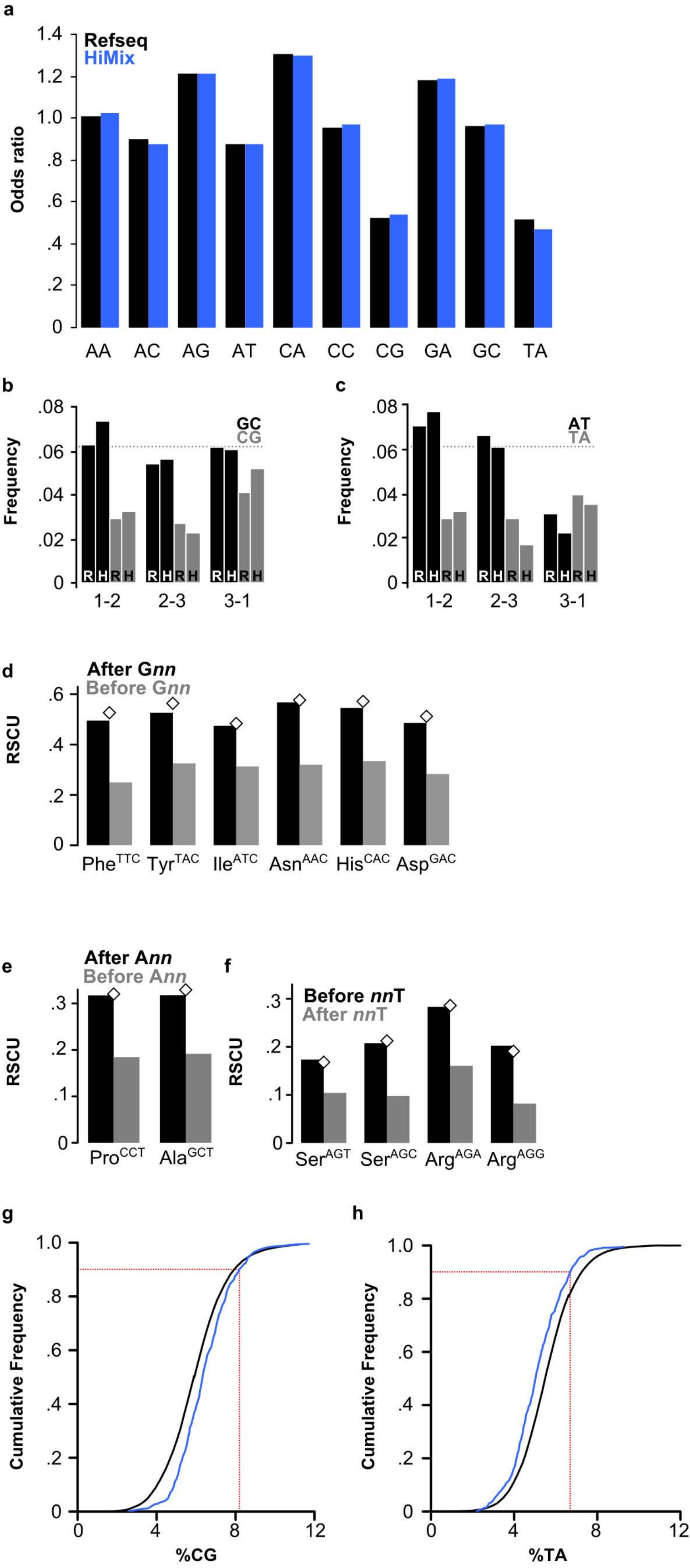


First position	U	C	A	G
U	UUU Phe 0.47 0.40	UCU Ser 0.21 0.23	UAU Tyr 0.44 0.38	UGU Cys 0.53 0.47
	UUC Phe 0.53 0.60	UCC Ser 0.18 0.21	UAC Tyr 0.56 0.62	UGC Cys 0.47 0.53
	UUA Leu 0.08 0.04	UCA Ser 0.17 0.14	UAA Stp 0.34 0.54	UGA Stp 0.48 0.32
	UUG Leu 0.13 0.13	UCG Ser 0.06 0.05	UAG Stp 0.18 0.14	UGG Trp 1.00 1.00
C	CUU Leu 0.14 0.14	CCU Pro 0.32 0.33	CAU His 0.43 0.41	CGU Arg 0.12 0.18
	CUC Leu 0.18 0.19	CCC Pro 0.23 0.27	CAC His 0.57 0.59	CGC Arg 0.16 0.20
	CUA Leu 0.07 0.05	CCA Pro 0.31 0.28	CAA Gln 0.29 0.22	CGA Arg 0.12 0.09
	CUG Leu 0.40 0.44	CCG Pro 0.14 0.12	CAG Gln 0.71 0.78	CGG Arg 0.12 0.09
A	AUU Ile 0.35 0.35	ACU Thr 0.28 0.28	AAU Asn 0.42 0.36	AGU Ser 0.17 0.15
	AUC Ile 0.48 0.57	ACC Thr 0.27 0.34	AAC Asn 0.58 0.64	AGC Ser 0.21 0.22
	AUA Ile 0.17 0.08	ACA Thr 0.32 0.27	AAA Lys 0.50 0.41	AGA Arg 0.29 0.24
	AUG Met 1.00 1.00	ACG Thr 0.12 0.11	AAG Lys 0.50 0.59	AGG Arg 0.19 0.20
G	GUU Val 0.23 0.22	GCU Ala 0.33 0.36	GAU Asp 0.49 0.47	GGU Gly 0.22 0.26
	GUC Val 0.23 0.26	GCC Ala 0.28 0.32	GAC Asp 0.51 0.53	GGC Gly 0.26 0.29
	GUA Val 0.11 0.09	GCA Ala 0.27 0.22	GAA Glu 0.38 0.34	GGA Gly 0.36 0.33
	GUG Val 0.44 0.42	GCG Ala 0.12 0.10	GAG Glu 0.62 0.66	GGG Gly 0.16 0.12

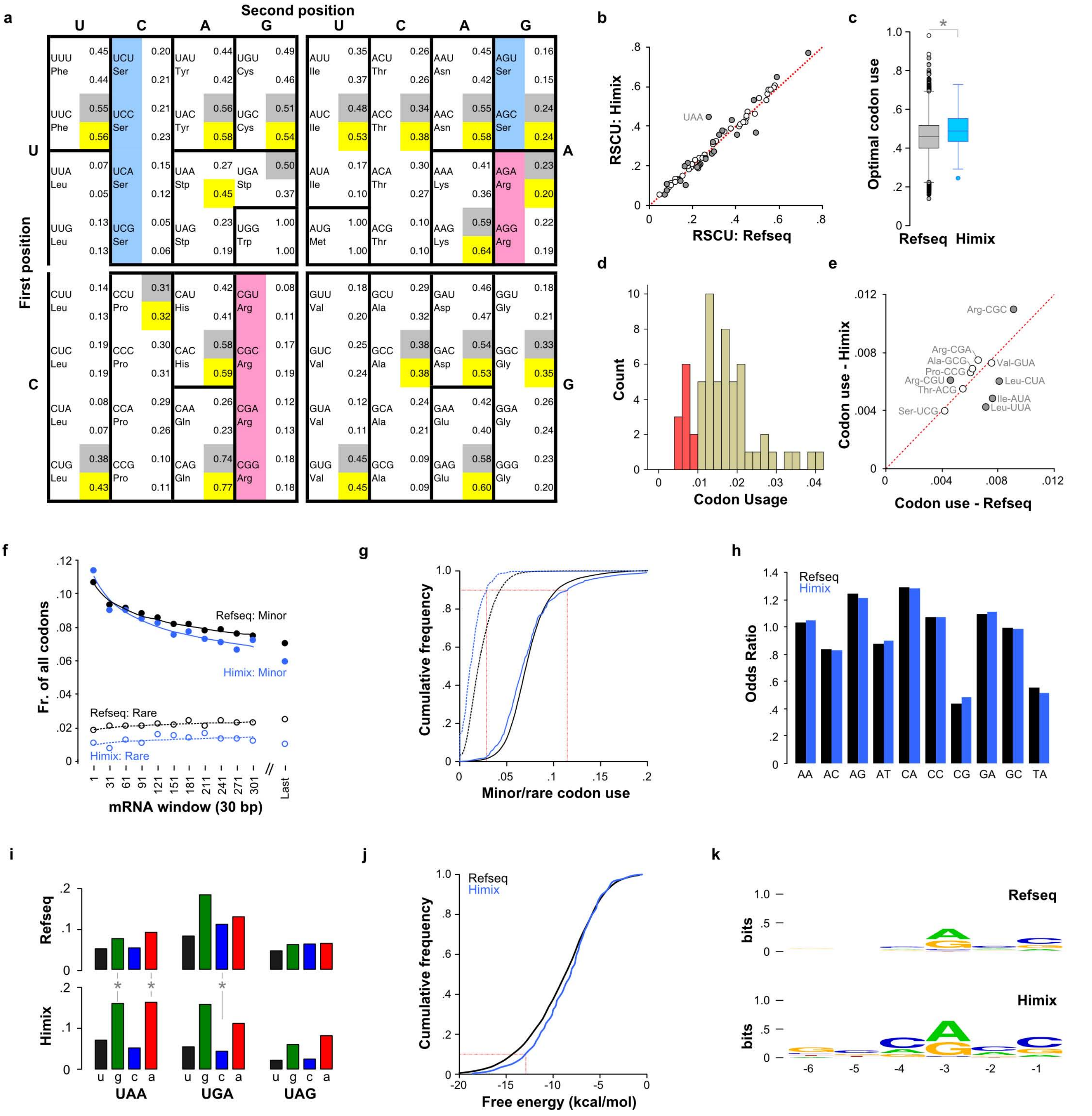
Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3



First position	U	C	A	G	U	C	A	G	
U	UUU Phe 0.45 0.44	UCU Ser 0.20 0.21	UAU Tyr 0.44 0.42	UGU Cys 0.49 0.46	AUU Ile 0.35 0.37	ACU Thr 0.26 0.26	AAU Asn 0.45 0.42	AGU Ser 0.16 0.15	A
	UUC Phe 0.55 0.56	UCC Ser 0.21 0.23	UAC Tyr 0.56 0.58	UGC Cys 0.51 0.54	AUC Ile 0.48 0.53	ACC Thr 0.34 0.38	AAC Asn 0.55 0.58	AGC Ser 0.24 0.24	
	UUA Leu 0.07 0.05	UCA Ser 0.15 0.12	UAA Stp 0.27 0.45	UGA Stp 0.50 0.37	AUA Ile 0.17 0.10	ACA Thr 0.30 0.27	AAA Lys 0.41 0.36	AGA Arg 0.23 0.20	
	UUG Leu 0.13 0.13	UCG Ser 0.05 0.06	UAG Stp 0.23 0.19	UGG Trp 1.00 1.00	AUG Met 1.00 1.00	ACG Thr 0.10 0.10	AAG Lys 0.59 0.64	AGG Arg 0.22 0.19	
C	CUU Leu 0.14 0.13	CCU Pro 0.31 0.32	CAU His 0.42 0.41	CGU Arg 0.08 0.11	GUU Val 0.18 0.20	GCU Ala 0.29 0.32	GAU Asp 0.46 0.47	GGU Gly 0.18 0.21	G
	CUC Leu 0.19 0.19	CCC Pro 0.30 0.31	CAC His 0.58 0.59	CGC Arg 0.17 0.19	GUC Val 0.25 0.24	GCC Ala 0.38 0.38	GAC Asp 0.54 0.53	GGC Gly 0.33 0.35	
	CUA Leu 0.08 0.07	CCA Pro 0.29 0.26	CAA Gln 0.26 0.23	CGA Arg 0.12 0.13	GUA Val 0.12 0.11	GCA Ala 0.24 0.21	GAA Glu 0.42 0.40	GGA Gly 0.26 0.24	
	CUG Leu 0.38 0.43	CCG Pro 0.10 0.11	CAG Gln 0.74 0.77	CGG Arg 0.18 0.18	GUG Val 0.45 0.45	GCG Ala 0.09 0.09	GAG Glu 0.58 0.60	GGG Gly 0.23 0.20	

Supplementary figure 4

Supplementary Figure 5.

CodonZ

Installation and Operation Guide

CodonZ is software for replacing the codons in a given gene with a set of codons designed to improve protein expression levels. This manual covers setting up and operating CodonZ.

CodonZ is written by Harold Burgess (NICHD, haroldburgess@mail.nih.gov).

12/22/2014 version 2.1

Table of Contents

1. Installation

1. CodonZ is written in IDL and packaged as runtime code that can operate on any computer with the free 'virtual machine' platform from IDL. CodonZ will need to be installed on a PC, running Windows XP or 7. Download CodonZ runtime code (contained in the file 'codonz2.sav') from:
https://science.nichd.nih.gov/confluence/display/burgess/Software

2. Create a directory C:\Program Files\CodonZ. Then place the codonz2.sav file in the CodonZ folder

3. Download the IDL virtual machine: go to the Exelis Visual Information Solutions website (http://www.exelisvis.com/) and if you do not already have an account, register as a new user (click on "My Account" and follow the instructions for setting up a new account). You will receive an email with a link to verify your account, which will allow you to login to the website. Approval and the ability to download can take an additional 24 hours, however. It seems that although an initial verification email is sent, there is no email notification for approval being granted, so periodically check whether downloads are now accessible once you have logged in to the Exelis website. Then download IDL (this will be more than 1 Gb). You may have to perform a full IDL installation, but you do not need to purchase a license to run IDL runtime code.

4. Install software that calculates the RNA secondary structure. Our initial analyses were performed using the UNAfold package. However you may now need to pay a fee to use this software. We have therefore modified CodonZ to use the (currently) free software package ViennaRNA. Either package is fine and you do not need both. CodonZ will detect which package you have installed. If you have both installed, it will use UNAfold by default.

4. 1. Installation of UNAfold. Download at:
http://mfold.rna.albany.edu/?q=DINAMelt/software
or
http://dinamelt.rit.albany.edu/download.php

Install to into a directory called C:\Program Files\UNAFold
Note that this may not be the default directory! For some computers the default will be C:\Program Files (x86) \UNAFold, and you need to change the installation parameters. Specifically, CodonZ will look for this file:
C:\Program Files\UNAFold\bin\hybrid-ss-min.exe

4. 2. Installation of ViennaRNA. Download at:
http://www.tbi.univie.ac.at/RNA/index.html#download

Install to the default directory. CodonZ will look for this file:
C:\Program Files\ViennaRNA Package\RNAfold.exe

2. Codon modification

2. 1. Quick start

1. Paste the amino acid sequence you want to encode into the box on the left. If the sequence contains spaces or line breaks, press Cleanup to remove these.

If you are starting with a DNA sequence, paste it into the box on the right. Press 'Cleanup' to remove spaces, numbers and line breaks. Then press '>>> Prot' to translate to the cognate amino acid sequence.

2. Under the Organism drop-down menu on top, select Zebrafish/Mouse as appropriate.

3. Press the Optimize to produce the DNA sequence with codons optimized according to default settings. CodonZ will repeatedly run UNAfold or ViennaRNA software for secondary structure prediction, so that the taskbar and/or windows will appear to flicker on screen. This is normal and will take a couple of minutes. A progress indicator will replace the information about the current DNA sequence.

4. Copy the new DNA sequence from the box on the right and send it for synthesis.

2. 2. Overview of user interface

Elements of the main interface are highlighted in Figure 1. The two major text boxes on the left and right are for protein and nucleotide sequence respectively. The small text box below the protein box is for nucleotide sequences that are to be avoided when generating nucleotide sequence. The text box below the nucleotide sequence box will be used in future implementations for specifying regions of the nucleotide sequence that should be filled with minor codons.

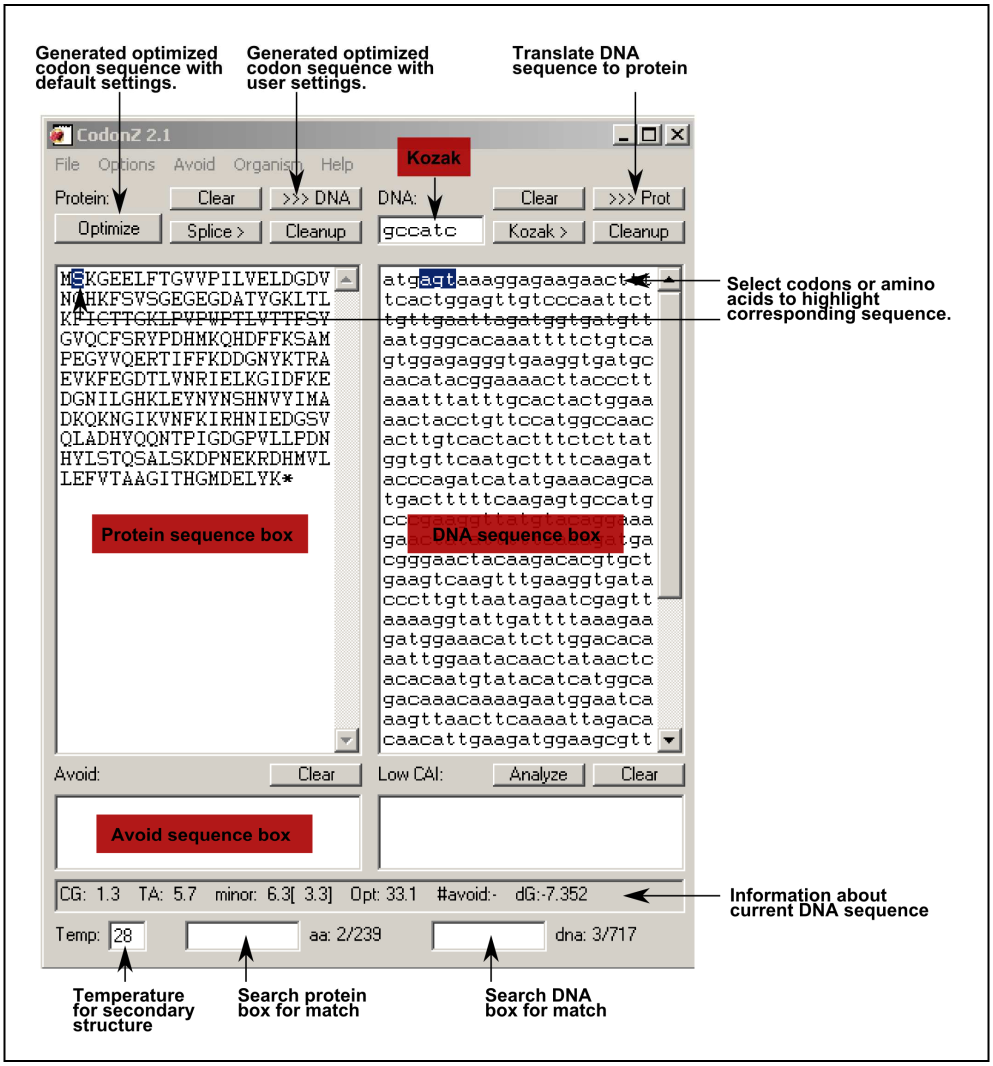
The Clear buttons above each box delete text in the corresponding box.

The Cleanup buttons remove any non-amino acid text from the left box, or any non-nucleotide text from the right box. This is useful if you are pasting sequences with white-space or numbers.

DNA sequence can be translated to protein with the >>> Prot button. The >>> DNA button generates a nucleotide sequence for the protein, using the user specified parameters.

The status bar below the Avoid and Low CAI boxes contains information about the current nucleotide sequence, in order:

- The percentage of CG dinucleotides
- The percentage of TA dinucleotides
- The percentage of minor [and rare] codons
- The percentage of optimal codons
- The number of sequences matching an entry in the Avoid box.
- The free energy of the minimum energy secondary structure for the specified Kozak and nucleotide sequence.


Generated optimized codon sequence with default settings.
Generated optimized codon sequence with user settings.
Translate DNA sequence to protein
CodonZ 2.1
File
Options
Avoid
Organism
Help
Kozak
Protein:
Clear
>>> DNA
DNA:
Clear
>>> Prot
Optimize
Splice >
Cleanup
Kozak >
Cleanup
Select codons or amino acids to highlight corresponding sequence.
Protein sequence box
DNA sequence box
Avoid:
Clear
Low CAI:
Analyze
Clear
Avoid sequence box
Information about current DNA sequence
Temp:
Temperature for secondary structure
Search protein box for match
Search DNA box for match

DNA: gccatc

```
MSKGEELFTGVVPILVELDGDV
NGHKFSVSGEGEGDATYGKLTL
KFICTTGKLPVPWPTLVTTFSY
GVQCFSRYPDHMKQHDFFKSAM
PEGYVQERTIFFKDDGNYKTRA
EVKFEGDTLVNRIELKGIDFKE
DGNILGHKLEYNYNSHNVYIMA
DKQKNGIKVNFKIRHNIEDGSV
QLADHYQQNTPIGDGPVLLPDN
HYLSTQSALSKDPNEKRDHMVL
LEFVTAAGITHGMDELYK*
```

```
atgagtaaaggagaagaacttt
tcactggagttgtcccaattct
tgttgaattagatggtgatgtt
aatgggcacaaattttctgtca
gtggagagggtgaaggtgatgc
aacatacggaaaacttaccctt
aaatttatttgcactactggaa
aactacctgttccatggccaac
acttgtcactactttctcttat
ggtgttcaatgcttttcaagat
acccagatcatatgaaacagca
tgactttttcaagagtgccatg
cccgaaggttatgtacaggaaa
gaactatatttttcaaagatga
cgggaactacaagacacgtgct
gaagtcaagtttgaaggtgata
cccttgttaatagaatcgagtt
aaaaggtattgattttaaagaa
gatggaaacattcttggacaca
aattggaatacaactataactc
acacaatgtatacatcatggca
gacaaacaaaagaatggaatca
aagttaacttcaaaattagaca
caacattgaagatggaagcgtt
```

CG: 1.3 TA: 5.7 minor: 6.3[3.3] Opt: 33.1 #avoid:- dG:-7.352

Temp: 28 aa: 2/239 dna: 3/717

Figure 1. Main user interface of CodonZ.

3. Options

Several options allow user control of codon choice for nucleotide sequence. These options are accessed via the dropdown menus at the top of the screen.

3. 1. Avoid CG/TA

When checked, CodonZ will attempt to avoid CG or TA dinucleotide sequences.

3. 2. Avoid swaps

When checked, CodonZ will maximize repeated use of same codon for a given amino acid. The first time an amino acid is specified, the optimal codon is chosen. If, in avoiding CG dinucleotides, a different codon (codon B) must be chosen, then the next time that amino acid appears, codon B will be selected rather than the optimal codon. This is a simple way to maximize tRNA cycling.

By default this option is not used for codon optimization.

3. 3. Avoid minor codons

Checking this options will force the algorithm to avoid codons which are used at less than 1% of total codon usage. Minor codons include the smaller set of 'rare' codons which are those which we operationally define as those which our analyses show are significantly avoided by highly expressed genes.

3. 5. Max folding

This option will adjust the secondary structure of the mRNA around the ATG to prevent tight folding. For this to work, you must have installed either the ViennaRNA or UNAfold package. If both are installed, select one using the Options menu. Results will not be identical but should be similar.

CodonZ performs a semi-random search to find the nucleotide sequence with the least secondary structure. Because the search is semi-random, slightly different solutions are likely to be found with repeated runs. The process takes a couple of minutes and during this time, a progress indicator replaces the DNA sequence information bar and should advance.

When you press Optimize, CodonZ goes through the following procedure.

1. Generate nucleotide sequence for the entire protein using options to avoid CG/TA dinucleotides, minor codons, restriction enzyme sites, splice sites and degradation sites. Save an internal copy of the resulting DNA sequence.

2. Keep the first 13 amino acids and delete all amino acids after those.

3. Clear all RE sites and turn off the avoid CG/TA and minor codons options.

4. Turn on the option to maximize folding energy (ie to find the minimum energy structure with the greatest free energy) and search for the best solution.

5. Replace the first 13 codons in the saved sequence with those maximizing folding energy.

3. 4. Kozak

The Kozak button simply cycles through different frequently used Kozak sequences in the Kozak box to allow you to quickly evaluate the effect of altering the Kozak sequence on the mRNA folding energy. A slightly different set of Kozak-like sequences are used by fish and mouse and the Kozak button will therefore cycle through different entries depending on the species selected.

For zebrafish, although gccatc is the most frequently used Kozak-like sequence for genes that are highly expressed, gccacc is sometimes a good choice. This sequence is still similar to the zebrafish consensus. If the start codon is followed by a codon starting with a 'G', then the sequence will be gccaccATGG, containing an NcoI site which can facilitate cloning.

3. 5. Sequences to avoid

After optimizing the sequence, CodonZ will try to remove any entry that appears in the Avoid box.

In removing these sequencing, the algorithm will avoid using minor and rare codons but introduce CG or TA dinucleotides if needed.

The most common sequences to avoid are:
- restriction sites that complicate cloning
- splice sites that may give rise to spurious transcripts

- AUUUA motifs that may destabilize RNA.

Note that the algorithm used for this process may not find an sequence that avoids all entries in the Avoid box, as in some cases, altering one codon to avoid an entry introduces a match to another entry.

3. 6. Splice

Most introns occur inside exons, rather than in untranslated regions. The Splice button assists in designing coding sequence that can accommodate the insertion of an intron that will be seamlessly excised without changing the coding sequence. Because a sequence matching the splice donor consensus is CAG.gtaagt and an good splice acceptor is cag.G, introns can be cloned into PstI (CTGCAG) sites which are in a CTGCAGG sequence

Example, target sequence is `##CTGCAGG##`

Cut with PstI gives `##CTGCA` and `GG##`
Add fragment with PstI overhangs: `Ggtaagtnnnnn...nnnnnctgca`

Ligation: `##CTGCAGgtaagtnnnnn...nnnnnctgcaGG##`
Exons are underlined, so after splicing you get `##CTGCAGG##`

Pressing the Splice will highlight amino acid sequences where you can alter the codon usage to create a CTGCAGG sequence. Edit the DNA sequence manually to create these sequences.

3. 8. Low CAI

This feature is under development. Clusters of rare codons can be needed to allow correct protein folding. You can identify regions of low codon adaptation using the 'Analyze' button and copy regions of low adaptation into the box below the DNA sequence area. During codon optimization, these sequence areas will be populated by minor codons.

4. Selecting a species specific codon use

4. 1. For zebrafish or mouse

To specify the codon usage database for analyzing DNA in the DNA window, or that will be used when translating from protein sequence, select zebrafish or mouse from the Organism menu. For these species, the codon usage is determined by our analysis of highly expressed genes. This analysis also includes which codons are designated as 'minor and rare' codons.

Because the stop codon is recognized by release factors that in many organisms have a specific 4 nucleotide preference, the DNA sequence will contain an extra base (in lowercase). Thus when optimizing using zebrafish and mouse codon frequencies, where the preference is TAAA, the DNA sequence will appear as 'TAAa'.

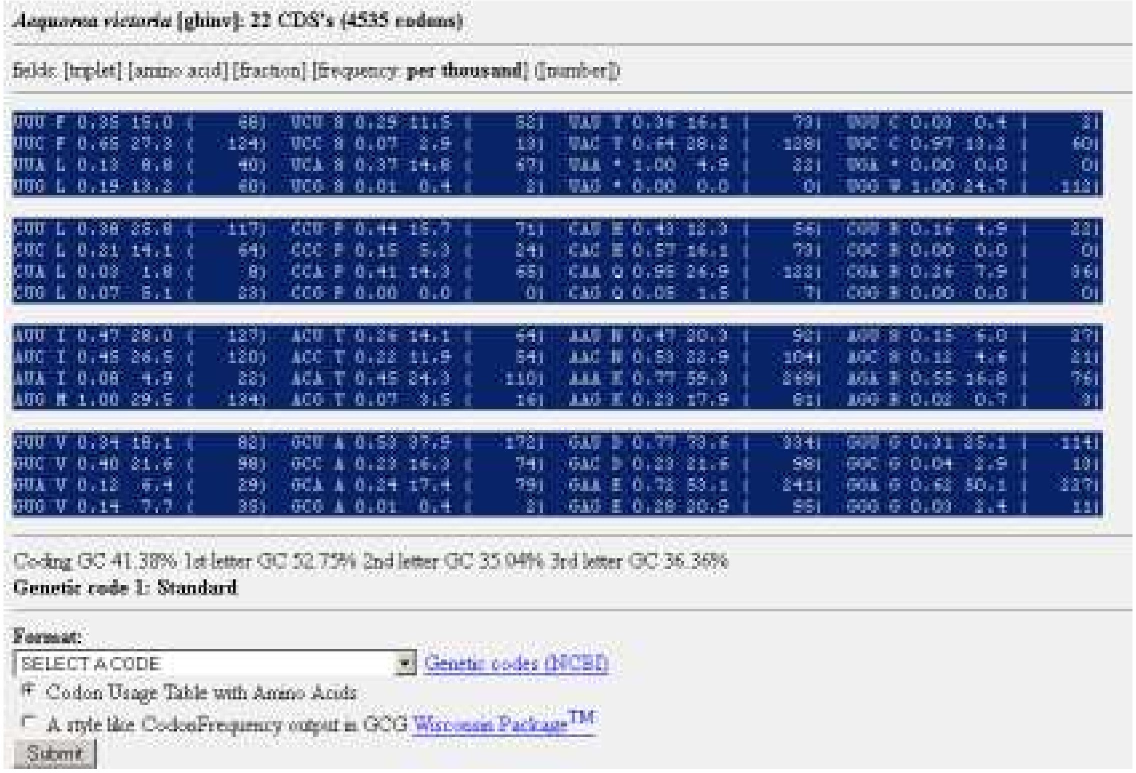
4. 2. For other organisms

For other organisms you can specify the codon usage table using these steps:

1. Find the codon usage for the organism at http://www.kazusa.or.jp/codon/

2. Under 'Format' on the webpage, select 'Standard' and 'Codon Usage Table with Amino Acids' then press Submit

3. Select the entire table from UUU to the final bottom right bracket (see image below) then copy to the clipboard using ctrl-C.



***Aequorea victoria* [gbinv]: 22 CDS's (4535 codons)**

fields: [triplet] [amino acid] [fraction] [frequency: **per thousand**] ([number])

```
UUU F 0.35 15.0 (    68)  UCU S 0.29 11.5 (    52)  UAU Y 0.36 16.1 (    73)  UGU C 0.03  0.4 (     2)
UUC F 0.65 27.3 (   124)  UCC S 0.07  2.9 (    13)  UAC Y 0.64 28.2 (   128)  UGC C 0.97 13.2 (    60)
UUA L 0.13  8.8 (    40)  UCA S 0.37 14.8 (    67)  UAA * 1.00  4.9 (    22)  UGA * 0.00  0.0 (     0)
UUG L 0.19 13.2 (    60)  UCG S 0.01  0.4 (     2)  UAG * 0.00  0.0 (     0)  UGG W 1.00 24.7 (   112)

CUU L 0.38 25.8 (   117)  CCU P 0.44 15.7 (    71)  CAU H 0.43 12.3 (    56)  CGU R 0.16  4.9 (    22)
CUC L 0.21 14.1 (    64)  CCC P 0.15  5.3 (    24)  CAC H 0.57 16.1 (    73)  CGC R 0.00  0.0 (     0)
CUA L 0.03  1.8 (     8)  CCA P 0.41 14.3 (    65)  CAA Q 0.95 26.9 (   122)  CGA R 0.26  7.9 (    36)
CUG L 0.07  5.1 (    23)  CCG P 0.00  0.0 (     0)  CAG Q 0.05  1.5 (     7)  CGG R 0.00  0.0 (     0)

AUU I 0.47 28.0 (   127)  ACU T 0.26 14.1 (    64)  AAU N 0.47 20.3 (    92)  AGU S 0.15  6.0 (    27)
AUC I 0.45 26.5 (   120)  ACC T 0.22 11.9 (    54)  AAC N 0.53 22.9 (   104)  AGC S 0.12  4.6 (    21)
AUA I 0.08  4.9 (    22)  ACA T 0.45 24.3 (   110)  AAA K 0.77 59.3 (   269)  AGA R 0.55 16.8 (    76)
AUG M 1.00 29.5 (   134)  ACG T 0.07  3.5 (    16)  AAG K 0.23 17.9 (    81)  AGG R 0.02  0.7 (     3)

GUU V 0.34 18.1 (    82)  GCU A 0.53 37.9 (   172)  GAU D 0.77 73.6 (   334)  GGU G 0.31 25.1 (   114)
GUC V 0.40 21.6 (    98)  GCC A 0.23 16.3 (    74)  GAC D 0.23 21.6 (    98)  GGC G 0.04  2.9 (    13)
GUA V 0.12  6.4 (    29)  GCA A 0.24 17.4 (    79)  GAA E 0.72 53.1 (   241)  GGA G 0.62 50.1 (   227)
GUG V 0.14  7.7 (    35)  GCG A 0.01  0.4 (     2)  GAG E 0.28 20.9 (    95)  GGG G 0.03  2.4 (    11)
```

Coding GC 41.38% 1st letter GC 52.75% 2nd letter GC 35.04% 3rd letter GC 36.36%
Genetic code 1: Standard

Format:

SELECT A CODE ▾ Genetic codes (NCBI)

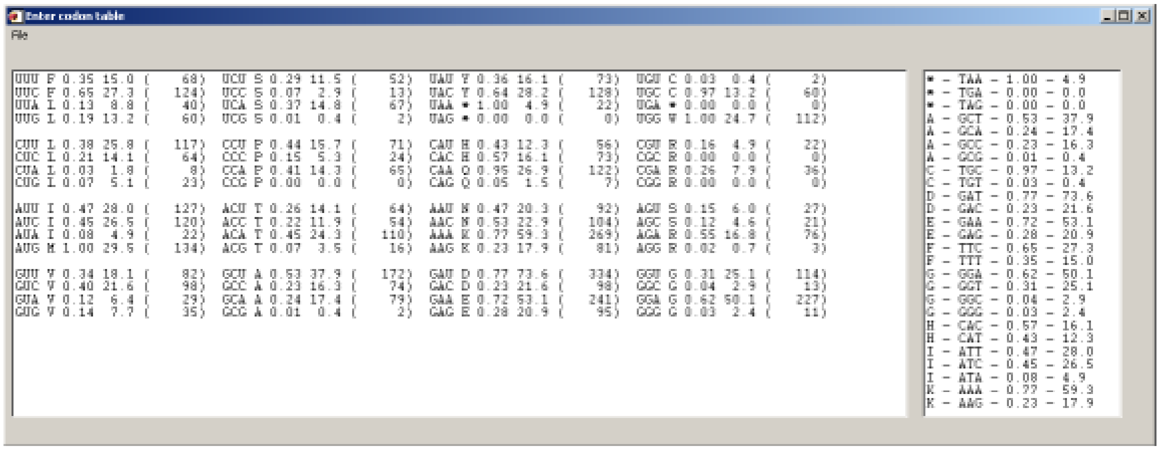
● Codon Usage Table with Amino Acids

○ A style like CodonFrequency output in GCG Wisconsin Package™

Submit

4. In CodonZ, select Organism → Enter New

5. Paste the table into the left hand box. If all goes well, you'll see the table reproduced in the right hand box:



```
UUU F 0.35 15.0 (    68)  UCU S 0.29 11.5 (    52)  UAU Y 0.36 16.1 (    73)  UGU C 0.03  0.4 (     2)
UUC F 0.65 27.3 (   124)  UCC S 0.07  2.9 (    13)  UAC Y 0.64 28.2 (   128)  UGC C 0.97 13.2 (    60)
UUA L 0.13  8.8 (    40)  UCA S 0.37 14.8 (    67)  UAA * 1.00  4.9 (    22)  UGA * 0.00  0.0 (     0)
UUG L 0.19 13.2 (    60)  UCG S 0.01  0.4 (     2)  UAG * 0.00  0.0 (     0)  UGG W 1.00 24.7 (   112)

CUU L 0.38 25.8 (   117)  CCU P 0.44 15.7 (    71)  CAU H 0.43 12.3 (    56)  CGU R 0.16  4.9 (    22)
CUC L 0.21 14.1 (    64)  CCC P 0.15  5.3 (    24)  CAC H 0.57 16.1 (    73)  CGC R 0.00  0.0 (     0)
CUA L 0.03  1.8 (     8)  CCA P 0.41 14.3 (    65)  CAA Q 0.95 26.9 (   122)  CGA R 0.26  7.9 (    36)
CUG L 0.07  5.1 (    23)  CCG P 0.00  0.0 (     0)  CAG Q 0.05  1.5 (     7)  CGG R 0.00  0.0 (     0)

AUU I 0.47 28.0 (   127)  ACU T 0.26 14.1 (    64)  AAU N 0.47 20.3 (    92)  AGU S 0.15  6.0 (    27)
AUC I 0.45 26.5 (   120)  ACC T 0.22 11.9 (    54)  AAC N 0.53 22.9 (   104)  AGC S 0.12  4.6 (    21)
AUA I 0.08  4.9 (    22)  ACA T 0.45 24.3 (   110)  AAA K 0.77 59.3 (   269)  AGA R 0.55 16.8 (    76)
AUG M 1.00 29.5 (   134)  ACG T 0.07  3.5 (    16)  AAG K 0.23 17.9 (    81)  AGG R 0.02  0.7 (     3)

GUU V 0.34 18.1 (    82)  GCU A 0.53 37.9 (   172)  GAU D 0.77 73.6 (   334)  GGU G 0.31 25.1 (   114)
GUC V 0.40 21.6 (    98)  GCC A 0.23 16.3 (    74)  GAC D 0.23 21.6 (    98)  GGC G 0.04  2.9 (    13)
GUA V 0.12  6.4 (    29)  GCA A 0.24 17.4 (    79)  GAA E 0.72 53.1 (   241)  GGA G 0.62 50.1 (   227)
GUG V 0.14  7.7 (    35)  GCG A 0.01  0.4 (     2)  GAG E 0.28 20.9 (    95)  GGG G 0.03  2.4 (    11)
```

```
* - TAA - 1.00 - 4.9
* - TGA - 0.00 - 0.0
* - TAG - 0.00 - 0.0
A - GCT - 0.53 - 37.9
A - GCA - 0.24 - 17.4
A - GCC - 0.23 - 16.3
A - GCG - 0.01 - 0.4
C - TGC - 0.97 - 13.2
C - TGT - 0.03 - 0.4
D - GAT - 0.77 - 73.6
D - GAC - 0.23 - 21.6
E - GAA - 0.72 - 53.1
E - GAG - 0.28 - 20.9
F - TTC - 0.65 - 27.3
F - TTT - 0.35 - 15.0
G - GGA - 0.62 - 50.1
G - GGT - 0.31 - 25.1
G - GGC - 0.04 - 2.9
G - GGG - 0.03 - 2.4
H - CAC - 0.57 - 16.1
H - CAT - 0.43 - 12.3
I - ATT - 0.47 - 28.0
I - ATC - 0.45 - 26.5
I - ATA - 0.08 - 4.9
K - AAA - 0.77 - 59.3
K - AAG - 0.23 - 17.9
```

6. Select File → Quit. Check that the correct codon usage table is loaded under Organism → Display.

5. Editing and proofing sequences

5. 1. Editing sequences

Gene synthesis companies can not generate every sequence. Manual editing of the sequence may be required to re-engineer regions flagged as having too much repetitive sequence or secondary structure.

To assist with this, use the Organism → Display function to open a window showing the codon frequency table currently in use. This table shows the amino acid, codon and relative synonymous codon usage frequency (RSCU). Codons where the RSCU is followed by a ! symbol are minor codons and should be avoided.

Find the nucleotide sequence that needs modifying using the nucleotide search box at the bottom of the window. This will highlight the corresponding amino acids in the protein window.

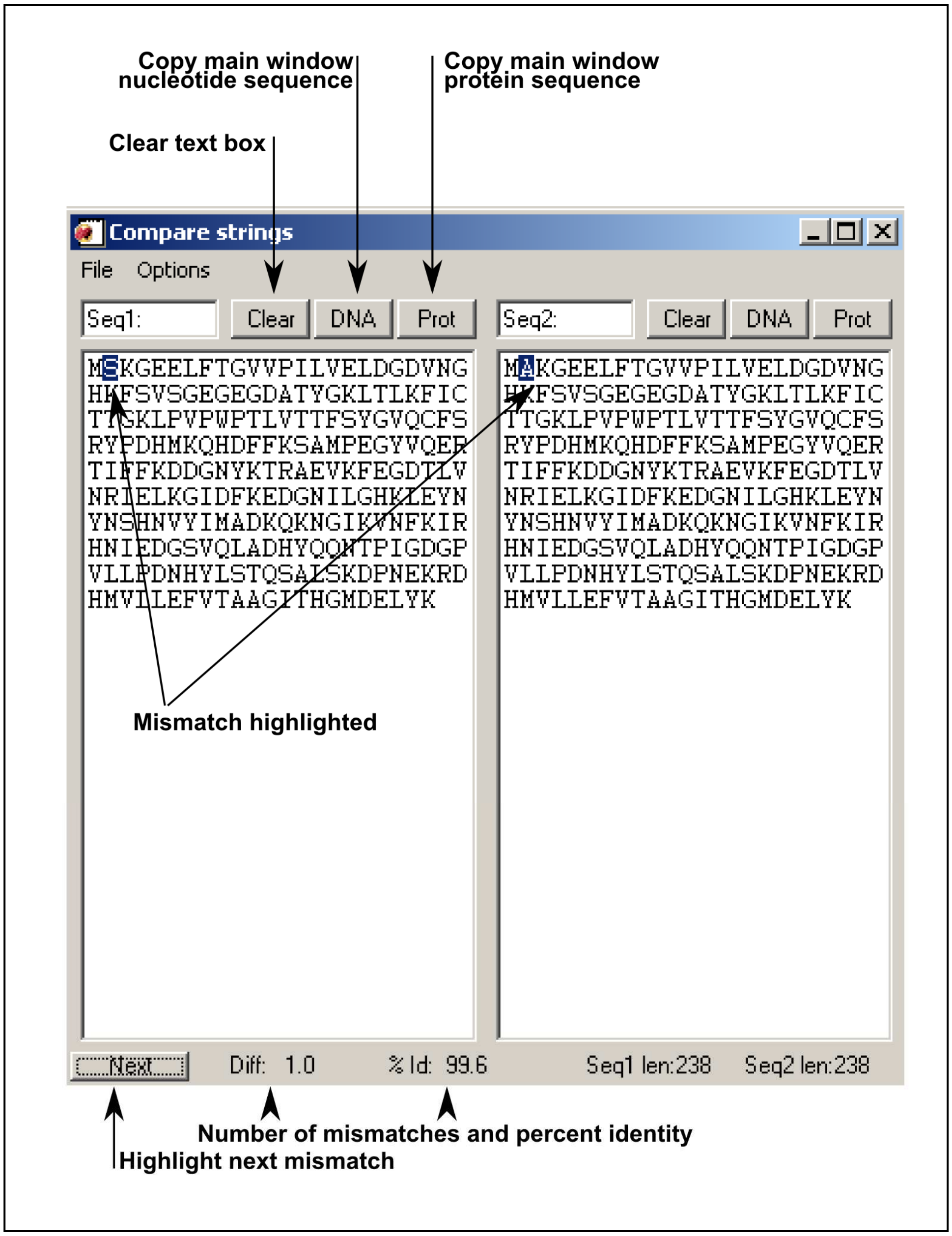
To be sure that you are editing in the correct frame, select a single amino acid in the protein window. This now highlights the corresponding codon in the right window. Choose a new codon from the window showing the codon frequency table.

Populate the Avoid box with entries from the Avoid drop down menu to ensure that you did not inadvertently create a splice sites or degradation sequence.

Finally, double-check that the new nucleotide sequence encodes the correct protein. The tool in the next section can assists with this.

5. 2. Checking sequences

To check the resulting sequences, open the File->Compare utility window. This allows users to copy nucleotide or protein sequences from the main window, and compare two sequences for mismatches. Press the 'DNA' or 'Prot' button above each of the text boxes to copy the sequence from the main window to the corresponding box.



Copy main window nucleotide sequence

Copy main window protein sequence

Clear text box

Seq1:

```
MSKGEELFTGVVPILVELDGDVNG
HKFSVSGEGEGDATYGKLTLKFIC
TTGKLPVPWPTLVTTFSYGVQCFS
RYPDHMKQHDFFKSAMPEGYVQER
TIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYN
YNSHNVVIMADKQKNGIKVNFKIR
HNIEDGSVQLADHYQQNTPIGDGP
VLLPDNHYLSTQSALSKDPNEKRD
HMVLLEFVTAAGITHGMDELYK
```

Seq2:

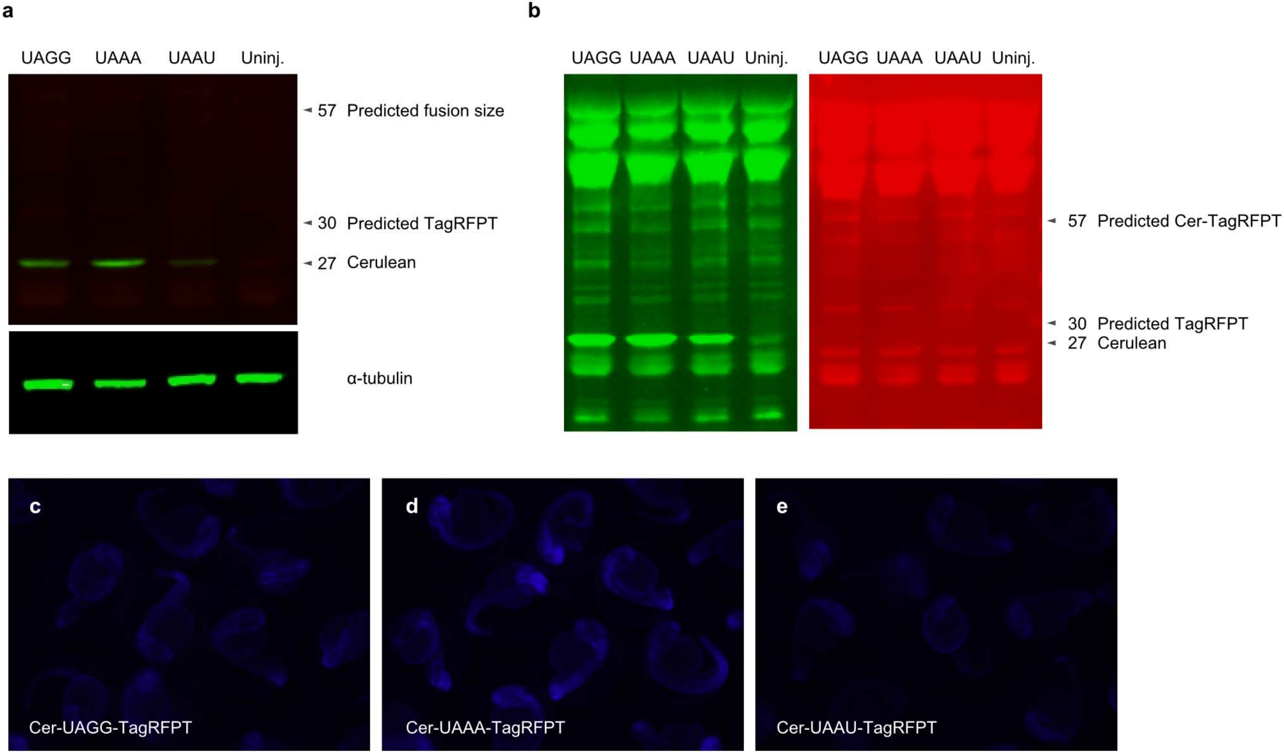
```
MAKGEELFTGVVPILVELDGDVNG
HKFSVSGEGEGDATYGKLTLKFIC
TTGKLPVPWPTLVTTFSYGVQCFS
RYPDHMKQHDFFKSAMPEGYVQER
TIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYN
YNSHNVVIMADKQKNGIKVNFKIR
HNIEDGSVQLADHYQQNTPIGDGP
VLLPDNHYLSTQSALSKDPNEKRD
HMVLLEFVTAAGITHGMDELYK
```

Mismatch highlighted

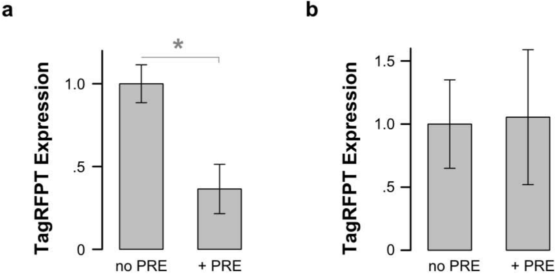
Diff: 1.0 % Id: 99.6 Seq1 len:238 Seq2 len:238

Number of mismatches and percent identity

Highlight next mismatch



Supplementary Figure 6



Supplementary figure 7

Experiment	Plasmid	RNA/ DNA	Conc. (ng/µL)	Line Injected	Protein Collected (dpf)
Cer mod.	pCS2 Cer	RNA	100	w.t.	1
	pCS2 Cer.zf1	RNA	100	w.t.	1
TagRFPT mod.	pCS2 TagRFPT	RNA	10	w.t.	1
	pCS2 TagRFPT.zf1	RNA	10	w.t.	1
Cre Mod.	pCS2 Cre	RNA	10	bActin:floxGFP-lyntRFP	3
	pCS2 Cre.zf1	RNA	10	bActin:floxGFP-lyntRFP	3
Gal4ff Mod.	pCS2 Gal4ff	RNA	5	14xUAS:BGi-NLS-emGFP	1
	pCS2 Gal4ff.zf1	RNA	5	14xUAS:BGi-NLS-emGFP	1
NfsB Mod.	pCS2 NfsB.zf1	RNA	50	w.t.	NA
	pCS2 NfsB.zf1	RNA	50	w.t.	NA
Tol1 Mod.	pCS2 tol1	RNA	80	w.t.	5
	pCS2 tol1.zf1	RNA	80	w.t.	5
	Coinjected: bActin:floxGFP-lynTagRFPT	DNA	20		
Intron	14xUAS:GCaMP3-v2a-mCherry	DNA	20	Et(SCP1:Gal4)y271	5
	14xUAS:BGint-GCaMP3-v2a-mCherry	DNA	20	Et(SCP1:Gal4)y271	5
	HuC:Cer	DNA	20	w.t.	5
	HuC:UBCint-Cer	DNA	20	w.t.	5
	HuC:ZGCint-Cer	DNA	20	w.t.	5
	Coinjected: tol1 transposase	RNA	80		
Stop tetra.	pCS2 Cer-TAAT-TagRFPT	RNA	100	w.t.	1
	pCS2 Cer-TAAA-TagRFPT	RNA	100	w.t.	1
	pCS2 Cer-TAGG-TagRFPT	RNA	100	w.t.	1
PolyA	pCS2 TagRFPT-SV40	RNA	30	w.t.	1
	pCS2 TagRFPT-p10 UTR	RNA	30	w.t.	1
	pCS2 TagRFPT-afp UTR	RNA	30	w.t.	1
	pCS2 TagRFPT-bGlobin UTR	RNA	30	w.t.	1
	pCS2 TagRFPT-rps26 UTR	RNA	30	w.t.	1
	pCS2 TagRFPT-gnb2l1 UTR	RNA	30	w.t.	1
	pSP64T TagRFPT	RNA	30	w.t.	1
Combinations	HuC:Cer-sv40	DNA	10	w.t.	7
	HuC:UBCint-Cer-sv40	DNA	10	w.t.	7
	HuC:UBCint-Cer-afp	DNA	10	w.t.	7
	HuC:UBCint-Cer-bGlobin UTR	DNA	10	w.t.	7
	HuC:UBCint-Cer.zf1-sv40	DNA	10	w.t.	7
	Coinjected: tol1 transposase	RNA	80		
PRE mRNA	pCS2 TagRFPT	RNA	10	w.t.	1
	pCS2 TagRFPT-PRE	RNA	10	w.t.	1
PRE transgene	14xUAS:BGi-NTR-TagRFPT	DNA	20	HuC: gal4	5
	14xUAS:BGi-NTR-TagRFPT-PRE	DNA	20	HuC: gal4	5
	Co-injected: tol1 transposase	RNA	80		

Supplementary Table 1.

	Cer		Cre		Gal4FF		NfsB		TagRFPT		Tol1	
	Std.	Mod.	Std.	Mod.	Std.	Mod.	Std.	Mod.	Std.	Mod.	Std.	Mod.
Optimal codon frequency	70.0	79.2	33.4	82.0	39.4	76.1	35.3	74.8	68.2	79.2	38.4	81.7
Minor codon frequency	4.2	0.0	18.6	0.6	12.8	0.0	13.8	0.0	1.6	0.0	9.2	0.8
Rare codon frequency	0.0	0.0	3.5	0.6	6.7	0.0	1.8	0.0	0.0	0.0	5.2	0.0
Initiator structure (kcal/mol)	-23.1	-5.8	-2.2	1.0	-7.2	-5.1	-7.0	-5.6	-8.2	-3.0	-6.6	-1.3
Stop tetranucleotide	UAAu	UAAa	UAGg	UAAa	UAAu	UAAa	UAAu	UGAu	UGAg	UAAa	UAGu	UAAa
Kozak-like sequence	gccacc	gccacc	gccacc	gccacc	gccacc	gccacc	gccacc	gccacc	gccacc	gccacc	uaaaau	gccatc
CG frequency	8.8	0.1	6.4	0.1	4.1	0.0	7.7	0.0	4.8	0.0	1.4	0.4
TA frequency	1.7	2.9	4.7	2.4	5.2	1.9	3.5	3.7	2.7	3.1	4.1	1.1

Supplementary Table 2.