Gene- and genome-based analysis of significant codon patterns in yeast, rat and mice genomes with the CUT Codon UTilization tool

Citation: Doyle, Francis et al. "Gene- and Genome-Based Analysis of Significant Codon Patterns in Yeast, Rat and Mice Genomes with the CUT Codon UTilization Tool." Methods 107 (September 2016): 98–109 © 2016 Elsevier Inc

As Published: http://dx.doi.org/10.1016/J.YMETH.2016.05.010

Publisher: Elsevier BV

Persistent URL: http://hdl.handle.net/1721.1/117601

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

HHS Public Access

Author manuscript

Methods. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

Methods. 2016 September 1; 107: 98–109. doi:10.1016/j.ymeth.2016.05.010.

Gene- and Genome-Based Analysis of Significant Codon Patterns in Yeast, Rat and Mice Genomes with the CUT Codon Utilization Tool

Frank Doyle[#1], **Andrea Leonardi**[#1], **Lauren Endres**[#2], **Scott A. Tenenbaum**[1], **Peter C. Dedon**[3,4], and **Thomas J. Begley**[1,5,#]

[1]State University of New York – SUNY Polytechnic Institute, College of Nanoscale Science and Engineering, Albany, NY

[2]State University of New York – SUNY Polytechnic Institute, College of Arts and Sciences, Utica, NY

[3]Department of Biological Engineering and Center for Environmental Health Science, Massachusetts Institute of Technology, Cambridge, MA

[4]Singapore-MIT Alliance for Research and Technology, Singapore

[5]RNA Institute, University at Albany, State University of New York

[#] These authors contributed equally to this work.

Abstract

The translation of mRNA in all forms of life uses a three-nucleotide codon and aminoacyl-tRNAs to synthesize a protein. There are 64 possible codons in the genetic code, with codons for the ~20 amino acids and 3 stop codons having 1- to 6-fold degeneracy. Recent studies have shown that families of stress response transcripts, termed modification tunable transcripts (MoTTs), use distinct codon biases that match specifically modified tRNAs to regulate their translation during a stress. Similarly, translational reprogramming of the UGA stop codon to generate selenoproteins or to perform programmed translational read-through (PTR) that results in a longer protein, requires distinct codon bias (*i.e.*, more than one stop codon) and, in the case of selenoproteins, a specifically modified tRNA. In an effort to identify transcripts that have codon usage patterns that could be subject to translational control mechanisms, we have used existing genome and transcript data to develop the gene-specific Codon UTilization (CUT) tool and database, which details all 1-, 2-, 3-, 4- and 5-codon combinations for all genes or transcripts in yeast (*Saccharomyces cerevisiae*), mice (*Mus musculus*) and rats (*Rattus norvegicus*). Here, we describe the use of the CUT tool and database to characterize significant codon usage patterns in specific genes and

[#]Correspondence should be addressed to Thomas Begley, SUNY Polytechnic Institute, College of Nanoscale Science and Engineering, State University of New York, tbegley@sunypoly.edu.

Author Contributions The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

groups of genes. In yeast, we demonstrate how the CUT database can be used to identify genes that have runs of specific codons (e.g., AGA, GAA, AAG) linked to translational regulation by tRNA methyltransferase 9 (Trm9). We further demonstrate how groups of genes can be analyzed to find significant dicodon patterns, with the 80 Gcn4-regulated transcripts significantly ($P < 0.00001$) over-represented with the AGA-GAA dicodon. We have also used the CUT database to identify mouse and rat transcripts with internal UGA codons, with the surprising finding of 45 and 120 such transcripts, respectively, which is much larger than expected. The UGA data suggest that there could be many more translationally reprogrammed transcripts than currently reported. CUT thus represents a multi-species codon-counting database that can be used with mRNA-, translation- and proteomics-based results to better understand and model translational control mechanisms.

Keywords

anticodon; codon bias; gene expression; modification tunable transcripts; stress response; translation; RNA modification

1. Introduction

Translation is the process by which proteins are synthesized from messenger RNA (mRNA) in the ribosome, using both transfer RNAs (tRNAs) and other trans-acting factors. The codons in mRNA consist of three nucleotides and interact with the tRNA anticodon during translation to specify the insertion of one of ~20 amino acids in a growing peptide chain, or in the case of stop codons, to terminate translation. tRNA molecules consist of 70 to 90 nucleotides folded into a cloverleaf structure containing an anticodon loop and a 3'-CAA that is covalently linked to a specific amino acid. tRNA is initially transcribed to contain adenosine (A), guanosine (G), uridine (U) and cytidine (C), but it is the most heavily modified nucleic acid in the cell and contains an average of 9–11 post-transcriptional modifications at >30 conserved sites [1, 2]. These RNA modifications have been shown play important roles in promoting tRNA stability, optimizing codon-anticodon interactions and preventing translation errors [3, 4].

In mRNA there are 64 possible three-nucleotide combinations that comprise the full set of codons. AUG is generally used as the start codon and the translation initiation site, but UUG, GUG and CUG have also been identified as occasional start sites in some transcripts in several organisms [5]. The ribosome has three sites associated with bound tRNA: one where the aminoacyl-tRNA enters (A site), one that has the tRNA bound to the growing polypeptide chain (P site), and lastly an E site where tRNA exits the ribosome. The polypeptide chain is catalytically transferred from the tRNA at the P site to the tRNA at the A site, at which point the free tRNA exits through the E site and the ribosome moves along to the next codon [6]. Excluding stop codons, there are 61 possible codons that can occupy each site or more than 226,918 three-codon combinations that can associate with A-, P- and E-sites. Translation elongation continues until the ribosome encounters one of three stop codons, UAA, UAG or UGA. There are instances, described below, where a UGA stop codon, in conjunction with trans- and cis-acting factors, can signal for the addition of a

unique amino acid and these instances highlight the use of codon-based translational regulatory mechanisms.

The regulation of gene expression by specific patterns of codon usage has been established for both prokaryotic and eukaryotic systems [7–12]. In bacterial systems where transcription is coupled to translation, the expression of tryptophan metabolic enzymes from the Trp operon is linked to the translation of the UGG-UGG dicodon, where UGG codes for tryptophan, in a short leader peptide. The transcription and subsequent translation of tryptophan metabolic enzymes requires pausing at the UGG-UGG dicodon due to low levels of charged tryptophan tRNA, which promotes specific folding of the upstream transcript (the attenuator) to promote the transcription of the entire Trp operon [7]. Translational regulation has also been demonstrated in budding yeast for some stress response genes over-represented with codons linked to leucine, arginine and glutamic acid, by the regulated levels of modified cytidines and uridines in the wobble position of the tRNA anticodon. tRNA methyltransferase (Trm) 4 and Trm9 complete the formation of 5-methylcytidine (m^5C) and 5-methoxycarbonylmethyluridine (mcm^5U) and 5-methoxycarbonylmethyl-2-thiouridine (mcm^5s^2U), respectively, in the anticodon of tRNAs specific for Leu, Arg and Glu [8, 9, 13], Further, the levels of m^5C (H_2O_2, ROS agents), mcm^5U (HU, MMS, alkylating agents) and mcm^5s^2U (HU, MMS, alkylating agents) modifications have been shown to change in response to stress [8, 9, 14, 15]. Reporter, targeted, and systems-based studies comparing transcripts and proteins have demonstrated that the translation of UUG (Leu), AGA (Arg) and GAA (Glu) codons are dependent on specific wobble base modifications and that some stress response genes that over-use these codons have decreased translation in the absence of the corresponding Trm and wobble base modification [9, 10, 14, 16].

Links between the regulation of translation and distinct codon usage patterns has also been demonstrated for mammalian transcripts that use UGA stop codon reprogramming or programmed translational read-through (PTR). Stop codon recoding is used to incorporate selenocysteine into selenoproteins, with the corresponding activities involved in ROS detoxification, selenium utilization and thyroid function [12, 17]. Selenocysteine has been referred to as the 21st amino acid, with no dedicated triplet codon [18, 19]. As such, decoding for selenocysteine is unconventional and requires a “recoding” of the UGA stop codon for incorporation in a process termed stop codon reprogramming [17, 19, 20]. Similar to the yeast Trm9 example noted earlier, stop codon recoding utilizes specifically modified wobble uridine bases, mcm^5U and 5-methoxycarbonylmethyl-2′-O-methyluridine (mcm^5Um), to promote optimal anticodon-codon interactions [17, 21, 22], as well as 3'-UTR regulatory sequence and trans-acting factors. The 3'-UTR found in transcripts encoding selenoproteins contains a selenocysteine insertion sequence (SECIS) that helps identify the internal UGA codons being re-coded for selenocysteine [23, 24]. PTR has also been shown to occur at UGA stop codons, as in the recent notable example of human vascular endothelial growth factor-A (*VEGFA*) isoform VEGF-Ax, in which Ax denotes an extended form [25]. VEGFA encodes a growth factor that promotes endothelial cell migration and growth, with roles in angiogenesis and tumorigenesis, among others [26]. VEGFA contains another stop codon that is 22 codons (66 nts) down from the first UGA stop codon, with the nucleotide sequence well conserved in humans, rats and mice. PTR of the first UGA stop

codon inserts serine and promotes the formation of a 22 amino acid extended version (VEGF-Ax), with PTR requiring a cis-acting element found in the Ax region that is 63 nts downstream of the canonical stop codon and corresponds to an hnRNP1 A2/B1 element [25]. With similarities to stop codon recoding, PTR uses cis-acting sequences, but the precise tRNA involved in PTR has not been clearly identified. PTR has also been shown for Ago1 and Mtch2 mRNA, and is predicted for Nr1d1, Adamts4 and Tox [25]. Other examples of translational read-through (Mdh1 and LdhB) have been reported to include stop codons outside of UGA, with targeted and genome-wide approaches identifying real and potential targets in humans, fruit flies, yeast and viruses [27].

Codon usage is one of the many parameters involved in the translational regulation of specific transcripts. Monocodon usage data is readily available to describe genome-based trends [28] and, in the case of yeast, to describe individual genes [13]. Bacterial studies on the Trp and other operons for other metabolic processes have demonstrated that dicodons in individual genes can have regulatory effects [29]. To investigate the potential regulatory effects of gene-specific dicodons and other codon combinations, we have developed a Codon UTilization (CUT) tool and database to analyze all genes or transcripts banked for a species, and we applied this tool to the analysis of yeast, mouse and rat genome sequence data (http://pare.sunycnse.com/cut/index.jsp). We demonstrate in yeast that specific dicodon and quadruple codon runs can be linked to optimal translation by Trm9-catalyzed tRNA. Further, we show that the AGA-GAA dicodon is over-represented in 80 Gcn4-regulated transcripts, which provides a novel link between the regulation of translation initiation in the Gcn2-eIF2α-Gcn4 pathway and regulation of translational elongation by Trm9. Finally, we demonstrate how CUT can be used as an exploratory tool in identifying 45 mouse and 120 rat transcripts that contain an internal UGA codon, which greatly expands the potential mammalian targets that could be undergoing stop codon recoding or some form of PTR.

2. Materials and Methods

2.1 CUT tool algorithms and database design

The CUT tool was implemented using a standard “Model View Controller” (MVC) software design pattern as a Java 2 Enterprise Edition (J2EE) package deployed on a JBoss application server. The “Model” is comprised of a MySQL relational database and Enterprise Java Bean (EJB), the “View” is provided by Java Server Pages (JSPs), and the “Controller” is formed by Java Servlets. Offline Java code was written to populate the database by parsing GTF annotation files (obtained from SGD and the UCSC table browser) and using this data to assemble coding sequences for each protein coding gene or transcript from the relevant genome sequences. The coding sequences then undergo a brief quality “control” check (e.g., ensure complete reading frame and accepted start codons). Accepted coding sequences were then analyzed for specific codon and codon combination usage. Codon sequence bias (for every contiguous combination of length 1 to 5) in each transcript is calculated as a Z-score (number of standard deviations above/below the population mean) by comparison to the full population of transcripts within a particular annotation set. The procedure for calculating the expected codon frequency for each gene/transcript was based on the observed global frequency of each mono-, di-, tri-, quadra- and quint-codon sequence

in the set of all gene/transcripts in a given species. In addition, we treated all alternative isoforms of a given gene as a separate transcript, which resulted in the exon sequences found in each isoform being counted. We reasoned that multiple isoforms could be present in the cell at the same time, to support our use of codon sequences specific to each isoform as separate variables that compose the expected frequency value. Z-scores are a simple measure of whether a gene/transcript is using a codon sequence more or less then the global frequency. Specifically Z-scores detail how many standard deviations above or below the global frequency the specific codon sequence is used in the gene/transcript, which can be misleading when dealing with expected frequencies that approach zero. For example the Z-score measures for quintuplet codon combinations should be used with caution as the expected frequency of all quintuplet codons approaches zero and a single use in a gene/transcript results in a Z-score > 100. We note that the number of times each quintuplet codon is used in a gene is detailed in the CUT database and these can be more informative measures then Z-scores. The online tool does not support right clicking due its use of dynamically generated gene/transcript pages.

2.2 Gene specific codon and dicodon analysis

Species-containing genes or transcript data stored in the CUT database can be accessed using the search tab found at http://pare.sunycnse.com/cut/controller?action=search. The organism, genome, annotation and gene of interest must be specified in the pull down menu, with the organism designation leading to single genome and annotation choices. The gene/tx pull-down can be searched with partial or full gene names or transcript (NM_numbers) identifiers using the search gene button. Results on each search are displayed below the search gene button and their CUT data can be accessed by clicking on the gene/tx hyperlink, which will bring users to a page that displays the gene sequence and provides access to monocodon to quintuplecodon data and graphs for each gene. In addition there is a download spreadsheet tab for each gene that allows for the export of all monocodon and dicodon data that is linked to the gene-specific codon sequence, Z-score, expected frequency and actual frequency. The resulting download is produced in tab-delimited form. Resulting data can be imported into a graphing program (*i.e.*, Excel) and data can be organized to identify the most and least used moncodons and dicodons in the gene sequence. On each gene/transcript webpage the hyperlinks detailing the monocodon to quintupletcodons used in each gene can be clicked to bring the user to bar graphs detailing the sequence, number of times used and Z-score.

2.3 Gene codon painting and immunoblot analysis

Codon *painting* can be performed using the codon search input box found on each gene/transcript page tied to the database. Once the specific codon pattern is searched (using the input field to specify sequence, search button to search and then highlight selected tab to paint), resulting output is highlighted to paint sequence occurrences in each target sequence. Protein expression analysis can be performed in the organism of choice. In the past in mouse and yeast models reported here we have analyzed specific proteins in the presence or absence of specific anticodon wobble base modifications [8–11, 13, 30]. For example, in yeast specific protein levels in wild-type (By4741) and *trm9*Δ cells were determined as previously described using corresponding C-terminal TAP tagged strains [31] and

immunoblots. Briefly, each TAP-tag strain was made *trm9*Δ using a PCR amplified cassette derived from YML014W, with selection occurring on G418. Protein extracts from each strain were then analyzed by immunoblots using an anti-TAP antibody, with equal loading probed using an anti-β-tubulin antibody, as previously described [13].

2.4 CUT downloads, gene parsing and dicodon statistical analysis

Using the Genomes tab, users can access a page that allows for the download of all gene-specific moncodon and dicodon data for yeast, mouse and rat genes/transcripts. Data on genome wide monocodon and dicodon frequencies for all species-containing genes/transcripts can be downloaded from the CUT database (http://pare.sunycnse.com/cut/controller?action=listAnnot) by clicking on the Download Stats button. The Heatmap Files icon that is shown on each gene/transcript page can be used to download all monocodon and dicodon Z-scores for all genes/transcripts in a specific organism. The resulting download is a compressed folder that when extracted will provide a tab-delimited file detailing Z-scores for all monocodons and dicodons used in each gene. Outside of CUT these lists can be used to compile sets of genes (*i.e.*, visual basic programmed search of gene list in Excel) and, by using random sampling approaches, determine if summed Z-scores for sets of genes are higher than average values. For example, visual basic based scripts can be programmed in Excel to randomly order all yeast genes, pick 80 random genes to obtain and then sum dicodon Z-score values, with this process performed for N = 300 occurrences. The average value and standard deviation for the summed Z-score can then be compared to an actual value to determine a measure of significance (Z-score), similar to our previously described studies [13, 32].

2.5 Multi-codon searches to identify species-containing genes with specific in frame sequences

Multi-codon searches to identify species-containing genes with specific sequence can also be performed in the CUT database using the search tab (http://pare.sunycnse.com/cut/controller?action=search). The organism, genome and annotation information must be specified in the pull down menu, with the organism designation only providing a single genome and annotation choice. The “containing codon sequence” input can be specified with a codon string (*i.e.*, UGANNN, with N being either U, G, A or C), followed by a click on the search gene button. Results of each search are displayed below the search gene button and their CUT data can be accessed by clicking on the gene/tx hyperlink, which will bring users to a page that displays gene sequence and other information, as described above, for each gene.

3. Results and Discussion

The CUT algorithm was developed to methodically count the codon content in gene or transcript sequences beginning with the start codon. To facilitate bulk gene analysis, data was imported (Fig. 1) as either gene annotations (SGD, 6664 yeast genes) or as RefSeq annotated transcripts (30,392 mouse and 16,711 rat transcripts). A key difference between the yeast and mammalian data imports is that genes are represented as single entities, while multiple transcripts are present for specific genes, respectively, with the latter including

splice variants. Each entry was methodically analyzed to catalog the number of each possible monocodon as well as each possible dicodon (64×64), tricodon (64×64×64), quadruplecodon (64×64×64×64) and quintuplecodon (64×64×64×64×64) combinations (http://pare.sunycnse.com/cut/). For each gene/transcript, there are 64 to over 1 billion possible features to catalog and store, but no gene had codon sequences that represent the entire combination space. The CUT database was designed using a table based format (Fig. 2) and allows for individual gene queries and the retrieval of all of the codon combination data noted earlier, and describes whether a codon or codon combination is over-represented in a gene sequence using a Z-score metric. Search functions that are associated with CUT include queries based on Z-score cutoffs, codon combinations, and gene identity. Bulk download of codon usage frequencies and Z-scores for all monocodon and dicodon combinations is enabled for individual genes as well as for all genes or transcripts specific to each species.

We have previously reported that the yeast translation elongation factor 3 (*YEF3*) transcript is translationally regulated by Trm9-catalyzed uridine modifications on tRNA and that *YEF3* overuses GAA (91 total count, 3.4 Z-score) and AGA (41 total count, 3.6 Z-score) codons [13, 33]. In *trm9*Δ cells, we have shown that Yef3 protein levels are dramatically reduced, with little difference observed in transcript levels [10, 13]. *YEF3* was used to query the database (Fig. 3A) to demonstrate results that report monocodon usage patterns in bar graph format (Fig. 3B). In addition, *YEF3* was searched to detail dicodon usage pattern numbers (Fig. 3C), such as GAA-GAA, which is represented 8 times and has a Z-score of 3.2. Z-scores that detail pattern usage can be exported for all represented dicodons in a gene. For example, data specific to *YEF3* have been exported, sorted from low to high, and plotted in Fig. 3D. Dicodons overused in *YEF3* are found on the far right side of Fig. 3D and have Z-scores >8. It is interesting to note that 22 of the 25 most overused dicodons in *YEF3* contain either GAA or AGA, with both codons decoded by tRNAs modified by Trm9. Notably, the dicodon combinations AGA-AGA, GAA-AGA, AGA-GAA and GAA-GAA were in the top 25 most over-represented list for *YEF3*, having Z-scores of 4.8, 4.1, 3.3 and 3.2, respectively. We propose that dicodon patterns are important because they report on the specific tRNAs that will need to be sequentially accessed for translation. In the case of AGA-AGA, GAA-AGA, AGA-GAA, GAA-GAA, they are all decoded by tRNAs modified by Trm9, such that these dicodons should be difficult to translate if there is a deficiency in Trm9-dependent uridine modifications.

We have also used CUT to explore codon usage patterns in yeast genes (Fig. 4) other then *YEF3*. Our goal was to determine if we could highlight significant dicodon and quintuplecodon patterns that correlate with decreased protein levels in *trm9*Δ cells, relative to wild-type cells. Specifically, we have used the codon painting function of CUT to highlight the AGA codon in the gene *ENT2* (Fig. 4A), with the finding that this gene contains two AGA-AGA doublets (Z-score = 2.0). The Ent2 protein is required for actin patch assembly and endocytosis. Based on over-use of AGA-AGA dicodons, we predicted that Ent3 protein levels would decrease in *trm9*Δ cells, a prediction that was confirmed by immunoblot analysis (Fig. 4B). As a negative control we have included immunoblot data on Met6 protein levels from wild-type and *trm9*Δ cells (Supplemental Figure S1). The Met6 gene is over-represented with AGA codons (31 total, Z-score = 2.5) but contains 0 AGA-

AGA dicodons. Met6 protein levels are not affected by a Trm9 deficiency, further supporting the idea that complex codon patterns are an important determinant of MoTTs.

The GAA codon has also been shown to have its translation linked to Trm9. We searched the CUT database to identify genes that use five GAA codons in a row and identified *NAB3*. In Fig. 4C, we painted *NAB3* to identify GAA codons, with the finding that there are nine instances of five GAA codons in a row (Z-score = 28). These nine instances were found in runs of 7 and 10 GAA codons. Nab3 is an RNA-binding protein that is part of the Nrd1 complex involved in some 3'-end processing. The GAA codon runs found in Nab3 are extremely rare and, based on previously published reporter data, the Nab3 protein should be decreased in *trm9*Δ cells compared to wild-type, which we confirmed by immunoblot analysis (Fig. 4D).

Translation of the AAG codon has been linked to Trm9 [9]. We have previously used an AAG-AAG-AAG-AAG reporter system to observe a modest but statistically insignificant decrease in reporter activity in *trm9*Δ cells, relative to wild-type [9]. The reporter system is an artificial construct and we wanted to determine if any endogenous yeast genes had AAG codon combinations greater than four and, if so, was the corresponding protein expression linked to Trm9. We used the CUT database to find the *HMO1* gene, which was significantly over-represented with AAG-AAG-AAG-AAG-AAG. This quintuplecodon occurred four times, with a Z-score of 55 (Fig. 4E). In fact, all four AAG-AAG-AAG-AAG-AAG quintuple codons are found in runs of 8 AAG codons. We predicted that, due to the presence of 8 AAG codons in a row, the levels of the Hmo1 protein should be strongly dependent on Trm9, with was again supported by immunoblot data (Fig. 4F). The *HMO1* gene is an interesting case because it contains 23 AAG codons in total, which is an average amount. The significance of the AAG codon in *NAB3* becomes apparent when dicodons are analyzed, with the AAG-AAG dicodon occurring 10 times and has a Z-score of 8.0. *NAB3* represents a good example of how complex codon patterns can be more informative for modeling translation than simple monocodon based descriptors.

In some cases, regulation of gene expression requires the coordination of many genes and, based on monocodon patterns, we have previously reported that groups of genes have the potential to be regulated by changes in tRNA wobble base modifications. The CUT database can be used to download all dicodon data for all genes in a species and determine if groups of genes over-use some dicodon combinations. In response to nutrient starvation, the Gcn2-eIF2α-Gcn4 pathway is activated by translational regulation [34–39], with subsequent Gcn4-based transcriptional regulation of many genes. Computational studies have identified Gcn4-binding sites in 80 transcripts [40]. We identified the 80 Gcn4-regulated genes in the CUT database download detailing dicodon usage, and then summed the Z-scores in each dicodon category, to get a group measure of dicodon over-usage (Fig. 5). The five dicodons with the highest summed Z-scores (>65) for the group of 80 Gcn4-regulated transcripts were GGU-GCC (Gly-Gly), GCA-GGU (Ala-Gly), GCU-GUG (Ala-Val), GGU-GUU (Gly-Val) and AGA-GAA (Arg-Glu). We were intrigued that the Trm9-regulated codons AGA and GAA were found in the list of top-scoring dicodons for the 80 Gcn4 transcripts, so we used a random sampling of 80 transcripts (300 times from all yeast genes) to determine significance (Fig. 5, **inset**). We note that significance of a group score was performed

outside of CUT and this functionality is planned for later versions of the tool. The average summed AGA-GAA dicodon Z-score for 80 randomly-sampled genes was 30 with a standard deviation of 7.2, making a highly significant ($P < 10^{-5}$) Z-score of 68 for Gcn4-regulated transcripts. While we can directly link the AGA-GAA dicodon to regulation by Trm9, the other top scoring Gly-, Ala- and Val-based codons are dependent on other tRNAs. It is interesting that the Gcn2-eIF2α-Gcn4 pathway, which begins with regulation of translation initiation, can now be linked to regulation of translation elongation through the AGA-GAA dicodon.

Next we determined if the CUT tool and database could used to identify interesting codon trends in mammalian transcripts. Specifically, we have applied the same algorithms and developed the same database query forms used to study yeast genes and analyzed all transcripts available in the RefSeq datasets for mice and rat genes (Fig. 6A). We highlight mouse *Brca1* and *Gpx1* genes, which encode key stress response proteins whose activities play important roles in DNA repair and ROS-detoxification, respectively [41, 42]. We downloaded dicodon Z-score data for all *Brca1* (Fig. 6B) and *Gpx1* entries (Fig. 6A) and then visualized the gene specific data in a scatter plot. Note that the maximum value for the Y-axis is 25 for *Brca1* and 65 for *Gpx1*. As can be observed in both *Brca1* and *Gpx1* plots (Fig. 6B–C), most of the data stays on a similar horizontal plane near the Y-axis (Z-score) values of 0 to 10 (*Brca1*) and 0 to 5 (*Gpx1*). There are a number of dicodon outliers for *Brca1* that have Z-scores (~22), including AGU-AAA and UUA-CCG (Fig. 6B). A similar analysis of *Gpx1* was performed (Fig. 6C), with the most dramatic outlier found at a Z-score of 55 for UGA-GGC, which was twice the score of any dicodon found in *Brca1*. The high Z-score for the UGA-GGC dicodon in *Gpx1* was expected as the corresponding protein contains selenocysteine, which is incorporated by stop codon recoding and utilizes a non standard internal UGA. Gpx1 and the UGA-GGC dicodon are examples of extreme codon usage, as there are only 24 reported selenoproteins in mice [43], and UGA is not usually followed by another codon in most transcripts, because UGA usually signals for the end of translation.

In addition to using CUT to identify significant codon patterns in individual or groups of genes, the program is also designed to be used as an exploration tool. We decided to use CUT to identify all mouse and rat transcripts that contain an internal UGA codon (Fig. 7), as this should identify transcripts that have the potential to undergo stop codon recoding (*i.e.*, selenoproteins) or to participate in some form of translational read-through to generate an extended protein with a different function (e.g., VegF-Ax) [25] or targeting to a specific organelle (Mdh1) [27]. We searched both the mouse and rat data loaded in CUT using the sequence UGA-NNN, where NNN represents any of the 64 codons. Our search identified 45 mouse (Table I) and 120 rat (Table II) transcripts with internal UGA codons, which was more than the 25 – 30 expected hits for each organism. We had expected 23–24 selenoproteins and a small number of other reported transcripts that undergo translational read-through in mice and rats. The number of internal UGA transcripts noted for each of our identified targets ranged from 1 to 20 for mouse transcripts and 1 to 36 for rat transcripts. While the transcripts that contain more than 1 UGA codon could be RefSeq annotation errors, we note that transcripts for selenoproteins can include 1 to 10 internal UGA codons

[12, 43], with Sepp1 corresponding to the latter in both our mouse and rat datasets (Tables I and II).

Specific to the mouse UGA-NNN search using the CUT tool, we identified all 24 known mouse selenoproteins as well as an additional 21 transcripts with internal UGA codons. We speculate that the 21 mouse transcripts have their UGA codon translated, with this speculation supported for 3 below detailed proteins. Specific to the rat UGA-NNN search using the CUT tool, we identified 23 of 24 known rodent selenoproteins, with the exception of RGD1560938/Seli/Ept1, which does not have an internal UGA codon. Remarkably, we identified an additional 97 transcripts with internal UGA codons. All identified mouse and rat transcripts were compared to the NCBI database using a nucleotide blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to ensure correct gene identification, with criteria of >99% identical gene BLAST being employed. The biochemical functions of selenoproteins include reduction of ROS, selenium transport and selenocysteine synthesis. Additionally, some selenoproteins play major roles in cancer prevention, thyroid metabolism, male fertility, and immune, muscle, and central nervous system functions [4]. The roles of some selenoproteins such as selenoprotein O, T and H (identified by CUT as 2700094K13Rik), are still unknown.

Mouse- and rat-specific transcripts that possess internal UGA codons but do not contain selenocysteine were also most likely identified in our UGA-NNN search. For example, VegFA was identified in our mouse and rat searches. The human version of VegF, known as *VegF-Ax*, was reported by Fox and co-workers to use translational read-through of the UGA codon to form a longer protein with anti-angiogenic activity [25]. In addition, studies have provided evidence for UGA read-through targets identified in our search, and they included *Ago1* and *Mtch2* [25], as well as *LdhB* and *Mdh1* (mouse and rat) [27]. In some cases, specific *cis*-acting sequence elements are believed to facilitate UGA read-through, with the resulting C-terminally extended proteins taking on physiologically distinct roles compared to those produced using the upstream (or canonical) UGA. For VegF-Ax (a VegF-A isoform with 22 additional C-terminal amino acids), the ribonuclear protein A2/B1 was identified as a *trans*-acting factor critical for VegF-Ax production. In parallel, for selenocysteine-decoding transcripts, recognition of the SECIS element by the RNA binding protein Sbp2 is critical for UGA translation as selenocysteine [44]. 3'-UTR and other bioinformatics parameters that can be used to further parse the identified UGA-containing transcripts are discussed below.

LdhB and Mdh1 can both be considered stress-response enzymes critical for intracellular redox homeostasis and they have CU bases immediately following the UGA stop codon. It has been shown that changing the CU dinucleoteide results in decreased read-through efficiency, as does changing the stop codon [27]. Lactate dehydrogenase B (LdhB) is one subunit of the lactate dehydrogenase enzyme, which converts lactate to pyruvate and vice versa via the interconversion of NADH and NAD^+. The C-terminally extended LdhB, termed LdhBx, has 6 extra amino acids and contains a peroxisomal targeting signal type 1 (PTS1) sequence in the 3' extension. LdhBx has also been shown to co-import LdhA, another lactate dehydrogenase subunit, into peroxisomes [45]. Similarly, two protein isoforms of malate dehydrogenase 1 (Mdh1) exist, one of which is created via translational

read-through to create a 20 amino acid C-terminally extended protein. Mdh1 catalyzes the reversible oxidation of malate to oxaloacetate using NADH/NAD^+, with the extended protein also containing a PTS1 [2]. It is therefore likely that peroxisomal localization of both LdhB and Mdh1 is dependent on stop codon read-through.

Genome-wide analysis and published reports have identified two additional targets of translational read-through, Ago1 and Mtch2, which contain 37 and 10 amino acids, respectively. Argonaute 1 (Ago1) is involved in RNA interference and RNA silencing by binding to micro-RNAs or siRNAs and repressing translation of complementary mRNA [46]. Mitochondrial carrier 2 (Mtch2) is a key player in the mitochondrial death pathway via the recruitment of proapoptotic truncated BID (tBID) [47]. It is still unclear how translational read-through alters the function of these two proteins as they have only recently been identified as targets for PTR. Ago1 and Mtch2 do not share the same CU dicodon following UGA that is necessary for translational read-through of LdhB and Mdh1. Instead, both Ago1 and Mtch2 are followed by an “AG” dicodon. Interestingly, the read-through efficiency of these two transcripts is highest among the five known PTR targets, at 13% for Mtch2 and 24% for Ago1, compared to 11% for VegF-Ax [25] and 1–2% for LdhB and Mdh1 [45].

We identified 97 rat transcripts with internal UGA codons that are not currently classified as selenoproteins. Similar to the mouse results we speculate that these UGA codons are translated. Considering the links between toxicant-induced tRNA modifications and the translational decoding of stress response transcripts,[48] several of these gene transcripts stand out. For example, *Bag4* belongs to a family of *Bag*-related chaperone proteins that have been shown to regulate cell death and growth decisions by interacting with both anti-apoptotic factors, as well as growth stimulatory kinases such as PI3K [49–52]. Other internal UGA-containing transcripts can be linked to cellular stress responses through their involvement in DNA damage-induced cell cycle checkpoint controls: the *Cdc25b* phosphatase targets cyclin dependent kinase 2 for dephosphorylation allowing entry into mitosis, and is itself targeted by *Chk1* in response to DNA damage [53, 54]; the *Taok1* kinase mediates chromosome-microtubule interactions during M phase of the cell cycle, guarding against chromosomal instability [55]. Last, *Fam120a* (also *Ossa/C9orf10*) is an RNA binding protein implicated in tumor cell resistance to oxidative stress by mechanisms thought to involve the activation of anti-apoptotic signals [56]. It will be interesting to determine the extent of translational recoding during stress responses, as it could provide new activities to optimize the response.

A total of 32 rat olfactory receptors were identified as having at least one internal UGA codon, and several had greater than five. Rat olfactory proteins comprise seven-transmembrane receptors and sense smell via direct interaction with odorant molecules and they belong to the largest multi-gene family in rats [57]. Perhaps differential UGA recoding represents a further means of diversification for olfactory receptors in rats, contributing to their advanced ability to differentiate between a myriad of odorants. It is further remarkable to find such a high number of internal UGA-containing transcripts in rat compared to mouse. This finding raises the possibility that rat uses UGA recoding to diversify protein expression from an individual transcript to an extent greater than that observed in other rodents. It

remains to be seen whether the internal UGA of these transcripts is decoded as selenocysteine, or serine, as is the case for *VegF-Ax* [25, 58]. The first step has been taken in our methodical identification of transcripts with internal UGA codons as candidates for recoding; the next step will be to employ proteomic analysis of candidate transcripts in order to determine which amino acid is used. Additionally it will be interesting to investigate whether the UGA containing mouse and rat transcript genes are encoded by transcripts that display a further codon bias and whether their regulation is influenced by tRNA modifications, according to the MoTT hypothesis [48, 59].

4. Conclusions

We have demonstrated the utility of the CUT tool and database by applying and analyzing gene specific dicodon patterns in yeast, mice and rat genes and transcripts. Development of the CUT database has permitted the first genome-wide exhaustive analysis of internal UGA stop codons. Our results reveal a diversity of potential new transcripts that undergo translational recoding, which can provide a means to expand the genetic code and increase the functional diversity of gene products without requiring the addition of new genetic material. While our UGA-NNN search has identified many transcripts with the potential to be translationally decoded, bioinformatics analyses of functional classification, 3'-UTR sequence space, and conservation between species will need to be performed to add further proof. One approach that could be employed to validate that the UGA-NNN transcripts are translated could be mass spectrometry based proteomics to identify the corresponding peptide. In addition, experimental evidence that provides direct support for translational decoding in the form of new peptides that contain selenocysteine or that correspond to amino acid sequences downstream of standard stop codons will be needed to confirm that the UGA codon was translationally decoded. Nevertheless the CUT tool thus provides a powerful new approach to gain insight into the complexity of mono- and multi-codon usage across yeast and mammalian genomes which can be used to elucidate the biological function of codon usage patterns in the translational control of gene expression. The version 1 CUT tool and database provides simple and complex codon usage data for all yeast, mouse and rat genes/transcripts in their respective genomes. The future evolution of CUT will include the addition of specific modules that include human, bacterial and other model organism data sets. Complex codon usage patterns have the potential to be regulatory in many species, and the CUT tool and database approach has the potential to be applied to all sequenced and gene-defined genomes. Future capabilities that include advanced search functions based on gene function and codon count criteria, statistical analysis to identify significant codon patterns in groups of genes and genome-genome comparisons are also planned.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank past and present lab members and colleagues for helping us compile and think about the data. We would like to thank Fraulin Joseph for his work on immunoblots. Funding for the these studies was provided by NIH ES017010 (TJB), National Science Foundation CHE-1308839 (PCD), National Research

Foundation of Singapore through the Singapore-MIT Alliance for Research and Technology Infectious Disease research program (PCD).

References

[1]. Globisch D, Pearson D, Hienzsch A, Bruckl T, Wagner M, Thoma I, Thumbs P, Reiter V, Kneuttinger AC, Muller M, Sieber SA, Carell T. Angew Chem Int Ed Engl. 2011; 50:9739–9742. [PubMed: 21882308]

[2]. Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowik A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother KM, Helm M, Bujnicki JM, Grosjean H. Nucleic Acids Res. 2013; 41:D262–267. [PubMed: 23118484]

[3]. Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. Cell. 2012; 149:202–213. [PubMed: 22464330]

[4]. Agris PF, Vendeix FA, Graham WD. J Mol Biol. 2007; 366:1–13. [PubMed: 17187822]

[5]. Tikole S, Sankararamakrishnan R. Journal of biomolecular structure & dynamics. 2006; 24:33–42. [PubMed: 16780373]

[6]. Weinger JS, Parnell KM, Dorner S, Green R, Strobel SA. Nature structural & molecular biology. 2004; 11:1101–1106.

[7]. Oxender DL, Zurawski G, Yanofsky C. Proc Natl Acad Sci U S A. 1979; 76:5524–5528. [PubMed: 118451]

[8]. Chan CT, Pang YL, Deng W, Babu IR, Dyavaiah M, Begley TJ, Dedon PC. Nat Commun. 2012; 3:937. [PubMed: 22760636]

[9]. Patil A, Dyavaiah M, Joseph F, Rooney JP, Chan CT, Dedon PC, Begley TJ. Cell Cycle. 2012; 11:3656–3665. [PubMed: 22935709]

[10]. Deng W, Babu IR, Su D, Yin S, Begley TJ, Dedon PC. PLoS Genet. 2015; 11:e1005706. [PubMed: 26670883]

[11]. Endres L, Begley U, Clark R, Gu C, Dziergowska A, Malkiewicz A, Melendez JA, Dedon PC, Begley TJ. PLoS One. 2015; 10:e0131335. [PubMed: 26147969]

[12]. Gladyshev VN, Hatfield DL. J Biomed Sci. 1999; 6:151–160. [PubMed: 10343164]

[13]. Begley U, Dyavaiah M, Patil A, Rooney JP, Direnzo D, Young CM, Conklin DS, Zitomer RS, Begley TJ. Mol Cell. 2007; 28:860–870. [PubMed: 18082610]

[14]. Chan TYC, Deng W, Li F, DeMott MS, Babu IR, Begley TJ, Dedon PC. Chem Res Toxicol. 2015; 28:978–988. [PubMed: 25772370]

[15]. Chan CT, Dyavaiah M, DeMott MS, Taghizadeh K, Dedon PC, Begley TJ. PLoS Genet. 2010; 6:e1001247. [PubMed: 21187895]

[16]. Chan CTY, Pang YLJ, Deng W, Babu IR, Dyavaiah M, Begley TJ, Dedon PC. Nat Commun. 2012; 3:937. [PubMed: 22760636]

[17]. Moustafa ME, Carlson BA, El-Saadani MA, Kryukov GV, Sun QA, Harney JW, Hill KE, Combs GF, Feigenbaum L, Mansur DB, Burk RF, Berry MJ, Diamond AM, Lee BJ, Gladyshev VN, Hatfield DL. Mol Cell Biol. 2001; 21:3840–3852. [PubMed: 11340175]

[18]. Bock A, Forchhammer K, Heider J, Leinfelder W, Sawers G, Veprek B, Zinoni F. Mol Microbiol. 1991; 5:515–520. [PubMed: 1828528]

[19]. Lee BJ, Worland PJ, Davis JN, Stadtman TC, Hatfield DL. J Biol Chem. 1989; 264:9724–9727. [PubMed: 2498338]

[20]. Gladyshev VN, Hatfield DL. Curr Protoc Protein Sci. 2001; Chapter 3(Unit 3):8. [PubMed: 18429173]

[21]. Novoselov SV, Calvisi DF, Labunskyy VM, Factor VM, Carlson BA, Fomenko DE, Moustafa ME, Hatfield DL, Gladyshev VN. Oncogene. 2005; 24:8003–8011. [PubMed: 16170372]

[22]. Songe-Moller L, van den Born E, Leihne V, Vagbo CB, Kristoffersen T, Krokan HE, Kirpekar F, Falnes PO, Klungland A. Mol Cell Biol. 2010; 30:1814–1827. [PubMed: 20123966]

[23]. Berry MJ, Banu L, Harney JW, Larsen PR. EMBO J. 1993; 12:3315–3322. [PubMed: 8344267]

[24]. Korotkov KV, Novoselov SV, Hatfield DL, Gladyshev VN. Mol Cell Biol. 2002; 22:1402–1411. [PubMed: 11839807]

[25]. Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, Willard B, Graham LM, DiCorleto PE, Fox PL. Cell. 2014; 157:1605–1618. [PubMed: 24949972]

[26]. Hoeben A, Landuyt B, Highley MS, Wildiers H, Van Oosterom AT, De Bruijn EA. Pharmacological reviews. 2004; 56:549–580. [PubMed: 15602010]

[27]. Stiebler AC, Freitag J, Schink KO, Stehlik T, Tillmann BA, Ast J, Bolker M. PLoS Genet. 2014; 10:e1004685. [PubMed: 25340584]

[28]. Ikemura T. Mol Biol Evol. 1985; 2:13–34. [PubMed: 3916708]

[29]. Keller EB, Calvo JM. Proc Natl Acad Sci U S A. 1979; 76:6186–6190. [PubMed: 392514]

[30]. Patil A, Chan CT, Dyavaiah M, Rooney JP, Dedon PC, Begley TJ. RNA Biology. 2012; 9:990–1001. [PubMed: 22832247]

[31]. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. Nature. 2003; 425:737–741. [PubMed: 14562106]

[32]. Begley TJ, Rosenbach AS, Ideker T, Samson LD. Mol Cell. 2004; 16:117–125. [PubMed: 15469827]

[33]. Tumu, S.; Patil, A.; Towns, WL.; Dyavaiah, M.; Begley, TJ. Database. 2012. bas002

[34]. Hinnebusch AG, Natarajan K. Eukaryot Cell. 2002; 1:22–32. [PubMed: 12455968]

[35]. Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ. Mol Cell Biol. 2001; 21:4347–4368. [PubMed: 11390663]

[36]. Goossens A, Dever TE, Pascual-Ahuir A, Serrano R. J Biol Chem. 2001; 276:30753–30760. [PubMed: 11408481]

[37]. Yang R, Wek SA, Wek RC. Mol Cell Biol. 2000; 20:2706–2717. [PubMed: 10733573]

[38]. Hinnebusch AG. Mol Microbiol. 1993; 10:215–223. [PubMed: 7934812]

[39]. Dever TE, Feng L, Wek RC, Cigan AM, Donahue TF, Hinnebusch AG. Cell. 1992; 68:585–596. [PubMed: 1739968]

[40]. Schuldiner O, Yanover C, Benvenisty N. Current genetics. 1998; 33:16–20. [PubMed: 9472075]

[41]. Welcsh PL, Owens KN, King I. Trends Genet. 2000; 16:69–74. [PubMed: 10652533]

[42]. de Haan JB, Bladier C, Griffiths P, Kelner M, O'Shea RD, Cheung NS, Bronson RT, Silvestro MJ, Wild S, Zheng SS, Beart PM, Hertzog PJ, Kola I. J Biol Chem. 1998; 273:22528–22536. [PubMed: 9712879]

[43]. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN. Science. 2003; 300:1439–1443. [PubMed: 12775843]

[44]. Driscoll DM, Copeland PR. Annu Rev Nutr. 2003; 23:17–40. [PubMed: 12524431]

[45]. Schueren F, Lingner T, George R, Hofhuis J, Dickel C, Gartner J, Thoms S. eLife. 2014; 3:e03640. [PubMed: 25247702]

[46]. Kim DH, Villeneuve LM, Morris KV, Rossi JJ. Nature structural & molecular biology. 2006; 13:793–797.

[47]. Katz C, Zaltsman-Amir Y, Mostizky Y, Kollet N, Gross A, Friedler A. J Biol Chem. 2012; 287:15016–15023. [PubMed: 22416135]

[48]. Endres L, Dedon PC, Begley TJ. RNA Biol. 2015; 12:603–614. [PubMed: 25892531]

[49]. Boiani M, Daniel C, Liu X, Hogarty MD, Marnett LJ. J Biol Chem. 2013; 288:6980–6990. [PubMed: 23341456]

[50]. Rahman P, Huysmans RD, Wiradjaja F, Gurung R, Ooms LM, Sheffield DA, Dyson JM, Layton MJ, Sriratana A, Takada H, Tiganis T, Mitchell CA. J Biol Chem. 2011; 286:29758–29770. [PubMed: 21712384]

[51]. Takayama S, Xie Z, Reed JC. J Biol Chem. 1999; 274:781–786. [PubMed: 9873016]

[52]. Eichholtz-Wirth H, Fritz E, Wolz L. Cancer Lett. 2003; 194:81–89. [PubMed: 12706861]

[53]. Sanchez Y, Wong C, Thoma RS, Richman R, Wu Z, Piwnica-Worms H, Elledge SJ. Science. 1997; 277:1497–1501. [PubMed: 9278511]

[54]. Draetta G, Eckstein J. Biochim Biophys Acta. 1997; 1332:M53–63. [PubMed: 9141461]

[55]. Shrestha RL, Tamura N, Fries A, Levin N, Clark J, Draviam VM. Open biology. 2014; 4:130108. [PubMed: 24898139]

[56]. Tanaka M, Sasaki K, Kamata R, Hoshino Y, Yanagihara K, Sakai R. Mol Cell Biol. 2009; 29:402–413. [PubMed: 19015244]

[57]. Buck LB. Cell. 2004; 116:S117–119. 111 p following S119. [PubMed: 15055598]

[58]. Eswarappa SM, Fox PL. Cancer Res. 2015; 75:2765–2769. [PubMed: 26122849]

[59]. Dedon PC, Begley TJ. Chem Res Toxicol. 2014; 17:7.

Highlights

- CUT tool and database can be used to identify yeast, mouse and rat genes with distinct codon usage patterns, which include 1- to 5 codon runs or combinations.
- CUT database can be used to download dicodon usage data for all 6,664 yeast genes and 16,711 rat and 30,392 mouse transcripts.
- The CUT database was used to identify 45 mouse and 120 rat transcripts with internal UGA stop codons.
- Identified UGA containing transcripts correspond to known selenoproteins and transcripts that undergo programmed translational read-through (PTR), as well as many new transcripts with potential to be translationally regulated.

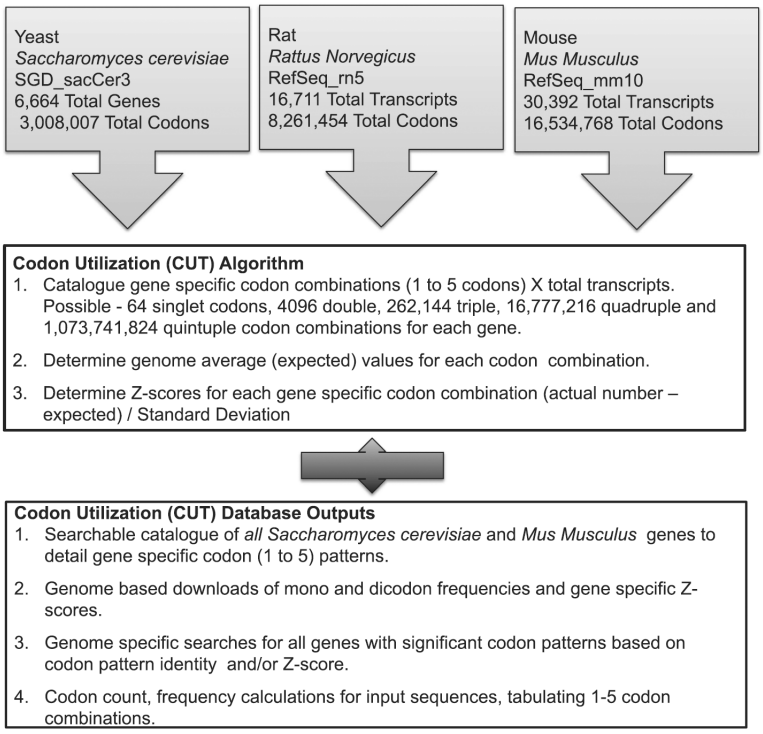

Figure 1. Algorithm outputs and database design for CUT

Species specific inputs for yeast, rat and mouse genes are listed in the upper panel, with these gene (yeast) and transcript (rat and mouse) entries analyzed (middle panel) to catalog and provide searchable fields. Data provided in the CUT database includes species- and gene specific information on codon usage patterns and measures of the significance of said codon patterns in a specific gene (Z-score), relative to species specific genome measures.

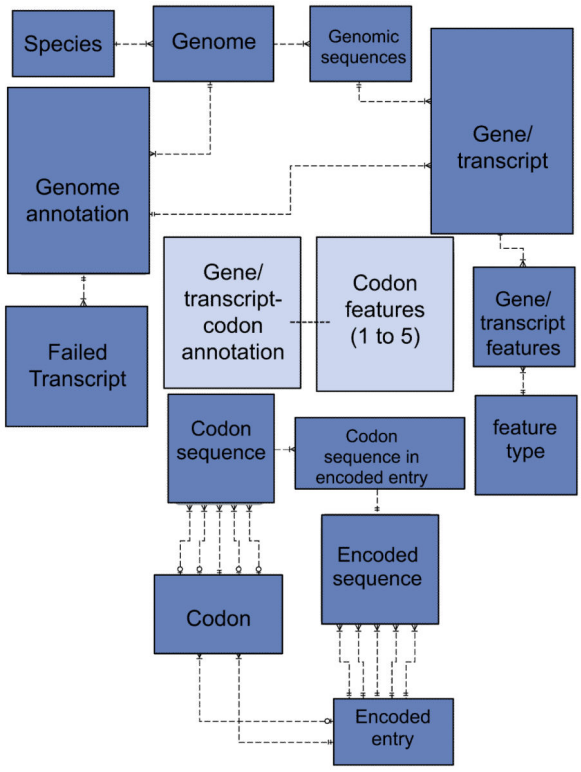

Figure 2. CUT database overview

The design of the database utilized tables that track the codon sequence usage for each transcript and for each annotation set (genome).

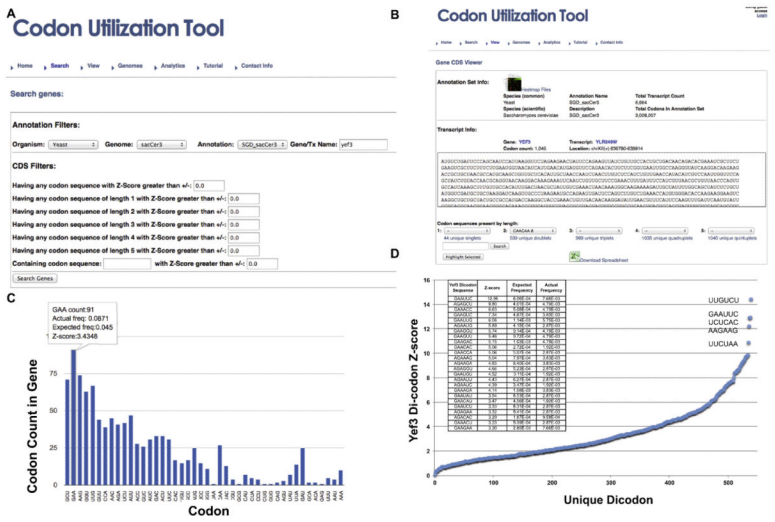

Figure 3. Design features and data outputs using CUT

(A) Species specific data can be accessed and searched to provide data on specific genes, or identify genes with specific codon patterns (1 to 5 codons) and Z-scores, with a data search of the Budding Yeast gene *YEF3* demonstrated. **(B)** Single codon patterns for the *YEF3* highlight the number of highly used codons and their gene specific over usage (Z-score). For example in *YEF3* the GAA codon is used 91 times and based on actual and expected frequencies a significant Z-score is observed. **(C)** As exampled using the *YEF3* gene and the GAA-GAA dicodon that is found 8 times, CUT can be used to determine the number of times a dicodon is represented in a gene, with **(D)** all dicodons from a gene being able to be exported into a tab delimited file and analyzed to determine the most significant dicodon combinations. The top five most overused dicodons in *YEF3* are represented on the graph, with the inset table detailing the next 25.

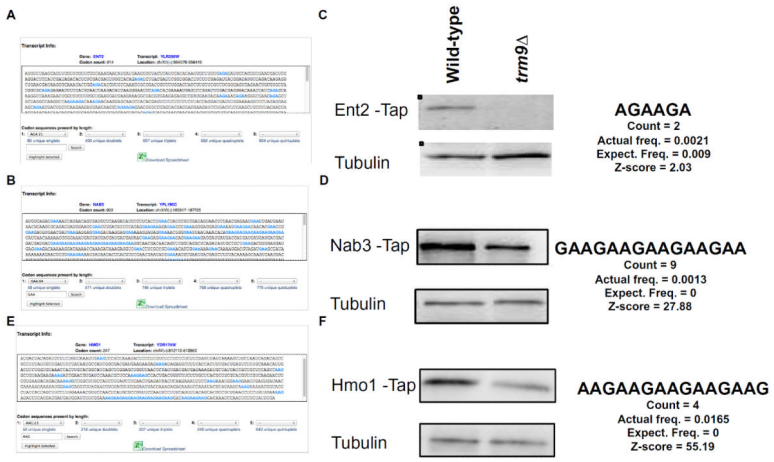

Figure 4. Codon painting of specific genes

CUT can be used to identify codon patterns on a specific gene to identify the general position of specific codon runs in **(A)** *ENT2*, **(C)** *NAB3* and **(E)** *HMO1*. AGA, GAA and AAG codons have all been linked to Trm9 regulated translation. Immunoblot analysis of the levels of specific proteins with **(B)** AGA-AGA for Ent2, **(D)** GAA-GAA-GAA-GAA-GAA for Nab3 and **(F)** AAG-AAG-AAG-AAG-AAG for wild-type and *trm9*Δ cells.

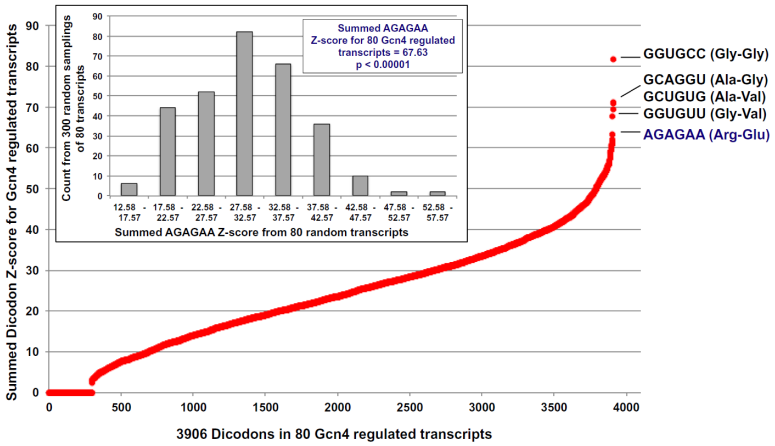

Figure 5. Dicodon trends in similarly regulated group of genes

All 64×64 dicodon patterns have been cataloged for each organism's genes/transcripts in CUT. Dicodon data on all species-containing genes can be downloaded and specific groups of genes can be analyzed for dicodon trends. Specifically, 80 transcripts that have been linked to regulation by Gcn4 were analyzed by summing each dicodon Z-score, with these values plotted for all 4,096 possible combinations. The AGA-GAA summed dicodon Z-score was determined to be significantly increased ($p < 10^{-5}$) for the Gcn4-regulated genes (inset panel). Significance was determined by random sampling of 80 transcripts 300-times from the 6,664 yeast genes to determine the average summed Z-score and standard deviation for the AGA-GAA dicodon in a group. The summed Z-score for the 80 Gcn4-regulated transcripts (68) was compared to the average score range of 80 random transcripts (28 to 33).

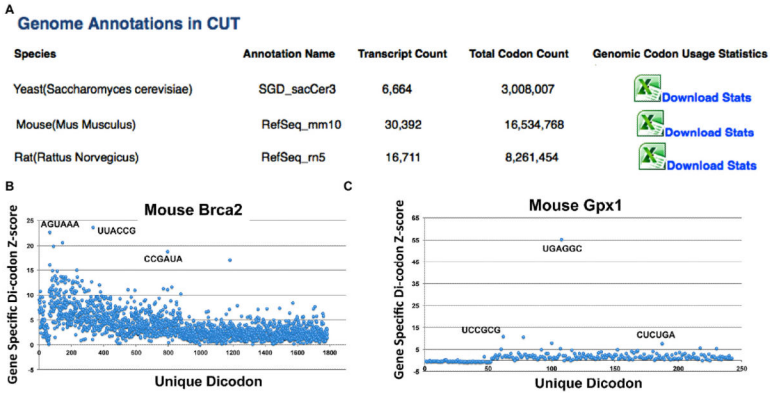

Figure 6. CUT tool organism downloads and dicodon data specific to mammalian genes

(A) Species specific information and gene specific monocodon and dicodon data for yeast, mouse and rat annotated genes and transcripts can be downloaded by clicking on the Download Stats Icon. Unique dicodon Z-scores indicating whether they are significantly over-used in mouse **(B)** *Brca2* and **(C)** *Gpx1*. Note that the Y-scale is different when comparing *Brca2* and *Gpx1* data, as the latter translationally recodes UGA.

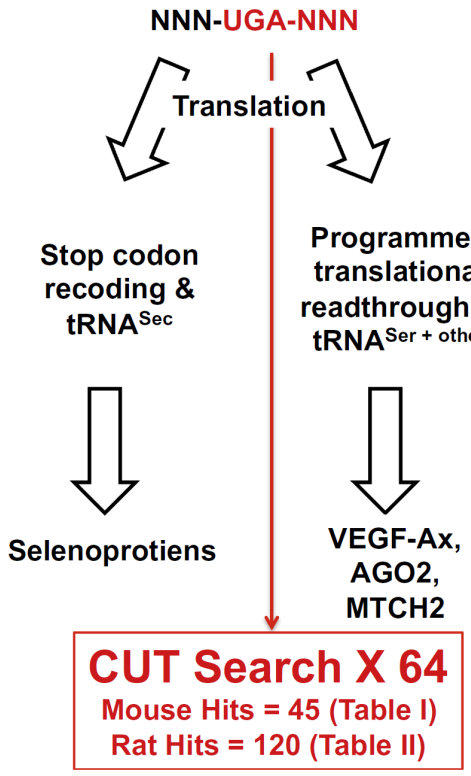

Figure 7. Translational recoding in mammals and CUT-based search for targets

In black, two potential mechanisms that can translate a UGA codon are detailed. In red, we describe the search space used to identify the 45 and 120 transcripts that contain internal UGA codons in mice and rats, respectively.

Table I

CUT-identified Mouse Transcripts with Internal UGA Codons

Mouse Gene Name	#of *internal* UGA codons	Transcript ID
4933416I08Rik	9	NM_027700
2700094K13Rik	1	NM_001033166
Ago1	1	NM_001317173
Art2a-ps	1	NM_007490
BC089491	1	NM_175033
Clec7a	1	NM_020008
Dio1	1	NM_007860
Dio2	2	NM_010050
Dio3	1	NM_172119
Ept1	1	NM_027652
Gm10058	1	NM_001109969
Gm10096	1	NM_001102678
Gm10147	1	NM_001099919
Gm10230	1	NM_001099347
Gm10486	1	NM_001109970
Gm14819	1	NM_001110250
Gpx1	1	NM_008160
Gpx2	1	NM_030677
Gpx3	1	NM_008161
Gpx4	1	NM_008162
Ldhb	1	NM_001316322
Mdh1	1	NM_001316675
Mia3	20	NM_177389
Msrb1	1	NM_013759
Mtch2	1	NM_001317241
Oas1b	1	NM_001083925
Olfr421-ps1	1	NM_146720
Olfr915	1	NM_146785
Pira6	6	NM_011093
Ptprv	1	NM_007955
Selk	1	NM_019979
Selm	1	NM_053267
Selo	1	NM_027905
Selt	1	NM_001040396
Sep15	1	NM_053102
Sephs2	1	NM_009266
Sepn1	1	NM_029100

Mouse Gene Name	#of *internal* UGA codons	Transcript ID
Sepp1	10	NM_009155
Sepw1	1	NM_009156
Smcp	3	NM_008574
Txnrd1	1	NM_015762
Txnrd2	1	NM_013711
Txnrd3	1	NM_153162
Vegfa	1	NM_001317041
Vimp	1	NM_024439

Table II

CUT-identified Rat Transcripts with Internal UGA Codons

Rat Gene Name	# of internal UGA codons	Transcript ID
Abcb11	6	NM_031760
Aass	12	NM_001100963
Abca1	8	NM_178095
Abi2	12	NM_173143
Acat2	1	NM_001006995
Acp1	4	NM_001313735
Aldh1a7	6	NM_017272
Atn1	10	NM_017228
Bag4	9	NM_001025130
Batf3	1	NM_021865
Cacna1a	9	NM_012918
Ccd25b	10	NM_133572
Ces1e	9	NM_031565
Ces2a	3	NM_144743
Cntnap5a	5	NM_001047865
Cstf3	7	NM_001077672
Cyp3a9	1	NM_147206
Dio1	1	NM_021653
Dio2	2	NM_031720
Dio3	1	NM_017210
Doc2b	8	NM_031142
Dync2h1	2	NM_023024
Eml1	8	NM_001025741
Fam120a	2	NM_001191816
Fdsp	1	NM_031840
Gnb5	1	NM_031770
Gpx1	1	NM_030826
Gpx2	1	NM_183403
Gpx3	1	NM_022525
Gpx4	1	NM_001039849
Grcc10	3	NM_001198725
Hdx	26	NM_001134568
Impg1	5	NM_023958
Itpr3	12	NM_013138
Kb23	1	NM_001008813
Kbtbd7	1	NM_001302944
Kif1c	14	NM_145877

Rat Gene Name	# of internal UGA codons	Transcript ID
Kng2	4	NM_001102418
Krt75	1	NM_001008828
Krt77	1	NM_001008807
Ldhb	1	NM_001316333
LOC100911576	4	NM_001271241
LOC498592	6	NM_001166307
LOC500684	12	NM_001047959
Lrp1b	2	NM_001107843
Ly6c	3	NM_020103
Mdh1	1	NM_001316877
Mfsd14a	17	NM_001106467
Msrb1	1	NM_001044285
Mybpc1	14	NM_001100758
Ndrg3	3	NM_001013923
Nkr-plc	1	NM_001040189
Olr1000	1	NM_001000077
Olr1052	4	NM_001001363
Olr1064	6	NM_001001076
Olr1065	5	NM_001000498
Olr1065	5	NM_001000498
Olr1226	8	NM_001000442
Olr1227	6	NM_001000443
Olr1228	5	NM_001000964
Olr1229	10	NM_001000444
Olr1237	8	NM_001000811
Olr1238	7	NM_001001013
Olr1239	8	NM_001000811
Olr1240	6	NM_001000448
Olr1241	6	NM_001000449
Olr1247	7	NM_001000807
Olr1254	7	NM_001001085
Olr1256	6	NM_001001086
Olr1257	10	NM_001000596
Olr1273	6	NM_001000458
Olr1273	4	NM_001000458
Olr1274	2	NM_001000801
Olr1275	4	NM_001000800
Olr1475	2	NM_001000027
Olr19	9	NM_001000117

Rat Gene Name	# of internal UGA codons	Transcript ID
Olr20	7	NM_001000118
Olr440	6	NM_001000282
Olr440	6	NM_001000282
Olr5	2	NM_001000112
Olr520	1	NM_001000930
Olr703	1	NM_001000359
Olr917	5	NM_001001354
Olr943	1	NM_001001368
Plaur	5	NM_017350
Plb1	17	NM_138898
Psmd10	4	NM_053925
Ptpn18	4	NM_001013111
Rabif	12	NM_001007678
Rasgrf1	1	NM_001105753
Rbm25	4	NM_001108984
Reln	1	NM_080394
RGD1305537	1	NM_001108822
RGD1307621	1	NM_001108025
RGD1563348	1	NM_001114939
SelI	1	NM_001134754
Selk	1	NM_207589
SelM	1	NM_001115013
SelO	1	NM_001085485
Selt	1	NM_001014253
Selv	1	NM_001166396
Sephs2	1	NM_001079889
Sepp1	10	NM_0191921
Sep15	1	NM_001166396
Sepw1	1	NM_013027
Slc21a4	7	NM_030837
Snap25	1	NM_001270576
Snx9	1	NM_001127637
Stau2	2	NM_001007149
Stxbp5l	26	NM_001271250
Taok1	9	NM_173327
Trpm7	36	NM_053705
Txnrd1	1	NM_031641
Txnrd2	1	NM_022584
Txnrd3	1	NM_001184712

Rat Gene Name	# of internal UGA codons	Transcript ID
Usp3	3	NM_001025424
Vegfa	1	NM_001317043
Vimp	1	NM_173120
Vom2r73	1	NM_001099486
Zfp14	21	NM_001100991