Published online 19 October 2022

https://doi.org/10.1093/nargab/lqac076

Analysis of nested alternate open reading frames and their encoded proteins

Kommireddy Vasu[1], Debjit Khan[1], Iyappan Ramachandiran[1], Daniel Blankenberg[2,*] and Paul L. Fox[1,*]

[1]Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA and [2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA

Received July 03, 2022; Revised August 14, 2022; Editorial Decision September 18, 2022; Accepted September 27, 2022

ABSTRACT

Transcriptional and post-transcriptional mechanisms diversify the proteome beyond gene number, while maintaining a sequence relationship between original and altered proteins. A new mechanism breaks this paradigm, generating novel proteins by translating alternative open reading frames (Alt-ORFs) within canonical host mRNAs. Uniquely, 'alt-proteins' lack sequence homology with host ORF-derived proteins. We show global amino acid frequencies, and consequent biochemical characteristics of Alt-ORFs nested within host ORFs (nAlt-ORFs), are genetically-driven, and predicted by summation of frequencies of hundreds of encompassing host codon-pairs. Analysis of 101 human nAlt-ORFs of length ≥150 codons confirms the theoretical predictions, revealing an extraordinarily high median isoelectric point (pI) of 11.68, due to anomalous charged amino acid levels. Also, nAlt-ORF proteins exhibit a >2-fold preference for reading frame 2 versus 3, predicted mitochondrial and nuclear localization, and elevated codon adaptation index indicative of natural selection. Our results provide a theoretical and conceptual framework for exploration of these largely unannotated, but potentially significant, alternative ORFs and their encoded proteins.

INTRODUCTION

The mechanistic link between genes and enzymes—the 'one-gene, one-enzyme' (or polypeptide) hypothesis—proposed by Beadle and Tatum (1) is a foundational component underlying the central dogma of molecular biology, and a key concept linking the fields of genetics and biochemistry. However, subsequent studies revealed a far more complex relationship between genes and proteins, namely, many genes generate a plethora of proteins by diverse physiological mechanisms in three major categories. In the first, the mRNA is modified at the transcriptional level by, for example, alternative splicing (2), mRNA editing (3) and altered polyadenylation (including coding region polyadenylation) (4–6). Alternatively, mRNA translation can expand the proteome by alternative translation initiation and by stop codon readthrough (7,8). Finally, diverse post-translational mechanisms are common, including amino acid modification and proteolytic cleavage. Importantly, these mechanistically distinct processes all share a common characteristic, namely, the primary sequences of the 'original' and modified proteins are closely related proteoforms. There is recent evidence and appreciation for a novel mechanism of proteome expansion, namely, translation of mRNAs in alternative reading frames (9,10). In this mechanism, an alternate start codon, out-of-frame with the 'canonical' initiation site of the reference coding sequence (cds), is followed by a stop codon in-frame with the novel start codon. The out-of-frame start and stop sites can be contained within the 5′- or 3′-untranslated region (UTR), overlap the UTR and reference cds, or be entirely nested within the reference cds (10). Most importantly, the primary sequences of proteins encoded by out-of-frame ORFs, or Alt-ORFs, are completely unrelated to the reference proteins, thus distinguishing this mechanism from other mechanisms of proteome expansion.

The extent of proteome expansion by out-of-frame mRNA translation has been explored, primarily by sequence analysis. An exhaustive compendium of potential Alt-ORFs of length >30 codons in 10 species has been described (OpenProt), and a human collection made accessible in a web-based database (HAltORF) (11,12). Analysis of overlapping viral genes and their encoded proteins revealed that at least two-thirds are expressed by translational mechanisms (13). The most common mechanism took advantage of alternative start codons, but ribosomal frameshifting and internal ribosome entry sites were

[*]To whom correspondence should be addressed. Tel: +1 216 444 8053; Fax: +1 216 444 9404; Email: foxp@ccf.org
Correspondence may also be addressed to Daniel Blankenberg. Tel: +1 216 444 4336; Email: blanked2@ccf.org

also utilized. Endogenous expression of several Alt-ORFs have been experimentally validated by mass spectrometric detection, by specific targeting antibodies, and by ribosome profiling (14–16). In some cases, information about the function of the Alt-ORF-derived proteins has been reported. The presence of functional Alt-ORF-derived proteins can contribute to the otherwise unexplained activity of single nucleotide polymorphisms that are synonymous in the host ORF, but non-synonymous in an altered reading frame (17).

To our knowledge there has not been a theoretical analysis of the ensemble of Alt-ORFs, and in particular, the overarching properties of their protein products. Among the categories of Alt-ORFs, nested Alt-ORFs (nAlt-ORFS) are unique in that the sequence, unlike Alt-ORFs containing untranslated regions, consists entirely of authentic codons, albeit out-of-frame (Figure 1A). Here, we take advantage of codon-related properties unique to nAlt-ORFs to interrogate specific characteristics of the transcripts and their polypeptide products, particularly the common features of the proteins including their physical and structural properties as well as intracellular localization.

MATERIALS AND METHODS

Code implementation and availability

The application nAltORFs was developed to investigate alternate protein coding regions (nAlt-ORFs) nested within annotated host coding sequences. The ‘find_nested_alt_orfs.py’ script identifies nAlt-ORFs based upon user-provided inputs, including a set of protein-coding genes (BED12 format), a matching reference genome (twobit format), a minimum translated peptide length threshold, an NCBI translation table numerical identifier, and an additional optional set of protein-coding genes for exclusion (BED12). The algorithm follows a stepwise approach, with values provided as used in this present study: (1) A list of annotated coding genes comprising about 20 000 unique transcripts, and their genomic locations, were obtained from the UCSC Table Browser in Gencode38, along with the associated two bit-formatted reference genome (2). The second ATG position in the cds was identified (3). ORFs with second ATG in frame 1 are excluded (4). Potential ORFs with the second ATG in frame 2 or frame 3 are selected (5). ORFs with translated peptide lengths greater than or equal to 150 are included (6). The sub-region is reported, otherwise the next regions are analyzed from the input BED file (7). Regions wholly in or spanning untranslated regions are excluded (8). Output files corresponding to a FASTA file of the original cds, a FASTA file for the potential nAlt-ORF in amino acid sequence, and FASTA files for the potential nAlt-ORFs in nucleotide sequence are generated. The output was manually curated to remove non-coding RNA, transcription-readthrough transcripts, duplicates, and pseudogenes. Using the above-described inclusion criteria, 101 putative human nAlt-ORFs with length ≥150 codons were ascertained (Supplementary Table S1).

Two additional Python scripts were created as part of nAltORFs for investigating codon and codon-pair usage. The script ‘codon_freq_from_bicodons.py’ takes the raw codon-pair table from CoCoPUTs (e.g. https://dnahive.fda.gov/dna.cgi?cmd=objFile&ids=537&filename=Refseq_Bicod.tsv&raw=1), along with the taxonomy ID and genome of interest, and calculates the frequency of codons in reading frames 2 and 3, based upon the entry for each codon-pair. Counts and frequencies are reported in two different tabular files, one by codon and a second by translated amino acid. The translation table as reported by CoCoPUTs formatted raw data is used, with the exception of a reported translation table identifier of ‘0’ being dynamically mapped to CoCoPUTs Table 1 Standard, SGC0, as reported for *Homo sapiens* genomic dataset. The final script, ‘bicodon_counts_from_fasta.py’, separately reports codon and codon-pair count information based upon a user-provided FASTA file. This script also takes as input taxonomy ID, organelle, division, assembly, species, and translation table, with the output formats mimicking that of the CoCoPUTs raw files. This allows simplified reuse of the previous script and comparison to the values provided by CoCoPUTs.

For control ORFs, a list of 101 ENSEMBL transcript IDs was generated with Regulatory Sequence Analysis Tools (RSAT) suite using the random gene selection form (RSAT: http://rsat.sb-roscoff.fr/random-genes_form.cgi) with the server command:

$RSAT/perl-scripts/random-genes -n 101 -g 1 -org Homo_sapiens_GRCh38 -feattype mRNA. Nucleotide sequences were manually curated and filtered to remove one transcript that was a host mRNA of a nAlt-ORF, to yield a final list of 100 control ORFs (Supplementary Table S2).

Bioinformatic analysis

Individual transcripts and their genetic location were collected from Ensembl genome GRCh38.p13 (https://www.ensembl.org) and ORFs detected using SMS2 (https://www.bioinformatics.org/sms2/orf_find.html). Amino acid sequences of the ORFs were determined and used for isoelectric point analysis using DTASelect algorithm in SMS2. Protein Pi distribution was analyzed using GraphPad Prism software. Codon-pairs that encode AUG were extracted in reading frames 2 and 3 from niche region of the 101 host-ORFs. Codon-pairs that encode the upstream start site after the authentic initiation site were aligned using using WebLogo 2.8.2 (https://weblogo.berkeley.edu/logo.cgi).

Codon adaptation index (CAI), a measure of synonymous codon usage relative to an efficiently translated reference set were determined for host, nAlt-ORF, and random gene set. For generating random out-of-frame ORFs, stop codons were removed in all the three reading frames. CAI values were calculated using codon usage of 93 487 CDSs from the human transcriptome using CAIcal web server (http://genomes.urv.es/CAIcal/).

Analysis of protein localization and function

Gene ontology analysis was done using PANTHER protein class and Reactome pathway annotation datasets (18). Briefly, host genes harboring nAlt-ORFs were mapped using UniProt ID Mapping to IDs in the PANTHER annotation data set (https://www.pantherdb.org/). Gene IDs

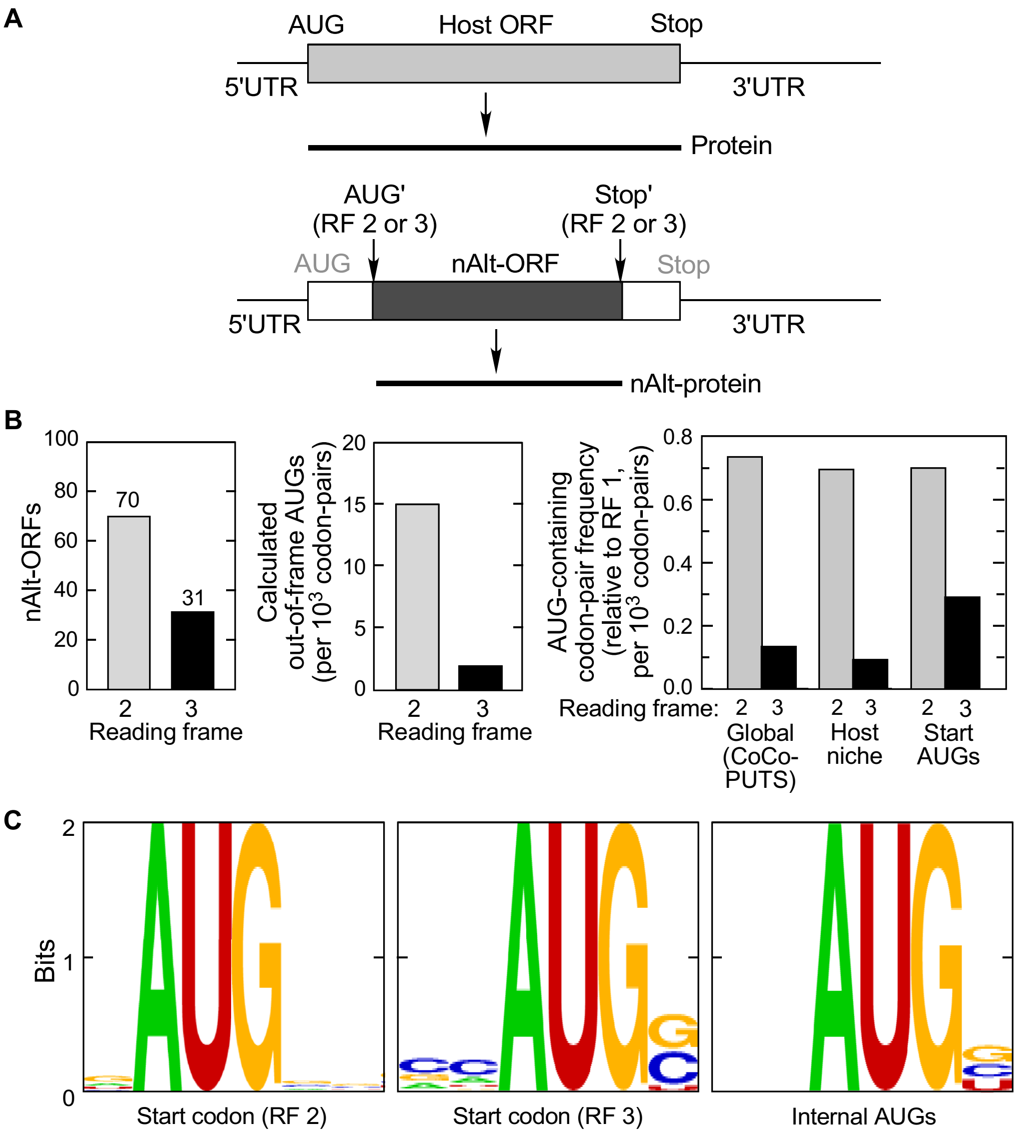


Figure 1. Asymmetric reading frame preference of nAlt-ORFs. (**A**) Schematic of nAlt-ORFs. Translation-initiation at the canonical start codon (AUG) generates the parental host ORF. Translation-initiation of an nAlt-ORF begins at a downstream start codon (AUG') within the host ORF, in an altered reading frame (RF) and ending at a stop codon (stop') in-frame with AUG' and within the host ORF. (**B**) Distribution of human nAlt-ORFs in reading frames 2 and 3 (left). Out-of-frame AUGs calculated from CoCoPUTs human codon-pair database (center). AUG-containing codon-pair frequency in alternate reading frames in global group of human CDSs (from CoCoPUTs), host niche sequences corresponding to nAlt-ORF regions, and start sites in nAlt-ORFs normalized to reading frame 1 (right). (**C**) Sequence alignment of codon-pairs surrounding the start sites in RF2 (left), RF3 (center) and internal AUG codons of nAlt-ORFs, visualized as sequence logos.

not mapped to PANTHER sequence IDs were manually corrected using Ensembl Biomart conversion tool (https://www.ensembl.org/biomart/martview/). Mapped IDs were subjected to statistical over-representation test using all human genes as a reference list and default settings. P-values were further corrected using Benjamini and Hochberg false-discovery rate correction. Data was considered significant at $P < 0.05$. GOSlim categories under umbrella GO categories, i.e. 'protein class' and 'reactome pathway', were tested.

The input protein sequences were converted into FASTA format using BioWord. Intracellular localization of nAlt-ORFs was predicted using YLoc web server with default parameters. YLoc uses Bayesian analysis with entropy-based discretization to generate predicted location, probability, and confidence scores. Locations with high confidence score were tabulated and Pearson correlation coefficients determined by GraphPad Prism software.

Modeling and visualization

Protein structure prediction was done using AlphaFold v2.1.0 with a Jupyter Notebook. Briefly, py3dmol, OpenMM and pdbfixer and AlphaFold (https://github.com/deepmind/alphafold) were installed. Monomers were modeled using a single ORF input sequence. For each protein, the number of sequences used for multiple sequence alignment and the average predicted local distance difference test (pLDDT) score, respectively, are shown. Predicted structures were energy minimized and visualized using Pymol or Chimera.

Quantification and statistical analysis

For statistical analysis of non-sequencing data, GraphPad Prism software processed and presented the data as indicated in the figure legends. Statistical significance was calculated by unpaired two-tailed parametric *t* test for the column graphs. *P* values <0.05 are considered significant (NS, not significant). For divergence in sub-cellular localization and codon-pair analysis and Pearson or Spearman correlation coefficients were determined by GraphPad Prism software. For GO analysis, *P*-values were further corrected using Benjamini and Hochberg false-discovery rate correction provided by the PANTHER database. Transformation of the codon usage in various datasets by z score was used for clustering analysis.

RESULTS

Asymmetric reading frame selection in nAlt-ORFs

To improve the likelihood of selection of nAlt-ORF generating non-random polypeptides, a minimum length of 150 codons was selected for inclusion. The probability of a ≥150 codon stretch without a stop codon is low; assuming equal codon utilization, the probability is less than $(61/64)^{150}$ or $<7.5 \times 10^{-4}$. Although small peptides certainly can exhibit activity, we assumed large proteins are more likely to be non-random and functional. To further constrain the collection, only nAlt-ORFs that utilized the first out-of-frame AUG following the canonical start codon were selected. This initiation site would be the first encountered by a leaky scanning ribosome that bypasses the canonical codon, and thus has higher likelihood of expression. An analysis of overlapping proteins in viral genomes supports this mechanism as more than two-thirds are generated by alternative start codon utilization (13). Query of 20 000 unique transcripts in the UCSC Table Browser, in both alternative reading frames, followed by manual curation to remove non-coding RNA, transcription-readthrough transcripts, duplicates, and pseudogenes, yielded a collection of 101 putative human nAlt-ORFs with length ≥150 codons (Supplementary Table S1). The longest nAlt-ORFs encoded proteins of lengths 529, 396 and 323 amino acids. Endogenous expression of two predicted nAlt-ORFs derived from *FUS* and *ATXN1* genes has been validated experimentally, and their protein products characterized (14,15).

An imbalance in the number of nAlt-ORFs in each reading frame was observed, with more than twice as many in reading frame 2 compared to frame 3 (Figure 1B, left). The ratio was particularly surprising given that 'slippery' sequences that drive programmed ribosomal frameshifting favor –1 nt shifts to generate reading frame 3 (19). We recognized that sequential codon-pairs that encode AUG in alternate reading frames can provide information to permit global estimation of relative frequency of alternative reading frames. According to this analysis, AUG in alternate reading frames can be specified in two (and only two) ways: by the codon-pair $N_1AU\text{-}GN_2N_3$ in reading frame 2 in the host mRNA, or by the codon-pair $N_1N_2A\text{-}UGN_3$ in reading frame 3, where N_1, N_2 and N_3 can be any nucleotide. The frequency of sequential codon pairs is neither random nor an exact function of the constituent codons (20). Importantly, we took advantage of the tabulated frequency of occurrence of all 3712 codon-pairs (64^2 minus 64 × 3 stop codons for each codon-pair position) in the human genome in the CoCoPUTs database (20). Based on the assumption that codon-pair frequency in nAlt-ORFs is consistent with the frequency in the human genome, the predicted frequency of AUG in reading frame 2 equals the sum of the frequencies of all 64 (= 4^3) $N_1AU\text{-}GN_2N_3$ codon-pairs:

$$\begin{aligned}&\text{Frequency of codon AUG in reading frame 2} =\\&\Sigma\ \text{Codon-pair frequencies } N_1AU\text{-}GN_2N_3\\&\forall N_1, N_2, N_3 = \text{U, C, A, G}\end{aligned}\qquad(1)$$

Similarly, the frequency of AUG in reading frame 3 equals the sum of the frequencies of the codon-pairs $N_1N_2A\text{-}UGN_3$ for all 64 combinations:

$$\begin{aligned}&\text{Frequency of codon AUG in reading frame 3} =\\&\Sigma\ \text{Codon-pair frequencies } N_1N_2A\text{-}UGN_3\\&\forall N_1, N_2, N_3 = \text{U, C, A, G}\end{aligned}\qquad(2)$$

The predicted start codon frequency was calculated using a Python script in which frequencies from the CoCoPUTS human codon-pair database were summed. The predicted ratio of AUG in reading frame 2 compared to frame 3 was 7.5:1, considerably higher than the ~2.3:1 ratio observed in the nAlt-ORF group (Figure 1B, center). The finding of substantially lower ratio observed in the collection of nAlt-ORFs could result from ribosome -1 slippage thereby partially offsetting the genetic 'force' driving the shift toward reading frame 2. Alternatively, the lower-than-predicted ratio can be due to discrepant codon-pair usage in nAlt-ORFs compared to the human genome. We investigated whether the frequency of AUG-containing codon-pairs in reading frame 2 of the nAlt-ORF group is unusually low, or alternatively, frequency in reading frame 3 is unexpectedly high. AUG-containing codon-pair frequency in three groups was evaluated in both reading frames and normalized to reading frame 1, namely, (i) the 'global' collection of codon-pairs in human CDSs derived from CoCoPUTs, (ii) a 'host niche' consisting of the sequences in the 101 host mRNAs corresponding to the nAlt-ORFs, and (iii) the start codon in nAlt-ORFs. In all three groups, AUGs in reading frame 2 were 70–74% relative to reading frame 1 (Figure 1B, right). In contrast, normalized to reading frame 1, reading frame 3 showed a divergence in AUG frequency: 9–13% in the host niche and global codon-pairs, but 30% in nAlt-ORF start codons. Thus, the abundance of start codons in reading frame 3 codon-pairs is the principal contributor to the observed anomalously high number of start codons in reading frame 3 of nAlt-ORFs.

The specific AUG-containing codon-pairs were investigated for a clue to the prominent selection in reading frame 3. A sequence logo plot of the codon-pairs revealed a dominant CCAUGG in frame 3 not present in frame 2, nor in the AUGs in codon-pairs in the -1 frame of the host niche interior (Figure 1C). Remarkably, CCAUGG was the most abundant codon-pair, present in 5/31 codon-pairs in reading frame 3 of the nAlt-ORFs. This sequence is similar to the Kozak consensus initiation sequence A/GCCAUGG where the A/G in the -3 position and G in the +4 position are most critical (21). However, only about 5% of eukaryotic mRNAs exhibit perfect consensus sequences; in the absence of a -3–position purine, the +4–G is essential and can drive initiation with an efficiency of about 60–90% of the exact sequence. Interestingly, 14/31 of the codon-pairs in reading frame 3 contain G in the +4 position. Possibly, nAlt-ORFS bearing the near-consensus Kozak sequence exhibit relatively high expression. Moreover, the analysis suggests a rationale for preferential selection of frame 3. With only a single nt upstream of the AUG in frame 2, inclusion of CC upstream of the AUG would require a third codon. Also, the +4 'G' in frame 3 is in the wobble-base position of the host ORF, but not in frame 2, thus providing more flexibility in its utilization.

Selective intracellular localization and elevated isoelectric point of nAlt-ORF-derived proteins

There was not an apparent concentration of nAlt-ORFs on any individual chromosome (Figure 2A), consistent with random distribution. nAlt-ORFs were not detected in chromosome Y (and 4), but the relatively small sample size precludes speculation on the relevance to sex-linkage. Protein class and pathway analyses of the 101 nAlt-ORF protein products was determined by gene ontology (GO) enrichment analysis using protein analysis through evolutionary relationships (PANTHER) classification system (22). As a protein class, the DNA-binding, helix-turn-helix transcription factor family was significantly enriched (Figure 2B). Likewise, three pathways were enriched, namely, metabolism, metabolism of lipids, and developmental biology. Intracellular localization of the nAlt-ORF proteins was determined by the YLoc Bayesian algorithm which considers potential sorting signals, sequence motifs, as well as amino acid physical properties including hydrophobicity, charge, and volume (23). Two notable features were observed, namely, a ~3-fold higher fraction of nAlt-ORF proteins in mitochondria compared to the host ORF proteins, a ~20% increase in nuclear proteins, and a ~10-fold lower level of cytoplasmic proteins (Figure 2C). There is little or no relationship between the predicted localization of the nAlt-ORF and host proteins (Figure 2D). A potential clue to the atypical localization of nAlt-ORF proteins was provided by a previous investigation of intracellular localization as a function of protein pI (24). Proteins with a pI of ~12 were preferentially localized in the nucleus and mitochondria, and at extremely low levels in the cytoplasm (Figure 2E, left). The intracellular distribution of nAlt-ORF proteins closely matched that of the global proteins exhibiting a pI of 12 (Figure 2E, right). Indeed, the vast majority of nAlt-ORF proteins exhibited calculated p*I*s >10, with a median pI of 11.68 (mean = 11.25), consistent with their atypical intracellular distribution (Figure 3A). There was not a significant difference in mean Pi between nAlt-ORFs in frames 2 and 3 (Supplementary Figure S1). These pIs were very distinct from the host proteins which were primarily distributed between 4 and 10 (Figure 3B), similar to the pI range of proteins from a group of randomly selected mRNAs (Figure 3C, Supplementary Table S2).

Anomalous charged amino acid composition of nAlt-ORF-derived proteins

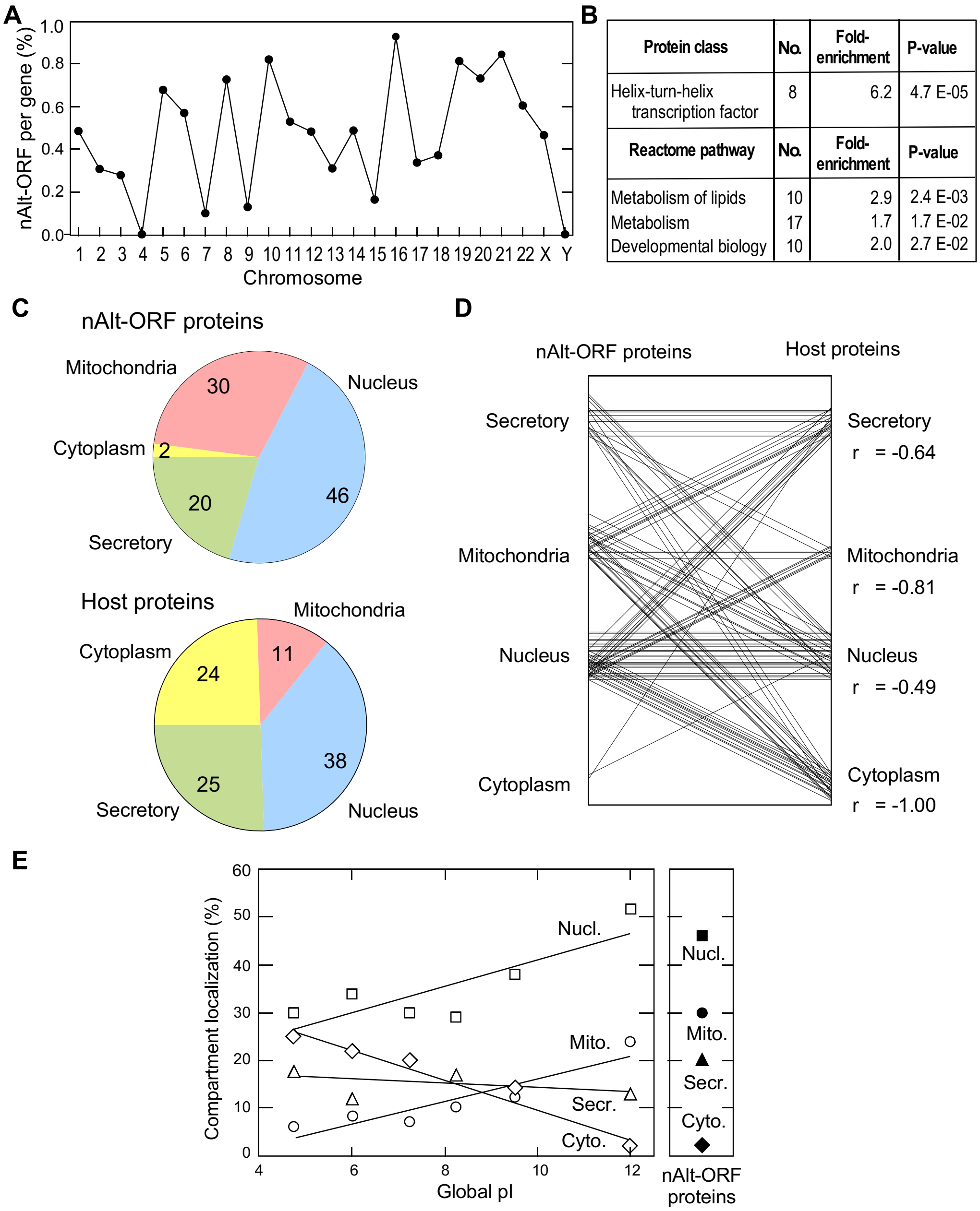
The underlying basis of the extraordinarily high p*I* in nAlt-ORF-derived proteins was explored in detail. The frequencies of the four charged amino acids that are primary contributors to p*I* were determined. Two characteristics contributed importantly to the elevated p*I* of nAlt-ORF proteins: (1) the overall frequencies of Asp and Glu are about one-third of the global mean frequency, and (2) the overall frequency of Arg was more than double the global mean frequency (Figure 4A, B). The overall positive-to-negative amino acid ratio, i.e. (Lys + Arg)/(Asp + Glu), in the nAlt-ORF protein group is 4.09, whereas the ratio in global human proteins is 0.98, thereby accounting for the extremely high mean p*I* in the nAlt-ORFs and the near-neutral p*I* globally. The positive-to-negative charge ratio in the host ORF proteins is 1.03, similar to the global ratio (Figure 4C), indicating the frequency and ratio of charged amino acids in nAlt-ORF-derived proteins is host-independent.

We investigated the genetic origin of the anomalous frequencies of charged amino acids translated from nAlt-ORFs, analogous to the anomalous initiation codon reading frame ratio. The expected frequency of any amino acid equals the sum of the expected frequencies of each of the 1–6 codons that specify the amino acid. Using the CoCoPUTs database, the expected global mean frequencies of amino acids in each alternate reading frame were calculated by summing all amino acid-specifying codon-pairs: Two codon-pairs for Asp, Glu, and Lys and six for Arg, were calculated separately and frequencies in each reading frame summed after weighting by their predicted ratio, i.e. 7.5:1 for reading frame 2 to reading frame 3, to account for discrepant utilization. Thus, the expected global mean frequency of each amino acid in an alternate reading frames is determined as the sum of the frequencies of all codons $X_1X_2X_3$ that specify the amino acid, as determined by nts N_1, N_2, and N_3 in the surrounding codon-pairs. The total global frequency for an amino acid is given by the sum in each reading frame, corrected by the calculated 7.5:1 ratio:

$$\begin{aligned}&\text{Frequency of codon } X_1X_2X_3 \text{ in reading frame 2} =\\&[\Sigma\ \forall \text{ codons } X_1X_2X_3,\ (\Sigma \text{ codon-pair frequencies } N_1X_1X_2\text{-}X_3N_2N_3)] * (7.5/8.5)\\&\forall N_1, N_2, N_3 = \text{U, C, A, G}\end{aligned} \tag{3}$$

$$\begin{aligned}&+ \text{Frequency of codon } X_1, X_2, X_3 \text{in reading frame 3} =\\&[\Sigma\ \forall \text{ codons } X_1X_2X_3,\ (\Sigma \text{codon-pair frequencies } N_1N_2X_1\text{-}X_2X_3N_3)] * (1.0/8.5)\\&\forall N_1, N_2, N_3 = \text{U, C, A, G}\end{aligned} \tag{4}$$

The calculated global mean frequencies of the charged amino acids in alternate reading frames are similar, but not identical, to the actual frequencies in the nAlt-ORF proteins, with low Asp and Glu content, and high Arg content,



Protein class	No.	Fold-enrichment	P-value
Helix-turn-helix transcription factor	8	6.2	4.7 E-05
Reactome pathway	**No.**	**Fold-enrichment**	**P-value**
Metabolism of lipids	10	2.9	2.4 E-03
Metabolism	17	1.7	1.7 E-02
Developmental biology	10	2.0	2.7 E-02

Figure 2. nAlt-ORFs exhibit preferential mitochondrial and nuclear localization. (**A**) Chromosomal distribution of nAlt-ORFs. (**B**) Gene ontology analysis of host proteins using PANTHER protein class and Reactome pathway annotation data sets. (**C**) Relative distribution of subcellular localization in nAlt-ORF (top) and host ORF proteins (bottom), as determined by the YLoc Bayesian algorithm. (**D**) Weak association of nAlt-ORF localization with parental host ORFs quantitated by Pearson correlation coefficients, *r*. (**E**) Proteome-wide compartment localization of human proteins as a function of mean pI (left, adapted from (24)), and relative localization nAlt-ORFs (right).

similar to the observed frequencies (compare Figure 4D to A), with a positive-to-negative charge ratio of 3.27. An analysis limited to the ensemble of 101 nAlt-ORFs very nearly matched the actual frequencies determined in these proteins (Figure 4E), with a positive-to-negative charge ratio of 3.71 indicating global codon-pair frequencies are not identical to the frequencies in the nAlt-ORFs considered here. The substantial difference in charged amino acid content between the actual nAlt-ORF ensemble (Figure 4A) and the global group, calculated from CoCoPUTs data (Figure 4D) might reflect a difference in codon-pair usage. A Spearman *r* of 0.21 confirms the unique codon-pair usage in nAlt-ORFs compared to global frequency (Figure 4F). Likewise, codon-pairs in the nAlt-ORF were poorly correlated with the host-ORF (Spearman $r = 0.42$). Finally, host-ORF codon pairs correlated more closely to the global pairs (Spearman $r = 0.71$), than did the host sequence in reading frame 1 (host-niche, Spearman $r = 0.51$); removal of the niche from the host-ORF (host- Δniche) improved the correlation slightly (Spearman $r = 0.75$). The differen-

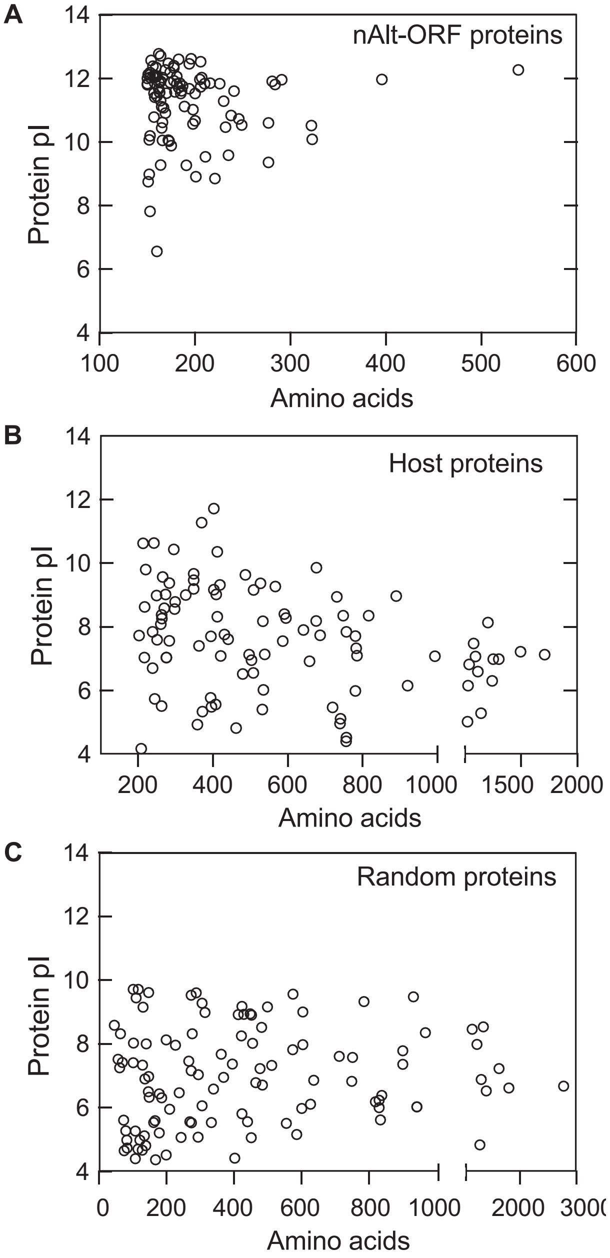


Figure 3. Isoelectric point of nAlt-ORFs. (A–C) Relationship between isoelectric point and protein length of (**A**) nAlt-ORF-derived proteins, (**B**) host ORF-derived proteins, and (**C**) proteins translated from a randomly selected group of human ORFs.

tial utilization of codon-pairs between the host-niche and its sequence complement, the host-Δniche, (Spearman $r = 0.75$), despite both occupying reading frame 1, suggests the niche is subject to evolutionary adaptation.

Genetic mechanisms underlying anomalous usage of charged amino acids

The differential utilization of charged amino acids in alternate reading frames is illuminated by examining the relationship between mean amino acid utilization and the number of codons encoding each amino acid, i.e. 2 for Asp, Glu, and Lys and 6 for Arg. A graphical analysis of global amino acid utilization in reading frame 1 as a function of codon number shows a near-linear relationship between codon number and frequency (Figure 5A), as observed by others (25,26). Importantly, among the more prominent exceptions to this relationship, Asp and Glu are substantially higher than predicted by codon number, and Arg substantially lower. To calculate the global amino acid utilization in both alternate reading frames, we again took advantage of the CoCoPUTS database. In each reading frame, for each codon, 64 (4^3) codon-pairs are summed. Thus, for amino acids encoded by two codons, 128 codon-pairs are summed for each of two reading frames, or 256 in total. For Arg, which is encoded by 6 codons, 384 codon-pairs are summed for each reading frame or, remarkably, a total of 768 codon-pairs. The codon-pairs are generally weakly related to the original codon, i.e. the original codon is not encoded by either codon of the codon-pair. Thus, the large number of codon-pairs involved tends to lessen the anomalous relationship between amino acid utilization and codon number. This moderation is clearly seen in the calculations for Arg, Asp, and Glu in reading frames 2 and 3, as well as after the calculation considering both reading frames (Figure 5A–D). In short, utilization of Asp, Glu, and Arg is closer to the expected linear relationship than is observed in the actual proteins in reading frame 1, consistent with the differential amino acid utilization observed.

The calculation based on global codon-pair usage indeed shows markedly elevated Arg and reduced Asp and Glu content in out-of-frame proteins (Figure 5D) compared to actual global usage (Figure 5A), i.e. 8.62% versus 5.67%. However, the calculated mean Arg content is substantially lower than the actual content in the nAlt-ORF group (12.69%, Figure 4A). We considered the possibility that this anomaly is due to differential codon usage, and investigated usage of the six individual codons encoding for Arg. Codon frequency in reading frames 2 and 3 were determined separately using Equations (3) and (4), respectively, and weighted by their calculated ratio of occurrence. The major increase in calculated usage in frames 2 and 3 (CoCoPuts) compared to actual global codon usage was seen for codons AGA and AGG - usage was 2.7- and 2.3-fold higher, respectively (Figure 5E). Two Arg codons, CGC and CGG, are markedly enriched in nAlt-ORFs compared to the global out-of-frame ORFs calculated by CoCoPUTs. Examination of the most highly used codon-pairs containing these codons are notably G- and C-rich (Figure 5F). Intriguingly, the CG-dinucleotide, present in the four least abundant Arg codons, occurs with a frequency about 2.6-fold higher in nAlt-ORFs compared to global CDS usage.

We considered the possibility that codon-pair usage in nAlt-ORFs drives differential usage of codons in general, i.e. not just Arg codons. Codon usage in global, nAlt, and host ORFs were compared by hierarchical clustering in a heatmap (Figure 6A). A near-complementarity of usage of nAlt-ORF compared to global ORFs is apparent; a Pearson r score of 0.26 confirmed the extreme difference between the groups (Figure 6B). Even more striking is the clear dichotomy in codon clustering in nAlt-ORFs; there is a pre-

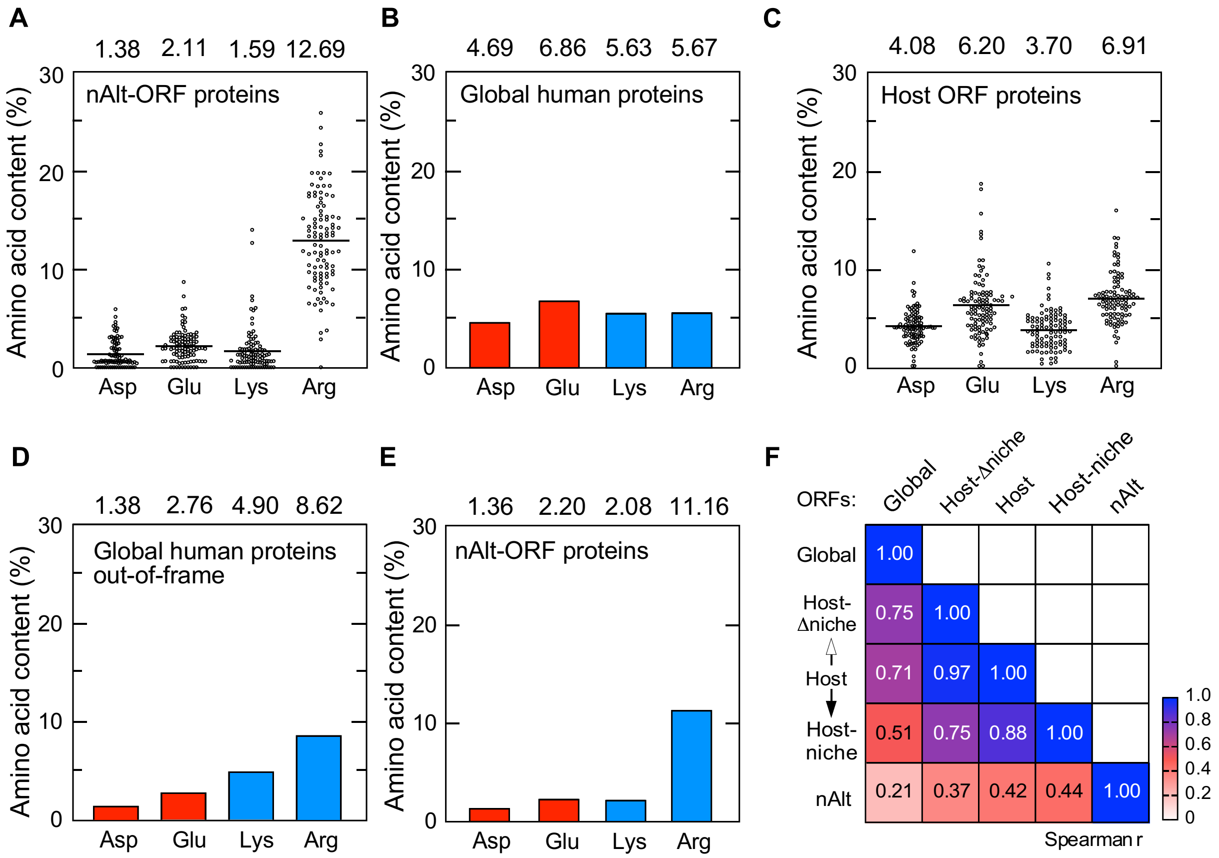


Figure 4. nAlt-ORFs exhibit anomalous content of charged amino acids Asp, Glu, Arg. (A–C) Charged amino acid content is calculated using DTASelect algorithm in SMS2 for (**A**) nAlt-ORF proteins, (**B**) global human proteome, and (**C**) host ORF data set. (D, E) Amino acid frequencies of the charged residues were calculated using Equations (3) and (4), from codon-pair frequencies for each reading frame, and weighted by the ratio of 7.5:1 for reading frame 2 to 3. (**D**) Global collection of human CDSs derived from CoCoPUTs data. (**E**) Host niche sequences corresponding to the region in nAlt-ORFs. (**F**) Spearman correlation coefficients were determined to reveal differences in codon-pair usage in the global CoCoPUTs, host ORFs, host ORF niche region, host ORF sequences lacking the niche region and nAlt-ORF sequences.

ponderance of G/C in the second position in high-usage codons, i.e. Z-score >0, and A/U in the second position in low-usage codons, i.e. Z-score ≤ 0 (Figure 6A). Overall, second position G/C utilization in nAlt-ORFs is more than twice that of A/U (Figure 6C). In contrast, about equal base utilization in the second position is observed in host ORFs, and a minor reversal of utilization is seen globally. The central role of the second position is highlighted by a reconsideration of the genetic code 'wheel' in which the second position replaces the first as the dominant, organizing circle (Figure 6D) (26). Briefly, a second position 'A' specifies primarily hydrophilic amino acids, excepting Arg specified by a second position 'G'. Thus, the low level of second position 'A' accounts for the low frequency of Lys, Asp, Glu and other hydrophilic amino acids in the nAlt-ORF proteins, and the high level of second position 'G' specifies the high frequency of Arg and other semi-polar amino acids.

To assess the 'quality' of nAlt-ORF codon usage, the codon adaptation index (CAI) was calculated using CAIcal which quantitates the relative use of optimal synonymous codons by comparison to a set of highly expressed genes, and can be used to predict the translation rate and level of expression of a gene (28,29). The CAI is scaled from 0 to 1, with an index of 1.0 indicating a very efficiently translated mRNA. The host ORFs exhibited a compact range of CAI scores with a mean of 0.81, consistent with efficient translatability (Figure 6E). The nAlt-ORFs exhibited a broader range of CAI scores with a mean of 0.74, somewhat lower than the parental ORFs, but still indicative of efficient translatability. As a control, alternate reading frames in the random group of 100 ORFs was examined (following manual removal of internal stop codons). The CAI was 0.70, significantly lower than the nAlt-ORF group. Based on yeast expression data, that shows a logarithmic relationship between protein abundance and CAI (30), the CAI differential corresponds to an ∼24% increase in translatability of the nAlt-ORFs. More importantly, the difference suggests codon usage in genes containing nAlt-ORFs is subject to non-neutral natural selection during evolution.

Computed structures of nAlt-ORF-derived proteins

Experimental evidence is required to validate the physiological significance of any individual nAlt-ORF. In some cases, evidence has been provided by mass spectrometric detection of peptide constituents, or by specific antibody recognition (Supplementary Table S1). Although essential, the approach is hindered by selection of appropriate cells or tis-

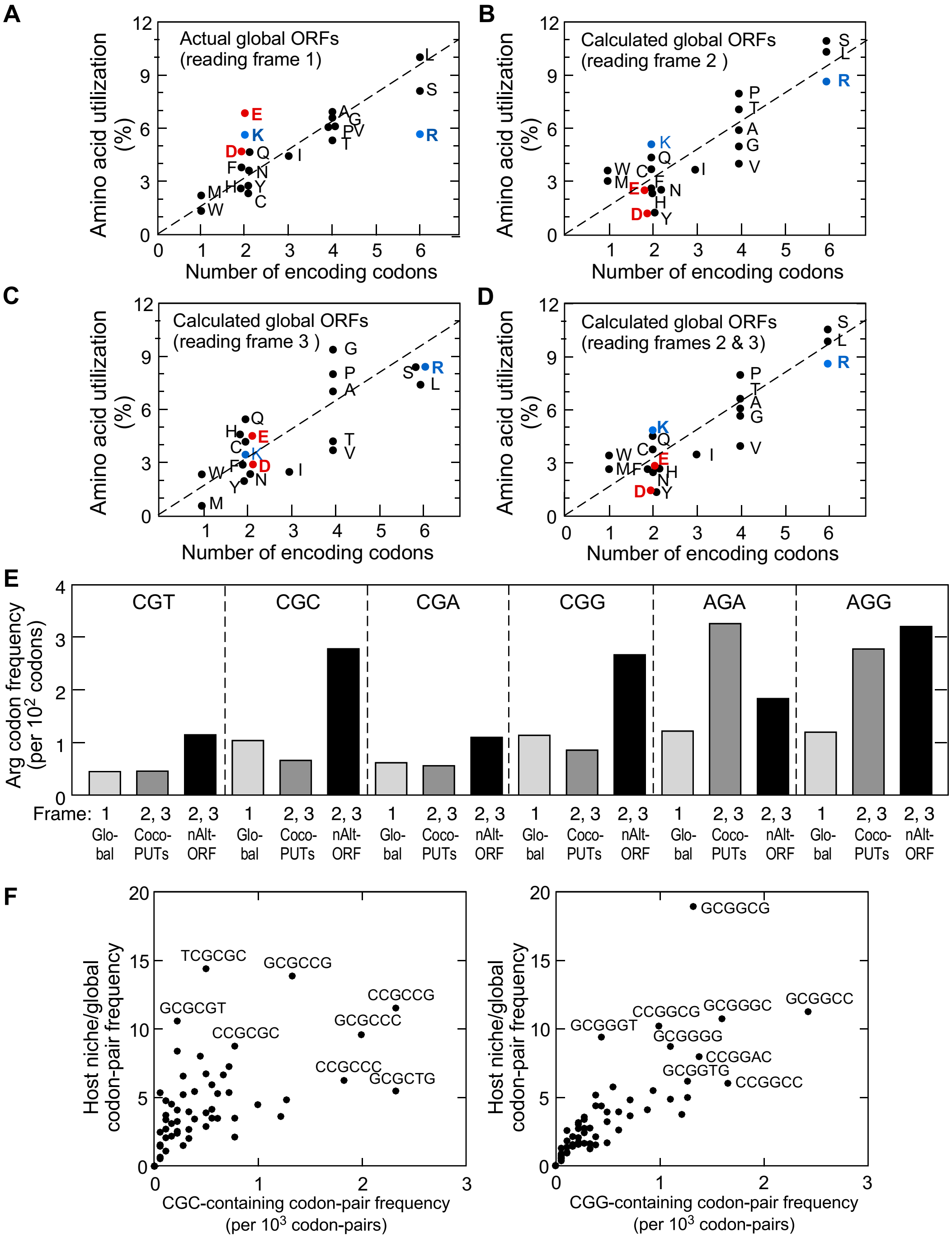


Figure 5. Genetic mechanisms underlying anomalous Asp, Glu and Arg utilization in nAlt-ORFs. (A–D) Amino acid utilization as a function of number of encoding codons. The amino acid frequencies were calculated from codon usage table of human 93 487 CDSs (**A**), from codon-pair frequencies in human CDSs using Equation (3) for reading frame 2 (**B**) Equation (4) for reading frame 3 (**C**), and by weighting reading frame 2 to 3 by the reading frame usage ratio of 7.5:1 (**D**). (**E**) Frequency of six Arg codons globally in RF 1 (light gray bars), global calculation from CoCoPuts codon-pair data in RF 2 and 3 with correction for RF usage ratio (medium gray bars), and calculation for nAlt-ORFs in RF 2 and 3 with correction for RF usage ratio (black bars). (**F**) Frequencies of the most abundant CGC- (left) and CGG- (right) containing codon pairs displayed as ratio of host niche-to-global.

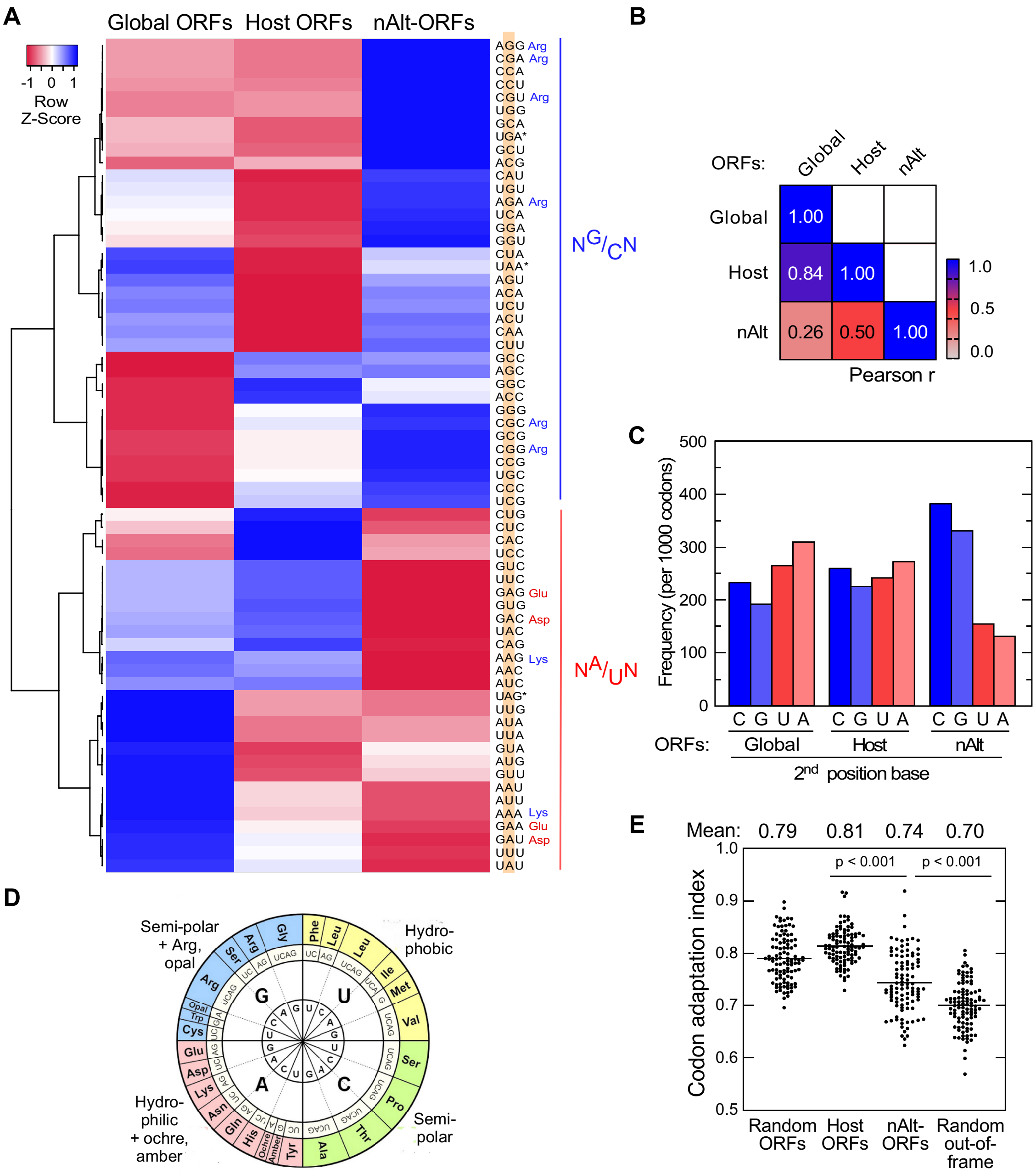


Figure 6. Differential codon usage in nAlt-ORFs. (**A**) Hierarchical clustering of codon usage in global, host, and nAlt ORFs. Charged codons are highlighted at right. (**B**) Pearson correlation coefficients for the codon usage in global, host, and nAltORFs data sets. (**C**) Second base frequencies plotted for codons in global, host, and nAlt ORF data sets. (**D**) Codon wheel representation of 64 codons arranged according to the second position. (**E**) Individual and mean CAI values of randomly selected ORFs, host ORFs, nAlt-ORFs, and out-of-frame sequences of the randomly selected ORFs.

sues, appropriate conditions (potentially including pathological condition), and detection limits. A less constrained approach, albeit also less rigorous, takes advantage of recent computational advances in molecular modeling. The sequences for several nAlt-ORF proteins were submitted to the AlphaFold server for structural analysis (31). The nAlt-ORF derived from Fus (Alt-Fus) was experimentally validated as a mitochondria-localized protein (14). The predicted structure is highly compact, consisting primarily of six α-helices of moderate length, i.e. 3–6 turns, and a cluster of basic residues (Figure 7A). Alt-KCNQ1, a predicted nuclear protein, likewise exhibits a highly compact globular form, consisting primarily of 5 short α-helices and 2 short β-sheets forming a ring of basic residues (Figure 7B). Alt-CHD1, predicted to be a secreted protein, is highly elongated with a remarkable predicted structure consisting almost entirely of α-helices, including a helix-turn-helix structure of 9 and 15 turns (Figure 7C). Although Alt-NOL3 is predicted to contain four α-helices, it is primarily nonstructured (Figure 7D) and similar in conformation to the

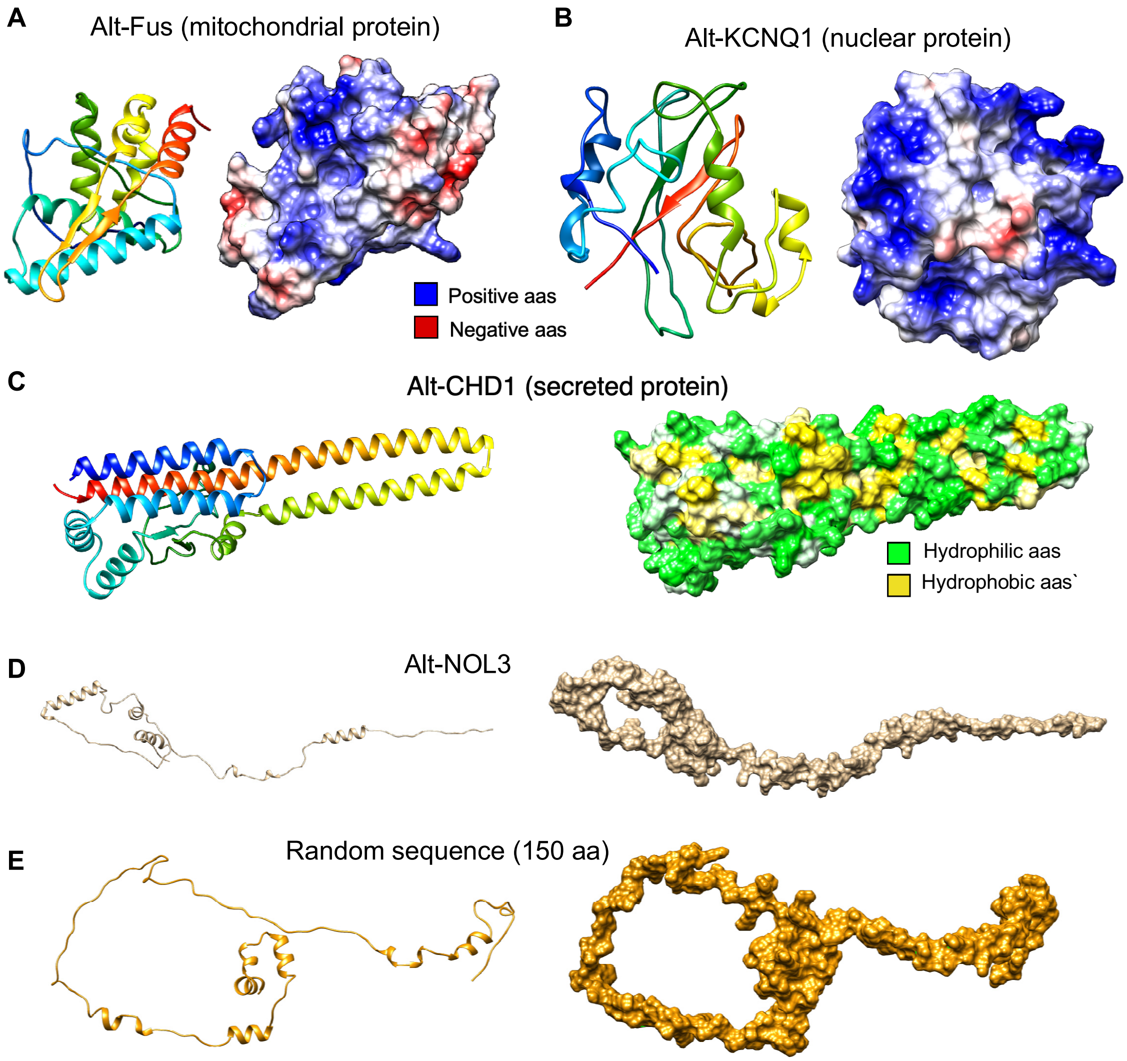


Figure 7. Computed structures of nAlt-ORFs. AlphaFold-generated three-dimensional structures displayed as ribbon (left) and surface (right) models. The number of sequences used for multiple sequence alignment and the average predicted local distance difference test (pLDDT) score, respectively, are indicated in parentheses following each gene. (**A**) Alt-Fus (38, 53.1) (**B**) Alt-KCNQ1 (178, 59.4), (**C**) Alt-CHD1 (15, 58.7), (**D**) Alt-NOL3 (91, 50.8) and (**E**) a randomly generated amino acid sequence. The ribbon diagrams (A-C) are colored in rainbow order from N-terminus to C-terminus. Surface representations of the models (A, B) are depicted with per residue estimation of electrostatic potential over the protein surface, ranging from negative (red) to positive (blue), or (C) hydrophobic (yellow) and hydrophilic (green), as computed by Chimera.

structure of a randomly generated sequence that contained 4 short α-helices, but is largely non-compact (Figure 7E).

DISCUSSION

A remarkable and unanticipated observation is the extraordinarily high p*I* of proteins encoded by the 101 nAlt-ORFs considered here. In addition to the markedly higher median pI in nAlt-ORFs, the range is also highly compacted, i.e. the vast majority of nAlt-ORF proteins exhibit p*I*'s between 10 and 13, with a median of 11.68, whereas the host-ORFs and randomly selected groups exhibit p*I*'s spread uniformly between 4 and 10. Focusing on the extremes 89/101 nAlt-ORF proteins have p*I*'s >10, whereas only 6/101 host ORFs and 0/100 random proteins exhibit p*I*'s >10. Quantitation of the amino acid content in the nAlt proteins revealed that the high median p*I* can be accounted for by the extremely high Arg content, 12.7% compared to 5.7% globally, and low Asp and Glu content, 1.4% and 2.1%, respectively, versus global content of 4.7% and 6.9%, respectively. The content of charged amino acids in the host ORFs was similar to the global content, indicating the anomalous content in the nAlt-ORFs is independent of the host ORFs. Although some differences were reported, a similar alteration in the content of charged amino acids was shown in a collection of dual-coding alternate reading frames generated by alternate splicing (27). Together, these findings raise the question of the mechanism underlying the major difference in charged amino acid content in nAlt-ORF proteins.

Two unrelated mechanisms contribute to anomalous Asp, Glu, and Arg content in nAlt-ORF-derived proteins: (i) codon usage in reading frame 1 is modulated when shifted into reading frames 2 and 3, and (ii) anomalous dinucleotide usage in Alt-OFs drives anomalous codon frequency. The first mechanism moderates the anomalous global codon usage of the three charged amino acids: Asp and Glu codons—2.18% and 2.51% for Asp, and 2.90% and 3.96% for Glu—all higher than the mean codon usage of 1.64% (1/61). Likewise, global codon usage of the six Arg codons—0.45%, 1.04%, 0.22%, 1.14%, 1.22% and 1.20%—are lower than mean codon usage. When these codons, with adjacent nucleotides, are shifted into reading frames 2 and 3, the anomalous usage is moderated, accounting for the diminished levels of Asp and Glu in nAlt-ORFS, but only partially accounting for the exceptionally high Arg content (5.67% globally, 8.62% calculated in frames 2 and 3, and 12.69% in nAlt-ORF proteins). The first mechanism applied to Arg is clearly visualized as the calculated increases in AGA and AGG codon utilization in reading frames 2, 3 (Figure 5E). The second mechanism accounts for the remaining increase in Arg content in nAlt-ORF proteins. Specifically, usage of the four CG-containing Arg codons are expressed at 2- to 4-fold higher levels in nAlt-ORF proteins compared to the calculated amount based on global codon-pair usage (Figure 5E). High CGn codon usage in nAlt-ORF proteins was unexpected given that CG is the least frequent dinucleotide in coding sequences of the human genome (Supplementary Figure S2A). Thus, the ∼2.6-fold higher CG dinucleotide frequency in nAlt-ORFs is the second major contributor to Arg codon utilization in these proteins. The reason for low CG dinucleotide level in human coding sequences is not known. One possibility is that mutation to CpG doublets occurs at a normal rate, but despite being synonymous with other codons are eliminated by natural selection due to a disadvantageous property (25). The high mutation rate of CpG by deamination of 5-methylcytosine to thymidine is potentially one such property. Possibly, the advantage of high Arg content in Alt-ORF proteins essentially neutralizes any disadvantage presented by CG-containing codons. Alternatively, the Alt-ORFs have appeared rather recently during evolution, and have not had the opportunity to back-mutate into less injurious codons. This possibility is supported by the proposed comparatively rapid rate of evolution of Alt-ORF sequences compared to CDSs (32).

To our knowledge, ours is the first analysis of nAlt-ORF proteins to recognize the significance of codon-pair frequencies as critical determinants of amino acid utilization of nAlt-ORFs, and takes advantage of the CoCoPUTs codon-pair usage table. Importantly, every codon in reading frames 2 and 3 in a nested ORF spans a codon-pair (Equations 3 and 4). Thus, every codon in each altered reading frame is specified by combinations of 3 surrounding nucleotides, i.e. 64 (4^3) combinations in each of the two alternative reading frames, equivalent to summation of 256 and 768 codon-pair frequencies for amino acids encoded by 2 and 6 codons, respectively. The recognition that codon-pair frequencies are at a variance from expectation based on codon usage bias has led to multiple applications (20). Codon-pair optimization has been used to improve translation rate and gene expression of recombinant proteins. For example, production of interferon-gamma was nearly twice as efficient following optimization of codon-pair usage compared to optimization by simple codon bias (33). In contrast, codon-pair deoptimization has been used for generation of attenuated vaccines (34). Lastly, codon-pair is under consideration as a mechanism underlying diseases caused by synonymous mutations (20,35). From this viewpoint, one can consider codon-pairs as a secondary 6-letter genetic code overlying the primary 3-letter code, specifying amino acid utilization in nAlt-ORFs.

Codon-pair analysis explains specific characteristics of Alt-ORF protein, and predicts new properties. For example, codon-pair analysis of AUG initiation codons predicts a marked bias towards reading frame 2, estimating ∼88% of nAlt-ORFs encoded in that frame. Importantly, in a previous compilation of 17 096 predicted human Alt-ORF proteins of length ≥24 amino acids, 83% were found in reading frame 2, confirming the predictive power of our analytic approach (12). Interestingly, a theoretical analysis of CUG codons based on codon-pair frequency predicts a +1 to –1 ratio of ∼2:1, substantially lower than the ∼7.5:1 ratio for AUG start codons in reading frame 2 compared to frame 3 (Supplementary Figure S3). As described above, the codon-pair-based analysis predicts an elevated pI, or equivalently, high ratio of basic-to-acidic amino acids. This prediction is supported by experimentally validated nAlt-ORFs, namely, Alt-FUS, Alt-ataxin-1, and Alt-PrP, that exhibit pIs of 11.53, 11.54 and 9.24, respectively (14–16). The extraordinarily high pI of the nAlt-ORFs has implications regarding cellular localization. Due to respiratory chain-driven outward transport of protons, the mitochondria inner membrane and matrix are highly electronegative, and mitochondrially-targeted proteins generally have a high pI (36,37). Moreover, the canonical mitochondrial targeting signal is a short peptide primarily dominated by basic residues with few acidic residues (37). Thus, the data suggest a predilection for nAlt-ORFS to localize at the mitochondrial inner membrane or matrix. Expression of few nAlt-ORFs have been experimentally validated. Among the best characterized is Alt-FUS, a 170-amino acid protein expressed in an alternative reading frame of *FUS* cds which encodes a nuclear RNA-binding protein (14). In contrast, Alt-FUS is expressed primarily in mitochondrial puncta following a cristae-like formation, suggesting binding to the inner membrane. A second Alt-ORF, Alt-PrP, is encoded out-of-frame by the *PRNP* gene that encodes the prion protein PrP (16). Alt-PrP did not appear in our search as it is only 64 amino acids long. Nonetheless, it likewise is observed primarily in mitochondria, but unlike Alt-FUS, Alt-PrP is localized in the outer mitochondrial membrane. A nAlt-ORF is expressed by *ATXN1*, the ataxin-1 gene associated with spinocerebellar ataxia type 1 (15). Alt-ATXN1 is a 186-amino acid protein that localizes in the nucleus bound to polyadenylated mRNA, which carries a dense negative charge. Thus, in these examples of endogenously expressed, well-characterized Alt-ORFs, their highly basic nature is a strong determinant of their localization, and potentially their function.

Experimental evidence is the gold-standard for endogenous expression of a given nAlt-ORF; however, the approaches described here can provide insights into the likelihood of expression and relative expression level. Certainly, a relationship to the Kozak sequence is informative, and our analysis suggests that nAlt-ORFs in reading frame 3, although in the minority, are likely to undergo translation-initiation more efficiently. A high CAI predicts a relatively fast rate of elongation—the three nAlt-ORFs in our collection with the highest CAIs are KRTAP9-1 (keratin associated protein 9–1), GRINA (glutamate ionotropic receptor NMDA type subunit associated protein 1), and SLC4A2 (solute carrier family 4 member 2), with CAI's of 0.92, 0.87 and 0.85, respectively. The group of nAlt-ORFs delineated here have a surprisingly high 8-to-1 overall ratio of Arg-to-Lys residues. The low Lys level reduces the likelihood of ubiquitination-mediated proteasomal degradation. Consistent with this observation, the Lys residue number in Alt-Atxn, Alt-PrP and Alt-FUS is 0, 0, and 3, respectively. A recently reported nAlt-ORF, Alt-B2R, encoded by the human bradykinin B2 receptor gene, also has no Lys residues; Alt-B2R is not included in our list because it is not translated from the first AUG after the canonical initiation site. Lastly, the folding structures determined by AlphaFold are likely to be informative as compact, structured proteins are likely to be less susceptible to proteolysis and thus more stable than extended, unstructured proteins.

Limitations of the approach and conclusions described here should be considered. Nested Alt-ORFs with length ≥150 codons, and initiating after the first AUG following the canonical ORF start codon were selected for inclusion in our study. The selection was based on likely preferential generation by leaky scanning of ribosomes—the predominant mode of translation of the mRNA from the 5′ terminus. The application of codon-pairs in our analysis precludes the inclusion of 5′ and 3′ untranslated regions. These criteria were applied to maximize the likelihood that the Alt-ORFs selected are likely to be expressed and functional. However, there is the possibility of a bias in which characteristics of smaller ORFs, ORFs within (or spanning) UTRs, or ORFs generated from downstream start codons might not be the same. Pseudogenes were also excluded from analysis since these genes are generally subjected to function-independent evolutionary selection pressures and might not exhibit Alt-ORF properties consistent with those generated by authentic, 'expressed' genes. Finally, any functional classification of Alt-ORFs assumes that the function is related to the parental ORF, a possibility consistent with several previous studies (14,15,32)

The approaches and results described here can be applied to other data sets. For example, the codon-pair calculations described here are independent of mRNA length, as supported by the relatively high pI of Alt-PrP despite its 64-amino acid length (16). The codon-pair calculations can also be used in evaluation of amino acid frequencies in other species as CoCoPUTs tables are available for al fully-sequenced species (20). As a caveat, the approach is unlikely to be predictive of Alt-ORFs that include substantial sequences within UTRs since these regions do not contain defined codons or codon-pairs. Nonetheless, these results provide a theoretical foundation that will facilitate design and analysis of experiments to extend our understanding of this novel class of proteins.

DATA AVAILABILITY

All data reported in this paper will be shared by the lead contact upon request. Any additional information required to reanalyze the data reported here is available upon request from the senior author. nAltORFs is implemented in Python and is compatible with Python v3.8 and higher. The development version of nAltORFs can be obtained from the GitHub repository (https://github.com/BlankenbergLab/nAltORFs). Released versions of nAltORFS are available from the Python Package Index (PyPI) as the nAlt-ORFs package, and is accessible at https://pypi.org/project/nAltORFs/). The current, stable version of nAltORFs (0.1.2) can be installed via PyPI using the following command: pip install nAltORFs. A conda distribution package has been added to bioconda: https://anaconda.org/bioconda/naltorfs. Anaconda users can install nAltORFs from the bioconda channel: conda install -c bioconda nAltORFs. Furthermore, Galaxy tools (38) have been created for each of the three nAltORFs commands, enabling web-based graphical user interface (GUI) usage and workflow access. These Galaxy tools (https://github.com/galaxyproject/tools-iuc/tree/master/tools/naltorfs) have been reviewed by the Intergalactic Utilities Commission (https://galaxyproject.org/iuc/) and are installable from the ToolShed (39). The Galaxy tools generated to identify nAlt-ORFs through the leaky scanning mechanism are freely accessible to researchers to analyze the mouse transcriptome in addition to the human transcriptome. Using these tools and others already available within Galaxy, researchers can explore nAlt-ORFs of their organisms of interest without having to install any software.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank all the Fox laboratory members for helpful comments and suggestions.

FUNDING

National Institutes of Health; National Institutes of Diabetes and Digestive Disease [R01 DK124203, R01 DK123236 to P.L.F.]; National Institutes of Aging [R01 AG067146 to P.L.F.]; National Institutes of Neurosciences [R01 NS124547 to P.L.F.].

Conflict of interest statement. Daniel Blankenberg has a significant financial interest in GalaxyWorks, a company that may have a commercial interest in the results of this research and technology. This potential conflict of interest has been reviewed and is managed by the Cleveland Clinic. The other authors declare no competing financial interests.

REFERENCES

1. Beadle,G.W. and Tatum,E.L. (1941) Genetic control of biochemical reactions in *neurospora*. *Proc. Natl. Acad. Sci. U.S.A.*, **27**, 499–506.
2. Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.

3. Gott,J.M. (2003) Expanding genome capacity via RNA editing. *C. R. Biol.*, **326**, 901–908.
4. Yao,P., Potdar,A.A., Arif,A., Ray,P.S., Mukhopadhyay,R., Willard,B., Xu,Y., Yan,J., Saidel,G.M. and Fox,P.L. (2012) Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell*, **149**, 88–100.
5. Di Giammartino,D.C., Nishida,K. and Manley,J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, **43**, 853–866.
6. Pelechano,V., Wei,W. and Steinmetz,L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.
7. Lee,S., Liu,B., Lee,S., Huang,S.X., Shen,B. and Qian,S.B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424–2432.
8. Eswarappa,S.M., Potdar,A.A., Koch,W.J., Fan,Y., Vasu,K., Lindner,D., Willard,B., Graham,L.M., DiCorleto,P.E. and Fox,P.L. (2014) Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell*, **157**, 1605–1618.
9. Brunet,M.A., Lucier,J.F., Levesque,M., Leblanc,S., Jacques,J.F., Al-Saedi,H.R.H., Guilloy,N., Grenier,F., Avino,M., Fournier,I. *et al.* (2021) OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res.*, **49**, D380–D388.
10. Orr,M.W., Mao,Y., Storz,G. and Qian,S.B. (2020) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.*, **48**, 1029–1042.
11. Brunet,M.A., Brunelle,M., Lucier,J.F., Delcourt,V., Levesque,M., Grenier,F., Samandi,S., Leblanc,S., Aguilar,J.D., Dufour,P. *et al.* (2019) OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.*, **47**, D403–D410.
12. Vanderperre,B., Lucier,J.F. and Roucou,X. (2012) HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database (Oxford)*, **2012**, bas025.
13. Pavesi,A., Vianelli,A., Chirico,N., Bao,Y., Blinkova,O., Belshaw,R., Firth,A. and Karlin,D. (2018) Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS One*, **13**, e0202513.
14. Brunet,M.A., Jacques,J.F., Nassari,S., Tyzack,G.E., McGoldrick,P., Zinman,L., Jean,S., Robertson,J., Patani,R. and Roucou,X. (2021) The FUS gene is dual-coding with both proteins contributing to FUS-mediated toxicity. *EMBO Rep.*, **22**, e50640.
15. Bergeron,D., Lapointe,C., Bissonnette,C., Tremblay,G., Motard,J. and Roucou,X. (2013) An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J. Biol. Chem.*, **288**, 21824–21835.
16. Vanderperre,B., Staskevicius,A.B., Tremblay,G., McCoy,M., O'Neill,M.A., Cashman,N.R. and Roucou,X. (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J.*, **25**, 2373–2386.
17. Brunet,M.A., Levesque,S.A., Hunting,D.J., Cohen,A.A. and Roucou,X. (2018) Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res.*, **28**, 609–624.
18. Mi,H., Muruganujan,A., Huang,X., Ebert,D., Mills,C., Guo,X. and Thomas,P.D. (2019) Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.*, **14**, 703–721.
19. Plant,E.P. and Dinman,J.D. (2006) Comparative study of the effects of heptameric slippery site composition on -1 frameshifting among different eukaryotic systems. *RNA*, **12**, 666–673.
20. Alexaki,A., Kames,J., Holcomb,D.D., Athey,J., Santana-Quintero,L.V., Lam,P.V.N., Hamasaki-Katagiri,N., Osipova,E., Simonyan,V., Bar,H. *et al.* (2019) Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *J. Mol. Biol.*, **431**, 2434–2441.
21. Kozak,M. (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.*, **196**, 947–950.
22. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
23. Briesemeister,S., Rahnenfuhrer,J. and Kohlbacher,O. (2010) YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*, **38**, W497–W502.
24. Kurotani,A., Tokmakov,A.A., Sato,K.I., Stefanov,V.E., Yamada,Y. and Sakurai,T. (2019) Localization-specific distributions of protein pI in human proteome are governed by local pH and membrane charge. *BMC Mol. Cell Biol.*, **20**, 36.
25. King,J.L. and Jukes,T.H. (1969) Non-Darwinian evolution. *Science*, **164**, 788–798.
26. Saier,M.H. Jr (2019) Understanding the genetic code. *J. Bacteriol.*, **201**, e00091-19.
27. Kovacs,E., Tompa,P., Liliom,K. and Kalmar,L. (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 5429–5434.
28. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
29. Puigbo,P., Bravo,I.G. and Garcia-Vallve,S. (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct*, **3**, 38.
30. Futcher,B., Latter,G.I., Monardo,P., McLaughlin,C.S. and Garrels,J.I. (1999) A sampling of the yeast proteome. *Mol. Cell. Biol.*, **19**, 7357–7368.
31. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Zidek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.
32. Samandi,S., Roy,A.V., Delcourt,V., Lucier,J.F., Gagnon,J., Beaudoin,M.C., Vanderperre,B., Breton,M.A., Motard,J., Jacques,J.F. *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife*, **6**, e27860.
33. Chung,B.K., Yusufi,F.N., Mariati,Yang, Y. and Lee,D.Y. (2013) Enhanced expression of codon optimized interferon gamma in CHO cells. *J. Biotechnol.*, **167**, 326–333.
34. Eschke,K., Trimpert,J., Osterrieder,N. and Kunec,D. (2018) Attenuation of a very virulent marek's disease herpesvirus (MDV) by codon pair bias deoptimization. *PLoS Pathog.*, **14**, e1006857.
35. McCarthy,C., Carrea,A. and Diambra,L. (2017) Bicodon bias can determine the role of synonymous SNPs in human diseases. *BMC Genomics*, **18**, 227.
36. Wisnovsky,S., Lei,E.K., Jean,S.R. and Kelley,S.O. (2016) Mitochondrial chemical biology: new probes elucidate the secrets of the powerhouse of the cell. *Cell Chem. Biol.*, **23**, 917–927.
37. Claros,M.G., Brunak,S. and von Heijne,G. (1997) Prediction of N-terminal protein sorting signals. *Curr. Opin. Struct. Biol.*, **7**, 394–398.
38. Jalili,V., Afgan,E., Gu,Q., Clements,D., Blankenberg,D., Goecks,J., Taylor,J. and Nekrutenko,A. (2020) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.*, **48**, W395–W402.
39. Blankenberg,D., Von Kuster,G., Bouvier,E., Baker,D., Afgan,E., Stoler,N., Galaxy,T., Taylor,J. and Nekrutenko,A. (2014) Dissemination of scientific software with galaxy toolshed. *Genome Biol.*, **15**, 403.