Màster Universitari
Anàlisi de Dades Òmiques / Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

The impact and function of C-terminal degrons and protein misfolding in the degradation of truncated proteins

by

Mònica Sánchez Guixé

Supervisor: Francisco Martínez Jiménez, PhD; Biomedical Genomics Laboratory, Institut de Recerca Biomèdica de Barcelona

Academic tutor: Josep Maria Serrat Jurado, Facultat de Ciències i Tecnologia, Universitat de Vic

Biosciences Department

University of Vic – Central University of Catalonia

15-09-2020

ABSTRACT

Premature Termination Codons (PTCs) are mainly generated after nonsense and frameshift alterations. The Nonsense Mediated Decay (NMD) system detects and degrades mRNAs containing PTCs. However, the efficiency of this surveillance mechanism is limited and some PTC-containing mRNAs can escape from degradation, thus potentially generating truncated protein forms. The Ubiquitin Proteasome System (UPS) is involved in the degradation of proteins either by detecting specific amino acid sequence motifs (degrons) or by detecting misfolding features of the protein. Degrons located at the c-terminal region of proteins (c-degrons) potentiate a rapid degradation of proteins, however, their impact on the degradation of proteins with truncated c-terminus remains unknown.

In this study, we employed two cancer datasets as natural experiments to study the degradation of truncated proteins: a dataset of more than 600 primary tumor biopsies from the Clinical Proteomics Tumor Analysis Consortium (CPTAC), and a dataset with more than 300 cancer-derived cell lines from CCLE with integrated genome, transcriptome and proteome information. In order to select those mutations producing truncated proteins, we annotated the NMD efficiency probability of all the PTCs in both datasets and classified them between NMD-skipping or NMD-triggering. To our surprise, not only NMD-skipping but also NMD-triggering showed decreased protein stability. We annotated all c-degrons instances in the predicted truncated proteins and analysed the changes in protein stability, but no significant differences nor tendencies were observed. We further explored whether the length of the truncated protein could impact protein stability and yet that was the case for NMD-skipping protein products but not for NMD-triggering, where protein stability decrease was independent of the mutation localization in the protein.

In conclusion, c-degrons failed to explain the overall protein stability decrease of truncated proteins. However, major protein sequence loss increases destabilisation but only for proteins from NMD-skipping PTCs. Results presented in this study suggest that truncated proteins could follow different degradation pathways depending on NMD efficiency.

INTRODUCTION

Protein homeostasis is an important process for the correct function and survival of cells. Multiple mechanisms take part in maintaining the correct folding, function and turnover of proteins. Among these, the Ubiquitin Proteasome System (UPS) controls the removal of proteins by targeting them for degradation (1).

The UPS consists of a multi-step reaction in which ubiquitin (Ub, small 8 KDa protein) polypeptide chains are conjugated in specific lysine residues of proteins to be degraded. Polyubiquitinated proteins are then recognised and hydrolysed into small peptides by the 26S subunit of the proteasome. The process of ubiquitination is performed by three enzymatic components: E1s (Ub-activating enzyme) and E2s (Ub-carrier or conjugating proteins), which carry the ubiquitin; and E3s (Ub-protein ligase), which serve as bridges between the targeted protein and the E2 enzyme to catalyze the addition of the Ub (**Fig. S1**) (1,2).

E3-ligases recognize proteins for degradation through two main mechanisms: 1) detecting degrons, which are 6-10 amino acid sequence motifs; and 2) detecting protein misfolding features. Degron detection is a system that controls the normal protein turnover of proteins, as different proteins have different half lifes and removal rates (2). On the other hand, the clearance of misfolded proteins is a protective system to avoid the aggregation of such misfolded forms which eventually can cause toxicity and cell death. In this case, proteins are targeted for degradation through the detection -either by specialized E3 ligases or through chaperones- of exposed hydrophobic residues which, in native protein conformation, would face the inner core of the folded protein (**Fig. S1**) (3).

Degrons hold high specificity for their cognate E3-ligases, depending on the nature and localization in the protein sequence (N-terminal, C-terminal or internal regions) (**Fig. S1**) (4). In 2018, two studies described a set of c-terminal degrons (c-degron) instances to be detected by cullin-RING E3-ligases (5,6). These c-degrons appear to be evolutionarily depleted from the eukaryotic proteome, suggesting an E3-ligase modulation of the proteome composition. Thus, c-degrons are suspected to be related to aberrant protein products (through proteolytic cuts or truncating alterations) detection and consequent degradation (4).

Premature Termination Codons (PTCs) appear mainly as a consequence of nonsense and frameshift mutations. mRNAs containing PTCs can be detected and degraded through the Nonsense Mediated Decay (NMD) (7). However, NMD is of limited efficacy depending on the exonic localisation of the PTC, as those proximal to the start codon, in the last exon, in a long exon —more than 200 nucleotides—, and within the last 50 nucleotides of the penultimate exon are predicted to be skipped by NMD, and truncated proteins can potentially be synthetized (**Fig. S1**) (8). The molecular mechanisms that control protein degradation of these protein products are still to be fully understood. Whether degron instances or aberrant misfolding affect the stability of truncated proteins remains to be explored.

In this study, we analyzed the implication of c-degron alterations and protein misfolding in the degradation of truncated proteins. We employed a dataset of 679 primary tumor biopsies from the Clinical Proteomics Tumor Analysis Consortium (CPTAC), and a dataset of 368 cancer-derived cell lines from the Cancer Cell Line Encyclopedia (CCLE) with integrated genome (WES), transcriptome (RNA-seq) and proteome (mass-spectrometry) information. We annotated the predicted NMD efficacy (9) from all detected PTCs and analyzed the stability of the truncated proteins by comparing the protein stability change of the mutated *vs.* the WT (methodology developed by Martínez-Jiménez *et al.* (10)). We annotated all c-degron instances in truncated proteins using the consensus c-degron motifs (4–6) and compared their protein stability change. Finally, to approximate the implication of protein misfolding in the degradation of the truncated proteins, we compared the stability change of truncated forms of different length, as we expect higher misfolding in truncating forms with higher protein sequence loss.

MATERIAL AND METHODS

A complete workflow of the data analysis is represented in **Figure S2**.

Data collection

Genomic (somatic mutations from the tumor, Whole Exome Sequencing), transcriptomic (RNA-seq) and proteomic (mass spectrometry) data from CPTAC of 7 tumor types (Lung Adenocarcinoma -LUAD-, Uterine Corpus Endometrial Carcinoma -UCEC-, Breast Cancer -BRCA-, Brain Cancer -GBM-, Clear Cell Renal Cell Carcinoma -CCRCC-, Colon Adenocarcinoma -COAD- and Ovarian Cancer -OV-) were downloaded from the CPTAC data portal (*https://cptac-data-portal.georgetown.edu/datasets*) on January 14th 2020.

Datasets containing genomic, transcriptomic and proteomic data from CCLE of 368 cell lines were downloaded from the Broad Institute data portal (*https://portals.broadinstitute.org/ccle/data*) on January 14th 2020.

RNA fold change, protein expression and stability change

Transcript counts (FPKM in CPTAC and RSEM and CCLE) were $\log_2$ and $\log_{10}$-transformed respectively. The fold change mRNA for each gene and sample was calculated by subtracting the median $\log_{10}$-FPKM of the WT samples of each tumor type to each $\log_{10}$-FPKM of each sample:

$$CPTAC:\ Fold-change\ mRNA\ =\ log_{10}(FPKM_{sample})\ -\ median\ (log_{10}(FPKM_{WT}))_{\ tumor\ type}$$

$$CCLE:\ Fold-change\ mRNA\ =\ log_{10}(RSEM_{sample})\ -\ median(log_{10}(FPKM_{WT}))_{\ tumor\ type}$$

Protein expression was normalized by subtracting the median protein expression of the WT per tumor type from the protein expression of each sample divided by the standard deviation:

$$Normalized\ protein\ expression\ =\ \frac{protein\ expression_{sample} - median(protein\ expression_{WT})_{\ tumor\ type}}{s.d.(protein\ expression)}$$

Protein stability change was calculated as previously described by (10). First, a relationship between protein and mRNA levels is established with the *WT* samples (regression line). Then, stability change is defined as the y-axis residual from the protein expression value to the regression line given a specific mRNA level (*raw residual*) corrected by the standard deviations (*s.d.*) (see **Fig. 1C** for an example with *ARID1A* gene):

$$Stability\ change\ =\ \frac{raw\ residual \cdot s.d.(mRNA)}{s.d.(protein\ expression)}$$

Variant Allele Frequency annotation

CPTAC Variant Allele Frequency (VAF) data were extracted from the Variant Call Format (VCF) files from CPTAC data portal.

No VAF data was available for CCLE dataset, although we expect cell samples to harbor higher purity than CPTAC given the *in vitro* culture of the cell lines *vs.* the heterogeneous nature of tumor biopsies.

Copy Number Alteration cutoff

Copy Number Alterations (CNA) with a ratio higher or equal to 2, or lower or equal to -2 were excluded from the analysis in CCLE dataset. CNA information was not available in CPTAC dataset.

NMD-score **annotation**

Genome-wide predictions of NMD efficacy (*NMD-score*) were downloaded from *https://figshare.com/articles/NMDetective/7803398* (9) on April 6th 2020. The corresponding *NMD-score* for each PTC of CPTAC and CCLE datasets were annotated according to the CDS position of the mutation for both nonsense and frameshift alterations using the canonical transcript sequences.

Mutated coding sequences, *in silico* translation and c-degron annotation

WT coding sequences of canonical transcripts were downloaded from ENSEMBL using BioMart (*https://m.ensembl.org/info/data/biomart/index.html*). Mutated sequences from nonsense and frameshift alterations were generated by replacing the reference with the alternate nucleotides at the specified CDS position in each gene from the VCF files. *In silico* translation of the mutated coding sequences was performed and the resulting protein sequences were annotated. Truncated protein sequences were analyzed for the presence of consensus c-degrons (4) and the corresponding c-degrons were annotated. C-terminal amino acids (i.e., last amino acid) were annotated for all sequences.

Relative mutation position annotation

Mutation positions in each protein were transformed as relative positions in the sequence (from 0, N-terminal, to 1, C-terminal). Proteins were grouped in quarters (i.e., Q1: 0 - 0.25; Q2: 0.25 - 0.50; Q3: 0.50 - 0.75; Q4: 0.75 - 1) in order to compare proteins with higher (Q1) or lower (Q4) protein sequence loss upon truncation.

Statistical analysis

All statistical analysis was performed using the matplotlib library 3.3.0 in Python 3.8.3. Mann-Whitney-Wilcoxon two-sided tests were performed for two group comparisons, and Bonferroni correction was applied when more than 2 groups were compared. Pearson correlation was performed for regression analysis. Kruskal-Wallis H-test was performed to evaluate possible differences among c-degrons. P-value annotation: ns, 0.05 < p-value ≤ 1; *, 0.01 < p-value ≤ 0.05; **, 0.001 < p-value ≤ 0.01; ***, 0.0001 < p-value ≤ 0.001; ****, p-value ≤ 0.0001.

RESULTS

Truncating proteins show decreased stability

Proteogenomic data from 679 tumor samples of 7 tumor types from CPTAC dataset and 368 cell line samples from CCLE dataset (**Fig. 1A**) were used in this analysis. Each sample is sequenced at genomic (Whole Exome Sequencing, WES), transcriptomic (RNA-seq) and proteomic (Mass Spectrometry) levels (**Fig. 1B**). To evaluate the changes at protein level, we calculated the protein stability change from the transcript and protein quantifications (see methods section for further details).

Stability change was defined as the y-axis distance from an observed protein quantification to the expected. **Figure 1C** shows the representation of stability change in *ARID1A* gene among UCEC tumors, where most of the nonsense and frameshift indels show a decrease in protein stability. Overall, truncating alterations from *ARID1A* presented a significant decrease in stability change compared to the wild type counterparts, while non-truncating alterations did not show a significant difference.

In order to explore the protein stability change in truncating alterations, we calculated the Area Under Curve (AUC) of the ROC curves for each gene of stability change classifying wild type (WT) and truncating (nonsense and frameshift) alterations (x-axis), and the difference of the stability change of the truncating alterations vs. WT (y-axis) (**Fig. 1D**). The resulting volcano plot showed that most of the truncating alterations harbor a decreased protein stability in both CPTAC and CCLE datasets. Among all genes, *ARID1A* showed the most prominent difference in stability change with a high number of truncating mutations in both datasets (80 in CPTAC and 69 in CCLE).

Furthermore, some genes presented a correlation between the VAF and stability change in CPTAC dataset (**Fig. S3**), indicating that the relative abundance of the WT form in mutated samples could affect the quantifications in RNA-seq and mass spectrometry. Therefore, we validated each step of our analysis with a CPTAC sample selection with high VAF (>0.3).

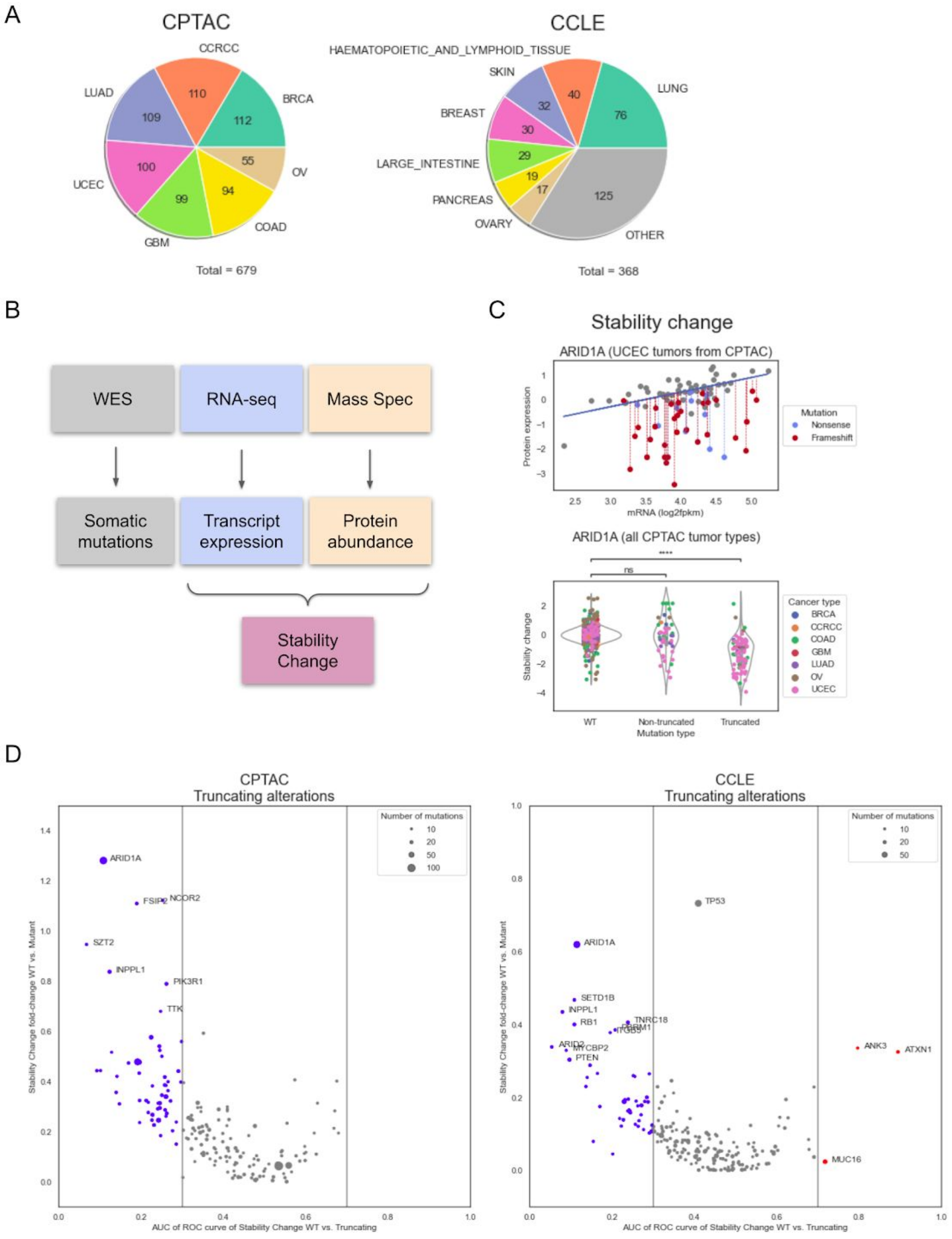

A
CPTAC
CCRCC
LUAD
BRCA
OV
COAD
GBM
UCEC
110
112
55
94
99
100
109
Total = 679
CCLE
HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
SKIN
LUNG
BREAST
LARGE_INTESTINE
PANCREAS
OVARY
OTHER
40
76
32
30
29
19
17
125
Total = 368
B
WES
RNA-seq
Mass Spec
Somatic mutations
Transcript expression
Protein abundance
Stability Change
C
Stability change
ARID1A (UCEC tumors from CPTAC)
Protein expression
mRNA (log2fpkm)
Mutation
Nonsense
Frameshift
ARID1A (all CPTAC tumor types)
****
ns
Stability change
Cancer type
BRCA
CCRCC
COAD
GBM
LUAD
OV
UCEC
WT
Non-truncated
Truncated
Mutation type
D
CPTAC
Truncating alterations
Stability Change fold-change WT vs. Mutant
AUC of ROC curve of Stability Change WT vs. Truncating
Number of mutations
10
20
50
100
ARID1A
FSIP2
NCOR2
SZT2
INPPL1
PIK3R1
TTK
CCLE
Truncating alterations
Stability Change fold-change WT vs. Mutant
AUC of ROC curve of Stability Change WT vs. Truncating
Number of mutations
10
20
50
TP53
ARID1A
SETD1B
INPPL1
RB1
TNRC18
PBRM1
ITGB3
ARID2
MYCBP2
PTEN
ANK3
ATXN1
MUC16

Figure 1. Truncating alterations show decreased protein stability. A) 679 tumor samples from CPTAC and 368 cell line samples from CCLE were used in this analysis. B) Genomic (WES), transcriptomic (RNA-seq) and proteomic (mass spectrometry) data was available for both datasets. Stability change was calculated from mRNA and protein quantification data. C) Stability change representation as an mRNA-protein relationship in *ARID1A* gene with UCEC tumors. Regression line indicates the mRNA-protein ratio for the WT. Grey points indicate WT, blue indicate nonsense mutations and red indicate frameshift indels. The residual from the observed protein expression to the expected (red/blue dashed lines in y-axis) is defined as the stability change. Bottom panel compares the stability change between the WT and the non-truncating (Non_trunc) and truncating (Trunc, nonsense and frameshift) alterations. Colors indicate tumor type of the CPTAC dataset. D) Genes with at least 15 truncating (nonsense and frameshift) alterations in CPTAC (left panel) and CCLE (right panel) were selected. Area Under Curve (AUC, x-axis) was calculated from ROC curves of stability change to classify between WT and truncating alterations. Stability change fold-change (y-axis) was calculated subtracting the stability change median of the truncating alterations from the median stability change from the WT per gene. Point size indicates the number of truncating mutations per gene. Represented in blue are the AUC inferior to 0.3 (truncating alterations that decrease stability) and in red the AUC superior to 0.7 (truncating alterations that increase stability). Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction.

Both NMD-skipping and NMD-triggering PTCs show decreased protein stability

We employed the *NMD score* from a published work that uses NMDetectiveA, an algorithm to calculate the likelihood of a PTC-bearing mRNA to be detected and degraded by NMD (9) (**Fig. 2A**). We annotated the *NMD score* for each of the PTCs present in CPTAC and CCLE datasets, where scores above 0.52 indicate NMD-triggering PTCs and scores below 0.25 denote those skipping NMD (NMD-skipping).

We evaluated the mRNA fold-change and stability change of each mutation type (synonymous, missense, nonsense and frameshift) and WT mRNAs in CPTAC dataset (**Fig. 2B**). Results corroborated that mRNAs with NMD-skipping PTCs (nonsense and frameshift) do not present a significant difference in mRNA fold-change while mRNAs with NMD-triggering PTCs have a significant down-regulation of the mRNA levels. Intriguingly, a decrease in protein stability was observed in both NMD subsettings. This was unexpected, as protein synthesis (and therefore, decrease in protein stability change) is assumed only on those mRNAs that skip NMD surveillance, while mRNAs targeted by NMD would be degraded and therefore no protein product is expected. However, this data points to the presence of truncated proteins, with decreased stability, regardless of NMD efficiency. These observations were further confirmed in the CPTAC subsetting with high VAF and in CCLE dataset (**Fig. S4**).

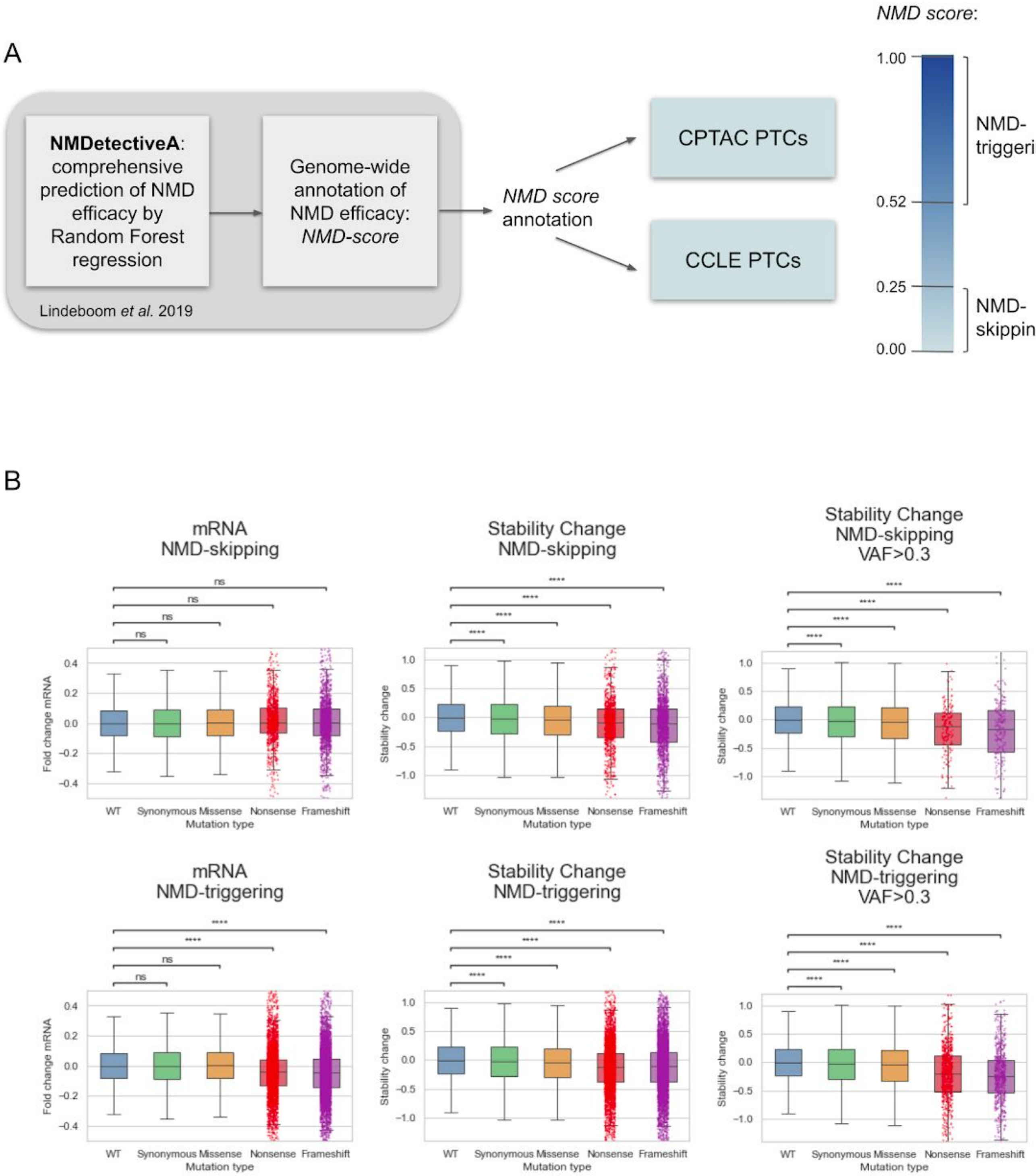


Figure 2. Both NMD-skipping and NMD-triggering PTCs show decreased protein stability. A) Genome-wide annotation of NMD efficacy (NMD-score) was obtained from Lindeboom *et al.* (9). NMD-score for each PTC in CPTAC and CCLE datasets were annotated. NMD-scores above 0.52 were classified as NMD-triggering and NMD-scores below 0.25 were classified as NMD-skipping. NMD-scores between 0.25 and 0.52 were not included in the analysis. Thresholds were extracted from Lindeboom *et al.* (9). B) mRNA fold-change (left panel), stability change (middle panel) and stability change on samples with high VAF (>0.3) (right panel) comparing WT, synonymous, missense and NMD-skipping nonsense and frameshift (above) or NMD-triggering nonsense and frameshift (below) alterations. Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction.

C-degrons fail to explain the overall stability change in truncated proteins

We hypothesized that the decreased stability observed across truncated proteins could be caused by the appearance of new c-degrons (**Fig. 3A**). Using the canonical transcript sequences, we generated the mutated coding sequences of all nonsense and frameshift alterations, performed an *in silico* translation, and annotated the truncated proteins containing a c-degron (**Fig. 3B**). Results showed no significant differences in stability change between truncated proteins with and without c-degrons, neither in NMD-skipping nor NMD-triggering subsettings in CPTAC (**Fig. 3C**) and CCLE (**Fig. S5A**) datasets. We further evaluated separately the truncated forms by mutation type (nonsense and frameshift), either all or selecting mutations with high VAF (**Fig. S5C**) and only NMD-triggering frameshift mutations in CPTAC with high VAF were significant, although this difference was not detected in any other comparison.

We next analyzed the stability change among the different c-degrons, but Kruskal-Wallis H-test did not show significant differences neither in CPTAC (**Fig. 3D**), CCLE (**Fig. S5B**) nor selecting high VAF CPTAC mutations (**Fig. S5D**). However, there was certain tendency for variability in stability change among c-degron types, especially in NMD-skipping subsetting, where some c-degrons showed an increased protein stabilization (e.g. RXXG) while others presented a decreased stability (e.g. AX). Therefore, we wondered whether the affinity for degradation was different for each c-degron. For that, we evaluated if the variability between c-degrons was consistent between mutation type (nonsense and frameshift) and between datasets (CPTAC and CCLE) (**Fig. 3E**). In NMD-skipping subsetting, results showed that only proteins with c-degrons from nonsense mutations have a significant correlation between CPTAC and CCLE datasets (p-value = 0.0082). RXXG was the only consistent c-degron across regressions, showing an increased stabilization, while the other c-degrons did not show similar tendencies in all four comparisons. In NMD-triggering subsetting, no significant correlation nor tendency was found in any regression evaluated. In summary, we observed that the generation of c-degrons cannot solely explain the observed changes in stability of truncated protein products.

In order to explore other c-terminal signals that could determine higher degradation, we annotated the last amino acid of each truncated protein and analyzed the stability change per amino acid in the c-terminal (**Fig. S6A**). Similarly as with c-degrons, there was certain variability across amino acids, especially in NMD-skipping subsetting, where histidine (H) and arginine (R) showed higher stability while proline (P) and threonine (T) presented lower stability. However, when evaluating if this variability was consistent between mutation type (nonsense and frameshift) and datasets (CPTAC and CCLE), no significant correlation nor tendency in any of the four regressions was found (**Fig. S6B**).

The same correlation analysis was performed with the high VAF mutations in CPTAC, and no significant correlation nor tendency was found neither for c-degrons (**Fig. S7A**) nor for last amino acid (**Fig. S7B**).

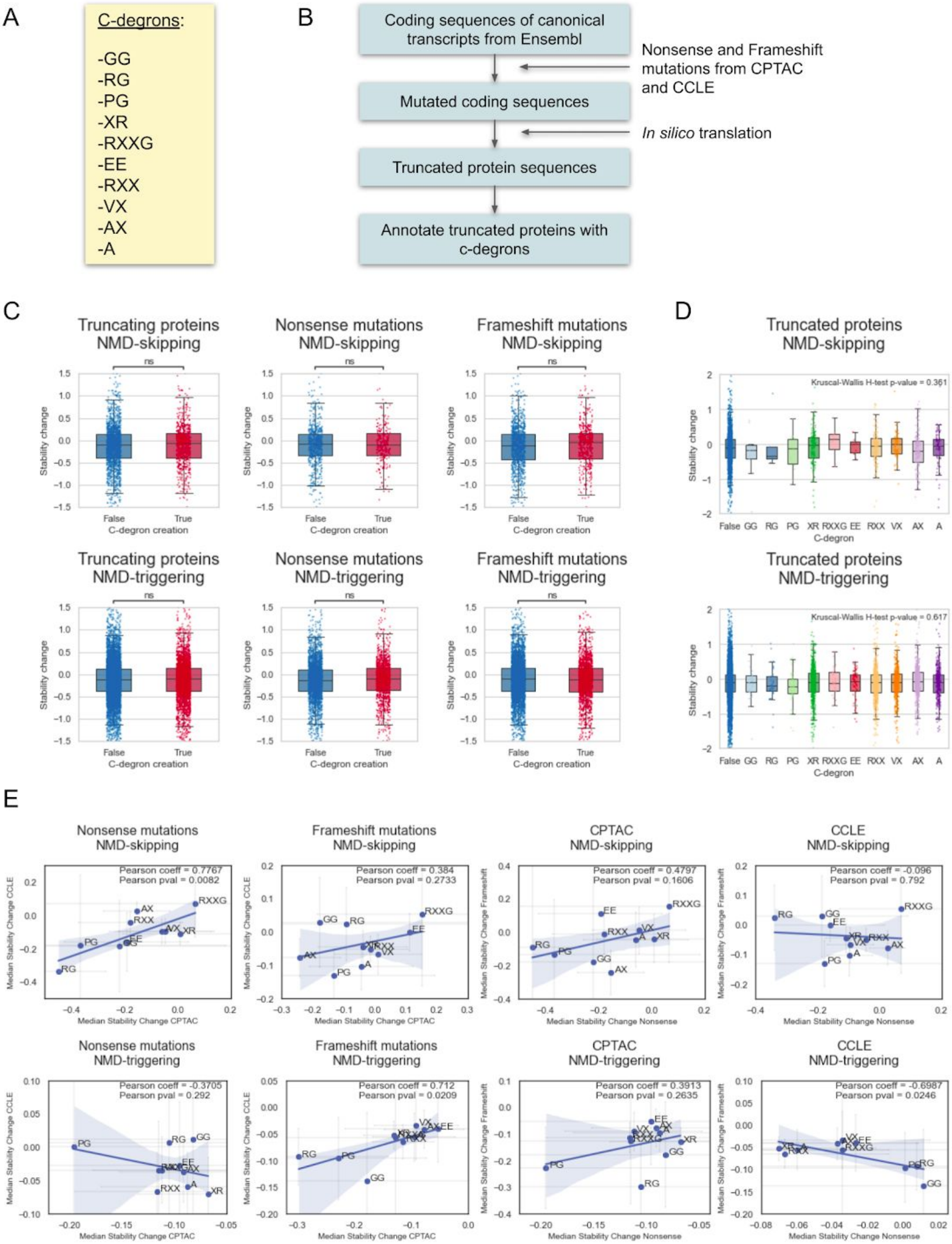

A
C-degrons:
-GG
-RG
-PG
-XR
-RXXG
-EE
-RXX
-VX
-AX
-A
B
Coding sequences of canonical transcripts from Ensembl
Nonsense and Frameshift mutations from CPTAC and CCLE
Mutated coding sequences
In silico translation
Truncated protein sequences
Annotate truncated proteins with c-degrons
C
Truncating proteins NMD-skipping
Nonsense mutations NMD-skipping
Frameshift mutations NMD-skipping
Truncating proteins NMD-triggering
Nonsense mutations NMD-triggering
Frameshift mutations NMD-triggering
ns
Stability change
False
True
C-degron creation
D
Truncated proteins NMD-skipping
Kruscal-Wallis H-test p-value = 0.361
Truncated proteins NMD-triggering
Kruscal-Wallis H-test p-value = 0.617
Stability change
False GG RG PG XR RXXG EE RXX VX AX A
C-degron
E
Nonsense mutations NMD-skipping
Pearson coeff = 0.7767
Pearson pval = 0.0082
Frameshift mutations NMD-skipping
Pearson coeff = 0.384
Pearson pval = 0.2733
CPTAC NMD-skipping
Pearson coeff = 0.4797
Pearson pval = 0.1606
CCLE NMD-skipping
Pearson coeff = -0.096
Pearson pval = 0.792
Nonsense mutations NMD-triggering
Pearson coeff = -0.3705
Pearson pval = 0.292
Frameshift mutations NMD-triggering
Pearson coeff = 0.712
Pearson pval = 0.0209
CPTAC NMD-triggering
Pearson coeff = 0.3913
Pearson pval = 0.2635
CCLE NMD-triggering
Pearson coeff = -0.6987
Pearson pval = 0.0246
Median Stability Change CCLE
Median Stability Change CPTAC
Median Stability Change Frameshift
Median Stability Change Nonsense

Figure 3. C-degrons fail to explain the overall stability change in truncated proteins. A) The list of the 10 canonical c-degrons was obtained from Varshavsky *et al*. (4). B) C-degron annotation workflow. Coding sequences from WT canonical transcripts were downloaded brom Ensembl using BioMart; using the information (chromosome, genomic position, reference and alternate nucleotides) of nonsense and frameshift mutations in CPTAC and CCLE, we created the mutated coding sequences, which were subsequently translated *in silico*. C-degrons and c-terminal amino acids were annotated from the obtained protein sequences. C) Stability change comparison between mutated proteins with (True) or without (False) c-degrons in all truncating proteins (left), or nonsense (middle) and frameshift (right) separately, in NMD-skipping subsetting (above) or triggering (below). D) Stability change comparison between truncated proteins without c-degrons (False) and truncated proteins for each of the c-degron instances. E) Median stability change correlation for each c-degron between in CPTAC (x-axis) and CCLE (y-axis) datasets of nonsense mutations (left) and frameshift (middle-left) alterations; median stability change correlation for each c-degron between nonsense (x-axis) and frameshift (y-axis) alterations of CPTAC (middle right) and CCLE (right) datasets; for NMD-skipping (above) and triggering (below) subsettings. Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction. Pearson correlations were performed for regression plots. Error bars from regression plots indicate CI95.

Short truncating proteins from mRNAs that skip NMD degradation present lower protein stability

We then asked whether the impact on protein misfolding caused by their truncation could explain their decreased stability. However, evaluating the misfolding of a protein from sequencing data can be challenging. For that, we employed a simplistic approximation to predict the proper folding of a protein. Since a correctly folded protein needs the complete sequence with its domains and regions, we hypothesized that truncated proteins lacking major elements of the sequence would tend to present higher misfolding and therefore degradation. Hence, we aimed to compare the change in stability between different positions of the truncating mutation in the protein sequence, as we expect to see higher misfolding in proteins missing most of the protein composition (thus, truncations further from c-terminal). For this, we annotated all mutation positions of each gene relative to the length of the protein, and grouped the truncating alterations in 4 regions or quarters (Q1-4), where Q1 would be the most shortened proteins and Q4 would contain those proteins with the least truncation (**Fig. 4A**).

Results showed that, in NMD-skipping subsetting, proteins with a truncation in the first quarter of the protein (Q1) showed a decrease in protein stability compared with the other protein quarters. This difference was significant between Q1 and Q3/Q4 in CPTAC dataset, although this significance is lost when analyzing the mutation types separately, albeit the tendency is maintained (**Fig. 4B**). On the other hand, in NMD-triggering subsetting, this decreased protein stability was not present. In fact, stability change of truncated proteins presented very similar levels across relative mutation locations. These results were as well observed in CCLE dataset (**Fig. S8A**) and in CPTAC dataset with high VAF (**Fig. S8B**).

To further demonstrate this increased destabilization of shorter truncated proteins in NMD-skipping subsetting, we compared the stability change of each protein quarter between synonymous, missense, nonsense and frameshift mutations. This analysis showed decreased protein stability in the first quarter, although the difference was as well significant for the last quarter and for frameshift mutations in the second quarter. Still, Q1 presented the highest decrease in stability (**Fig. 4C**). Overall, we observed that shorter protein products showed higher protein destabilization in NMD-skipping but not in NMD-triggering subsetting.

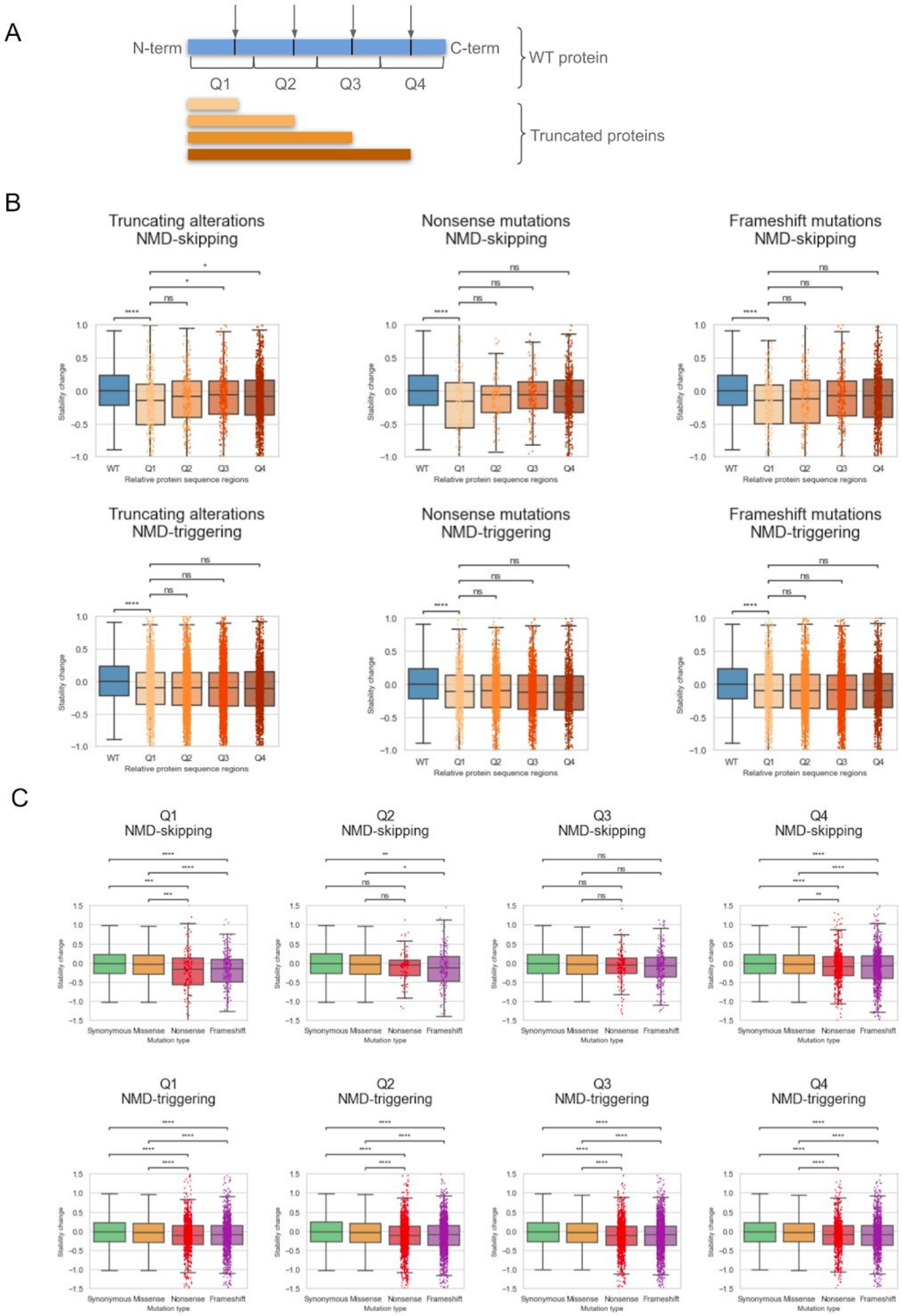

A
N-term
C-term
WT protein
Q1
Q2
Q3
Q4
Truncated proteins
B
Truncating alterations NMD-skipping
Nonsense mutations NMD-skipping
Frameshift mutations NMD-skipping
Truncating alterations NMD-triggering
Nonsense mutations NMD-triggering
Frameshift mutations NMD-triggering
****
ns
*
ns
Stability change
1.0
0.5
0.0
−0.5
−1.0
WT
Q1
Q2
Q3
Q4
Relative protein sequence regions
C
Q1 NMD-skipping
Q2 NMD-skipping
Q3 NMD-skipping
Q4 NMD-skipping
Q1 NMD-triggering
Q2 NMD-triggering
Q3 NMD-triggering
Q4 NMD-triggering
****
***
**
*
ns
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Synonymous
Missense
Nonsense
Frameshift
Mutation type

Figure 4. Truncated proteins from NMD-skipping mutations with high protein sequence loss present lower stability. A) Mutation positions are relativized to the length of the protein and grouped into 4 regions or quarters: Q1 (0 - 0.25), Q2 (0.25 - 0.50), Q3 (0.50 - 0.75) and Q4 (0.75 - 1). Q1 represents truncated proteins with high sequence loss and Q4 represents truncated proteins with minor sequence loss. B) Stability change comparison between WT and truncated proteins from the 4 defined quarters in all truncating alterations (left), and in nonsense (middle) and frameshift (right) separately, in NMD-skipping (above) and NMD-triggering (below). C) Stability change comparison between synonymous, missense, nonsense and frameshift mutations located in Q1 (left), Q2 (middle left), Q3 (middle right) and Q4 (right), for NMD-skipping (above) and NMD-triggering (below) truncating mutations. Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction.

DISCUSSION

The detection of PTCs is highly regulated in the cell to avoid the presence of abnormal protein forms that could harm the correct function of cells. NMD system detects and degrades mRNA containing PTCs; however, this system is not 100% efficient and some PTCs may escape from NMD surveillance, and therefore the truncated protein may be translated (8,11). Here, we have analyzed some possible mechanisms for the degradation of these truncated products.

First, we observed that truncating mutations produce an overall decrease in protein stability. These results indicate that there is a mechanism, beyond NMD surveillance, that negatively regulates protein levels of truncated protein forms, either by decreasing protein synthesis and/or by increasing degradation (12). At first, we speculated that this stability change was due to the protein products from mRNAs skipping NMD, but in fact, we observed decreased stability in both skipping and triggering cases. This indicates that mRNAs with NMD-triggering PTCs are not fully degraded; these might harbor some degree of protein translation and consequent protein regulation. This is in accordance with some studies pointing that NMD, despite detecting efficiently a PTC, it does not completely degrade the mRNA, and that a protein degradation is occurring through UPF1 (13–16).

The fact that NMD-triggering mutations generate a truncated protein suggests that the processes involved in the degradation of proteins from NMD-skipping and triggering PTCs are not the same —or that the rules driving the degradation do not necessarily match. Our data shows that NMD-skipping truncating mutations generate a protein that follows a degradation pathway related to protein misfolding, as shorter protein products are highly degraded compared to longer lengths. This is not the case for NMD-triggering truncating mutations: protein degradation is independent of the localization of the truncation. Whether the degradation of proteins from NMD-triggering PTCs is related to the E3-ligase activity of UPF1 or other NMD elements needs to be explored (14,15).

In this work we did not observe an implication of the c-terminal composition in the overall degradation of truncated proteins. However, whether this could be the rule for some specific cases, or be part of the degradation process in collaboration with other degradation mechanisms remains unknown.

Moreover, it is intriguing whether truncated forms of c-degron-bearing WT proteins might present higher stability and if this could be implicated in some carcinogenic process. We previously described that alterations in internal degrons can impair E3-ligase recognition causing an increased protein stability in oncogenes (10). Therefore, analyzing long truncated proteins that disrupt c-degrons (or internal degrons) but still maintain their main function could bring novel mechanisms in carcinogenesis.

In conclusion, this analysis showed that c-degrons do not determine the overall protein stability change in truncated proteins. Instead, we observed a decreased protein stability in shorter truncating forms for proteins from NMD-skipping mutations. Interestingly, this data showed protein stability change as well in NMD-triggering mutations, indicating a possible second layer of quality control after mRNA surveillance, which follows a different degradation pattern as proteins from NMD-skipping mutations. Further studies exploring the molecular mechanisms of these degradation processes are highly encouraged in order to finally decipher the cellular pathways implicated in the removal of truncated protein forms.

REFERENCES

1. Lecker SH, Goldberg AL, Mitch WE. Protein degradation by the ubiquitin-proteasome pathway in normal and disease states. J Am Soc Nephrol JASN. julio de 2006;17(7):1807-19.
2. Mészáros B, Kumar M, Gibson TJ, Uyar B, Dosztányi Z. Degrons in cancer. Sci Signal. 14 de marzo de 2017;10(470).
3. Shiber A, Ravid T. Chaperoning Proteins for Destruction: Diverse Roles of Hsp70 Chaperones and their Co-Chaperones in Targeting Misfolded Proteins to the Proteasome. Biomolecules. 17 de julio de 2014;4(3):704-24.
4. Varshavsky A. N-degron and C-degron pathways of protein degradation. Proc Natl Acad Sci. 8 de enero de 2019;116(2):358-66.
5. Koren I, Timms RT, Kula T, Xu Q, Li MZ, Elledge SJ. The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons. Cell. 14 de junio de 2018;173(7):1622-1635.e14.
6. Lin H-C, Yeh C-W, Chen Y-F, Lee T-T, Hsieh P-Y, Rusnac DV, et al. C-Terminal End-Directed Protein Elimination by CRL2 Ubiquitin Ligases. Mol Cell. mayo de 2018;70(4):602-613.e3.
7. Maquat LE. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. RNA. julio de 1995;1(5):453-65.
8. Lindeboom RGH, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat Genet. 2016;48(10):1112-8.
9. Lindeboom RGH, Vermeulen M, Lehner B, Supek F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nat Genet. 2019;51(11):1645-51.
10. Martínez-Jiménez F, Muiños F, López-Arribillaga E, Lopez-Bigas N, Gonzalez-Perez A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. Nat Cancer. enero de 2020;1(1):122-35.
11. Litchfield K, Reading JL, Lim EL, Xu H, Liu P, Al-Bakir M, et al. Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. Nat Commun. 30 de julio de 2020;11(1):3800.
12. Karamyshev AL, Karamysheva ZN. Lost in Translation: Ribosome-Associated mRNA and Protein Quality Controls. Front Genet [Internet]. 2018 [citado 21 de agosto de 2020];9. Disponible en: https://www.frontiersin.org/articles/10.3389/fgene.2018.00431/full
13. Kuroha K, Tatematsu T, Inada T. Upf1 stimulates degradation of the product derived from aberrant messenger RNA containing a specific nonsense mutation by the proteasome. EMBO Rep. noviembre de 2009;10(11):1265-71.

14. Feng Q, Jagannathan S, Bradley RK. The RNA Surveillance Factor UPF1 Represses Myogenesis via Its E3 Ubiquitin Ligase Activity. Mol Cell. 20 de julio de 2017;67(2):239-251.e6.
15. M Plank T-D, Wilkinson MF. RNA Decay Factor UPF1 Promotes Protein Decay: A Hidden Talent. BioEssays News Rev Mol Cell Dev Biol. enero de 2018;40(1).
16. Raxwal VK, Simpson CG, Gloggnitzer J, Entinze JC, Guo W, Zhang R, et al. Nonsense-mediated RNA Decay Factor UPF1 is Critical for Post-transcriptional and Post-translational Gene Regulation in Arabidopsis. Plant Cell. 14 de julio de 2020;

APPENDIX

Code availability

The necessary code to perform this analysis is in the following github repository: https://github.com/msguixe/TFM_MSG. It includes the post-processing and data analysis parts of the workflow, performed by M. Sánchez-Guixé (see **Fig. S2** for more information).

Supplementary Figures

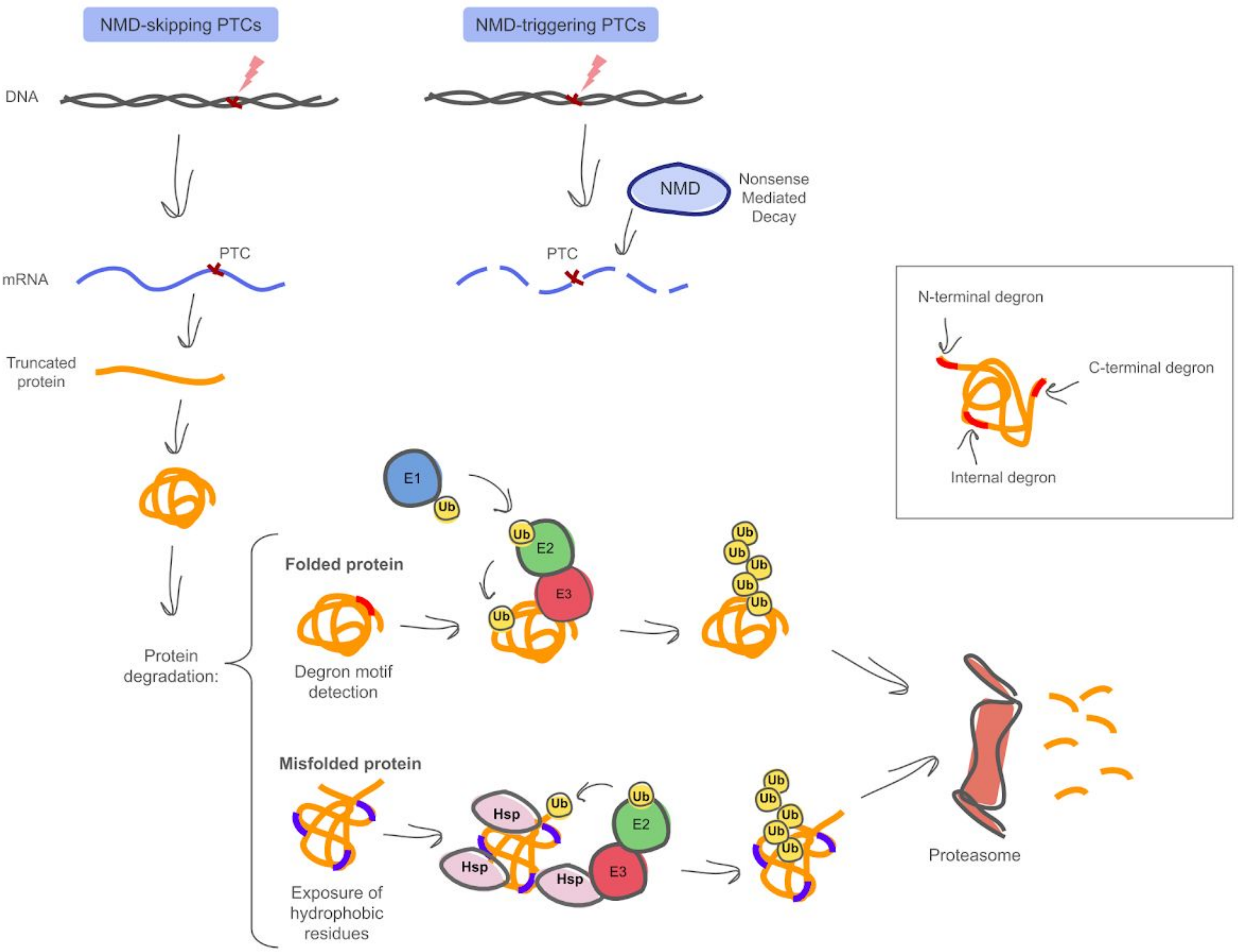


Figure S1. Cellular pathways for the degradation of PTC-containing mRNA and truncated proteins. Premature Termination Codons (PTCs) are the consequence of nonsense and frameshift (indel) mutations. Nonsense Mediated Decay (NMD) targets and degrades PTC-bearing mRNAs (NMD-triggering PTCs), however some PTCs can escape from NMD degradation (NMD-skipping), thus potentially generating a truncated protein. Protein degradation can be performed either by detecting degron sequences (normal protein turnover) or by detecting misfolding features such as the exposure of hydrophobic amino acids (clearance of misfolded proteins). Proteins targeted for degradation are then ubiquitinated (Ub) by E1 and E2 enzymes, which bind to specialized E3 ligases. Ubiquitinated proteins are sent to the proteasome for degradation.

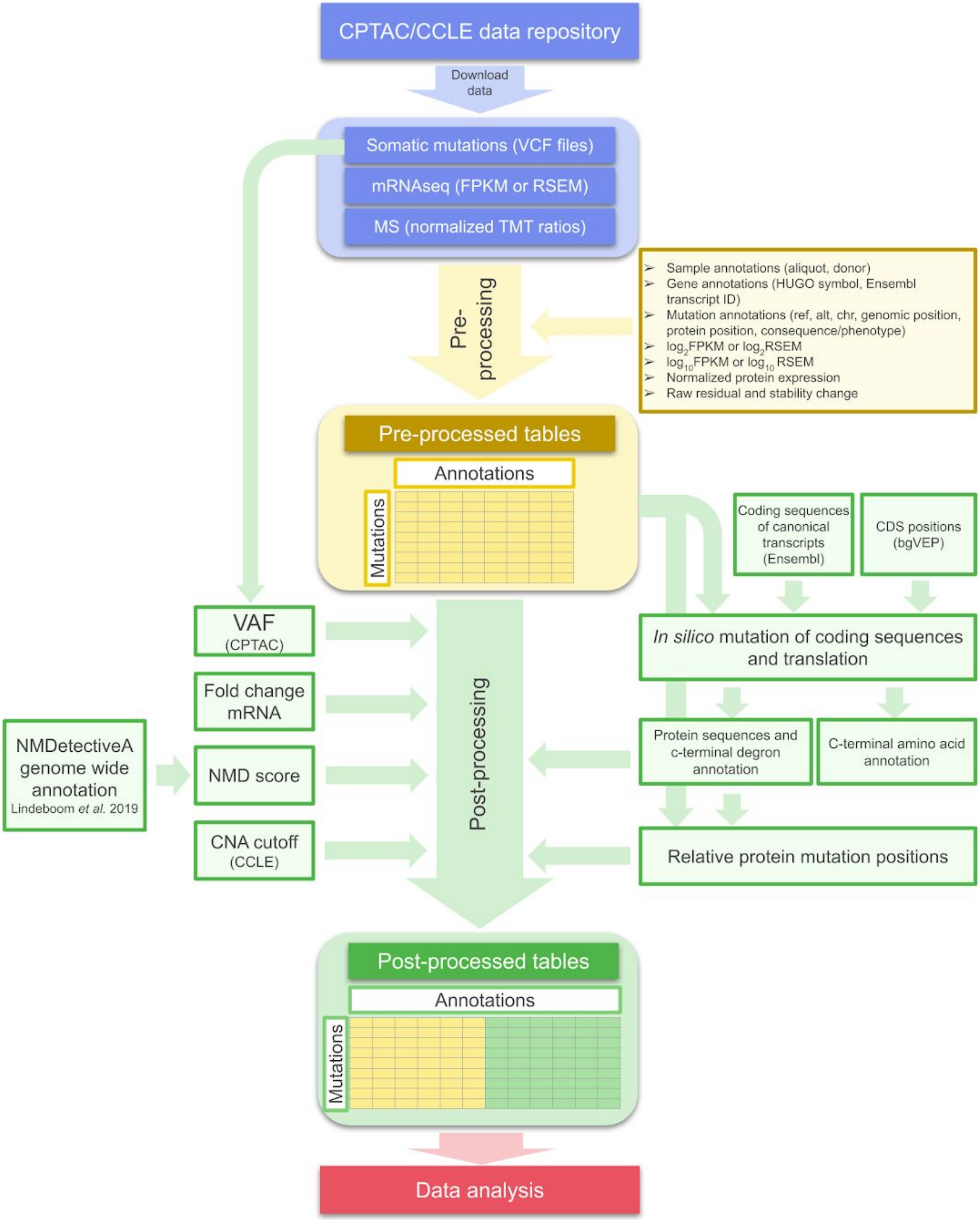

CPTAC/CCLE data repository
Download data
Somatic mutations (VCF files)
mRNAseq (FPKM or RSEM)
MS (normalized TMT ratios)
Pre-processing
➢ Sample annotations (aliquot, donor)
➢ Gene annotations (HUGO symbol, Ensembl transcript ID)
➢ Mutation annotations (ref, alt, chr, genomic position, protein position, consequence/phenotype)
➢ $\log_2$FPKM or $\log_2$RSEM
➢ $\log_{10}$FPKM or $\log_{10}$ RSEM
➢ Normalized protein expression
➢ Raw residual and stability change
Pre-processed tables
Annotations
Mutations
Coding sequences of canonical transcripts (Ensembl)
CDS positions (bgVEP)
In silico mutation of coding sequences and translation
VAF (CPTAC)
Fold change mRNA
NMDetectiveA genome wide annotation Lindeboom et al. 2019
NMD score
CNA cutoff (CCLE)
Post-processing
Protein sequences and c-terminal degron annotation
C-terminal amino acid annotation
Relative protein mutation positions
Post-processed tables
Annotations
Mutations
Data analysis

Figure S2. Workflow diagram of the data analysis. Three files from CPTAC and CCLE datasets were downloaded from their respective data repositories: somatic mutations information (VCF files), mRNAseq (FPKM or RSEM counts) and Mass-Spectrometry (MS, normalized TMT expression) (shown in blue). These files were pre-processed to annotate the necessary parameters for the downstream analysis in a unique table per dataset (Pre-processed tables) (shown in yellow). These tables were then post-processed through several steps: 1) addition of VAF information from VCF files (only for CPTAC dataset), 2) calculation of mRNA fold change, 3) addition of NMD-score (extracted from NMDetectiveA tables from Lindeboom *et al.* 2019), 4) exclusion of altered CNA (only for CCLE dataset), 5) addition of protein sequences and c-terminal degron/amino acids annotations and 6) addition of the relative protein mutation positions (Post-processed tables) (shown in green). These tables with all the integrated information were then used for the data analysis and figure plots (in red). Important note: M. Sánchez Guixé performed the post-processing (green) and data analysis (red) parts of the workflow. For more information about the code used in these steps please visit the github repository (github/msguixe/TFM_MSG).

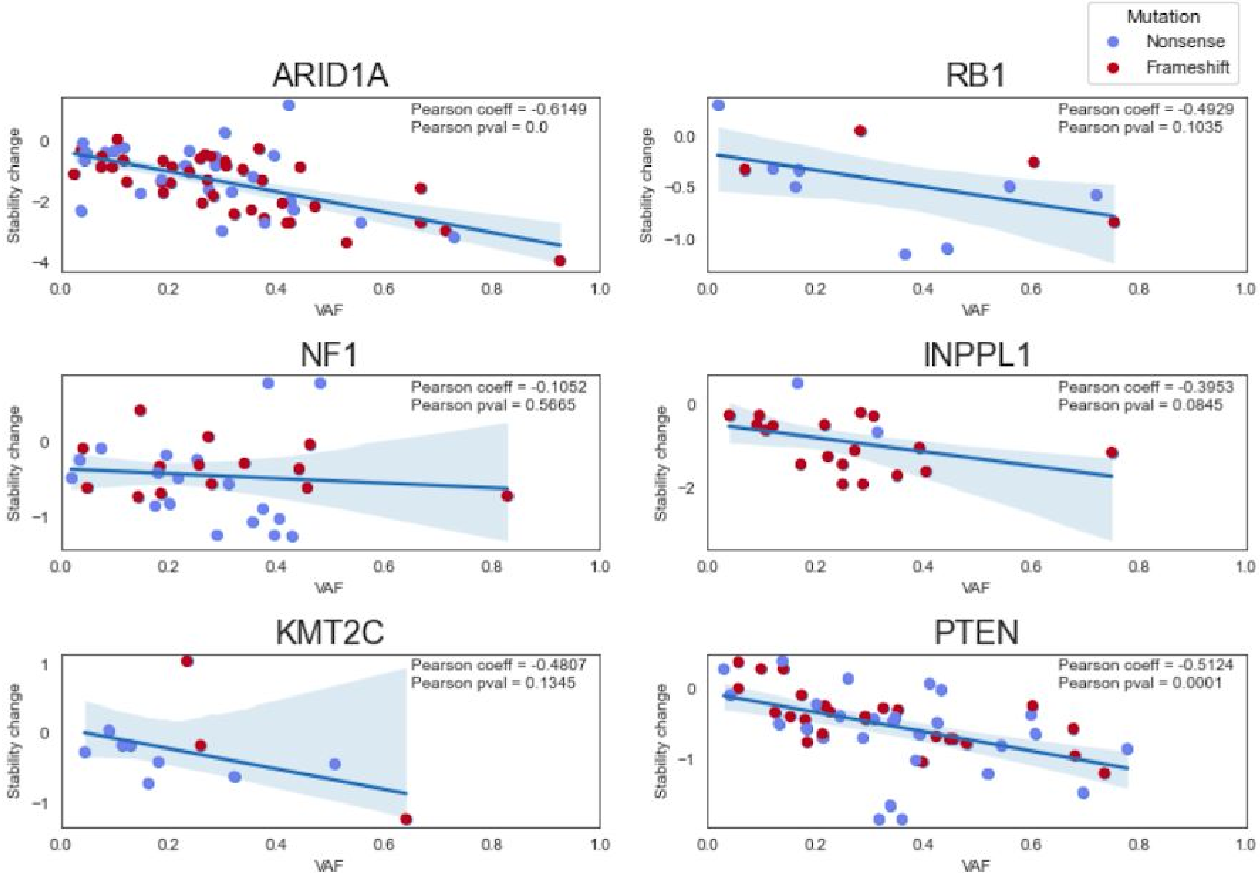
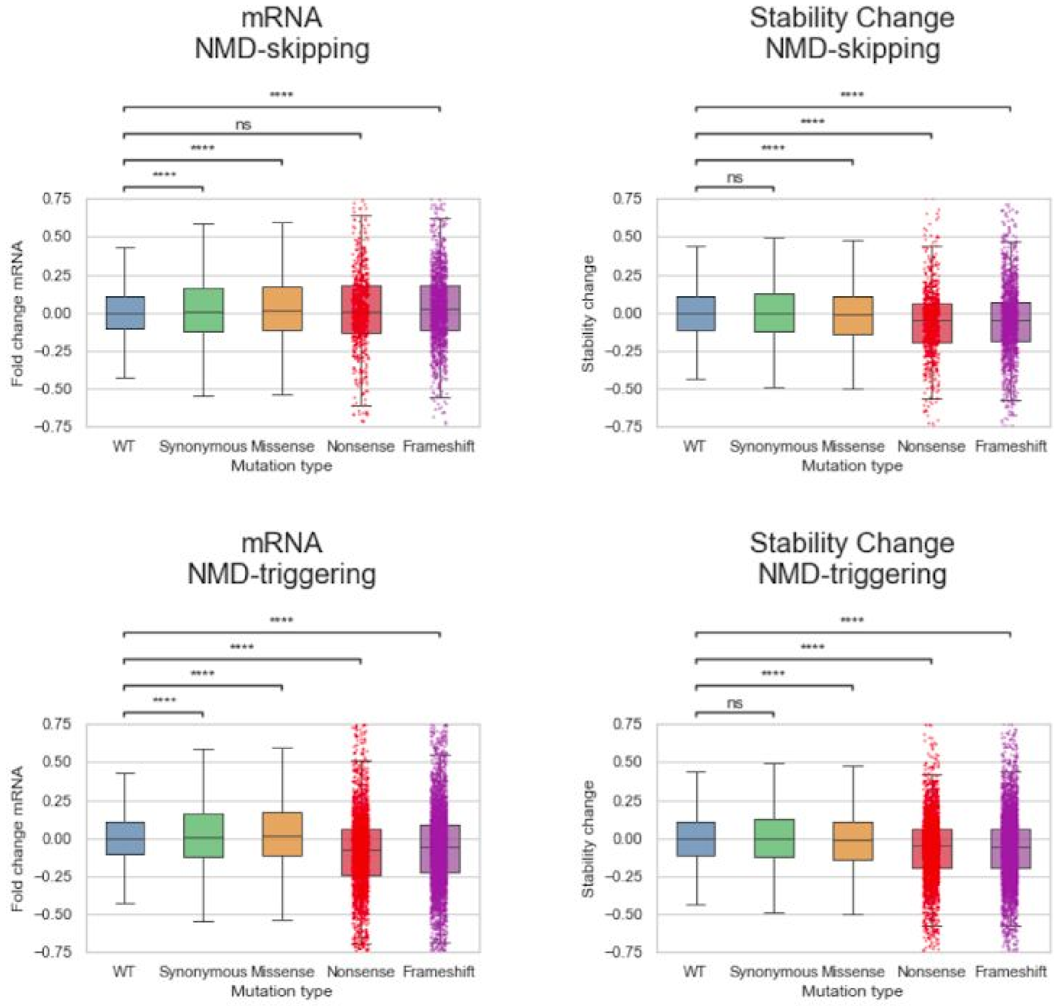


Figure S3. VAF by stability change. Stability change across different VAF values for *ARID1A*, *RB1*, *NF1*, *INPPL1*, *KMT2C* and *PTEN* genes. Blue points indicate nonsense mutations and red points indicate frameshift mutations. Statistical analysis: Pearson correlation.



mRNA NMD-skipping

Stability Change NMD-skipping

mRNA NMD-triggering

Stability Change NMD-triggering

Figure S4. NMD-triggering PTCs show decreased protein stability in CCLE. mRNA fold-change (left) and stability change (right) comparing WT, synonymous, missense and NMD-skipping nonsense and frameshift (above) or NMD-triggering nonsense and frameshift (below) alterations. Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction.

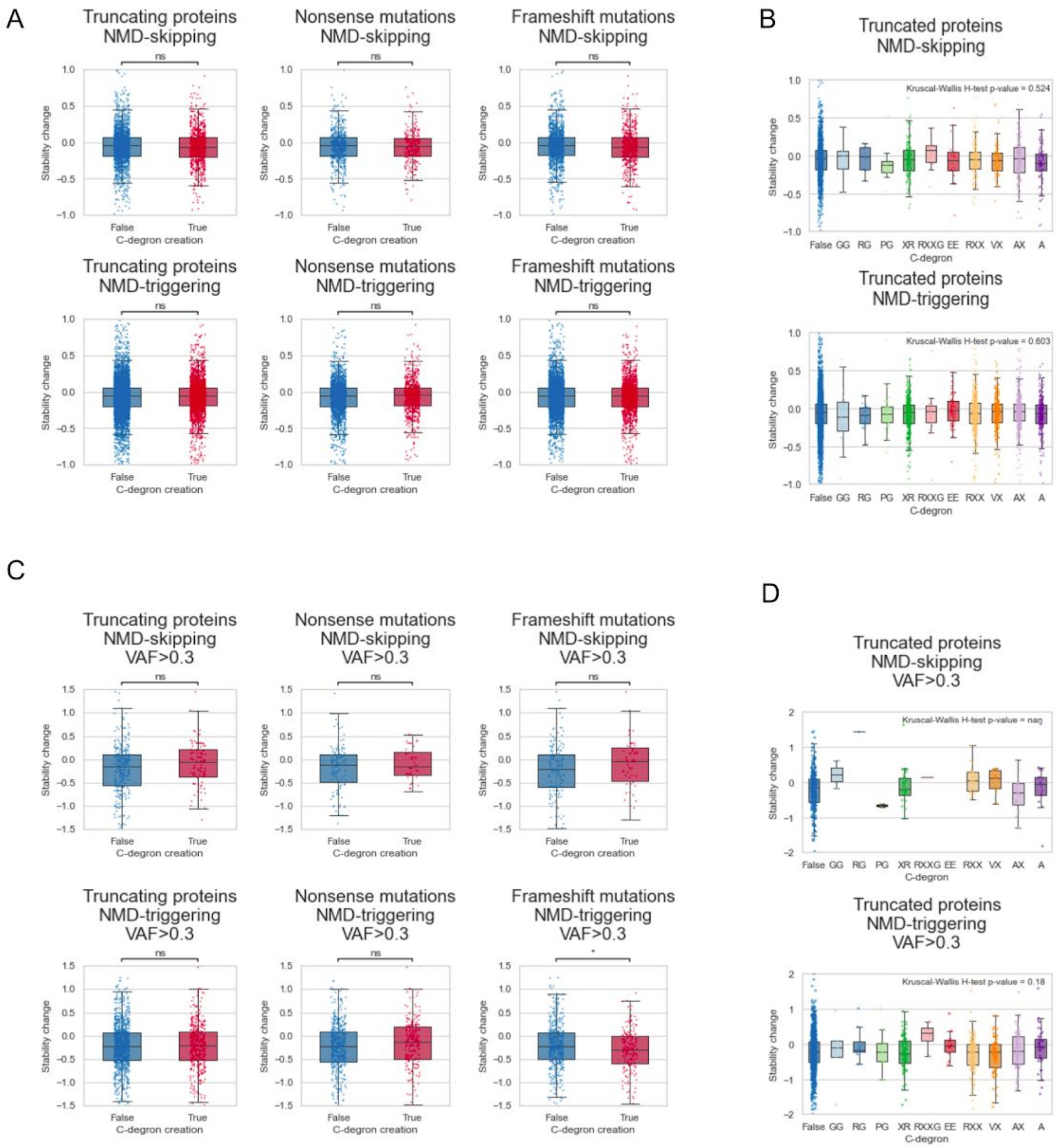


Figure S5. C-degron analysis in CCLE and CPTAC with high VAF. Stability change comparison in CCLE dataset (A) and CPTAC dataset with high VAF (C) between mutated proteins with (True) or without (False) c-degrons in all truncating proteins (left), or nonsense (middle) and frameshift (right) separately, in NMD-skipping (above) or triggering (below) subsettings. Stability change comparison in CCLE (B) and CPTAC with high VAF (D) between truncated proteins without c-degrons (False) and with each of the c-degron instances. Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction for c-degron creation plots and Kruskal-Wallis H-test for differences between c-degrons.

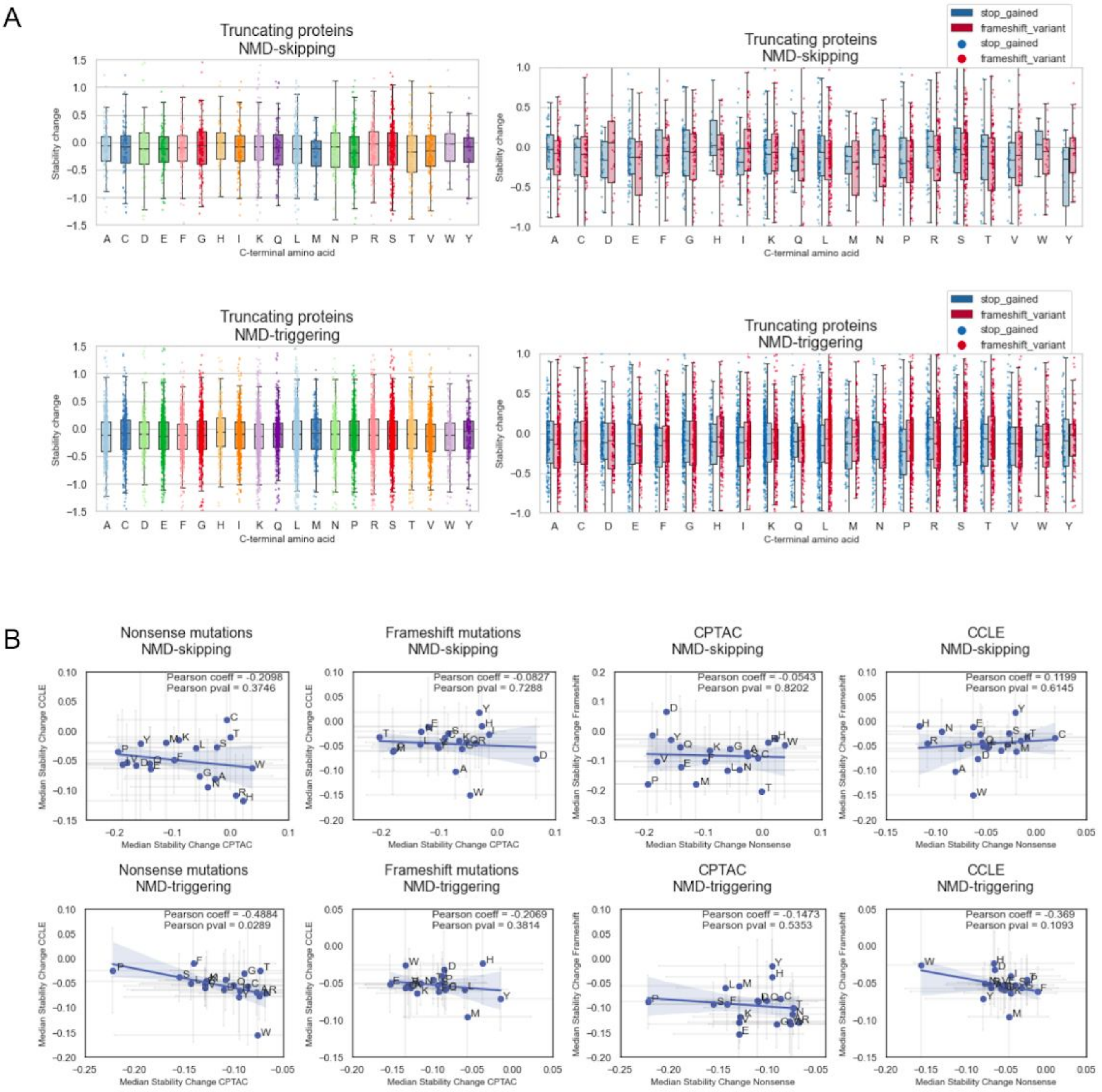


Figure S6: Stability change of truncated proteins by the c-terminal amino acid. A) Stability change comparison between truncated proteins grouped by the c-terminal amino acid for all truncating alterations (left) and for nonsense and frameshift separately (right), in NMD-skipping (above) and NMD-triggering (below) subsettings. B) Median stability change correlation for each c-terminal amino acid between in CPTAC (x-axis) and CCLE (y-axis) datasets of nonsense mutations (left) and frameshift (middle-left) alterations; median stability change correlation for each c-degron between nonsense (x-axis) and frameshift (y-axis) alterations in CPTAC (middle right) and CCLE (right) datasets; for NMD-skipping (above) and triggering (below) subsettings. Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction. Pearson correlations were performed for regression plots. Error bars from regression plots indicate CI95.

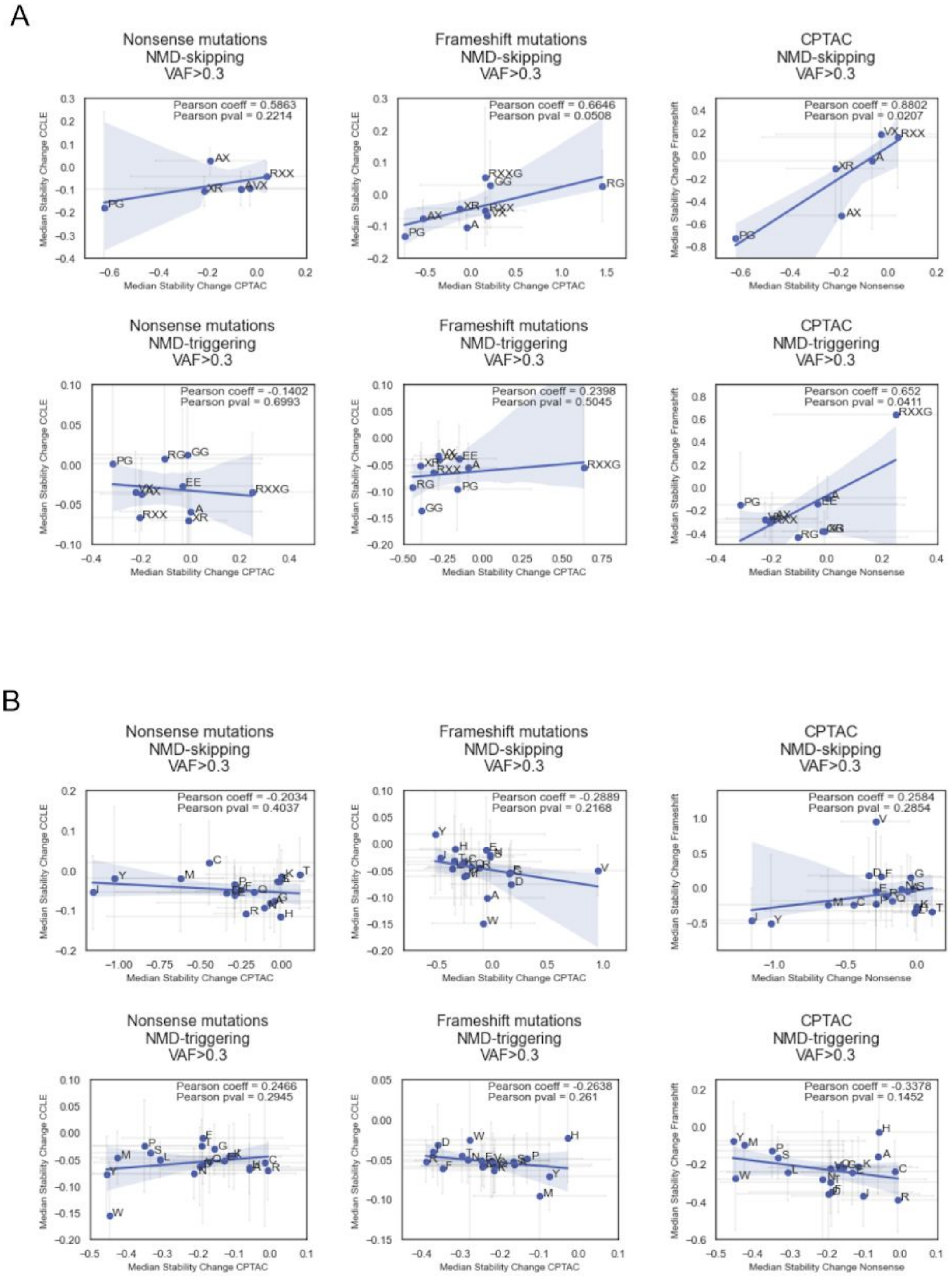


Figure S7. C-degron and c-terminal amino acid stability change correlation on mutations with high VAF in CPTAC dataset. Median stability change correlation for each c-degron (A) and c-terminal amino acid (B) between in CPTAC with high VAF (x-axis) and CCLE (y-axis) datasets of nonsense mutations (left) and frameshift (middle) alterations; median stability change correlation for each c-degron between nonsense (x-axis) and frameshift (y-axis) alterations in CPTAC with high VAF (right); for NMD-skipping (above) and triggering (below) subsettings. Statistical test: Pearson correlation. Error bars indicate CI95.

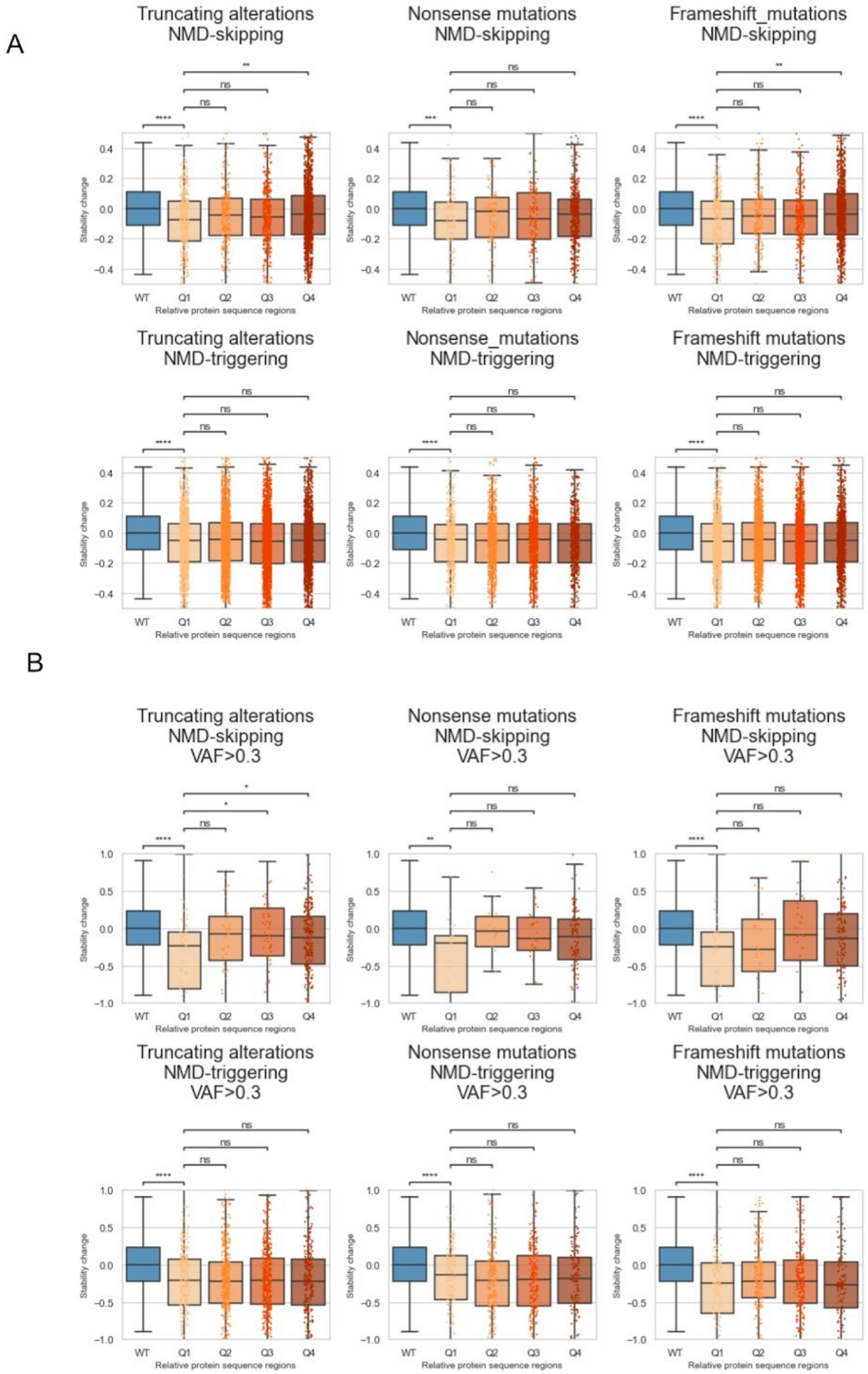

A
Truncating alterations NMD-skipping
**** ns ns **
Nonsense mutations NMD-skipping
*** ns ns ns
Frameshift_mutations NMD-skipping
**** ns ns **
Truncating alterations NMD-triggering
**** ns ns ns
Nonsense_mutations NMD-triggering
**** ns ns ns
Frameshift mutations NMD-triggering
**** ns ns ns
Stability change
0.4
0.2
0.0
-0.2
-0.4
WT Q1 Q2 Q3 Q4
Relative protein sequence regions
B
Truncating alterations NMD-skipping VAF>0.3
**** ns * *
Nonsense mutations NMD-skipping VAF>0.3
** ns ns ns
Frameshift mutations NMD-skipping VAF>0.3
**** ns ns ns
Truncating alterations NMD-triggering VAF>0.3
**** ns ns ns
Nonsense mutations NMD-triggering VAF>0.3
**** ns ns ns
Frameshift mutations NMD-triggering VAF>0.3
**** ns ns ns
Stability change
1.0
0.5
0.0
-0.5
-1.0
WT Q1 Q2 Q3 Q4
Relative protein sequence regions

Figure S8. Relative mutation position in CCLE and CPTAC with high VAF. Stability change comparison between WT and truncated proteins from the 4 defined quarters in all truncating alterations (left), and in nonsense (middle) and frameshift (right) separately, in NMD-skipping (above) and NMD-triggering (below) in CCLE (A) and CPTAC with high VAF (B). Statistical test: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction.

Abbreviations

ARID1A	AT-Rich Interaction Domain 1A
AUC	Area Under Curve
BRCA	Breast Cancer
CCLE	The Cancer Cell Line Encyclopedia
CCRCC	Clear Cell Renal Cell Carcinoma
C-degron	C-terminal Degron
CDS	Coding Sequence
CNA	Copy Number Alteration
COAD	Colon Adenocarcinoma
CPTAC	The Clinical Proteomics Tumor Analysis Consortium
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GBM	Brain Cancer / Glioblastoma
LUAD	Lung Adenocarcinoma
MS	Mass Spectrometry
NMD	Nonsense Mediated Decay
OV	Ovarian Cancer
PTC	Premature Termination Codon
RNAseq	RNA sequencing
ROC	Receiver Operating Characteristic
RSEM	RNA-Seq expression levels with EM algorithm
TMT	Tandem Mass Tag
Ub	Ubiquitin
UCEC	Uterine Corpus Endometrial Carcinoma
UPF1	UPF1 RNA Helicase And ATPase
UPS	Ubiquitin Proteasome System
VAF	Variant Allele Frequency
VCF	Variant Call Format
WES	Whole Exome Sequencing
WT	Wild-type

Table of contents