Supplemental Information

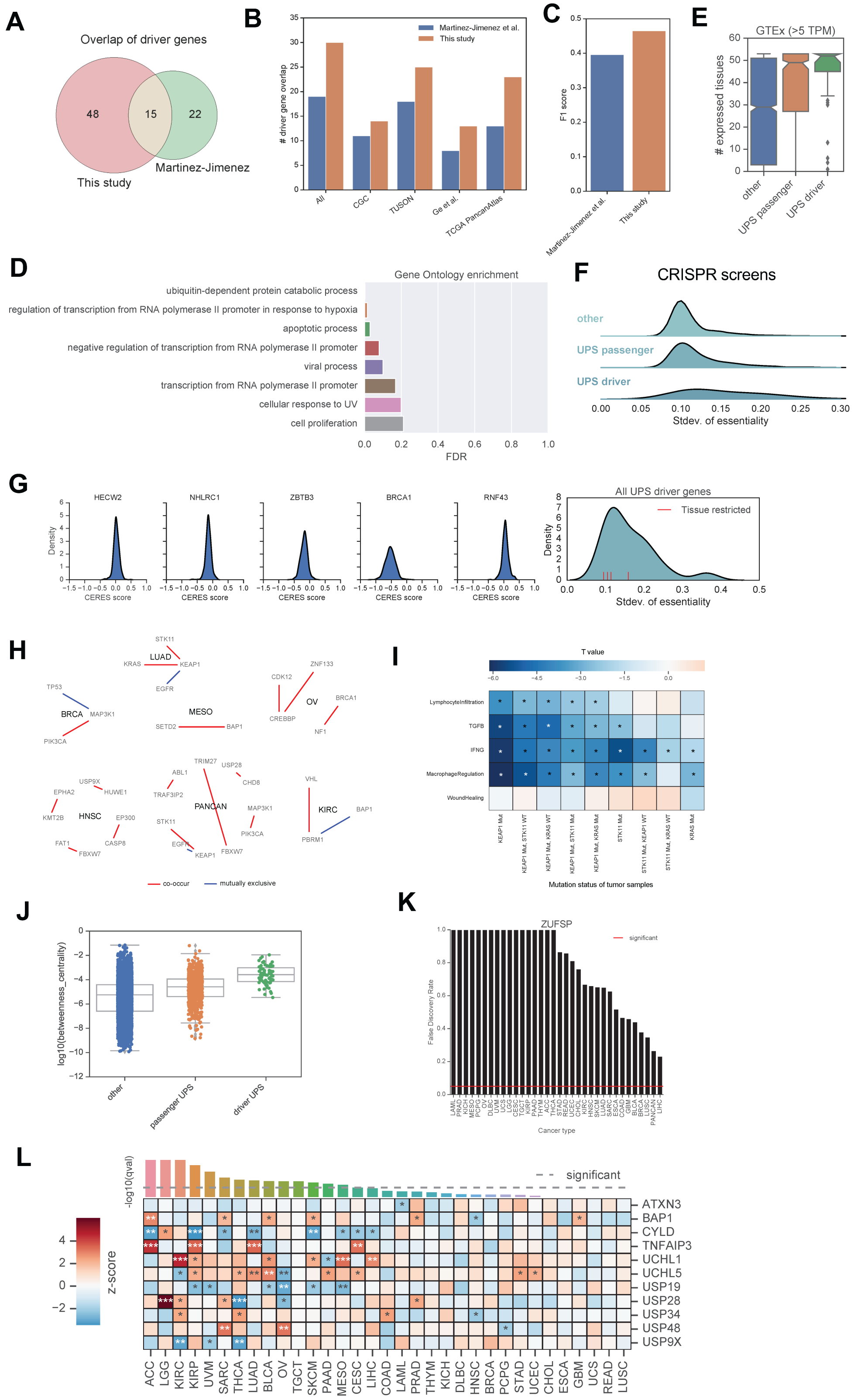

Figure S1. Putative UPS drivers are contextually related to other pathways. Related to Figure 2.

(A) Venn diagram displaying the overlap of significantly mutated genes in Martinez-Jimenez et al. with this study.

(B) Number of genes overlapping with prior studies statistically implicating driver genes.

(C) Our current study has higher performance (F1 score) than Martinez-Jimenez et al., as benchmarked against agreement with all prior studies. The F1 score is a metric that balances the contribution of sensitivity and precision.

(D) Bar graph displaying the Gene Set Enrichment Analysis (GSEA) of the 63 UPS driver genes.

(E) Putative driver genes implicated in the UPS pathway are expressed at the RNA-level [Transcript Per Million (TPM) > 5] in a greater number of normal tissues in GTEx than compared to “passenger genes” in the UPS pathway (p=0.005, Mann-Whitney U test) or all other genes (p=2.2e-10, Mann-Whitney U test). The boxplots represent the distribution of the number of expressed tissues across multiple genes. Boxplot whiskers represent the quartile +/- 1.5 times the interquartile range.

(F) Conversely, putative driver genes implicated in the UPS pathway have more variability in cell line essentiality (CERES scores) according to CRISPR screens from ~500 cell lines in DepMap (UPS passenger: p=3e-08; non-UPS genes: p=1.2e-13; Mann-Whitney U test), suggesting cell type specificity.

(G) We examined all UPS driver genes with tissue restricted expression, as defined by being expressed (>5 TPM) in less than half of tissues from the Gene-Tissue Expression (GTEx) consortium. Only five genes met this definition of restricted tissue expression (HECW2, NHLRC1, ZBTB3, BRCA1, and RNF43). Left, distribution of gene essentiality scores from DepMap (https://depmap.org/portal/). CERES scores are negative when a gene is essential for cell viability and positive when gene knockout provides a selective advantage. Right, variability of essentiality scores from DepMAP across cell lines for all driver genes identified in this study. Red ticks mark genes with tissue restricted expression.

(H) Co-mutational analysis of UPS driver genes with previously identified driver genes from the TCGA PancanAtlas consortium. A red line indicates significant co-occurrence, while a blue line indicates significant mutual exclusivity (Mantel-Haenzel test, q<0.25, adjusted for mutation burden).

(I) Heatmap displaying the association (t statistic) of mutations in driver genes KEAP1, STK11 and/or KRAS with 5 immune-related biomarkers (lymphocyte infiltration, TGFB, IFNG response, macrophage regulation and wound healing) for lung adenocarcinoma. Asterisk indicates association is significant (q<0.1). See the “Correlation with immune-related gene expression signatures” section in STAR methods for details. Columns with a gene symbol marked as “Mut” indicates tumors had a mutation in that gene, while those marked as “WT” (wildtype) indicate the absence of a mutation. KEAP1 showed an association with immune-related biomarkers that was independent of STK11 co-mutation.

(J) Boxplot of the betweenness centrality metric in a protein-protein interaction network (BioGrid) of genes that were either found to be UPS driver genes, UPS passenger genes, or all other genes. In this case, betweenness centrality measures how often a protein is situated on shortest paths within a PPI network (see STAR methods for further details). UPS driver genes had a significantly higher betweenness centrality than compared to UPS passenger genes (p=2e-11, Mann-Whitney U test) and all other genes (p=1e-26). Boxplot shows quartiles with whiskers representing the quartile +/- 1.5 times the interquartile range.

(K) Plot of the false discovery rate from the 20/20+ method of the recently identified deubiquitinating enzyme *ZUFSP* across all available cancer types and pan-cancer analysis (“PANCAN”, aggregate of all cancer types together). *ZUFSP* did not meet the stringent requirements for statistical significance (q<0.05) in any of the analyses.

(L) Heatmap displaying the association between deubiquitinating enzyme (DUB) gene expression and overall patient survival (Cox PH regression) for various cancer types in the TCGA (x-axis). The significance for the combined model of 11 DUB gene expression is shown in the top bar plot (q<0.05). *=p<0.05, **p<0.01, ***p<0.001.

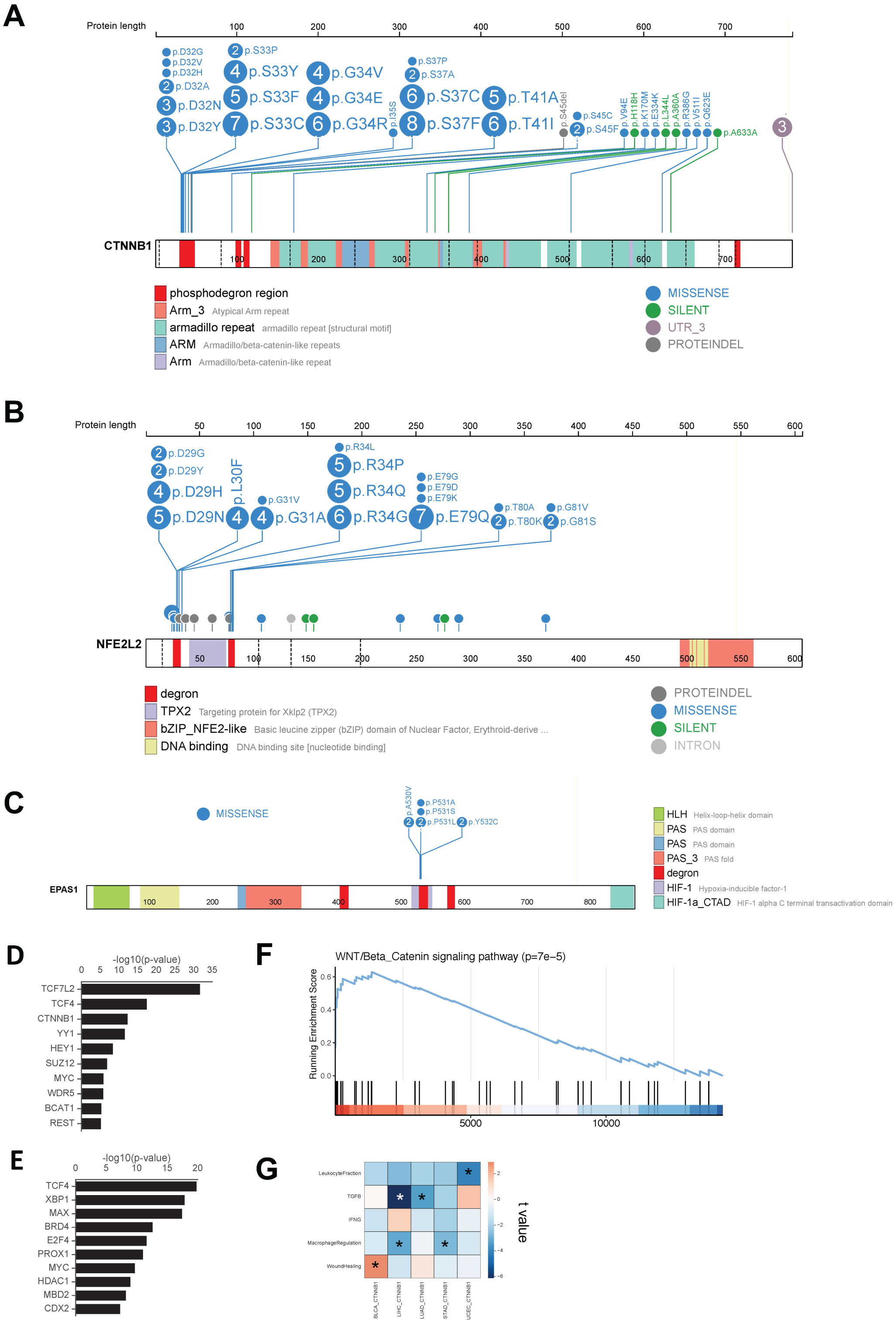
A
Protein length
100
200
300
400
500
600
700
p.D32G
p.D32V
p.D32H
2 p.D32A
3 p.D32N
3 p.D32Y
2 p.S33P
4 p.S33Y
5 p.S33F
7 p.S33C
4 p.G34V
4 p.G34E
6 p.G34R
p.I35S
p.S37P
2 p.S37A
6 p.S37C
8 p.S37F
5 p.T41A
6 p.T41I
p.S45del
p.S45C
2 p.S45F
p.V94E
p.H118H
p.K170M
p.E334K
p.L344L
p.A360A
p.R386G
p.V511I
p.Q623E
p.A633A
3
CTNNB1
phosphodegron region
Arm_3 Atypical Arm repeat
armadillo repeat armadillo repeat [structural motif]
ARM Armadillo/beta-catenin-like repeats
Arm Armadillo/beta-catenin-like repeat
MISSENSE
SILENT
UTR_3
PROTEINDEL
B
Protein length
50
100
150
200
250
300
350
400
450
500
550
600
2 p.D29G
2 p.D29Y
4 p.D29H
5 p.D29N
4 p.L30F
p.G31V
4 p.G31A
p.R34L
5 p.R34P
5 p.R34Q
6 p.R34G
p.E79G
p.E79D
p.E79K
7 p.E79Q
p.T80A
2 p.T80K
p.G81V
2 p.G81S
NFE2L2
degron
TPX2 Targeting protein for Xklp2 (TPX2)
bZIP_NFE2-like Basic leucine zipper (bZIP) domain of Nuclear Factor, Erythroid-derive ...
DNA binding DNA binding site [nucleotide binding]
PROTEINDEL
MISSENSE
SILENT
INTRON
C
MISSENSE
p.A530V
p.P531A
p.P531S
2 p.P531L
2 p.Y532C
EPAS1
100
200
300
400
500
600
700
800
HLH Helix-loop-helix domain
PAS PAS domain
PAS PAS domain
PAS_3 PAS fold
degron
HIF-1 Hypoxia-inducible factor-1
HIF-1a_CTAD HIF-1 alpha C terminal transactivation domain
D
-log10(p-value)
0 5 10 15 20 25 30 35
TCF7L2
TCF4
CTNNB1
YY1
HEY1
SUZ12
MYC
WDR5
BCAT1
REST
E
-log10(p-value)
0 5 10 15 20
TCF4
XBP1
MAX
BRD4
E2F4
PROX1
MYC
HDAC1
MBD2
CDX2
F
WNT/Beta_Catenin signaling pathway (p=7e−5)
Running Enrichment Score
0.6
0.4
0.2
0.0
5000
10000
G
LeukocyteFraction
TGFB
IFNG
MacrophageRegulation
WoundHealing
BLCA_CTNNB1
LIHC_CTNNB1
LUAD_CTNNB1
STAD_CTNNB1
UCEC_CTNNB1
*
2
0
−2
−4
−6
t value

Figure S2. Detailed examples of degrons enriched for mutations. Related to Figure 3. Lollipop diagram of TCGA mutations in: *CTNNB1* for uterine and endometrial carcinoma **(A)**, *NFE2L2* for lung squamous cell carcinoma **(B)**, and *EPAS1* for Pheochromocytoma and Paraganglioma **(C)**. The color of circles distinguishes the type of mutation, while colored rectangles are annotations of the protein. Numbered text indicates the position along the protein sequence. **(D)** Inferred up-regulated transcription factors (RABIT analysis) based on *CTNNB1* mutant versus wildtype differential expression. **(E)** Inferred up-regulated transcription factors (RABIT analysis) based on *APC* mutant versus wildtype differential expression. **(F)** Gene set enrichment analysis of the differential gene expression profile of CTNNB1 degron mutant versus wildtype tumors shows strong enrichment for the WNT signaling pathway hallmark from MSigDB. **(G)** Heatmap displaying the association (t-statistic) of degron mutations with inferred immune-related biomarkers based on gene expression using a Wald test (Methods). *=q<0.1.

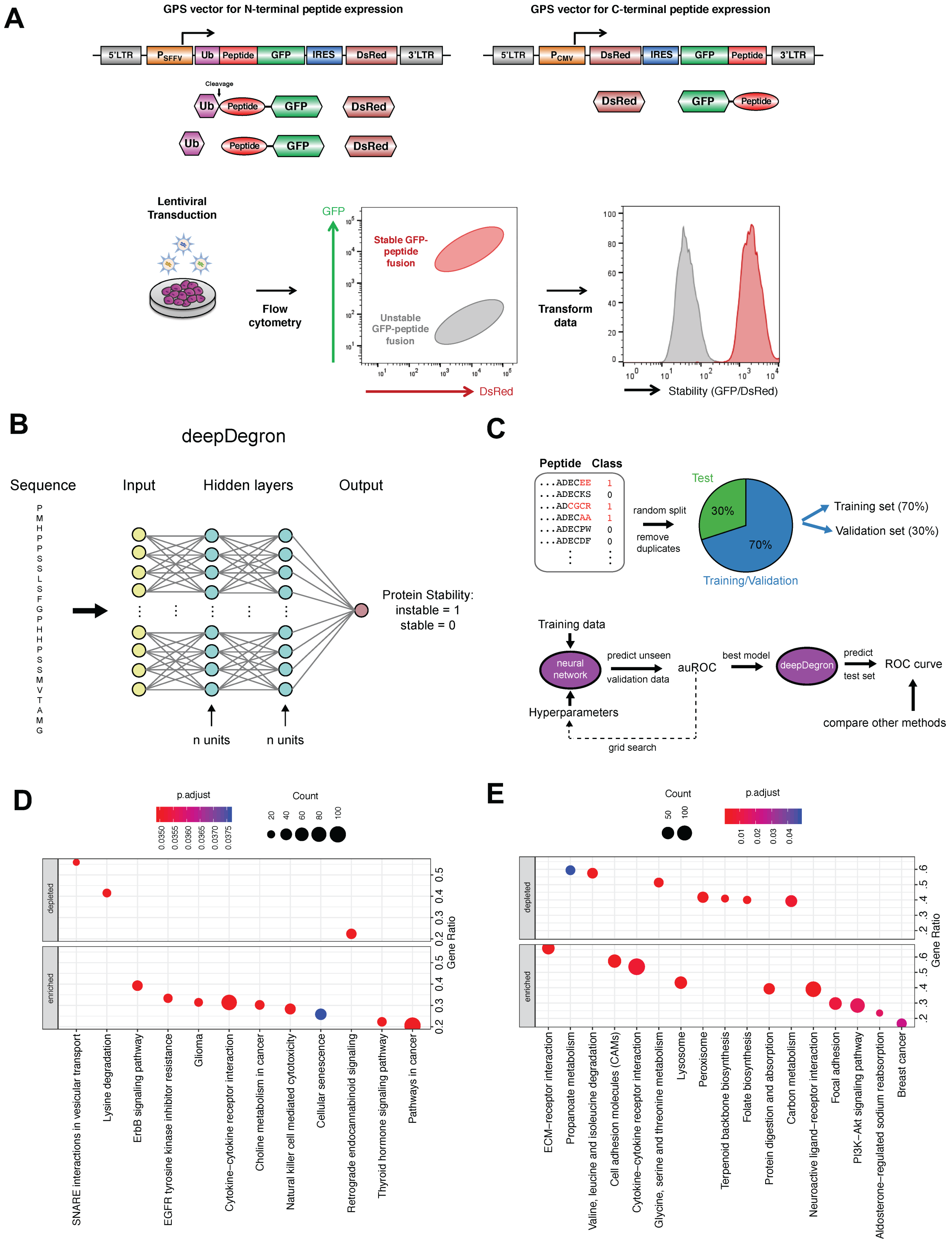
A
GPS vector for N-terminal peptide expression
5'LTR
P_{SFFV}
Ub
Peptide
GFP
IRES
DsRed
3'LTR
Cleavage
Ub
Peptide
GFP
DsRed
Ub
Peptide
GFP
DsRed
GPS vector for C-terminal peptide expression
5'LTR
P_{CMV}
DsRed
IRES
GFP
Peptide
3'LTR
DsRed
GFP
Peptide
Lentiviral Transduction
Flow cytometry
GFP
Stable GFP-peptide fusion
Unstable GFP-peptide fusion
DsRed
Transform data
Stability (GFP/DsRed)
B
deepDegron
Sequence
Input
Hidden layers
Output
P M H P P S S L S F G P H H P S S M V T A M G
Protein Stability:
instable = 1
stable = 0
n units
n units
C
Peptide Class
...ADECEE 1
...ADECKS 0
...ADCGCR 1
...ADECAA 1
...ADECPW 0
...ADECDF 0
random split
remove duplicates
Test
30%
70%
Training/Validation
Training set (70%)
Validation set (30%)
Training data
neural network
Hyperparameters
predict unseen
validation data
auROC
grid search
best model
deepDegron
predict
test set
ROC curve
compare other methods
D
p.adjust
0.0350 0.0355 0.0360 0.0365 0.0370 0.0375
Count
20 40 60 80 100
depleted
enriched
Gene Ratio
0.2 0.3 0.4 0.5
SNARE interactions in vesicular transport
Lysine degradation
ErbB signaling pathway
EGFR tyrosine kinase inhibitor resistance
Glioma
Cytokine−cytokine receptor interaction
Choline metabolism in cancer
Natural killer cell mediated cytotoxicity
Cellular senescence
Retrograde endocannabinoid signaling
Thyroid hormone signaling pathway
Pathways in cancer
E
Count
50 100
p.adjust
0.01 0.02 0.03 0.04
depleted
enriched
Gene Ratio
.2 .3 .4 .5 .6
ECM−receptor interaction
Propanoate metabolism
Valine, leucine and isoleucine degradation
Cell adhesion molecules (CAMs)
Cytokine−cytokine receptor interaction
Glycine, serine and threonine metabolism
Lysosome
Peroxisome
Terpenoid backbone biosynthesis
Folate biosynthesis
Protein digestion and absorption
Carbon metabolism
Neuroactive ligand−receptor interaction
Focal adhesion
PI3K−Akt signaling pathway
Aldosterone−regulated sodium reabsorption
Breast cancer

Figure S3. Overview of deepDegron. Related to Figure 4.

(A) Diagram of the Global Protein Stability (GPS) assay to measure protein stability.

(B) Neural network architecture of deepDegron to predict the instability conferred by peptides in the Global Protein Stability (GPS) assay.

(C) Performance evaluation procedure for deepDegron. Area under Receiver Operating Characteristic curve = auROC.

(D-E) Gene set enrichment analysis of the degron potential score for C-terminal and N-terminal peptides, respectively, from the human proteome.

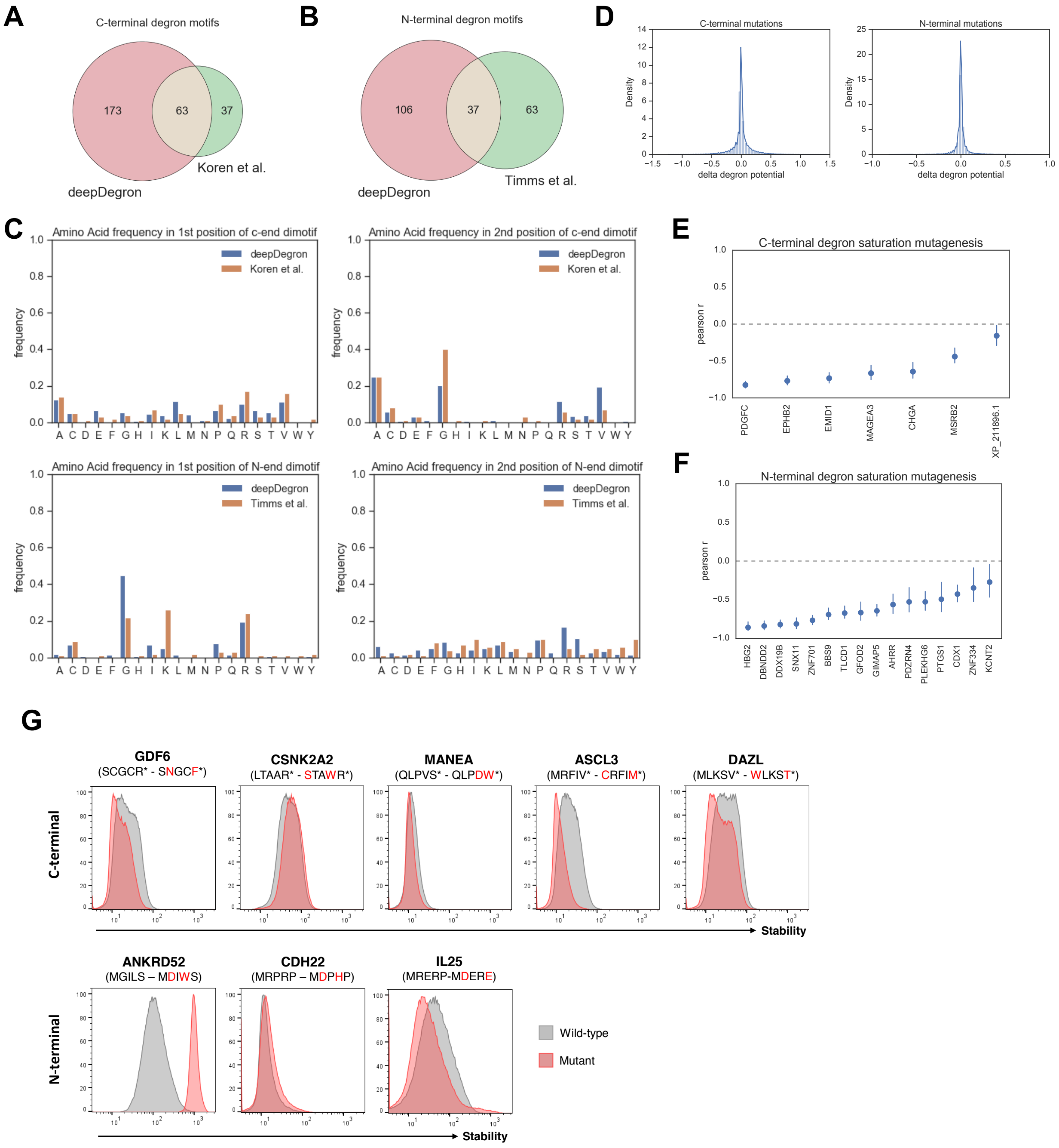
A
C-terminal degron motifs
173
63
37
deepDegron
Koren et al.
B
N-terminal degron motifs
106
37
63
deepDegron
Timms et al.
D
C-terminal mutations
Density
delta degron potential
N-terminal mutations
Density
delta degron potential
C
Amino Acid frequency in 1st position of c-end dimotif
Amino Acid frequency in 2nd position of c-end dimotif
deepDegron
Koren et al.
Amino Acid frequency in 1st position of N-end dimotif
Amino Acid frequency in 2nd position of N-end dimotif
deepDegron
Timms et al.
frequency
A C D E F G H I K L M N P Q R S T V W Y
E
C-terminal degron saturation mutagenesis
pearson r
PDGFC
EPHB2
EMID1
MAGEA3
CHGA
MSRB2
XP_211896.1
F
N-terminal degron saturation mutagenesis
pearson r
HBG2
DBNDD2
DDX19B
SNX11
ZNF701
BBS9
TLCD1
GFOD2
GIMAP5
AHRR
PDZRN4
PLEKHG6
PTGS1
CDX1
ZNF334
KCNT2
G
GDF6
(SCGCR* - SNGCF*)
CSNK2A2
(LTAAR* - STAWR*)
MANEA
(QLPVS* - QLPDW*)
ASCL3
(MRFIV* - CRFIM*)
DAZL
(MLKSV* - WLKST*)
C-terminal
Stability
ANKRD52
(MGILS – MDIWS)
CDH22
(MRPRP – MDPHP)
IL25
(MRERP-MDERE)
N-terminal
Stability
Wild-type
Mutant

Figure S4. Comparison of degron motifs found by deepDegron to those previously reported. Related to Figure 4.

Venn diagram comparing **(A)** C-terminal and **(B)** N-terminal degron motifs identified by deepDegron to those found by Koren et al. and Timms et al., respectively. **(C)** Comparison of overall amino acid usage in identified motifs. **(D)** Distribution of delta degron potential for mutations observed in our TCGA mutation dataset. **(E)** Pearson correlation coefficients for all saturation mutagenesis experiments of C-terminal peptides in Koren et al. Blue lines indicate 95% bootstrapped confidence intervals (1,000 iterations). **(F)** Pearson correlations for all saturation mutagenesis experiments of N-terminal peptides in Timms et al. **(G)** Remaining GPS stability measurements of C-terminal (top) or N-terminal (bottom) peptides derived from the indicated genes, comparing wild-type (gray histograms) and double mutant (red histograms) sequences. Mutated residues are indicated in red font. X-axis is proportional to GFP / DsRed signal as measured by flow cytometry (see STAR methods); Y-axis is normalized cell count.

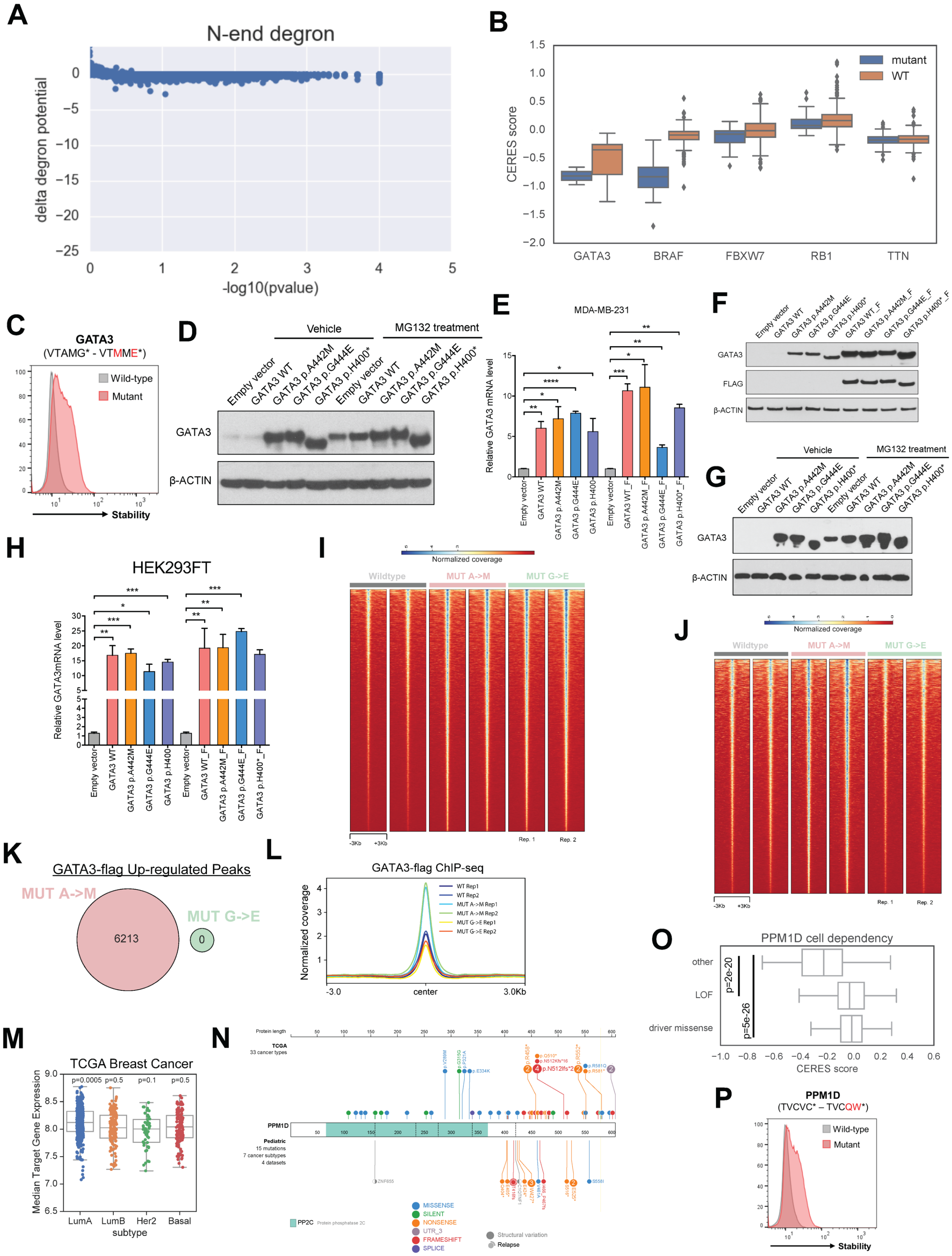
A
N-end degron
delta degron potential
-log10(pvalue)
B
CERES score
mutant
WT
GATA3
BRAF
FBXW7
RB1
TTN
C
GATA3
(VTAMG* - VTMME*)
Wild-type
Mutant
Stability
D
Vehicle
MG132 treatment
Empty vector
GATA3 WT
GATA3 p.A442M
GATA3 p.G444E
GATA3 p.H400*
GATA3
β-ACTIN
E
MDA-MB-231
Relative GATA3 mRNA level
Empty vector
GATA3 WT
GATA3 p.A442M
GATA3 p.G444E
GATA3 p.H400*
GATA3 WT_F
GATA3 p.A442M_F
GATA3 p.G444E_F
GATA3 p.H400*_F
F
GATA3
FLAG
β-ACTIN
G
Vehicle
MG132 treatment
GATA3
β-ACTIN
H
HEK293FT
Relative GATA3mRNA level
I
Normalized coverage
Wildtype
MUT A->M
MUT G->E
-3Kb
+3Kb
Rep. 1
Rep. 2
J
Normalized coverage
Wildtype
MUT A->M
MUT G->E
Rep. 1
Rep. 2
K
GATA3-flag Up-regulated Peaks
MUT A->M
MUT G->E
6213
0
L
GATA3-flag ChIP-seq
Normalized coverage
WT Rep1
WT Rep2
MUT A->M Rep1
MUT A->M Rep2
MUT G->E Rep1
MUT G->E Rep2
-3.0
center
3.0Kb
M
TCGA Breast Cancer
Median Target Gene Expression
p=0.0005
p=0.5
p=0.1
p=0.5
LumA
LumB
Her2
Basal
subtype
N
Protein length
TCGA
33 cancer types
PPM1D
Pediatric
15 mutations
7 cancer subtypes
4 datasets
p.N512Ifs*2
PP2C Protein phosphatase 2C
MISSENSE
SILENT
NONSENSE
UTR_3
FRAMESHIFT
SPLICE
Structural variation
Relapse
O
PPM1D cell dependency
other
LOF
driver missense
p=2e-20
p=5e-26
CERES score
P
PPM1D
(TVCVC* – TVCQW*)
Wild-type
Mutant
Stability

Figure S5. Detailed examination of genes whose truncating mutations are predicted to result in C-terminal degron loss. Related to Figure 5.

(A) Scatter plot showing the results of the mutational enrichment for N-end degron loss across all analyses (33 cancer types and pan-cancer). P-value resolution is limited to 0.0001.

(B) Essentiality of genes based on CRISPR screens of cells found in DepMap (lower CERES scores indicate a gene is more essential). Results are stratified by the mutation status of the gene (blue=putative driver mutation, orange=wildtype). GATA3 essentiality is shown for breast cancer cell lines and is compared to BRAF (an oncogene), FBXW7 (dominant-negative tumor suppressor), RB1 (tumor suppressor gene) and TTN (a highly mutated, common false positive gene in DNA sequencing studies). GATA3 showed a similar pattern of essentiality as BRAF. Boxplot shows quartiles with whiskers representing the quartile +/- 1.5 times the interquartile range.

(C) GPS stability measurements of the wild-type GATA3 C-terminal peptide (gray) versus the double mutant peptide (red).

(D) Protein expression of wildtype and mutant GATA3 with or without proteasome inhibitor treatment (MG132, 50 μM, 8 hours) in MDA-MB-231 cells, a triple-negative breast cancer cell line.

(E) RNA expression of GATA3 in MDA-MB-231 cells by qPCR. Error bars represent standard error.

(F) Protein expression of wildtype compared to mutant GATA3 in HEK293FT cells. F=FLAG tag.

(G) Protein expression of wildtype and mutant GATA3 with or without proteasome inhibitor treatment (MG132, 50 μM, 8 hours) in HEK293FT cells.

(H) RNA expression of GATA3 in HEK293FT cells by qPCR. Error bars represent standard error.

(I) Heatmap displaying normalized read coverage of peaks identified by GATA3-flag ChIP-seq.

(J) Heatmap displaying normalized read coverage of peaks identified by GATA3-flag ChIP-seq.

(K) Overlap of significantly up-regulated ChIP-seq peaks for GATA3 point mutants (p.A442M or p.G444E) with a FLAG-tag.

(L) Average read coverage profile for GATA3-flag ChIP-seq peaks (per million reads).

(M) Distribution of expression for genes nearby up-regulated peaks of mutant GATA3-flag, stratified by tumor subtype. Boxplot shows quartiles with whiskers representing the quartile +/- 1.5 times the interquartile range.

(N) Lollipop diagram of TCGA mutations in *PPM1D* (top) and St. Jude pediatric cancer mutations (bottom). The color of circles distinguishes the type of mutation, while colored rectangles are annotations of the protein. Numbered text indicates the position along the protein sequence.

(O) gRNA depletion from DepMap CRISPR screens for PPM1D stratified by TP53 mutation status. Driver missense mutations were defined by CHASMplus ($q<0.01$) and LOF (loss-of-function) consists of frameshift indels, nonsense, splice site and lost start mutations. Mann-Whitney U test was used. Boxplot shows quartiles with whiskers representing the quartile +/- 1.5 times the interquartile range.

(P) GPS stability measurements of the wild-type PPM1D C-terminal peptide (gray) versus the double mutant peptide (red).

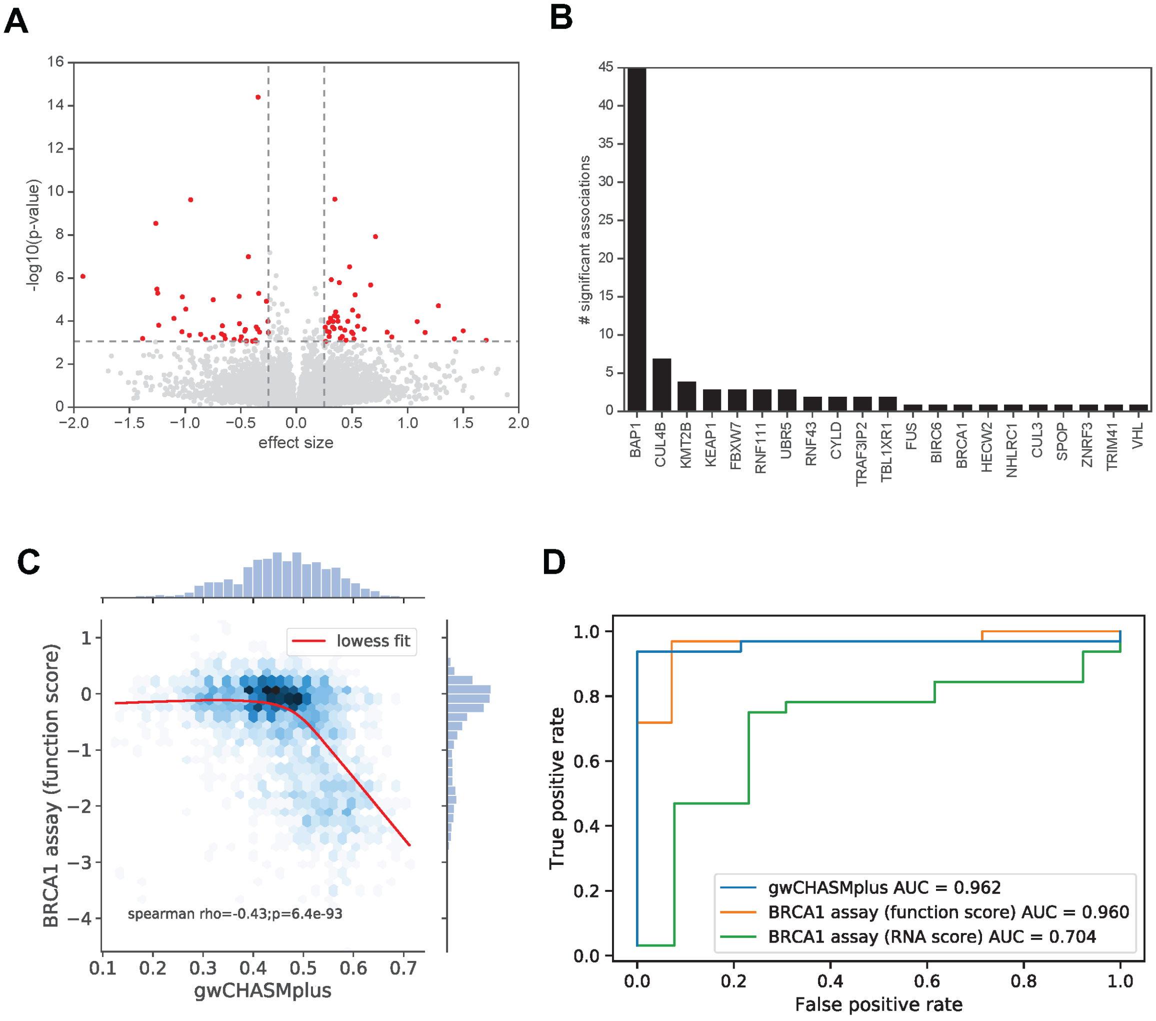
A
-log10(p-value)
16
14
12
10
8
6
4
2
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
effect size
B
# significant associations
45
40
35
30
25
20
15
10
5
0
BAP1
CUL4B
KMT2B
KEAP1
FBXW7
RNF111
UBR5
RNF43
CYLD
TRAF3IP2
TBL1XR1
FUS
BIRC6
BRCA1
HECW2
NHLRC1
CUL3
SPOP
ZNRF3
TRIM41
VHL
C
lowess fit
BRCA1 assay (function score)
1
0
−1
−2
−3
−4
spearman rho=-0.43;p=6.4e-93
0.1
0.2
0.3
0.4
0.5
0.6
0.7
gwCHASMplus
D
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
gwCHASMplus AUC = 0.962
BRCA1 assay (function score) AUC = 0.960
BRCA1 assay (RNA score) AUC = 0.704
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Figure S6. UPS-substrate inference by Reverse Phase Protein Arrays (RPPA) reveals few putative substrates. Related to Figure 6.

(A) Volcano plot showing the results of the correlation of UPS mutation status with protein abundance for all analyses. Effect size is the regression coefficient in the association.

(B) Number of significantly associated proteins (q<0.1) for each UPS gene across all cancer types.

(C) The small number of associated UPS genes is unlikely to be explained by labeling of driver mutated tumor samples. Predictions from CHASMplus are strongly negatively correlated with the function of BRCA1, an E3 ubiquitin ligase, from a prior saturation mutagenesis study.

(D) Performance of CHASMplus at distinguishing pathogenic from benign variants of BRCA1 from ClinVar.

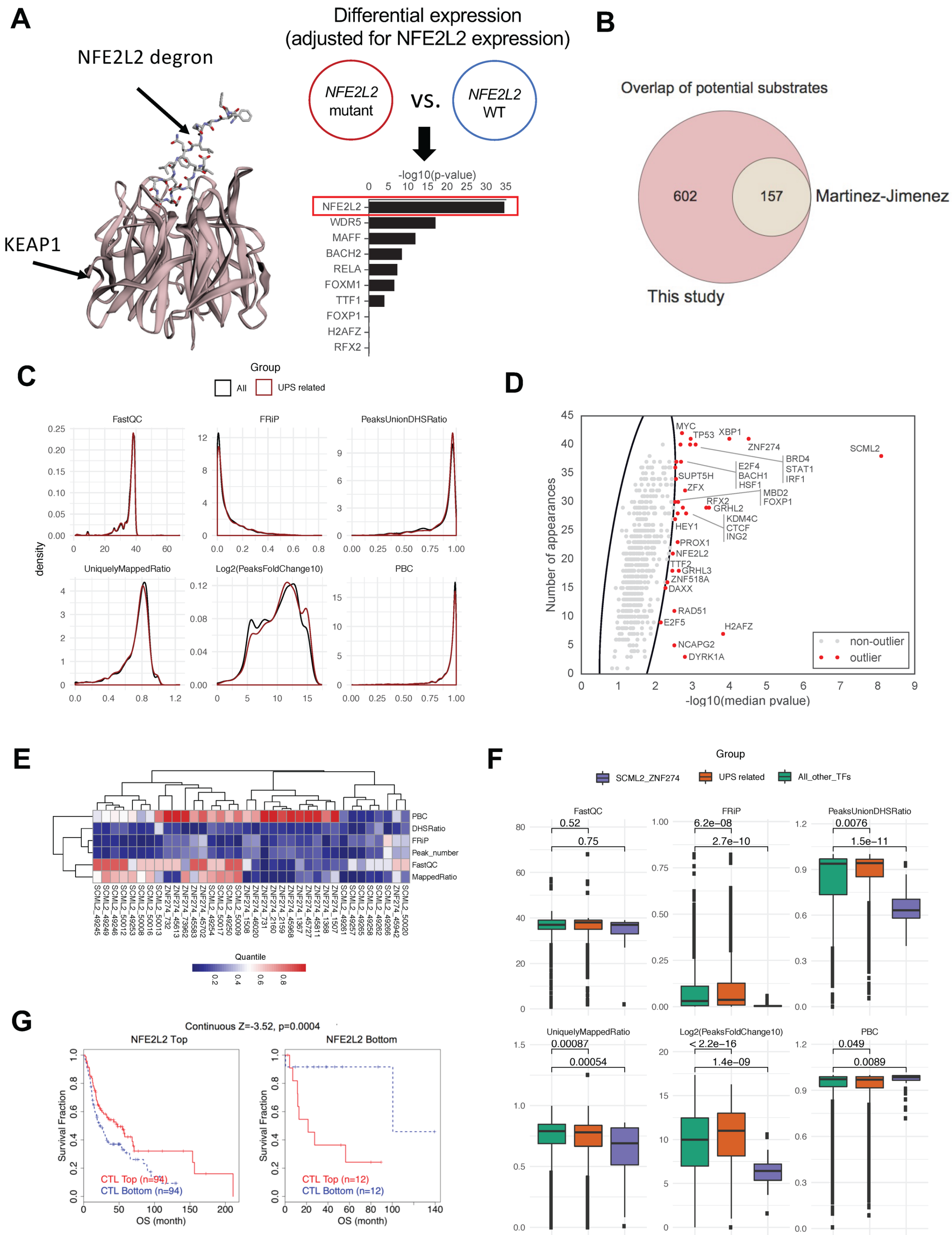
A
NFE2L2 degron
KEAP1
Differential expression
(adjusted for NFE2L2 expression)
NFE2L2 mutant
vs.
NFE2L2 WT
-log10(p-value)
0 5 10 15 20 25 30 35
NFE2L2
WDR5
MAFF
BACH2
RELA
FOXM1
TTF1
FOXP1
H2AFZ
RFX2
B
Overlap of potential substrates
602
157
Martinez-Jimenez
This study
C
Group
All
UPS related
density
FastQC
FRiP
PeaksUnionDHSRatio
UniquelyMappedRatio
Log2(PeaksFoldChange10)
PBC
D
Number of appearances
-log10(median pvalue)
MYC
TP53
XBP1
ZNF274
SCML2
BRD4
STAT1
IRF1
E2F4
BACH1
HSF1
SUPT5H
ZFX
MBD2
FOXP1
RFX2
GRHL2
KDM4C
CTCF
ING2
HEY1
PROX1
NFE2L2
TTF2
GRHL3
ZNF518A
DAXX
RAD51
E2F5
H2AFZ
NCAPG2
DYRK1A
non-outlier
outlier
E
PBC
DHSRatio
FRiP
Peak_number
FastQC
MappedRatio
SCML2_49245
SCML2_49249
SCML2_49246
SCML2_50012
SCML2_49253
SCML2_50008
SCML2_50016
SCML2_50013
ZNF274_732
ZNF274_45613
ZNF274_73962
ZNF274_45583
ZNF274_45702
SCML2_49254
SCML2_50017
SCML2_49250
SCML2_50009
ZNF274_1508
ZNF274_46020
ZNF274_731
ZNF274_2160
ZNF274_2159
ZNF274_45968
ZNF274_1367
ZNF274_45727
ZNF274_45811
ZNF274_1368
ZNF274_1507
SCML2_49261
SCML2_49257
SCML2_49265
SCML2_49258
SCML2_49262
SCML2_49266
ZNF274_45942
SCML2_50020
Quantile
0.2 0.4 0.6 0.8
F
Group
SCML2_ZNF274
UPS related
All_other_TFs
FastQC
0.52
0.75
FRiP
6.2e−08
2.7e−10
PeaksUnionDHSRatio
0.0076
1.5e−11
UniquelyMappedRatio
0.00087
0.00054
Log2(PeaksFoldChange10)
< 2.2e−16
1.4e−09
PBC
0.049
0.0089
G
Continuous Z=-3.52, p=0.0004
NFE2L2 Top
NFE2L2 Bottom
Survival Fraction
OS (month)
CTL Top (n=94)
CTL Bottom (n=94)
CTL Top (n=12)
CTL Bottom (n=12)

Figure S7. Quality control for UPS-TF substrate inference. Related to Figure 6 and Table 1.

(A) Proof-of-principle that a known transcription factor (NFE2L2) can be retrieved as the top hit when analyzing its own degron-containing mutations (left, PDB: 2FLU). A differential expression profile was generated based on comparing NFE2L2 mutant vs wildtype samples in lung squamous cell carcinoma (middle) after adjusting for covariates. Right, bar graph of the statistical significance (negative log-transformed p-values) of transcription factors identified as explaining the differential expression profile by RABIT. RABIT leverages thousands of uniformly processed ChIP-seq profiles to identify which genes are likely to be regulated by a transcription factor based on their binding location profiles.

(B) A Venn diagram displaying the greater extent of genes analyzed as potential UPS substrates in this study compared to Martinez-Jimenez et al., which did not analyze transcription factors.

(C) The distribution of ChIP-seq quality control metrics for transcription factors implicated with UPS genes are identical to those that were not associated. FastQC=median read quality scores from FastQC, FRiP=fraction of reads in peaks, PeaksUnionDHSRatio=proportion of peaks that overlap with open chromatin, UniquelyMappedRatio=proportion of reads that are uniquely mapped, Log2(PeaksFoldChange10)=log-transform of the number of peaks with at least 10 fold enrichment above background, and PBC=pcr bottleneck coefficient.

(D) Outlier analysis of the results reported by RABIT. RABIT selects a minimum set of transcription factors (based on corresponding ChIP-seq profiles) to explain a differential expression profile. Robust-covariance estimation analysis showed that several transcription factors were outliers in terms of the median p-value and the number of times they appeared in the RABIT results.

(E) Heatmap displaying ChIP-seq quality control (QC) metrics for the two biggest outliers: SCML2 and ZNF274.

(F) Statistical comparison of the distribution of QC metrics for SCML2 and ZNF274 showed substantially worse quality. Boxplot shows quartiles with whiskers representing the quartile +/- 1.5 times the interquartile range.

(G) Association of NRF2 protein abundance (encoded by the *NFE2L2* gene) with overall patient survival (Kaplan-Meier curves) has an interaction effect with a cytotoxic t lymphocyte (CTL) signature. Statistical significance assessed by Wald test of Cox Proportional Hazard model (Methods).