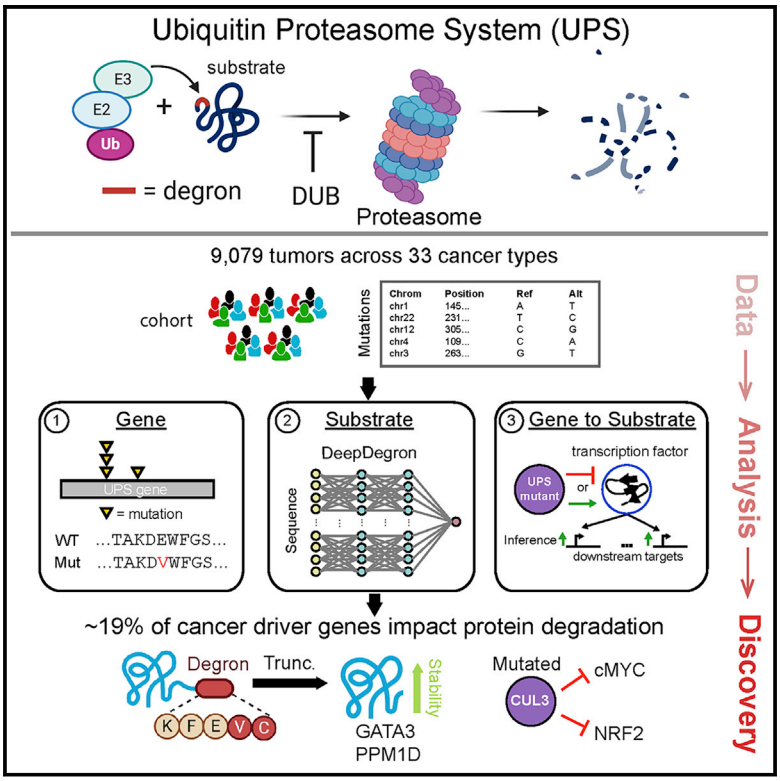
Molecular Cell

Article

Systematic characterization of mutations altering protein degradation in human cancers

Graphical Abstract



Highlights

- The ubiquitin-proteasome system (UPS) represents ~19% of mutated cancer driver genes
- A machine learning approach, deepDegron, reveals *de novo* degron motifs
- Truncating mutations in GATA3 and PPM1D increase protein expression via degron loss
- ChIP-seq data can help infer transcription factor substrates of mutated UPS in cancer

Authors

Collin Tokheim, Xiaoqing Wang, Richard T. Timms, ..., Stephen J. Elledge, Myles Brown, X. Shirley Liu

Correspondence

selledge@genetics.med.harvard.edu (S.J.E.), myles_brown@dfci.harvard.edu (M.B.), xsliu@ds.dfci.harvard.edu (X.S.L.)

In Brief

The mechanisms underlying oncogenic mutations in cancer remain incompletely understood. By leveraging machine learning, Tokheim et al. find that ~19% of cancer driver genes affect protein degradation, systematically revealing transcription factors as important substrates. They also validate an unconventional role of truncating mutations to increase stability of GATA3 and PPM1D.

Tokheim et al., 2021, Molecular Cell *81*, 1–17
March 18, 2021
https://doi.org/10.1016/j.molcel.2021.01.020

CellPress

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Molecular Cell

CellPress

Article

Systematic characterization of mutations altering protein degradation in human cancers

Collin Tokheim,[1,2,10] Xiaoqing Wang,[1,2,3,10] Richard T. Timms,[4,5,10,11] Boning Zhang,[1,2,3] Elijah L. Mena,[4,5] Binbin Wang,[6] Cynthia Chen,[7] Jun Ge,[6] Jun Chu,[8] Wubing Zhang,[1,6] Stephen J. Elledge,[4,5,*] Myles Brown,[3,9,*] and X. Shirley Liu[1,2,12,*]

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[4]Division of Genetics, Department of Medicine, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA 02115, USA
[5]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[6]Clinical Translational Research Center, Shanghai Pulmonary Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China
[7]The Harker School, San Jose, CA 95129, USA
[8]Key Laboratory of Xin'an Medicine, Ministry of Education, Anhui University of Chinese Medicine, Hefei, Anhui 230038, China
[9]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[10]These authors contributed equally
[11]Present address: Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK
[12]Lead contact
*Correspondence: selledge@genetics.med.harvard.edu (S.J.E.), myles_brown@dfci.harvard.edu (M.B.), xsliu@ds.dfci.harvard.edu (X.S.L.)
https://doi.org/10.1016/j.molcel.2021.01.020

SUMMARY

The ubiquitin-proteasome system (UPS) is the primary route for selective protein degradation in human cells. The UPS is an attractive target for novel cancer therapies, but the precise UPS genes and substrates important for cancer growth are incompletely understood. Leveraging multi-omics data across more than 9,000 human tumors and 33 cancer types, we found that over 19% of all cancer driver genes affect UPS function. We implicate transcription factors as important substrates and show that c-Myc stability is modulated by CUL3. Moreover, we developed a deep learning model (deepDegron) to identify mutations that result in degron loss and experimentally validated the prediction that gain-of-function truncating mutations in GATA3 and PPM1D result in increased protein stability. Last, we identified UPS driver genes associated with prognosis and the tumor microenvironment. This study demonstrates the important role of UPS dysregulation in human cancer and underscores the potential therapeutic utility of targeting the UPS.

INTRODUCTION

Cancer is fundamentally a disease of the genome, where only certain mutations drive a selective growth advantage for cancer cells, with most mutations being benign passengers that accumulate by chance. From the start of DNA sequencing studies of human tumors (Barbieri et al., 2012; Cancer Genome Atlas Research Network, 2008; Jones et al., 2008; Wood et al., 2007), it quickly became clear that genes involved in protein degradation are perturbed by mutations in cancer. For example, mutated *VHL* leads to elevated HIF-1/2a protein abundance, which allows cells to adapt to hypoxic conditions (Iliopoulos et al., 1996; Ivan et al., 2001; Iyer et al., 1998; Jaakkola et al., 2001). The ubiquitin-proteasome system (UPS) regulates degradation of over 80% of proteins in cells (Collins and Goldberg, 2017). UPS dysregulation has been implicated in nearly all hallmarks of cancer (Hanahan and Weinberg, 2011), such as *USP28* in the DNA damage response (Zhang et al., 2006), *KEAP1* in oxidative stress (Jaramillo and Zhang, 2013), and *FBXW7* in cell proliferation (King et al., 2013; Welcker and Clurman, 2008). Moreover, defects in the UPS have been linked to a variety of other human diseases or disorders (Atkin and Paulson, 2014; Das et al., 2006; Nalepa and Clapp, 2018; Staub et al., 1997); for example, loss-of-function mutations in *UBE3A* are implicated in Angelman syndrome, a neurodevelopmental disorder (Buiting et al., 2016). Despite the importance of UPS in human disease and especially cancer, a systems-level understanding of the UPS is still lacking.

The UPS operates through covalent attachment of ubiquitin (an 8-kDa protein) to lysine residues in substrate proteins, which is achieved through a relay of steps by passing ubiquitin from E1 enzymes to E2 enzymes (Stewart et al., 2016) and, ultimately with the help of E3 ubiquitin ligases, to substrates. Although ubiquitination can have many functions, polyubiquitination is often a

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

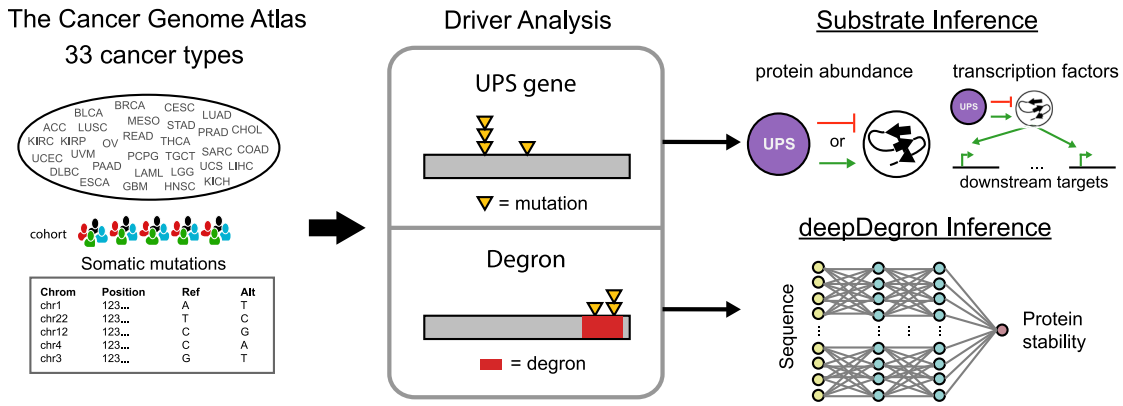


Figure 1. Study overview

Somatic mutations from 33 cancer types in The Cancer Genome Atlas (TCGA) (left) were analyzed to reveal significantly mutated genes in the ubiquitin-proteasome system (UPS) and its substrates with significant enrichment of mutations at known degron-related sites (center). A machine learning model, deepDegron (bottom right), was then used to find additional degron sites and to determine the effect of additional mutations. Last, leveraging the significantly mutated genes in the UPS pathway, we associated UPS pathway genes with protein abundance or inferred activity of TFs to implicate putative substrates (top right).

signal for protein degradation by the 26S proteasome (Collins and Goldberg, 2017). The key step of this process conferring regulatory specificity is performed by E3 ubiquitin ligases, which are thought to recognize short linear amino acid motifs, known as degrons, on substrate proteins (Mészáros et al., 2017). With over 600 E3 ubiquitin ligases encoded in the human genome, there are more than 10 million possible E3 ligase-substrate pairs. The transient nature of E3-substrate interactions makes experimental detection of these interactions using co-immunoprecipitation challenging (Ella et al., 2019; Mészáros et al., 2017). In addition, deubiquitinating enzymes (DUBs) act in the opposite direction, preventing degradation by removing ubiquitin from proteins (Reyes-Turcu et al., 2009; Ronau et al., 2016). Although many mechanistic steps of the UPS are well characterized, the regulatory logic of how E3 ubiquitin ligases and DUBs selectively recognize their target protein remains mostly unknown (Deshaies and Joazeiro, 2009). Because it is unclear which genes involved in ubiquitination act in a proteasome-dependent versus -independent manner, we include all E1, E2, E3, and DUB enzymes in our subsequent analyses.

Many tumors (37%–57%) harbor potentially clinically actionable mutations (Bailey et al., 2018; Zehir et al., 2017). Some of these are in genes that encode the UPS components; for instance, *BRCA1* (an E3 ubiquitin ligase) mutant tumors are sensitive to PARP inhibitors through a synthetic-lethal interaction (Robson et al., 2017). Traditionally, clinical actionability has been largely based on classic drug development of small molecules or antibodies that bind to an enzyme or receptor. Recent developments of protein degrader-based drugs, such as proteolysis-targeting chimeras (PROTACs) (Sakamoto et al., 2001; Winter et al., 2015), have promised to expand the scope of druggable targets in cancer through a novel mechanism of action. PROTACs that act by co-opting the cell's normal UPS machinery to degrade specific target proteins are in active development, and early PROTAC drugs are undergoing clinical trials for breast and prostate cancer (Scudellari, 2019). However, it is still not well understood how the UPS is usually perturbed in cancer and how PROTACs or other UPS-targeting drugs could counteract this effect. Thus, a comprehensive characterization of which mutated UPS genes may drive carcinogenesis and their corresponding dysregulated protein substrates is not only important for understanding cancer biology but also of potentially significant therapeutic utility.

Prior studies have been underpowered to identify significantly mutated genes within the UPS or lacked the capability to identify previously unknown substrates. Although Ge et al. (2018) found 23 mutated genes in the UPS to be statistically significant, their analysis mostly found already known genes and did not consider the affected substrates. Martínez-Jiménez et al. (2020) attempted to identify substrates of E3 ligases in cancer, but they only analyzed expression data for ~200 proteins (Li et al., 2013) and a handful of E3 ligases with already known degron motifs (Gouw et al., 2018). In contrast, our study considered all components of the UPS, including E1-activating enzymes, E2-conjugating enzymes, E3 ligases, and DUBs. In addition, we developed a machine learning method to systematically infer degron sequences *de novo* and identify mutated substrates that escape protein degradation. We aimed to provide the most systematic assessment of the role of protein degradation in human cancer to date, supported by experimental validation of our predictions.

In this study, to dissect the complex regulation of the UPS in cancer, we divided the problem into several steps: identifying mutated UPS genes, identifying mutated substrates, and linking mutated UPS genes to substrates (Figure 1). We employed integrative computational approaches to identify cancer driver genes in the UPS, associated these with candidate substrates through a multi-omics approach, and leveraged deep learning to model the effect of mutations on degrons. While investigating over 9,000 tumors in 33 cancer types, we found a significantly larger role of UPS dysregulation in carcinogenesis than appreciated previously, comprising approximately 19% of cancer driver genes. Predictions of mutations leading to degron loss in GATA3 and PPM1D were then validated experimentally. Furthermore, UPS alterations are associated with prognosis and immune

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

CellPress

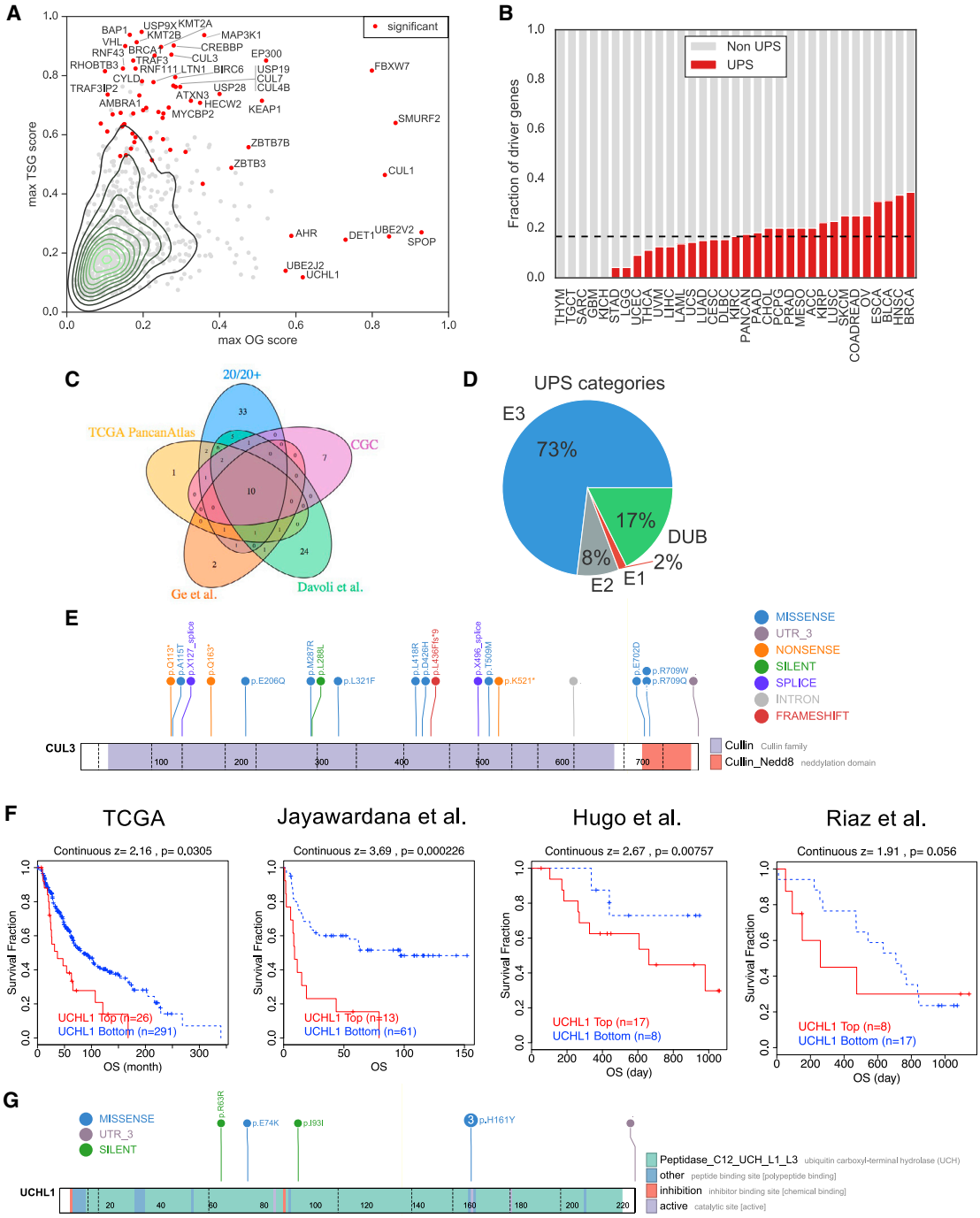


Figure 2. Landscape of cancer driver genes in the UPS

(A) Driver gene analysis was performed by the 20/20+ method. A scatterplot for each UPS gene (dots) is shown with the maximum oncogene (OG) score (x axis) and maximum tumor suppressor gene (TSG) score (y axis) across 33 cancer types and a pan-cancer analysis. Red indicates that the gene was found to be statistically significant in at least one analysis.

(B) Fraction of putative cancer driver genes that occur in the UPS pathway (red bar). A dashed line indicates the median across all analyses.

(C) Venn diagram showing the overlap of putative cancer driver genes in this study (20/20+) with previous studies: TCGA PancanAtlas consortium, ubiquitin pathway analysis by Ge et al. (2018), Davoli et al. (2013), and a curated list of cancer driver genes, in general, from the Cancer Gene Census (CGC).

(D) Pie diagram displaying the percentage of UPS driver genes in terms of molecular function.

(E) Lollipop diagram of CUL3 mutations in head and neck squamous cell carcinoma in TCGA. Exon-exon junctions are displayed as dashed lines. The colors of circles distinguish the types of mutation, and colored rectangles are Uniprot domain annotations of the protein.

(legend continued on next page)

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

infiltration of the tumor microenvironment (TME). Our results could provide insights into rational selection of protein degrader drugs to counteract the effects of UPS dysregulation in human cancer.

RESULTS

Expanded landscape of putative cancer driver genes in the UPS

An understanding of the UPS requires assessment of the genes comprising the pathway and the protein substrates they regulate (Figure 1). To establish a landscape of the former in cancer, we evaluated whether UPS genes (Table S1) were somatically mutated more often than expected by chance across a large cohort of tumors from The Cancer Genome Atlas (TCGA). The rationale is that driver mutations in a UPS gene would confer a selective growth advantage to a clonal cell population, leading to cancer, which leaves a statistically distinguishable signal compared with mutations that happen by chance. Using the 20/20+ method we developed previously to identify mutated cancer driver genes (Tokheim et al., 2016; STAR methods), we found a total of 63 unique UPS genes as putative drivers ($q < 0.05$; Figure 2A; Table S1), covering 28 of 33 analyzed cancer types (Figure 2B). The putative UPS drivers are enriched for curated cancer driver genes in the Cancer Gene Census ($p = 2e{-}11$, two-tailed Fisher's exact test) (Sondka et al., 2018), driver genes defined by the TCGA consortium ($p = 6e{-}25$) (Bailey et al., 2018), and biological processes relevant to cancer (Figure S1D). Moreover, unlike a recent study (Martínez-Jiménez et al., 2020), which only includes E3 ubiquitin ligases, our analysis includes E2-conjugating enzymes, E1-activating enzymes, and deubiquitinases, which led to a greater number of putative UPS driver genes with better agreement with prior literature (Figures S1A-S1C). Notably, compared with the results from the TCGA consortium, the putative UPS drivers represented ~16% of all driver genes, including 33 genes not reported previously (Figure 2C; Table S1), which suggests a substantial role of the UPS in carcinogenesis. Reflective of their occurrence in diverse cancer types, UPS driver genes showed substantial variability in gene dependencies across cell lineages from CRISPR knockout (KO) (Figures S1E–S1G) and contextually co-occurred with other mutations (Figures S1H–S1J). Last, we stratified mutated UPS genes by oncogene or tumor suppressor gene scores from 20/20+ and observed the majority to be tumor suppressors (Figure 2A). In some cases, a tumor suppressor gene may also have a high "oncogene" score because of the presence of recurrent hotspot mutations in addition to truncating mutations, suggesting that the hotspot mutations have a dominant-negative effect (Davis et al., 2014).

The 63 putative UPS driver genes spanned E3 ubiquitin ligases (n = 46), E2-conjugating enzymes (n = 5), E1-activating enzymes (n = 1), and DUBs (n = 11) (Figure 2D). Identified components of E3 ubiquitin ligases represent not only target recognition subunits but also cullin scaffold proteins (n = 5). These included *CUL3*, which exhibited widely distributed loss-of-function mutations and a recurrent mutation (p.R709) near the activating neddylation site (Figure 2E). Although DUBs were fewer in number, expression of driver DUB genes had prognostic value in 16 of 33 analyzed cancer types (Figure S1L; STAR methods). For example, high expression of *UCHL1* was significantly associated with worse overall survival in individuals with metastatic melanoma in the TCGA (Figure 2F), consistent with our prediction of *UCHL1* being an oncogene in melanoma because of a recurrent H161Y mutation at its active site (Figure 2G). The *UCHL1* gene expression association was also replicated in an independent metastatic melanoma cohort ($p = 0.0002$, Cox Proportional Hazards model) (Jayawardana et al., 2015). We reasoned that because *UCHL1* expression is associated with a poor prognosis in melanoma, it might also be relevant in recent immunotherapy trials in melanoma. Indeed, *UCHL1* expression was also associated with worse overall survival in a study of anti-PD-1 treatment ($p = 0.008$) (Hugo et al., 2016) and approached significance in another study with nivolumab for treatment-naive individuals ($p = 0.06$) (Riaz et al., 2017). This underscores that E3 ubiquitin ligases and DUBs might have important roles in cancer progression.

Degron annotations limit the number of significantly mutated UPS substrates

Although alterations affecting genes in the UPS pathway would be expected to lead to multiple changes in downstream protein substrates, mutations in the substrates themselves could provide greater specificity for cancer cells by affecting much fewer proteins. We therefore hypothesized that we could identify substrate mutations under positive selection in tumors by finding enriched missense mutations at known degron-related sites (STAR methods). From the PhosphoSitePlus database (Hornbeck et al., 2015), we found that mutations were enriched in annotated ubiquitination sites in the *SF3B1* gene in breast cancer and in the *KIT* gene in cutaneous melanoma ($q < 0.1$; Table S2). Mutations were also enriched at annotated degron sites (Mészáros et al., 2017) located in *CTNNB1* (Figure S2A), *SPRY1*, *NFE2L2* (Figure S2B), and *EPAS1* (Figure S2C) and phosphodegron sites located in *CTNNB1* and *CCND1* ($q < 0.1$; Table S2; Figure 3A). An example is CCND1-mutant endometrial tumors (Figure 3B), which as expected showed higher protein expression (Figure 3C, left) and greater cell cycle progression than wild-type tumors (Figure 3C, right). Surprisingly, mutations outside of the phosphodegron also displayed a similar trend, largely consisting of truncating mutations that also eliminate the phosphodegron (Figure 3B) while being predicted to escape nonsense-mediated decay (NMD) (Lindeboom et al., 2016). Likewise, *CTNNB1*-mutant tumors were also associated with a functional effect, including altered transcriptional activity (Figures S2D and S2E), activation of WNT signaling (Figure S2F), and an altered TME (Figure S2G), consistent with previous reports (Hatzis et al., 2008; Spranger et al.,

(F) Kaplan-Meier curves of the relationship between *UCHL1* expression and overall survival in 4 melanoma datasets.
(G) Lollipop diagram of *UCHL1* mutations in the TCGA skin cutaneous melanoma cohort. Numbered circles indicate that a mutation was found in more than one tumor.
See also Figure S1.

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

CellPress

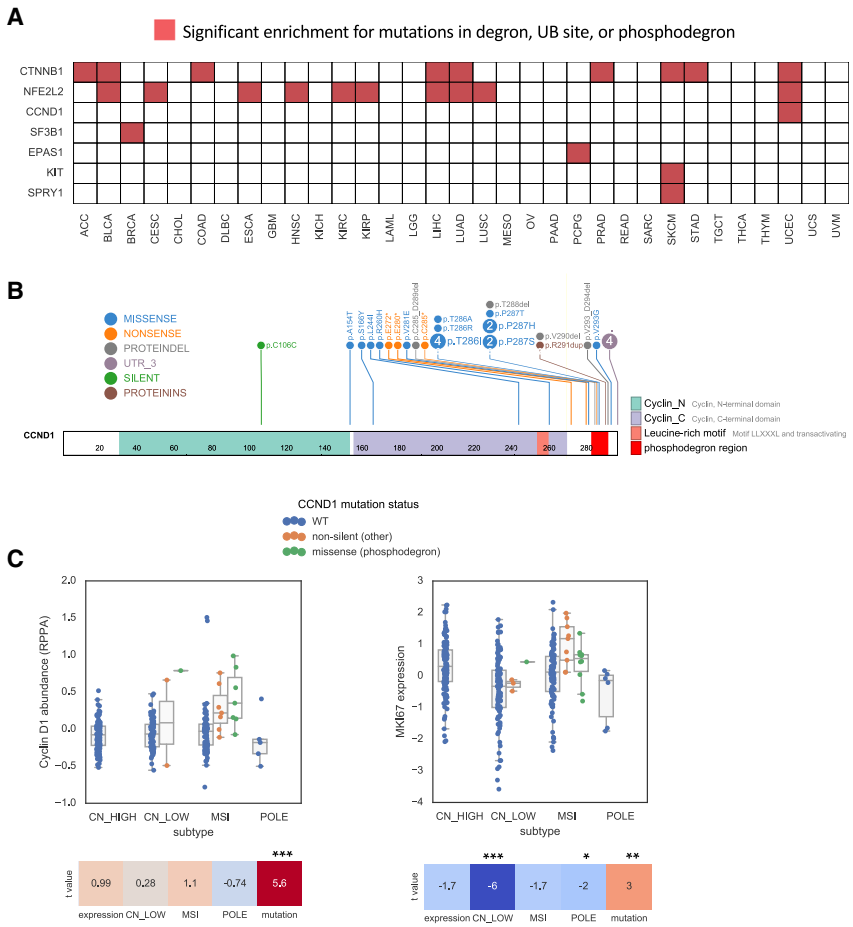


Figure 3. Somatic mutations are enriched at known degron sites

(A) Heatmap displaying genes that are enriched for mutations at literature-annotated degron sites (Mészáros et al., 2017), ubiquitination sites (PhosphositePlus), or phosphodegrons (PhosphoSitePlus). Red indicates significant enrichment ($q < 0.1$) for a given gene (y axis) and cancer type (x axis) in TCGA.

(B) Lollipop diagram of *CCND1* mutations in uterine corpus endometrial carcinoma (UCEC) in TCGA.

(C) Boxplots showing the association of *CCND1* mutations with Cyclin D1 protein abundance ($p = 4e{-}8$, Wald test) and a marker of cell cycle progression (MKI67, $p = 0.003$) in UCEC. The heatmap shows t statistics of the association after adjustment for RNA expression and tumor subtype. Tumor subtypes: CN_LOW, copy number low; MSI, microsatellite instable; POLE, POLE mutated gene. RPPA, reverse-phase protein arrays. All boxplots show the distribution quartiles with whiskers representing the quartile ± 1.5 times the Interquartile Range (IQR).

See also Figure S2.

2015). In total, the significantly mutated genes affecting the UPS, either in the pathway directly or on their substrates, comprise 19% of all cancer driver genes relative to the TCGA PancanAtlas consortium analysis (Bailey et al., 2018). However, the smaller number of genes with mutations in degrons in cancer is likely due to the considerable sparsity of known degron annotations (Mészáros et al., 2017). Therefore, the true proportion of cancer driver genes affecting the UPS is very likely to be higher than 19%.

deepDegron infers degron sequences

Although a few UPS substrate mutations can be implicated in cancer based on known degrons, systematic investigation requires better degron annotation. To address this challenge, we developed a protein sequence-based model, deepDegron, that leverages data from recently published global protein stability (GPS) analysis of N-terminal and C-terminal sequences from the human proteome (Koren et al., 2018; Timms et al., 2019) to predict degrons (Figure S3). GPS uses fluorescence-activated cell sorting (FACS) to quantify protein stability based on the abundance of a fluorescent reporter protein (GFP, green) fused to a short peptide compared with a control reporter with no fusion partner attached (DsRed, red) (Figure S3A). Because the peptides consisted of known sequences and could contain degrons, we reasoned that deepDegron could learn the sequence rules of the degron effect on protein stability. deepDegron is a feedforward neural network with one input layer, two hidden layers with a rectified linear unit activation function, and an output layer (Figure S3B; STAR methods). Hyperparameters were determined by performance on a leave-out dataset, such as the number of units in each layer, dropout rate, training epochs, and peptide sequence encoding (Figure S3C). On a held-out test set, deepDegron achieved high performance when predicting the results of the GPS assay (Figures 4A and 4B). This was higher than the previously proposed rule-based alternatives (Koren et al., 2018), such as the numbers of bulky amino acids, acidic residues, or top 100 motifs, and better than a combination thereof (logistic regression) (Figures 4A and 4B).

Protein stability is likely affected by general biophysical characteristics of the attached peptide in the GPS assay, such as hydrophobicity and intrinsic disorderedness (van der Lee et al., 2014). However, we were most interested in understanding the specific sequence motifs that might mediate degron recognition by specific ubiquitin ligases. Therefore, to infer degrons, we trained two deep learning models: one containing position information from the primary sequence and another without position information ("bag of amino acids" representation; Figure 4C). We hypothesized that the difference between these two models could approximate a degron potential score, where high scores demonstrate position-specific features to be more informative than general degradation properties.

To identify the degron motifs learned by deepDegron, we performed *de novo* motif enrichment analysis from the human N and C terminomes (Table S4; STAR methods). Our analysis revealed numerous previously known motifs (Figure S4), such as -GG

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

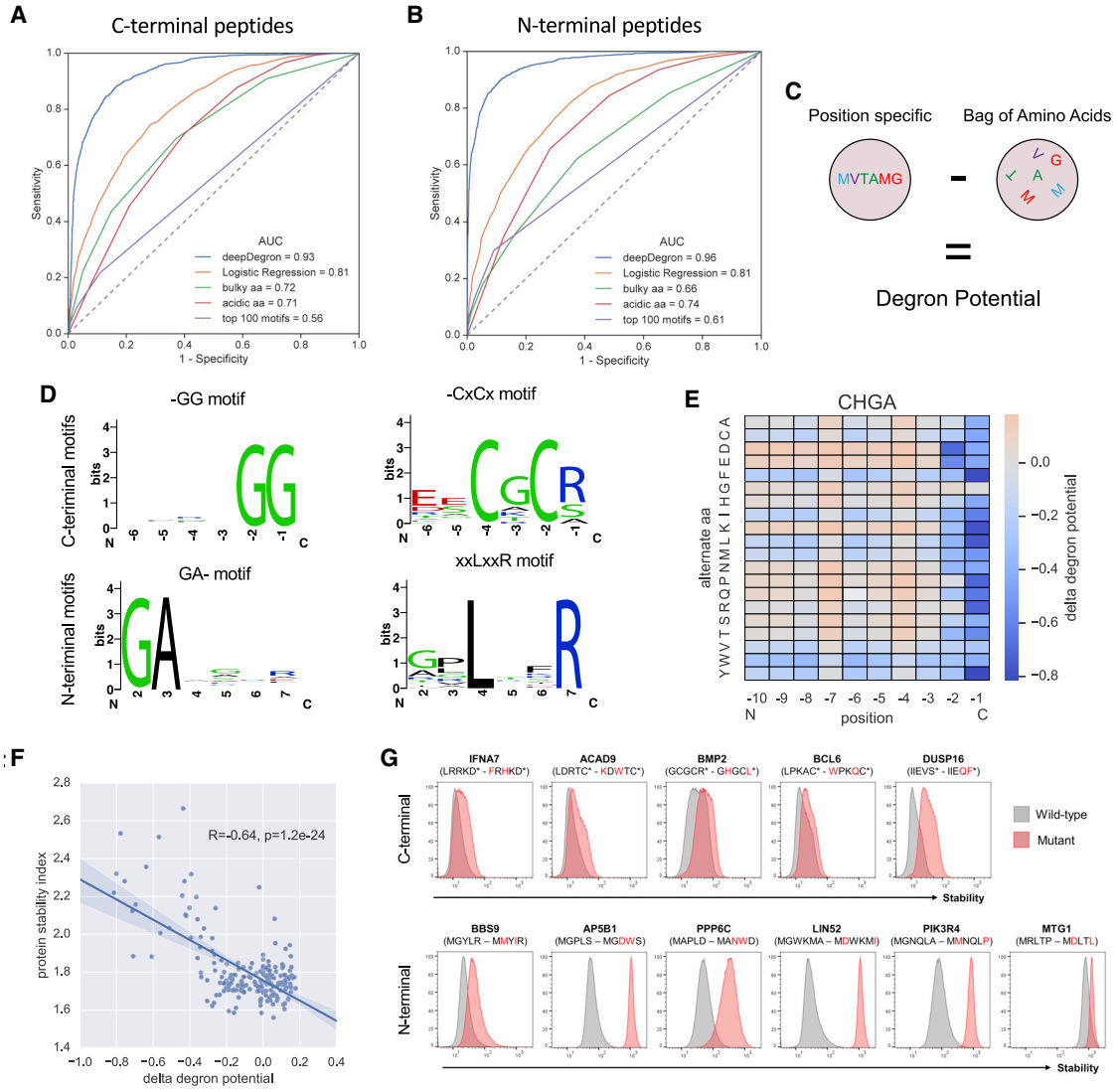


Figure 4. deepDegron accurately predicts the effect of primary sequence on protein stability

(A) Performance of deepDegron when predicting the stability of C-terminal peptides from the global protein stability (GPS) assay according to the area under the receiver operating characteristic curve (AUC; maximum = 1.0, random = 0.5) (see deepDegron and Dataset in STAR methods).

(B) Receiver Operating Characteristic (ROC) curve for the N-terminal peptide GPS assay.

(C) Diagram showing that the degron potential score is computed based on the difference between a deepDegron model that uses the position of the amino acids versus one that does not ("bag of amino acids").

(D) Sequence logo visualizations of select motifs identified by deepDegron ($q < 0.05$, binomial test; STAR methods).

(E) DeepDegron-predicted change in degron potential (delta degron potential) for various mutations of the C-terminal peptide encoded by *CHGA*.

(F) Correlation between the change in degron potential and the protein stability index according to a saturation mutagenesis study of CHGA.

(G) 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) sequences. The x axis is proportional to the GFP/DsRed signal, as measured by flow cytometry (STAR methods); the y axis is normalized cell count.

See also Figures S3 and S4.

and GA- in C-end and N-end degrons, respectively, but also previously unknown motifs, such as C-terminal C[A/G]C[R] and N-terminal [P]LxxR (Figure 4D). Although previous models have emphasized the effect of di-amino acid motifs on C- and N-end degrons (Koren et al., 2018; Timms et al., 2019), the discovered motifs suggest that additional complexity might exist with a longer extended degron, albeit with partial degeneracy at these residues, as evidenced by the sequence logo plots (Figure 4D). To assess whether deepDegron could accurately predict the effect of mutations on degrons, we evaluated its performance relative to saturation mutagenesis experiments (Koren et al., 2018; Timms et al., 2019). For example, the deepDegron

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

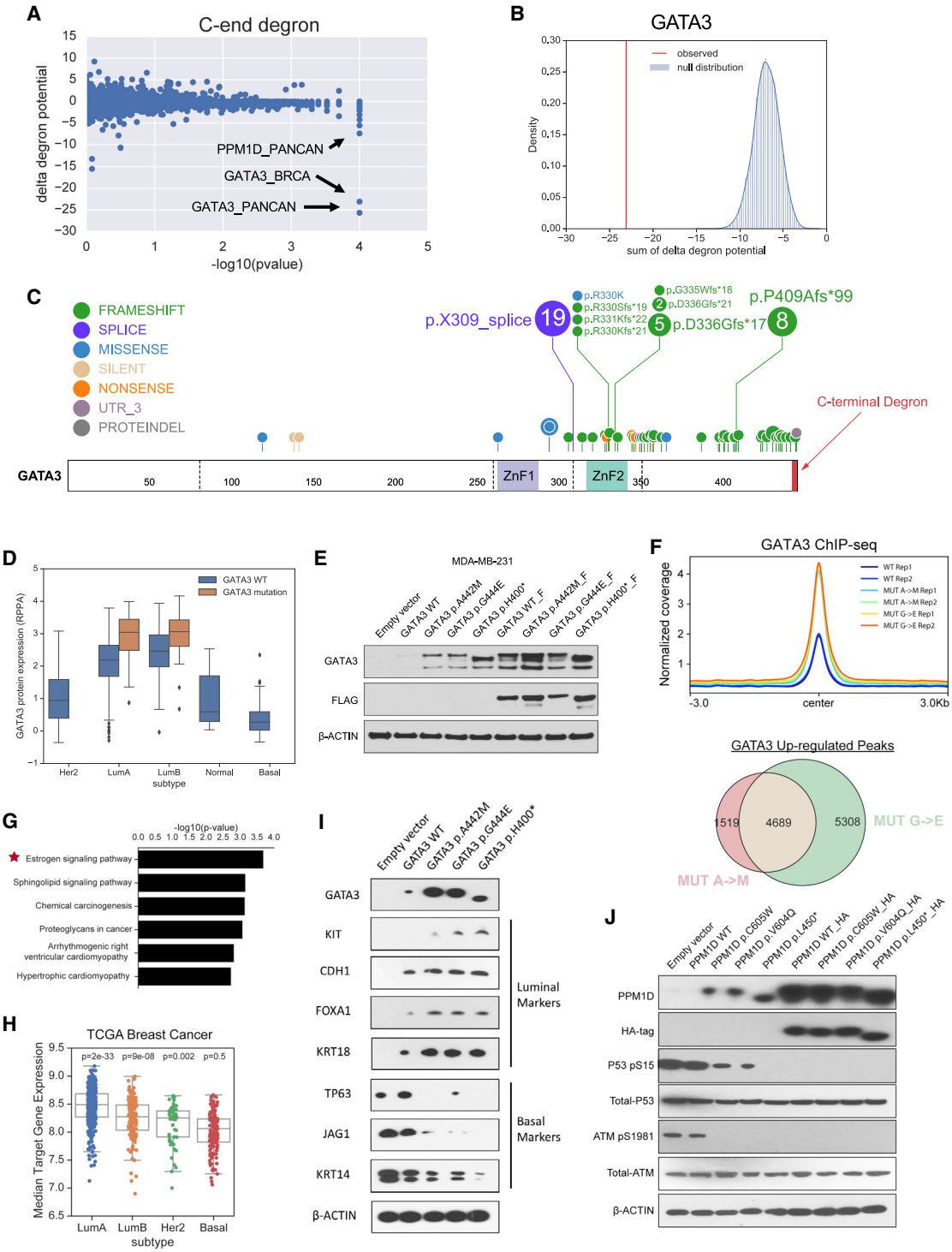


Figure 5. deepDegron finds C-terminal degrons disrupted by mutations in cancer

(A) Scatterplot showing the results of the mutational enrichment for C-end degron loss across all analyses (33 cancer types and pan-cancer). The p value resolution is limited to 0.0001.

(B) Example of *GATA3* in breast cancer, which shows that the change in degron potential (red) is considerably more negative than the background model (blue).

(C) Lollipop diagram of TCGA mutations in *GATA3* for breast cancer. Colored rectangles represent zinc-finger domains 1 (ZnF1) and 2 (ZnF2).

(D) Boxplot showing the association of *GATA3* mutations with GATA3 protein abundance in TCGA breast cancer (top left). All boxplots show the distribution quartiles with whiskers representing the quartile ± 1.5 times the Interquartile Range (IQR).

(legend continued on next page)

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

model scored most mutations in the C-end -RG motif as disrupting a degron in the CHGA protein (Figure 4E), as demonstrated by a strong negative change in degron potential when the last two amino acids are mutated. Indeed, compared with the experimental results, the predicted change in degron potential was, as expected, negatively correlated with protein stability (Figure 4F). Moreover, this negative correlation was observed for all saturation mutagenesis experiments performed on N-terminal and C-terminal peptides (Figures S4E and S4F). These results suggest that deepDegron is capable of capturing the sequence-level rules of degrons.

To experimentally validate the new degron predictions by the deepDegron model, we used the GPS stability assay. We selected 21 significant degron motifs for testing, comprising 9 predicted N-terminal degrons and 12 predicted C-terminal degrons (STAR methods). GPS was used to examine the stability of the terminal 23-mer peptide derived from each of the 21 proteins, comparing the wild-type sequence with a mutant version containing two point mutations in the putative degron motif. The precise mutations were chosen to maximize the decrease in degron potential, as determined by deepDegron (Table S4; STAR methods). We found that mutation of 8 of 12 (67%) C-terminal degrons and 8 of 9 (89%) N-terminal degrons resulted in protein stabilization (Figures 4G, S4G, S5C, and S5N; Table S4G). These results underscored the potential power of deepDegron as a tool for degron discovery.

deepDegron identifies mutations likely disrupting degrons in cancer

Given the strong concordance between deepDegron's predictions and the available experimental data, we reasoned that we could systematically apply deepDegron to identify mutations that may disrupt degrons in cancer. We thus computed the change in degron potential between the mutated and wild-type sequence in TCGA (delta degron potential) and assessed whether there was enrichment for mutations predicted to disrupt a degron in genes (STAR methods). Our analysis revealed that mutations in *GATA3* and *PPM1D* had the most significantly disrupted degrons across all analyzed cancer types ($q < 0.1$, Figures 5A and S5A; Table S5). Indeed, for breast cancer, in which *GATA3* was identified as significant, the change in degron potential (-23) had far more of an effect than expected by chance (Figures 5B and S4D).

GATA3 is an essential transcription factor (Figure S5B) that regulates luminal differentiation of mammary tissue (Kouros-Mehr et al., 2006) and cooperates with *ESR1* to mediate estrogen response (Eeckhoute et al., 2007; Theodorou et al., 2013). Heterozygous *GATA3* mutations typically occur in the estrogen receptor (ER)+ subtype of breast cancer (luminal A or luminal B) and show a clear bias for frameshift and splice site mutations near the 3′ end of the gene (Figure 5C). Notably, the mutations are clustered on the last exon-exon junction so that they are not expected to cause NMD (Lindeboom et al., 2016). According to the deepDegron model, the -AxG sequence (x = any amino acid) at the C terminus of wild-type GATA3 is strongly predictive of its degron potential, and frameshift or splice site mutations would eliminate this motif. Consistent with the predicted loss-of-degron effect for these mutations, we found that *GATA3*-mutant tumors in TCGA had elevated protein abundance according to reverse-phase protein arrays (RPPAs) ($p = 9e{-}9$, Wald test; Figure 5D). To experimentally confirm the minimal degron region, we generated a double point mutant in the C-terminal -AxG motif of GATA3 and measured protein stability by GPS assay. Similar to clinical tumor samples, we found that the double point mutant of the GATA3 C terminus had significantly higher protein expression compared with the wild-type sequence (Figure S5C). Moreover, individual substitution of either amino acid led to increased protein expression in the context of the full-length GATA3 protein, as assessed by immunoblot, suggesting that both residues are critical for degron recognition (Figure 5E). Given that GATA3 was also upregulated substantially upon treatment with the proteasome inhibitor MG132 (Figure S5D), the identified -AxG motif is likely a degron that mediates protein degradation of GATA3 via the UPS. Additionally, RNA expression was not elevated substantially in the mutants (Figure S5E), ruling out potential transcriptional effects. These findings were further confirmed in a second cell line (HEK293FT), underscoring the robustness of our finding that mutations lead to degron loss in GATA3 (Figures S5F–S5H).

Next we sought to evaluate whether *GATA3* mutations mediate their effect on breast cancer through elevated protein expression. If so, then these mutations should shift a basal-like breast cancer cell line (MBA-MD-231) toward a gene expression program of ER+ breast cancer. We therefore compared the genome-wide binding sites of mutated GATA3 with wild-type GATA3 by chromatin immunoprecipitation sequencing (ChIP-seq) (Figures S5I and S5J; Tables S5B–S5E). As a control, we created *GATA3* constructs that would be stable regardless of point mutation status by adding a FLAG tag to the C terminus of GATA3 (Figure 5E). Addition of residues blocks the function of the C-end degron because the location at the extreme C terminus is required (Koren et al., 2018). Notably, GATA3 mutations led to a consistent overall gain in binding compared with wild-type GATA3 ($p < 1e{-}16$, Fisher's exact test), but only without a FLAG tag control (Figures 5F, S5K, and S5L). Upregulated binding sites were preferentially near estrogen signaling genes (Figure 5G), but no pathway was enriched in the presence of a FLAG tag control (false discovery rate [FDR] < 0.1). Moreover, genes closest to upregulated binding sites displayed substantially higher expression in ER+

(E) Western blot of the protein expression of GATA3 mutants compared with the control. F, FLAG tag.
(F) Top: average read coverage profile for peaks. Bottom: overlap of upregulated ChIP-seq peaks for GATA3 mutants.
(G) Pathway enrichment analysis of upregulated peaks for GATA3 mutants.
(H) Distribution of expression for genes near upregulated peaks stratified by tumor subtype.
(I) Western blot showing the effect of mutating the GATA3 degron on markers for luminal and basal-like breast cancer.
(J) Western blot analysis of the PPM1D (WIP1) mutant versus the control. HA, hemagglutinin.
See also Figure S5.

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

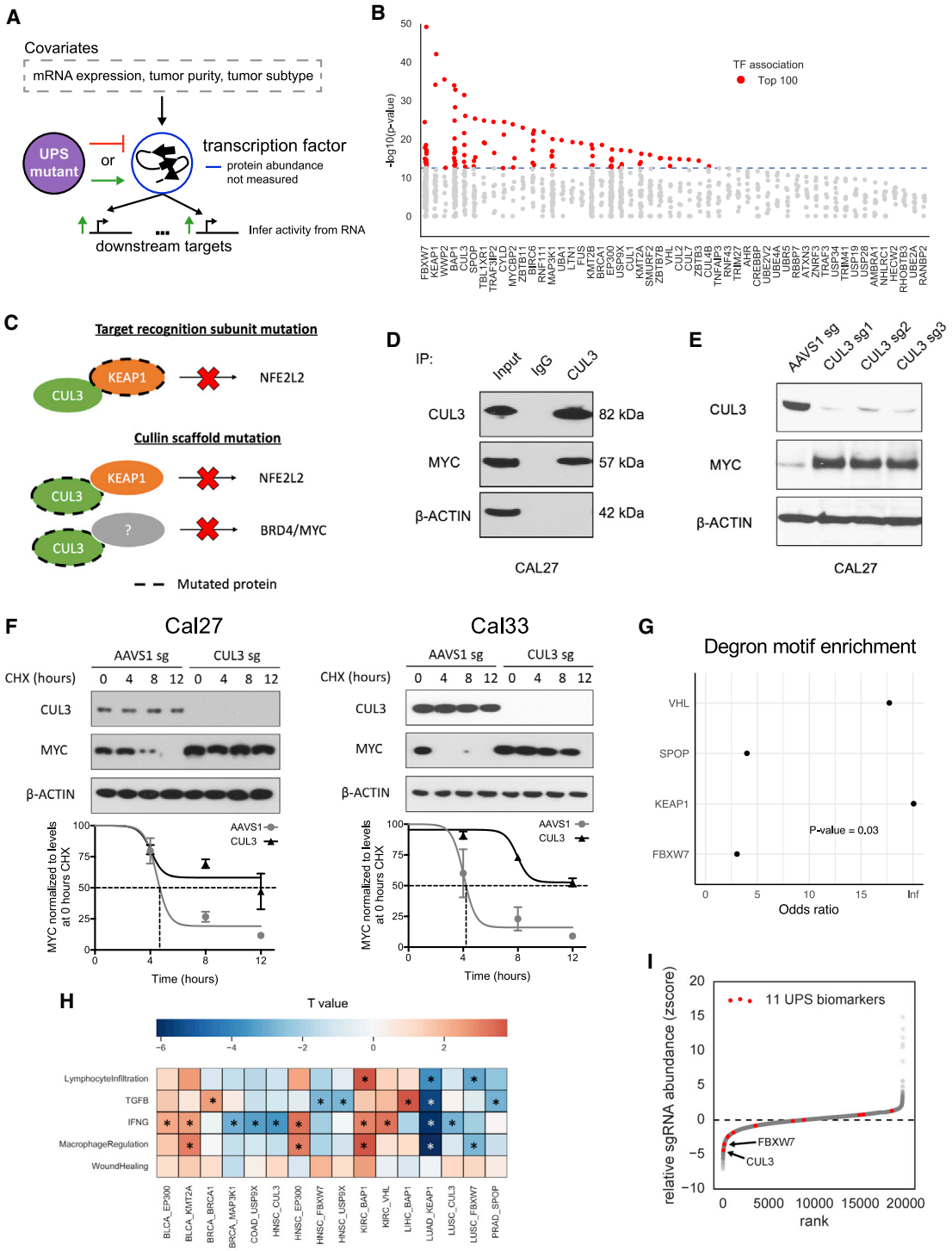


Figure 6. UPS-substrate inference finds associations with markers of the immune TME

(A) Diagram depicting the strategy for associating UPS genes with putative TF substrates.

(B) Scatterplot showing the significance of each transcription factor (TF) association for a particular UPS gene (x axis).

(C) Diagram of inferred substrate relationships of KEAP1 and CUL3.

(D) Western blot showing co-immunoprecipitation of CUL3 with c-Myc.

(E) Western blot showing increased c-Myc protein abundance in CUL3 KO cells.

(legend continued on next page)

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

compared with basal-like subtypes of breast cancer (Figure 5H), which was not the case with the FLAG tag control (Figures S5K–S5M). Last, mutation of the GATA3 degron shifted protein expression biomarkers toward an ER+ state in the basal-like MDA-MB-231 breast cancer cell line (Figure 5I). GATA3 mutations in breast cancer, at least in part, mediate their effect by increasing protein stability through elimination of a degron.

Similar to the GATA3 prediction, deepDegron also predicted that truncating mutations in *PPM1D* will disrupt a C-terminal -VC degron motif (Figure S5N). *PPM1D* encodes the Ser/Thr phosphatase WIP1, which negatively regulates p53 (Bulavin et al., 2002; Emelyanov and Bulavin, 2015) and has been reported to be amplified frequently in breast cancer (Li et al., 2002; Rauta et al., 2006). Consistent with an oncogenic role through negative regulation of *TP53*, PPM1D is more essential in *TP53* wild-type compared with *TP53* mutant cell lines from CRISPR screens reported in DepMap (Figure S5O). Furthermore, *PPM1D*-truncating mutations observed in TCGA were mutually exclusive with *TP53* mutations ($p = 0.04$, one-sided Mantel-Haenszel test), suggesting that they might redundantly affect the same pathway. Supporting our prediction of a mechanism involving degron loss, a double point mutant of the -VC motif in the WIP1 C-terminal peptide displayed elevated protein stability by GPS (Figure S5P). Point mutation of either amino acid residue also led to increased protein expression of full-length WIP1, according to western blot analysis (Figure 5J), suggesting that both amino acids are critical. Functionally, the higher protein expression of mutant WIP1 resulted in greater dephosphorylation of known downstream targets in the DNA damage response pathway (Figure 5J), including p53 (Lu et al., 2005; Shreeram et al., 2006). Although *in vivo* evidence of WIP1 protein expression is not available in TCGA, similar truncating mutations have been reported to lead to greater protein stability of WIP1 and to chemotherapy resistance in acute myeloid leukemia (Hsu et al., 2018; Kahn et al., 2018). Our finding of truncating mutations leading to C-end degron loss in WIP1 (the *PPM1D* gene) is consistent with this clinical phenomenon.

Integrative analysis of UPS driver genes identifies putative transcription factor (TF) substrates

Having analyzed UPS substrates and UPS driver genes in isolation, we next wanted to explore pairing of UPS genes with their substrates. One approach is to correlate the presence of putative driver mutations in UPS components with protein abundance measurements of potential substrates from RPPAs (Li et al., 2013) after adjusting for RNA expression and other covariates (STAR methods). Although we could confirm known UPS-substrate relationships, such as targeting of CCNE1 by FBXW7 (Koepp et al., 2001; Strohmaier et al., 2001), we could only find a small number of associations (Table S6). This is unlikely to be due to incorrect labeling of cancer driver mutations (STAR methods) because our predictions were significantly correlated with a previous saturation mutagenesis experiment performed on the E3 ubiquitin ligase *BRCA1* (Figure S6). Rather, RPPA only contains abundance measurements for a limited number of proteins ($n = 198$).

To expand our analyses (Figures S7A and S7B), we reasoned that TFs might be a substrate particularly amenable for analysis because RNA expression of a TF's target genes might serve as a proxy for TF protein activity (Figure 6A). We generated differential expression profiles comparing tumor samples carrying wild-type versus putative driver mutations in the UPS genes (STAR methods). RNA expression of the TF was then adjusted as a covariate, presumably leaving effects of the TF based on the protein level. TF regulator analysis using RABIT (Jiang et al., 2015) was then performed to infer substrate TFs based on their target genes defined by thousands of uniformly processed TF ChIP-seq profiles from the Cistrome database (Zheng et al., 2019).

As a proof of principle, we first tested whether a known TF, *NFE2L2*, could be retrieved by analyzing its own degron mutations. Indeed, *NFE2L2* was identified correctly as the top hit (Figure S7A) to explain the differentially expressed genes in tumors containing *NFE2L2* mutations. Applying the method globally to UPS-substrate inference, we found 494 cancer-specific associations (Table S7) to be significant at a conserved family-wise error rate of 0.05 (Bonferroni method, corresponding to $p < 7.8e{-}7$). Because some could be downstream effects, we decided to focus on the top 100 associations (Figure 6B), where, at most, 5 associations per UPS gene are shown in Table 1. Importantly, there was no indication of systematic differences in ChIP-seq quality in our significant results (Figures S7C–S7F; STAR methods), suggesting that technical artifacts are likely low.

Numerous UPS-substrate associations we identified have been validated previously, such as FBXW7 and c-Myc (encoded by the *MYC* gene; King et al., 2013), SPOP and androgen receptor (AR) (An et al., 2014), and BRCA1 and ERα (encoded by the *ESR1* gene; Eakin et al., 2007; Ma et al., 2010). In some cases, although not finding the direct target, our analysis found proteins that were regulated by the direct target, such as the UPS gene CYLD and downstream RELA (Kovalenko et al., 2003), or interaction partners in the same protein complex, such as VHL-ARNT, where ARNT forms a dimer with the known VHL target HIF-1α (Tanimoto et al., 2000). For example, as indicated by our results and previously (Shibata et al., 2008), KEAP1 directly regulates NRF2 (encoded by the *NFE2L2* gene) and is known to form a complex with CUL3, a scaffold protein for many substrate recognition subunits (Figure 6C). As expected, *CUL3* mutations also showed modulation of NRF2, but, unlike KEAP1, were associated with *MYC* or

(F) Quantification of c-Myc protein half-life upon CUL3 KO in Cal27 and Cal33 cells. Cycloheximide (CHX), a protein translation inhibitor, was given at a concentration of 100 μg/mL. Error bar, ±1 SEM.
(G) Enrichment analysis for degron motifs in associated TFs for 4 E3 ubiquitin ligases that have a previously reported degron motif (Fisher's exact test).
(H) Heatmap displaying the association (t statistic) of mutations in UPS driver genes with 5 immune-related biomarkers, *, FDR < 0.1.
(I) *Z* score measuring the relative abundance of cancer cells with a gene KO when they are co-cultured with T cells, where negative values indicate sensitivity to T cell killing.
See also Figures S6 and S7.

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Table 1. Mutated UPS driver genes are associated with TF activity

UPS gene	TF (Cancer Type)
FBXW7	EP300 (LUSC, CESC), KLF4 (HNSC, LUSC), MYC (READ, BLCA), GRHL2 (HNSC), XBP1 (UCS)
KEAP1	NFE2L2 (PANCAN, LUAD)
WWP2	EP300 (HNSC), KDM4C (HNSC)
BAP1	XBP1 (BRCA, CHOL, MESO), YY1 (LIHC, PANCAN), CDK9 (PANCAN), MITF (UVM), TAF1 (PANCAN)
CUL3	NFE2L2 (PANCAN, LUSC, KIRP), MYC (PANCAN, KIRP, HNSC), BRD4 (KIRP)
SPOP	AR (PRAD), EP300 (UCEC), NKX3-1 (PRAD), ARRB1 (PRAD)
TBL1XR1	XBP1 (BRCA), BRD4 (BRCA), MBD2 (BRCA)
TRAF3IP2	MYC (BLCA), EED (BLCA)
CYLD	RELA (HNSC), EP300 (HNSC), FOS (HNSC)
MYCBP2	PROX1 (COAD), MAX (COAD), MYC (COAD)
ZBTB11	EED (HNSC)
BIRC6	HNF4A (ESCA), FOS (HNSC), EP300 (HNSC), MAX (ESCA), KDM4C (HNSC)
RNF111	EP300 (HNSC)
MAP3K1	XBP1 (PANCAN), ESR1 (PANCAN), WDR5 (BRCA), EP300 (CESC), GRHL2 (CESC)
UBA1	RUNX1 (LAML)
LTN1	TTF1 (LUAD)
FUS	EP300 (BLCA)
KMT2B	STAT1 (HNSC), RFX1 (PANCAN), REST (COAD), CDX2 (COAD), HEY1 (PANCAN)
BRCA1	ESR1 (BRCA)
EP300	IRF4 (BRCA), XBP1 (HNSC, CESC), MAX (HNSC), SMARCA4 (PANCAN), KLF5 (CESC)
USP9X	XBP1 (PCPG), GRHL2 (HNSC), GTF2B (BRCA), IRF2 (COAD)
CUL1	CDK8 (BLCA)
KMT2A	FLI1 (LIHC), NR2F2 (LIHC), FOXO1 (BLCA)
SMURF2	FOXP1 (SKCM)
ZBTB7B	MYOD1 (UCS), GABPA (UCS)
VHL	STAT1 (KIRC), ARNT (KIRC)
CUL2	SUPT5H (BLCA)
CUL7	CTCF (BRCA)
ZBTB3	MYH11 (LAML)
CUL4B	STAT1 (LGG)

See also Figure S7.

BRD4 in our analysis. This would suggest an effect based on a different substrate recognition subunit. Supportive of this hypothesis, *NFE2L2* showed co-essentiality with *KEAP1* ($p = 2.6e{-}7$, Wald test) and *CUL3* ($p = 0.02$, Wald test) in cancer cell lines from DepMap CRISPR screens, whereas *MYC* is co-essential with *CUL3* ($p = 0.0004$, Wald test) but not with *KEAP1*. As expected for a direct regulatory relationship, CUL3 and c-Myc co-immunoprecipitated (Figure 6D), and KO of *CUL3* resulted in elevated protein expression (Figure 6E) and increased the protein half-life of c-Myc in CAL27 cells (Figure 6F). The increased c-Myc protein half-life was also reproducible in a second cell line (Figure 6F, right), which suggests that the of role of CUL3 in degrading c-Myc is robust. Although a direct assessment of the overall sensitivity or specificity of our approach is not possible because of limited known examples, we did find that, for the four E3 ubiquitin ligases with reported degron motifs (Mészáros et al., 2017), there was a significant enrichment of degron motifs in our results ($p = 0.03$; Figure 6G). This suggests that, overall, our analysis is enriched for direct targets.

UPS driver genes correlate with an altered immune TME

We noticed that many of the TFs in our analysis are related to interferon response (*STAT1*, *IRF2*, and *IRF4*) or are potentially immunomodulatory (*RELA*, *XBP1*, and *MYC*) (Cubillos-Ruiz et al., 2017; Grivennikov et al., 2010; Kortlever et al., 2017; Wellenstein and de Visser, 2018). We therefore sought to examine whether mutations in our putative UPS driver genes are associated with an altered immune TME. By correlating the tumor mutation status with previous immune-related signatures from TCGA (Thorsson et al., 2018), we found that 11 UPS genes had a significant correlation ($q < 0.1$; Figure 6G; Table S3). Many of the associations were related to interferon gamma (IFNG) response, so we examined whether they were hits in a previous CRISPR screen of cancer cells co-cultured with T cells (Pan et al., 2018). In this CRISPR screen, KO of genes in cancer cells that regulate T cell-mediated killing are expected to affect the fitness of those cancer cells *in vitro*, leading to altered representation of corresponding guide RNAs (STAR methods). Indeed, guide RNAs targeting CUL3 ($q = 0.0001$) and FBXW7 ($q = 0.002$) exhibited significant depletion in the CRISPR

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

screen (Figure 6H), suggesting increased sensitivity to T cell killing. Because CUL3 mutations in human tumors are correlated with a weak IFNG response, this would suggest that CUL3 mutations might only be advantageous for cancer cells in a low-IFNG environment or that CUL3 mutations might attenuate the cancer cell response to IFNG. In either scenario, we reasoned that an altered cancer cell IFNG response could change the anti-tumor efficacy of cytotoxic T lymphocytes (CTLs). Indeed, for head and neck squamous cell carcinoma, we found that a proxy for CUL3-activitiy based on NRF2 (encoded by *NFE2L2*) protein abundance altered the association between a CTL biomarker and overall survival (Figure S7H). Future experiments are needed to clarify which CUL3 substrate recognition adaptor protein and its corresponding substrate mediate this effect. These analyses revealed a potential immunomodulatory role of UPS in IFNG response in cancer.

DISCUSSION

Although the tumor transcriptome has been studied extensively by RNA sequencing (RNA-seq), how dysregulated pathways lead to altered proteomic states is far less understood. This is in spite of early DNA sequence studies of human cancers implicating driver genes that affect protein degradation through the UPS (Barbieri et al., 2012). In this study, we addressed this issue by performing a systematic analysis of the UPS and its corresponding substrates in 33 human cancer types. This revealed a much larger role of the UPS in cancer than appreciated previously, constituting over 19% of cancer driver genes. Moreover, our study includes the technical innovation of modeling degron loss by deep learning (deepDegron) and associating potential TF substrates of UPS genes by their inferred activity from TF ChIP-seq targets. By considering all components of the whole UPS pathway, *de novo* degrons from machine learning, and TF substrates, our study increased the significantly mutated UPS genes compared with Ge et al. (2018) by ~3-fold and expanded analysis of UPS substrates by ~4-fold compared with Martínez-Jiménez et al. (2020). These approaches could also be leveraged by researchers of other diseases to interpret the role of protein degradation.

Our study has several limitations. First, although our analysis of TF substrates of the UPS did identify *bona fide* direct targets, it was unavoidable to also find TFs that are regulated by or reside in the same protein complex as the actual substrate. Thus, careful considerations should be given to the possibility of related proteins when interpreting results. Second, although our analysis had the power to identify UPS driver genes mutated at low frequencies, for some of these genes, there were simply not enough mutations to make confident associations with potential substrates. Larger multi-omics studies or larger tumor profiling cohorts will be better powered to make such connections in the future. Last, although our study is an important advance in trying to systematically understand the UPS in cancer, we are far from having a complete landscape. One reason is that, because of the lack of systematic protein stability assays, we were not able to infer mutated degrons in the middle of proteins. The other reason is that, because mass spectrometry-based proteomics that can assess upward of 10,000 proteins have only been conducted on limited samples for limited cancer types (Mertins et al., 2016; Zhang et al., 2016), we could only associate potential substrates that are TF or on the RPPA panel (~200 proteins).

Although truncating mutations are commonly associated with tumor suppressor genes and a loss-of-function effect (Vogelstein et al., 2013), we found that truncating mutations may actually be gain-of-function mutations in oncogenes for more cases than appreciated previously. For example, *CCND1*, *GATA3*, and *PPM1D* have truncating mutations clustered near the 3′ end of their respective genes, which are predicted to lead to degron loss and showed evidence of higher protein abundance. This is somewhat surprising because it has been suggested previously that truncated proteins are degraded rapidly by protein quality control mechanisms (Goldberg, 2003). Indeed, clinical databases such as OncoKB (Chakravarty et al., 2017) have assumed GATA3-truncating mutations as likely being loss-of-function mutations, but our evidence suggests otherwise. Experimental point mutants of the GATA3 degron recapitulated the increased protein abundance seen for truncating mutations. Notably, truncation of the N-terminal part of proteins that lead to degron loss is appreciated for fusion genes, such as TMPRSS2:ETV1 (Vitari et al., 2011) and TMPRSS2:ERG (Gan et al., 2015). It is also possible that truncation of other types of inhibitory sequences could produce similar pro-oncogenic phenotypes, so not all cases of clustered truncating mutations may result in degron loss. Nonetheless, we expect that, as more degron motifs are discovered, there will be a concordant increase in identifying gain-of-function truncating mutations.

Our finding that most driver genes in the UPS are tumor suppressors suggests that therapeutic targeting of upregulated substrates may be a more efficacious strategy than targeting the UPS driver genes themselves. Indeed, mutations in the E3 ligase SPOP abrogate AR protein degradation (An et al., 2014), and targeted therapies against (non-mutated) AR are effective in prostate cancer (Watson et al., 2015). The advent of PROTACs may be a key advance because unaffected UPS genes could be co-opted to replace the function of mutated tumor suppressor genes. Moreover, this approach conceptually could be applied to target substrates that have escaped UPS recognition through mutations that result in degron loss. Numerous questions about the UPS remain to be answered. Are mutations in UPS driver genes selected preferentially because of their effect on single or multiple substrates? What are all of the substrates of each specific UPS driver gene? Why are UPS genes drivers in one cancer type but not in another? Future studies with increasing scales of tumor proteome-wide profiles may resolve these questions and capture a comprehensive picture of how the UPS modulates cancer initiation and progression. A better understanding of the UPS will undoubtedly provide insights guiding the development of novel cancer therapies that target protein degradation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2021.01.020.

ACKNOWLEDGMENTS

This study was partially supported by the Breast Cancer Research Foundation (BCRF-20-100 to X.S.L.) and the NIH (AG11085 to S.J.E.). C.T. is a Damon Runyon fellow supported by the Damon Runyon Cancer Research Foundation (DRQ-04-20). R.T.T. is supported by a Pemberton-Trinity fellowship and a Sir Henry Wellcome postdoctoral fellowship (201387/Z/16/Z). S.J.E. is an investigator with the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

C.T. and X.S.L. conceived the study. C.T., B.W., J.G., C.C., W.Z., R.T.T., S.J.E., and X.S.L. drafted and edited the manuscript. C.T. developed the computational methods. R.T.T. and S.J.E. contributed GPS data. X.W., B.Z., J.C., R.T.T., and E.L.M. performed experiments. C.T., B.W., J.G., W.Z., and C.C. analyzed the results.

DECLARATION OF INTERESTS

S.J.E. is a member of the *Molecular Cell* advisory board. X.S.L. is a cofounder, board member, SAB, and consultant of GV20 Oncotherapy and its subsidiaries and the SAB of 3DMedCare; a consultant for Genentech; a stockholder of BMY, TMO, WBA, ABT, ABBV, and JNJ; and receives research funding from Takeda and Sanofi. M.B. is a consultant to and receives sponsored research support from Novartis. M.B. serves on the SAB of H3 Biomedicine, Kronos Bio, and GV20 Oncotherapy.

Received: April 14, 2020
Revised: December 1, 2020
Accepted: January 17, 2021
Published: February 9, 2021

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

An, J., Wang, C., Deng, Y., Yu, L., and Huang, H. (2014). Destruction of full-length androgen receptor by wild-type SPOP, but not prostate-cancer-associated mutants. Cell Rep. *6*, 657–669.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data (Babraham Bioinformatics).

Atkin, G., and Paulson, H. (2014). Ubiquitin pathways in neurodegenerative disease. Front. Mol. Neurosci. *7*, 63.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al.; MC3 Working Group; Cancer Genome Atlas Research Network (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371–385.e18.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–689.

Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center

(LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B *57*, 289–300.

Buiting, K., Williams, C., and Horsthemke, B. (2016). Angelman syndrome - insights into a rare neurogenetic disorder. Nat. Rev. Neurol. *12*, 584–593.

Bulavin, D.V., Demidov, O.N., Saito, S., Kauraniemi, P., Phillips, C., Amundson, S.A., Ambrosino, C., Sauter, G., Nebreda, A.R., Anderson, C.W., et al. (2002). Amplification of PPM1D in human tumors abrogates p53 tumor-suppressor activity. Nat. Genet. *31*, 210–215.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Caruana, R., and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning (ACM), pp. 161–168.

Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis. Oncol. *2017*, PO.17.00011.

Collins, G.A., and Goldberg, A.L. (2017). The Logic of the 26S Proteasome. Cell *169*, 792–806.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Cubillos-Ruiz, J.R., Bettigole, S.E., and Glimcher, L.H. (2017). Tumorigenic and Immunosuppressive Effects of Endoplasmic Reticulum Stress in Cancer. Cell *168*, 692–706.

Das, C., Hoang, Q.Q., Kreinbring, C.A., Luchansky, S.J., Meray, R.K., Ray, S.S., Lansbury, P.T., Ringe, D., and Petsko, G.A. (2006). Structural basis for conformational plasticity of the Parkinson's disease-associated ubiquitin hydrolase UCH-L1. Proc. Natl. Acad. Sci. USA *103*, 4675–4680.

Davis, R.J., Welcker, M., and Clurman, B.E. (2014). Tumor suppression by the Fbw7 ubiquitin ligase: mechanisms and opportunities. Cancer Cell *26*, 455–464.

Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962.

Deshaies, R.J., and Joazeiro, C.A. (2009). RING domain E3 ubiquitin ligases. Annu. Rev. Biochem. *78*, 399–434.

Eakin, C.M., Maccoss, M.J., Finney, G.L., and Klevit, R.E. (2007). Estrogen receptor alpha is a putative substrate for the BRCA1 ubiquitin ligase. Proc. Natl. Acad. Sci. USA *104*, 5794–5799.

Eeckhoute, J., Keeton, E.K., Lupien, M., Krum, S.A., Carroll, J.S., and Brown, M. (2007). Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. Cancer Res. *67*, 6477–6483.

Ella, H., Reiss, Y., and Ravid, T. (2019). The Hunt for Degrons of the 26S Proteasome. Biomolecules *9*, 230.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al.; MC3 Working Group; Cancer Genome Atlas Research Network (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. *6*, 271–281.e7.

Emelyanov, A., and Bulavin, D.V. (2015). Wip1 phosphatase in breast cancer. Oncogene *34*, 4429–4438.

Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.

Gan, W., Dai, X., Lunardi, A., Li, Z., Inuzuka, H., Liu, P., Varmeh, S., Zhang, J., Cheng, L., Sun, Y., et al. (2015). SPOP Promotes Ubiquitination and Degradation of the ERG Oncoprotein to Suppress Prostate Cancer Progression. Mol. Cell *59*, 917–930.

Ge, Z., Leighton, J.S., Wang, Y., Peng, X., Chen, Z., Chen, H., Sun, Y., Yao, F., Li, J., Zhang, H., et al.; Cancer Genome Atlas Research Network (2018). Integrated Genomic Analysis of the Ubiquitin Pathway across Cancer Types. Cell Rep. *23*, 213–226.e3.

Goldberg, A.L. (2003). Protein degradation and protection against misfolded or damaged proteins. Nature *426*, 895–899.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F., and Hansen, L.K. (1999). On clustering fMRI time series. Neuroimage *9*, 298–310.

Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L.B., Davey, N.E., Deng, Z., et al. (2018). The eukaryotic linear motif resource - 2018 update. Nucleic Acids Res. *46* (D1), D428–D434.

Grivennikov, S.I., Greten, F.R., and Karin, M. (2010). Immunity, inflammation, and cancer. Cell *140*, 883–899.

Haigis, K.M., Cichowski, K., and Elledge, S.J. (2019). Tissue-specificity in cancer: The rule, not the exception. Science *363*, 1150–1151.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hatzis, P., van der Flier, L.G., van Driel, M.A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I.J., Koster, J., Santo, E.E., Welboren, W., et al. (2008). Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol. Cell. Biol. *28*, 2732–2744.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. arXiv, arXiv:1512.03385 https://arxiv.org/abs/1512.03385.

Hellmann, M.D., Nathanson, T., Rizvi, H., Creelan, B.C., Sanchez-Vega, F., Ahuja, A., Ni, A., Novik, J.B., Mangarin, L.M.B., Abu-Akeel, M., et al. (2018). Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. Cancer Cell *33*, 843–852.e4.

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520.

Hsu, J.I., Dayaram, T., Tovy, A., De Braekeleer, E., Jeong, M., Wang, F., Zhang, J., Heffernan, T.P., Gera, S., Kovacs, J.J., et al. (2018). PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. Cell Stem Cell *23*, 700–713.e6.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell *165*, 35–44.

Iliopoulos, O., Levy, A.P., Jiang, C., Kaelin, W.G., Jr., and Goldberg, M.A. (1996). Negative regulation of hypoxia-inducible genes by the von Hippel-Lindau protein. Proc. Natl. Acad. Sci. USA *93*, 10595–10599.

Ivan, M., Kondo, K., Yang, H., Kim, W., Valiando, J., Ohh, M., Salic, A., Asara, J.M., Lane, W.S., and Kaelin, W.G., Jr. (2001). HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. Science *292*, 464–468.

Iyer, N.V., Kotch, L.E., Agani, F., Leung, S.W., Laughner, E., Wenger, R.H., Gassmann, M., Gearhart, J.D., Lawler, A.M., Yu, A.Y., and Semenza, G.L. (1998). Cellular and developmental control of O2 homeostasis by hypoxia-inducible factor 1 alpha. Genes Dev. *12*, 149–162.

Jaakkola, P., Mole, D.R., Tian, Y.M., Wilson, M.I., Gielbert, J., Gaskell, S.J., von Kriegsheim, A., Hebestreit, H.F., Mukherji, M., Schofield, C.J., et al. (2001). Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science *292*, 468–472.

Jaramillo, M.C., and Zhang, D.D. (2013). The emerging role of the Nrf2-Keap1 signaling pathway in cancer. Genes Dev. *27*, 2179–2191.

Jayawardana, K., Schramm, S.J., Haydu, L., Thompson, J.F., Scolyer, R.A., Mann, G.J., Müller, S., and Yang, J.Y. (2015). Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. Int. J. Cancer *136*, 863–874.

Jiang, P., Freedman, M.L., Liu, J.S., and Liu, X.S. (2015). Inference of transcriptional regulation in cancers. Proc. Natl. Acad. Sci. USA *112*, 7731–7736.

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., Li, Z., Traugh, N., Bu, X., Li, B., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat. Med. *24*, 1550–1558.

Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., et al. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science *321*, 1801–1806.

Jonsson, P., Bandlamudi, C., Cheng, M.L., Srinivasan, P., Chavan, S.S., Friedman, N.D., Rosen, E.Y., Richards, A.L., Bouvier, N., Selcuklu, S.D., et al. (2019). Tumour lineage shapes BRCA-mediated phenotypes. Nature *571*, 576–579.

Kahn, J.D., Miller, P.G., Silver, A.J., Sellar, R.S., Bhatt, S., Gibson, C., McConkey, M., Adams, D., Mar, B., Mertins, P., et al. (2018). *PPM1D*-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. Blood *132*, 1095–1105.

King, B., Trimarchi, T., Reavie, L., Xu, L., Mullenders, J., Ntziachristos, P., Aranda-Orgilles, B., Perez-Garcia, A., Shi, J., Vakoc, C., et al. (2013). The ubiquitin ligase FBXW7 modulates leukemia-initiating cell activity by regulating MYC stability. Cell *153*, 1552–1566.

Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv, arXiv:14126980 https://arxiv.org/abs/1412.6980.

Koepp, D.M., Schaefer, L.K., Ye, X., Keyomarsi, K., Chu, C., Harper, J.W., and Elledge, S.J. (2001). Phosphorylation-dependent ubiquitination of cyclin E by the SCFFbw7 ubiquitin ligase. Science *294*, 173–177.

Koren, I., Timms, R.T., Kula, T., Xu, Q., Li, M.Z., and Elledge, S.J. (2018). The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons. Cell *173*, 1622–1635.e14.

Kortlever, R.M., Sodir, N.M., Wilson, C.H., Burkhart, D.L., Pellegrinet, L., Brown Swigart, L., Littlewood, T.D., and Evan, G.I. (2017). Myc Cooperates with Ras by Programming Inflammation and Immune Suppression. Cell *171*, 1301–1315.e14.

Kouros-Mehr, H., Slorach, E.M., Sternlicht, M.D., and Werb, Z. (2006). GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. Cell *127*, 1041–1055.

Kovalenko, A., Chable-Bessia, C., Cantarella, G., Israël, A., Wallach, D., and Courtois, G. (2003). The tumour suppressor CYLD negatively regulates NF-kappaB signalling by deubiquitination. Nature *424*, 801–805.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Communications of the ACM *60*, 84–90.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, J., Yang, Y., Peng, Y., Austin, R.J., van Eyndhoven, W.G., Nguyen, K.C., Gabriele, T., McCurrach, M.E., Marks, J.R., Hoey, T., et al. (2002). Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23. Nat. Genet. *31*, 133–134.

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W., Yang, J.Y., Broom, B.M., Verhaak, R.G., Kane, D.W., et al. (2013). TCPA: a resource for cancer functional proteomics data. Nat. Methods *10*, 1046–1047.

Li, S., Wan, C., Zheng, R., Fan, J., Dong, X., Meyer, C.A., and Liu, X.S. (2019). Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks. Nucleic Acids Res. *47* (W1), W206–W211.

Lindeboom, R.G., Supek, F., and Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat. Genet. *48*, 1112–1118.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.; Cancer Genome Atlas Research Network (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400–416.e11.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lu, X., Nannenga, B., and Donehower, L.A. (2005). PPM1D dephosphorylates Chk1 and p53 and abrogates cell cycle checkpoints. Genes Dev. *19*, 1162–1174.

Ma, Y., Fan, S., Hu, C., Meng, Q., Fuqua, S.A., Pestell, R.G., Tomita, Y.A., and Rosen, E.M. (2010). BRCA1 regulates acetylation and ubiquitination of estrogen receptor-alpha. Mol. Endocrinol. *24*, 76–90.

Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N., and Gonzalez-Perez, A. (2020). Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. Nat. Cancer *1*, 122–135.

Masica, D.L., Douville, C., Tokheim, C., Bhattacharya, R., Kim, R., Moad, K., Ryan, M.C., and Karchin, R. (2017). CRAVAT 4: Cancer-Related Analysis of Variants Toolkit. Cancer Res. *77*, e35–e38.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al.; NCI CPTAC (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

Mészáros, B., Kumar, M., Gibson, T.J., Uyar, B., and Dosztányi, Z. (2017). Degrons in cancer. Sci. Signal. *10*, eaak9982.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Nalepa, G., and Clapp, D.W. (2018). Fanconi anaemia and cancer: an intricate relationship. Nat. Rev. Cancer *18*, 168–185.

Oshiro, T.M., Perez, P.S., and Baranauskas, J.A. (2012). How many trees in a random forest? In Machine Learning and Data Mining in Pattern Recognition, P. Perner, ed. (Springer), pp. 154–168.

Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res. *47* (D1), D529–D541.

Pan, D., Kobayashi, A., Jiang, P., Ferrari de Andrade, L., Tay, R.E., Luoma, A.M., Tsoucas, D., Qiu, X., Lim, K., Rao, P., et al. (2018). A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. Science *359*, 770–775.

Qin, Q., Mei, S., Wu, Q., Sun, H., Li, L., Taing, L., Chen, S., Li, F., Liu, T., Zang, C., et al. (2016). ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC Bioinformatics *17*, 404.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44* (W1), W160-5.

Rauta, J., Alarmo, E.L., Kauraniemi, P., Karhu, R., Kuukasjärvi, T., and Kallioniemi, A. (2006). The serine-threonine protein phosphatase PPM1D is frequently activated through amplification in aggressive primary breast tumours. Breast Cancer Res. Treat. *95*, 257–263.

Reyes-Turcu, F.E., Ventii, K.H., and Wilkinson, K.D. (2009). Regulation and cellular roles of ubiquitin-specific deubiquitinating enzymes. Annu. Rev. Biochem. *78*, 363–397.

Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., Sims, J.S., Hodi, F.S., Martín-Algarra, S., Mandal, R., Sharfman, W.H., et al. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell *171*, 934–949.e16.

Robson, M., Im, S.A., Senkus, E., Xu, B., Domchek, S.M., Masuda, N., Delaloge, S., Li, W., Tung, N., Armstrong, A., et al. (2017). Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. N. Engl. J. Med. *377*, 523–533.

Ronau, J.A., Beckmann, J.F., and Hochstrasser, M. (2016). Substrate specificity of the ubiquitin and Ubl proteases. Cell Res. *26*, 441–456.

Rousseeuw, P.J., and van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics *41*, 212–223.

Sakamoto, K.M., Kim, K.B., Kumagai, A., Mercurio, F., Crews, C.M., and Deshaies, R.J. (2001). Protacs: chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. Proc. Natl. Acad. Sci. USA *98*, 8554–8559.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al.; Cancer Genome Atlas Research Network (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell *173*, 321–337.e10.

Scudellari, M. (2019). Protein-slaying drugs could be the next blockbuster therapies. Nature *567*, 298–300.

Shibata, T., Ohta, T., Tong, K.I., Kokubu, A., Odogawa, R., Tsuta, K., Asamura, H., Yamamoto, M., and Hirohashi, S. (2008). Cancer related mutations in NRF2 impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. Proc. Natl. Acad. Sci. USA *105*, 13568–13573.

Shreeram, S., Demidov, O.N., Hee, W.K., Yamaguchi, H., Onishi, N., Kek, C., Timofeev, O.N., Dudgeon, C., Fornace, A.J., Anderson, C.W., et al. (2006). Wip1 phosphatase modulates ATM-dependent signaling pathways. Mol. Cell *23*, 757–764.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv, arXiv:14091556 https://arxiv.org/abs/1409.1556.

Singh, A.A., Schuurman, K., Nevedomskaya, E., Stelloo, S., Linder, S., Droog, M., Kim, Y., Sanders, J., van der Poel, H., Bergman, A.M., et al. (2018). Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. Life Sci. Alliance *2*, e201800115.

Skoulidis, F., Byers, L.A., Diao, L., Papadimitrakopoulou, V.A., Tong, P., Izzo, J., Behrens, C., Kadara, H., Parra, E.R., Canales, J.R., et al. (2015). Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. Cancer Discov *5*, 860–877.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705.

Spranger, S., Bao, R., and Gajewski, T.F. (2015). Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity. Nature *523*, 231–235.

Staub, O., Gautschi, I., Ishikawa, T., Breitschopf, K., Ciechanover, A., Schild, L., and Rotin, D. (1997). Regulation of stability and function of the epithelial Na+ channel (ENaC) by ubiquitination. EMBO J. *16*, 6325–6336.

Stewart, M.D., Ritterhoff, T., Klevit, R.E., and Brzovic, P.S. (2016). E2 enzymes: more than just middle men. Cell Res. *26*, 423–440.

Strohmaier, H., Spruck, C.H., Kaiser, P., Won, K.A., Sangfelt, O., and Reed, S.I. (2001). Human F-box protein hCdc4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line. Nature *413*, 316–322.

Tanimoto, K., Makino, Y., Pereira, T., and Poellinger, L. (2000). Mechanism of regulation of the hypoxia-inducible factor-1 alpha by the von Hippel-Lindau tumor suppressor protein. EMBO J. *19*, 4298–4309.

Theodorou, V., Stark, R., Menon, S., and Carroll, J.S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res. *23*, 12–22.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al.; Cancer Genome Atlas Research Network (2018). The Immune Landscape of Cancer. Immunity *48*, 812–830.e14.

Timms, R.T., Zhang, Z., Rhee, D.Y., Harper, J.W., Koren, I., and Elledge, S.J. (2019). A glycine-specific N-degron pathway mediates the quality control of protein *N*-myristoylation. Science *365*, eaaw4912.

Tokheim, C., and Karchin, R. (2019). CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. Cell Syst. *9*, 9–23.e8.

Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proc. Natl. Acad. Sci. USA *113*, 14330–14335.

Torkamani, A., and Schork, N.J. (2008). Prediction of cancer driver mutations in protein kinases. Cancer Res. *68*, 1675–1682.

van der Lee, R., Lang, B., Kruse, K., Gsponer, J., Sánchez de Groot, N., Huynen, M.A., Matouschek, A., Fuxreiter, M., and Babu, M.M. (2014). Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell Rep. *8*, 1832–1844.

Vitari, A.C., Leong, K.G., Newton, K., Yee, C., O'Rourke, K., Liu, J., Phu, L., Vij, R., Ferrando, R., Couto, S.S., et al. (2011). COP1 is a tumour suppressor that causes degradation of ETS transcription factors. Nature *474*, 403–406.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Watson, P.A., Arora, V.K., and Sawyers, C.L. (2015). Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. Nat. Rev. Cancer *15*, 701–711.

Welcker, M., and Clurman, B.E. (2008). FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. Nat. Rev. Cancer *8*, 83–93.

Wellenstein, M.D., and de Visser, K.E. (2018). Cancer-Cell-Intrinsic Mechanisms Shaping the Tumor Immune Landscape. Immunity *48*, 399–416.

Winter, G.E., Buckley, D.L., Paulk, J., Roberts, J.M., Souza, A., Dhe-Paganon, S., and Bradner, J.E. (2015). DRUG DEVELOPMENT. Phthalimide conjugation as a strategy for in vivo target protein degradation. Science *348*, 1376–1381.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108–1113.

Yao, Z., Yaeger, R., Rodrik-Outmezguine, V.S., Tao, A., Torres, N.M., Chang, M.T., Drosten, M., Zhao, H., Cecchi, F., Hembrough, T., et al. (2017). Tumours with class 3 BRAF mutants are sensitive to the inhibition of activated RAS. Nature *548*, 234–238.

Yen, H.C., Xu, Q., Chou, D.M., Zhao, Z., and Elledge, S.J. (2008). Global protein stability profiling in mammalian cells. Science *322*, 918–923.

Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713.

Zhang, D., Zaugg, K., Mak, T.W., and Elledge, S.J. (2006). A role for the deubiquitinating enzyme USP28 in control of the DNA-damage response. Cell *126*, 529–542.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.Y., Petyuk, V.A., Chen, L., Ray, D., et al.; CPTAC Investigators (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell *166*, 755–765.

Zheng, N., and Shabek, N. (2017). Ubiquitin Ligases: Structure, Function, and Regulation. Annu Rev Biochem *86*, 129–157.

Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer, C.A., and Liu, X.S. (2019). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. *47* (D1), D729–D735.

Zhou, X., Edmonson, M.N., Wilkinson, M.R., Patel, A., Wu, G., Liu, Y., Li, Y., Zhang, Z., Rusch, M.C., Parker, M., et al. (2016). Exploring genomic alteration in pediatric cancer using ProteinPaint. Nat. Genet. *48*, 4–6.

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal anti-mouse/human GATA3	Cell Signaling Technology	Cat# 5852; RRID:AB_10835690
Rabbit monoclonal anti-human PPM1D/WIP1	Abcam	Cat# ab31270; RRID:AB_10585435
Rabbit monoclonal anti-human Phospho-p53 (Ser15)	Cell Signaling Technology	Cat# 9284; RRID:AB_331464
Rabbit monoclonal anti-human p53	Cell Signaling Technology	Cat# 9282; RRID:AB_331476
Mouse monoclonal anti-human Phospho-ATM (Ser1981)	Cell Signaling Technology	Cat# 4526; RRID:AB_2062663
Rabbit monoclonal anti-human/mouse ATM	Cell Signaling Technology	Cat# 2873; RRID:AB_2062659
Mouse monoclonal anti-human c-Myc	Santa Cruz Biotechnology	Cat# sc-40 AC; RRID:AB_2857941
Rabbit monoclonal anti-human/mouse CUL3	Cell Signaling Technology	Cat# 2759; RRID:AB_2086432
Rabbit monoclonal anti-human/mouse β-Actin	Cell Signaling Technology	Cat# 4970; RRID:AB_2223172
Rabbit monoclonal anti-human IgG XP Isotype Control	Cell Signaling Technology	Cat# 3900; RRID:AB_1550038
Mouse monoclonal anti DYKDDDDK Tag	Cell Signaling Technology	Cat# 8146; RRID:AB_10950495
Rabbit monoclonal anti-HA-tag	Cell Signaling Technology	Cat# 3724; RRID:AB_1549585
Goat anti-Mouse IgG Secondary Antibody, HRP	Thermo Fisher Scientific	Cat# 31430; RRID:AB_228307
Donkey anti-Rabbit IgG Secondary Antibody, HRP	Thermo Fisher Scientific	Cat# 31458; RRID:AB_228213
Rabbit monoclonal anti-human/mouse KIT	Cell Signaling Technology	Cat# 3074; RRID:AB_1147633
Rabbit monoclonal anti-human/mouse CDH1	Cell Signaling Technology	Cat# 13116; RRID:AB_2687616
Rabbit monoclonal anti-human FOXA1	Cell Signaling Technology	Cat# 53528; RRID:AB_2799438
Mouse monoclonal anti-human KRT18	Sigma Aldrich	Cat# WH0003875M1; RRID:AB_1842192
Rabbit monoclonal anti-human/mouse TP63	Abcam	Cat# ab124762; RRID:AB_10971840
Rabbit monoclonal anti-human/mouse JAG1	Cell Signaling Technology	Cat# 70109; RRID:AB_2799774
Mouse monoclonal anti-human KRT14	Santa Cruz	Cat# sc-53253; RRID:AB_2134820
Bacterial and virus strains		
XL10-Gold Ultracompetent Cells	Agilent	Cat#200314
Endura ElectroCompetent Cells	Lucigen	Cat#60242-2
Chemicals, peptides, and recombinant proteins		
PBS	GIBCO	Cat#14190250
DMEM, high glucose, pyruvate	GIBCO	Cat#11995065
Lonza BioWhittaker L-Glutamine (200mM)	Lonza	Cat#BW17605E
Fetal bovine serum	VWR	Cat#9706
Penicillin-Streptomycin	GIBCO	Cat#15140122
PolyJet *In Vitro* DNA Transfection Reagent	SignaGen Laboratories	Cat#SL100688
E-Gel Low Range Quantitative DNA Ladder	Invitrogen	Cat#NP0008
E-Gel EX Agarose Gels, 2%	Life Technologies	Cat#G402002
NuPAGE 3-8% Tris-Acetate Protein Gels, 1.5 mm, 10-well	Life Technologies	Cat#EA0378BOX
NuPAGE LDS Sample Buffer	Life Technologies	Cat#NP0008
Pierce ECL Western Blotting Substrate	Thermo Fisher Scientific	Cat#32106
Precision Plus Protein Dual Color Standards	Bio-Rad Laboratories	Cat#161-0394
X-tremeGENE HP DNA Transfection Reagent	Sigma-Aldrich	Cat#6366236001
Polybrene	Sigma-Aldrich	Cat#107689-10G
Puromycin dihydrochloride	Thermo Fisher Scientific	Cat#A1113803
BamHI-HF	New England Biolabs	Cat#R3136S
EcoRI-HF	New England Biolabs	Cat#R3101S
FastDigest Esp3I	Thermo Fisher Scientific	Cat#FD0454
Q5 DNA Polymerase	New England Biolabs	Cat#M0491L

(*Continued on next page*)

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Nuclease-Free Water	Ambion	Cat#AM9938
Pierce Homobifunctional Cross Linkers	Life Technologies	Cat#20593
2-Mercaptoethanol	Sigma Aldrich	Cat#M6250-10ML
Dynabeads Protein A	Thermo Fisher Scientific	Cat#10004D
Dynabeads Protein G	Thermo Fisher Scientific	Cat#10002D
EDTA	Sigma Aldrich	Cat#E8008-100ML
Protease/Phosphatase Inhibitor Cocktail (100X)	Cell Signaling Technology	Cat#5872S
Quick-Load 1 kb Plus DNA Ladder	New England Biolabs	Cat#N0469S
LB Broth	Mp Biomedicals	Cat#244610
L-Broth Agar Large Capsules	Mp Biomedicals	Cat#MP 113001236
RIPA buffer	Invitrogen	Cat#R0278
Pierce 16% Formaldehyde (w/v), Methanol-free	Life Technologies	Cat#28906
Opti-MEM I Reduced Serum Medium, no phenol red	Thermo Fisher Scientific	Cat#11058021
Cycloheximide powder	Cell Signaling Technology	Cat#2112
Critical commercial assays		
QIAprep Spin Miniprep Kit	QIAGEN	Cat#27106
RNeasy Plus Mini Kit	QIAGEN	Cat#74134
QIAquick PCR Purification Kit	QIAGEN	Cat#28104
QIAquick gel extraction kit	QIAGEN	Cat#28704
Gibson Assembly Master Mix	New England Biolabs	Cat#E2611L
iScript cDNA Synthesis Kit	Bio-Rad Laboratories	Cat#1708891
SsoAdvanced Univ SYBR Grn Suprmx	Bio-Rad Laboratories	Cat#1725272
Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific	Cat#Q32854
Qubit RNA HS Assay Kit	Thermo Fisher Scientific	Cat#Q32855
GenElute HP Plasmid Maxiprep Kit	Sigma-Aldrich	Cat#NA0410-1KT
Ampure xp	Beckman Coulter	Cat#A63881
BCA Assay Kit	Thermo Fisher Scientific	Cat#23225
SMARTer® ThruPLEX® DNA-Seq Kit	Takara Bio	Cat#R400675
Experimental models: cell lines		
Human: HEK293FT	Thermo Fisher Scientific	Cat#R70007
Human: MDA-MB-231	American Type Culture Collection	Cat#ATCC-HTB-26
Human: CAL27	Ravi Uppaluri lab	N/A
Human: CAL33	Ravi Uppaluri lab	N/A
Deposited data		
GATA3 ChIP-Seq	This paper	GSE162003
Original gel images	This paper	https://doi.org/10.17632/kgfzbpv2w4.1
Oligonucleotides		
Primers for PCR, see Table S5	Invitrogen	N/A
Recombinant DNA		
hWIP1-FLAG	Addgene	RRID:Addgene_28105
pHAGE-GATA3	Addgene	RRID:Addgene_116747
lentiCRISPR v2 puro	Addgene	RRID:Addgene_98290
pMD2.G	Addgene	RRID:Addgene_12259
psPAX2	Addgene	RRID:Addgene_12260
pHAGE-CMV-DsRed-IRES-GFP	Koren et al., 2018	N/A
pHAGE-SFFV-GFP-IRES-DsRed	Timms et al., 2019	N/A
pLenti-EF1a-PGK-Puro	Kai Wucherpfennig lab	N/A
pLenti-EF1a-GATA3-WT	This paper	N/A

(*Continued on next page*)

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pLenti-EF1a-GATA3-A442M	This paper	N/A
pLenti-EF1a-GATA3-G444E	This paper	N/A
pLenti-EF1a-GATA3-H400	This paper	N/A
pLenti-EF1a-GATA3-WT-Fg	This paper	N/A
pLenti-EF1a-GATA3-A442M-Fg	This paper	N/A
pLenti-EF1a-GATA3-G444E-Fg	This paper	N/A
pLenti-EF1a-GATA3-H400-Fg	This paper	N/A
pLenti-EF1a-PPM1D-WT	This paper	N/A
pLenti-EF1a-PPM1D-V604Q	This paper	N/A
pLenti-EF1a-PPM1D-C605W	This paper	N/A
pLenti-EF1a-PPM1D-L450	This paper	N/A
pLenti-EF1a-PPM1D-WT-HA	This paper	N/A
pLenti-EF1a-PPM1D-V604Q-HA	This paper	N/A
pLenti-EF1a-PPM1D-C605W-HA	This paper	N/A
pLenti-EF1a-PPM1D-L450-HA	This paper	N/A
Software and algorithms		
GraphPad Prism 7	GraphPad Software	https://www.graphpad.com
DeepDegron	This Paper	https://github.com/ctokheim/deepDegron
Transcription factor inference	This Paper	https://github.com/ctokheim/tf_association
Other		
Corning Filter System (0.45um)	Corning Life Sciences	Cat#431096
milliTUBE 1 ml AFA Fiber	Covaris Inc.	Cat#520130
NITROCEL MEMB 0.45um	Bio-Rad Laboratories	Cat#1620115
Multiplate 96-Well PCR Plates	Bio-Rad Laboratories	Cat#MLL9601
QUBIT ASSAY TUBES SET	Life Technologies	Cat#Q32856
Microseal 'B' Adhesive Seals	Bio-Rad Laboratories	Cat#MSB-1001

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents (including code) should be directed to and will be fulfilled by the Lead Contact, X. Shirley Liu (xsliu@ds.dfci.harvard.edu).

Materials availability

Materials associated with the paper are available upon request to Lead Contact, X. Shirley Liu (xsliu@ds.dfci.harvard.edu).

Data and code availability

Original data have been deposited to Mendeley Data: https://dx.doi.org/10.17632/kgfzbpv2w4.1. The accession number for the GATA3 ChIP-seq reported in this paper is GEO: GSE162003. The DeepDegron code is available on github: https://github.com/ctokheim/deepDegron. The code for associating UPS genes with putative transcription factor substrates is also available on github: https://github.com/ctokheim/tf_association. Jupyter notebooks for data analysis are stored on github (https://github.com/ctokheim/Tokheim_Mol_Cell_2021).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Lines

Human embryonic kidney 293FT cell line (HEK293FT) was obtained from Thermo Fisher Scientific. HEK293T cells were grown in DMEM supplemented with 10% fetal bovine serum, 2% penicillin/streptomycin, 1% L-glutamine and 100 mM sodium pyruvate according to standard protocol and maintained at 37°C with 5% CO2.

Human breast cancer MDA-MB-231 cell line was obtained from American Type Culture Collection (ATCC). MDA-MB-231 cells were grown in DMEM supplemented with 10% fetal bovine serum, 2% penicillin/streptomycin, 1% L-glutamine and 100 mM sodium pyruvate according to standard protocol and maintained at 37°C with 5% CO2.

Human oral squamous cell carcinoma cell lines CAL27 and CAL33 were kindly provided by Ravi Uppaluri laboratory. Cells were cultured in in DMEM supplemented with 10% fetal bovine serum, 2% penicillin/streptomycin, 1% L-glutamine and 100 mM sodium pyruvate according to standard protocols and maintained at 37°C with 5% CO2. Cell lines were stored in liquid nitrogen at early passages and were cultured within 20 doublings.

METHOD DETAILS

Mutation dataset

We used somatic mutations from 33 cancer types called by the MC3 group in The Cancer Genome Atlas (TCGA) (https://gdc.cancer.gov/about-data/publications/pancanatlas; v0.2.8), which were formed by the consensus of multiple mutation calling algorithms in a unified pipeline (Ellrott et al., 2018). We then filtered the dataset according to quality control metrics for both mutations and tumor samples. Specifically, the following filters were applied: 1) mutations should have passed all QC metrics by the MC3 group (i.e., "PASS" in the "filter" column), except for the allowance of whole genome amplified samples in ovarian cancer and AML where the majority of tumor samples used a whole genome amplification step; 2) tumor samples which failed pathology review were excluded; 3) for statistical power reasons, we excluded hypermutated tumors (Lawrence et al., 2014; Tokheim et al., 2016), defined as having a greater number of mutations than 1.5x the interquartile range above the 3rd quartile (Tukey's condition) for the respective tumor's cancer type. Because this procedure also excludes outliers for cancer types with overall low tumor mutation burden, we also required the tumor sample to have greater than 1,000 mutations to be excluded. These filters resulted in 1,457,702 mutations for final analysis.

Gene and Protein Expression Data

Gene expression estimates from RNA-seq were quantified from the RSEM v2 pipeline (Li and Dewey, 2011) of TCGA. The data was downloaded from the Genomic Data Commons website (https://api.gdc.cancer.gov/data/3586c0da-64d0-4b74-a449-5ff4d9136611). RNA expression values were log normalized (i.e., log2(RSEM+1)) and centered with median value of zero per gene. Normalized protein expression from Reverse Phase Protein Arrays (RPPA) was also download from the Genomic Data Commons website (http://api.gdc.cancer.gov/data/fcbb373e-28d4-4818-92f3-601ede3da5e1).

Ubiquitin-Proteasome System (UPS) pathway genes

We curated a set of UPS genes from two previous publications (Ge et al., 2018; Mészáros et al., 2017), which included E1 activating enzymes, E2 conjugating enzymes, E3 ubiquitin ligases and deubiquitinating enzymes. We used only those annotated with literature support from Ge et al. (2018) and the E3 ubiquitin ligases reported by Mészáros et al. (2017). Additionally, we removed a gene, *CDH1*, that was erroneously labeled as involved with ubiquitination due to conflicting symbols with a known UPS gene (*FZR1*, known at the protein-level as Cdh1). This resulted in a set of 775 genes for further analysis (Table S1).

Driver gene analysis

To ascertain which genes in the UPS pathway might promote cancer development and progression, we analyzed whether genes in the UPS were significantly mutated in human cancers by the method 20/20+ (Tokheim et al., 2016). 20/20+ was ran using default parameters except for usage of 100,000 simulations, as described previously (https://github.com/KarchinLab/2020plus; v1.2.0) (Bailey et al., 2018), on each of the 33 cancer types individually and all cancer types aggregated together (known as a "pan-cancer" analysis). Briefly, 20/20+ is a random forest method that scores the propensity of a gene to be an oncogene, a tumor suppressor gene or, in general, a cancer driver gene (scores range from 0 to 1). P values for each score are then computed based on a Monte Carlo simulation procedure that generates a background distribution of mutations accounting for nucleotide sequence context (probabilistic2020 python package, v1.2.0). Here, to increase statistical power to identify lowly mutated driver genes in the ubiquitin pathway, we performed a restricted hypothesis test on only the 775 UPS genes annotated above. Genes were deemed significant at a false discovery rate of 0.05 (Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995)) and those with high effect size (score > 0.5 out of 1.0).

Lollipop diagram visualization

Mutations on protein sequence were visualized using the ProteinPaint tool (https://pecan.stjude.cloud/proteinpaint; Zhou et al., 2016). Mutations were submitted according to their genomic coordinates and mutations that do not match the default reference transcript used by ProteinPaint are not shown. Height corresponds to the number of mutations while the x axis represents the codon position along the protein sequence. Protein domains are shown as colored boxes along the protein sequence.

Expression and essentiality analysis of putative driver genes

The putative UPS driver genes were characterized by their tissue expression from GTEx (Battle et al., 2017) and cancer cell line essentiality in CRISPR screens from DepMap (~500 cell lines) (Meyers et al., 2017). The 63 driver genes were compared to both other UPS genes not found as drivers and all other non-UPS genes using a Mann-Whitney U test. Version 7 of TPM expression values from

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

GTEx were used (https://gtexportal.org/home/). Additionally, CERES scores (Meyers et al., 2017), which quantify how essential a gene is in CRISPR screens, were obtained from the 2019 Q1 data release of DepMap. Negative CERES scores indicate a gene is essential in a particular cancer cell line.

Recent evidence suggests context-specific roles for UPS driver genes (Haigis et al., 2019), such as PARP inhibitors being selectively effective in BRCA1 mutant tumors in traditionally BRCA-associated cancer types (breast, ovary, prostate and pancreas) (Jonsson et al., 2019). We therefore examined the specificity of UPS driver genes in both cell-lineage and genetic mutation contexts. According to the Genotype-Tissue Expression (GTEx) data (Battle et al., 2017), we noticed that putative UPS driver genes were expressed in most normal tissues, and more broadly expressed than other UPS genes or non-UPS genes ($p < 0.05$; Figure 2H). However, from CRISPR screens across ~500 cancer cell lines from the DepMap (Meyers et al., 2017), UPS driver genes showed significantly higher variability in the gene dependency scores (CERES scores) across cell lines compared to other genes ($p < 0.05$), suggesting substantial cell type-specific essentiality (Figure 2I). One possible explanation for the variable gene essentiality despite widespread expression might, in part, arise from cells uniquely expressing or modifying certain important substrates of the UPS (Figure S1F). This is consistent with previous literature that substrate recognition by E3 ubiquitin-ligases, such as c-CBL and β-TrCP, can depend on signaling pathways which mark degrons by phosphorylation (Zheng and Shabek, 2017).

Mutational co-occurrence

We analyzed whether non-silent mutations in putative driver genes in the ubiquitin pathway would tend to co-occur in the same tumor samples with mutations in 299 driver genes identified previously by the TCGA (Bailey et al., 2018). We used the Mantel-Haenszel test to identify pairs of genes with an odds ratio significantly different from 1.0 at an FDR threshold of 0.25. To control for the confounding effect of tumor mutation burden, we adjusted for high (> 500 mutations; half the hypermutator threshold) and low (≤500 mutations) tumor mutation burden samples in our analysis. In the pan-cancer analysis, we also adjusted for the cancer type of the tumor labeled by TCGA.

Next, we sought to examine whether mutations in UPS driver genes would contextually co-occur or be mutually exclusive with mutations in other driver genes in the same tumor. This revealed 13 of the UPS driver genes with an enriched co-mutational pattern with other driver genes previously identified by TCGA (Bailey et al., 2018; Figure S1H; Table S1). For example, we found KEAP1-KRAS-STK11 to be co-mutated in lung adenocarcinoma (LUAD) tumors, which have been reported to form a biologically distinct subtype of *KRAS* mutant LUAD (Skoulidis et al., 2015). Previously, mutations in *STK11* have been implicated in a T cell exclusion phenotype for these tumors and ultimately responsible for resistance to immune checkpoint inhibition (Hellmann et al., 2018). Instead, we found that mutation of the E3 ligase *KEAP1*, regardless of *STK11* status, correlates with lower immune infiltration in TCGA (Figure S1I), suggesting that *KEAP1* has additional immunomodulatory roles. The interaction with other driver genes might be partially related to UPS driver genes being preferentially situated centrally in a protein-protein interaction network (Figure S1J), a property previously noted for other driver genes (Davoli et al., 2013). In summary, the 63 putative UPS driver genes we identified showed context-specificity with regard to both cell type and genetic mutations.

Global Protein Stability (GPS) Assays

GPS experiments were performed as described in Koren et al., 2018 and Timms et al., 2019. Individual sequences encoding example 23-mer peptides were PCR-amplified from either the N-terminome (Timms et al., 2019) or C-terminome (Koren et al., 2018) oligonucleotide libraries and cloned into lentiviral GPS expression vectors. Lentivirus was packaged through the transfection of HEK293T cells (ATCC® CRL-3216) grown in Dulbecco's Modified Eagle's Medium (DMEM) (Life Technologies) supplemented with 10% fetal bovine serum (HyClone) and penicillin/streptomycin (Thermo Fisher Scientific). HEK293T at around 70% confluency were transfected with the GPS vector plus four packaging plasmids (encoding Gag-Pol, Rev, Tat and VSV-G) using PolyJet *In Vitro* DNA Transfection Reagent (SignaGen Laboratories) as recommended by the manufacturer. The media was changed after 24 hours, and the viral supernatant collected a further 24 hours later. Following centrifugation (800 x *g*, 5 min) to remove cellular debris, the viral supernatant was applied to target HEK293T cells. After a further 48 hours, stability measurements were made by flow cytometry using a BD LSRII instrument (Becton Dickinson); at least 10,000 $DsRed^+$ cells were collected in each case. The resulting data were analyzed using FlowJo software.

deepDegron

deepDegron is a feed forward neural network trained on the Global Protein Stability (GPS) assay (Yen et al., 2008), which at proteome-scale measures the conferred stability or instability of peptides when attached to GFP in HEK293T cells. Importantly, the GPS assay also contains an internal control DsRed (located on the same transcript) which does not contain an attached peptide. FACS is then used to sort cells based on the red (DsRed) to green (GFP) ratio into separate bins and subsequently barcodes are sequenced to quantify the representation of peptides in each bin.

Dataset

Data from the GPS assay related to N-terminal (Timms et al., 2019) and C-terminal (Koren et al., 2018) peptides were collected from their respective publications and analyzed separately. In the case of the C-terminal data, we analyzed the full 23-mer peptide screen. While for the N-terminal data, we only analyzed peptides with an initiator methionine (24-mer), but since the methionine was always

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

the same at the first position, we did not include the methionine in our model (23-mer). To establish a classification task for the deepDegron model, we binarized each peptide into two classes based on the mode of the read count distribution across bins in the GPS assay. If a peptide's modal bin was in the lower half of the red to green ratio it was assigned as instable (class = 1) and the remaining were assigned as stable (class = 0). If a gene had multiple peptides in the GPS assay, we only used the first occurrence for further analysis.

Neural network

deepDegron, a two hidden-layer feed forward neural network, was trained using the Keras python package with the tensorflow backend (https://github.com/ctokheim/deepDegron). ReLu activation functions were used for hidden layers and the sigmoid function was used for the final output node, which generally performs well for neural network models (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Training was performed using the Adam optimizer using the default learning rate, given it has previously been suggested that Adam gives superior results compared to other optimizers (Kingma and Ba, 2014).

Training, validation and test sets

We randomly separated out 30% of the sequences for purpose of evaluation as a test set. For the remaining 70% of the data, we randomly split again 70% (49% overall) of that data into a training set and 30% (21% overall) as a validation set for hyperparameter selection.

Hyperparameters

Like most machine learning algorithms, neural networks benefit from fine tuning hyperparameters of the model. Here, we utilized grid search over hyperparameters for both feature engineering and neural network parameters. For feature engineering, we considered position-specific one-hot encoding of various lengths of the peptide from the terminal-ends (l = 6, 12, 18 or 23) with the remaining portion of the peptide sequence encoded only in terms of the count of each amino acid type (i.e., position agnostic). This was intended to limit the number of learned parameters of the model, if certain regions of the peptide were more important. Additionally, given previous evidence of the importance of dimer amino acid motifs at the very end of protein sequence (Koren et al., 2018), we also allowed for the one-hot encoding of di-amino acid motifs (di = True or False). For neural network parameters, we considered different number of nodes for each layer (n = 8 or 16). Additionally, we considered various levels of dropout regularization (d = 0, 0.25 or 0.5) for connections between the input and 1st hidden layer since it contained the greatest number of parameters in the model. Lastly, we also considered the number of epochs used for training (e = 20, 40 or 60).

Evaluation

The optimal hyperparameters were selected according to the highest area under the Receiver Operating Characteristic curve (auROC) on the validation dataset. The C-terminal deepDegron hyperparameters that were selected are: n = 8, d = 0.0, e = 20, l = 6 and di = True. While the N-terminal deepDegron hyperparameters that were selected are: n = 16, d = 0.5, e = 20, l = 6 and di = True.

The deepDegron models were then compared to a Random Forest model (scikit-learn with 1,000 trees as performance usually only increases with this parameter (Oshiro et al., 2012)), which empirically performs well on many machine learning tasks (Caruana and Niculescu-Mizil, 2006), and previously proposed rule-based alternatives (Koren et al., 2018), such as the number of acidic residues (D, E), number of bulky hydrophobic residues (F, W, Y) or the number of top 100 motifs. Evaluations for all models were performed on the held-out test set and compared using the auROC metric.

Degron Potential Calculation

We calculated a degron potential score to correct for protein stability likely reflecting both amino acid order effects (e.g., a degron motif exists) versus general amino acid properties. To do this, in addition to the model outlined in the deepDegron section (Methods), we trained a second model (''bag of amino acids'') containing the same hyperparameters that only has the count of each amino acid in the peptide sequence as features (20 features). We then calculated a degron potential score as the difference in prediction between the position specific model and the ''bag of amino acids'' model.

Motif Analysis

Motif analysis was conducted by measuring enrichment for sequence motifs among top degron potential scored peptides from deepDegron. First, we ranked all peptide sequences by degron potential score from high to low likelihood of containing a degron. Second, we performed area auROC analyses to calculate at which point the top degron potential sequences would cease to have meaningful enrichment. To determine this cutoff, we computed at various cutoffs a delta auROC score, which we defined as the difference in auROC between the two deepDegron models (position specific versus ''bag of amino acid'' model) tested on sequences where the top-ranking X and bottom-ranking X sequences were removed. The delta auROC was calculated and plotted over various cutoffs of X ranging from 0 to 8000 with an increment of 20. We then used the elbow-method (Goutte et al., 1999) based on the point of maximal curvature to delineate the transition (X*) from a performance gap existing to nearly equivalent performance. Since curvature is only well defined for continuous functions, we used an algorithmic approximation from the kneed python package with default parameters (v0.4.1; https://github.com/arvkevi/kneed; V. Satopaa et al., 2011, IEEE, conference). Third, we calculated

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

the background probability p that a particular peptide would contain particular motifs of length 2 (with or without gaps) and 3. We only considered motifs within the proximal 6 amino acids to either the N terminus or the C terminus, as our performance evaluation above suggested most gains were in this region. Additionally, since the number of possible motifs grows exponentially with motif length, we only considered gapped and position-specific motifs for length 2 motifs. Fourth, using a binomial model with background probability p, we measured whether motifs had significantly more motifs c than expected for the top X* sequences. Fifth, we corrected for multiple hypotheses by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and declared significant motifs at false discovery rate threshold of 0.05. Lastly, to identify potentially extended motifs outside those identified by our analysis, we generated sequence logo visualizations by compiling all the top sequences that contained the motif and inputting these sequences into the WebLogo software (Crooks et al., 2004).

Monte Carlo simulations

To establish a background distribution of mutations, we performed Monte Carlo simulations as described previously (Tokheim et al., 2016). Briefly, for single nucleotide variants, we moved mutations uniformly at random within the same gene but matched the same nucleotide context as the observed mutation (C*pG, CpG*, TpC*, G*pA, A, C, G, T). Indels were moved within the same gene without regard for the flanking sequence, as mutational signatures for indels are less known than for single nucleotide variants (Alexandrov et al., 2013). Based on the simulated mutations, we then recategorized the effect of the variant. For example, a mutation may have originally been a nonsense mutation but when moved to a new position it may be a missense mutation in a known degron site. Test statistics for degron enrichment were then computed and this simulation procedure was repeated 10,000 times. P values were computed based on the resulting empirical distribution, i.e., the fraction of simulations with test statistics that were as or more extreme then the observed value.

Mutation enrichment at known degrons, ubiquitination sites or phosphodegrons

Known degron sites were collected from a recent literature review (Mészáros et al., 2017), while ubiquitination sites and phosphodegrons (phosphorylation sites annotated as involved with ''protein degradation'') were obtained from the PhophoSitePlus database (Hornbeck et al., 2015). For each cancer type, we analyzed whether the number of missense mutations found in annotated sites of a gene were higher than expected based on an empirical background distribution established through Monte Carlo simulations (see section above). In the case of the phosphodegron analysis, we also considered the flanking 3 amino acids on either side of the phosphorylation site. Genes were deemed significant at a False Discovery Rate (FDR) of 0.1. Based on manual review of the literature, one significant result (BRAF, ubiquitination site enrichment due to K601E mutations) was excluded from further analysis due to previously literature suggesting a distinct mechanism of action (Yao et al., 2017).

Calculation of degron impact bias

Because known degron sites are limited, we also assessed for genes containing a significant enrichment of mutations predicted to lead to degron loss by deepDegron. First, we computed the change in degron potential (delta degron potential) between the mutated and reference protein sequence for each mutation in the 33 cancer types available from TCGA. Second, we computed a gene-wise test statistic as the sum of delta degron potential for all mutations within a gene. Scores considerably less than zero indicate degron loss. Third, to evaluate the statistical significance, we performed Monte Carlo simulations (described above) to compute a p value corresponding to seeing a score equal to or lower than the observed value (i.e., degron loss). Like for the known degron case, significant enrichment was defined at an FDR of 0.1 and, additionally, required the delta degron potential to indicate a preferential loss of a degron (delta degron potential below -1).

Selection of deepDegron motifs for experimental validation

To validate deepDegron predictions for degrons, we selected 10 novel motifs for experimental validation. For this we used the GPS assay to compare the protein stability of GFP fused to either the wild-type peptide, or one containing point mutations in the predicted degron motifs. Since some motifs partially overlapped, we prioritized motifs based on statistical significance ($q < 0.05$) and independence from other tested motifs. Motifs were equally divided between predicted C-terminal (-LxRxx, -MxxxV, -CxxR, -VS, and -LxxAx; x = any amino acid) and N-terminal degrons (GxL-, xPL-, RxR-, GxxxA- and RxxP-). To avoid introducing generally stabilizing amino acids that are independent of a degron motif, point mutations were selected based on maximally decreasing the degron potential of the sequence while maintaining the score of the ''bag of amino acid'' model within a range of 0.1 from the original sequence. The selected double mutants for each motif are listed in Table S4G. The same selection procedure for point mutants was carried out for the degron motifs of *GATA3* and *PPM1D*.

Generation of lentiviral expression vectors

Plasmids (hWIP1-FLAG, pHAGE-GATA3) were obtained from Addgene. Overexpression vector pLenti-EF1a-PGK-Puro was kindly provided by Kai Wucherpfennig laboratory. Different forms of wild-type or mutated GATA3/PPM1D sequence were amplified by PCR and subcloned into a pLenti-EF1a-PGK-Puro empty vector via Gibson assembly to generate different overexpression vectors (GATA3 and PPM1D). Next, small amount (1 μl) of the Gibson assembly reactions was transformed into competent cells. Competent cells were incubated on ice for 30 minutes, then subjected to heat shock in a water bath or electroporated by a Gene Pulser Xcell

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Electroporator (Bio-Rad Laboratories) and returned to ice for 2 minutes. LB media (1 ml) was added to the competent cells and the cells were allowed to recover at 37°C for 60 minutes on a shaker; subsequently 30 μL of the mixture (LB+ competent cells) was plated on LB-agar plates containing 100 μg/ml ampicillin and incubated at 37°C overnight (12-16 hours).

Generation of CRISPR/Cas9 Knock-out cells

Construction of lenti-CRISPR/Cas9 vectors targeting AAVS1 (Control) or CUL3 was performed following the protocol associated with the backbone vector lentiCRISPR_V2 (Addgene). The sgRNA sequences used are listed in the Key resources table. CAL27 cells were infected with lentivirus expressing sgRNAs targeting AAVS1 or CUL3. After puromycin selection for 3 days, cells were expanded for at least 7 days and collected. CUL3 knockout was verified by western blot analysis.

Viral library production

The pLenti-EF1a-GATA3/pLenti-EF1a-PPM1D expression constructs and the empty pLenti-EF1a-PGK-Puro vector were transfected into the 293FT cell line at 80%–90% confluency in 10 cm tissue culture plates. Viral supernatant was collected at 48 and 72 hours post-transfection, filtered via a 0.45 mm filtration unit (Corning). The supernatant was subsequently aliquoted and stored in −80°C freezer until use.

Viral transduction of cells

Cells were cultured in complete growth medium according to standard protocols. For viral transduction, a total of 3×10^5 cells were transduced with lentivirus containing gene cDNA construct described above at a high level of multiplicity of infection (MOI) in 10 cm tissue culture plates. After puromycin selection for 3 days, surviving cells were allowed to grow for another 7 days to overexpress specific genes. Immunoblotting and PCR were performed to confirm the expression of specific genes.

Co-immunoprecipitation of CUL3 protein

Human oral squamous cell carcinoma CAL27 cells were lysed in Tris buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.5% NP-40, 5% glycerol, with protease and phosphatase inhibitors) for 30 min with gentle rocking at 4°C. Cell lysate was spun down by a centrifuge in cold room at 12,000 rpm for 10 minutes and then supernatant was collected and incubated with CUL3 antibody coupled to Protein A/G agarose beads (Pierce Biotechnology) at 4°C overnight (12 hours). Beads were washed extensively in Tris lysis buffer containing 0.5 M NaCl and then eluted in LDS-sample buffer (Invitrogen) containing 1% 2-mercaptoethanol. Cell lysate was supplemented with 4X SDS loading buffer (0.2 M Tris-HCl, 0.4 M DTT, 8.0% SDS, 6 mM Bromophenol blue, 4.3 M Glycerol) and heated at 95°C for 15 minutes before western blot analysis.

Western Blot of protein expression in human cells

Pellets from 5×10^6 cells were collected and digested by 500 μL RIPA Buffer (Invitrogen). Samples were incubated on ice for at least 15 minutes and centrifuged at 12,000 rpm for 10 minutes at 4°C, then subjected to BCA analysis (Thermo scientific). Approximately 40-60 μg of total protein from each sample was loaded for western blot analysis.

Measurement of protein half-life

Cancer cells (1×10^6) were seeded onto 100mm Petri dishes in complete growth medium according to standard protocols and incubated in a CO2 incubator. After 24 hours incubation, remove the medium and add complete medium with 100 μg/ml cycloheximide (CHX; dissolved in DMSO) into each dish. Cells were exposed to cycloheximide for 0, 4, 8 or 12 hours to inhibit the protein synthesis according to the experimental design. Then, cell lysates were collected at different time points and MYC protein levels were examined by western blot using an anti-MYC antibody. Western bands of MYC and β-ACTIN were quantified in triplicates using ImageJ software.

Real-time reverse transcription-PCR

RNA was extracted using RNeasy Plus Mini Kit (QIAGEN) from HEK293FT and MDA-MB-231 cells. Then, RNA was reverse transcribed into cDNA using iScriptTM cDNA Synthesis Kit (Bio-Rad Laboratories). Approximately 50 ng cDNA from each sample was mixed with gene-specific primers (Table S5) and SsoAdvancedTM universal SYBR® Green supermix (Bio-Rad Laboratories) following the manufacturer's protocol. Reactions were performed on a CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories).

ChIP sequencing of GATA3

MDA-MB-231 cells were plated in 15 cm tissue culture plates and cultured for 3 days. For GATA3 ChIP-sequencing, approximately 1×10^7 cells per condition were harvested and crosslinked by a two-step fixation, including 2 mM disuccinimidyl glutarate (DSG, Life Technologies) treatment for 45 minutes and followed by 10 minutes fixation using 1% methanol-free formaldehyde at room temperature (Eeckhoute et al., 2007; Singh et al., 2018). Cells were lysed in 1% SDS lysis buffer and sheared to 200-700 bp in size using the Covaris E220 ultrasonicator (PIP 140, DF 5%, CPB 200). Approximately 50 mg of sheared chromatin per condition were diluted and then incubated overnight with 5 ug GATA3 antibody (14074, Cell Signaling). Precipitates were then washed with following buffers:

RIPA 0 buffer (0.1% SDS, 10 mM Tris-HCl pH 7.4, 1% Triton X-100, 1 mM EDTA, 0.1% sodium deoxycholate), RIPA 0.3 buffer (0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate, 10 mM Tris-HCl pH 7.4, 1 mM EDTA, 0.3 M NaCl) and LiCl buffer (250 mM LiCl, 1 mM EDTA, 5% NP-40, 0.5% sodium deoxycholate, 10 mM Tris-HCl). DNA sequencing libraries were prepared using the Smarter Thruplex DNaseq kit (Takara Bio Inc.) according to the manufacturer's protocol. Libraries were sequenced on an Illumina HiSeq 2500 with 150 bp paired-end reads.

Data analysis of GATA3 ChIP-seq

Chromatin Immunoprecipitation sequencing (ChIP-seq) of GATA3 was analyzed using the ChiLin pipeline (Qin et al., 2016). Briefly, the Sentieon Bwa-mem aligner was used to map reads to the hg38 reference genome (https://support.sentieon.com/manual/). ChIP-seq peak calling was then performed using MACS2 v2.1.4 (Zhang et al., 2008), with the following parameters: "-SPMR -B -q 0.01 –keep-dup 1." Mapped reads were then down sampled to 4 million for subsequent quality control analysis. Quality control consisted of five metrics (Table S5): 1) the average read quality according to FastQC (Andrews, 2010); 2) the fraction of uniquely mapped reads; 3) a PCR bottleneck coefficient, which is the fraction of locations with one uniquely mapped read; 4) fraction of reads in peaks according to MACS2 (Zhang et al., 2008) (more, the better); 5) overlap of peaks with DNA hypersensitivity sites. All samples were of adequate quality.

To provide a consistent peak set across multiple samples for downstream analysis, we merged overlapping peaks using bedtools v2.29.2 (Quinlan and Hall, 2010). Differential peak analysis between wild-type GATA3 and mutant GATA3 was then performed using DESeq2 with the default Wald test (Love et al., 2014). Peaks were regarded as significant at Benjamini-Hochberg False Discovery Rate of 0.1 (Table S5). A heatmap visualizing the peaks was then generated using the deeptools package (v3.3.0) (Ramírez et al., 2016). KEGG pathway enrichment of the upregulated GATA3 peaks was then conducted using Cistrome GO (Li et al., 2019).

Labeling of driver mutations

Even implicated cancer driver genes contain a mixture of driver and passenger mutations when examined across multiple patients' tumors (Torkamani and Schork, 2008). Therefore, we restricted our subsequent analysis of putative substrates or immune-related biomarkers to likely driver mutations in the implicated set of 63 ubiquitin pathway genes. For tumor suppressor genes, we regarded any loss-of-function mutation (frameshift insertions or deletions, nonsense mutations, splice site mutations, lost start mutations, or lost stop mutations) as likely oncogenic, which is consistent with variant annotation guidelines from curated databases such as OncoKB (Chakravarty et al., 2017). However, the interpretation of missense mutations is often more difficult. We therefore used missense mutations that were previously reported to be drivers by CHASMplus at an FDR of 0.01(Tokheim and Karchin, 2019).

Comparison of CHASMplus to saturation mutagenesis

To understand the accuracy of the driver mutation labeling by CHASMplus, we compared predictions to a recent saturation mutagenesis study (Findlay et al., 2018) of the functional effect of all BRCT and RING domain variants in BRCA1, an E3 ubiquitin ligase. The study used a multiplexed functional assay in a homology-directed repair (HDR) sensitive cell line (HAP1) to measure the impact of BRCA1 mutations. Scores for CHASMplus were obtained from OpenCRAVAT (https://opencravat.org/; Masica et al., 2017) and then assessed for their spearman correlation with the functional HDR scores. Additionally, CHASMplus scores were assessed for their performance at distinguishing ClinVar labeled pathogenic versus benign variants in BRCA1 based on the area under the Receiver Characteristic Curve. ClinVar labels were obtained from Findlay et al. (2018) (n = 46).

Quality control of Cistrome ChIP-seq data

First, we examined the overall distribution of 5 quality control (QC) metrics for ChIP-seq from putative substrates identified by Rabit compared to all transcription factors in the Cistrome database. The 5 QC metrics were: 1) the average read quality according to FastQC; 2) the fraction of uniquely mapped reads; 3) a PCR bottleneck coefficient; 4) fraction of reads in peaks according to MACS2 (Zhang et al., 2008) (more, the better); 5) overlap of peaks with DNA hypersensitivity sites. By kernel density estimation, we observed that the putative substrates had a nearly identical distribution of QC scores across all 5 metrics (Figure S7), suggesting that there is no systematic QC problem in our analysis.

Next, we wanted to investigate whether only a few transcription factors might appear as outliers. To do this, we analyzed the number of times a transcription factor appeared in the Rabit result and its corresponding median log-transformed p value. We reasoned that poor-quality ChIP-seq might consistently, across many analyses, appear as highly significant, possibly due to technical artifacts. Outlier analysis was carried out through robust covariance estimation (scikit learn python package) (Rousseeuw and van Driessen, 1999), assuming a Gaussian distribution and a significant contamination rate of 0.05 (Figure S7). After manual examination of the outliers, we identified the genes *SCML2* and *ZNF274* as having significantly worse ChIP-seq quality than compared to other transcription factors in the Cistrome database (Figure S7). We therefore exclude these two transcription factors from further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Gene ontology enrichment analysis

We performed gene ontology enrichment analysis for putatively identified driver genes using DAVID (Huang et al., 2009) with the 775 UPS genes as the background. Biological process terms were deemed significant at an FDR of 0.25 (Figure S1).

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Overlap with previously implicated driver genes

We compared our putative UPS driver genes to a previous study that found significantly mutated UPS genes (Ge et al., 2018; Davoli et al., 2013), the Cancer Gene Census (downloaded January 7, 2017) (Sondka et al., 2018), and the set of driver genes defined by the TCGA PancanAtlas consortium (Bailey et al., 2018). Gene list enrichment was assessed using a one-tailed fisher exact test with a background consisting of all UPS genes.

Boxplots

All boxplots show the distribution quartiles with whiskers representing the quartile ± 1.5 times the Interquartile Range (IQR).

Protein-protein interaction network and Betweenness Centrality

Protein-protein interaction network data was download from the BioGrid website (v3.5.178) (Oughtred et al., 2019). The betweenness centrality measures how often a node in a network is situated on the shortest path between two other nodes in a network. Nodes with higher betweenness centrality are often hubs within a network. Betweenness centrality was computed for the BioGrid (Oughtred et al., 2019) protein-protein interaction network (downloaded 11/22/2019) using the networkx python package. Formally, for all possible pairs of nodes (s and t) in a network with nodes V, the betweenness centrality of a node (n) is the fraction of shortest paths (σ) that go through that node (Equation 1).

$$Betweeness\ Centrality = \sum_{s,t \in V} \frac{\sigma_{st}(n)}{\sigma_{st}} \quad \text{(Equation 1)}$$

Where $\sigma_{st}(n)$ is the number of shortest paths between node s and t that go through node n and σ_{st} is the total number of shortest paths.

Association of mutations with protein abundance

Using linear regression, we correlated the mutation status of each of the 63 putative driver genes with protein abundance from Reverse Phase Protein Arrays (RPPA) in TCGA. Only non-silent mutations were considered. A Wald test was performed after adjustment for tumor purity by ABSOLUTE (downloaded from https://gdc.cancer.gov/about-data/publications/pancanatlas) (Carter et al., 2012), tumor subtype (Sanchez-Vega et al., 2018) and RNA expression of the potential substrate (FDR < 0.1 and effect size > 0.25; Figure S6). The adjustment for RNA expression of potential substrates is important because it helps distinguish between direct UPS effects mediated through the protein-level from upstream effects at the transcriptional level.

Transcription factor substrate analysis

Conceptually, alterations in UPS genes should be able to explain the downstream target gene expression of a transcription factor by modulation through protein abundance or activity (Figure S7). To analyze this, first, we computed the differential expression between tumor samples containing putative driver mutations in a gene of interest versus those that did not (t test), while adjusting for tumor purity by ABSOLUTE (downloaded from https://gdc.cancer.gov/about-data/publications/pancanatlas) (Carter et al., 2012) and tumor subtypes (Sanchez-Vega et al., 2018). The generated differential expression profile was then analyzed by Rabit (Jiang et al., 2015) to associate top transcription factor (TF) regulators. Rabit infers transcriptional regulators based on TF binding sites using thousands of ChIP-seq profiles from the Cistrome database (Zheng et al., 2019) while adjusting for background covariates such as CpG density. For computational tractability reasons, we then corrected for transcription factor RNA expression only for the top 10 hits according to p value, by repeating the above analysis but with the TF RNA expression included as a covariate. A second round of Rabit analysis was then conducted using the TF adjusted differential expression profiles. While results were only carried out for the top 10 hits in each analysis, multiple testing correction (Bonferroni method) was carried out with consideration of all TFs as possible (family wise error rate < 0.05). Note, analysis was only performed for the cancer types implicated by driver analysis for the specific UPS gene. Code used for this analysis is available on GitHub (https://github.com/ctokheim/tf_association).

Gene co-essentiality analysis from DepMap

The correlation between two gene's dependency scores (CERES score) from CRISPR screens in DepMap was analyzed through a linear regression model. The cell culture type (adherent, suspension, etc.) and a CRISPR quality control metric (SSMD of control genes) was added as covariates. The statistical significance of the correlation was assessed by a Wald test.

Correlation with immune-related gene expression signatures

Using a linear regression model, we correlated the mutation status (see section: labeling of driver mutations) of each identified UPS driver gene or significantly mutated substrate with at least 5 putative driver mutations to several immune-related gene expression biomarkers from Thorsson et al. (2018) (signatures: leukocyte fraction, IFNG response, TGFB response, macrophage regulation and wound healing). A t test was used to assess significance after adjusting for tumor subtypes and the non-silent mutation rate of a tumor. Associations were deemed significant at an FDR threshold of 0.1.

Please cite this article in press as: Tokheim et al., Systematic characterization of mutations altering protein degradation in human cancers, Molecular Cell (2021), https://doi.org/10.1016/j.molcel.2021.01.020

Correlation with T cell co-culture CRISPR screen

Data from a previous T cell co-culture CRISPR screen (Pan et al., 2018) across two conditions were used to assess whether UPS genes correlated with immune-related gene expression signatures might affect T cell mediated killing of cancer cells. The two conditions used in the screen were: 1) Pmel T cells which recognize endogenously expressed gp100 antigen on a B16 melanoma cell line while in the presence of IFNG compared to a non-antigen-specific T cell; 2) OT1 T cells that recognize B16 cells with media supplemented with or without the ovalbumin antigen. The log fold change of the single guide RNA (sgRNA) and the estimate of significance (z-score) were obtained through the TIDE website (http://tide.dfci.harvard.edu/; Jiang et al., 2018). The z-scores from the two conditions (Pmel and OT1) were combined using Stouffer's method to generate a meta-analysis z-score and corresponding p value.

Association with overall patient survival

Curated overall survival information for TCGA was obtained from the genomic data commons (https://gdc.cancer.gov/about-data/publications/pancanatlas; Liu et al., 2018). Using a Cox proportional-hazard model, a Wald test was used to assess the statistical significance of any association with survival. Tumor purity and subtype were included as covariates. Kaplan-Meier curves were generated using the TIDE website (http://tide.dfci.harvard.edu/; Jiang et al., 2018).