The functional landscape of the human ubiquitinome

Julian van Gerwen[1,2], Maximilian Fottner[3], Shengbo Wang[4], Bede Busby[5], Ellen Boswell[6], Paul Schnacke[3], Andrea C. Carrano[7], Malina A. Bakowski[7], Emily R. Troemel[7], Romain Studer[4], Marta Strumillo[4], Maria-Jesus Martin[4], J. Wade Harper[8], Kathrin Lang[3], Andrew R. Jones[6,#], Eric J. Bennett[7], Juan Antonio Vizcaíno[4,#], Inigo Barrio-Hernandez[9,#], Pedro Beltrao[1,2,#]

1 - Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Switzerland
2 - Swiss Institute of Bioinformatics, Lausanne, Switzerland
3 - Laboratory for Organic Chemistry Department of Chemistry and Applied Biosciences, ETH Zurich
4 - European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
5 - Ministry for Primary Industries, Biosecurity New Zealand
6 - Department of Biochemistry, Cell and Systems Biology, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7BE, U.K.
7 - UC San Diego, Department of Cell and Developmental Biology
8 - Department of Cell Biology, Harvard Medical School, Boston, MA, USA
9 - Instituto de Agrobiotecnología, Mutilva Baja, Spain
corresponding authors: andrew.jones@liverpool.ac.uk juan@ebi.ac.uk inigo.barrio@csic.es pbeltrao@ethz.ch

Abstract

Protein ubiquitination regulates cell biology through diverse avenues, from quality control-linked protein degradation to signaling functions such as modulating protein-protein interactions and enzyme activation. Mass spectrometry-based proteomics has allowed proteome-scale quantification of hundreds of thousands of ubiquitination sites (ubi-sites), however the functional importance and regulatory roles of most ubi-sites remain undefined. Here, we assembled a human reference ubiquitinome of 108,341 ubi-sites by harmonizing public proteomics data. We identified a core subset of ubi-sites under evolutionary constraint through alignment of ubiquitin proteomics data from six non-human species, and determined ultra-conserved ubi-sites recurring at regulatory hotspots within protein domains. Perturbation proteomics revealed that these highly conserved ubi-sites are more likely to regulate signaling functions rather than proteasomal degradation. To further prioritize functional ubi-sites with roles in cellular signaling, we constructed a functional score for more than 100,000 ubi-sites by integrating evolutionary, proteomic, and structural features using machine learning. Our score identifies ubi-sites regulating diverse protein functions and rationalizes mechanisms of genetic disease. Finally, we employed chemical genomics to validate the functional relevance of high-scoring ubi-sites and leveraged genetic code expansion to demonstrate that ubiquitination of K320 in the RNA-regulator ELAVL1 disrupts RNA binding. Our work reveals systems-level principles of the ubiquitinome and provides a powerful resource for studying protein ubiquitination.

Introduction

Ubiquitination is a reversible post-translational modification (PTM) critical for cell biology. It involves the attachment of the 76-amino acid protein ubiquitin to specific lysine residues on substrate proteins, catalysed by the concerted activity of E1, E2, and E3 enzymes, and opposed by deubiquitinating enzymes[1–3]. Ubiquitin can be attached as a single moiety or as polyubiquitin chains of varying linkage types and lengths, each dictating diverse regulatory outcomes[1,4]. The best characterised of these fates is the targeting of aberrant proteins for proteasomal degradation, but ubiquitin can also modulate protein-protein interactions, induce protein conformational changes, and alter protein subcellular location, among other functions[1,4]. These functions place ubiquitin as a central player both in protein quality control and diverse signaling processes including DNA repair[5], cell cycle progression[6], and endocytosis[7]. Given the centrality of ubiquitination in cellular physiology and its frequent dysregulation in disease states[6], elucidating the molecular and functional consequences of ubiquitination events remains a critical goal in cell biology.

Our understanding of protein ubiquitination has been revolutionised over the last two decades by mass spectrometry-based proteomics, which allows the unbiased identification and quantification of ubiquitination sites at proteome scale using selective enrichment strategies[8–10]. Pioneering proteomics studies have mapped the dynamic responses of thousands of ubi-sites across a wide range of cellular perturbations[11–15] and have elucidated systems-level properties such as ubiquitin site occupancy and turnover[16]. However, the discovery afforded by proteomics has rapidly outpaced functional characterisation - of the approximately 100,000 human ubi-sites identified in aggregated mass spectrometry experiments, only approximately 1,000 have an experimentally determined regulatory role, as reported in the PhosphositePlus repository[17]. This constitutes a fundamental barrier to understanding the regulatory mechanisms of ubiquitin at the scale of the cell.

A parallel challenge has been addressed more extensively in the study of protein phosphorylation, a PTM of similar scale that also suffers from a widespread lack of functional characterisation[18]. Analysis of the evolutionary conservation of phosphorylation has proven useful in differentiating evolutionarily constrained sites critical for cell function from rapidly evolving sites potentially representing non-functional cellular noise[19–21]. Generalising this approach, we previously constructed a machine learning-based functional score that integrates diverse information to prioritise important regulatory phosphosites at the scale of the human proteome[22]. Overall, in comparison to phosphorylation, systems-wide analyses of the conservation, regulation and functionality of the human ubiquitinome are lacking.

Here, we performed a proteome-wide characterisation of the functional significance of human ubiquitin sites. We first assembled a high-confidence reference ubiquitinome of more than 100,000 ubi-sites by re-analyzing 11 published proteomics datasets. Leveraging published and novel ubiquitin proteomics data across six species and 103 cellular perturbations we uncovered shared ubiquitinome responses to cellular stress, identified thousands of evolutionary conserved sites, and found that these sites tend to mediate non-degradative signaling functions rather than proteasomal degradation. We then constructed a machine-learning based functional score

which prioritises cell-critical ubi-sites performing diverse regulatory roles and implicated in genetic disease susceptibilities. Finally, we used chemical genomics to identify cellular roles for high-scoring ubi-sites in β-tubulin, the glycolytic enzyme GPI, and the GTPase RAB18, and employed genetic code expansion to demonstrate that ubiquitination of K320 in the post-transcriptional regulator ELAVL1 inhibits RNA binding. Our work constitutes a significant advance in our understanding of the human ubiquitinome and a powerful resource for the study of ubiquitin site function.

Results

Constructing a human reference ubiquitinome

Mass spectrometry-based proteomics coupled to selective enrichment workflows has enabled proteome-wide identification of ubi-sites[9–11]. Such data is continuously collated in PTM-centric resources such as PhosphositePlus, providing a reference of possible ubi-sites[17]. However, this approach can lead to inaccurate PTM identifications, if false discovery rates are not statistically controlled across multiple studies and datasets[22]. Hence, we first sought to generate a high-confidence reference set of human ubi-sites by systematically re-analyzing public mass spectrometry-based proteomics datasets (Fig. 1a).

We began by curating 11 recent ubiquitin proteomics datasets from the PRIDE data repository[23] that covered a range of cell types and experimental conditions (Table S1[10,11,16,24–31]). We then extracted and re-analyzed raw proteomics data using both community-standard and bespoke analysis tools (see Methods), resulting in an average of 1,108,216 peptide-spectrum matches (PSMs) at 1% global false-discovery rate per dataset, 23,854 localized ubi-sites at 5% false-localisation rate per dataset, and 12,190,377 PSMs and 108,341 ubi-sites across datasets (Fig. 1a). We implemented a scoring scheme to address the increased risk of false localizations arising from aggregating multiple datasets, classifying ubi-sites with multiple high-stringency identifications as Gold (FLR < 1% in ≥ 2 datasets), ubi-sites with one high-stringency identification as Silver (FLR < 1% in 1 dataset), and those supported by data using only standard stringency as Bronze (1 < FLR ≤ 5% in ≥ 1 datasets)[32]. Of the 108,341 total unique ubi-sites, 41,070 were identified with Gold-level confidence and 35,369 with Silver-level confidence, underscoring the robustness and reliability of our pipeline (Fig. 1b). Ubi-sites were found on 11,079 total proteins, most of which contained at least one Gold-level ubi-site (Fig. 1b).

We next compared our reference ubiquitinome to existing PTM repositories. While PhosphositePlus contained a similar number of sites (98,078 as of 25 Nov 2024), these sites were generally observed less frequently across datasets than those in our ubiquitinome, with over half detected in only a single contributing dataset (Fig. 1c). Since PhosphositePlus aggregates a much larger number of datasets than ours (more than 150 published and internal datasets), this high proportion of single-source sites likely reflects a greater prevalence of false positives, as previously demonstrated for protein phosphorylation[22]. Overall, our dataset still showed strong concordance with both PhosphositePlus (71,268 shared sites) and PTMAtlas, a recent PTM database generated from reanalysis of public proteomics data (75,930 common sites, 106,038 sites in PTMatlas) testifying to the robustness of our ubi-site identifications (Fig.

S1a). Furthermore, ubiquitinated proteins in our dataset spanned almost the entire range of abundances in the human proteome with only minor differences compared to non-ubiquitinated proteins and across confidence levels (Fig. 1d-e[33]). This observation indicates that our dataset is only minimally affected by the bias of mass spectrometry to high-abundance molecules.

Within our reference ubiquitinome we found a wide range in the number of ubi-sites detected per protein (Fig. 1f). This was not strongly associated with protein abundance, protein length, or lysine content, suggesting this observation is not driven by biases in detectability by mass spectrometry or the availability of modifiable lysines (Fig. 1g). Therefore, extensive ubiquitination within a single protein likely reflects a regulated biological phenomenon. The 3,486 proteins with more than 10 ubi-sites were enriched in DNA repair and cell cycle pathways (Fig. 1j) and at diverse cellular locations including the centrosome and spindle (Fig. 1k), marking these as focal points for extensive ubiquitin regulation. These observations are consistent with a large body of literature describing extensive ubiquitin-dependent regulation of DNA repair[5], and with previous findings that highly ubiquitinated proteins are enriched at the centrosome and spindle[10].

Overall, we have generated a large-scale, high-confidence atlas of human ubiquitin sites that can provide rich insight into the human ubiquitination landscape. To ensure that these data meet FAIR guidelines[34], the outputs have been made available in PRIDE (PRIDE number PXD068989, Table S2).

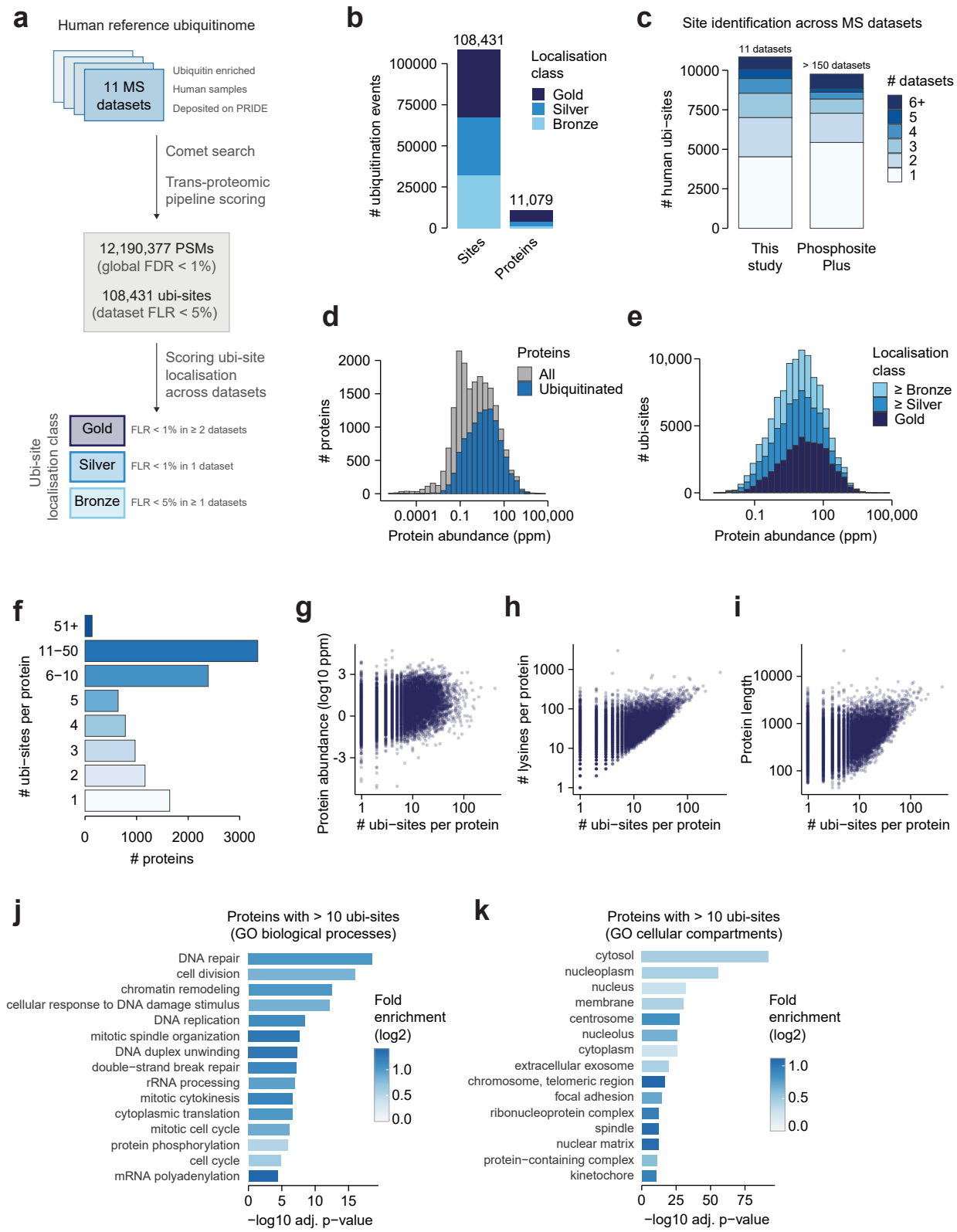

Figure 1: Constructing a human reference ubiquitinome

A) Pipeline for re-analyzing ubiquitin-proteomics spectral data and aggregating it into a reference

ubiquitinome. B) The number of ubiquitinated sites and proteins in the human reference ubiquitinome, stratified by their confidence level. A protein's confidence level is calculated as the maximum across all of its sites. C) The number of human ubiquitination sites in our reference ubiquitinome and PhosphositePlus[17], stratified by the number of datasets in which each site is quantified. D) The abundance of all proteins in the human reference ubiquitinome compared to the entire proteome, using abundance measurements from PaxDB[33]. E) The protein abundance of ubiquitination sites at different levels of identification confidence. F) The number of identified ubi-sites per protein. G-I) The relationship between the number of ubi-sites identified per protein and E) protein abundance, F) protein length, and G) the number of lysines in a protein sequence. J-K) The top 15 significantly overrepresented J) GO biological process terms and K) GO cellular compartment terms among proteins with more than 10 ubi-sites, relative to all proteins in the human reference ubiquitinome (one-sided Fisher's exact test, Benjamini-Hochberg p-value adjustment).

Characterizing the conservation of human ubi-sites

The vast majority of sites within our reference human ubiquitinome are functionally uncharacterised. As an initial step towards addressing this knowledge gap, we aimed to map the relative functional importance of ubi-sites. Evolutionary conservation is a powerful tool to identify functionally important regions within proteins and has been applied previously to PTMs such as phosphorylation[20,21]. However, a proteome-wide analysis of the ubi-site conservation is lacking.

To assess ubi-site conservation, we assembled published and in-house ubiquitin proteomics data from six species spanning a range of evolutionary distances from *Homo sapiens* (Fig. 2a, Table S3 and S4[35–41]). Human ubi-sites can range from very low conservation - where the acceptor lysine residue is absent at the orthologous position in other species - to very high conservation - where the acceptor lysine is conserved and ubiquitinated in related and distant species (Fig. 2b). Additionally, conserved lysines without detected ubiquitination provide intermediate evidence of conservation, as non-detection can result from incomplete mass spectrometry coverage or conditional regulation (Fig. 2b). Hence, we classified the conservation of human ubi-sites considering both ubi-site detection and ubi-acceptor lysine conservation (Fig. 2c, see Methods).

We detected high conservation for only a small fraction of the human ubiquitinome; of the 100,609 analyzed sites, only 17,805 displayed ubiquitination in *Mus musculus* or *Rattus norvegicus* (conservation level 4), and 3,814 displayed additional ubiquitination in the more divergent species *Gallus gallus, Drosophila melanogaster, Caenorhabditis elegans,* or *Saccharomyces cerevisiae* (conservation level 5, Fig. 2d). These values likely underestimate true conservation, as our pan-species proteomics data contains far fewer ubi-site identifications from non-human species compared to human. Regardless, the two highest conservation levels were enriched for ubi-sites annotated with regulatory, non-degradation functions in PhosphoSitePlus (Fig. 2e, 2.27- and 3.21-fold enrichment for levels 4 and 5), and also showed enrichment—albeit to a lesser degree—for degradation-related functions (Fig. S2a, 2.00- and 2.00-fold enrichment for levels 4 and 5). This confirms that evolutionary constraint identifies functionally relevant modifications with known regulatory roles.

Next, we employed an orthogonal approach to identify conserved ubi-sites. Following our previous work on protein phosphorylation[21] and related work by others[42,43], we searched for small regions in protein domains that are repeatedly ubiquitinated in domain instances within and across genomes, which we term “ubi-site hotspots” (Fig. 2f, see Methods). Such modifications have likely been conserved to regulate domain functions for millions of years of protein evolution. We identified 36 ubiquitination hotspots within 29 InterPro protein domains, comprising a total of 1,973 human ubiquitin sites (Fig. 2g, Table S5, Data S1, see methods). These hotspots were enriched in ubiquitin sites annotated with non-degradation functions (2.58-fold, one-sided Fisher’s exact test p-value = 0.00062) and to a lesser extent degradation functions (1.41-fold, one-sided Fisher’s exact test p-value = 0.23), demonstrating that domain-level conservation identifies functionally important modifications.

To further illustrate the value of ubi-site hotspot analysis, we highlight representative protein domains as case studies. We identified three ubi-site hotspots within the eukaryotic protein kinase domain (IPR000719) - a critical functional unit for cellular signaling - the most prominent of which is structurally proximal to the kinase activation loop (Fig. 2h). This hotspot contains seven characterised regulatory ubi-sites mostly annotated to activate their respective kinase, such as K132 on the cell cycle checkpoint kinase Chk1 and K202 on the mitotic kinase AurB. (Fig. 2h, Table S5). Of the approximately 500 human protein kinases[44], 163 contain a ubi-site in this hotspot, suggesting this may represent a general regulatory mechanism controlling kinase signaling. Second, we identified two hotspots in the small GTPase domain (IPR001806), which switches between inactive GDP-bound and active GTP-bound conformations to modulate protein-protein interactions (Fig. 2i). The most prominent hotspot is adjacent to the GTP/GDP-binding sites and mono-ubiquitination of HRAS K117 in this hotspot has been shown to promote nucleotide exchange[45], highlighting the regulatory potential of this hotspot. Other illustrative examples include hotspots at the tip of the C2H2-type zinc finger domain (IPR013087), at the GEF-GTPase interface of the Dbl homology domain (IPR000219), and proximal to the ubiquitin-binding cleft in the catalytic domain of USP deubiquitinases (IPR001394, Fig. S3a-c).

Based on the above findings, we anticipate that highly conserved ubi-sites and hotspot ubi-sites represent a wealth of modifications with regulatory significance.

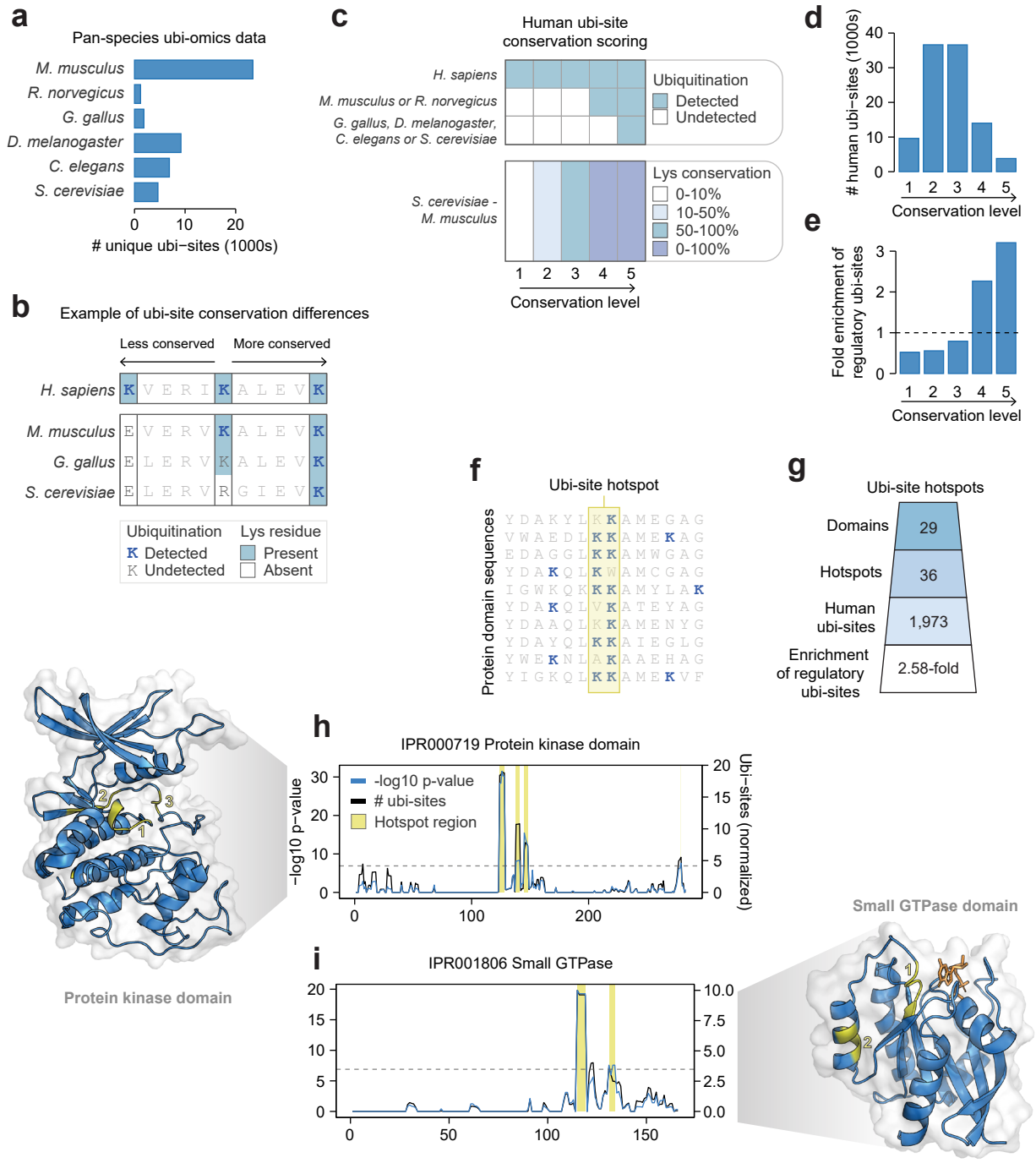

Figure 2: Characterizing the conservation of human ubi-sites

A) The number of ubi-sites identified in a compilation of ubiquitin-proteomics data across six non-human species. B) A toy sequence alignment demonstrating differences in the conservation of lysines and their ubiquitination. C) Rules for scoring the conservation of human ubi-sites based on the detection of ubiquitin and the presence of lysine at the orthologous position in other species. Lys conservation indicates the percentage of residues at the given alignment position that are lysine. D) The number of human ubi-sites at each conservation level. E) The enrichment of ubiquitin sites annotated to perform non-degradation regulatory functions at each conservation level (PhosphositePlus[17]). F) Conceptual illustration of ubi-site hotspots: regions within aligned protein

domain sequences enriched in ubiquitination. G) The number of identified ubi-site hotspots, hotspot domains, and human hotspot sites. The enrichment of human hotspot sites annotated to perform non-degradation regulatory functions (see Methods) is shown. H-I) Identification of ubi-site hotspots in two protein domains. The black line indicates the average number of ubi-sites observed across the domain sequence alignment within a rolling window, normalised by subtracting the number of ubi-sites expected by chance. The blue line indicates the p-value associated with the enrichment of ubi-sites at each alignment position. The horizontal line indicates a Bonferroni-corrected p-value cut-off of 0.01 (uncorrected p-value $< 1.29\times10^{-7}$). Positions with a -log10 p-value above this cut-off and average number of phosphosites per window higher than 2 are classified as hotspot regions and highlighted with a yellow bar. Hotspot regions are mapped onto representative structures for each domain in yellow. Multiple hotspots are labeled in their order in the domain sequence.

Characterizing the regulation of human ubi-sites

Beyond conservation, the dynamic regulation of ubi-sites across cellular perturbations can provide further insight into their functional importance and regulatory roles. To enable analysis of ubi-site regulation, we analyzed quantitative human ubiquitin proteomics data from 14 publications and generated an additional 10 proteomics datasets in-house (Fig. 3a, Table S3[12–15,46–55]). The resulting compilation encompasses 103 control-matched perturbation conditions, 62,094 human ubi-sites, and 452,191 quantified fold changes (Fig. 3b). These perturbations include diverse cellular stressors such as DNA damaging agents and ER stressors, as well as inhibitors of deubiquitinases (DUBs), neddylation, or the proteasome (Fig. 3c).

To understand the systems-level organisation of the ubiquitinome and its response to these perturbations, we first correlated and clustered the perturbation conditions (Fig. 3d). We generally observed medium-to-high correlations within the same perturbation across separate studies (Fig. S4a, average Pearson's r = 0.498 between studies and r = 0.605 within studies). This highlights the reproducibility of the assembled data, especially considering studies did not necessarily employ the same treatment dose, treatment duration, cell line, or mass spectrometer.

Correlations in the response to different perturbations indicate potential functional relationships, such as the positive association between DNA damage-inducing UV radiation, protein folding stress, and ER stress (cluster III in Fig. 3d, example in Fig. 3e). These stressors all induce the unfolded protein response[56,57], which triggers specific regulatory ubi-sites, including several sites on the ribosomal proteins RPS2 and RPS3 identified in our data[14] (Fig. 3e). The unfolded protein response also globally inhibits protein synthesis[56], leading to a widespread decrease of ubiquitination ordinarily resulting from the degradation of newly synthesised proteins[13]. In addition, the responses to UV and proteasome inhibition were negatively associated (Fig. 3d, f), particularly when the proteasome had already been inhibited before UV induction (Fig 3d, Fig. S4b). This may be explained by two mechanisms: first, degradative ubi-sites from newly synthesised proteins may decrease upon the UV-induced unfolded protein response but accumulate with acute proteasome inhibition[13]; second, this degradative chain buildup may limit ubiquitin availability for non-degradative DNA damage signals such as K6 and K63 chains[1,12,13,55]

We next assessed whether distinct patterns of ubi-site regulation are associated with differing levels of ubi-site conservation. Compared to all ubi-sites, highly conserved sites tended to be more up-regulated upon DNA damage, translation inhibition, and infection (Fig. 3g). In contrast, treatment with the proteasome inhibitors MG-132, bortezomib or epoxomicin enhanced the abundance of most poorly conserved sites, whereas many highly conserved sites failed to increase (Fig. 3g, h, Fig. S4c-d). Given that proteasome inhibition should cause the accumulation of ubi-sites actively targeting proteins for degradation (Fig. 3i), this observation implies that conserved ubi-sites are typically less likely to promote proteasomal degradation compared to the entire ubiquitinome. As an orthogonal measure of conservation we inspected our domain hotspots (Fig. 2f-j), and found that their ubi-sites were also less up-regulated upon proteasome inhibition (Fig. 3j, Fig. S4e-f). Taken together, these results suggest that ubi-sites that have been conserved throughout evolution in order to support organismal fitness are less likely to promote proteasomal degradation, and instead may regulate protein function in signaling contexts such as the DNA damage response and infection.

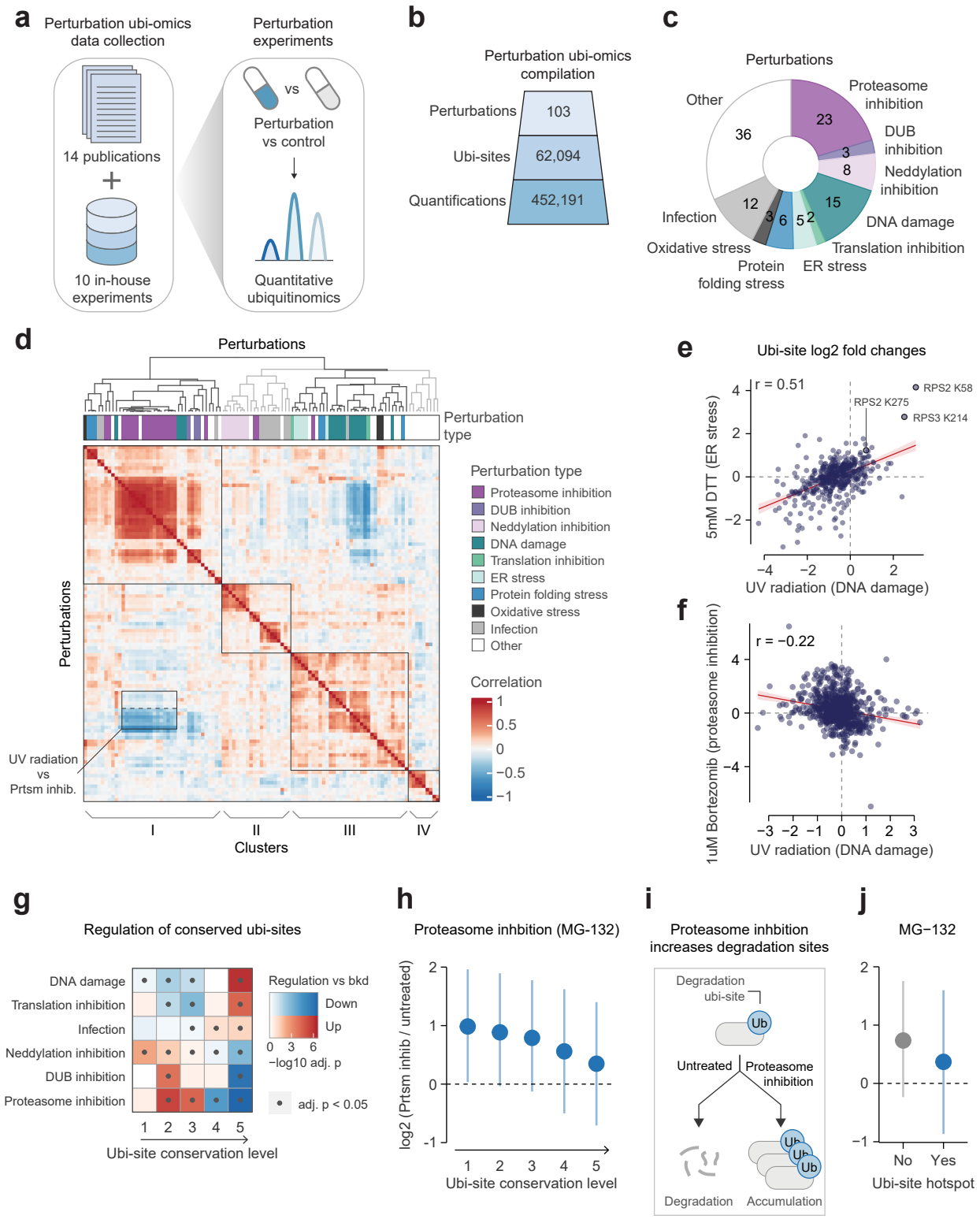

Figure 3: Characterizing the regulation of human ubi-sites

A) Ubiquitin proteomics experiments measuring cellular perturbations in human cells were collated from 14 publications and 10 in-house experiments. B) The number of perturbation conditions, quantified ubi-sites, and quantified ubi-site fold change values across compiled ubiquitin proteomics data. C) The

distribution of different types of perturbations in compiled ubiquitin proteomics data. D) Pairwise Pearson's correlations between perturbation conditions based on ubi-site log2-fold changes. Perturbations were divided into four clusters by hierarchical clustering using Pearson's correlation as the distance metric, indicated by different colours in the dendrogram and labels below the heatmap. The negative correlation between UV radiation conditions and proteasome inhibition conditions is outlined, where the bottom half of the box indicates UV radiation of cells with prior proteasome inhibition, while the top half indicates UV radiation of untreated cells. E-F) Examples illustrating E) apositive (UV radiation vs. ER stress) and F) negative (UV radiation vs. proteasome inhibition) correlations of ubi-site responses between perturbation types. Pearson correlation coefficients (r) are shown. Ribosomal ubi-sites implicated in the unfolded protein response are labeled[14]. G) Two-sided gene-set tests (Limma v3.56.2, geneSetTest, type = "t", alternative = "either") were performed to identify conditions in which ubi-sites at each conservation level were up- or down-regulated relative to all other ubi-sites. P-values were adjusted by the Benjamini-Hochberg procedure, converted into log10-values, and averaged within each perturbation group. Only perturbation groups with at least one significant enrichment (adj. $p < 0.05$) are shown. H) Mean log2-fold changes (± s.d.) of ubi-sites under proteasome inhibition (MG-132 treatment), grouped by evolutionary conservation level. Final values were obtained by normalising and averaging log2 fold-changes from multiple experiments (see Methods). The number of ubi-sites in each group from left-to-right are 1451, 6404, 7869, 5234, and 2144. I) Schematic explaining that proteasome inhibition should enhance ubi-sites actively targeting proteins for proteasomal degradation. J) As in H), for sites in ubi-site domain hotspots. The number of ubi-sites in each group from left-to-right are 23,758 and 503.

Prioritizing highly functional regulatory ubi-sites

Motivated by our analyses of ubi-site conservation and regulation, we sought a more general strategy to prioritise ubi-sites performing non-degradative regulatory roles important for cellular function. Mirroring our previous work on phosphorylation[22], we compiled 16 features potentially discriminative for ubi-site functionality, including conservation and structural context (Fig. 4a, Table S6). We then integrated these features into a classifier trained to discriminate the 213 non-degradative ubi-sites annotated in PhosphositePlus for which we could calculate features (Table S7, see Methods) from the 105,943 unannotated sites, under the assumption that many unannotated sites are likely non-essential for cellular function. The resulting classifier output can be interpreted as a "functional score" that prioritises known sites important for the cell; any unannotated sites that also score highly have putative functional importance.

Among several widely used classifiers - logistic regression, gradient boosting, and random forest - logistic regression demonstrated equivalent performance to the other more complex methods, achieving an average ROC AUC of 0.77 in five-fold train-test splits (Fig. S5a-b). Hence, we used logistic regression for final predictions. We then generated predictions by aggregating median scores from 25 logistic regression models trained through five time-repeated five-fold cross-validation, allowing us to establish robust functional scores (Table S6).

Many unannotated sites feature high functional scores equivalent to annotated sites, implying that a substantial fraction of the unexplored ubiquitinome may be important for cellular function (Fig. 4b). In particular, a functional score cutoff of 0.65 recovers approximately 50% of annotated regulatory ubi-sites while classifying around 12,281 previously unannotated sites as functionally significant. Logistic regression feature weights revealed that the functional score

leveraged nearly all features to varying extents, with ubi-site presence within protein domains or regions, quantification in proteomics data, and conservation exerting the strongest positive influence (Fig. 4c). This underscores the value of our integrative machine-learning approach .

To validate the relevance of high-scoring ubi-sites for organismal fitness we inspected clinical missense mutations from ClinVar, which can pinpoint protein residues critical to human health[58]. Ubi-sites with higher functional scores were more likely to co-localise with pathogenic mutations rather than benign or uncertain mutations, consistent with these sites performing critical biological functions (Fig. 4d). This included structurally conservative lysine-to-arginine substitutions that are likely to exert pathogenic effects by disrupting ubiquitination rather than protein structure, such as K128R in the GTPase RALA, K158R in the kinase CSNK2A1, and K536R in SMC1A (Fig. 4d). K536 is located within the hinge domain that mediates dimerization with SMC3 in the Cohesin complex[59], hence ubiquitination of this site may regulate complex assembly, and its disruption could underlie the cerebral congenital muscular hypertrophy associated with the K536R variant (Fig. 4e).

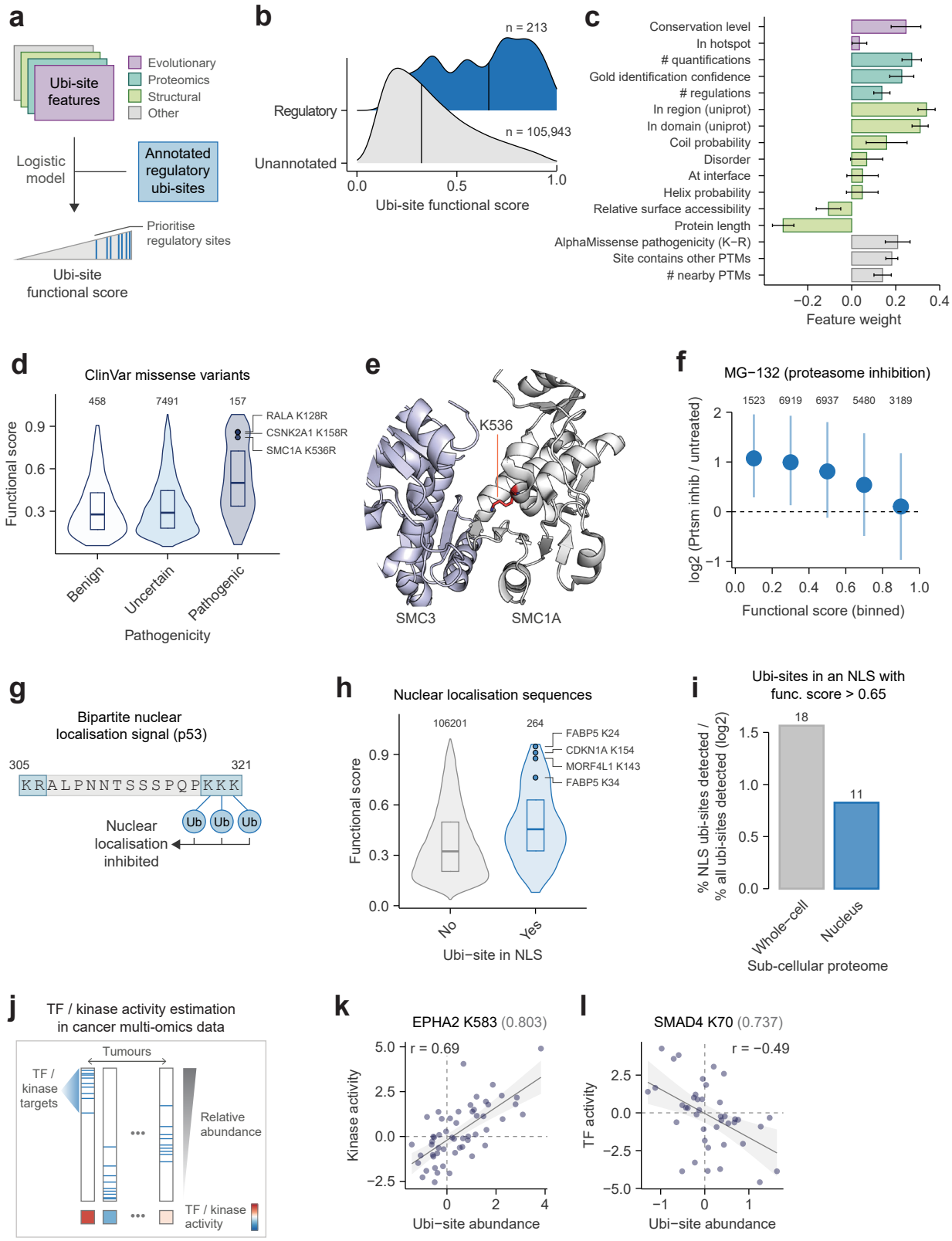

Figure 4: Prioritizing highly functional regulatory ubi-sites

A) Overview of the workflow for generating a ubi-site functional score. Sixteen features capturing distinct biological dimensions were integrated into a logistic regression model trained to separate ubi-sites annotated for non-degradation regulatory functions from unannotated sites. B) Distribution of functional scores among ubi-sites annotated for non-degradation regulatory functions compared to unannotated ubi-sites. Numbers of unique ubi-sites per group are shown. C) Feature weights derived from logistic regression coefficients. Features with positive weights increase the functional score while those with negative weights decrease it. Error bars indicate standard deviation across multiple logistic regression models. D) Functional scores of ubi-sites co-localising with missense variants annotated in ClinVar as benign, uncertain, or pathogenic[58]. Numbers of unique ubi-sites per group are indicated. Examples of high-scoring pathogenic lysine-to-arginine mutations are shown. E) The interface of the SMC3 and SMC1A hinge domains within a Cryo-EM structure of the cohesin complex (PDB: 6WG4[59]). K536 on SMC1A is indicated in red. F) Mean log2-fold changes (± s.d.) of ubi-sites under proteasome inhibition (MG-132 treatment) binned into functional score ranges. Numbers of quantified ubi-sites per bin are shown. Final values were obtained by normalising and averaging log2 fold-changes from multiple experiments (see Methods). G) Illustration of the p53 bi-partite nuclear localization signal (NLS), where ubiquitination inhibits nuclear localization[60]. H) Functional scores for ubi-sites within NLSs compared to other ubi-sites. Numbers of unique ubi-sites per group are indicated. Examples of high-scoring NLS-embedded ubi-sites are shown. I) The detection of NLS-embedded ubi-sites with functional score > 0.65 in a whole-cell proteome or nucleus-specific proteome from *Elia et al.*[12], presented as numbers of unique sites (above barplot) and enrichment relative to all ubi-sites (barplot).
J) Conceptual workflow for estimating kinase or transcription factor (TF) activity in tumors based on relative abundance of kinase/TF targets. K-L) Examples illustrating correlations between ubi-site abundance and K) kinase (EPHA2 K583) or L) TF (SMAD4 K70) activity across tumors. Pearson correlation coefficients (R), linear regression confidence intervals (95%), and ubi-site functional scores (plot titles) are shown.

Highly functional ubi-sites perform diverse regulatory roles

Having confirmed the biological relevance of our functional score, we turned to investigate the molecular functions and regulatory properties it captures. A recent landmark proteomics study measured ubi-site turnover and occupancy[16]; sites with a higher functional score exhibited faster turnover rates - consistent with dynamic regulatory roles - while no clear trend was observed for occupancy (Fig. S6a-b). Mirroring our earlier findings on ubi-site conservation, lower-scoring sites were typically up-regulated upon proteasome inhibition by MG-132, bortezomib or epoxomicin, while higher-scoring sites showed limited regulation (Fig. 4f, Fig. S6c-d). This suggests that the functionally critical ubi-sites identified by our approach are less likely to be involved in proteasome-mediated protein turnover and quality control, instead acting as regulatory switches that modulate protein function through other avenues.

One non-degradative regulatory role of ubiquitin is to modulate the subcellular localisation of proteins. Specifically, ubiquitination of nuclear localisation signals (NLSs) - which are characteristically rich in lysines - can inhibit nuclear import, exemplified by the bipartite NLS in p53[60] (Fig. 4g). We identified 264 ubi-sites in nuclear localisation signals annotated in UniProt[61]; these NLS-embedded ubi-sites were enriched in higher functional scores, supporting the ability of our functional score to identify critical regulatory sites likely to affect protein localisation (Fig. 4h, Table S8). High-scoring NLS ubi-sites sites include K24 and K34 in FABP5, which could inhibit the ability of FABP5 to deliver fatty acids to nuclear receptors, and K154 and K143 in the DNA repair proteins CDKN1A and MORF4L1, which could modulate nuclear localisation of

these proteins in response to genotoxic stress (Fig. 4h). In a subcellular ubiquitin proteomics dataset we found that NLS-embedded ubi-sites with high functional scores (score > 0.65) were substantially less likely to be detected in the nuclear extract compared to whole-cell lysate[12], providing evidence that ubiquitination of these sites may repress nuclear localisation (Fig. 4i).

Finally, we assessed whether our functional score can identify ubi-sites modulating protein activity. We estimated the activities of protein kinases and transcription factors (TFs) based on the relative abundance of known targets in phosphoproteomics and transcriptomics data from 85 lung squamous cell carcinoma tumours in the CPTAC consortium (Fig. 4j), which can be directly compared to ubiquitinomes measured on the same samples[62]. From 255 ubi-sites in kinases and 144 ubi-sites in TFs, we filtered for ubi-sites associated with the activity of their parent protein, controlling for associations confounded by protein abundance and sites regulating activity indirectly through protein degradation (see Methods). Although this filtering produced only six ubi-sites, the majority had high functional scores, indicating these associations may result from ubiquitin-dependent activity regulation (Table S8). For instance, K583 on EPHA2 is positively associated with the kinase's activity (Fig. 4k). This site is located in the auto-inhibitory juxtamembrane domain which is typically relieved by tyrosine phosphorylation[63], and it is possible that K583 ubiquitination may play a similar activating role. By contrast, K70 in the TF SMAD4 is negatively associated with its activity; this site resides in the MH1 domain which binds DNA, which may underlie its potential inhibitory effect on SMAD4-mediated transcription (Fig. 4l). Overall, our functional score prioritises ubi-sites important for organismal fitness that can regulate protein function through diverse avenues.

Exploring cellular roles of functional ubi-sites with yeast chemical genomics

We next sought to characterise the cellular roles of high-scoring ubi-sites. We employed our established workflow for PTM-centric chemical genomics, exploiting the ease of genetic manipulation in yeast[64]. First, we selected 8 high-scoring human ubi-sites from proteins performing diverse functions, and mutated the 9 orthologous yeast lysine sites to arginine using CRISPR-Cas9 (Fig. 5a-b, see Methods). We then screened site mutants for fitness effects across 41 stress conditions alongside knock-out or knock-down mutations of the corresponding gene, providing rich phenotypic information on ubi-site mutant effects (Fig. 5a, Table S9). Where possible we also mutated nearby lysines that were not conserved in humans as negative controls (Table S9).

Three of the nine ubi-site mutations impacted cell growth under one or more stress conditions (Fig. 5b). Phenotypes of ubi-site mutants diverged from those observed upon knockout or knockdown of the corresponding proteins, suggesting these mutations disrupt regulatory functions of ubi-sites rather than simply destabilising the entire protein. For example, most growth effects caused by mutating K371 in the glycolytic enzyme PGI1 were in the opposite direction to PGI1 knockout, consistent with an inhibitory function (Fig. 5c). PGI1 is enzymatically active as a homodimer[65] and K371 is present at the dimer interface (Fig. 5d), suggesting that ubiquitination of K371 may block dimerisation and inhibit enzymatic activity. Similarly, arginine mutation of K58 in TUB2 (β-tubulin) conveyed resistance to nocodazole and benzimidazole

treatment whereas TUB2 knock-down resulted in sensitivity, and K58 lies at the interface between adjacent TUB2 monomers within the microtubule, also consistent with an inhibitory role for this site (Fig. 5e-f). Importantly, we cannot exclude the possibility that ubiquitination targets these proteins for degradation, and the reversal of gene-level phenotypes in arginine mutants results from increased protein abundance.

Finally, mutating K122 in the GPTase YPT1 caused several fitness effects including resistance to the ER stress-inducing agent tunicamycin, consistent with YPT1's role in combating ER stress through unfolded protein response (Fig. 5g)[66]. Mechanistically, ubiquitination of K122 may affect nucleotide exchange, since this site lies in the conserved GTPase ubi-site hotspot discussed previously (Fig. 2i, 5h). Finally, it is pertinent that mutation of K126 - a nearby lysine not conserved in humans - exhibited no phenotype under any tested condition, confirming the specificity of our approach (Fig, 5g). Ultimately, these results demonstrate that our functional score can identify ubi-sites regulating critical cellular processes through diverse regulatory mechanisms.

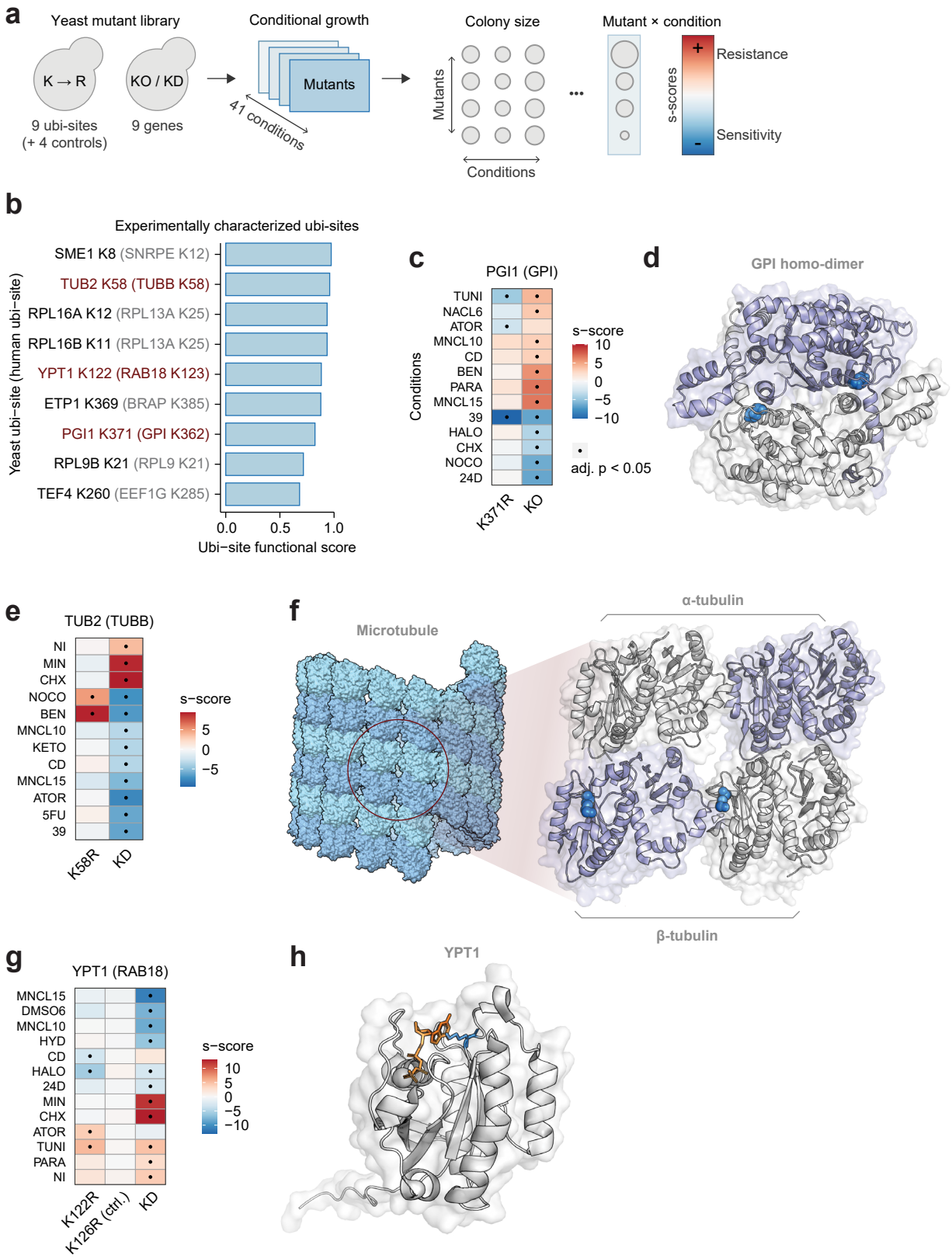

Figure 5: Exploring cellular roles of functional ubi-sites with yeast chemical genomics

A) Experimental workflow for characterizing human ubi-site orthologs in yeast. Ubi-site lysine-to-arginine mutants, negative control mutants, and knockout/knockdown mutants of the corresponding genes were screened for conditional growth under 41 stress conditions. Growth phenotypes were quantified as s-scores, representing sensitivity or resistance under a given condition relative to other genotypes. B) Yeast ubi-sites (black text) and their human ubi-site orthologs (grey text in brackets) selected for experiments. Human ubi-site functional scores are indicated. Ubi-site mutants with stress-specific growth phenotypes are indicated in red. C) Significant growth phenotypes (adj. $p < 0.05$) for K371R or knock-out of yeast PGI1, corresponding to human GPI. D) Experimental structure of the human GPI homo-dimer (PDB: 9FHF[67]). The orthologous site of K371 is indicated in blue.
E) Significant growth phenotypes (adj. $p < 0.05$) for K58R or knock-down of yeast TUB2, corresponding to human TUBB (β-tubulin). F) Left: Structural model of the microtubule with α-subunits in light-blue and β-subunits in dark-blue (PDB: 3J2U[68]). Right: Experimental structure of adjacent α-tubulin and β-tubulin subunits, with K58 indicated in blue (PDB: 6DPU[69]). G) Significant growth phenotypes (adj. $p < 0.05$) for K122R, K126R (a negative control), or knock-out of yeast YPT1, corresponding to human RAB18. H) Experimental structure of yeast YPT1 (PDB: 1YZN[70]). K122 is indicated in blue and the GTP analog GppNHp in orange.

Ubiquitination of ELAVL1 K320 disrupts RNA binding

We next aimed to leverage our functional score to uncover mechanistic insight into unstudied ubi-sites. We identified multiple high-scoring ubi-sites in RNA-binding protein and post-transcriptional regulator ELAVL1/HuR (Fig. 6a). K320 - the highest scoring ubi-site - resides at the interface between the third RNA recognition motif of ELAVL1 (RRM3) and 6×U RNA in a recently resolved crystal structure, suggesting that ubiquitination of K320 may modulate the interaction between ELAVL1 and RNA (Fig. 6a-b[71]). Supporting a non-degradational regulatory role, proteasome inhibition does not increase K320 ubiquitination (Fig. 6c). Furthermore, K320 ubiquitination is enhanced following UV exposure and during early Salmonella infection, implicating this site in ELAVL1's known involvement in cellular stress and inflammation[72] (Fig. 6c). These lines of evidence motivated further molecular dissection of K320 ubiquitination

While it is straightforward to abolish ubiquitination by site mutagenesis, it is challenging to attach ubiquitin to a target lysine residue in a precise manner[73]. We employed our recently established workflow UbyW (Ubiquitylation by UBE2W)[73] to site-specifically ubiquitylate the RRM3 domain of human ELAVL1 at position 320 (Fig. 6d). Briefly, we introduce the non-canonical amino acid LisoK at position 320 by amber suppression utilizing the active uptake of its propeptide precursor G-LisoK into *E. coli*[74]. Co-expressed UBE2W then recognizes the introduced neo-N-terminus of LisoK and conjugates ubiquitin onto it. This protein-ubiquitin conjugate recapitulates the native lysine-ubiquitin isopeptide bond with a single leucine extension at the C-terminus of ubiquitin (Fig. 6d).

Using this workflow we successfully generated and purified ubiquitinated ELAVL1 RRM3 (K320LisoK), as confirmed by SDS-PAGE and LC-MS (Fig. 6e-f, Fig. S7a-c). We then employed fluorescence anisotropy to assess binding to 10xU RNA, representing the U-rich sequences found in ARE-containing RNA[75]. Wildtype ELAVL1 RRM3 bound readily to 10xU RNA with a Kd of 1.4 μM, reflecting published results (Fig. 6g, Kd = 0.65 μM for the 11-mer RNA

5'-AUUUUUAUUUU-3' in Ripin et al.[76]). By contrast, we observed virtually no RNA binding of ubiquitinated ELAVL1 RRM3 (K320LisoK), confirming our hypothesis that ubiquitination of this site modulates RNA binding (Fig. 6g). Overall, these results underscore the ability of our functional score to pinpoint regulatory ubiquitination events and guide elucidation of their molecular mechanisms.

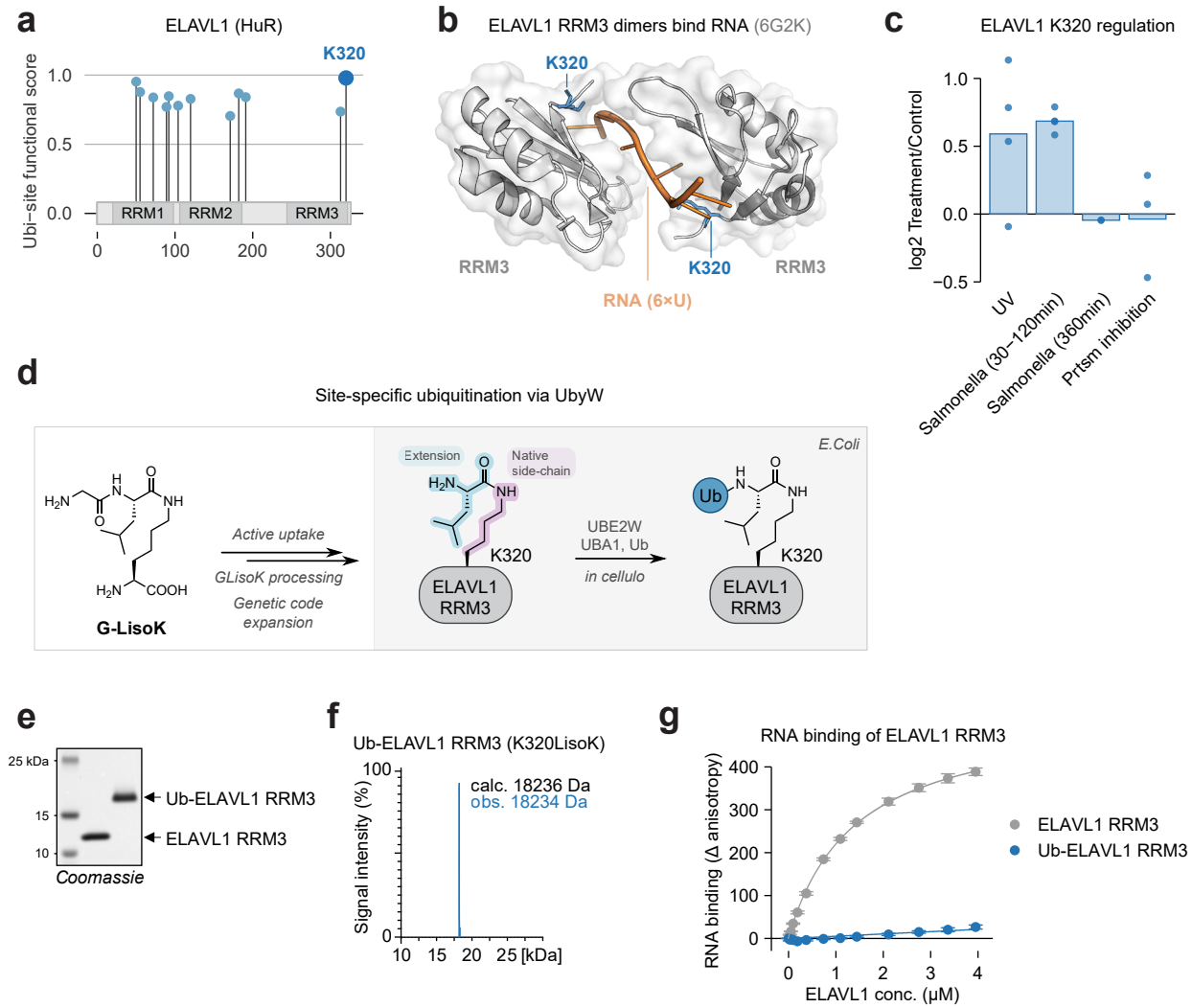

Figure 6: Ubiquitination of ELAVL1 K320 disrupts RNA binding
A) Functional scores of ubi-sites on human ELAVL1 (HuR). B) Crystal structure of two ELAVL1 RRM3 domains binding 6×U RNA (6G2K[71]). C) Log2 fold changes of ELAVL1 K320 ubiquitination in ubiquitin proteomics data. D) Workflow for generating site-specifically ubiquitinated ELAVL1 RRM3 by UbyW[77]. E) Coomassie-stained SDS-PAGE of purified ELAVL1 constructs: Wildtype ELAVL1 RRM3 and ubiquitinated ELAVL1 RRM3 (K320LisoK). The full blot is shown in Fig. S7. F) LC-MS measurement of purified Ub-ELAVL1 RRM3. The non-deconvoluted spectrum is shown in Fig. S7. G) The binding of ELAVL1 RRM3 constructs to 10×U RNA was assessed by fluorescence anisotropy.

Discussion

Ubiquitination modulates protein function through diverse avenues in order to regulate a broad spectrum of critical cellular processes. However, 99% of catalogued human ubi-sites remain functionally uncharacterised[1,4,6,17]. Here, we have presented a proteome-wide functional analysis of the human ubiquitinome, comprising the generation of a high-confidence reference atlas of human ubi-sites and systems-wide analyses of ubi-site conservation and regulation, and culminating in a machine-learning functional score that prioritises ubi-sites pivotal for cellular function. We exemplified the utility of this score by exploring genetic disease susceptibilities linked to ubiquitination, identifying ubi-sites involved in diverse regulatory mechanisms, and validating a subset of high-scoring sites experimentally.

We demonstrated that more evolutionarily conserved ubi-sites are less frequently linked to proteasomal degradation. This supports the notion that site specificity is less critical in degradative contexts. Indeed, the mere presence of a degradative ubiquitin chain is sufficient for proteasome-mediated recognition and subsequent degradation, provided the protein can be efficiently unraveled into the proteasome barrel[78]. It is likely that many different positions of ubiquitination can facilitate this process, resulting in minimal selective pressure for ubiquitination at an exact location. On the other hand, the specific site of ubiquitination is likely important for signaling functions including altering a protein-protein interaction, or modulating functionalized protein regions such as a nuclear localisation signal. Importantly, our results do not imply that protein degradation and quality control are unimportant for cellular homeostasis. Rather, there is likely functional redundancy at the level of the individual ubiquitination sites involved, rendering any single site less critical in isolation.

It is important to acknowledge some limitations of our functional score. Firstly, the score was trained to predict annotated regulatory ubi-sites, which are few in number and potentially shaped by biases. For instance, researchers may be more likely to study ubi-sites that are conserved, detectable by mass spectrometry, and present in protein domains. Second, many of the proteomics datasets used in this study employed antibodies recognising the di-glycine remnant produced by tryptic digestion of lysine-conjugated ubiquitin. However, this antibody exhibits minor selectivity for surrounding amino acids[35], fails to capture N-terminally linked ubiquitin, and also captures the ubiquitin-like modifications NEDD8 and ISG15. Capture of NEDD8 and ISG15 should not have a major impact on our data given that these modifications have been estimated to represent between 2% and 6% of the di-glycine remnant pool[13,37]. Regardless, future work would benefit from employing ubiquitin-specific enrichment methods such as the UbiSite protocol[10]. Third, in constructing our human reference ubiquitinome, we prioritized data quality by focusing on high-confidence site identifications. As a result, the atlas is unlikely to capture the full landscape of ubiquitination, particularly at sites restricted to specific cell types or induced only under particular perturbations.

Another feature of ubiquitination we were unable to address is the chain architecture of ubiquitin modifications, since this cannot be resolved at site resolution by current ubiquitin proteomics methods. Ubiquitin chain architectures confer distinct functional outcomes[1,4], and it is possible that a given ubi-site performs different functions of differing importance depending on the chain

type it carries. Finally, while our analysis centred on canonical lysine ubiquitination, a recent body of work has identified and characterised the ubiquitination of non-canonical residues such as serine and threonine[79,80]. As large-scale proteomics data capturing these non-canonical ubiquitination events becomes available, our approach can be readily extended to capture this expanded landscape of ubiquitin signaling.

Overall, we have established a high-confidence reference ubiquitinome and provided a map of its evolutionary, regulatory, and functional terrain. By functionally ranking approximately 100,000 human ubi-sites, our resource opens new paths to a range of future mechanistic and functional studies across the community.

Acknowledgements

First of all, we would like to thank all researchers who deposited the ubiquitin-enriched datasets in the PRIDE database. We would like to thank Steven Gygi and Woong Kim from Harvard Medical School for their assistance in generating proteomics data. J.A.V. would like to acknowledge funding from BBSRC [grant numbers BB/S01781X/1, BB/Y513829/1], EPSRC [grant number EP/Y035984/1], Wellcome Trust [grant number 223745/Z/21/Z] and EMBL core funding. A.J. acknowledges support from BBSRC grant [BB/S017054/1], and Doctoral Training Partnership (DTP) scholarship for EB [BB/T008695/1]. K.L. would like to acknowledge funding from ETH Zurich and the European Research Council (ERC under the European Union's Horizon 2020 research and innovation programme, grant agreement no. 101003289-Ubl-tool to K.L). E.J.B acknowledges research funding provided by the NIH (R35GM148339, R01GM136994, DP2-GM119132).

Author contributions

Conceptualization: P.B., I.B.H., E.J.B. Data curation: S.W., E.B., R.S., I.B.H. Formal analysis: J.vG., I.B.H., R.S., M.S., S.W., E.B. Funding acquisition: P.B., J.A.V., A.R.J., K.L., E.J.B. Investigation: M.F., B.B., P.S., A.C.C., M.A.B., E.R.T. Supervision: P.B., J.A.V., A.R.J., K.L., E.J.B., J.W.H., M.J.M. Visualization: J.vG. Writing – original draft: J.vG. Writing – review & editing: All authors

Declaration of interests

The authors declare no competing interests.

Data availability

Proteomics data generated in this manuscript have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository[23] with the dataset identify PXD068906. The human reference ubiquitinome has been deposited to the ProteomeXchange with the dataset identify PXD068989.

Code availability

All code used to analyse data and produce figures is available on request and will be made public before publication.

Methods

Computational analysis

Unless otherwise stated, computational analysis was performed using R (v4.3.1) in VSCode (v1.99.3). Visualisations were created using the ggplot2 package (v3.4.2).

Creation of a human reference ubiquitinome

Selection of datasets

Datasets from the PRIDE database[23] were selected for data re-analysis based on the following criteria: (i) human-derived samples enriched for lysine ubiquitination; (ii) data generated using Thermo Fisher Scientific instruments; and (iii) availability of metadata, either through the original publication or by direct communication with the authors.

After preliminary curation, 27 ubiquitination-enriched datasets were identified, of which 11 met the selection criteria, including employing different cell lines (e.g., HEK293, HeLa, HCT116). To ensure comparability, only datasets employing the diGly/UbiSite enrichment methods were included[8,10]. A summary of the resulting datasets, including PRIDE accession numbers, biological samples, and key characteristics, is provided in Table S1.

Proteomics raw data processing

Raw files from each dataset were converted to mzML format using ThermoRawFileParser (version 1.3.4) and analyzed independently. Dataset PXD037009 was split into 5 subsets because of its large size and the different experimental conditions included. An initial subset of raw files from each dataset was processed using Fragpipe with an open search to identify modifications. Modifications detected in > 1% of the peptide-spectrum matches (PSMs) were retained as parameters of the search. Peptide and protein identification, including post-translational modifications (PTMs), was performed using the Comet search engine (version 2024) on a Linux-based high-performance computing cluster. Default parameters were applied, with the following exceptions: missed cleavages were set to 4, and variable modifications were set to include PTMs from the Fragpipe open search results, with oxidation of methionine and N-terminal protein acetylation (excluding N-terminal peptide acetylation) included for all datasets. For all datasets, trypsin was used as a setting for digestion. The search database consisted of the UniProt human reference proteome (one protein per gene, downloaded in April 2024) and cRAP protein sequences as contaminants database (https://www.thegpm.org/crap/, obtained April 2024). Decoys were generated using the reverse decoy method via FragPipe (https://fragpipe.nesvilab.org/).

Statistical validation of PSMs and distinct peptide sequences was conducted using PeptideProphet and iProphet from the Trans-Proteomic Pipeline (TPP, version 7.1.0[81]). High-confidence PSM matches were obtained, and PTM site localization was computed using PTMProphet (TPP), generating a unified mzidentML format file.

Post-processing

The searching result files from the TPP were processed using a custom Python script (mzidFLR; https://github.com/PGB-LIV/mzidFLR), as previously described[82] and also applied in prior PTMeXchange projects[83]. First, a global false discovery rate (FDR) was calculated at the PSM level at a 1% threshold to retain high-confidence matches. The results were converted to a site-based format, with PTM localization scores assigned to each ubiquitination site. Contaminants, decoy hits, and non-ubiquitinated PSMs were excluded, retaining only ubiquitinated sites for downstream analysis.

Second, to estimate the probability of correct PTM localization, the PTM localization probability (from PTMProphet) was multiplied by the PSM probability (from PeptideProphet). Redundancy in PSMs site-based evidence was addressed by collapsing the data to the peptidoform level using a binomial adjustment[32]: the probability of a site being ubiquitinated was calculated by comparing the number of ubiquitination events to the total evidence for that site across all PSMs. False localization rates (FLRs) were estimated by introducing decoy alanine residues, yielding a 1% false discovery rate (FDR) PSM file and a PTM site-specific FLR file for quality control (QC) analysis for each individual dataset. In the case where a data set was analyzed in experimental blocks, the results were collapsed by retaining the evidence with the lowest FLR for a peptidoform. Peptide C-terminal ubiquitinated Lys were removed since it is expected that trypsin cannot cleave those. To combine results across datasets, a meta-analysis approach was employed to control FLR inflation. PTM sites were classified into Gold, Silver, and Bronze categories, with thresholds set as follows: Gold (≥ 2 datasets with FLR<1%), Silver (1 dataset with FLR<1%) and Bronze (≥ 1 datasets with FLR<5% and no datasets with FLR < 1%). When a peptide could be mapped to multiple protein sequences, the first protein in alphabetical order was selected to avoid redundancy.

Data set deposition

To ensure that the reanalyzed data meets FAIR guidelines[34], the outputs have been made available in PRIDE with the identifier PXD068989.

GO enrichment of highly ubiquitinated proteins

Gene ontology (GO) terms were extracted using UniProt protein accessions and the R packages org.Hs.eg.db (v3.17.0) and annotationDBI (v1.62.2). GO term enrichment was performed by one-side Fisher's exact test on proteins with more than 10 ubi-sites, with the background as all proteins with at least one ubi-site. Only GO terms containing at least three process-regulating ubi-sites were tested. P-values were adjusted for within each ontology by the Benjamini-Hochberg procedure.

PhosphositePlus annotations of regulatory ubi-sites

PhosphositePlus annotations of regulatory ubi-sites were downloaded (25 November 2024[17]). Ubiquitin site positions were corrected by manually aligned surrounding sequences into the respective protein sequence in the UniProt human proteome (19 February 2025). The “ON FUNCTION” column was used to assign proteins to degradative functions (“protein degradation”) or other functions.

Ubiquitin proteomics experiments

For unpublished proteomics datasets, experiments were performed as described in Table S3. Mammalian cells were cultured as previously described[55]. Preparation of cell extracts for diGly-peptide immunoaffinity enrichment and MS/MS analysis were performed as previously described[13,55]. DiGly-peptides were quantified either by SILAC or label-free quantification as indicated in Table S3.

Ubi-site conservation analysis

Protein multiple sequence alignments were extracted from EnsemblCompara GeneTrees[84] (release 86) and subsetted for *H. sapiens, M. musculus, R. norvegicus, E. caballus, G. gallus, X. tropicalis, Danio rerio, D. melanogaster, C. elegans,* and *S. cerevisiae*. Ubi-site identifications from *H. sapiens, M. musculus, R. norvegicus, G. gallus, D. melanogaster, C. elegans,* and *S. cerevisiae* were mapped onto multiple sequence alignments. The lysine conservation at each alignment position was defined as the percentage of non-human residues that were lysine. The conservation level of each human ubi-site was scored using the following criteria:

Level 1: No ubiquitination in species apart from *H. sapiens*, lysine conservation < 10%.

Level 2: No ubiquitination in species apart from *H. sapiens*, lysine conservation > 10% and lysine conservation ≤ 50%.

Level 3: No ubiquitination in species apart from *H. sapiens*, lysine conservation > 50%.

Level 4: Ubiquitination detected in only two groups of species: 1) *H. sapiens* and 2) either *M. musculus or R. norvegicus*. Any lysine conservation level was allowed.

Level 5: Ubiquitination detected in three groups of species: 1) *H. sapiens*, 2) *M. musculus or R. norvegicus*, and 3) *G. gallus, D. melanogaster, C. elegans,* or *S. cerevisiae*. Any lysine conservation level was allowed.

Ubi-site hotspot analysis

Ubi-site hotspots were identified using our pan-species ubiquitin proteomics data, including sites from *A. thaliana* that were not used in conservation analysis (Table S3 and Table S4). First, InterPro protein domains were extracted from UniProt protein identifiers using an API call (https://www.ebi.ac.uk/interpro/api/entry/interpro/protein/uniprot). Results were filtered for domains with a Pfam domain identifier. All domain sequences containing at least one ubi-site in any species were extracted. Domains with at least 15 instances and 50 ubi-sites were taken forward for further analysis. Domain sequences were aligned using MAFFT (v7.526, using the G-INS-i option[85]).

Hotspot analysis was then performed using domain alignments and ubi-site identifications as previously described[21]. Briefly, we first counted the observed ubi-sites at each position in a given domain alignment using a rolling window with a fixed size of 5 positions. To generate a null

background model we randomly selected lysine residues within the alignment. Permutations were repeated 100 times and for each position in the alignment an expected median and standard deviation of phosphorylation were calculated. The observed values were converted to *z*-scores using the permutation information and then to *p*-values using the survival function of the normal distribution. Only enrichment over random was considered, and Bonferroni correction was used to account for multiple testing globally. To avoid identification of hotspots with a low effect size, a cut-off of an average of 2 ubi-sites per position was used. Finally, contiguous positions were merged to identify domain regions of interest with added ±2 positions on either side that were defined as phosphorylation hotspot regions. For each domain with a significant hotspot, the InterPro representative PDB structure was extracted/downloaded. The corresponding domain sequence was aligned to all other domain sequences using MAFFT (v7.526, using the G-INS-i option[85]). Hotspot analysis results were then mapped onto the representative sequence for visualisation. Hotspot analysis was performed with python v3.13.2.

Quantitative ubiquitin proteomics data

Data preparation

Public proteomics data was accessed from the appropriate publications. Within each experimental condition log2 fold changes were calculated and averaged across different peptides corresponding to the same ubi-site, and across technical and biological replicates. Conditions with fewer than 500 quantified ubi-sites were filtered out.

Data analysis

Pairwise correlations between perturbation conditions were calculated using ubi-site log2-fold changes and Pearson's correlation. The correlation heatmap was visualised using the R package ComplexHeatmap[86] (2.16.0). Hierarchical clustering was performed on the heatmap using Pearson's correlation as the distance metric. The clustering dendrogram was cut to create four clusters.

Log2-fold changes from different proteasome inhibitor experiments were combined to increase coverage. Specifically, within each proteasome inhibitor compound (MG-312, Bortezomib, and Epoxomicin) conditions were normalized by subtracting the condition median, dividing by the condition interquartile range, multiplying by the median of condition interquartile ranges, and adding the median of condition medians. Normalized log2-fold changes were then averaged across conditions, ignoring missing values.

Machine learning for ubi-site functional scores

Feature collection

Conservation level: Conservation levels were converted into integers (1-5). Missing values from unaligned sites were imputed as the overall mean.

In hotspot: The presence of sites in ubi-site hotspots was encoded as a one-hot.

Gold identification confidence: Sites with Gold identification confidence were labeled with a one-hot encoding.

 # regulations: Ubi-sites were defined as regulated in a given ubiquitin proteomics condition if they occurred in the top 5% or bottom 5% of log2-fold change values. The number of regulated conditions was counted at the site level, excluding proteasome inhibition conditions due to their distinct regulatory nature.

 # quantifications: The number of quantified conditions was counted at the site level, including proteasome inhibition conditions.
In domain, In region: Protein domains and regions were extracted from UniProt using the GetFamily_Domains function from the R package UniprotR (2.3.0). Ubi-sites present in domains or regions were labeled with one-hot encodings.
Disorder, helix probability, coil probability, relative surface accessibility: NetSurfP3.0 was run on protein sequences to predict structural properties (v3.0[87]). Sequences larger than 5,000 residues could not be run and were hence excluded.
Protein length: Protein sequence length was calculated using canonical UniProt protein sequences.
At interface: Ubi-sites present at computationally predicted interfaces for 108,931 interacting protein pairs were labeled with a one-hot encoding (DockQ > 0.23[88]).
Site contains other PTMs: Ubiquitinated lysines annotated to contain other PTMs in PhosphositePlus (25 November 202417) were labeled with a one-hot encoding.
 # nearby PTMs: The number of PTMs recorded in (25 November 2024[17]) in a 21-residue window around each ubi-site was counted.
AlphaMissense pathogenicity: Pathogenicity scores were calculated for lysine to arginine mutations of each ubi-site using AlphaMissense[89].

Feature processing
Several features were log-transformed to achieve more gaussian-like distributions (AlphaMissense pathogenicity, # nearby PTMs, # regulations, # quantifications, protein length). The subset of these features containing 0-values were transformed using the log(1 + x) function to avoid producing missing values. All features were centered and scaled using the preProcess function from the R package caret (version 6.0.94, method = c("center", "scale")). Apart from ubi-site conservation levels, no missing values were imputed. Ubi-sites with missing values for any other feature were removed.

Model training and evaluation
The positive class was defined as all sites annotated to perform non-degradation functions in PhosphositePlus, and the negative class as all sites without an annotated function. The protein XRCC5 (UniProt: P13010) was removed from training data to mitigate bias. In particular, this protein contained 19 ubi-sites annotated to regulate DNA repair, which was by far the most number of regulatory sites per protein. Given the number of positive cases (n = 213), this risks overfitting to general properties of XRCC5 and its ubi-sites. Furthermore, the evidence underlying these annotations comes from experiments on a protein construct harbouring lysine-to-arginine mutations of all 19 sites at once[90], which is weak evidence for the functional relevance of any one site in isolation.

Models were trained using the caret package in R. The model architectures used and their corresponding arguments for the caret "train" function were logistic regression (method = "glm", family = "binomial"), gradient-boosting machine (method "gbm") and random forest (method = "ranger"). In order to estimate the generalization error of these different architectures, models were trained and evaluated through train-test splits. Specifically, the data were divided into five

equal folds; in each iteration, one fold was held out as the test set while the remaining four (80%) were used for training, yielding five distinct train–test splits. The createFolds function from caret was used to preserve the class balance. To prevent data leakage due to the use of protein-level features, ubi-sites belonging to the same protein were grouped together during fold separation. During each of the five train-test iterations, models were trained on the training data using a subsampling approach. In particular, five random samples were drawn from the negative class at 10× the size of the positive class, to mitigate the impact of the class imbalance and to speed up training. Within each subsample, all positive class instances were retained, and a model was training with 5-fold cross-validation for hyperparameter optimization. The optimized model was then applied to the corresponding test set, and model accuracy was measured using the area under the ROC curve (ROC AUC). The generalization error of each classifier architecture was then estimated by aggregating the 25 ROC AUC values arising from five negative subsamples within 5 train-test splits.

Generating predictions

To generate predictions, 25 train-test splits were generated by repeating the train-test split procedure described above five times. A logistic regression model was trained on each training set. Negative class subsampling was not used due to the speed of training logistic regression models; instead all samples were weighted inversely proportional to the size of the positive/negative class in order to mitigate the impact of the class imbalance. Each model was applied to all ubi-sites, generating prediction probabilities for the positive class. The median probability was calculated across the 25 models for each site to generate final functional scores.

Transcription factor and kinase activity estimation from CPTAC data

Log2-transformed and normalized ubiquitin proteomics data were downloaded from the original publication (Table S3L[62]). The remaining CPTAC data were extracted using the cptac python package (cptac v1.5.14, python v3.12.4). The following pipelines were used: phosphoproteomics - bcm; proteomics - bcm; transcriptomics - bcm; whole-exome sequencing - harmonized. Non-tumour samples were removed. For phosphoproteomics, proteomics, and transcriptomics, data were transformed into log2 fold changes by subtracting the median within each tumour type for each quantified phosphosite/protein/transcript.

Transcription factor targets were extracted from the DoRothEa database (R package dorothea[91] v1.12.0, confidence scores A-C). Kinase substrates were downloaded from PhosphositePlus (25 November 2024[17]). Kinase and transcription factor activities were estimated using the VIPER algorithm (viper() function from the R package viper[92] v1.34.0) with minsize = 5, adaptive.size = FALSE, and eset.filter = F.

Ubi-sites associated with the activity of their parent protein were identified, applying a stringent multi-step procedure to exclude associations confounded by protein abundance and sites possibly regulating activity indirectly through protein degradation. In particular, the following steps were applied:

1. Retain ubi-sites significantly correlated with parent protein activity (Pearson's correlation, Benjamini-hochberg adjusted p-value < 0.05).

2. Remove ubi-sites up-regulated upon proteasome inhibition (mean log2FC under Bortezomib and MG-132 treatment > 0),
3. For ubi-sites positively correlated with parent protein activity, regress protein abundance out of ubi-site abundance and protein activity using a linear model ("lm" function from the R package "stats" (version 4.5.1)). Repeat correlation between ubi-site and protein activity. Remove ubi-sites with unadjusted p-value ≥ 0.05.

Nuclear localisation sequences

Nuclear localisation signals were extracted from UniProt using the GetFamily_Domains function from the R package UniprotR (2.3.0).

Yeast chemical genomics

Overview of lysine point-mutant generation

Lysine point-mutant strains were generated using CRISPR-Cas9 in the yeast strain DHY214[93] (a kind gift from Lars M. Steinmetz). Two confirmed mutant strains were used per point mutation to mitigate the risk of an off-target mutation causing the phenotype of interest, improving the reliability of the screen results. Point mutants were generated following a protocol from the Ellis lab (https://benchling.com/pub/ellis-crispr-tools), described in more detail below.

sgRNA design and cloning

Guide RNAs (gRNA) were designed to target sequences within 10 bp of an NGG or CCN PAM site. gRNAs are described in table S9. Forward and reverse oligonucleotides for gRNAs were ordered, phosphorylated using T4 PNK at 37 °C for 1 h, and annealed by heat denaturation and gradual cooling. Annealed oligonucleotides were ligated into BsmBI-digested entry vector pWS082 (a kind gift from Tom Ellis) using T7 ligase. Constructs were transformed into *E. coli* NEB 10β, and plasmids were recovered from GFP-negative colonies and confirmed by PCR.

Donor DNA construction

Donor fragments with 50 bp homology arms flanking the predicted Cas9 cleavage site (3 bp upstream of the PAM) were generated as overlapping oligonucleotides (20 bp overlap). Donor DNA sequences are described in table S9. Donors were assembled by PCR with Phire DNA polymerase, followed by ethanol precipitation and elution.

Plasmid preparation

sgRNA plasmids were digested with EcoRV and pooled when multiple guides were used. Cas9-sgRNA gap repair plasmids (pWS173, a kind gift from Tom Ellis) were linearized with BsmBI, gel-purified, and adjusted to 100 ng/µl.

Yeast transformation

Yeast transformations were performed as previously described with some modifications[64]. First, DHY214 yeast cells were grown to exponential phase, harvested by centrifugation and washed with LiAc 100 mM, 1× TE pH 8.0 buffer to produce chemically competent yeast cells. DNA was added to cells (100 ng linearized Cas9-sgRNA gap repair plasmid, 200 ng digested sgRNA plasmid, and 3-5 µg donor DNA) followed by transformation mix: 100 µl of 50% PEG 3350, 15 µl

of 1 M LiAc, 20 µl of single-stranded carrier DNA from salmon sperm and 30 µl of competent cells. The resulting mix was heat-shocked at 42 °C for 40 min with shaking at 600 rpm in a thermomixer. Outgrowth was performed in 1.5 ml tubes in YPAD medium at 30 °C, shaking at 1500 rpm on a thermomixer for ~4 h, after which colonies were plated onto YPD + NAT (100 µg/mL Nourseothricin). 4-8 colonies were sequenced to select for successful editing.

Chemical genomics screening

The BY4741 *MATa* haploid KO library, DaMP knock-down library[94], and the lysine point-mutant library were maintained in YPAD + G418, YPAD + G418 and YPAD, respectively, before screening in 384 colony format. All strains used in this study are described in Table S9 and were maintained on the appropriate selection media. chemical genomics screening was performed as previously described[64]. All chemical genomics growth conditions are described in Table S9. Colony imaging and calculation of s-scores and q-values were performed as previously described[64]. For downstream analysis and visualisation, s-scores and q-values were collapsed across replicates and timepoints by calculating the mean (s-scores) or geometric mean (q-values).

ELAVL1 ubiquitination experiments

General Methods

Codon optimized genes encoding ELAVL(244-326)-K320TAG and SUMO2-ELAVL(244-326)-wt were purchased as DNA Strings (Twist Bioscience) and cloned into the pBAD vectors via restriction cloning. All solvents and chemical reagents were purchased from Sigma Aldrich, Senn, Carbolution, Acros Organics, or Fisher Scientific and were used without further purification unless stated otherwise. Bolt 4-12 % Bis-Tris gradient gels (Invitrogen) were run (at 165 V for 40 min) using a BoltTM Mini Gel Tank system (Invitrogen). Gels were stained with Quick Coomassie Stain (Generon). PageRuler Prestained Plus Protein Ladder 10-250 kDa (ThermoFisher) was used as the protein marker. Protein and DNA concentrations were measured on a NanoPhotometer® NP60 (Implen). Protein LC-MS was performed on an Agilent Technologies 1260 Infinity LC-MS system with a 6310 Quadrupole spectrometer equipped with a Phenomenex Jupiter C4 300 A LC Column (150 x 2 mm, 5 µm). The solvent system consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). HPLC-purified 5'-FAM labeled 10xU RNA was obtained from Microsynth. Fluorescence anisotropy measurements were performed on a Jasco Fluorescence Spectrometer FP-8350 equipped with polarizers (Jasco). GLisoK was prepared as previously described[74].

Expression and purification of ELAVL-wt

Chemically competent E. coli K12 cells were transformed with pBAD H6-SUMO-ELAVL-wt (encoding for H6-SUMO-SGSG-ELAVL(244-326) wt (see plasmids in Table S10 and methods section "Protein sequences")). After recovery with 1 mL of SOC medium for 1 h at 37 °C, the cells were cultured overnight in 50 mL of 2× YT medium supplemented with ampicillin (100 µg/mL) at 37 °C, 200 rpm. The overnight culture was diluted to an OD600 of 0.05 in 200 mL of fresh autoinduction medium[95] supplemented with ampicillin (50 µg/mL) and cultured at 37 °C, 200 rpm overnight. The cells were harvested by centrifugation (4000 × g, 20 min, 4 °C) and resuspended in 15 mL of lysis buffer (20 mM Tris pH 7.5, 300 mM NaCl, 30 mM imidazole,

0.175 mg/mL PMSF). The cell suspension was incubated on ice for 30 min and sonicated with cooling in an ice-water bath. The lysed cells were centrifuged (14,000 × g, 20 min, 4 °C), the cleared lysate added to Ni Sepharose 6 Fast Flow (Cytiva) (0.5 mL of slurry per 1 L of culture) and the mixture was incubated with agitation for 1 h at 4 °C. After incubation, the mixture was transferred to an empty plastic column and washed with 10 CV (column volumes) of wash buffer (20 mM Tris pH 7.5, 300 mM NaCl, 30 mM imidazole). On beads SUMO-cleavage was performed by addition of SUMO protease (Sigma-Aldrich, Cat. No. SAE0067) in wash buffer supplemented with 1 mM TCEP, followed by incubation at 4 °C for 3 h. The flow through was collected and applied to size-exclusion chromatography (SEC) using a Superdex Increase 75 10/300 (GE Healthcare) with SEC buffer (20 mM Tris pH 7.5, 100 mM NaCl and 1 mM DTT). Fractions containing the ELAVL wt were pooled together and concentrated using Amicon centrifugal filter units with a 3 kDa MWCO (Millipore). Protein concentration was calculated from the measured A280 absorption (extinction coefficients were calculated with ProtParam (https://web.expasy.org/protparam/)).

Expression and purification of site-specifically ubiquitylated Ub-ELAVL(K320LisoK) via a UBE2W-based reconstituted ubiquitylation cascade in E. coli

Chemically competent E. coli K12 cells were cotransformed with pBAD H6-TEV-ELAVL K320TAG / UBE2W (encoding for H6-TEV-GS-ELAVL(244-326)-K320TAG and UBE2W (D. rerio)) and pEVOL Ub/E1/Ma aaRS/tRNA (encoding for ubiquitin, UBA1, Ma “IP” aaRS and Ma tRNA (see plasmids in Table S10 and methods section “Protein sequences”)). After recovery with 1 mL of SOC medium for 1 h at 37 °C, the cells were cultured overnight in 50 mL of 2× YT medium supplemented with ampicillin (100 µg/mL) and chloramphenicol (50 µg/mL) at 37 °C, 200 rpm. The overnight culture was diluted to an OD600 of 0.05 in 200 mL of fresh autoinduction medium[95] supplemented with ampicillin (50 µg/mL) and chloramphenicol (25 µg/mL) and GLisoK (500 µM) and cultured at 37 °C, 200 rpm overnight. The cells were harvested by centrifugation (4000 × g, 20 min, 4 °C) and resuspended in 15 mL of lysis buffer (20 mM Tris pH 7.5, 300 mM NaCl, 30 mM imidazole, 0.175 mg/mL PMSF). The cell suspension was incubated on ice for 30 min and sonicated with cooling in an ice-water bath. The lysed cells were centrifuged (14,000 × g, 20 min, 4 °C), the cleared lysate added to Ni Sepharose 6 Fast Flow (Cytiva) (0.5 mL of slurry per 1 L of culture) and the mixture was incubated with agitation for 1 h at 4 °C. After incubation, the mixture was transferred to an empty plastic column and washed with 10 CV (column volumes) of wash buffer (20 mM Tris pH 7.5, 300 mM NaCl, 30 mM imidazole) followed by elution in 150 µL fractions with wash buffer supplemented with 300 mM imidazole. Afterwards, TEV Protease (Sigma-Aldrich, Cat. No. T4455) was added to the elution and incubated at 4 °C for 3 h. TEV Protease was removed by reverse NiNTA purification followed by SEC using a Superdex Increase 75 10/300 (GE Healthcare) with a SEC buffer (20 mM Tris pH 7.5, 100 mM NaCl and 1 mM DTT) to separate ubiquitylated ELAVL from unmodified ELAVL. Fractions containing Ub-ELAVL(K320LisoK) were pooled together and concentrated using Amicon centrifugal filter units with a 10 kDa MWCO (Millipore). Protein concentration was calculated from the measured A280 absorption (extinction coefficients were calculated with ProtParam (https://web.expasy.org/protparam/)).

Kd determination via fluorescence anisotropy

Fluorescence anisotropy measurements were conducted on a Jasco Fluorescence Spectrometer FP-8350 equipped with polarizers (Jasco). Excitation and emission monochromators were set to 500 nm and 520 nm and measurements were performed at RT with a bandwidth of 5 nm. ELAVL/Ub-ELAVL(K320LisoK) were titrated to 100 nM 5'FAM labeled 10xU RNA in 20 mM Tris pH 7.5, 200 mM NaCl, 1 mM DTT and 0.01% NP40. All data processing was performed using GraphPad Prism 10 (GraphPad software). KD was determined using a single binding site model, and average values and error bars (s.d.) were calculated from three different experiments (n = 3).

Protein Sequences

H6-SUMO-SGSG-ELAVL-wt

MPGSHHHHHHGSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAK
RQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGSGSGWCIFIYNLGQDADEGIL
WQMFGPFGAVTNVKVIRDFNTNKCKGFGFVTMTNYEEAAMAIASLNGYRLGDKILQVSFKTNK
SHK

H6-TEV-GS-ELAVL-K320TAG

MPHHHHHHGENLYFQGSGWCIFIYNLGQDADEGILWQMFGPFGAVTNVKVIRDFNTNKCKGF
GFVTMTNYEEAAMAIASLNGYRLGDKILQVSF*TNKSHK

Ubiquitin

MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTL
HLVLRLRGG

UBA1

MAKNGSEADIDEGLYSRQLYVLGHEAMKRLQTSSVLVSGLRGLGVEIAKNIILGGVKAVTLHDQ
GTAQWADLSSQFYLREEDIGKNRAEVSQPRLAELNSYVPVTAYTGPLVEDFLSGFQVVVLTNTP
LEDQLRVGEFCHNRGIKLVVADTRGLFGQLFCDFGEEMILTDSNGEQPLSAMVSMVTKDNPGV
VTCLDEARHGFESGDFVSFSEVQGMVELNGNQPMEIKVLGPYTFSICDTSNFSDYIRGGIVSQV
KVPKKISFKSLVASLAEPDFVVTDFAKFSRPAQLHIGFQALHQFCAQHGRPPRPRNEEDAAELVA
LAQAVNARALPAVQQNNLDEDLIRKLAYVATGDLAPINAFIGGLAAQEVMKACSGKFMPIMQWLY
FDALECLPVDKEVLTEDKCLQRQNRYDGQVAVFGSDLQEKLGKQKYFLVGAGAIGCELLKNFA
MIGLGCGEGGEIIVTDMDTIEKSNLNRQFLFRPWDVTKLKSDTAAAAVRQMNPHIRVTSHQNRV
GPDTERIYDDDFFQNLDGVANALDNVDARMYMDRRCVYYRKPLLESGTLGTKGNVQVVIPFLT
ESYSSSQDPPEKSIPICTLKNFPNAIEHTLQWARDEFEGLFKQPAENVNQYLTDPKFVERTLRLA
GTQPLEVLEAVQRSLVLQRPQTWADCVTWACHHWHTQYSNNIRQLLHNFPPDQLTSSGAPFW
SGPKRCPHPLTFDVNNPLHLDYVMAAANLFAQTYGLTGSQDRAAVATFLQSVQVPEFTPKSGV
KIHVSDQELQSANASVDDSRLEELKATLPSPDKLPGFKMYPIDFEKDDDSNFHMDFIVAASNLR
AENYDIPSADRHKSKLIAGKIIPAIATTTAAVVGLVCLELYKVVQGHRQLDSYKNGFLNLALPFFGF
SEPLAAPRHQYYNQEWTLWDRFEVQGLQPNGEEMTLKQFLDYFKTEHKLEITMLSQGVSMLY
SFFMPAAKLKERLDQPMTEIVSRVSKRKLGRHVRALVLELCCNDESGEDVEVPYVRYTIR

Ma PylRS H227I Y288P

MMTVKYTDAQIQRLREYGNGTYEQKVFEDLASRDAAFSKEMSVASTDNEKKIKGMIANPSRHG
LTQLMNDIADALVAEGFIEVRTPIFISKDALARMTITEDKPLFKQVFWIDEKRALRPMLAPNLYSV
MRDLRDHTDGPVKIFEMGSCFRKESHSGMHLEEFTMLNLVDMGPRGDATEVLKNYISVVMKA
AGLPDYDLVQEESDVYKETIDVEINGQEVCSAAVGPIPLDAAHDVHEPWSGAGFGLERLLTIREK
YSTVKKGGASISYLNGAKIN

References

1. Komander, D. & Rape, M. The Ubiquitin Code. *Annu. Rev. Biochem.* **81**, 203–229 (2012).
2. Clague, M. J., Heride, C. & Urbé, S. The demographics of the ubiquitin system. *Trends Cell Biol.* **25**, 417–426 (2015).
3. Zheng, N. & Shabek, N. Ubiquitin Ligases: Structure, Function, and Regulation. *Annu. Rev. Biochem.* **86**, 129–157 (2017).
4. Yau, R. & Rape, M. The increasing complexity of the ubiquitin code. *Nat. Cell Biol.* **18**, 579–586 (2016).
5. Schwertman, P., Bekker-Jensen, S. & Mailand, N. Regulation of DNA double-strand break repair by ubiquitin and ubiquitin-like modifiers. *Nat. Rev. Mol. Cell Biol.* **17**, 379–394 (2016).
6. Rape, M. Ubiquitylation at the crossroads of development and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 59–70 (2018).
7. Mukhopadhyay, D. & Riezman, H. Proteasome-Independent Functions of Ubiquitin in Endocytosis and Signaling. *Science* **315**, 201–205 (2007).
8. Xu, G., Paige, J. S. & Jaffrey, S. R. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat. Biotechnol.* **28**, 868–873 (2010).
9. Li, C., Nelson, T. G., Vertegaal, A. C. O. & Thibault, P. Proteomic strategies for characterizing ubiquitin-like modifications. *Nat. Rev. Methods Primer* **1**, 53 (2021).
10. Akimov, V. *et al.* UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. *Nat. Struct. Mol. Biol.* **25**, 631–640 (2018).
11. Hansen, F. M. *et al.* Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. *Nat. Commun.* **12**, 254 (2021).
12. Elia, A. E. H. *et al.* Quantitative Proteomic Atlas of Ubiquitination and Acetylation in the DNA Damage Response. *Mol. Cell* **59**, 867–881 (2015).
13. Kim, W. *et al.* Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome. *Mol. Cell* **44**, 325–340 (2011).

14. Higgins, R. *et al.* The Unfolded Protein Response Triggers Site-Specific Regulatory Ubiquitylation of 40S Ribosomal Proteins. *Mol. Cell* **59**, 35–49 (2015).
15. Fiskin, E., Bionda, T., Dikic, I. & Behrends, C. Global Analysis of Host and Bacterial Ubiquitinome in Response to Salmonella Typhimurium Infection. *Mol. Cell* **62**, 967–981 (2016).
16. Prus, G., Satpathy, S., Weinert, B. T., Narita, T. & Choudhary, C. Global, site-resolved analysis of ubiquitylation occupancy and turnover rate reveals systems properties. *Cell* **187**, 2875-2892.e21 (2024).
17. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
18. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, eaau8645 (2019).
19. Landry, C. R., Levy, E. D. & Michnick, S. W. Weak functional constraints on phosphoproteomes. *Trends Genet.* **25**, 193–197 (2009).
20. Beltrao, P. *et al.* Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell* **150**, 413–425 (2012).
21. Strumillo, M. J. *et al.* Conserved phosphorylation hotspots in eukaryotic protein domain families. *Nat. Commun.* **10**, 1977 (2019).
22. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
23. Perez-Riverol, Y. *et al.* The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* gkae1011 (2024) doi:10.1093/nar/gkae1011.
24. Hendriks, I. A., Akimov, V., Blagoev, B. & Nielsen, M. L. MaxQuant.Live Enables Enhanced Selectivity and Identification of Peptides Modified by Endogenous SUMO and Ubiquitin. *J. Proteome Res.* **20**, 2042–2055 (2021).
25. Zecha, J. *et al.* Linking post-translational modifications and protein turnover by site-resolved

protein turnover profiling. *Nat. Commun.* **13**, 165 (2022).

26. Steger, M. *et al.* Time-resolved in vivo ubiquitinome profiling by DIA-MS reveals USP7 targets on a proteome-wide scale. *Nat. Commun.* **12**, 5399 (2021).
27. Laigle, V. *et al.* Quantitative Ubiquitylome Analysis Reveals the Specificity of RNF111/Arkadia E3 Ubiquitin Ligase for its Degradative Substrates SKI and SKIL/SnoN in TGF-β Signaling Pathway. *Mol. Cell. Proteomics MCP* **20**, 100173 (2021).
28. Trulsson, F. *et al.* Deubiquitinating enzymes and the proteasome regulate preferential sets of ubiquitin substrates. *Nat. Commun.* **13**, 2736 (2022).
29. Rusilowicz-Jones, E. V. *et al.* USP30 sets a trigger threshold for PINK1-PARKIN amplification of mitochondrial ubiquitylation. *Life Sci. Alliance* **3**, e202000768 (2020).
30. Roedig, J. *et al.* USP22 controls necroptosis by regulating receptor-interacting protein kinase 3 ubiquitination. *EMBO Rep.* **22**, e50163 (2021).
31. Heidelberger, J. B. *et al.* Proteomic profiling of VCP substrates links VCP to K6-linked ubiquitylation and c-Myc function. *EMBO Rep.* **19**, e44754 (2018).
32. Camacho, O. M., Ramsbottom, K. A., Collins, A. & Jones, A. R. Assessing Multiple Evidence Streams to Decide on Confidence for Identification of Post-Translational Modifications, within and Across Data Sets. *J. Proteome Res.* **22**, 1828–1842 (2023).
33. Huang, Q., Szklarczyk, D., Wang, M., Simonovic, M. & Von Mering, C. PaxDb 5.0: Curated Protein Quantification Data Suggests Adaptive Proteome Changes in Yeasts. *Mol. Cell. Proteomics* **22**, 100640 (2023).
34. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
35. Wagner, S. A. *et al.* Proteomic Analyses Reveal Divergent Ubiquitylation Site Patterns in Murine Tissues. *Mol. Cell. Proteomics* **11**, 1578–1585 (2012).
36. Buckley, S. M. *et al.* Regulation of pluripotency and cellular reprogramming by the ubiquitin-proteasome system. *Cell Stem Cell* **11**, 783–798 (2012).

37. Na, C. H. *et al.* Synaptic Protein Ubiquitination in Rat Brain Revealed by Antibody-based Ubiquitome Analysis. *J. Proteome Res.* **11**, 4722–4732 (2012).

38. Iesmantavicius, V., Weinert, B. T. & Choudhary, C. Convergence of ubiquitylation and phosphorylation signaling in rapamycin-treated yeast cells. *Mol. Cell. Proteomics MCP* **13**, 1979–1992 (2014).

39. Tong, Z., Kim, M.-S., Pandey, A. & Espenshade, P. J. Identification of candidate substrates for the Golgi Tul1 E3 ligase using quantitative diGly proteomics in yeast. *Mol. Cell. Proteomics MCP* **13**, 2871–2882 (2014).

40. Sap, K. A., Bezstarosti, K., Dekkers, D. H. W., Voets, O. & Demmers, J. A. A. Quantitative Proteomics Reveals Extensive Changes in the Ubiquitinome after Perturbation of the Proteasome by Targeted dsRNA-Mediated Subunit Knockdown in Drosophila. *J. Proteome Res.* **16**, 2848–2862 (2017).

41. Walton, A. *et al.* It's Time for Some 'Site'-Seeing: Novel Tools to Monitor the Ubiquitin Landscape in Arabidopsis thaliana. *Plant Cell* **28**, 6–16 (2016).

42. Dewhurst, H. M., Choudhury, S. & Torres, M. P. Structural Analysis of PTM Hotspots (SAPH-ire) – A Quantitative Informatics Method Enabling the Discovery of Novel Regulatory Elements in Protein Families. *Mol. Cell. Proteomics* **14**, 2285–2297 (2015).

43. Torres, M. P., Dewhurst, H. & Sundararaman, N. Proteome-wide Structural Analysis of PTM Hotspots Reveals Regulatory Elements Predicted to Impact Biological Function and Disease. *Mol. Cell. Proteomics* **15**, 3513–3528 (2016).

44. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **298**, 1912–1934 (2002).

45. Baker, R. *et al.* Differences in the Regulation of K-Ras and H-Ras Isoforms by Monoubiquitination. *J. Biol. Chem.* **288**, 36856–36862 (2013).

46. Wagner, S. A. *et al.* A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell. Proteomics MCP* **10**, M111.013284 (2011).

47. Emanuele, M. J. *et al.* Global identification of modular cullin-RING ligase substrates. *Cell* **147**, 459–474 (2011).

48. Udeshi, N. D. *et al.* Methods for quantification of in vivo changes in protein ubiquitination following proteasome and deubiquitinase inhibition. *Mol. Cell. Proteomics MCP* **11**, 148–159 (2012).

49. Povlsen, L. K. *et al.* Systems-wide analysis of ubiquitylation dynamics reveals a key role for PAF15 ubiquitylation in DNA-damage bypass. *Nat. Cell Biol.* **14**, 1089–1098 (2012).

50. Udeshi, N. D. *et al.* Refined preparation and use of anti-diglycine remnant (K-ε-GG) antibody enables routine quantification of 10,000s of ubiquitination sites in single proteomics experiments. *Mol. Cell. Proteomics MCP* **12**, 825–831 (2013).

51. Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat. Methods* **10**, 634–637 (2013).

52. Krönke, J. *et al.* Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301–305 (2014).

53. Theurillat, J.-P. P. *et al.* Prostate cancer. Ubiquitylome analysis identifies dysregulation of effector substrates in SPOP-mutant prostate cancer. *Science* **346**, 85–89 (2014).

54. Krönke, J. *et al.* Lenalidomide induces ubiquitination and degradation of CK1α in del(5q) MDS. *Nature* **523**, 183–188 (2015).

55. Gendron, J. M. *et al.* Using the Ubiquitin-modified Proteome to Monitor Distinct and Spatially Restricted Protein Homeostasis Dysfunction. *Mol. Cell. Proteomics* **15**, 2576–2593 (2016).

56. Hetz, C., Zhang, K. & Kaufman, R. J. Mechanisms, regulation and functions of the unfolded protein response. *Nat. Rev. Mol. Cell Biol.* **21**, 421–438 (2020).

57. Wu, S. *et al.* Ultraviolet Light Inhibits Translation through Activation of the Unfolded Protein Response Kinase PERK in the Lumen of the Endoplasmic Reticulum. *J. Biol. Chem.* **277**, 18077–18083 (2002).

58. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**,

D835–D844 (2020).

59. Shi, Z., Gao, H., Bai, X. & Yu, H. Cryo-EM structure of the human cohesin-NIPBL-DNA complex. *Science* **368**, 1454–1459 (2020).
60. Marchenko, N. D. *et al.* Stress-mediated nuclear stabilization of p53 is regulated by ubiquitination and importin-α3 binding. *Cell Death Differ.* **17**, 255–267 (2010).
61. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
62. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).
63. Liang, L.-Y., Patel, O., Janes, P. W., Murphy, J. M. & Lucet, I. S. Eph receptor signalling: from catalytic to non-catalytic functions. *Oncogene* **38**, 6567–6584 (2019).
64. Viéitez, C. *et al.* High-throughput functional characterization of protein phosphorylation sites in yeast. *Nat. Biotechnol.* **40**, 382–390 (2022).
65. Bruch, P., Schnackerz, K. D. & Gracy, R. W. Matrix-Bound Phosphoglucose Isomerase. Formation and Properties of Monomers and Hybrids. *Eur. J. Biochem.* **68**, 153–158 (1976).
66. Tsvetanova, N. G., Riordan, D. P. & Brown, P. O. The Yeast Rab GTPase Ypt1 Modulates Unfolded Protein Response Dynamics by Regulating the Stability of HAC1 RNA. *PLoS Genet.* **8**, e1002862 (2012).
67. Jónatansdóttir, Y. Y., Rolfsson, Ó. & Hjörleifsson, J. G. Human glycolysis isomerases are inhibited by weak metabolite modulators. *FEBS J.* febs.70049 (2025) doi:10.1111/febs.70049.
68. Asenjo, A. B. *et al.* Structural Model for Tubulin Recognition and Deformation by Kinesin-13 Microtubule Depolymerases. *Cell Rep.* **3**, 759–768 (2013).
69. Zhang, R., LaFrance, B. & Nogales, E. Separating the effects of nucleotide and EB binding on microtubule structure. *Proc. Natl. Acad. Sci.* **115**, (2018).
70. Eathiraj, S., Pan, X., Ritacco, C. & Lambright, D. G. Structural basis of family-wide Rab

GTPase recognition by rabenosyn-5. *Nature* **436**, 415–419 (2005).

71. Pabis, M. *et al.* HuR biological function involves RRM3-mediated dimerization and RNA binding by all three RRMs. *Nucleic Acids Res.* **47**, 1011–1029 (2019).
72. Srikantan, S. & Gorospe, M. HuR function in disease. *Front. Biosci. Landmark Ed.* **17**, 189–205 (2012).
73. Wanka, V., Fottner, M., Cigler, M. & Lang, K. Genetic Code Expansion Approaches to Decipher the Ubiquitin Code. *Chem. Rev.* **124**, 11544–11584 (2024).
74. Iype, T., Fottner, M., Böhm, P., Piedrafita, C. & Lang, K. Hijacking a bacterial membrane transporter for efficient genetic code expansion. Preprint at https://doi.org/10.1101/2025.03.06.641855 (2025).
75. Bakheet, T. ARED 3.0: the large and diverse AU-rich transcriptome. *Nucleic Acids Res.* **34**, D111–D114 (2006).
76. Ripin, N. *et al.* Molecular basis for AU-rich element recognition and dimerization by the HuR C-terminal RRM. *Proc. Natl. Acad. Sci.* **116**, 2935–2944 (2019).
77. Schnacke, P., Fottner, M., Kvasha, D., Willenborg, F. & Lang, K. Genetic Code Expansion Facilitates Programmable Ubiquitylation via UBE2W. Preprint at https://doi.org/10.1101/2025.09.22.676137 (2025).
78. Arkinson, C., Dong, K. C., Gee, C. L. & Martin, A. Mechanisms and regulation of substrate degradation by the 26S proteasome. *Nat. Rev. Mol. Cell Biol.* **26**, 104–122 (2025).
79. Lechtenberg, B. C. & Komander, D. Just how big is the ubiquitin system? *Nat. Struct. Mol. Biol.* **31**, 210–213 (2024).
80. Dikic, I. & Schulman, B. A. An expanded lexicon for the ubiquitin code. *Nat. Rev. Mol. Cell Biol.* **24**, 273–287 (2023).
81. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **22**, 615–624 (2023).
82. Ramsbottom, K. A. *et al.* Method for Independent Estimation of the False Localization Rate

for Phosphoproteomics. *J. Proteome Res.* **21**, 1603–1615 (2022).

83. Ramsbottom, K. A. *et al.* Meta-Analysis of Rice Phosphoproteomics Data to Understand Variation in Cell Signaling Across the Rice Pan-Genome. *J. Proteome Res.* **23**, 2518–2531 (2024).
84. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
85. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
86. Gu, Z. Complex heatmap visualization. *iMeta* **1**, e43 (2022).
87. Høie, M. H. *et al.* NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res.* **50**, W510–W515 (2022).
88. Jänes, J. *et al.* Predicted mechanistic impacts of human protein missense variants. Preprint at https://doi.org/10.1101/2024.05.29.596373 (2024).
89. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
90. Ishida, N. *et al.* Ubiquitylation of Ku80 by RNF126 Promotes Completion of Nonhomologous End Joining-Mediated DNA Repair. *Mol. Cell. Biol.* **37**, e00347-16 (2017).
91. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
92. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
93. Roy, K. R. *et al.* Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.* **36**, 512–520 (2018).
94. Breslow, D. K. *et al.* A comprehensive strategy enabling high-resolution functional analysis

of the yeast genome. *Nat. Methods* **5**, 711–718 (2008).

95. Muzika, M. *et al.* Chemically-defined lactose-based autoinduction medium for site-specific incorporation of non-canonical amino acids into proteins. *RSC Adv.* **8**, 25558–25567 (2018).
96. Wen, B. *et al.* DeepMVP: deep learning models trained on high-quality data accurately predict PTM sites and variant-induced alterations. *Nat. Methods* https://doi.org/10.1038/s41592-025-02797-x (2025) doi:10.1038/s41592-025-02797-x.
97. Ren, R. *et al.* Structural basis for transcription factor ZBTB7A recognition of DNA and effects of ZBTB7A somatic mutations that occur in human acute myeloid leukemia. *J. Biol. Chem.* **299**, 102885 (2023).

Supplementary figures

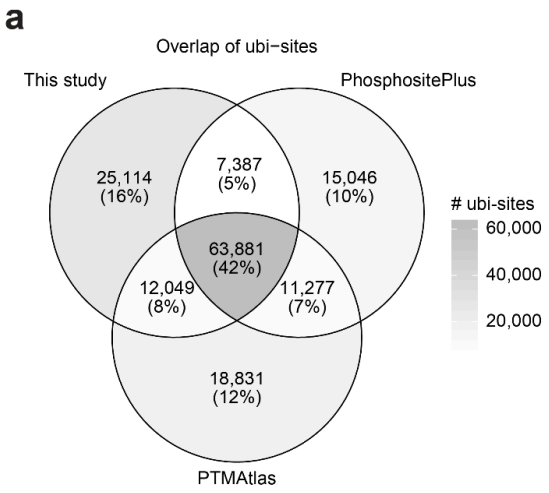

Figure S1: Additional analysis of reference ubiquitinome
A) The overlap of unique ubi-sites identified in our reference ubiquitinome, PhosphositePlus[17], and PTMAtlas[96] is shown.

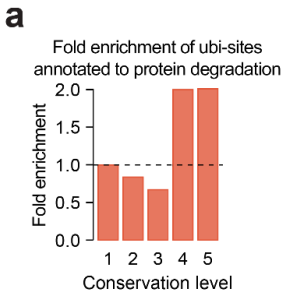

Figure S2: Additional analysis of conserved ubi-sites
A) The enrichment of ubiquitin sites annotated to perform degradation regulatory functions at each conservation level (PhosphositePlus[17]).

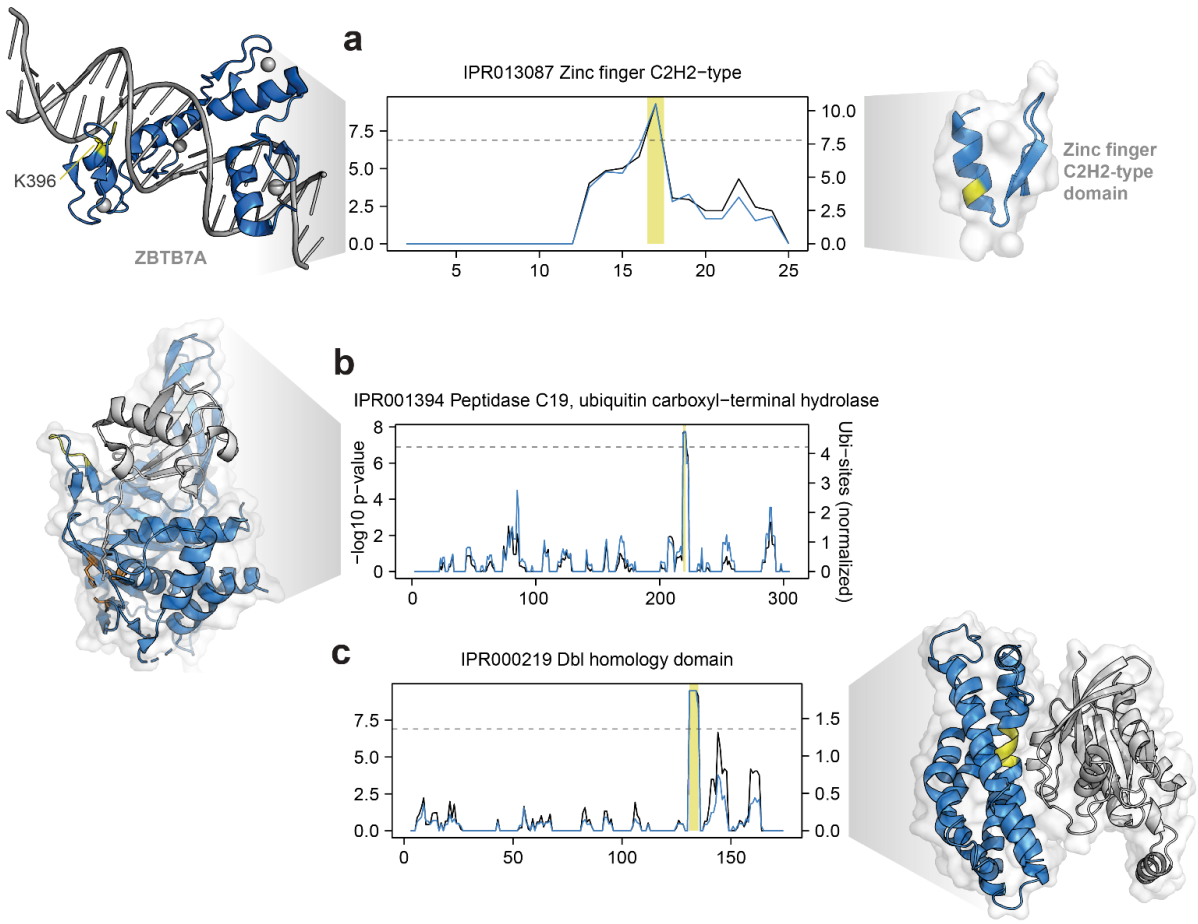

Figure S3: Additional ubi-site hotspots

A-C) Identification of ubi-site hotspots in three protein domains. The black line indicates the average number of ubi-sites observed across the domain sequence alignment within a rolling window, normalised by subtracting the number of ubi-sites expected by chance. The blue line indicates the p-value associated with the enrichment of ubi-sites at each alignment position. The horizontal line indicates a Bonferroni-corrected p-value cut-off of 0.01 (uncorrected p-value < 1.29×10^{-7}). Positions with a -log10 p-value above this cut-off and average number of phosphosites per window higher than 2 are classified as hotspot regions and highlighted with a yellow bar. Hotspot regions are mapped onto representative structures for each domain in yellow. In A) an example of a hotspot ubi-site located at the interface of a zinc finger with DNA is shown (K396 in ZBTB7A, PDB: 8E3E[97]). In B) the catalytic triad of the protease is coloured orange and a ubiquitin molecule covalently bound to the catalytic triad is coloured white. In C) the interaction partner Rac1 is coloured in white.

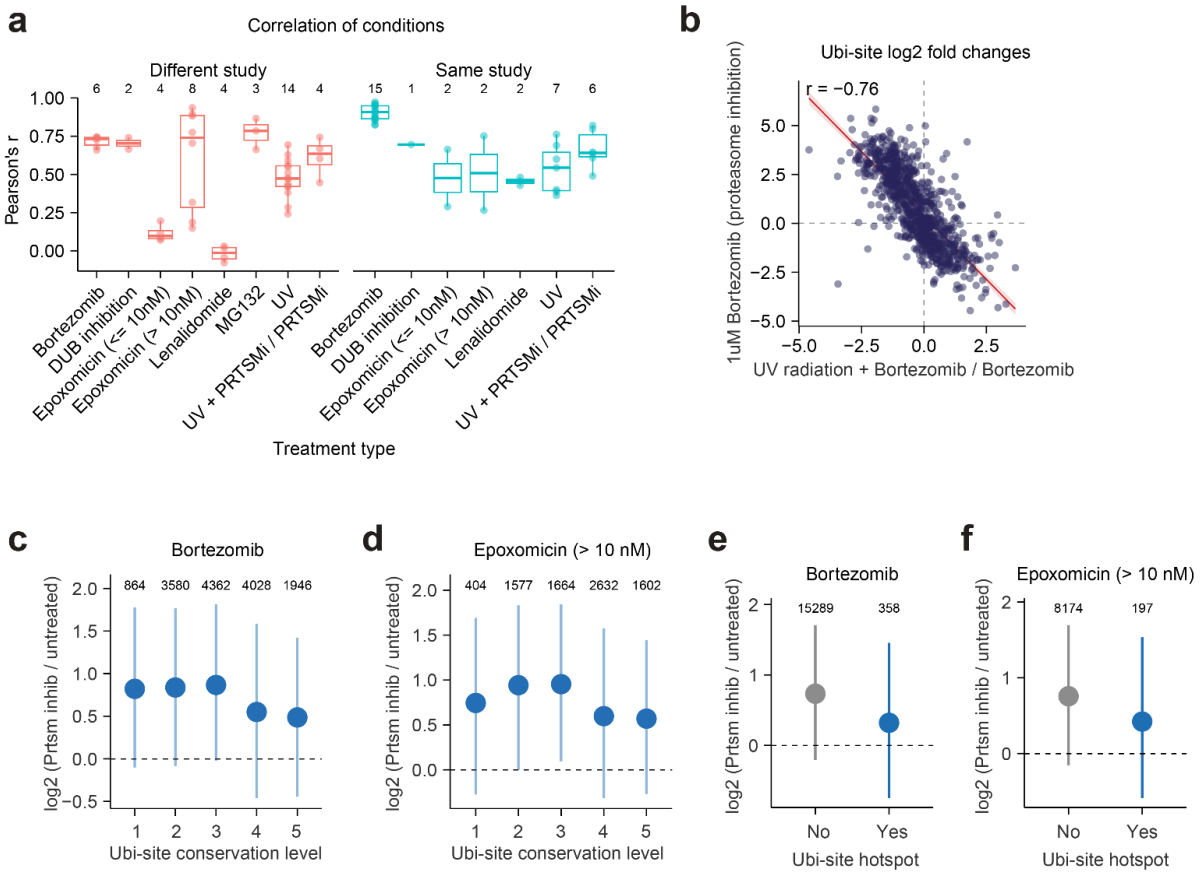

Figure S4: Additional analysis of quantitative ubiquitin proteomics data

A) Pearson's correlation was performed using ubi-site fold changes between pairs of conditions representing the same treatment. Pairs were separated based on whether the conditions came from the same study or different studies. The number of pairs in each group is indicated above boxplots. B) The correlation in ubi-site fold changes between Bortezomib treatment and UV radiation on a background of Bortezomib treatment. C-D) Mean log2-fold changes (± s.d.) of ubi-sites under proteasome inhibition (bortezomib or epoxomicin treatment), grouped by evolutionary conservation level. Final values were obtained by normalising and averaging log2 fold-changes from multiple experiments (see Methods). The number of ubi-sites in each group is indicated. Only epoxomicin treatments at > 10 nM were used since lower doses do not cause proteome-wide inhibition of protein degradation[55]. E-F) As in C-D), for sites in ubi-site domain hotspots.

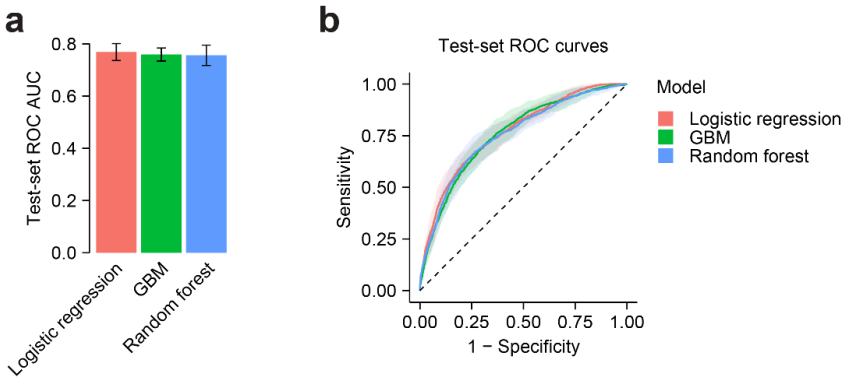

Figure S5: Training a machine-learning based ubiquitin site functional score
A-B) Three classifier architectures were trained to separate unannotated ubi-sites from ubi-sites annotated to non-degradation roles. Classifiers were trained and evaluated through a five-fold train-test split, with hyperparameter optimization performed during each training iteration via five-fold cross-validation. A) shows ROC curve AUCs on each test set (error bars indicate standard deviation) while B) shows averaged ROC curves (shading indicates standard deviation).

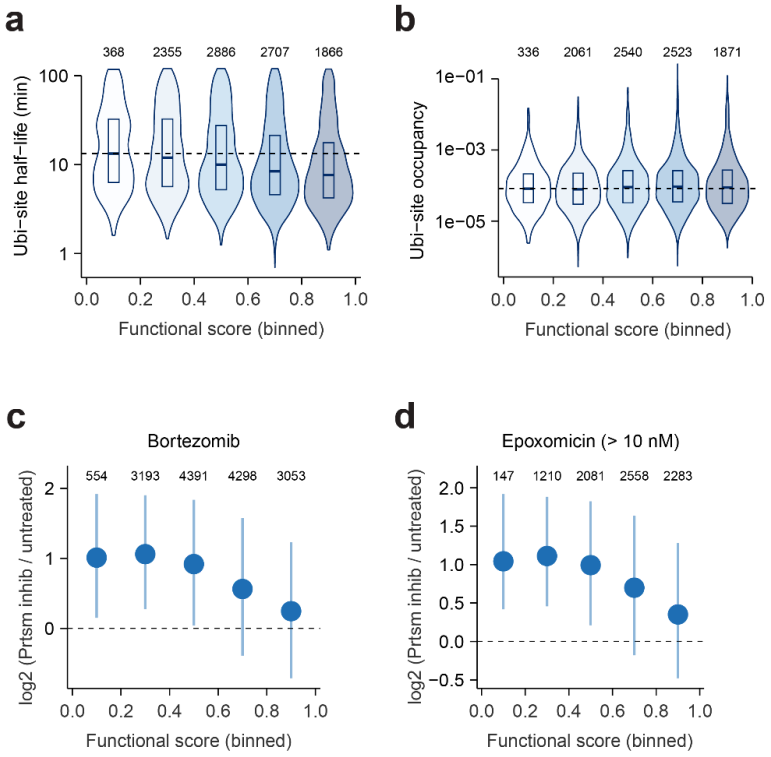

Figure S6: Additional characterisation of the ubi-site functional score
A-B) Measurements of A) ubi-site half-life and B) ubi-site occupancy in *Prus et al.*[16], binned into functional score ranges. Numbers of quantified ubi-sites per bin are shown. The median of the lowest functional score bin is shown as a dotted line. C-D) Mean log2-fold changes (± s.d.) of ubi-sites under

proteasome inhibition (bortezomib or epoxomicin treatment), binned into functional score ranges. Numbers of quantified ubi-sites per bin are shown. Final values were obtained by normalising and averaging log2 fold-changes from multiple experiments (see Methods). Only epoxomicin treatments at > 10 nM were used since lower doses do not cause proteome-wide inhibition of protein degradation[55].

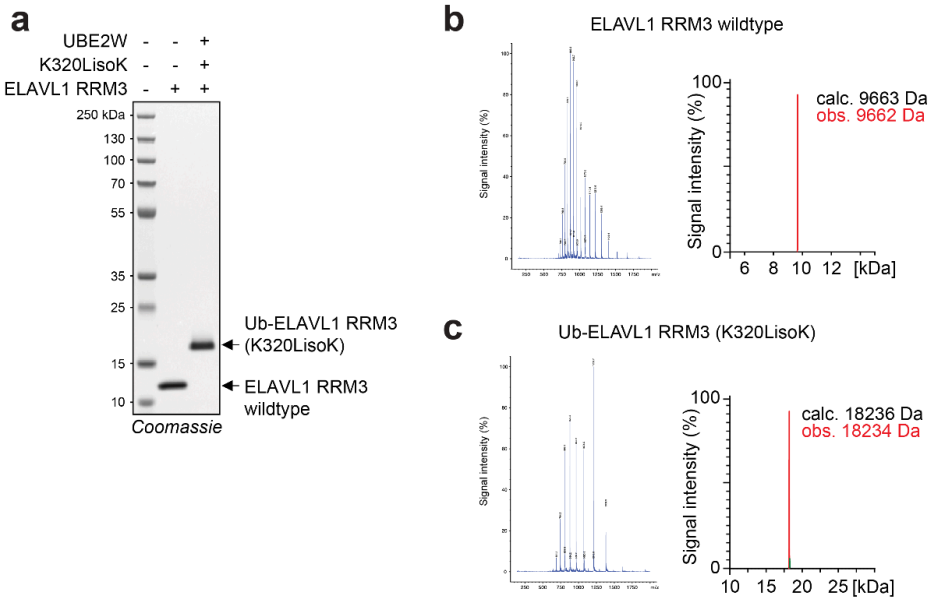

Figure S7: Confirmation of ELAVL1 RRM3 constructs
A) Coomassie-stained SDS-PAGE of purified ELAVL1 constructs: Wildtype ELAVL1 RRM3 and ubiquitinated ELAVL1 RRM3 (K320LisoK). The cropped blot is shown in Fig. 5e. B-C) LC-MS measurement of purified B) ELAV1 RRM3 and C) Ub-ELAVL1 RRM3. Left: non-deconvoluted spectra, right: deconvoluted spectra.

Supplementary tables and data

Table S1: Datasets in the human reference ubiquitinome

Table S2: The human reference ubiquitinome

Table S3: Pan-species ubiquitin proteomics datasets
All_Experiments_nonHs: Meta-data from proteomics experiments in non-human species. Quantitative_Conditions: Meta-data of human quantitative proteomics data.

Table S4: Pan-species ubiquitin proteomics sites

Table S5: Ubiquitin site domain hotspots

hotspotSummary: Summary of hotspots. hotspotSites: All ubi-sites found in hotspots. hotspotRegulatorySites: Hotspot ubi-sites with regulatory annotations in PhosphositePlus.

Table S6: Functional scores and features used for training

Table S7: Ubiquitin sites annotated to degradative and non-degradative functions in PhosphositePlus

Table S8: Ubiquitin sites in nuclear localisation signals or associated with protein activity

Table S9: Yeast chemical genetics experimental data
conditionDescriptions: Descriptions of conditions used in growth screens. MutantDescriptions: Yeast strains used in growth screens. Sscores: S-scores and Q-values from chemical genetics experiments. primers: Primer sequences used for CRISPR-Cas9 generation of lysine mutants. donors: Donor sequences used for CRISPR-Cas9 generation of lysine mutants.

Table S10: Plasmids used in this study

Data S1: Ubi-site hotspots
Visualisation of all ubi-site hotspots with a hotspot found in a representative crystal structure.