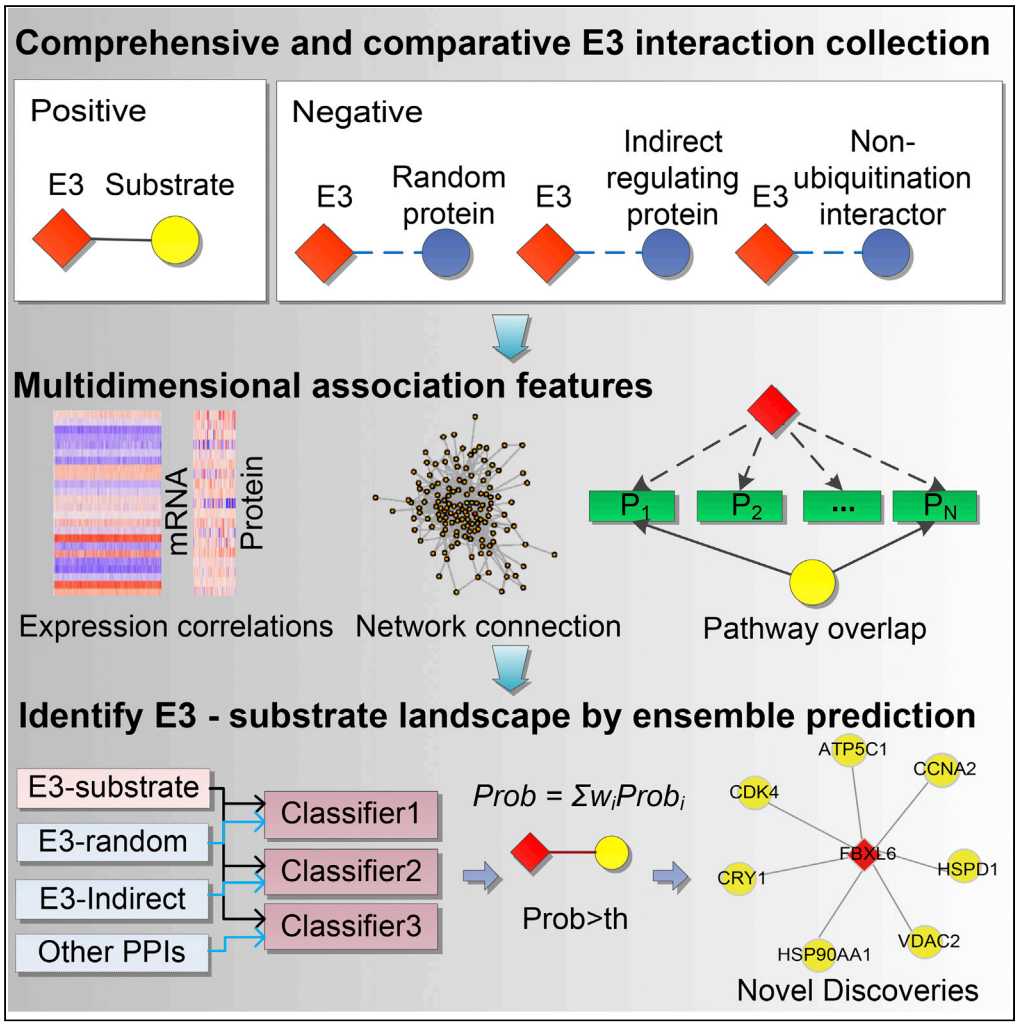
iScience

CellPress

Article

A Multidimensional Characterization of E3 Ubiquitin Ligase and Substrate Interaction Network



Di Chen, Xiaolong Liu, Tian Xia, ..., Huan Qi, Hui He, Hai-long Piao

hpiao@dicp.ac.cn

HIGHLIGHTS

Systematically describe E3 ligase and substrate association patterns

Build a computational model for predicting E3 ligase and substrate interactions

Predict and validate promising substrates for F box family E3 ligases

Construct a valuable data resource on potential E3-substrate interactions

Chen et al., iScience 16, 177–191
June 28, 2019
https://doi.org/10.1016/j.isci.2019.05.033

iScience

CellPress

Article

A Multidimensional Characterization of E3 Ubiquitin Ligase and Substrate Interaction Network

Di Chen,[1,3] Xiaolong Liu,[1,3] Tian Xia,[1] Dinesh Singh Tekcham,[1] Wen Wang,[1] Huan Chen,[1] Tongming Li,[1] Chang Lu,[1,2] Zhen Ning,[1,2] Xiumei Liu,[1] Jing Liu,[1] Huan Qi,[1] Hui He,[2] and Hai-long Piao[1,4,*]

SUMMARY

E3 ubiquitin ligases (E3s) play a critical role in molecular and cellular mechanisms. However, a large number of E3-substrate interactions (ESIs) remain unrevealed. Here, we integrated the increasing omics data with biological knowledge to characterize and identify ESIs. Multidimensional features were computed to obtain the association patterns of ESIs, and an ensemble prediction model was constructed to identify ESIs. Comparison with non-ESI cases revealed the specific association patterns of ESIs, which provided meaningful insights into ESI interpretation. Reliability of the prediction model was confirmed from various perspectives. Notably, our evaluations on leucine-rich repeat family of F box (FBXL) family were consistent with a proteomic study, and several substrates for SKP2 and an orphan E3 FBXL6 were experimentally verified. Moreover, a cancer hallmark ESI landscape was studied. Taken together, our study catches a glimpse at the omics-driven ESI association patterns and provides a valuable resource (http://www.esinet.dicp.ac.cn/home.php) to assist ubiquitination research.

INTRODUCTION

Protein ubiquitination refers to ubiquitin conjugation at a target substrate through three enzymes, including ubiquitin-activating enzyme (E1), ubiquitin-conjugating enzyme (E2), and ubiquitin protein ligase (E3), and mainly results in degradation of a specific substrate (Scheffner et al., 1995). It is one of the most prevalent post-translational modifications in eukaryotic cells (Suresh et al., 2016). Dysregulations of ubiquitination will induce serious diseases, such as cancer (Hoeller and Dikic, 2009; Mani and Gelmann, 2005). During protein ubiquitination, E3s play a key role by specifically recognizing the target substrates. Given the specificity and diversity, E3s are regarded as potential therapeutic targets in cancer (Bassermann et al., 2014). However, apart from several well-defined E3s, e.g., MDM2 (Rayburn et al., 2005) and NEDD4 (Ye et al., 2014), most of them remain poorly characterized, making it a great challenge to fully understand the ubiquitination system.

Typically, substrates for E3s are discovered by biochemical experiments (e.g., two-hybrid screen or co-immunoprecipitation) in a case-by-case manner (Chan et al., 2012; Maddika et al., 2011), which are commonly time- and resource-consuming. Recently, some high-throughput methods (Tan et al., 2013; Yumimoto et al., 2012) have been utilized on ESI recognition. However, the screened results are still mixed with plenty of false discoveries, and proteome-wide ESI identifications are still far behind. It is indispensable to build effective computational methods to assist ESI recognition. A platform of UbiBrowser (Li et al., 2017) has been constructed to predict ESIs, providing references for deciphering a proteome-wide ESI network. However, it mainly depends on mechanism-agnostic enrichment of pairwise protein features among known E3-substrate pairs, overlooking the expression or functional correlations between E3s and substrates. In contrast to the limited knowledge of E3-substrate interactome, extensive omics data are available, owing to the efforts on collection and organization of high-throughput biological data by public repositories, like The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) (Cancer Genome Atlas Research Network et al., 2013) and Clinical Proteomic Technology Assessment for Cancer (CPTAC, https://cptac-data-portal.georgetown.edu/cptacPublic/) (Edwards et al., 2015). They have been widely employed in functional description of cancer-driving factors or interactions (Xue et al., 2017; Ye et al., 2016). Association-centric heuristic methods are commonly the optimal option for omics-based analysis (Stuart et al., 2003; Wang et al., 2016), and functional related or interacting components can be identified. However,

[1]CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China

[2]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Dalian Medical University, Dalian Medical University, Dalian 116000, China

[3]These authors contributed equally

[4]Lead Contact

*Correspondence: hpiao@dicp.ac.cn

https://doi.org/10.1016/j.isci.2019.05.033

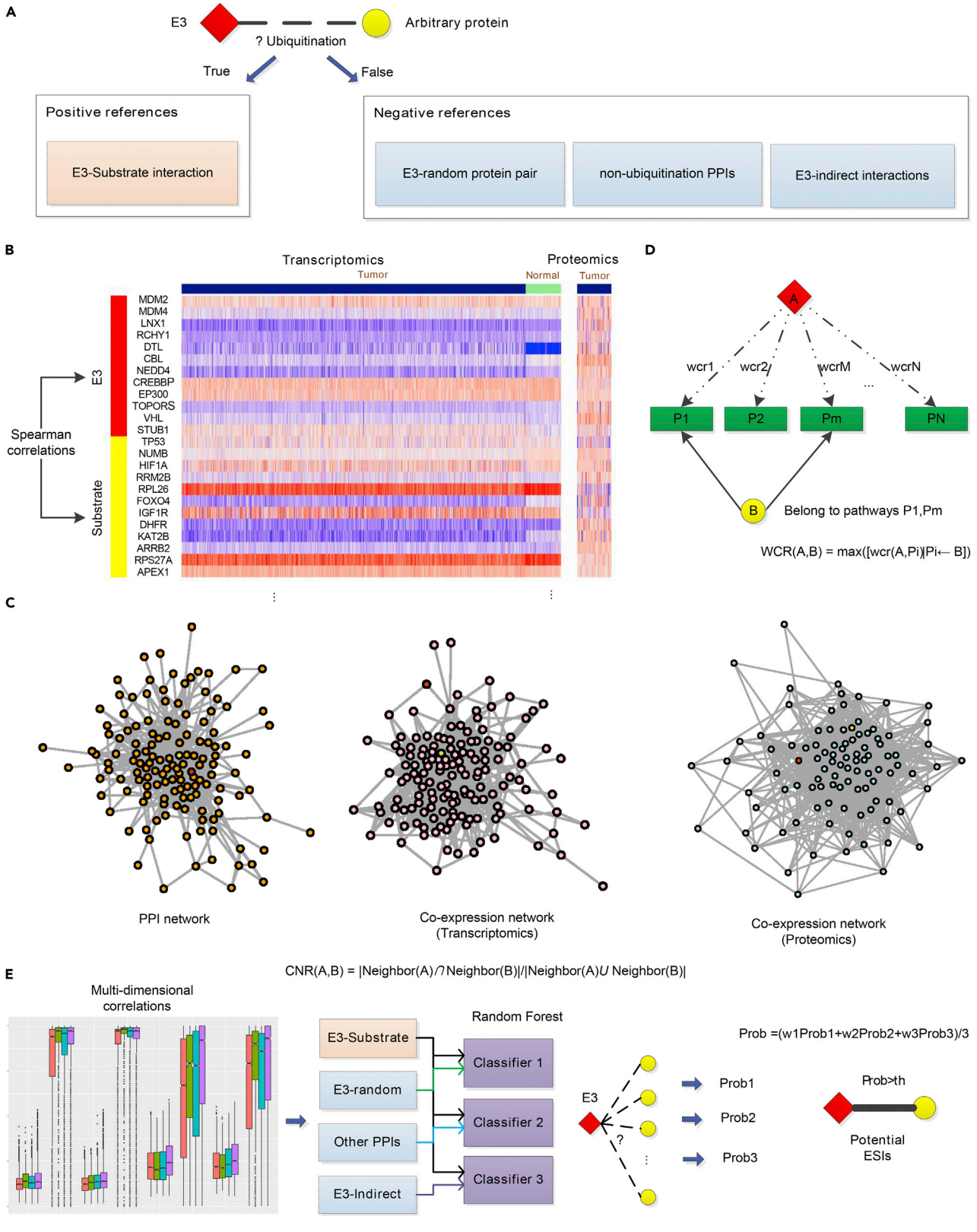

A
E3
? Ubiquitination
Arbitrary protein
True
False
Positive references
E3-Substrate interaction
Negative references
E3-random protein pair
non-ubiquitination PPIs
E3-indirect interactions
B
Transcriptomics
Tumor
Normal
Proteomics
Tumor
E3
Substrate
Spearman correlations
MDM2
MDM4
LNX1
RCHY1
DTL
CBL
NEDD4
CREBBP
EP300
TOPORS
VHL
STUB1
TP53
NUMB
HIF1A
RRM2B
RPL26
FOXO4
IGF1R
DHFR
KAT2B
ARRB2
RPS27A
APEX1
D
A
wcr1
wcr2
wcrM
wcrN
P1
P2
Pm
PN
B
Belong to pathways P1,Pm
WCR(A,B) = max([wcr(A,Pi)|Pi← B])
C
PPI network
Co-expression network (Transcriptomics)
Co-expression network (Proteomics)
E
CNR(A,B) = |Neighbor(A)∩Neighbor(B)|/|Neighbor(A)∪ Neighbor(B)|
Multi-dimensional correlations
Random Forest
E3-Substrate
E3-random
Other PPIs
E3-Indirect
Classifier 1
Classifier 2
Classifier 3
E3
?
Prob1
Prob2
Prob3
Prob =(w1Prob1+w2Prob2+w3Prob3)/3
Prob>th
Potential ESIs

Figure 1. Workflow of a Comparative Characterization on ESIs

(A) The association patterns of ESIs were obtained by comparison between positive and negative reference datasets, where three different negative reference datasets (NRDs) were prepared for comprehensiveness.

(B) Omics-based associations. The Spearman correlations between E3s and substrates were calculated based on both transcriptomics and proteomics.

(C) Network-based associations. The ESIs were mapped onto three networks including one PPI network and two co-expression networks constructed, respectively, based on transcriptomics and proteomics data. The associations were calculated from the shared neighbors on the networks.

(D) Pathway-based associations. Pathways are taken as mediators between E3 and substrates, and the pathway-based association evaluates whether one E3 is tightly related with the pathways their substrates belong to, or vice versa.

(E) Diagram of ESI prediction model. It was an ensemble of three types of random forest (RF) classifiers trained with different categories of NRDs.

See also Tables S1, S2, S3, and S4.

given the reversibility and dynamics of ubiquitination, it is still unclear whether E3-substrate interaction (ESI) can be correctly described by omics-based associations. In addition, progressively accumulated biological knowledge, like the annotated pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017) and experimentally verified protein-protein interaction (PPI) records in BioGrid (Chatr-Aryamontri et al., 2017), provide an alternative way to create mechanistic descriptions on ESIs.

Here, we put forward to characterize and predict ESIs in an integrative way. It turns out that integrating omics (especially proteomics data) with network or pathway information can help distinguish ESIs from various negative categories (even indirect regulating relations). A case study on leucine-rich repeat family of F box (FBXL) proteins demonstrated that our evaluations outperformed both the UbiBrowser and a proteomics-based approach. Moreover, we portrayed a cancer-oriented ESI landscape by identifying potential ESIs for cancer hallmark proteins. In addition, to facilitate the utility of our model, both confirmed and potential ESIs along with their association features were distributed on a website (http://www.esinet.dicp.ac.cn/home.php).

RESULTS

Workflow of a Comparative Characterization on ESIs

We developed a computational model for both describing and predicting ESIs by integrating omics with network and pathway (Figure 1, see also Transparent Methods). First, previously reported ESIs were collected as a positive reference dataset (PRD, Table S1), whereas three negative reference datasets (NRDs) covering randomly combined E3s and proteins ("E3-random," Table S2), non-ESI PPIs ("Other PPIs," Table S3), and pairwise E3-indirect regulatory proteins ("E3-Indirect," Table S4) were prepared as control (Figure 1A). For both PRD and NRDs, multidimensional association features between two proteins were calculated by integrating omics (Figure 1B), networks (Figure 1C), and pathways (Figure 1D). Then, specific ESI association patterns were recognized by comparing ESIs with three categories of NRDs. Finally, a prediction model was constructed through a weighted ensemble of three types of random forest classifiers (RFs) (Figure 1E).

Constructing an ESI Network

Initially, an ESI network composed of 1,806 previously reported ESIs involving 300 E3s and 1,089 substrates was constructed to describe the complex relations between E3s and substrates (Figure 2A). Similar to other biological networks (Goymer, 2008; Rolland et al., 2014), it exhibited an approximate scale-free topology (Barabasi, 2009) (linear model fitting R2 index = 0.78). Most hubs in the ESI network were E3s; they had significantly higher (p value = 4.07×10^{-7} based on t test) degrees than substrates, suggesting that a large number of substrates can be recognized by the same E3s. However, when mapping onto the global PPI network, such difference disappeared (p value = 0.9634 by t test); E3s and substrates exhibited similar degree distribution and their degrees were much higher than those on the ESI network (Figure 2B), emphasizing that the impacts rendered by ubiquitination may be spread to various processes by non-ESI interactions. Besides, a cancer hallmark subgraph (Figure 2C) was extracted. On this subgraph, nodes belong to genes recorded in the "Catalogue of somatic mutations in cancer" (COSMIC) (Forbes et al., 2017) and edges are annotated with the ubiquitinated consequences of substrates. It suggests that a large fraction of the reported ESIs will lead to degradation of substrates. Some E3s (e.g., BRCA1 and KEAP1) are tumor suppressors; their mutations might induce the accumulation of carcinogenic substrates (e.g., AKT, IKK-β, and JAK2), which will promote the development of cancer and affect key signaling pathways, such as AKT or nuclear factor (NF)-κB signaling pathway, confirming that ubiquitination can affect cancer development and progression in alternative ways. To be more comprehensive, further refining of the ESI network is indispensable.

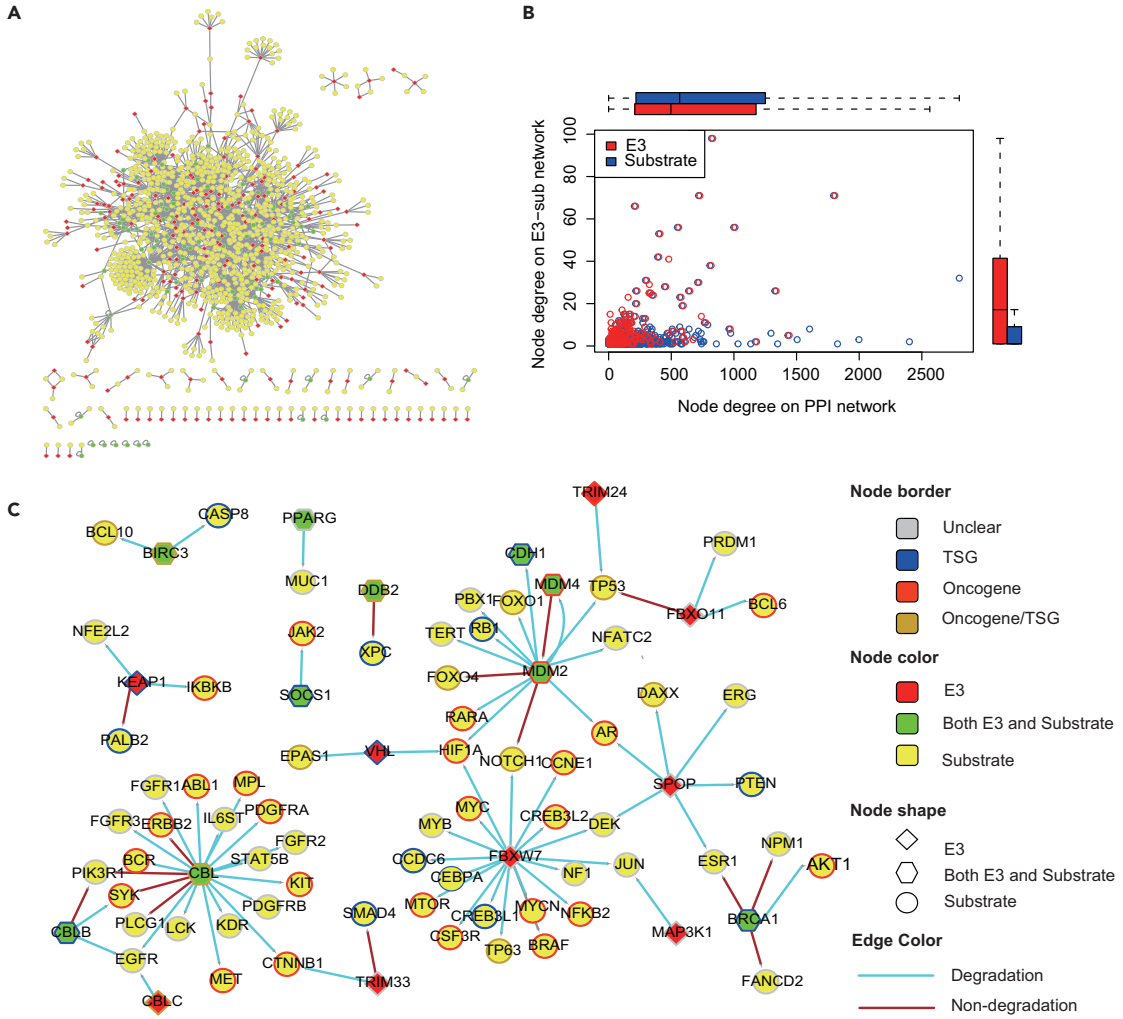


Figure 2. ESI Network

(A) An ESI network. E3s (red) and their corresponding substrates (yellow) are connected as edges, where green nodes stand for proteins acting as both E3s and substrates.

(B) Degree distribution of E3s (red bars) and substrates (blue bars) on the ESI network (horizontal axis) and the global PPI network (vertical axis). Centers of boxes represent median values. Bottom and top bounds of boxes represent 25th and 75th percentiles.Whiskers mark 1.5 times of the interquartile range

(C) Cancer hallmark ESI subgraph. Node borders were colored according to their functions in cancer; blue stands for tumor suppressor gene (TSG), red stands for oncogene, orange represents genes with both functions, whereas the functions of gray-border ones are still unclear. The edge color represents the ubiquitinated consequence of substrates; blue stands for degradation, whereas red refers to activity change or other non-degradation effects. Nodes are labeled with the corresponding gene symbols of proteins.

See also Table S1.

Omics-Based Associations Alone Are Not Capable of ESI Recognition

Proteomics data offer the most direct resource to capture the expression relevance between E3s and substrates, given that ESIs often results in substrate degradation. In addition, as ubiquitination is also involved in transcription regulation (Hammond-Martel et al., 2012), transcriptomics data were also employed to estimate transcriptional associations (see Transparent Methods). Notably unexpected, E3s and substrates were not negatively correlated in both mRNA and protein levels (Figure 3A, median levels for ESIs were all near zero). However, when compared with NRDs, ESIs exhibited differences in both omics: ESIs showed higher correlations than "E3-random" cases on average (Figure 3A). "Other PPIs" were more mutually correlated than ESIs; indirect regulations were similar with ESIs in terms of proteomics, but differences emerged in terms of transcriptomics. Still, more features are required because the slight differences make it difficult to separate ESIs from other possibilities.

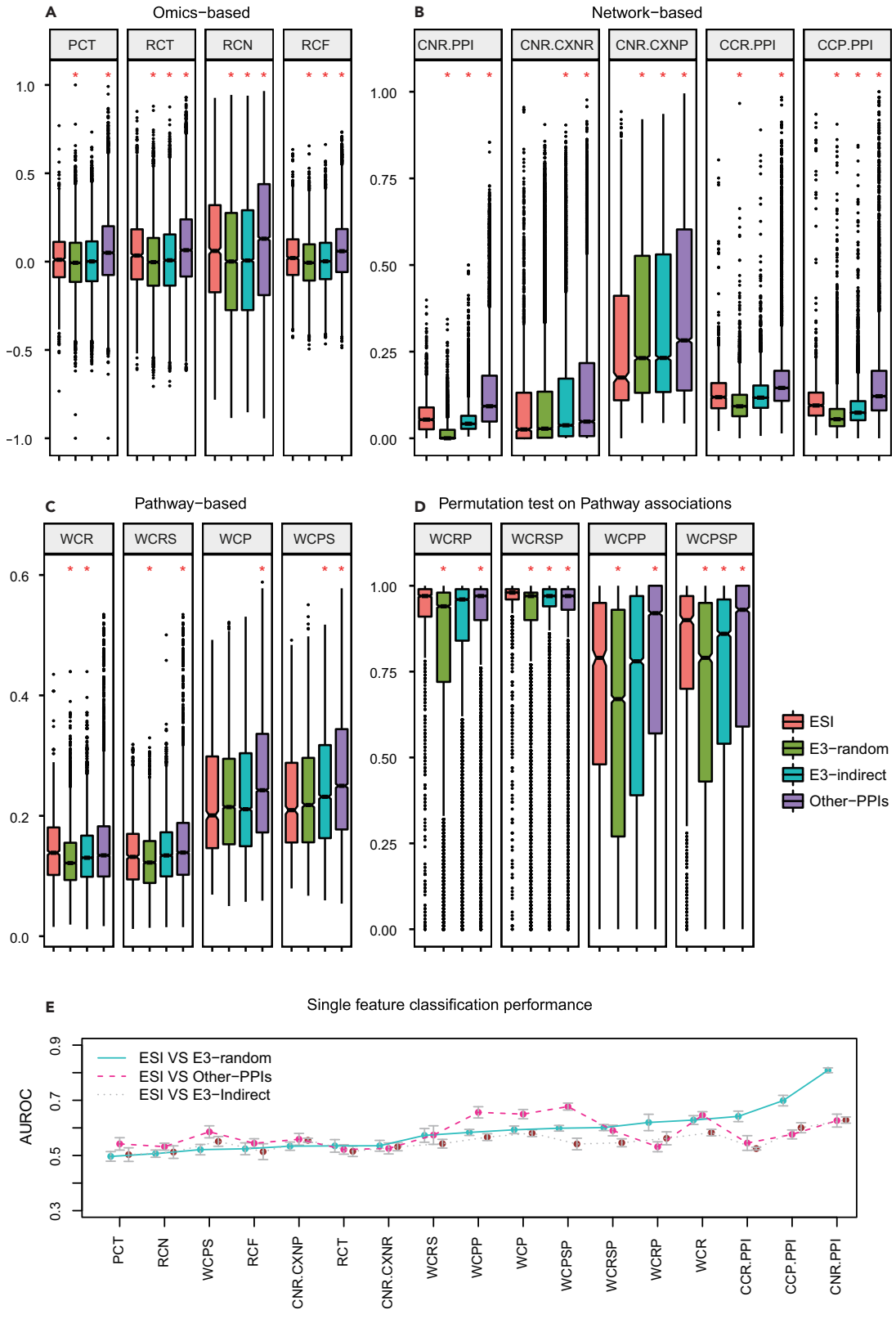

A
Omics−based
PCT
RCT
RCN
RCF
B
Network−based
CNR.PPI
CNR.CXNR
CNR.CXNP
CCR.PPI
CCP.PPI
C
Pathway−based
WCR
WCRS
WCP
WCPS
D
Permutation test on Pathway associations
WCRP
WCRSP
WCPP
WCPSP
ESI
E3−random
E3−indirect
Other−PPIs
E
Single feature classification performance
AUROC
ESI VS E3−random
ESI VS Other−PPIs
ESI VS E3−Indirect
PCT
RCN
WCPS
RCF
CNR.CXNP
RCT
CNR.CXNR
WCRS
WCPP
WCP
WCPSP
WCRSP
WCRP
WCR
CCR.PPI
CCP.PPI
CNR.PPI

Figure 3. Comparative Association Profiles of ESIs

(A) Omics-based associations. Different colors in the box plots represent different datasets. PCT refers to proteomics-based correlations between two proteins in tumor tissues. Three different forms of transcriptomics-based associations were calculated. RCT/RCN means the mRNA correlations in tumor/normal tissues; RCF refers to two proteins' correlations in terms of the fold change on mRNA of tumor tissues relative to matched normal tissues.
(B) Network-based associations. Five features were calculated: CNR.PPI, CNR.CXNR, and CNR.CXNP refer to the common neighbor rate (CNR) between two proteins in the PPI network, mRNA co-expression network, and protein co-expression network respectively; CCR.PPI/CCP.PPI stands for the CNR between co-expressed factors (mRNAs/proteins) in PPI network.
(C) Pathway-based associations. WCR/WCP evaluates the transcriptomics/proteomics-based correlations between E3 and the pathways their substrates belong to, whereas WCRS/WCPS estimates the correlations between E3s' pathways and substrates based on transcriptomics/proteomics.
(D) Permutation test on each form of pathway associations. WCRP, WCRSP, WCPP and WCPSP respectively represent the permutation test based P-values obtained for WCR, WCRS, WCP and WCPS. Data distributions for A-D are all discribed by box-plots. Centers of boxes represent median values. Bottom and top bounds of boxes represent 25th and 75th percentiles. Black lines mark 1.5 times of the interquartile range. Dots represent points falling outside this range.
(E) The AUROCs of knn-classifiers trained based on each single feature. The gray bars stand for standard deviation of AUROCs for five repetitive classifiers. Performances of classifiers based solely on omics were only around 0.5, but inclusion of network and pathway knowledge improved the performance. Statistical significance ($p < 0.01$, Wilcoxon test) for difference between each kind of NRDs and ESIs was marked by *.
See also Figures S1–S3.

E3s and Substrates Share Neighbors on PPI Network but Lack Connections on Co-expression Networks

Omics-based associations simply investigated on the "one-to-one" expression relations, ignoring the fact that cascades of other proteins also take part in the ubiquitination process. Consequently, three network systems were applied to measure the relations between E3s and substrates systematically (see Transparent Methods). As a result, ESIs and NRDs showed significantly different distributions on the network-based associations (Figure 3B). No matter on which kind of network, the average correlations of ESIs were significantly (p value<0.01) less than "Other PPIs." On the PPI network, E3s and substrates are more likely to be connected by common neighbors than "E3-random" or "E3-indirect" (Figure 3B, CNR.PPI). However, this tendency was changed on the co-expression network; especially on the protein co-expression network (Figure 3B, CNR.CXNP), the ESIs were less connected compared with all other NRDs, even the random cases, reflecting that the ESIs interrupt the expression correlations despite the probably transient and dynamic properties. When the co-expression networks and PPI network were integrated, the profile (Figure 3B, CCR.PPI and CCP.PPI) was similar with that based on PPI network. The network-based associations can assist to differentiate the substrates of E3s from indirect or random situations in a relatively more accurate manner than the omics-based associations.

Pathway-Based Associations Improve ESI Discriminability and Provide Hypothesis on Their Upstream or Downstream Processes

As a special kind of bionetworks, a pathway refers to a cascade of molecular interactions with interdependent functions. A total of 306 human pathways in KEGG (Kanehisa et al., 2017) were considered as intermediary to examine the functional associations between E3s and substrates (see Transparent Methods). With regard to mRNA-based pathway associations (i.e., WCR, WCRS), ESIs only exhibited modest higher associations than "E3-random" and "E3-Indirect" (Figure 3C); however, permutation tests on these associations indicated that specificities (i.e., WCRP, WCRSP) of the pathway-based associations of ESIs were much higher than those of the others (Figure 3D). It suggests that although E3s were not highly correlated or co-expressed with their substrates compared with random or indirect regulatory cases in mRNA level (Figures 3A and 3B), they may be closely related with the other members of the pathways their substrates belong to, or vice versa. Proteomic level presented different tendency; ESIs exhibited lower associations than all three NRDs on average, but the specificities were higher than "E3-random" and "E3-Indirect." The "Other PPIs" exhibited both considerably higher pathway associations and permutation scores than ESIs. It is probably a consequence of the degradation effect or dynamic nature of ESIs. Moreover, the pathway-based associations promoted to discriminate ESIs from NRDs; some of them (Figure 3E, WCP, WCR, WCPP, WCPSP) even rendered better performances than network-based features, especially in distinguishing ESIs from "Other PPIs."

The pathway-based associations can also capture which pathway was highly correlated with an ESI (see Transparent Methods). For example, pathways, like cell cycle (Shabbeer et al., 2013; Zheng et al., 2016), p53 signaling pathway (Li et al., 2012; Zhang et al., 2014), and transforming growth factor-β signaling pathway (Gen et al., 2017), all of which have been confirmed as ubiquitination-mediated pathways, were highly correlated with the E3s in proteomics, and a large fraction of substrates can be located on these pathways (Figure S1), similar results were observed in the transcriptomics-based pathway associations

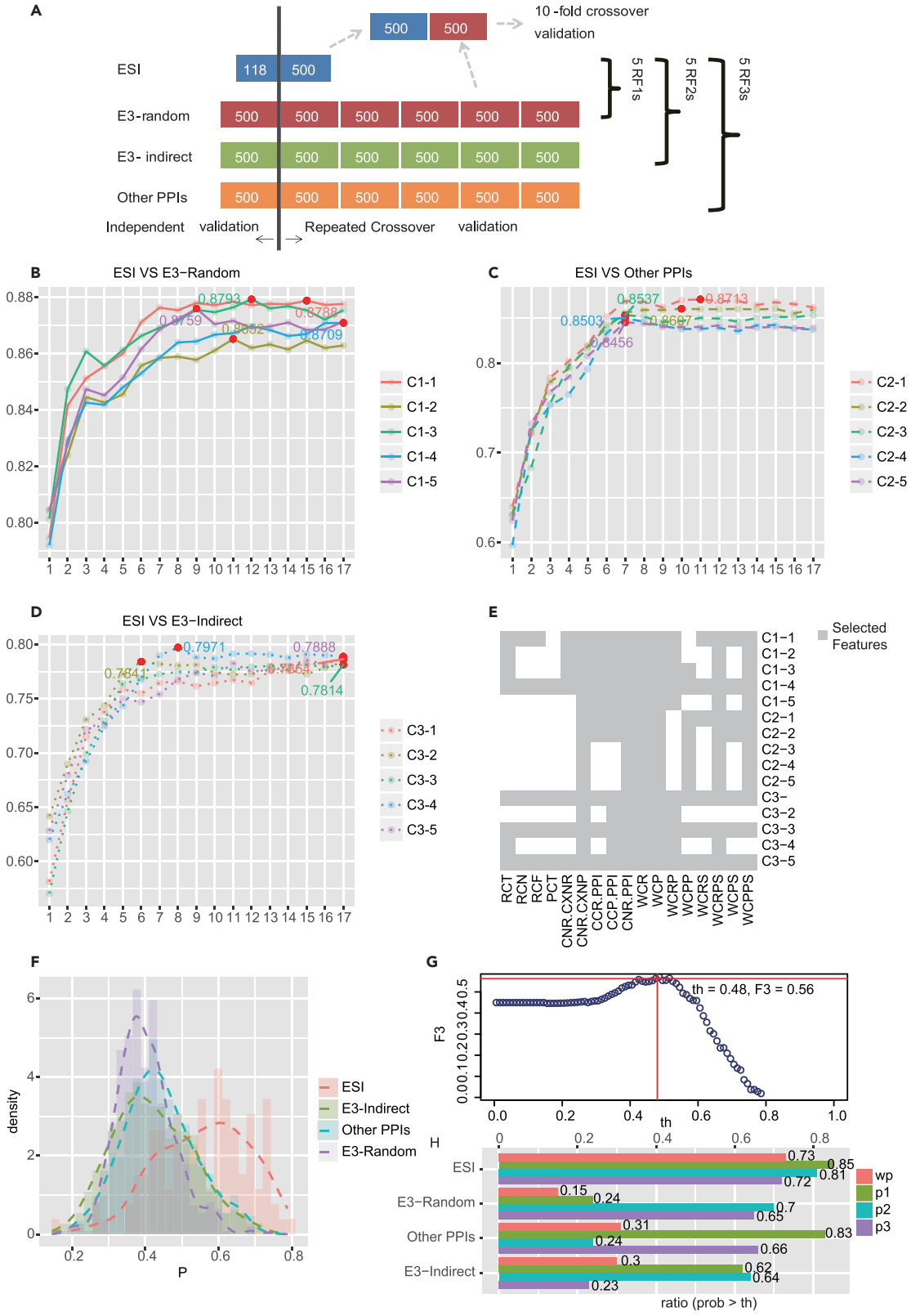

A
ESI
E3-random
E3- indirect
Other PPIs
118
500
10 -fold crossover validation
5 RF1s
5 RF2s
5 RF3s
Independent validation
Repeated Crossover validation
B
ESI VS E3−Random
C1-1
C1-2
C1-3
C1-4
C1-5
C
ESI VS Other PPIs
C2-1
C2-2
C2-3
C2-4
C2-5
D
ESI VS E3−Indirect
C3-1
C3-2
C3-3
C3-4
C3-5
E
Selected Features
RCT
RCN
RCF
PCT
CNR.CXNR
CNR.CXNP
CCR.PPI
CCP.PPI
CNR.PPI
WCR
WCP
WCRP
WCPP
WCRS
WCRPS
WCPS
WCPPS
F
density
P
ESI
E3-Indirect
Other PPIs
E3-Random
G
th = 0.48, F3 = 0.56
F3
th
H
ESI
E3−Random
Other PPIs
E3−Indirect
wp
p1
p2
p3
ratio (prob > th)

Figure 4. Classification Models for Distinguishing ESIs from NRDs

(A–D) (A) Sample assignment for five rounds of crossover validations and an independent validation. (B–D) Performance of three types of RFE-RFs. The same ESIs were used as positive samples, whereas "E3-random" (B), "Other PPIs" (C), and "E3-Indirect" (D) pairs were taken as negatives. The horizontal and vertical axes represent the number of selected features and the corresponding average AUROC got from a 10-fold crossover validation. For each type of RFE-RFs, five classifiers were constructed by replacing negative samples.

(E) Selected features for each classifier in (B–D).

(F) Density plot of the predicted probabilities across four types of samples in the independent validation dataset.

(G) Threshold selection for positive predictions. Performances under different thresholds (th) were evaluated by F3-score.

(H) Histogram of positively predicted samples across different categories of samples in the independent validation.

See also Table S5.

(Figure S2), implying that our analysis might provide some hypotheses on the upstream or downstream processes for certain ESIs by referring to the highly correlated pathways.

Pan-Cancer Analysis Show Concordance on the E3-Substrate Association Patterns

The analyses above utilized omics data from Breast Invasive Carcinoma (BRCA) in TCGA. To be more comprehensive, we asked whether the ESI association patterns were consistent across different cancers. We recalculated the features based on additional 10 cancers. Parallel analyses based on distinct cancers manifested that the differential trends between ESIs and NRDs on most available features were consistent with those shown in BRCA (Figures 3A–3D and S3A–S3J). The concordance also indicates that the multi-omics data in TCGA can be applicable to gain insights into biological mechanism in a more universal manner, even though they are originally cancer oriented.

Classification Models Based on the Multidimensional Associations Perform Well in Discriminating ESIs from NRDs

The considerable differences between ESIs and NRDs in the multidimensional association space prompt us to construct a classification model wherein three types of reverse-feature-reduction based random forest classifiers (RFE-RFs) were combined to identify ESIs (Figure 4A, see also Transparent Methods). The performance was quantified by the area under the receiver operating characteristic curve (AUROC). The classification model showed a satisfying and stable performance (Figures 4B–4D), especially when "E3-random" and "Other PPIs" categories (average AUROCs around 0.87 and 0.85, respectively) were taken as negatives, whereas the performance was somewhat declined for "E3-Indirect" (average AUROCs around 0.78). The observation indicated it is more difficult to separate true substrates from indirect regulatory proteins than other negative cases. For different types of RFE-RFs, features like CNR.PPI, WCR, and WCP (Figure 4E) were consistently selected for all classifiers, further confirming the importance of network and pathway-based associations.

Moreover, ESIs were predicted with significantly higher probabilities than all three types of NRDs (Figure 4F). Both crossover and independent validations show that our model can help separate true and false ESIs. Although false-positives still existed, the misclassification rate was significantly reduced by inclusion of three negative categories. In general, random cases are always chosen as negative controls in interactome prediction, based on which the classifier generated much more false-positives (24% of "E3-Random," 83% of "Other PPIs," and 62% of "E3-indirect" were misclassified as ESIs, Figure 4H). Similar conditions were observed when only "Other PPIs" or "E3-Indirect" was taken as the control. However, when three types of classifiers were integrated, the false discovery rate was significantly reduced, with 73% ESIs being accurately predicted under the threshold of 0.48 (Figures 4G and 4H), whereas only 15% "E3-Random," 31% "Other PPIs," and 30% "E3-indirect" being misclassified.

We also trained the models by omics data from other cancers (Table S5). We observed that models trained based on ovarian serous cystadenocarcinoma(OV) and BRCA, where both transcriptomics and proteomics data were available, obtained better performance than others with only transcriptomics data, confirming the importance of multi-omics integration. As data of BRCA were more comprehensive than OV (only 155 ESIs can be assigned with both omics), we mainly utilized data from BRCA in the following study.

Prediction and Validation on Potential Substrates of FBXL Family

To further estimate the model quality, we applied it on the FBXL proteins, of which certain post-transcriptional modifications are often required for the substrates (Skaar et al., 2013). In particular, to avoid circularity of training and predicting samples, we reconstructed the classification model (Figure S4) by removing known ESIs of FBXLs from the training process. Both known and potential substrates were identified by our

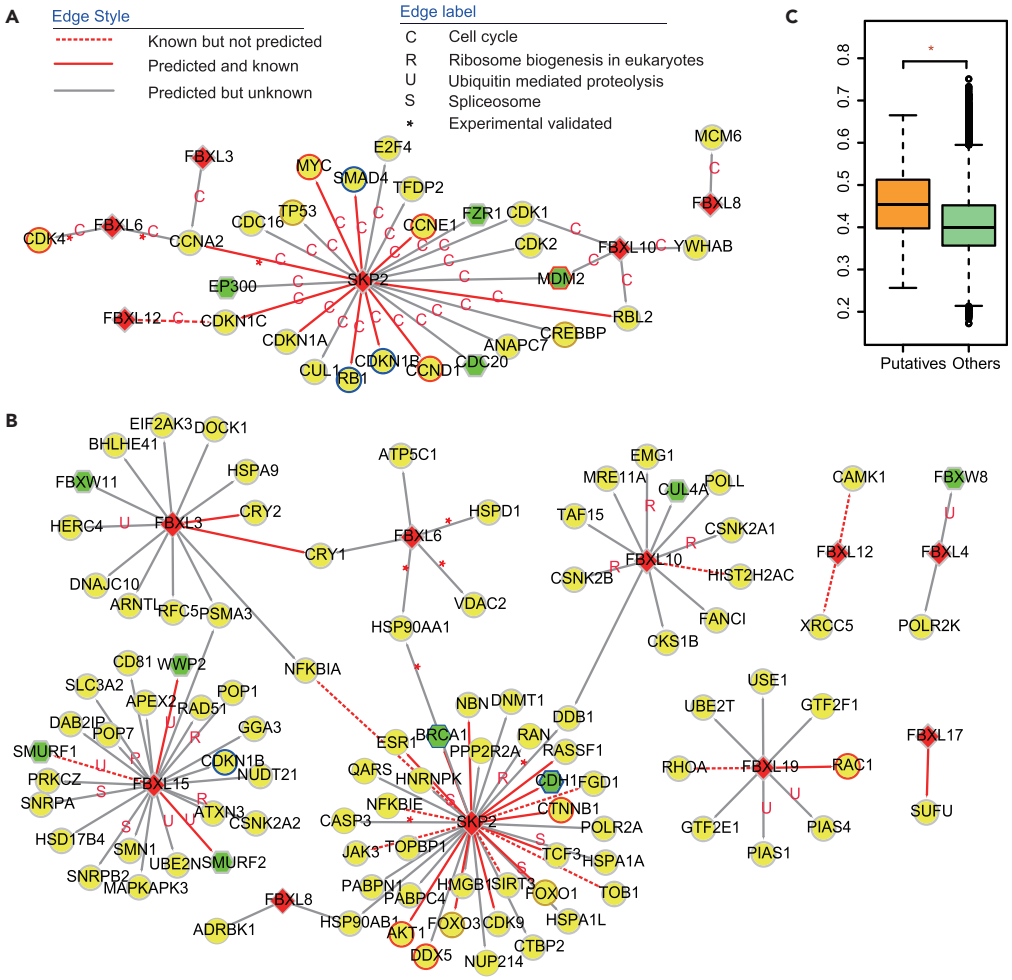


Figure 5. Evaluations on Substrates of FBXL Family E3s

(A and B) Known (red edge) and predicted (gray edge) ESIs for FBXLs. The meaning of node colors and shapes is the same as in Figure 2. For clarity, only the high-confidence substrates (with a predicted probability higher than 0.65 and ranked within top-20 for corresponding FBXL) were shown. (A) A large part of the predicted ESIs are mapped onto cell cycle pathway by pathway-based associations. (B) In addition to substrates on the cell cycle pathway, the FBXLs were also predicted to target on other substrates across different pathways.

(C) Boxplot of the predicted results on candidate ESIs. The putative ESIs identified by PAC (orange box) were observed with significant higher probabilities than the remaining ones (green box). *$p < 0.01$ (Wilcoxon test, two sided, unpaired). See also Figure S4 and Table S6.

model. Considering 89 confirmed ESIs of FBXLs, 39 pairs that can be assigned with the multidimensional features were calculable by our model, and 27 of the 39 (69%) cases were correctly recalled, whereas only 13 among the 39 cases were predicted by UbiBrowser (Table S6).

Multiple FBXLs were predicted to target on substrates belonging to the pathways of cell cycle, ribosome biogenesis in eukaryotes, ubiquitin-mediated proteolysis, and spliceosome (Figures 5A and 5B). The mostly influenced pathway is the cell cycle pathway, where multiple known substrates like MYC, SMAD4, and CCNE1 were correctly linked to SKP2, and other unreported ones like CCNA2 and CDK1 were predicted to be recognized by FBXL3, FBXL6, and FBXL10, indicating the potential cross-regulatory mechanism on the cell cycle pathways by FBXLs. This is consistent with the fact that F box proteins play key roles in cell cycle regulation (Zheng et al., 2016).

To test on the quality of our model, we compared our predictions to a proteomics-based experimental study (Tan et al., 2013), where a well-designed parallel adaptor capture (PAC) proteomic method was

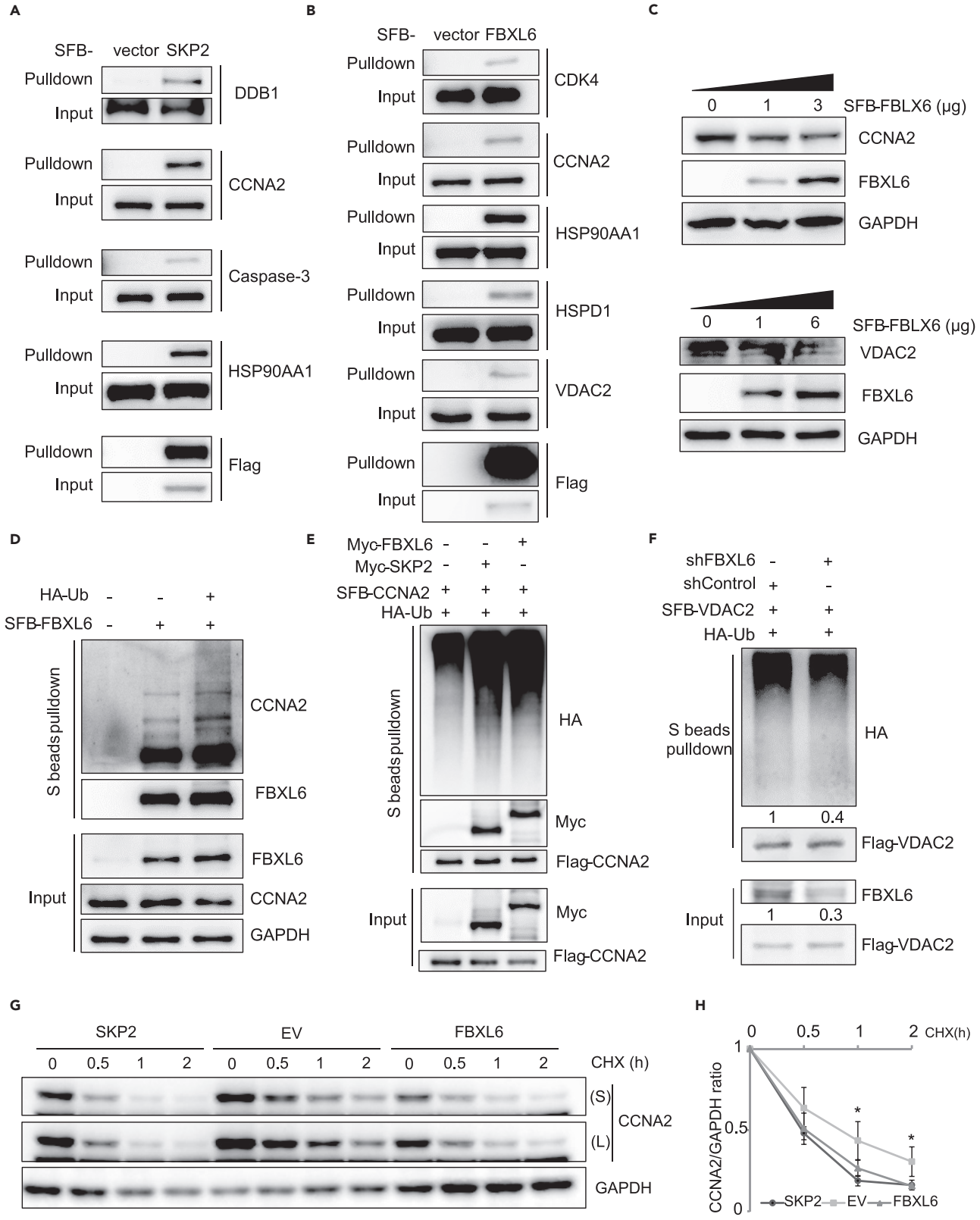

A
SFB- vector SKP2
Pulldown
Input
DDB1
CCNA2
Caspase-3
HSP90AA1
Flag
B
SFB- vector FBXL6
CDK4
CCNA2
HSP90AA1
HSPD1
VDAC2
Flag
C
0 1 3 SFB-FBLX6 (μg)
CCNA2
FBXL6
GAPDH
0 1 6 SFB-FBLX6 (μg)
VDAC2
FBXL6
GAPDH
D
HA-Ub - - +
SFB-FBXL6 - + +
S beads pulldown
CCNA2
FBXL6
Input
FBXL6
CCNA2
GAPDH
E
Myc-FBXL6 - - +
Myc-SKP2 - + -
SFB-CCNA2 + + +
HA-Ub + + +
S beads pulldown
HA
Myc
Flag-CCNA2
Input
Myc
Flag-CCNA2
F
shFBXL6 - +
shControl + -
SFB-VDAC2 + +
HA-Ub + +
S beads pulldown
HA
1 0.4
Flag-VDAC2
Input
FBXL6
1 0.3
Flag-VDAC2
G
SKP2 EV FBXL6
0 0.5 1 2 0 0.5 1 2 0 0.5 1 2 CHX (h)
(S)
(L)
CCNA2
GAPDH
H
0 0.5 1 2 CHX(h)
CCNA2/GAPDH ratio
1
0.5
0
SKP2 EV FBXL6

Figure 6. Experimental Validations on High-Confident Predictions

(A and B) Validations on interactions between substrates and FBXLs. HEK293T cells were transfected with SFB-tagged SKP2 or FBXL6. SKP2 and FBXL6 were purified with S protein beads and immunoblotted with antibodies against substrates. (A) Interactions between SKP2 and DDB1, Cyclin A2 (CCNA2), Caspase-3 (CASP3), as well as HSP90AA1 were examined. (B) Interactions between FBXL6 and CDK4, CCNA2, HSP90AA1, HSPD1, as well as VDAC2 were examined.

(C) Increased FBXL6 correlates with reduced CCNA2 and VDAC2 expression. HEK293T cells were transfected with SFB-FBXL6 at different doses (0, 1, and 3 or 6 μg) for 48 h. Expression of CCNA2 and VDAC2 were analyzed by western blot.

(D) FBXL6 mediated the ubiquitination of CCNA2. SFB-FBXL6 vector was transfected into HEK293T cells along with or without haemagglutinin-tagged ubiquitin (HA-Ub). Cells were cultured for 48 h, and cell lysates were probed with CCNA2 antibody.

(E) FBXL6 and SKP2 mediated the ubiquitination of CCNA2. HEK293T cells were transfected with SFB-CCNA2 and HA-Ub along with Myc-FBXL6 or Myc-SKP2 for *in vivo* ubiquitination assay. CCNA2 was purified with S protein beads and ubiquitination of CCNA2 by FBXL6 and SKP2 were analyzed.

(F) HA-Ub and SFB-VDAC2 were co-transfected with control or FBXL6-shRNA into HEK293T cells. Cells were treated with 20 μM MG132 for 4 h before collection. VDAC2 was purified with S protein beads and western blotted with antibodies against HA, FLAG, and FBXL6.

(G) HEK293T cells were transfected with SFB-FBXL6, SFB-SKP2, along with CCNA2-V5 for *in vitro* half-life assay under 100 μg/mL of cycloheximide (CHX) treatment. Cells were collected at different time points and immunoblotted with antibodies against CCNA2 and GAPDH. S/L: short/long exposure.

(H) Quantification of relative CCNA2 levels in (G) and replicated experiments were performed using ImageJ, and data are represented as mean ± SEM. *$p < 0.05$ (t test, two sided, unpaired) for the comparisons of both SKP2 and FBXL6 to empty vector (EV). The experiments were repeated three times, and the most representative image is shown.

See also Figure S5.

applied to discriminate putative substrates for FBXLs. Among all the candidate pairs we predicted, the intersection with putative ones identified by PAC show significant (Wilcoxon test, p value = 1.036×10^{-6}) higher probabilities than the others (Figure 5C), confirming the reliability of our model. Besides, among all 89 confirmed ESIs of FBXLs, 27 cases were correctly identified by our model, whereas only 5 and 22 can be discriminated by the PAC method and UbiBrowser, respectively (Table S6).

Furthermore, we experimentally checked on several high-confidence substrates of SKP2 and FBXL6 (see Transparent Methods). For SKP2, its interactions with four predicted substrates were validated, where the interaction with CCNA2 was already reported by previous studies (Nakayama et al., 2000) and the other three proteins including CASP3, DDB1, and HSP90AA1 have not been discovered (Figure 6A, S5A, and S5B). For FBXL6, an orphan E3 without any known substrates, its interactions with five predicted substrates including CDK4, CCNA2, HSP90AA1, HSPD1, and VDAC2 were validated (Figures 6B and S5A–S5D). Moreover, we found that FBXL6 inhibited CCNA2 and VDAC protein expression in a dose-dependent manner (Figure 6C). In addition, FBXL6 and SKP2 significantly increased both endogenous and exogenous CCNA2 protein polyubiquitylation (Figures 6D and 6E). In contrast, knockdown of FBXL6 expression by transducing shRNA_FBXL6 significantly inhibited polyubiquitylation of VDAC2 (Figures 6F and S5E). Also, we examined CCNA2 protein levels in the presence of cycloheximide. As expected, SKP2 and FBXL6 significantly decreased the stability of CCNA2 protein (Figures 6G and 6H). These experimental results prove that our predictions can provide credible references on identification of promising ESIs.

Inferences on Potential E3s for Cancer Hallmark Proteins

To be more comprehensive, the ultimate prediction model (Figure S6) was reconstructed by incorporating all known ESIs into the training procedure. Finally, about 2,80,000 pairs of proteome-wide potential ESIs were inferred by our model. To investigate whether cancer hallmark proteins might be ubiquitinated by certain E3s (see Table S7 for all investigated E3s), candidate pairs with COSMIC-recorded proteins as substrates were retrieved. Among them, 79 pairs were predicted as high-confidence ($p > 0.75$, Figure 7A) cases for cancer hallmark proteins, where 22 of them were previously revealed and another 57 were predicted (19 of them have been reported or examined in previous literatures, Table S8). A number of crucial pathways for cancer, e.g., NF-κB signaling pathway, Notch signaling pathway, and apoptosis, were influenced by these E3s (Figure 7A). Besides, these predicted hallmark ESIs show the "multi-to-multi" relations between E3s and substrates again. This information is important for E3s that may be taken as promising therapeutic targets for cancers. It is essential to make sure that drugs targeting on certain E3s will not lead to undesirable outcomes by disturbing unexpected ESIs for the multifunctional E3s.

Besides, we observed that BRCA1, a tumor suppressor with E3 activity, was also predicted as the substrate for multiple E3s when the prediction was conducted based on data of TCGA-BRCA (Figure 7A). However, some interactions were not high-confidence ones any more in other cancers (Figure 7B), implying the assumption that predicted results for mutant substrates can be cancer type specific.

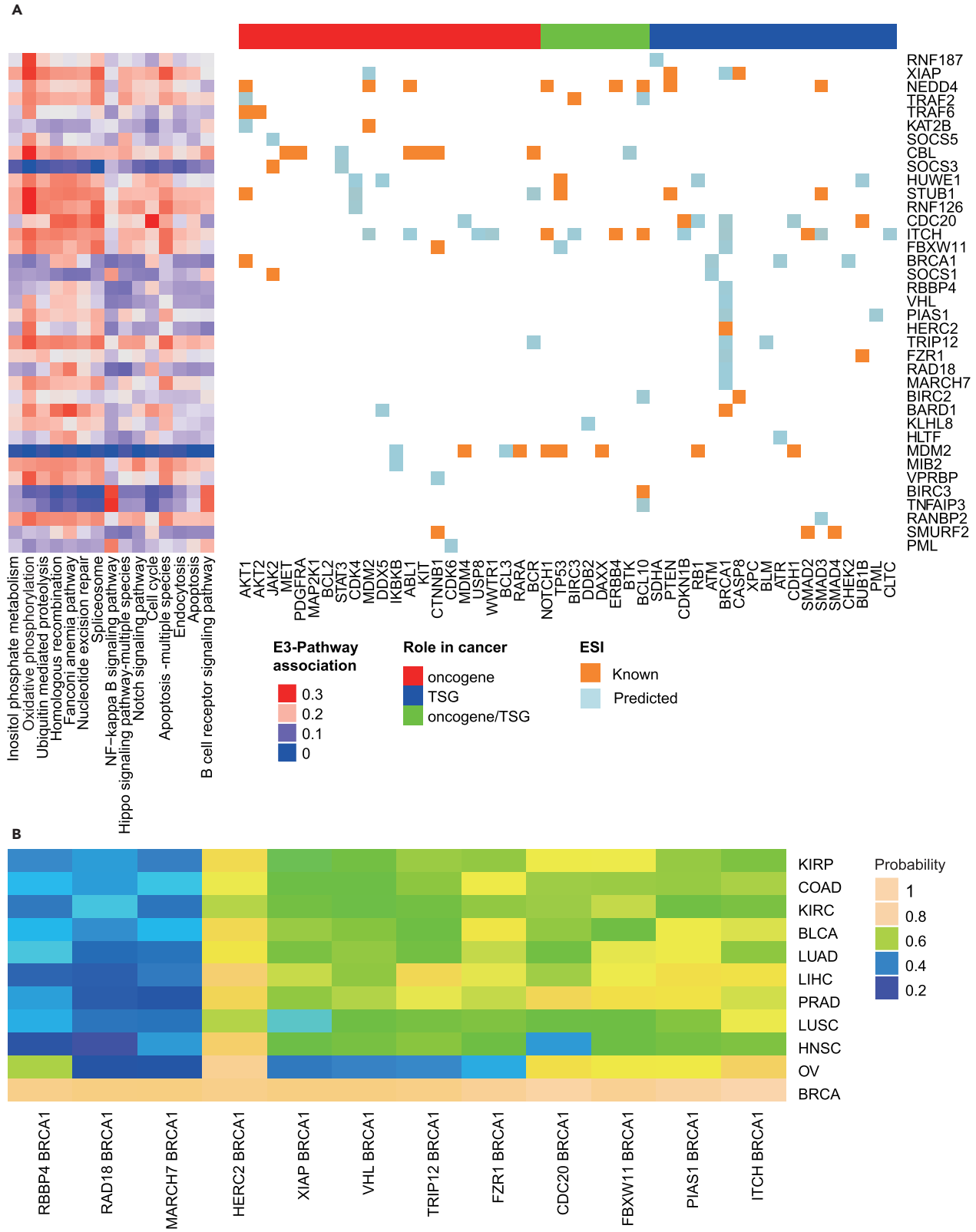

A
RNF187
XIAP
NEDD4
TRAF2
TRAF6
KAT2B
SOCS5
CBL
SOCS3
HUWE1
STUB1
RNF126
CDC20
ITCH
FBXW11
BRCA1
SOCS1
RBBP4
VHL
PIAS1
HERC2
TRIP12
FZR1
RAD18
MARCH7
BIRC2
BARD1
KLHL8
HLTF
MDM2
MIB2
VPRBP
BIRC3
TNFAIP3
RANBP2
SMURF2
PML
AKT1 AKT2 JAK2 MET PDGFRA MAP2K1 BCL2 STAT3 CDK4 MDM2 DDX5 IKBKB ABL1 KIT CTNNB1 CDK6 MDM4 USP8 WWTR1 BCL3 RARA BCR NOTCH1 TP53 BIRC3 DDB2 DAXX ERBB4 BTK BCL10 SDHA PTEN CDKN1B RB1 ATM BRCA1 CASP8 XPC BLM ATR CDH1 SMAD2 SMAD3 SMAD4 CHEK2 BUB1B PML CLTC
Inositol phosphate metabolism
Oxidative phosphorylation
Ubiquitin mediated proteolysis
Homologous recombination
Fanconi anemia pathway
Nucleotide excision repair
Spliceosome
NF−kappa B signaling pathway
Hippo signaling pathway-multiple species
Notch signaling pathway
Cell cycle
Apoptosis -multiple species
Endocytosis
Apoptosis
B cell receptor signaling pathway
E3-Pathway association
0.3
0.2
0.1
0
Role in cancer
oncogene
TSG
oncogene/TSG
ESI
Known
Predicted
B
KIRP
COAD
KIRC
BLCA
LUAD
LIHC
PRAD
LUSC
HNSC
OV
BRCA
Probability
1
0.8
0.6
0.4
0.2
RBBP4 BRCA1
RAD18 BRCA1
MARCH7 BRCA1
HERC2 BRCA1
XIAP BRCA1
VHL BRCA1
TRIP12 BRCA1
FZR1 BRCA1
CDC20 BRCA1
FBXW11 BRCA1
PIAS1 BRCA1
ITCH BRCA1

Figure 7. High-Confidence ESIs Predicted for Cancer Hallmark Substrates

(A) The left side shows the highly correlated pathways by which the E3s (rows of the right-side matrix) may exert their ubiquitination effects. The right side represents the high-confidence (prob>0.7) ESIs with cancer hallmark proteins as substrates, where the rows and columns stand for E3s and substrates, respectively.

(B) Predictions for BRCA1 across different cancers.

See also Figure S6, Tables S7 and S8.

DISCUSSION

Progressive accumulation in multi-omics data (Cancer Genome Atlas Research Network et al., 2013; Edwards et al., 2015) and prior biological knowledge (Chatr-Aryamontri et al., 2017; Kanehisa et al., 2017) allow for a data-driven investigation on ESIs. Here, we aimed to construct an ESI landscape and describe the association profiles between E3s and substrates by integrating different data resources.

Our study provides a glance at the association patterns of ESIs by combining multi-omics data and biological knowledge, where three types of negative control were taken into consideration. An initial scale-free ESI network composed of 1,806 reported ESIs was constructed, where plenty of cancer hallmark genes act as hubs, and the numbers of interacting substrates for different E3s vary considerably. It may suggest a general rule that some E3s are with a broad-spectrum function and that they can regulate various types of substrates, whereas the others only have effects on certain substrates. Unexpectedly, although a large fraction of ESIs will lead to substrate degradation (Figure 2C), E3s and substrates did not show significantly negative correlations in the omics-based associations (Figure 3A), suggesting that no consistent expression relation exists for ESIs and indicating the heterogeneity of ubiquitination mechanism. Nonetheless, integrating omics with network or pathway information greatly enhanced the differences between ESIs and NRDs, and the degradation effect emerged. Notably, the pathway-based associations can help identify downstream or upstream processes such as DNA replication (Faircloth et al., 2011) and mammalian target of rapamycin signaling pathway (Xia et al., 2016). The effectiveness and consistency of omics-driven association features for ESI identification also indicate the reusability of cancer omics-resources.

Different from the models utilized by UbiBrowser (Li et al., 2017) where only one negative category was considered, our model combined three negative categories covering "E3-random," "Other PPIs," and "E3-Indirect." Both crossover and independent validations have demonstrated the model's effectiveness, and the inclusion of three negative categories has significantly reduced the false discovery rate. Moreover, the false discoveries (e.g., SKP2-CUL1, FBXL19-UBE2T, Figures 5A and 5B) can be reduced further by removing E1s, E2s, and scaffold proteins in E3 complexes from candidate substrates, as most of them are already defined by studies on well-characterized E3s, like FBXW7 and MDM2. Taken together, our predictions can provide assistance on ESI recognition, but further experimental investigations are also required to confirm the interactions.

Our model also exhibits effectiveness for substrates that require certain forms of modifications before ubiquitination. Taking FBXL family as an illustration, our predictions output consistent results with a previous proteomic experimental study (Tan et al., 2013) and with a significantly improved recall (27/89 versus 5/89). When compared with UbiBrowser, our model also showed higher recall (27/89 versus 22/89), and the superiority was even higher in terms of calculable cases (27/39 versus 13/39), suggesting the advantage of our model, especially for substrates with knowledge background. Furthermore, a number of un-reported predictions of ESIs for SKP2 and FBXL6 were validated by experimental investigations. All the above imply the effectiveness of our prediction model.

In conclusion, our study provides a data-driven way to portray the ESI landscape, offering meaningful hypothesis on the latent E3-substrate association patterns, promising ESIs or even regulatory mechanism. For convenience, a website (http://www.esinet.dicp.ac.cn/home.php) is developed for browsing the multi-scaled association features of confirmed (1,806 pairs) and proteome-wide predicted (about 2,80,000 pairs) ESIs. It provides a valuable resource and assistance for further studies on protein ubiquitination.

Limitations of the Study

The data-driven prediction model's application scope is limited by currently available data resources, especially the proteomics data that is indispensable for most prediction-dependent multidimensional features. Although we utilized a relatively comprehensive proteomics database CPTAC wherein 10,602

proteins (about 400 E3s were included) were measured, some E3s like FBXL2, FBXL5, and FBXL22 were still not covered. We cannot calculate the multidimensional features for these items, thus lacking predictions on these E3s. As more proteomics data and network or pathway knowledge are being accumulated, we envision a continual optimization of the prediction results.

In addition, the fate of ubiquitinated substrates depends on the linkage of ubiquitin chains (Senft et al., 2018). Although most of the reported ESIs can lead to substrate degradation by K48 or K11 linkage, ubiquitination can also induce other outcomes, like activity or stability change. Our present prediction model mainly focuses on whether there is an interaction between an E3 and a substrate; however, the linkage-dependent ubiquitinated consequence was not considered. In our future studies, we will attempt to deduce the linkage information as well.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.05.033.

ACKNOWLEDGMENTS

We thank members of the Dr. Piao laboratory for helpful discussion. This study is supported by National Natural Science Foundation of China grants (Nos. 31701156 and 81672440). This project is funded by China Postdoctoral Science Foundation (No. 2017M611281), Innovation program of science and research from the DICP, CAS (DICP TMSR201601, DICP ZZBS201803), and the 100 Talents Program of Chinese Academy of Sciences. Dedicated to the 70th anniversary of Dalian Institute of Chemical Physics, CAS.

AUTHOR CONTRIBUTIONS

Conceptualization, H.-l.P., D.C., and X.L.; Methodology, D.C. and X.L.; Formal analysis, D.C.; Validation, X.L.; Investigation, T.X., D.S.T., W.W., H.C., T.L., C.L., Z.N., X.L., J.L., H.Q., and H.H.; Writing – Original Draft, D.C., X.L., and H,-l.P; Writing – Review & Editing, D.C., X.L., and H,-l.P.; Funding Acquisition, D.C. and H.-l.P.; Resources, D.C.; Data Curation, X.L.; Supervision, H.-l.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 21, 2017
Revised: April 14, 2019
Accepted: May 22, 2019
Published: June 28, 2019

REFERENCES

Barabasi, A.L. (2009). Scale-free networks: a decade and beyond. Science *325*, 412–413.

Bassermann, F., Eichner, R., and Pagano, M. (2014). The ubiquitin proteasome system - implications for cell cycle control and the targeted treatment of cancer. Biochim. Biophys. Acta *1843*, 150–162.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

Chan, C.H., Li, C.F., Yang, W.L., Gao, Y., Lee, S.W., Feng, Z.Z., Huang, H.Y., Tsai, K.K.C., Flores, L.G., Shao, Y.P., et al. (2012). The Skp2-SCF E3 ligase regulates Akt ubiquitination, glycolysis, herceptin sensitivity, and tumorigenesis. Cell *149*, 1098–1111.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. *45*, D369–D379.

Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., and Ketchum, K.A. (2015). The CPTAC data portal: a resource for cancer proteomics research. J. Proteome Res. *14*, 2707–2713.

Faircloth, L.M., Dorn, E., Carlile, C., and Cook, J.G. (2011). H2B K123 ubiquitination and H3 K4 methylation play a positive role in DNA replication. FASEB J. *25*, 896.5–896.5.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Gen, Y., Yasui, K., Kitaichi, T., Iwai, N., Terasaki, K., Dohi, O., Hashimoto, H., Fukui, H., Inada, Y., Fukui, A., et al. (2017). ASPP2 suppresses invasion and TGF-beta1-induced epithelial-mesenchymal transition by inhibiting Smad7 degradation mediated by E3 ubiquitin ligase ITCH in gastric cancer. Cancer Lett. *398*, 52–61.

Goymer, P. (2008). Network biology - why do we need hubs? Nat. Rev. Genet. *9*, 650.

Hammond-Martel, I., Yu, H., and Affar, E. (2012). Roles of ubiquitin signaling in transcription regulation. Cell. Signal. *24*, 410–421.

Hoeller, D., and Dikic, I. (2009). Targeting the ubiquitin system in cancer therapy. Nature *458*, 438–444.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.

Li, L.L., Li, W., Xiao, L.B., Xu, J., Chen, X., Tang, M., Dong, Z.G., Tao, Q., and Cao, Y. (2012). Viral oncoprotein LMP1 disrupts p53-induced cell cycle arrest and apoptosis through modulating K63-linked ubiquitination of p53. Cell Cycle *11*, 2327–2336.

Li, Y., Xie, P., Lu, L., Wang, J., Diao, L.H., Liu, Z.Y., Guo, F.F., He, Y.Z.G., Liu, Y., Huang, Q., et al. (2017). An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. Nat. Commun. *8*, 347.

Maddika, S., Kavela, S., Rani, N., Palicharla, V.R., Pokorny, J.L., Sarkaria, J.N., and Chen, J. (2011). WWP2 is an E3 ubiquitin ligase for PTEN. Nat. Cell Biol. *13*, 728–733.

Mani, A., and Gelmann, E.P. (2005). The ubiquitin-proteasome pathway and its role in cancer. J. Clin. Oncol. *23*, 4776–4789.

Nakayama, K., Nagahama, H., Minamishima, Y.A., Matsumoto, M., Nakamichi, I., Kitagawa, K., Shirane, M., Tsunematsu, R., Tsukiyama, T., Ishida, N., et al. (2000). Targeted disruption of Skp2 results in accumulation of cyclin E and p27(Kip1), polyploidy and centrosome overduplication. EMBO J. *19*, 2069–2081.

Rayburn, E., Zhang, R., He, J., and Wang, H. (2005). MDM2 and human malignancies: expression, clinical pathology, prognostic markers, and implications for chemotherapy. Curr. Cancer Drug Targets *5*, 27–41.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Scheffner, M., Nuber, U., and Huibregtse, J.M. (1995). Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade. Nature *373*, 81–83.

Senft, D., Qi, J., and Ronai, Z.A. (2018). Ubiquitin ligases in oncogenic transformation and cancer therapy. Nat. Rev. Cancer *18*, 69–88.

Shabbeer, S., Omer, D., Berneman, D., Weitzman, O., Alpaugh, A., Pietraszkiewicz, A., Metsuyanim, S., Shainskaya, A., Papa, M.Z., and Yarden, R.I. (2013). BRCA1 targets G2/M cell cycle proteins for ubiquitination and proteasomal degradation. Oncogene *32*, 5005–5016.

Skaar, J.R., Pagan, J.K., and Pagano, M. (2013). Mechanisms and function of substrate recruitment by F-box proteins. Nat. Rev. Mol. Cell Biol. *14*, 369–381.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Suresh, B., Lee, J., Kim, K.S., and Ramakrishna, S. (2016). The importance of ubiquitination and deubiquitination in cellular reprogramming. Stem Cells Int. *2016*, 6705927.

Tan, M.K., Lim, H.J., Bennett, E.J., Shi, Y., and Harper, J.W. (2013). Parallel SCF adaptor capture proteomics reveals a role for SCFFBXL17 in NRF2 activation via BACH1 repressor turnover. Mol. Cell *52*, 9–24.

Wang, Z., Zhang, C.B., Liu, X., Wang, Z.L., Sun, L.H., Li, G.Z., Liang, J.S., Hu, H.M., Liu, Y.W., Zhang, W., et al. (2016). Molecular and clinical characterization of PD-L1 expression at transcriptional level via 976 samples of brain glioma. OncoImmunology *5*, 5–11.

Xia, D., Qu, L.J., Li, G., Hongdu, B.Q., Xu, C.T., Lin, X., Lou, Y.X., He, Q.H., Ma, D.L., and Chen, Y.Y. (2016). MARCH2 regulates autophagy by promoting CFTR ubiquitination and degradation and PIK3CA-AKT-MTOR signaling. Autophagy *12*, 1614–1630.

Xue, M.Z., Liu, H.Y., Zhang, L.W., Chang, H.Y., Liu, Y.W., Du, S.W., Yang, Y.Q., and Wang, P. (2017). Computational identification of mutually exclusive transcriptional drivers dysregulating metastatic microRNAs in prostate cancer. Nat. Commun. *8*, 14917.

Ye, H., Zhang, X.H., Chen, Y.Q., Liu, Q., and Wei, J. (2016). Ranking novel cancer driving synthetic lethal gene pairs using TCGA data. Oncotarget *7*, 55352–55367.

Ye, X., Wang, L., Shang, B., Wang, Z., and Wei, W. (2014). NEDD4: a promising target for cancer therapy. Curr. Cancer Drug Targets *14*, 549–556.

Yumimoto, K., Matsumoto, M., Oyamada, K., Moroishi, T., and Nakayama, K.I. (2012). Comprehensive identification of substrates for F-box proteins by differential proteomics analysis. J. Proteome Res. *11*, 3175–3185.

Zhang, L., Huang, N.-J., Chen, C., Tang, W., and Kornbluth, S. (2012). Ubiquitylation of p53 by the APC/C inhibitor Trim39. Proc. Natl. Acad. Sci. U S A *109*, 20931–20936.

Zheng, N.N., Wang, Z.W., and Wei, W.Y. (2016). Ubiquitination-mediated degradation of cell cycle-related proteins by F-box proteins. Int. J. Biochem. Cell Biol. *73*, 99–110.

ISCI, Volume *16*

Supplemental Information

A Multidimensional Characterization of E3 Ubiquitin Ligase and Substrate Interaction Network

Di Chen, Xiaolong Liu, Tian Xia, Dinesh Singh Tekcham, Wen Wang, Huan Chen, Tongming Li, Chang Lu, Zhen Ning, Xiumei Liu, Jing Liu, Huan Qi, Hui He, and Hailong Piao

Supplemental Figures

Figure S1. E3-pathway correlations based on proteomics. Related to Figure 3**.** The core matrix exhibits the associations between E3s and pathways. For clarity, only the top-50 frequent pathways which can link E3s to substrates' pathways based on proteomics were listed out in the rows, and the corresponding top-50 correlated E3s were listed as columns. The left-most column shows the pathway category information. NUW annotates the Number of Ubiqutinated substrates on the pathWay. wcpM refers to the mean wcp (see Materials and methods) score across the top-50 E3s for a corresponding pathway.

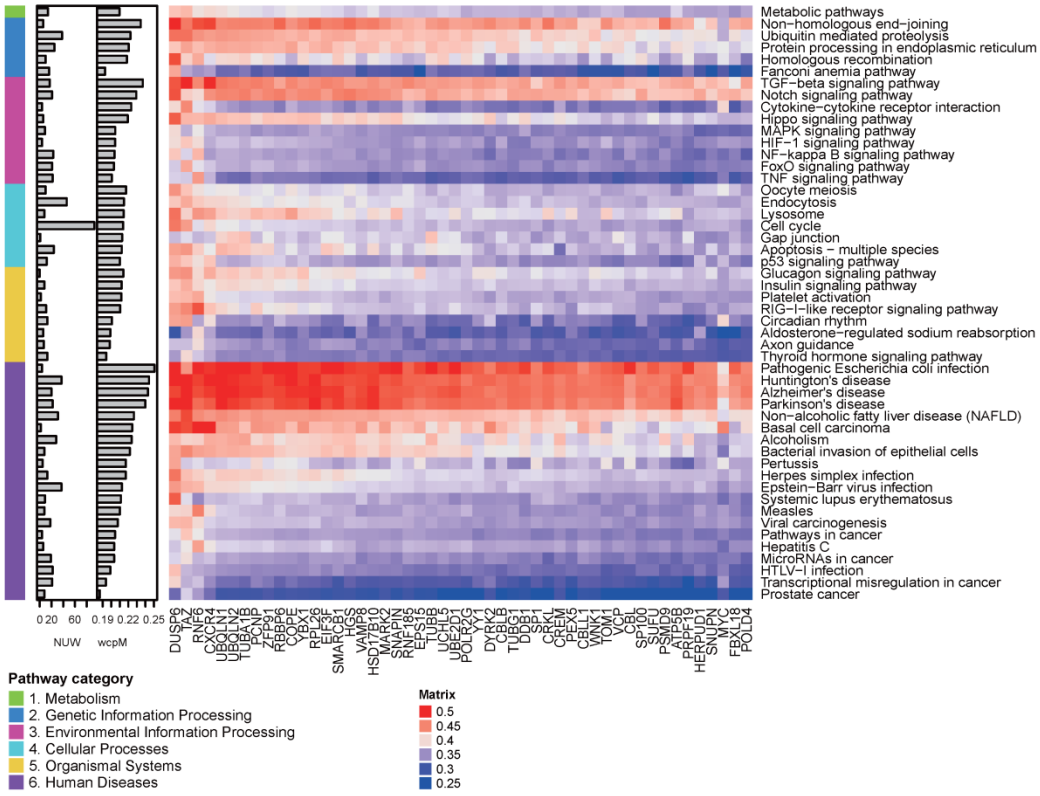


Figure S2. E3-pathway correlations based on transcriptomics. Related to Figure 3**.** The core matrix exhibits the associations between E3s and pathways. For clarity, only the top-50 frequent pathways which can link E3s to substrates' pathways based on transcriptomics are listed out in the rows, and the corresponding top-50 correlated E3s are listed as the columns. The left-most column shows the pathway category information. "NUW" annotates the Number of Ubiqutinated substrates on the pathWay. "wcrM" refers to the mean wcr (see Materials and methods) score across the top-50 E3s for a corresponding pathway.

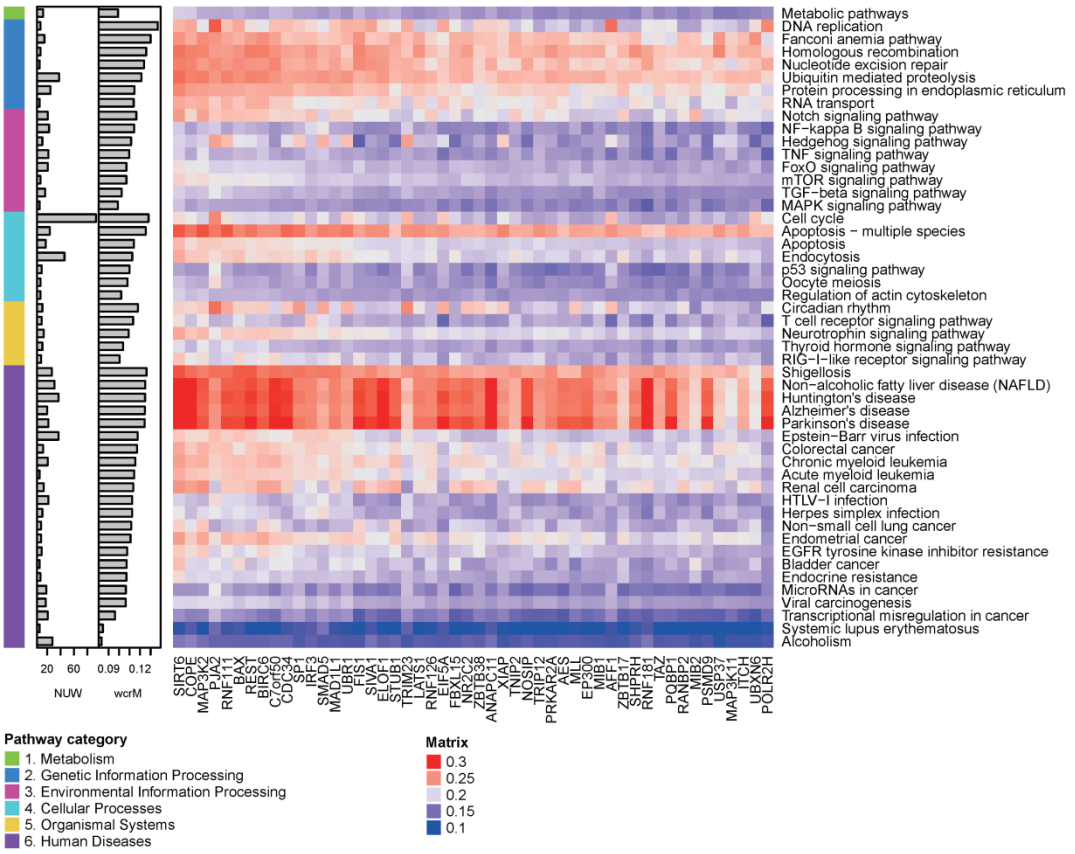


Figure S3. Pan-cancer based multidimensional association patterns of ESIs. Related to Figure 3**.** The association features were computed based on additional 10 tumors including Ovarian Serous Cystadenocarcinoma (OV) (A), Bladder Urothelial Carcinoma (BLCA) (B), Colon Adenocarcinoma (COAD) (C), Head and Neck Squamous Cell Carcinoma (HNSC) (D), Prostate Adenocarcinoma (PRAD) (E), Kidney Renal Clear Cell Carcinoma (KIRC) (F), Kidney Renal Papillary Cell Carcinoma (KIRP) (G), Liver Hepatocellular Carcinoma (LIHC) (H), Lung Adenocarcinoma (LUAD) (I), Lung Squamous Cell Carcinoma (LUSC) (J). Centers of boxes represent median values. Bottom and top bounds of boxes represent 25th and 75th percentiles. Whiskers mark 1.5 times of the interquartile range. Dots represent points falling outside this range. *: p<0.01, Wilcoxon-test for the differences comparing to the ESI dataset.

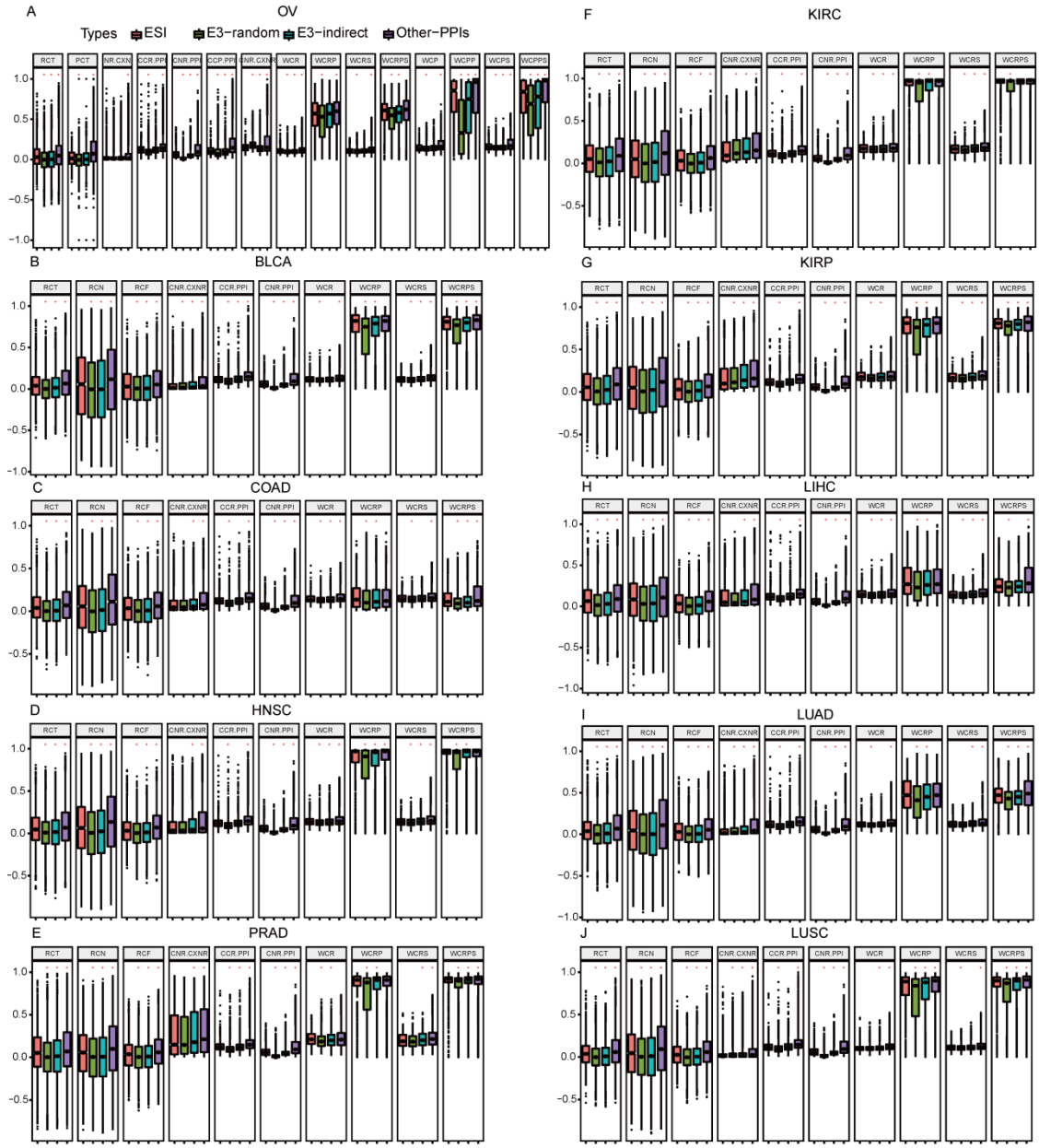


Figure S4. Performance of classifiers trained for FBXLs. Related to Figure 5**.** The same ESIs were used as positive samples, while (A) "E3-random", (B) "Other PPIs", and (C) "E3-Indirect" pairs were respectively taken as negative samples, and samples with FBXL proteins were removed. The horizontal and vertical axes represent the number of selected features and the corresponding average AUROC got from a 10-fold crossover validation. For each type of RFE-RFs, 5 classifiers were constructed by replacing negative samples. (D) The best performance features for each classifier in (A), (B) and (C).

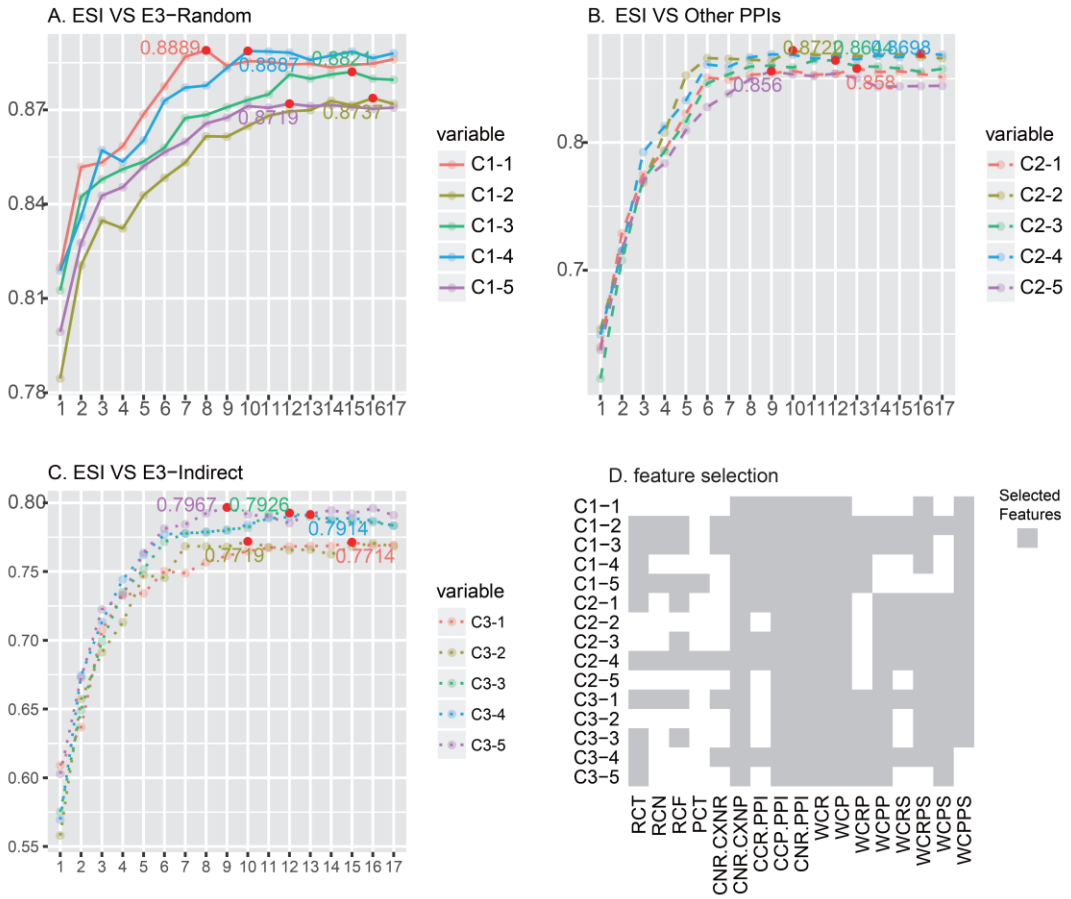


Figure S5. Experimental validations of high-confident ESIs. Related to Figure 6. (A) Validations on interactions between CCNA2 and FBXLs. HEK293T cells were transfected with SFB-tagged RFP, SKP2, FBXL6 and FBXL15. SFB-tagged proteins were purified with S protein beads and immunoblotted with antibodies against substrate of CCNA2 and FLAG. (B-D) Validations on interactions between substrates (B) Casp3-V5 was cotransfected with empty vector or SFB-SKP2 or SFB-FBXL6 into HEK293T cells. Western blotted with antibodies against FLAG and Caspase3. CDK4-V5 (C) or VDAC2 (D) were cotransfected with empty vector or SFB-FBXL6 into HEK293T cells. Western blotted with antibodies against FLAG and V5. (E) SFB-VDAC2 cotransfected with MYC-FBXL6, HA-Ub into HEK293T cells. Cells were treated with 20 μM MG132 for 4 h before collection. VDAC2 was purified with S-beads and western blotted with antibodies against HA, Myc and FLAG. All experiments have three biological replicates, and the most representative image is shown.

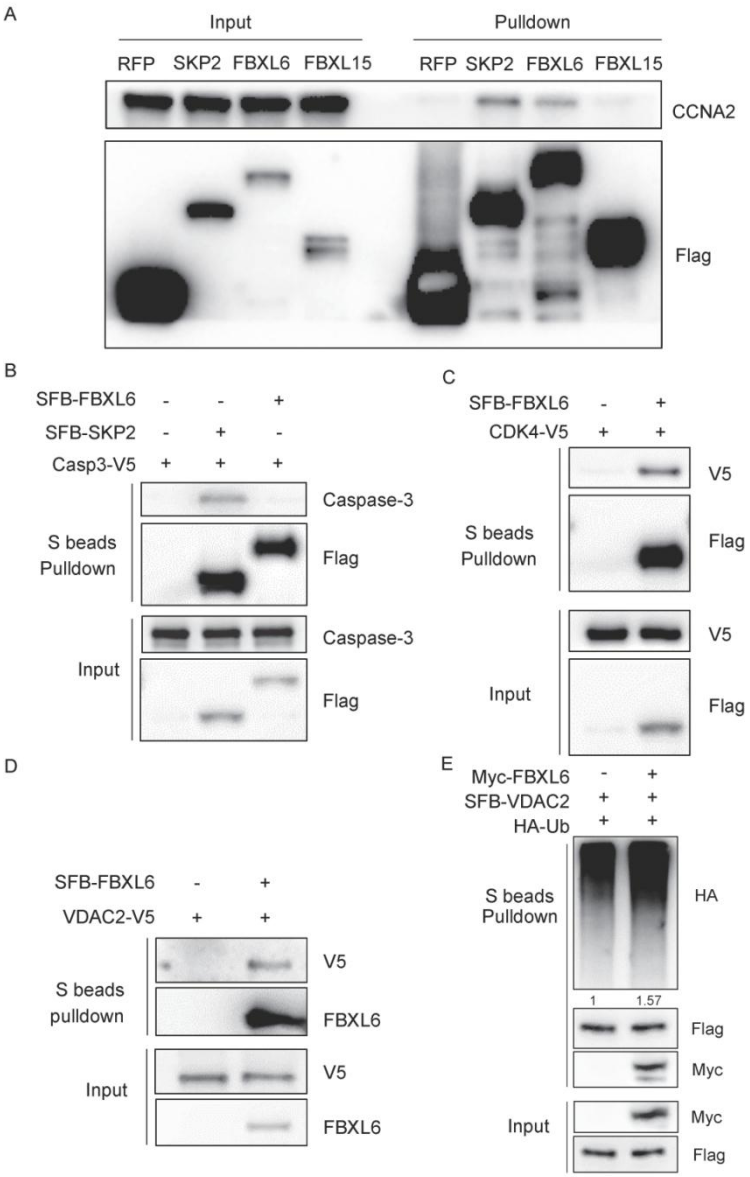
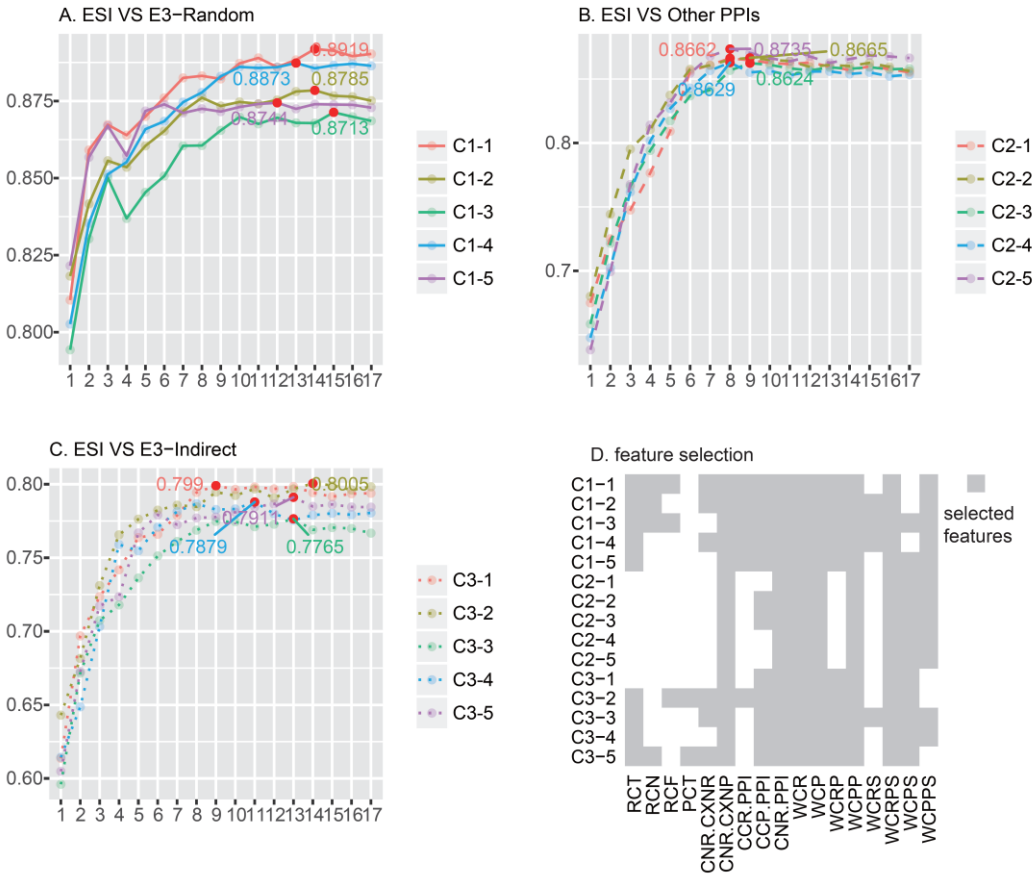


Figure S6. Performance of the final classifiers. Related to Figure 7**.** All calculable ESIs were included as positive samples and the same number of (A) "E3-random", (B) "Other PPIs", and (C) "E3-Indirect" pairs were respectively taken as negative samples. The horizontal and vertical axes represent the number of selected features and the corresponding average AUROC got from a 10-fold crossover validation. For each type of RFE-RFs, 5 classifiers were constructed by replacing different negative samples. (A) The best performance features for each classifier in (A), (B) and (C).



Supplemental Tables

Table S5. Performance of classifiers based on different cancers. Related to Figure 4**.** NC stands for the number of calculable ESIs based on the data of certain cancer. AUC1, AUC2 and AUC3 are respectively got from RFs trained by "E3-random","Other PPIs" and "E3-Indirect" samples. T: Transcriptomics; P: Proteomics.

Cancer	NC	AUC1	AUC2	AUC3	Omics
BRCA	576	0.87	0.86	0.79	T+P
OV	155	0.91	0.86	0.78	T+P
BLCA	808	0.85	0.75	0.74	T
COAD	808	0.84	0.75	0.71	T
HNSC	809	0.85	0.76	0.70	T
LIHC	809	0.85	0.77	0.74	T
LUAD	808	0.85	0.78	0.75	T
LUSC	809	0.87	0.79	0.74	T
PRAD	810	0.85	0.77	0.74	T
KIRC	811	0.85	0.77	0.69	T
KIRP	811	0.84	0.77	0.73	T

Transparent Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-flag (M2)	Sigma-Aldrich	F3165-5MG
Anti-V5	Invitrogen	R960-25
Anti-CDK4	Cell Signaling Technology	12790
anti-caspase-3	Cell Signaling Technology	9662S
Anti-DDB1	Proteintech	11380-1-AP
anti-cyclin A2	Proteintech	18202-1-AP
anti-HSP90AA1	Proteintech	13171-1-AP
anti-HSPD1	Proteintech	15282-1-AP
anti-VDAC1/2	Proteintech	10866-1-AP
anti-HA	Proteintech	51064-2-AP
anti-Myc	Proteintech	16286-1-AP
Deposited Data		
Transcriptomics data	Cancer Genome Atlas Research et al., 2013	http://cancergenome.nih.gov/
Proteomics data	Edwards et al., 2015	https://cptac-data-portal.georgetown.edu/cptacPublic/
Experimental Models: Cell Lines		
HEK293T	American Type Culture Collection	HEK293T
Software and Algorithms		
R		https://www.r-project.org
Cytoscape	Shannon et al., 2003	http://www.cytoscape.org/
R package of Random forest	Liaw, 2002	
R package of caret	Kuhn, 2008	
R package of DMwR	Torgo, 2010	
Other		
E3Net	Han et al., 2012	http://pnet.kaist.ac.kr/e3net/
hUbiquitome	Du et al., 2011	http://202.38.126.151/hmdd/hubi/

Uniprot	The UniProt, 2017	http://www.uniprot.org/
BioGrid	Chatr-Aryamontri et al., 2017	https://thebiogrid.org/
KEGG	Kanehisa et al., 2017	http://www.kegg.jp/
UbiBrowser	Li et al., 2017	http://ubibrowser.ncpsb.org/home/index
Website of known and predicted ESIs	This paper	http://www.esinet.dicp.ac.cn/home.php
Association features of ESIs and three negative categories	This paper	http://www.esinet.dicp.ac.cn/download.php

ESI collection

To be as comprehensive as possible, a positive reference dataset of ESIs were mainly integrated from four databases: E3Net(Han et al., 2012), hUbiquitome (Du et al., 2011), Uniprot (The UniProt, 2017) and BioGrid (Chatr-Aryamontri et al., 2017). (1) E3Net mainly collected the ESIs from the MEDLINE literatures and UniProt by a text mining method, which extracted E3s and substrates with definite textual description on the specificities between E3s and their substrates. We extracted all of the E3-substrate pairs from E3Net as the main part of the positive dataset. (2) From another text-mining based database hUbiquitome, which collected experimentally validated protein ubiquitination cascades (E1-E2-E3-substrate cascades), we also extracted the E3-substrate interactions. (3) Integrating all the E3s recorded by E3Net, hUbiquitome as well as UbiNet (Nguyen et al., 2016), we got 589 possible E3s. To complement newly-updated ESIs, we searched the Uniprot (release-2017_06) for these E3s and extracted E3-substrate pairs which were with definite textual description that certain protein was ubiquitinated by certain E3. (4) Additional ESIs were extracted from the physical PPIs recorded in BioGrid (version: BIOGRID-MV-Physical-3.4.152) by strict filtering rules: with at least one E3, recorded as "physical interactions", recognized by "Low-Throughput" methods, modified by "Ubiquitination", and with explicit literature evidences. Besides, ESIs summarized in a previous review (Wang et al., 2014) were also included. In total, 1806 pairs of E3s and substrates (Table S1) in human were collected as the positive reference dataset.

Control set collection

Three categories of negative cases were prepared. The first one ("E3-random") contained 10000 pairs of randomly paired E3s and proteins (Table S2). The second one ("Other PPIs") incorporated 10000 pairs of randomly selected ubiquitination-independent PPIs from BioGrid (Table S3). The last one ("E3-Indirect") encompassed 10000 pairs of indirect E3 regulation

relations which were randomly sampled from the two-step distant E3s and their indirectly connected proteins on the PPI network (Table S4). During sampling, known ESIs and redundant cases were excluded.

Cancer hallmark subgraph extraction

The cancer hallmark subgraph (Figure 2C) was extracted as a representative profile to show part of the ESI network. To highlight the functional mechanism of ESIs in cancer, this subgraph was confined to two requirements. Firstly, the nodes should be recorded in the COSMIC database which focuses on cancer hallmark genes. Secondly, literature investigations were conducted to obtain the ubiquitination fate of the substrate on each edge, retained edges should be annotated with explicit ubiquitination fate, either degradation or non-degradation, for corresponding ESIs (see Table S1 for specific evidences of different edges).

Omics data preparation

For transcriptomics, normalized RNA-seq data of different cancers in TCGA (Cancer Genome Atlas Research et al., 2013) were downloaded by an R package – TCGAbiolinks (Colaprico et al., 2016). For each cancer, the RNA-Seq data contained mRNA expressions of 20501 genes across tumor and normal tissues, a fraction of the tumor tissues can be matched with normal ones which were taken from the same patients. Proteomics data for BRCA and OV were downloaded from CPTAC (Edwards et al., 2015), where the protein expressions were quantified by an isobaric peptide labeling approach (iTRAQ). For BRCA, 10599 proteins were quantified, while 6160 proteins were measured for OV.

Omic-based associations

Four forms of omic-based associations were considered. PCT (A, B) / RCT (A, B) calculated the spearman correlation coefficient (SCC) between A and B based on their protein / mRNA expressions in tumor tissues. RCN (A, B) calculated the SCC between A and B based on the mRNA expressions in normal tissues. RCF (A, B) calculated the SCC between A and B regarding the fold change of mRNA-expressions in tumor tissues relative to matched normal tissues.

Network-based associations

Three networks were considered: the PPI network was composed of 302868 experimentally-obtained PPIs in BioGrid; the mRNA/protein co-expression network was constructed by connecting two genes/proteins with SCCs larger than certain threshold (selected as 0.3, since the most significant difference between PRDs and NRDs will be observed based on the threshold of 0.3).

Two forms of network-based associations were computed. On one hand, the common neighbor rate (CNR) between protein *a* and *b* on network *G* was calculated as:

$$CNR_G(a,b) = \frac{|Neis(a,G)| \bigcap |Neis(b,G)|}{\sqrt{|Neis(a,G)| \cdot |Neis(b,G)|}}$$

where *Neis* (*,*G*) represented the neighbors of node * on network *G*. On the other hand, co-expression and PPI networks were integrated, and the common neighbor rate of co-expressed factors on PPI network was computed:

$$CCR.PPI(a,b) = \frac{|Neis(HN(a,G_{CXNR}),G_{PPI}) | \bigcap |Neis(HN(b,G_{CXNR}),G_{PPI})|}{\sqrt{|Neis(HN(a,G_{CXNR}),G_{PPI}) | \cdot |Neis(HN(b,G_{CXNR}),G_{PPI}) |}}$$

$$CCP.PPI(a,b) = \frac{|Neis(HN(a,G_{CXNP}),G_{PPI}) | \bigcap |Neis(HN(b,G_{CXNP}),G_{PPI})|}{\sqrt{|Neis(HN(a,G_{CXNP}),G_{PPI}) | \cdot |Neis(HN(b,G_{CXNP}),G_{PPI}) |}}$$

where G_{CXNR}/G_{CXNR} denoted the mRNA/protein co-expression network, *HN* (*,*G*) represented the highly co-expressed neighbors of node * on network *G*, and only the top-10 co-expressed factors were retained.

Pathway-based associations

Pathway-based associations were applied to estimate whether one E3 was highly correlated with certain pathway its substrates belong to, or vice-versa. The correlation between protein *a* and pathway *P* with *N* member genes was calculated as:

$$wcr(a,P) = \frac{\sum_{m \in P, m \neq a} |SCC_{mRNA}(a,m)|}{N}$$

$$wcp(a,P) = \frac{\sum_{m \in P, m \neq a} |SCC_{protein}(a,m)|}{N}$$

where *m* is a member gene of pathway *P*, *wcr* and *wcp* were respectively calculated from transcriptomics and proteomics data; particularly, if protein *a* is also a member of pathway *P*, it will be eliminated from the pathway during calculation.

Depending on pathways of E3 and substrate respectively, four forms of pathway-based associations were generated:

$$WCR(a,b) = \max\{wcr(a,P_i) | b \in P_i\}$$

$$WCP(a,b) = \max(wcp(a,P_i) | b \in P_i)$$

$$WCRS(a,b) = \max\{wcr(b,P_j) | a \in P_j\}$$

$$WCPS(a,b) = \max\{wcp(b,P_j) | a \in P_j\}$$

where *WCR* and *WCRS* integrated transcriptomics with pathways, *WCP* and *WCPS* integrated

proteomics with pathways.

Permutation test was applied to examine the specificity of pathway-based associations. Taken *WCR* (*a*, *b*) as an illustration, if pathway P_{max} has the highest correlation with protein *a* among all known pathways of protein *b*, we randomly permutated the member genes of P_{max} by the same number of other genes, and calculated a new *wcr* (*a*, $P_{max'}$), repeated this process 100 times, and the ratio of decreased scores was calculated as *WCRP* (*a*, *b*). The higher *WCRP* (*a*, *b*) was, the less likely *b* was related with *a* through a pathway with the same size of P_{max} by chance. Likewise, *WCPP*, *WCRSP*, *WCPSP* were calculated as well.

Feature selection

Integration of 4 omic-based, 5 network-based and 8 pathway-based association features generated a 17-dimensional feature description for all pairwise proteins. To identify which features are more effective in distinguishing ESIs from NDRs, we used recursive feature elimination (RFE) algorithm (Guyon et al., 2002), in which each feature was ranked according to its importance for feature selection. It was carried out based on an R package of caret (Kuhn, 2008).

Classification algorithm

Together with the RFE feature selection, Random forest (RF) (Breiman, 2001) was applied to examine feature importance and train classifiers. This algorithm was implemented by an R package of randomForest (Liaw, 2002). Basically, three types of reverse-feature-reduction based random forest classifiers (RFE-RFs) were trained, where the positive references were the same, but the negative ones were from three different categories respectively. To encompass more negative samples, for each type of RFE-RFs, 5 classifiers were built using different negative samples.

Cross-over validation and independent validation

The model performance was quantified by AUROC (Lasko et al., 2005) through both 10-fold cross-over validations and an independent validation. During the 10-fold cross-over validation, to keep sample balance, the same number of ESIs and certain type of NRDs were randomly selected from the collected ESIs and NRDs as the cross-over validation set which was separated into 10 independent parts, each single part should be used as the testing dataset once, and the other 9 were used to train the model, thus the training and testing were conducted 10 times. Since the number of NRDs was extremely larger than ESIs, the cross-over validation was repeated 5 times by replacing non-redundant NRDs. Regarding three categories of NRDs, three types of classifiers were trained and tested by the cross-over validations. From the rest samples, another dataset composed of certain number of ESIs, "E3-random" pairs, "Other PPIs", and "E3-indirect" pairs was collected as the independent validation set.

F3-Score

F-Score (Maratea et al., 2014) is utilized to estimate the performance of a probability threshold for discriminating ESIs in predictions. It is computed as:

$$\mathrm{F}_{\beta} = (1+\beta^2)\frac{\mathrm{pre}cision \cdot recall}{\beta^2 \cdot precision + recall}$$

where β was defined as 3 here, weighing recall higher than precision, thus *F3* was computed. The rationale lies in that the number of possible NRDs was extremely higher than ESIs which were only partially known, emphasis on recall will avoid the overlook of novel predicted ESIs.

Imputation methods for missing values

Before prediction, we calculated the multi-dimensional features for candidate samples; however, due to data limitation, some samples could not be fully-described by 17 dimensional features. For a sample with missing features, if its missing ratio was less than 30%, the KNN imputation algorithm which predicted missing values according to weighted average of the K nearest neighbors was employed to fill the missing value; otherwise, the samples were omitted from predictions. The KNN imputation (distances to others were calculated based on only known features) was performed by R package of DMwR (Torgo, 2010).

Predict novel ESIs

Candidate ESIs were prepared before prediction. For an investigated E3, 6686 genes which were included in at least one pathway and measured in transcriptomics or proteomics were paired with the E3 as candidate ESIs; for an investigated substrate, 589 E3s that we have collected (Table S7) were paired with the substrate as candidate ESIs.

Three types of classifiers with the purpose to distinguish ESIs from different categories of NRDs were trained before prediction, and each type contained 5 parallel repetitions (each repetition was trained by replacing non-redundant NRDs within the same category). For the candidate ESIs, we calculated the 17-dimensional association features, and filled up missing values. Then, an ensemble classification model was applied on these candidates to predict the probability of being ESI as:

$$\mathrm{Prob}(a,b) = \frac{1}{3}\sum_{i\in(1,2,3)}\sum_{k=1}^{5} auc_{i,k} \cdot \sum_{\mathrm{k}=1}^{5} prob_{i,k}$$

where *i* represents the type of classifiers, *k* represents the repetition, *auc* represents AUROC of a classifier, and $prob_{i,k}$ was the predicted probability for the input pair based on the *k*-th classifier of type *i*.

Experimental antibodies

Anti-flag (M2) (F3165-5MG, 1:3,000 dilution) monoclonal antibody was obtained from Sigma-Aldrich. Anti-V5 (R960-25, 1:1,000 dilution) monoclonal antibody was purchased from Invitrogen. Anti-CDK4 (12790, 1:1,000 dilution) and anti-caspase-3 (9662S, 1:1,000 dilution)

polyclonal antibodies were purchased from Cell Signaling Technology. Anti-DDB1 (11380-1-AP, 1:1,000 dilution), anti-cyclin A2 (18202-1-AP, 1:1,000 dilution), anti-HSP90AA1 (13171-1-AP, 1:1,000 dilution), anti-HSPD1 (15282-1-AP, 1:1,000 dilution), anti-VDAC1/2 (10866-1-AP, 1:1,000 dilution), anti-HA (51064-2-AP, 1:2,000 dilution) and anti-Myc (16286-1-AP, 1:2,000 dilution) polyclonal antibodies were purchased from Proteintech.

Plasmids constructs, cell culture and transfection

All the expression vectors used in this study (including SFB-SKP2, SFB-FBXL6, SFB-FBXL15, SFB-CCNA2, SFB-VDAC2, Caspase3-V5, CDK4-V5, VDAC2-V5, Myc-SKP2 and Myc-FBXL6) were generated by polymerase chain reaction (PCR) and then subcloned into pDONR221 vector as the entry clones using Gateway Technology (Invitrogen). Subsequently, the entry clones were recombined into gateway destination vectors fused with various tags (SFB, V5 or Myc). The FBXL6 shRNAs were purchased from Dharmacon. HEK293T cells were purchased from American Type Culture Collection (ATCC) and cultured at 37°C under humidified air containing 5% CO_2 in Dulbecco's modified Eagle's medium (Gibco, Rockville, MD) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin and streptomycin antibiotics. Plasmid transfection was performed with the polyethylenimine (PEI) reagent as previously described (Longo et al., 2013).

Tandem affinity purification of SFB-tagged protein complex

For affinity purification experiments, HEK293T cells were harvested after transfection for 48 h and subjected to lysis in NETN buffer (100 mM NaCl, 0.5 mM EDTA, 20 mM Tris-Cl, 0.5% Nonidet P-40) with protease and phosphatase inhibitors at 4°C for 1 h. Supernatants were obtained after centrifugation (13,000 rpm) at 4°C for 15 min and then incubated with S protein beads (Millipore) overnight at 4°C. The beads were washed three times with NETN buffer at room temperature. Bound protein complex were released from the beads after boiling with 4×SDS loading buffer for 10 min and applied for Western blot analysis.

In vivo ubiquitination assay and half-life assay

HEK293T cells were plated in 10-cm dishes and cultured until cell confluence came up to 80%. The cells were transfected with the indicated plasmids and incubated for 36 h and then treated with MG132 (20 μM) for 5 h prior to collection. The cells were lysed in NETN buffer. Lysates were incubated with S protein beads for affinity purification as described earlier. Western blot was applied onto the pulldown protein complex for ubiquitination analysis of predicted substrates. For the half-life assay, HEK293T cells were transfected with empty vector (EV), SFB-SKP2 or SFB-FBXL6 along with CCNA2-V5. Forty-eight hours later, cells were treated with cycloheximide (CHX, 100 μM) for the indicated time points, and cell lysates were probed with CCNA2 antibody.

Network visualization.

Network visualization was performed with Cytoscape (Shannon et al., 2003).

Statistical analysis.

Differences between ESIs and NRDs on each association feature were examined by Wilcoxon-test, and differences between node degrees of E3s and substrates were examined by t-test. Statistical test and other computations were all conducted by R.

Data and software availability.

Most data that support our conclusion have been contained in supplemental tables or website(http://www.esinet.dicp.ac.cn/home.php). The multidimensional association information can be browsed for both confirmed and inferred ESIs on the website. Other data produced by different stages of the computational processes are also available upon request. All computational codes are available on request.

Supplemental References

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5-32.

Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113-1120.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A.*, et al.* (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. *45*, D369-D379.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I.*, et al.* (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. *44*, e71.

Du, Y., Xu, N., Lu, M., and Li, T. (2011). hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. Database (Oxford) *2011*, bar055.

Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., and Ketchum, K.A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J. Proteome Res. *14*, 2707-2713.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Mach. Learn. *46*, 389-422.

Han, Y., Lee, H., Park, J.C., and Yi, G.S. (2012). E3Net: a system for exploring E3-mediated regulatory networks of cellular functions. Mol. Cell. Proteomics *11*, O111 014076.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1-26.

Lasko, T.A., Bhagwat, J.G., Zou, K.H., and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. J. Biomed. Inf. *38*, 404-415.

Liaw, A.W., M. (2002). Classification and Regression by randomForest. R News *2*, 18-22.

Longo, P.A., Kavran, J.M., Kim, M.S., and Leahy, D.J. (2013). Transient mammalian cell transfection with polyethylenimine (PEI). Methods Enzymol. *529*, 227-240.

Maratea, A., Petrosino, A., and Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. Inform. Sciences *257*, 331-341.

Nguyen, V.N., Huang, K.Y., Weng, J.T., Lai, K.R., and Lee, T.Y. (2016). UbiNet: an online resource for exploring the functional associations and regulatory networks of protein ubiquitylation. Database (Oxford) *2016*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498-2504.

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158-D169.

Torgo, L. (2010). Data Mining with R, learning with case studies (Chapman and Hall/CRC).

Wang, Z.W., Liu, P.D., Inuzuka, H., and Wei, W.Y. (2014). Roles of F-box proteins in cancer. Nat. Rev. Cancer *14*, 233-247.