UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation

Tianshun Gao[1], Zexian Liu[2], Yongbo Wang[1], Han Cheng[1], Qing Yang[1], Anyuan Guo[1], Jian Ren[3] and Yu Xue[1,*]

[1]Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Luoyu Road 1037, Wuhan, Hubei 430074, [2]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Huangshan Road 443, Hefei, Anhui 230027 and [3]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Xinggang West Road 135, Guangzhou, Guangdong 510275, China

ABSTRACT

In this work, we developed a family-based database of UUCD (http://uucd.biocuckoo.org) for ubiquitin and ubiquitin-like conjugation, which is one of the most important post-translational modifications responsible for regulating a variety of cellular processes, through a similar E1 (ubiquitin-activating enzyme)–E2 (ubiquitin-conjugating enzyme)–E3 (ubiquitin-protein ligase) enzyme thioester cascade. Although extensive experimental efforts have been taken, an integrative data resource is still not available. From the scientific literature, 26 E1s, 105 E2s, 1003 E3s and 148 deubiquitination enzymes (DUBs) were collected and classified into 1, 3, 19 and 7 families, respectively. To computationally characterize potential enzymes in eukaryotes, we constructed 1, 1, 15 and 6 hidden Markov model (HMM) profiles for E1s, E2s, E3s and DUBs at the family level, separately. Moreover, the ortholog searches were conducted for E3 and DUB families without HMM profiles. Then the UUCD database was developed with 738 E1s, 2937 E2s, 46 631 E3s and 6647 DUBs of 70 eukaryotic species. The detailed annotations and classifications were also provided. The online service of UUCD was implemented in PHP + MySQL + JavaScript + Perl.

INTRODUCTION

The 2004 Nobel Prize in Chemistry was awarded to Aaron Ciechanover, Avram Hershko and Irwin Rose for their seminal discovery of the ubiquitin conjugation that targets proteins for degradation in an ATP-dependent manner (1–4). Previously, ubiquitin was isolated as a heat-stable protein with 76 aa (3), while further analyses revealed that the 'ubiquitin kiss' functions as a molecular death-tag through the ubiquitination, which reversibly and covalently forms an isopeptide bond between the C-terminal (–GG) carboxyl group of a ubiquitin protein and the ε-amino group of lysine residues or, less commonly, other types of residues of a substrate protein (4,5). The protein substrates can be modified by conjugating with mono- or polyubiquitins (4). Beyond degradation, non-proteolytic functions of ubiquitin conjugation were also demonstrated (6,7). For example, the K63-linked ubiquitin chain is implicated in DNA repair and endocytosis (7), whereas monoubiquitination plays an important role in histone regulation, virus budding and DNA repair (6,7).

The enzymatic process of the ubiquitin conjugation is a sequential three-step cascade (4). First, the C-terminal glycine (G) of ubiquitin is activated by a ubiquitin-activating enzyme (E1) to form an E1–Ub thioester in an ATP-dependent manner. Then the activated ubiquitin is transferred to the active site cysteine (C) residue of a ubiquitin-conjugating enzyme (E2). Finally, a ubiquitin-protein ligase (E3) removes the ubiquitin from the E2 and forms a covalent link between ubiquitin and a protein substrate (4). Also, ubiquitination is reversible by numerous deubiquitination enzymes (DUBs) (8). Analogous to ubiquitin, other components of superfamily of ubiquitin-like modifiers, such as SUMO, ISG15, Rub1/Nedd8 and Atg8/12 are also adopted the similar catalytic procedures (9). The ubiquitin and ubiquitin-like conjugation regulate a large number of cellular processes, such as cell cycle, signal transduction, apoptosis and autophagy (4,9), while aberrant modification is implicated in numerous pathologies, such as neurodegenerative disorders, inflammatory diseases and cancers (7,10). In this regard, identification of E1s, E2s, E3s and DUBs is

fundamental for understanding the regulatory roles of ubiquitin and ubiquitin-like conjugation and provides candidate drug targets for further biomedical consideration (11).

After the first E1 ligase was isolated in 1981 (12), identification and classification of E1s, E2s, E3s and DUBs had emerged to be a great challenge. In 2003, with the completion of RIKEN FANTOM-2 project, the total ubiquitin-associated enzymes were identified as 835, 764, 320, 785, 162 and 145 for *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*, respectively (13). In particular, there were 16 E1s, 53 E2s, 527 E3s and 74 DUBs characterized in *H. sapiens* and roughly classified by InterPro domains (13). Later, the number of human putative DUBs was refined as 95 and further classification was based on phylogenetic relations (8). In 2003, the first public database of PlantsUBQ was constructed with E1s, E2s and E3s in *Arabidopsis* (14). Later, the SCUD database contained 11 E2s, 42 E3s, 20 DUBs and 940 substrates in *S. cerevisiae* (15). In 2009, the plantsUPS was developed with E1s, E2s and E3s in seven plants (16). Recently, the hUbiquitome database collected 1 E1, 12 E2s, 138 E3s, 17 DUBs and 279 substrates of *H. sapiens* from the scientific literature (17). Although a number of studies were performed, an integrative data resource is still not available.

In this work, we manually collected 26 E1s, 105 E2s, 1003 E3s and 148 DUBs for ubiquitin and ubiquitin-like conjugation systems from the scientific literature (Table 1). Based on previously established rationales (8,9,18–23), we classified E1s, E2s and DUBs into one, three and seven families, respectively. Also, we classified E3s into a hierarchical structure with five levels, including class, group, subgroup, family and single E3. Totally, we obtained 2 classes, 7 groups, 4 subgroups and 19 families for E3s. With HMMER (24), we totally constructed 1, 1, 15 and 6 hidden Markov model (HMM) profiles for E1s, E2s, E3s and DUBs at the family level, respectively. Then we used the HMM profiles to computationally characterize 738 E1s, 2797 E2s, 43 881 E3s and 6516 DUBs from 70 eukaryotic organisms, respectively. For families without HMM profiles, we additionally conducted an ortholog search to detect 140 E2s, 2750 E3s and 131 DUBs. The classification information was provided, while the detailed annotations from Ensembl (25) and UniProt (26)

databases were also integrated. Finally, a comprehensive database of ubiquitin and ubiquitin-like conjugation (UUCD) was developed with 56 949 enzymes for ubiquitin and ubiquitin-like conjugation across 70 eukaryotic species.

CONSTRUCTION AND CONTENT

Data collection and curation

Although there are up to 17 types of ubiquitin-like proteins that can covalently modify other molecules, only nine of them specifically recognize protein substrates (9). Thus, we searched the PubMed with multiple keywords, including 'ubiquitin', 'ubiquitination', 'ubiquitylation', 'SUMO', 'sumoylation', 'NEDD8', 'neddylation', 'Atg8', 'Apg8', 'Atg12', 'Apg12', 'Urm1', 'ISG15', 'isgylation', 'UFM1' and 'FAT10', respectively. Totally, we collected 26 E1s, 105 E2s, 1003 E3s and 148 DUBs for ubiquitin and ubiquitin-like conjugation in eukaryotes (Table 1 and Supplementary Table S1). From Ensembl (25) and UniProt (26) databases, the full-length sequences of the enzymes were obtained. The functional domain information was taken from the annotations in the UniProt database and further verified by searching the Pfam database (27). Moreover, the complete proteome sequences were downloaded for 70 eukaryotes, including 54 animals, 14 plants and 2 fungi, from Ensembl (release version 64, http://www.ensembl.org/), EnsemblPlants (release version 11, http://plants.ensembl.org/) and EnsemblFungi (release version 11, http://fungi.ensembl.org/), respectively (25).

The Classification

As previously described, the known E1s can be distinguished from other proteins by the occurrence of ThiF/MoeB domains (9). Thus, all E1s were classified into a single family. Also, the E2s were categorized into three families based on their functional domains, including UBC (ubiquitin-conjugating), UEV (ubiquitin-enzyme variant) and Other (unclassified) (18). Moreover, DUBs for ubiquitin were classified into five families, including UCH (ubiquitin C-terminal hydrolase), USP (ubiquitin-specific protease), OTU (ovarian tumor), Josephin (Machado–Joseph disease, MJD) and JAMM (JAB1/MPN/Mov34 metalloenzyme) (8). In addition, the DUBs

Table 1. The data statistics of known E1, E2, E3 and DUB proteins

Organism	E1	E2	E3	DUB	Total
Homo sapiens	8	39	475	91	613
Mus musculus	2	6	71	14	93
Drosophila melanogaster	0	5	46	4	55
Caenorhabditis elegans	5	6	57	2	70
Saccharomyces cerevisiae	3	16	78	19	116
Schizosaccharomyces pombe	2	4	39	6	51
Arabidopsis thaliana	5	27	182	10	224
Others	1	2	55	2	60
Total	26	105	1003	148	1282

From the scientific literature, we manually collected experimentally identified E1s, E2s, E3s and DUBs, respectively.

for ubiquitin-like conjugations, such as SENPs for SUMO, were classified into the family of ULP (ubiquitin-like protease). The Other family contained unclassified DUBs (8).

The classification of E3 ligases is complicated, because a considerable number of proteins participate in forming E3 ligase complexes as adaptors without E3 activities (19–23). In this regard, we first classified E3-associated proteins into two classes as E3 activity and E3 adaptor (19–22). Enzymes in the E3 activity class were further categorized into four groups and seven families, including HECT (with one family of HECT), RING (two families as RING and U-box), N-recognin (three families as UBR-box, N-domain and Other) and Other (with one family of Other). Proteins in the E3 adaptor class were classified into 2 groups and 12 families, including Cullin RING (10 families as Cullin, F-box, SKP1, SOCS/VHL/BC-box, Elon-B/C, BTB_3-box, BTB_Other, DWD, DDB1 and Other) and APC/C (two families as CDC20 and APC/C). As different Cullins determine the specific compositions of distinct Cullin RING E3 ligase complexes (21,23), we further classified adaptors in the Cullin RING group into five subgroups, including SCF (Cullin 1 and 7), ECS (Cullin 2 and 5), BCR (Cullin 3), DCX (Cullin 4A/B) and Other (with one family of Other).

Proteome-wide identification

As the number of proteins is limited for several families, we totally constructed 1, 1, 15 and 6 HMM profiles for E1, E2, E3 and DUB families, respectively. The functional domain sequences for each family were firstly aligned by MUSCLE (http://www.drive5.com/muscle/, version 3.8.31), a widely used program for multiple sequence alignment (28). Then we used the hmmbuild program in the HMMER 3.0 package (http://hmmer.janelia.org/) (24) to construct HMM models. Moreover, the hmmsearch program of HMMER 3.0 (24) was used to search all protein sequences in 70 eukaryotic species. The default parameters were chosen for the three tools, and additional filters were used to improve the accuracy (presented in the 'Discussion' section). Since one gene can generate multiple variant transcripts, the Ensembl Gene ID was adopted as the unique accession to avoid any redundancy. For one gene, only the protein with the lowest E-value was reserved. All HMM profiles are available at http://uucd.biocuckoo.org/faq.php.

Here, we took 1282 identified E1s, E2s, E3s and DUBs as the benchmark dataset to evaluate the prediction performance and robustness of the HMM identifications. For each family, the annotated sequences were taken as positive data (P), while all other proteins were regarded as negative data (N). The sensitivity (S_n) and specificity (S_p) can be calculated as below:

$$Sn = \frac{TP}{TP+FN} \quad \text{and} \quad Sp = \frac{TN}{TN+FP}$$

Both the self-consistency and leave-one-out validations were carried out. The receiver operating characteristic (ROC) curves were illustrated, and AROC (area under ROC) values were calculated for eight families,

respectively (Supplementary Figure S1). The results suggested that our predictions are accurate and robust (Supplementary Figure S1). To ensure that all curated proteins can be correctly predicted and classified ($S_n = 100\%$), we selected distinct cut-off values for different families (Supplementary Table S2).

For the families without HMM profiles, we conducted the orthology search for 70 eukaryotes by using the reciprocal best-hit approach (29), which can efficiently identify ortholog pairs if two proteins in two different proteomes reciprocally find each other as the best hit, by the blastall program in the BLAST package (30).

Totally, we computationally characterized 738 E1s, 2937 E2s, 46 631 E3s and 6647 DUBs from 70 eukaryotic species, whereas the heat map of the classifications and results for several major groups or families were illustrated by the ggplot2 program (http://had.co.nz/ggplot2/) in the R package (http://www.r-project.org/) (Figure 1). The detailed classifications and data statistics are available in Supplementary Table S3.

USAGE

The UUCD database was constructed in an easy-to-use manner. Here, we used human F-box/WD repeat-containing protein 1A (β-TrCP) as an example to describe the usage of UUCD. The Browse page is the major option for users to look through the UUCD conveniently. Two strategies were implemented for browsing the data, including by species and by classifications (Figure 2). In the option of 'Browse by species', the right tree represents the phylogenetic relations of eukaryotic species in Ensembl, while the left tree represents the Ensembl taxonomy categories, including primates, rodents, laurasiatheria, afrotheria and so on (25) (Figure 2A). By clicking on the 'Homo sapiens' button, the E1, E2, E3 and DUB families in *H. sapiens* can be shown (Figure 2A). Also, UUCD can be browsed by classifications (Figure 2B). The left tree represents the hierarchical categories, whereas known 3D structures of E1s, E2s, E3s or DUBs were taken from the PDB database (31) and present in right (Figure 2B). Since human β-TrCP belongs to the F-box family, users can click on the 'F-box' button to visualize the family information of 70 eukaryotic species (Figure 2B). By either clicking on the 'F-box' button in the page of human families (Figure 2A) or the 'Homo sapiens' button in the page of the F-box family (Figure 2B), the members in human F-box family can be available, while a brief description is shown for biological functions and regulatory roles of F-box containing proteins (Figure 2C). The UUCD ID (UUC-) was adopted for organizing the database, while the Ensembl Gene ID was used as the secondary accession (Figure 2C). The users can click on the 'UUC-HoS-00457' to view the detailed information of human β-TrCP (Figure 2D).

The UUCD can be searched with one or multiple keywords (Figure 3A). For example, if the keyword of 'TRCP' is inputted and submitted, the results will be shown in a tabular page, with the features of UUCD ID
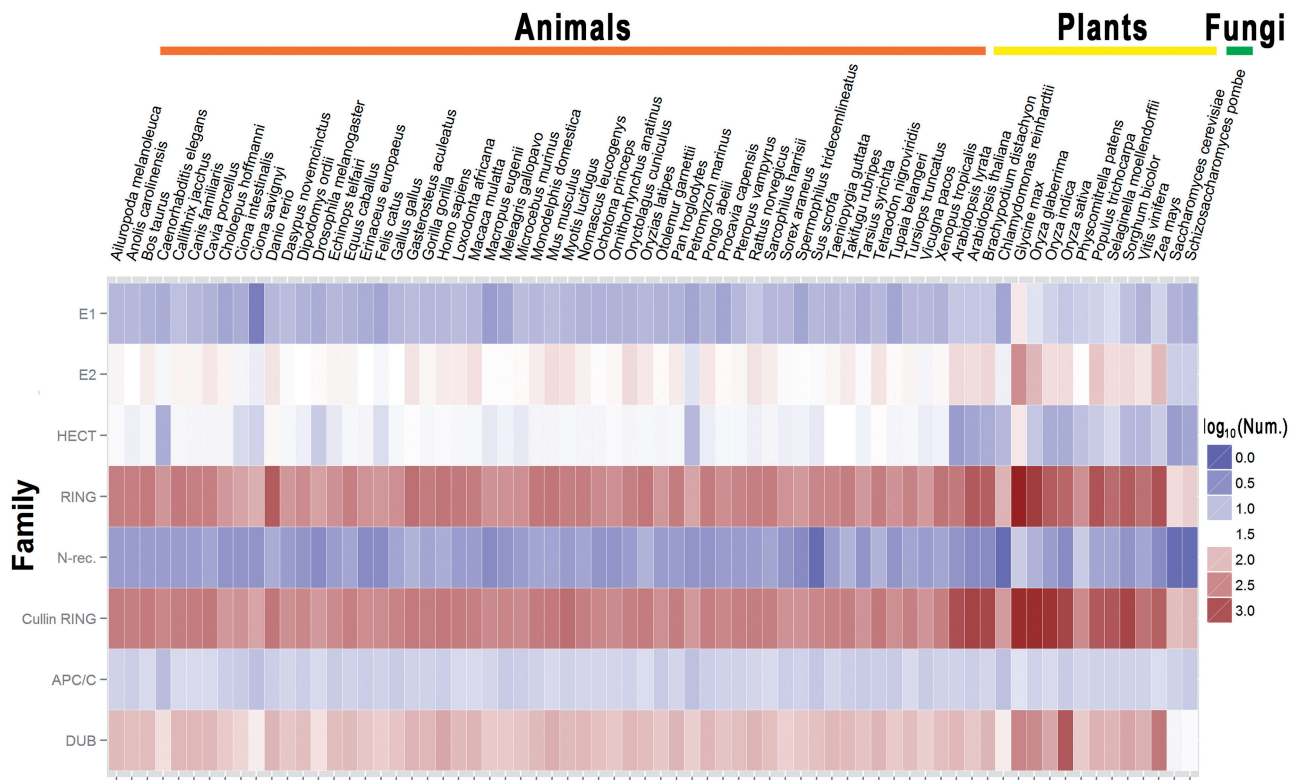
Figure 1. The heat map of the classifications and protein numbers of E1s, E2s, DUBs and several major groups for E3 ligases.

and protein/gene names/aliases (Figure 3A). Furthermore, three additional advance options were provided, such as (i) advance search, (ii) BLAST search and (iii) HMM search. In advance search option, up to three search terms can be specified and submitted to search the precise information (Figure 3B). The second option, BLAST search, was designed for querying the related information in UUCD by protein sequences. The blastall program of NCBI BLAST packages (30) was integrated in the database. Users can input a protein sequence in FASTA format to search identical or homologous proteins (Figure 3C). (iii) In HMM search option, a protein sequence in FASTA format can be inputted and scanned with 1, 1, 15 and 6 HMM profiles of E1, E2, E3 and DUB families, respectively. If the protein is determined as a ubiquitin-associated enzyme, the classification and detection information will be present (Figure 3D).
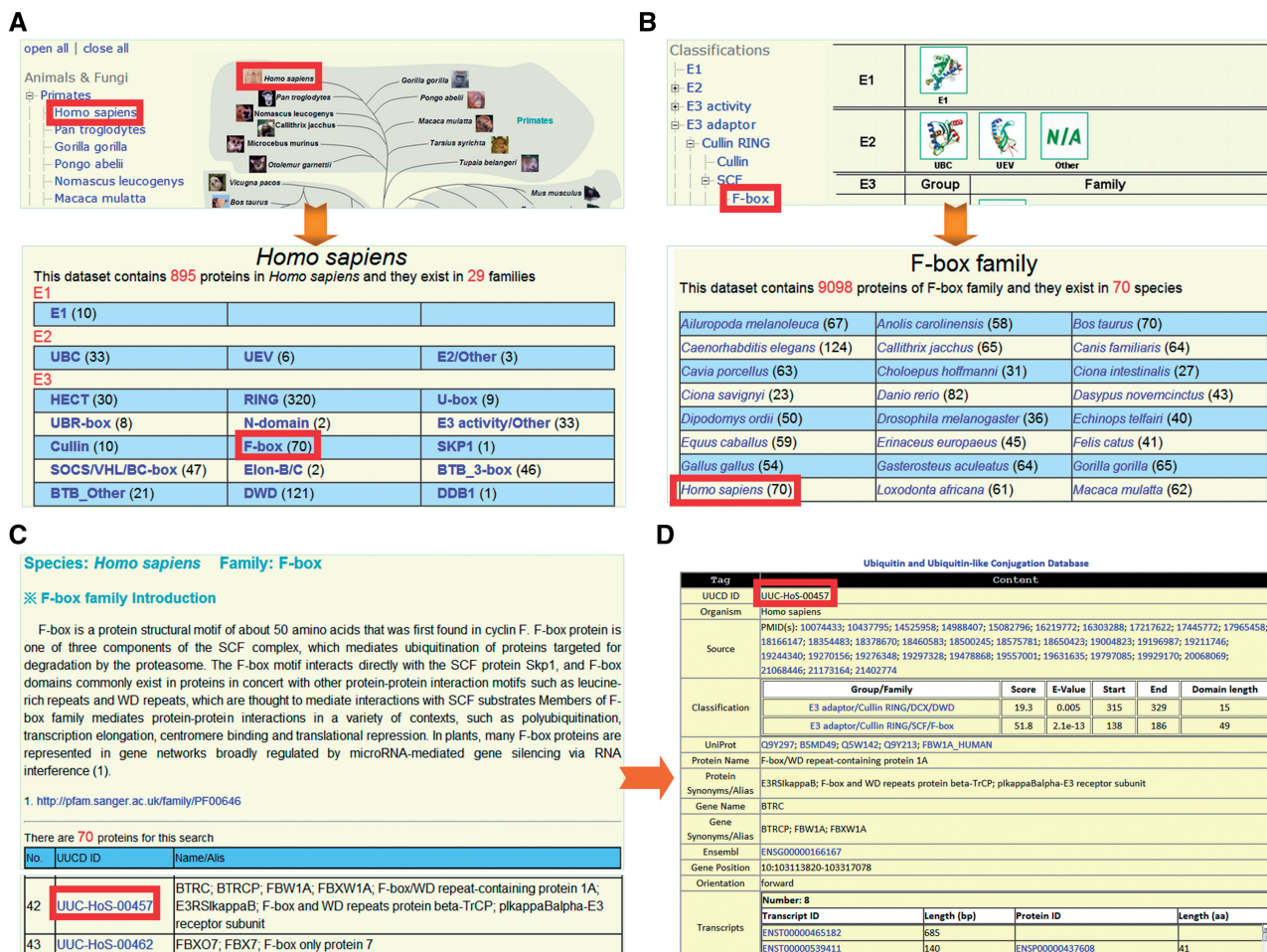
DISCUSSION

As one of the most important post-translational modifications of proteins, reversible ubiquitin and ubiquitin-like conjugation have been implicated in almost all aspects of biological processes and functions, and determines the cellular dynamics and plasticity (1,4,5,7–9). Aberrances of ubiquitin and ubiquitin-like conjugation systems are highly implicated in a variety of diseases and cancers (7,10). In this regard, identification and classification of E1s, E2s, E3s and DUBs is fundamental for dissecting the molecular

mechanisms and regulatory roles of the conjugation (4,8,9), analyzing the phylogenetic relations of enzymes (8,23), modeling E3-substrate networks (32) and providing potent candidates for drug design (11). Developing a comprehensive database with detailed annotation and classification information has emerged to be an urgent challenge.

Previously, a number of computational efforts were taken for systematically characterizing proteins in the ubiquitin-proteasome system (UPS) (8,13–17). For example, Semple *et al.* (13) performed a genome-wide study by identifying potential E1s, E2s, E3s and DUBs from four animals and two fungi. From the scientific literature, SCUD (15) and hUbiquitome (17) databases collected experimentally verified enzymes for *S. cerevisiae* and *H. sapiens*, respectively. The proteins in the two databases were fully covered by our benchmark dataset (Supplementary Table S1). To the best of our knowledge, currently the most comprehensive database was plantsUPS, which contains predicted E1s, E2s and E3s for up to seven plants, whereas the experimental information was not present (16).

By collecting known enzymes of ubiquitin and ubiquitin-like conjugation systems, we classified them based on distinguishable functional domains (Supplementary Table S3). Although several general databases such as Pfam (27), InterPro (33) and SMART (34) contain most of these domains, we re-constructed HMM profiles with our benchmark dataset to promise the prediction performance, and additional filters were used for accurate classification (Supplementary Figure S1). For example, the HMM profiles of UBC and UEV domains are highly similar and

A

open all | close all

Animals & Fungi
⊟ Primates
  — Homo sapiens
  — Pan troglodytes
  — Gorilla gorilla
  — Pongo abelii
  — Nomascus leucogenys
  — Macaca mulatta

Homo sapiens ... Gorilla gorilla / Pongo abelii / Pan troglodytes / Nomascus leucogenys / Callithrix jacchus / Macaca mulatta / Microcebus murinus / Tarsius syrichta / Otolemur garnettii / Tupaia belangeri / Vicugna pacos / Bos taurus / Mus musculus — Primates

Homo sapiens
This dataset contains **895** proteins in *Homo sapiens* and they exist in **29** families

E1

E1 (10)		

E2

UBC (33)	UEV (6)	E2/Other (3)

E3

HECT (30)	RING (320)	U-box (9)
UBR-box (8)	N-domain (2)	E3 activity/Other (33)
Cullin (10)	F-box (70)	SKP1 (1)
SOCS/VHL/BC-box (47)	Elon-B/C (2)	BTB_3-box (46)
BTB_Other (21)	DWD (121)	DDB1 (1)

B

Classifications
— E1
— E2
⊞ E3 activity
⊟ E3 adaptor
  ⊟ Cullin RING
    — Cullin
    ⊟ SCF
      — F-box

E1	E1	
E2	UBC	UEV	Other / N/A
E3	Group	Family

F-box family
This dataset contains **9098** proteins of F-box family and they exist in **70** species

Ailuropoda melanoleuca (67)	*Anolis carolinensis* (58)	*Bos taurus* (70)
Caenorhabditis elegans (124)	*Callithrix jacchus* (65)	*Canis familiaris* (64)
Cavia porcellus (63)	*Choloepus hoffmanni* (31)	*Ciona intestinalis* (27)
Ciona savignyi (23)	*Danio rerio* (82)	*Dasypus novemcinctus* (43)
Dipodomys ordii (50)	*Drosophila melanogaster* (36)	*Echinops telfairi* (40)
Equus caballus (59)	*Erinaceus europaeus* (45)	*Felis catus* (41)
Gallus gallus (54)	*Gasterosteus aculeatus* (64)	*Gorilla gorilla* (65)
Homo sapiens (70)	*Loxodonta africana* (61)	*Macaca mulatta* (62)

C

Species: *Homo sapiens*   Family: F-box

※ F-box family Introduction

F-box is a protein structural motif of about 50 amino acids that was first found in cyclin F. F-box protein is one of three components of the SCF complex, which mediates ubiquitination of proteins targeted for degradation by the proteasome. The F-box motif interacts directly with the SCF protein Skp1, and F-box domains commonly exist in proteins in concert with other protein-protein interaction motifs such as leucine-rich repeats and WD repeats, which are thought to mediate interactions with SCF substrates Members of F-box family mediates protein-protein interactions in a variety of contexts, such as polyubiquitination, transcription elongation, centromere binding and translational repression. In plants, many F-box proteins are represented in gene networks broadly regulated by microRNA-mediated gene silencing via RNA interference (1).

1. http://pfam.sanger.ac.uk/family/PF00646

There are **70** proteins for this search

No.	UUCD ID	Name/Alis
42	UUC-HoS-00457	BTRC; BTRCP; FBW1A; FBXW1A; F-box/WD repeat-containing protein 1A; E3RSIkappaB; F-box and WD repeats protein beta-TrCP; plkappaBalpha-E3 receptor subunit
43	UUC-HoS-00462	FBXO7; FBX7; F-box only protein 7

D

Ubiquitin and Ubiquitin-like Conjugation Database

Tag	Content					
UUCD ID	UUC-HoS-00457					
Organism	Homo sapiens					
Source	PMID(s): 10074433; 10437795; 14525958; 14988407; 15082796; 16219772; 16303288; 17217622; 17445772; 17965458; 18166147; 18354483; 18378670; 18460583; 18500245; 18575781; 18650423; 19004823; 19196987; 19211746; 19244340; 19270156; 19276348; 19297328; 19478868; 19557001; 19631635; 19797085; 19929170; 20068069; 21068446; 21173164; 21402774					
Classification		**Group/Family**	**Score**	**E-Value**	**Start**	**End**	**Domain length**
		E3 adaptor/Cullin RING/DCX/DWD	19.3	0.005	315	329	15
		E3 adaptor/Cullin RING/SCF/F-box	51.8	2.1e-13	138	186	49
UniProt	Q9Y297; B5MD49; Q5W142; Q9Y213; FBW1A_HUMAN					
Protein Name	F-box/WD repeat-containing protein 1A					
Protein Synonyms/Alias	E3RSIkappaB; F-box and WD repeats protein beta-TrCP; plkappaBalpha-E3 receptor subunit					
Gene Name	BTRC					
Gene Synonyms/Alias	BTRCP; FBW1A; FBXW1A					
Ensembl	ENSG00000166167					
Gene Position	10:103113820-103317078					
Orientation	forward					
Transcripts	Number: 8					
	Transcript ID		**Length (bp)**	**Protein ID**		**Length (aa)**
	ENST00000465182		685			
	ENST00000539411		140	ENSP00000437608		41

Figure 2. The browse option of UUCD. We provided two strategies for browsing the database: (**A**) by species and (**B**) by classifications. (**C**) For each family, a brief description and associated members of family were present. (**D**) The detailed information of human β-TrCP.

cannot be distinguished by hmmsearch (24). However, the active site cysteine residue locating ∼70–90 aa for ubiquitin coupling does not exist in UEV domains (18) (Supplementary Figure S2). This single rule was adopted for accurately separating UBC and UEV domains (Supplementary Figure S1). Also, the SOCS/VHL/BC-box family is less conserved and cannot be defined with a single domain (35). Thus, proteins containing at least one of SOCS, VHL or BC-box domains were classified into this family. Furthermore, recent analysis revealed a 32 aa '3-box' motif following BTB domain is essential for Cullin binding, although BTB proteins without the motif might interact with Cul3 in alternative mechanisms (36). The proteins predicted with both BTB domain and 3-box motif were classified into the BTB_3-box family (Supplementary Figure S3). In addition, experimental studies suggested that the highly conserved motif $EX_nHXHX_{10}D$ is essential for the DUB activity of the JAMM family (37,38). The predicted JAMM proteins without this motif were discarded.

Together with the ortholog search for families without HMM profiles, we systematically identified 738 E1s, 2937 E2s, 46 631 E3s and 6647 DUBs from 70 eukaryotic organisms, with an average number of 813.6 total enzymes per organism (Supplementary Table S3). Although there are 672.5 enzymes encoded in one animal, the average number of plant enzymes is >2-fold (1441.5) higher in animals (Supplementary Table S3). Also, the numbers of animal or plant enzymes in the same group or family can be greatly different (Figure 1). For example, we totally identified 495 E1s in 54 animals with an average number of 9.2 per species, whereas up to 226 E1s were detected in 14 plants with an average number of 16.1 (Supplementary Table S3). However, there were 1803 SOCS/VHL/BC-box proteins identified in animals (33.4 per organism), while only 19 SOCS/VHL/BC-box proteins were detected in plants (Supplementary Table S3). Taken together, our systematic analysis demonstrated the complexity and diversity of enzymes for ubiquitin and ubiquitin-like conjugation in eukaryotes.

For future plans, at least three additional types of ubiquitin-associated proteins will be collected and systematically identified. First, the polyubiquitin ligase (E4) was demonstrated to be responsible for assembling multiubiquitin chains (39). It is still not known how many E4 ligases are encoded in eukaryotic genomes. As the number of known E4s is quite limited, such information was not included in UUCD. Second, accumulative

Figure 3. The search and advance options. (**A**) The database can be queried with one or multiple keywords. (**B**) Advance search allows users to input up to three terms for the precise search. (**C**) Blast search option was designed for searching database with one protein sequence in FASTA format. (**D**) HMM search option will scan the inputted protein sequence with pre-constructed HMM profiles.

experiments revealed that a large number of ubiquitin-binding domain (UBD)-containing proteins (>150) can non-covalently interact with ubiquitin (40). Thus, UBD proteins can dynamically interact with modified substrates in a ubiquitination-dependent manner and participate in the ubiquitin network (40). Recently, systematic identification of ubiquitinated proteins with modified sites has emerged to be a hot topic (2). Such information will be carefully collected and curated in the near future. Taken together, although more information remains to be integrated, the comprehensive UUCD with E1s, E2s, E3s and DUBs across 70 eukaryotes can serve as a useful resource for further researches. The UUCD database will be continuously updated, when the proteome sequences of more species are available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1–3.

FUNDING

REFERENCES

1. Hershko,A. (2005) Nobel lecture: the ubiquitin system for protein degradation and some of its roles in the control of the cell-division cycle. *Angew. Chem. Int. Ed. Engl.*, **44**, 5932–5943.
2. Kim,W., Bennett,E.J., Huttlin,E.L., Guo,A., Li,J., Possemato,A., Sowa,M.E., Rad,R., Rush,J., Comb,M.J. *et al.* (2011) Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell*, **44**, 325–340.
3. Ciechanover,A., Heller,H., Elias,S., Haas,A.L. and Hershko,A. (1980) ATP-dependent conjugation of reticulocyte proteins with the polypeptide required for protein degradation. *Proc. Natl Acad. Sci. USA*, **77**, 1365–1368.
4. Ciechanover,A. (1994) The ubiquitin-proteasome proteolytic pathway. *Cell*, **79**, 13–21.
5. Cadwell,K. and Coscoy,L. (2005) Ubiquitination on nonlysine residues by a viral E3 ubiquitin ligase. *Science*, **309**, 127–130.
6. Hicke,L. (2001) Protein regulation by monoubiquitin. *Nat. Rev. Mol. Cell. Biol.*, **2**, 195–201.
7. Chen,Z.J. and Sun,L.J. (2009) Nonproteolytic functions of ubiquitin in cell signaling. *Mol. Cell*, **33**, 275–286.
8. Nijman,S.M., Luna-Vargas,M.P., Velds,A., Brummelkamp,T.R., Dirac,A.M., Sixma,T.K. and Bernards,R. (2005) A genomic and functional inventory of deubiquitinating enzymes. *Cell*, **123**, 773–786.

9. Schulman,B.A. and Harper,J.W. (2009) Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. *Nat. Rev. Mol. Cell Biol.*, **10**, 319–331.
10. Hoeller,D., Hecker,C.M. and Dikic,I. (2006) Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. *Nat. Rev. Cancer*, **6**, 776–788.
11. Bedford,L., Lowe,J., Dick,L.R., Mayer,R.J. and Brownell,J.E. (2011) Ubiquitin-like protein conjugation and the ubiquitin-proteasome system as drug targets. *Nat. Rev. Drug Discov.*, **10**, 29–46.
12. Ciechanover,A., Heller,H., Katz-Etzion,R. and Hershko,A. (1981) Activation of the heat-stable polypeptide of the ATP-dependent proteolytic system. *Proc. Natl Acad. Sci. USA*, **78**, 761–765.
13. Semple,C.A. (2003) The comparative proteomics of ubiquitination in mouse. *Genome Res.*, **13**, 1389–1394.
14. Stone,S.L., Hauksdottir,H., Troy,A., Herschleb,J., Kraft,E. and Callis,J. (2005) Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. *Plant Physiol.*, **137**, 13–30.
15. Lee,W.C., Lee,M., Jung,J.W., Kim,K.P. and Kim,D. (2008) SCUD: Saccharomyces cerevisiae ubiquitination database. *BMC Genomics*, **9**, 440.
16. Du,Z., Zhou,X., Li,L. and Su,Z. (2009) plantsUPS: a database of plants' Ubiquitin Proteasome System. *BMC Genomics*, **10**, 227.
17. Du,Y., Xu,N., Lu,M. and Li,T. (2011) hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. *Database*, Nov 30 (doi: 10.1093/database/bar055; epub ahead of print).
18. van Wijk,S.J. and Timmers,H.T. (2010) The family of ubiquitin-conjugating enzymes (E2s): deciding between life and death of proteins. *FASEB J.*, **24**, 981–993.
19. Sriram,S.M., Kim,B.Y. and Kwon,Y.T. (2011) The N-end rule pathway: emerging functions and molecular principles of substrate recognition. *Nat. Rev. Mol. Cell Biol.*, **12**, 735–747.
20. Deshaies,R.J. and Joazeiro,C.A. (2009) RING domain E3 ubiquitin ligases. *Annu. Rev. Biochem.*, **78**, 399–434.
21. Petroski,M.D. and Deshaies,R.J. (2005) Function and regulation of cullin-RING ubiquitin ligases. *Nat. Rev. Mol. Cell Biol.*, **6**, 9–20.
22. Patterson,C. (2002) A new gun in town: the U box is a ubiquitin ligase domain. *Sci. STKE*, **2002**, pe4.
23. Sarikas,A., Hartmann,T. and Pan,Z.Q. (2011) The cullin protein family. *Genome Biol.*, **12**, 220.
24. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
25. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
26. UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
27. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
28. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
29. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
30. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.
31. Velankar,S., Alhroub,Y., Best,C., Caboche,S., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Golovin,A., Gore,S.P. *et al.* (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **40**, D445–D452.
32. 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.
33. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
34. Letunic,I., Doerks,T. and Bork,P. (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.*, **40**, D302–D305.
35. Kamura,T., Maenaka,K., Kotoshiba,S., Matsumoto,M., Kohda,D., Conaway,R.C., Conaway,J.W. and Nakayama,K.I. (2004) VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases. *Genes Dev.*, **18**, 3055–3065.
36. Zhuang,M., Calabrese,M.F., Liu,J., Waddell,M.B., Nourse,A., Hammel,M., Miller,D.J., Walden,H., Duda,D.M., Seyedin,S.N. *et al.* (2009) Structures of SPOP-substrate complexes: insights into molecular architectures of BTB-Cul3 ubiquitin ligases. *Mol. Cell*, **36**, 39–50.
37. Cope,G.A., Suh,G.S., Aravind,L., Schwarz,S.E., Zipursky,S.L., Koonin,E.V. and Deshaies,R.J. (2002) Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science*, **298**, 608–611.
38. Verma,R., Aravind,L., Oania,R., McDonald,W.H., Yates,J.R. III, Koonin,E.V. and Deshaies,R.J. (2002) Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome. *Science*, **298**, 611–615.
39. Hoppe,T. (2005) Multiubiquitylation by E4 enzymes: 'one size' doesn't fit all. *Trends Biochem. Sci.*, **30**, 183–187.
40. Dikic,I., Wakatsuki,S. and Walters,K.J. (2009) Ubiquitin-binding domains - from structures to functions. *Nat. Rev. Mol. Cell Biol.*, **10**, 659–671.