Supplementary information

1. WD40-repeat domain structures and featured sites

The WD40-repeat domains are abundant in proteomes, especially in eukaryotes. As popular interaction molecules, they act as scaffolds to assemble various complexes that fulfill versatile functions, and thus play indispensable roles in many cellular processes. This type of domain is a β-propeller usually formed by 6-8 repeats, and the sequence of one single repeat contains 40-60 residues with conserved GH and WD dipeptide (Stirnimann, *et al.*, 2010; Wang, *et al.*, 2013; Xu and Min, 2011). Each WD40 repeat folds into a four-stranded antiparallel β-sheet (strand *d*, *a*, *b* and *c*, connected by loops), and is often stabilized by a strong side-chain hydrogen bond network (Wang, *et al.*, 2013; Wu, *et al.*, 2010) (**Figure S1A,B**), which is widely and uniquely presented in WD40 proteins. WD40 domain has three faces, *i.e.* top, side and bottom faces, to mediate the interactions with partners. The top face is better studied than others, and the potential hotspot residues on this face are exposed into the solvent by the β bulge between the strand *a* and strand *b* to participate in interactions. Previously, we defined a method to predict the potential hotspots residues on the top face (Wu, *et al.*, 2012). If binding-type residues (Arg, His, Lys, Asp, Glu, Trp, Tyr, Phe, Leu, Ile, Met, Asn, Gln)(Wu, *et al.*, 2012) occur at specific positions (**Figure S1C**, R_1, R_1-2, D-1; **Figure S2**, red asterisk), they will be predicted as potential hotspots at the top face. Based on these, the WDSP tool were thereafter developed to accurately predict the secondary structures of WD40 domains, hydrogen bond networks, and interaction hotspots (Wang, *et al.*, 2013). In this work, we first updated the WDSP tool, and then we applied the improved WDSP in an optimized pipeline for the WDSPdb curation.

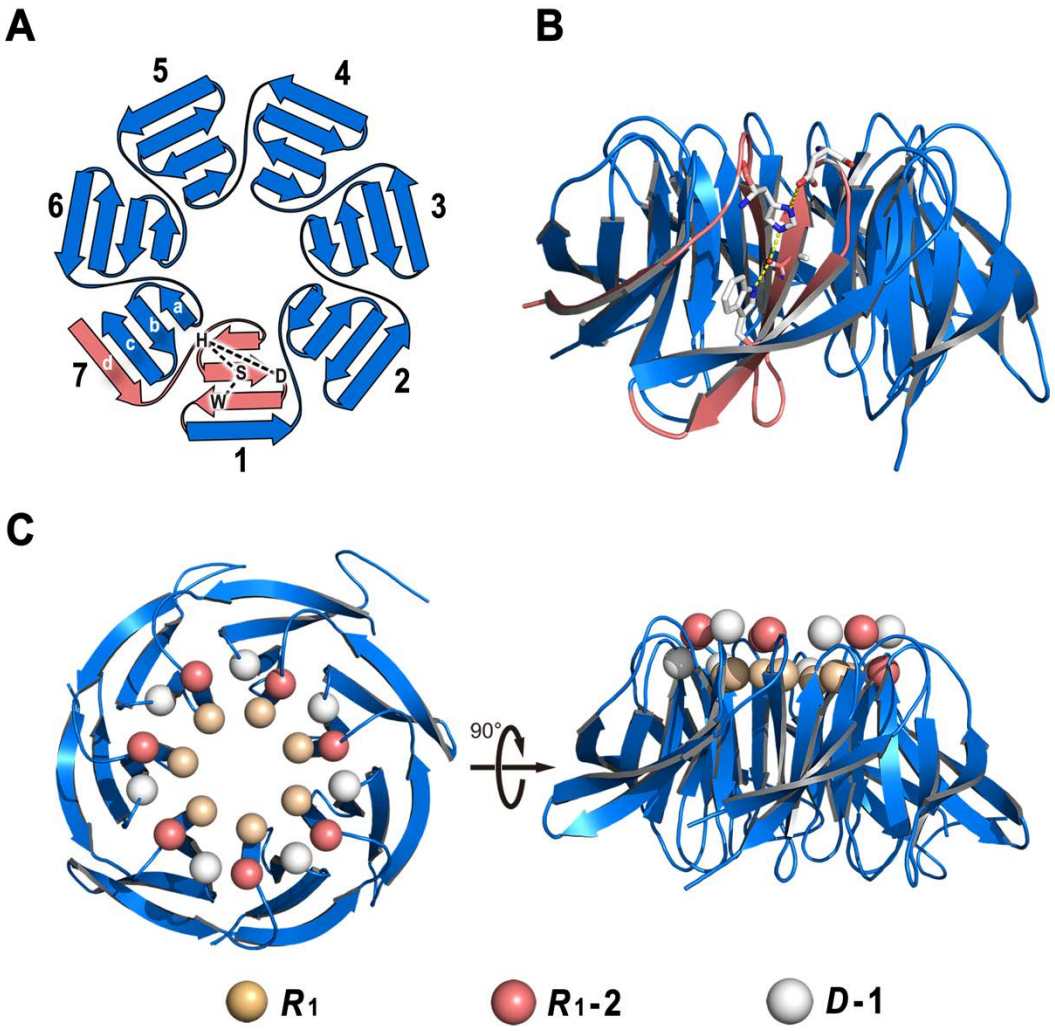

Figure S1 The schematic diagrams of WD40 domain structures and featured sites.

A) The 2D schematic diagram of one typical WD40 domain and repeat as well as the featured hydrogen bond network. The red colored β-strands compose one WD40 repeat. The continuous strand *d*, *a*, *b*, and *c* forms a WD40 repeat, while the continuous strand *a*, *b*, *c* from one repeat, and *d* from the next repeat forms a structural blade. The featured tetrad hydrogen bond network residues are marked.

B) The 3D schematic diagram of one typical WD40 domain. A typical repeat is coloured red, and a featured hydrogen bond network is shown as ball-and-sticks (PDB ID: 3FM0, CIAO1_HUMAN).

C) The 3D schematic diagram of potential interaction hotspots residues on the top face of WD40 domain. Their positions are R_1, R_1-2, D-1, respectively (PDB ID: 3FM0, CIAO1_HUMAN)

2. Update of WDSP tool

2.1 Typical WD40 proteins with experimental structures

In recent years, many experimental structures of WD40 proteins have been determined. The Swiss-Prot section of UniProtKB（release July 5, 2017）contained 97 proteins with typical WD40 structures deposited in PDB (**Table S1**). We performed pairwise global alignment among them, and those with sequence identity greater than 30% formed clusters. To obtain a non-redundant protein set, we kept only one protein within the same cluster. As a result, 65 proteins were retained. This set was used for building the position weight matrix in the update of WDSP tool. It also served as a "reference" set for setting up criteria for confidence category assignment (see section 3.3).

Table S1 WD40 proteins with typical structures*

Number	UniProt ACC	Gene name	PDB ID	WD40 domain start	WD40 domain end	Redundant or not
1	G0SFB5	YTM1	5CXB	100	486	
2	G0SCK6	ERB1	5CXB	436	801	Redundant
3	O00423	EML1	4CI8	258	813	
4	O14011	prp19	3JB9	196	488	
5	O14727	APAF1	3JBT	594	1245	
6	G0SC29	CTHT_0055700	4WJS	134	516	Redundant
7	O88879	Apaf1	3SFZ	595	910	Redundant
8	O36030	SPAC4F10.18	4GQ1	15	386	
9	O43172	PRPF4	3JCR	216	520	
10	P20053	PRP4	5GAN	162	465	Redundant
11	O43660	PLRG1	4YVD	192	490	
12	O13615	prp5	3JB9	152	449	Redundant
13	Q12417	PRP46	5GMK	128	427	Redundant
14	O43818	RRP9	4J0W	135	463	
15	O55029	Copb2	5A1U	5	300	
16	P41811	SEC27	2YNP	4	299	Redundant

17	O60508	CDC40	5MQF	275	579	
18	P40968	CDC40	5GMK	151	455	Redundant
19	O74910	raf1	4O9D	225	617	
20	O75530	EED	5U69	82	438	
21	Q921E6	Eed	2QXV	82	438	Redundant
22	O75717	WDHD1	5GVA	4	301	
23	O76071	CIAO1	3FM0	6	332	
24	Q05583	CIA1	2HES	5	325	Redundant
25	O89053	Coro1a	2AQ5	23	349	
26	P07834	CDC4	3V7D	366	742	
27	P16649	TUP1	1ERJ	329	706	
28	P25382	RSA4	4WJU	131	514	
29	P25635	PWP2	5WYK	5	705	
30	P35184	SQT1	4ZOX	55	428	
31	P38011	ASC1	3FRX	8	314	
32	O18640	Rack1	4V6W	6	312	Redundant
33	O24456	RACK1A	3DM0	6	323	Redundant
34	P38262	SIF2	1R5M	144	534	
35	P39108	PEX7	3W15	1	372	
36	P40362	UTP18	5WYK	234	590	
37	P42935	ELP2	5M2N	6	784	
38	P46680	AIP1	1PGU	5	611	
39	P53011	SEH1	3F3F	1	332	
40	P53196	RPN14	3VL1	15	413	
41	P55735	SEC13	3BG1	3	290	
42	P53024	SEC13	4L9O	1	274	Redundant
43	P61964	WDR5	4ERY	36	331	
44	P61965	Wdr5	2XL2	36	331	Redundant
45	Q498M4	Wdr5	4QQE	36	331	Redundant
46	Q9V3J8	wds	4CY3	62	358	Redundant
47	P62881	Gnb5	2PBI	95	395	
48	P54311	Gnb1	5TDH	46	340	Redundant
49	P62871	GNB1	5KDO	46	340	Redundant
50	P63005	Pafah1b1	1VYH	98	408	
51	P63151	PPP2R2A	3DW8	16	443	
52	P78406	RAE1	3MMY	30	356	
53	Q04660	ERB1	4U7A	421	806	
54	Q04724	TLE1	1GXR	467	767	
55	Q05946	UTP13	5WYK	4	645	
56	Q06506	RRP9	4J0X	142	567	
57	Q09028	RBBP4	4R7A	53	403	

58	P39984	HAT2	4PSW	35	381	Redundant
59	Q16576	RBBP7	3CFS	29	402	Redundant
60	Q24572	Caf1	2XYI	34	407	Redundant
61	Q11176	unc-78	1NR0	51	608	
62	Q12220	DIP2	5WYK	9	685	
63	Q12834	CDC20	4GGC	169	471	
64	P78972	slp1	4AEZ	167	464	Redundant
65	P53197	CDH1	4BH6	246	543	Redundant
66	Q9UM11	FZR1	4UI9	172	471	Redundant
67	Q13216	ERCC8	4A11	34	362	
68	Q16531	DDB1	3EI3	17	1041	
69	Q2YDS1	ddb2	3EI3	103	419	
70	Q3Y8L7	DAW1	5MZH	84	417	
71	Q58CQ2	ARPC1B	1K8K	1	358	
72	P78774	arc1	3DWL	2	376	Redundant
73	Q5IH81	EIF3I	5K0Y	1	315	
74	P40217	TIF34	3ZWL	1	321	Redundant
75	Q86YC2	PALB2	2W18	737	1184	
76	Q8CIE6	Copa	5A1U	2	318	
77	Q96WV5	SPBPJ4664.04	4J87	1	325	Redundant
78	Q8NFH3	NUP43	4I79	2	375	
79	Q8NHY2	COP1	5HQG	400	729	
80	P43254	COP1	5IGO	354	672	Redundant
81	Q8TAF3	WDR48	5K1A	20	325	
82	Q8TBZ3	WDR20	5K19	89	559	
83	Q8TEQ6	GEMIN5	5TEE	3	711	
84	Q969H0	FBXW7	2OVR	365	701	
85	Q96DI7	SNRNP40	5MQF	56	353	
86	O94620	cwf17	3JB9	39	338	Redundant
87	Q96EE3	SEH1L	5A9Q	2	306	
88	Q04491	SEC13	2PM7	1	282	Redundant
89	Q96MX6	WDR92	3I2N	8	350	
90	Q9BQA1	WDR77	4GQB	18	328	
91	Q6NUD0	wdr77	4G56	8	316	Redundant
92	Q9BZK7	TBL1XR1	4LG9	158	511	
93	Q8BHJ5	Tbl1xr1	5NAF	158	511	Redundant
94	Q9GZS3	WDR61	3OW8	6	302	
95	Q9HCU5	PREB	5TF2	4	382	
96	Q9UMS4	PRPF19	4LG8	208	503	
97	Q9Y297	BTRC	1P22	284	581	

*The proteins separated by dashed lines and shaded are sequence clusters, whose

sequence identity is greater than 30%. In each sequence cluster, the protein(s) marked as redundant are removed, and only one protein is retained in the non-redundant "reference" set.

2.2 Update of position weight matrix (PWM) of WD40-repeats

The position weight matrix (PWM, visualized as a sequence logo) is important for accurately and sensitively identifying the WD40 proteins when using the WDSP tool. Previously, we built the PWM using 33 WD40 structures, and here we updated it using this larger non-redundant "reference" set, which containing 65 WD40 structures with 534 repeats (**Table S1**). In brief, the WD40 repeat regions, the β-strands, and loops of WD40 repeats were manually determined by visual inspection of the domain structures. We built the multiple sequence alignment by preferentially and exactly aligning the corresponding β-strands. The loop regions between the β-strands have been manually cut to keep only conserved sites according to our previous experiences (**Figure S2**) (Wang, *et al.*, 2013). By counting the residue counts in each aligned position, the PWM can be built, and the sequence logo can be generated. In the sequence logo, the letter size represents the conservation level of each amino acid at that site.

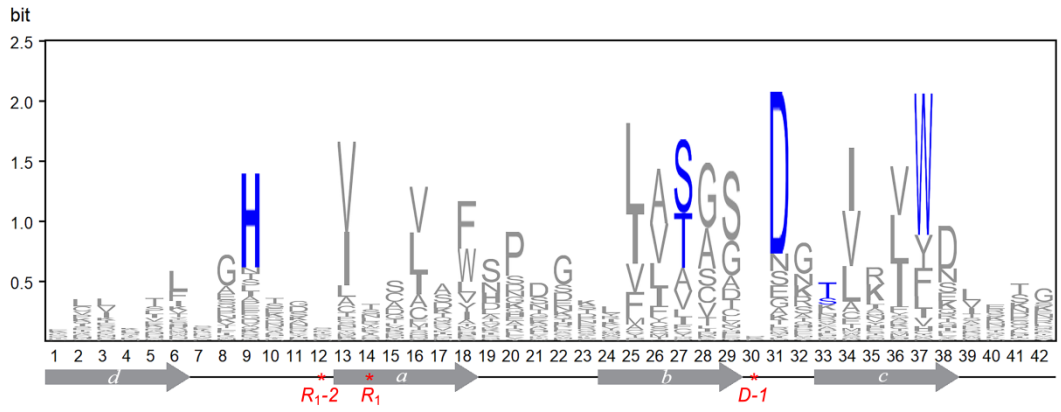

Figure S2 The updated sequence logo of the WD40 repeat.
The total height of the letters represents the conservation of each site. The secondary structures have been shown below the x-axis with arrows indicating the β-strands

(length of strand *d*/*a*/*b*/*c*=6) and lines representing the loops (length of loop *da*=6, *ab*=5, *bc*=3, *cd*=4). The positions highlighted by red asterisks are the potential hotspots positions on the top face involved in interactions. The blue residues may form side-chain hydrogen bond networks.

2.3 Update of the PSIPRED backend

PSI-blast based secondary structure PREDiction (PSIPRED) is one of the best methods for predicting the protein regular secondary structures (Buchan, *et al.*, 2013), and it was used as a backend of WDSP. That is, the predicted secondary structures from PSIPRED were used as the starting point of WDSP. Previously, WDSP utilized PSIRPED V3, and Swiss-Prot was used as the homolog-searching database. Here, we replaced the PSIPRED from V3 to V4 (bioinfadmin.cs.ucl.ac.uk/downloads/psipred). In the process of curation of WDSPdb described below, we chose UniRef90 as the homolog-searching database for annotating WD40 proteins from Swiss-Prot section to obtain better results. For annotating WD40 proteins from TrEMBL section we kept using Swiss-Prot as homolog-searching database to save computing resources, since proteins from TrEMBL is of a huge amount.

2.4 The performance of updated WDSP

To evaluate the capability of the updated WDSP in identifying WD40 repeats, we compared its predictions with the repeat regions derived from crystal structures using the aforementioned non-redundant “reference” set of WD40 proteins. To avoid over-fitting, we performed a five-fold cross validation. In brief, we randomly divided these 65 proteins into 5 parts, and used the WD40 repeats from 4 parts to generate the position weight matrix (PWM) in WDSP. Then the WDSP with this PWM was adopted to predict the last part of WD40 proteins. By comparing between predicted secondary structure and actual secondary structure, we can obtain its performance. This process was repeated 5 rounds, and each round we generated different PWM and

compared on different protein parts, so that each repeat was tested exactly for one time. The overall results can be used to calculate the average performance in the cross-validation.

The performance was also evaluated for other methods (old version of WDSP, UniProt REP, SMART, PROSITE, and Pfam) (Bateman, *et al.*, 2017; Finn, *et al.*, 2016; Letunic and Bork, 2018; Sigrist, *et al.*, 2013; Wang, *et al.*, 2013) using this non-redundant set of WD40 proteins as reference. Since they are less conserved than other repeats and loops, the predictions of strand *d* and loop *da* are much harder than other strands and loops. Hence in the comparison between predictions and actual secondary structures, we adopted two criteria, the loose and the strict. As for the loose criterion, we considered the identification correct when the actual repeat from strand *a* to strand *c* could be completely covered by the predicted repeat. As for the strict criteria, we considered that a whole actual repeat (from strand *d* to strand *c*) or a blade (from strand *a* to next strand *d*) should be found in the predicted repeat or blade with only two amino acids' shift at the boundaries tolerated.

We computed the *F*1 score, which offers a balanced measurement considering both recall and precision, to measure the detecting capability of different methods. The formula is defined as

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

$$Recall = \frac{TP}{TP + FN} = \frac{N_c}{N_r}$$

$$Precision = \frac{TP}{TP + FP} = \frac{N_c}{N_d}$$

, where N_c is the number of WD40 repeats identified correctly, N_r is the total number of actual repeats in the non-redundant set of 65 WD40 proteins, N_d is the total number of predicted WD40 repeats.

The results showed that both the old WDSP and updated WDSP perform much better than other methods, either the loose or strict criteria. Moreover, the updated WDSP is even better than the old version especially in predicting the less conservative strand *d* and loop *cd*. Other methods showed their defects in detecting WD40 repeats. (**Figure S**3).

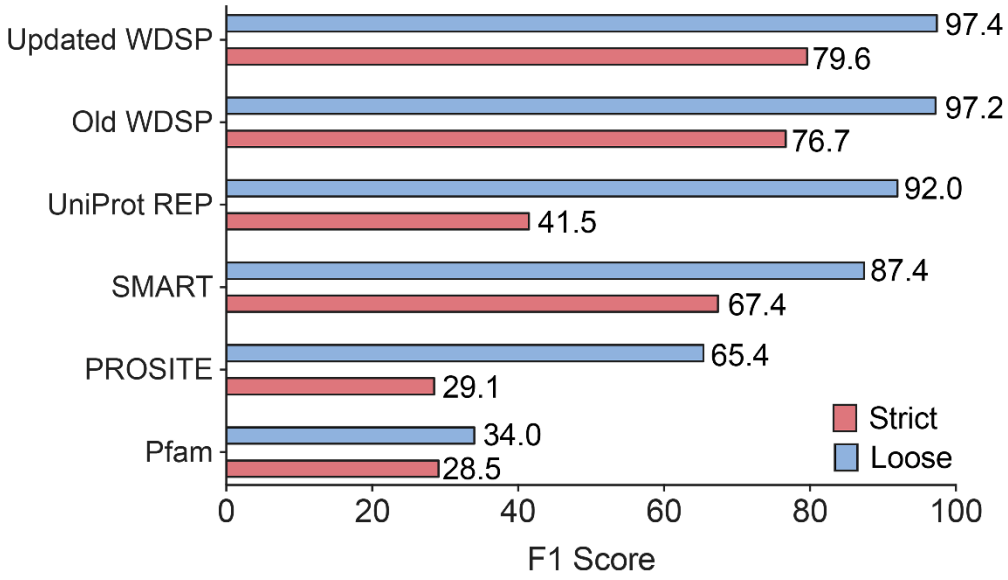

Figure S3 The performance of identifying WD40 repeats by Pfam, PROSITE, SMART, UniProt REP, the old WDSP, and updated WDSP under the loose and strict criteria, as measured by F1 scores.

3. The optimized pipeline of WDSPdb 2.0

3.1 Pre-screening the WD40 candidates

We used the "WD40" as keyword to search the InterPro database (Finn, *et al.*, 2017), and obtained 59 entries which contain 26 families, 24 domains, 5 homologous superfamilies, 3 repeats, and 1 conserved sites. By careful reading their descriptions, we chose those describing WD40 domain, WD40 repeat, or β-propeller. Since InterPro database is integrated from multiple other databases, each entry here may correspond to one or several contributing databases. We retrieved the HMM profiles from the contributing databases. If there is only multiple sequence alignment, but no profiles, we use HMMBuild from the HMMER package to construct the profile(Eddy, 2011). Some InterPro entries that didn't provide either profile or sequence alignment were discarded.

Finally, we collected 54 WD40-related Hidden Markov model-based profiles

(HMM-profiles) from several well-known databases, such as Pfam, SMART, PIRSF, *etc* (**Table S2**). Adopting HMMSearch in the HMMER package, we searched the protein sequences in UniProtKB（release July 5, 2017）using each of these profiles (Eddy, 1998). If a protein got an E-value no more than 10 (default threshold of HMMSearch) by at least one profile, we considered it as a WD40 candidate. All candidates were subjected to run WDSP for more detailed annotation, and all the annotations were stored in WDSPdb 2.0.

Table S2 54 WD40-related HMM profiles

Profile Name	InterPro ID	Database of contributing signature	Contributing signature ID
SM00320	IPR001680	SMART	SM00320
PS50082	IPR001680	PROSITE profiles	PS50082
PF00400	IPR001680	Pfam	PF00400
PF17005	IPR031544	Pfam	PF17005
PS00678	IPR019775	PROSITE profiles	PS00678
PF07676	IPR011659	Pfam	PF07676
PF12894	IPR024977	Pfam	PF12894
53623	IPR036322	SUPERFAMILY	SSF50978
48681	IPR036322	SUPERFAMILY	SSF50978
47077	IPR036322	SUPERFAMILY	SSF50978
48760	IPR036322	SUPERFAMILY	SSF50978
48349	IPR036322	SUPERFAMILY	SSF50978
48421	IPR036322	SUPERFAMILY	SSF50978
47024	IPR036322	SUPERFAMILY	SSF50978
54424	IPR036322	SUPERFAMILY	SSF50978
48759	IPR036322	SUPERFAMILY	SSF50978
47761	IPR036322	SUPERFAMILY	SSF50978
49440	IPR036322	SUPERFAMILY	SSF50978
48420	IPR036322	SUPERFAMILY	SSF50978
46612	IPR036322	SUPERFAMILY	SSF50978
49784	IPR036322	SUPERFAMILY	SSF50978
PF16756	IPR031920	Pfam	PF16756
PF14939	IPR032734	Pfam	PF14939
PF16529	IPR032401	Pfam	PF16529
SM01033	IPR012952	SMART	SM01033
PF08149	IPR012952	Pfam	PF08149
PTHR10856-SF23	IPR027337	PANTHER	PTHR10856:SF23
PF08154	IPR012972	Pfam	PF08154
PTHR10856-SF2	IPR027335	PANTHER	PTHR10856:SF2
PF15492	IPR029145	Pfam	PF15492
PTHR19846	IPR027106	PANTHER	PTHR19846

MF_03022	IPR026962	HAMAP	MF_03022
PTHR11227-SF27	IPR032911	PANTHER	PTHR11227:SF27
PTHR10856-SF10	IPR027333	PANTHER	PTHR10856:SF10
PTHR10856-SF17	IPR027339	PANTHER	PTHR10856:SF17
PTHR16038	IPR037379	PANTHER	PTHR16038
PTHR10856-SF18	IPR029508	PANTHER	PTHR10856:SF18
PTHR10856-SF20	IPR027331	PANTHER	PTHR10856:SF20
PF08596	IPR013905	Pfam	PF08596
PR00320	IPR020472	PRINTS	PR00320
PF15889	IPR031762	Pfam	PF15889
PIRSF006425	IPR014441	PIRSF	PIRSF006425
PTHR11227-SF23	IPR032909	PANTHER	PTHR11227:SF23
PTHR10856-SF24	IPR027340	PANTHER	PTHR10856:SF24
PF11615	IPR021653	Pfam	PF11615
PTHR10253	IPR037352	PANTHER	PTHR10253
SM01166	IPR015048	SMART	SM01166
PF08953	IPR015048	Pfam	PF08953
PTHR10856	IPR015505	PANTHER	PTHR10856
PTHR11024	IPR037363	PANTHER	PTHR11024
PF11635	IPR021665	Pfam	PF11635
PTHR10644-SF3	IPR031297	PANTHER	PTHR10644:SF3
PF11540	IPR025956	Pfam	PF11540
PIRSF037237	IPR017149	PIRSF	PIRSF037237

3.2 Data overview and comparison

After pre-screening by HMMSearch, we retained 594,402 entries from UniProtKB（release July 5, 2017）. We then utilized the updated WDSP to predict the detailed secondary structures of all the sequences except 83 from TrEMBL section, which were too long to be handled. Finally, a total of 594,319 entries and their secondary structures were stored in the WDSPdb 2.0 (5,601 in Swiss-Prot section, 588,718 in TrEMBL section).

The number of WD40 proteins is significantly increased in the WDSPdb 2.0 compared to that in the WDSPdb 1.0. As for the Swiss-Prot section, all of entries in version 1.0 have been included in version 2.0. As for the TrEMBL section, most of the entries in version 1.0 have been recorded in version 2.0 except 4,361 (**Figure S4**). The “missing” 4,361 proteins are caused by the updates between different UniProtKB

versions: Through tracing these entries in different UniProtKB version, we found most of them had been removed in UniProtKB（release July 5, 2017）, a few of them had been moved from TrEMBL section to Swiss-Prot, and a few of them had modified their protein sequences. All of the entries that were moved to Swiss-Prot section are included in WDSPdb 2.0 without any exception.

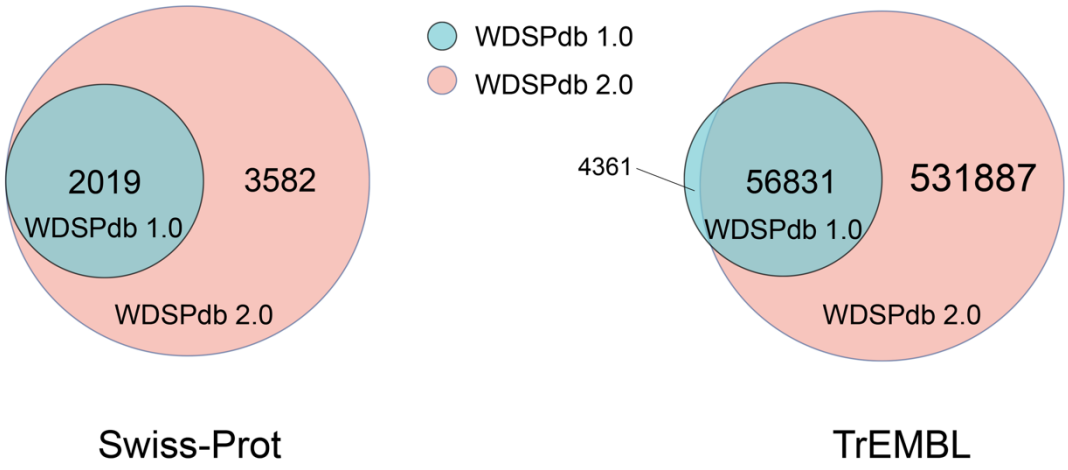

Figure S4 The comparison of the entries between WDSPdb 1.0 and 2.0.

3.3 Confidence category assignment

The WDSPdb 2.0 greatly expanded database capacity. In order to cater for different scientific study demands, we assigned a confidence category to each WD40 candidate according to the criteria described in this section.

In the descriptions in section 2.1 and **Table S1**, we have manually curated a non-redundant “reference” set of 65 typical WD40 proteins with experimental structures. Here, we further collected a set of 98 non-typical WD40 structures. After removing redundancy in a similar way, we finally obtained 80 non-typical WD40 structures (**Table S3**).

Table S3 The information of non-typical WD40 structures

Number	UniProt ACC	Gene name	PDB ID	Redundant or not
1	G0S4F3	NUP82	5CWW	

2	G0S7B6	NUP170	5HAX	
3	O51396	gyrA	1SUU	
4	O67108	gyrA	3NO0	Redundant
5	O74965	SPBC4B4.04	3WJ9	
6	O75326	SEMA7A	3NVQ	
7	O75694	NUP155	5IJN	
8	O95198	KLHL2	4CHB	
9	Q53G59	KLHL12	2VPJ	Redundant
10	Q9UH77	KLHL3	4CH9	Redundant
11	O95714	HERC2	3KCI	
12	P06103	PRT1	4U1F	
13	P0A855	tolB	2HQS	
14	Q8ZGZ1	tolB	4PWZ	Redundant
15	P0C581	par-1	4CZX	
16	P10493	Nid1	1NPE	
17	P11442	Cltc	1BPO	
18	P12293	moxF	1LRW	
19	P38539	NULL	2AD6	Redundant
20	P16027	moxF	1W6S	Redundant
21	P13650	gdhB	1CRU	
22	P14740	Dpp4	4FFV	
23	P22411	DPP4	1ORV	Redundant
24	P27487	DPP4	4A5S	Redundant
25	Q12884	FAP	1Z68	Redundant
26	P14925	Pam	3FVZ	
27	P21062	A39R	3NVX	
28	Q8JL80	EVM139	3NVN	Redundant
29	P21827	SRM1	3OF7	
30	P22219	VPS15	3GRE	
31	P23006	mauB	3C75	
32	P29894	mauB	2BBK	Redundant
33	P25171	Rcc1	3MVD	
34	P26449	BUB3	1YFQ	
35	P27169	PON1	1V04	
36	P27801	PEP3	4UUY	
37	P32523	PRP19	4ZB4	
38	P35729	NUP120	3F7F	
39	P38677	NCU04071	1JOF	
40	P39371	nanM	2UVK	
41	P40064	NUP157	4MHC	
42	P40368	NUP82	3PBP	
43	P40477	NUP159	1XIP	
44	P42658	DPP6	1XFD	

45	Q8N608	DPP10	4WJL	Redundant
46	P48147	PREP	3DDU	
47	P23687	PREP	2XDW	Redundant
48	P49951	CLTC	5M5T	
49	P52697	pgl	1RI6	
50	P53136	NSA1	5SUI	
51	P55884	EIF3B	5K1H	
52	P72181	nirS	1QKS	
53	P24474	nirS	1NIR	Redundant
54	P75804	yliI	2G8S	
55	P76116	yncE	3VGZ	
56	P77774	bamB	3Q7M	
57	Q9HXJ7	bamB	4HDJ	Redundant
58	P84888	aauB	2OIZ	
59	P84908	daip	5FZP	
60	P96086	tri	1K32	
61	P9WI79	pknD	1RWI	
62	Q00610	CLTC	4G55	
63	Q02793	SKI8	1SQ9	
64	Q04693	RSE1	5GM6	
65	Q06339	TFC6	2J04	
66	Q08282	PPM2	2ZWA	
67	Q12038	SRO7	2OAJ	
68	Q12308	TFC8	2J04	
69	Q15393	SF3B3	5IFE	
70	Q15493	RGN	3G4E	
71	Q3SYG4	BBS9	4YD8	
72	Q4W6G0	qgdA	1YIQ	
73	Q46444	qheDH	1KB0	Redundant
74	Q8GR64	qbdA	1KV9	Redundant
75	Q51705	nosZ	1FWX	
76	Q6CN23	HSV2	4V16	
77	Q70DK5	xghA	2CN3	
78	Q72U69	orfC	3BWS	
79	Q7MUW6	ptpA	2Z3Z	
80	Q7SIG4	NULL	1PJX	
81	Q8CJF7	Ahctf1	4I0O	
82	Q8DI95	tll1695	2XBG	
83	Q8FA95	yjiK	3QQZ	
84	Q8J0D2	NULL	1SQJ	
85	Q8MQJ9	brat	1Q7F	
86	Q8N122	RPTOR	5H64	
87	Q8UEU8	Atu1656	2OJH	
88	Q8WUM0	NUP133	1XKS	

89	Q99523	SORT1	3F6K	
90	Q9BVC4	MLST8	4JSN	
91	Q9ER30	Klhl41	2WOZ	
92	Q9FN03	UVR8	4NBM	
93	Q9UJX5	ANAPC4	5BPW	
94	Q9Y4B6	DCAF1	4PXW	
95	Q9YBQ2	APE_1547.1	3O4H	
96	Q9Z2X8	Keap1	3WN7	
97	Q14145	KEAP1	1ZGK	Redundant
98	Q9Z4J7	exaA	1FLG	

Next, we submitted them to the updated WDSP to obtain their WD40 repeats predictions and repeats scores. For each protein, we can obtain an average repeat score. It is shown that there is significantly difference in the average repeat scores between these two sets (p-value <0.0001, Mann Whitney test) (**Figure S5)**. When selecting an average repeat score as cut-off, the repeat whose average score is greater than the cut-off would be considered as typical WD40 proteins. By analyzing the repeat score distributions, we manually defined two cut-offs, 44 and 62. The average repeat score 62 is the best discriminable cut-off, *i.e.*, the junction point of distribution curves of typical and non-typical WD40 proteins' average repeat score (**Figure S5)**. The cut-off 44 was the lowest average repeat score in typical WD40 proteins. It was chosen as a second cut-off because the requirement of the average repeat scores greater than 44 can successfully identify all the typical WD40 proteins.

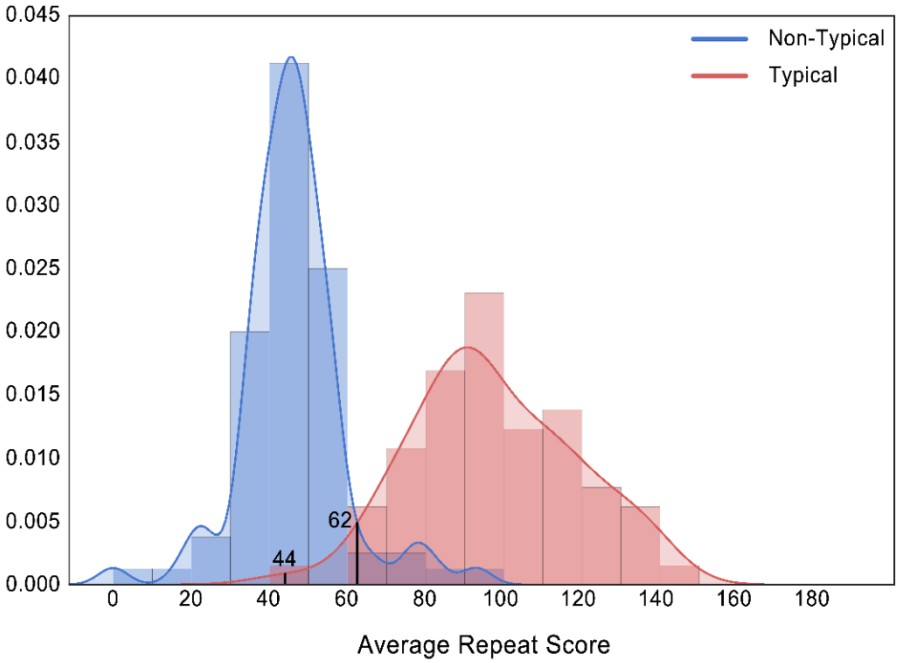

Figure S5 The average WD40 repeat scores distributions of typical and non-typical WD40 proteins.

Besides average repeat score, the unique hydrogen bond network is also an important signature of WD40 proteins. Furthermore, 6 or more WD40 repeats would be more easily to form a closed β-propeller structure. Taking all these features into account, we proposed the rules (**Figure S6**) of assigning four categories for all WD40 candidates:

1) "High":
    a) with PDB structure AND manually reviewed (*i.e.*, in "reference" set); OR
    b) the average repeat score >=62 AND repeats count >=6;
2) "Middle":
    a) the average repeat score >=44 AND <62 AND contains at least one hydrogen bond network AND repeat count >=6;
    b) the average repeat score >=62 AND repeat count <6;
3) "Low":
    a) the average repeat score<44; OR
    b) the average repeat score >=44 AND <62 AND (contains no hydrogen bond

network OR repeat count <6);

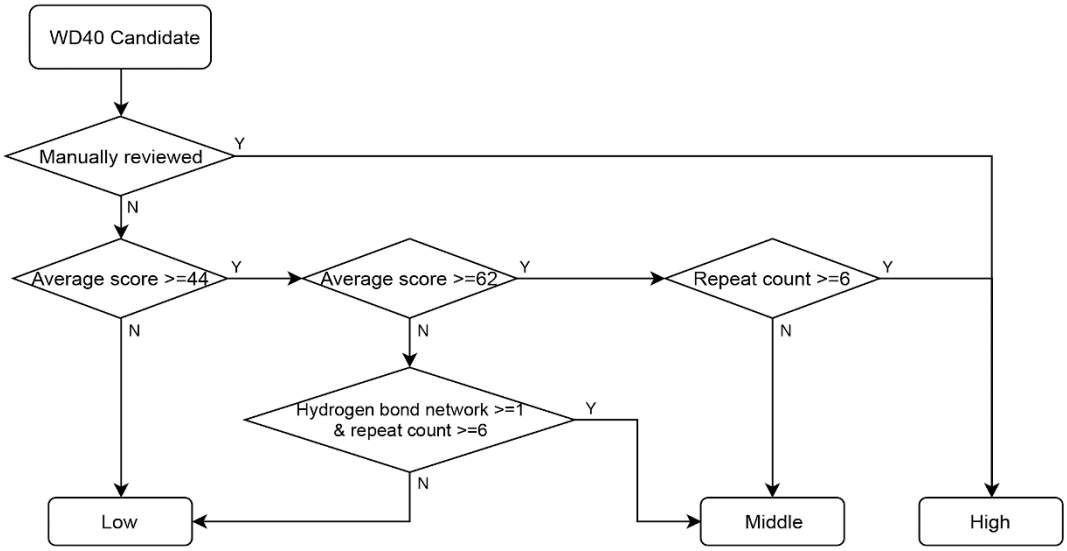

Figure S6 The flow chart of assigning the confidence category to WD40 candidates.

3.4 3D structure modeling

WD40 proteins prefer forming a domain with 6, 7 or 8 repeats, but may encounter more complicated situations when the repeat number is more than 8 or less than 6. Hence, in order to ensure the quality and accuracy, we only modeled proteins in which repeat number is 6, 7, or 8 using MODELLER V9.20 software (Webb and Sali, 2016). Here, 97 typical WD40 structures served as a template pool in 3D homology modeling. Even if a protein has experimental structure of WD40 domain, we also run modeling by using itself as template to fix chain-breaking problems and so on. For those that don't have experimental structures, we selected a template from the template pool by a customized sequence alignment that just considers β strands sequence in WD40 domain. That is, we deleted all of loops between adjacent β strands, connected all of the remained β strands, pairwise aligned target sequences with the templates sequences and selected the PDB structures with highest similarity as template (**Figure S7A**). Since loops of WD40 proteins are flexible and less conservative, such operations could help select the best template while avoiding the interference brought by less conserved loops. After selecting the best template, we built a customized pairwise alignment

between the target and the template. That is, the β-strands were manually aligned since they were confined to the fixed lengths, and the corresponding loop regions were aligned by using global alignment algorithm (**Figure S7B**). The alignment and template PDB structure were submitted to MODELLER for 3D structure modeling.

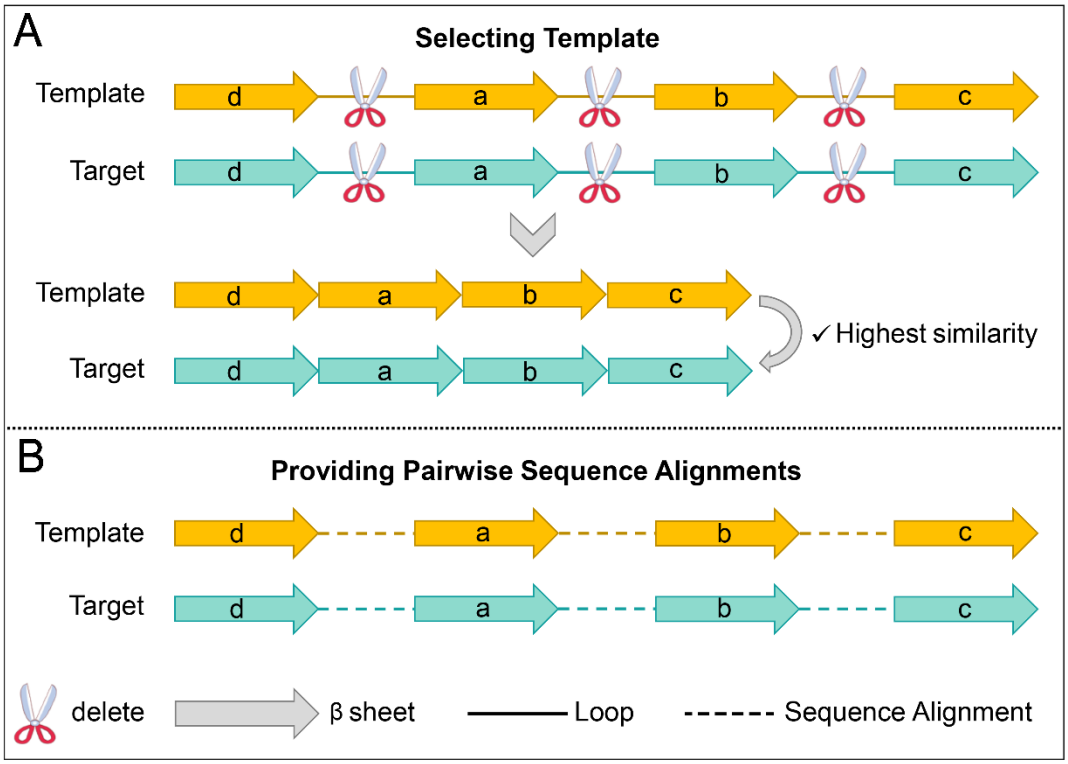

Figure S7 The schematic diagram of selecting template and building pairwise sequence alignments for structure modeling.

3.5 WD40 variants mapping

The single amino acid variants are a good bridge to investigate the relationship between genotypes and phenotypes. In order to facilitate the analysis of WD40 proteins, we collected missense variants from several sources. These sources specialize in collecting different genomic alterations, such as germ-line disease-related alterations (ClinVar (Landrum, *et al.*, 2018), published clinical samples), cancer-related alterations (Cosmic (Forbes, *et al.*, 2017)), alterations detected by whole genome or exome sequencing (1000 Genomes (1000 Genomes Project Consortium, 2015), ExAC (Lek, *et al.*, 2016)).

Each source contains its own reference sequences to store substitutions. We

mapped these substitutions to UniProtKB proteins through sequence alignment. As for the 1000 Genomes and ExAC, we adopted VEP (McLaren, *et al.*, 2016) tool to translate codon changes to amino acid changes first, then mapped them by sequence alignment.

In a similar way, we also collected cancer driver mutations from IntOGen database (Release May, 2016)(Gonzalez-Perez, *et al.*, 2013), cancer highly recurrent mutations, which are originally annotated by the Cancer Hotspots (Chang, *et al.*, 2016), from cBioPortal (Cerami, *et al.*, 2012), and mutations have been experimentally shown to affect the PPIs from the IntAct (Kerrien, *et al.*, 2012).

All these types of variants were mapped to human WD40 proteins with the confidence category of “High” from the Swiss-Prot section through the method described above. For PPI-influencing variants in IntAct, WD40 proteins from other species were also considered. Finally, we have mapped 37,184 variants to 252 WD40 proteins, and integrated them into WDSPdb 2.0. For these variants, we provide the annotations from the source databases, such as the specific cancer types associated with the cancer driver mutations obtained from IntOGen, the interacted partners which perturbed by the mutations from IntAct. We also present the special annotations including exact secondary structures locations, featured site identifications (hydrogen bond network, potential hotspots on the top face). In addition, we provide the pathogenic predictions from dbNSFP v3.5 (Liu, *et al.*, 2011; Liu, *et al.*, 2016). These comprehensive annotations would help users to conduct detailed and systematic pathogenic molecular mechanism studies.

4. Update of the website interface

We redesigned the website by using the Django web framework (version: 2.0.6) and Bootstrap (version: 3.3.7), a front-end library, to provide cleaner web structure, more organized web pages, and more powerful web functions. The new website brings more convenience to users when browsing or searching the database.

Django is an open-source web framework built by the Python programing

language. It follows the MVC (Model-View-Controller) architectural pattern, which provides security and extensibility while keeping the site structure clean. Database interfaces are encapsulated into the model layer to ensure that users cannot directly operate the database, which prevents the web site from attacks, such as SQL injection. Django has also flexible routing capabilities that allow web developers to design their own URLs. Based on this feature, we designed a REST API to help users to send requests to WDSPdb 2.0 to download the secondary structure annotations predicted by WDSP.

Bootstrap is the world's most popular front-end component library, which provides an abundant and powerful toolkit for developing with HTML, CSS and JS. By importing its library, we not only designed bootstrap-style web pages, but also added more user-friendly and advanced functions to our site. For example, we used the grid system of bootstrap to make the page layout more concise and easier to read. In addition, WDSPdb 2.0 adopted the bootstrap-table, a JavaScript plug-in, which enables user-customized data display and download in multiple formats (.txt, .csv, .json, *etc*).

Moreover, we replaced the structure visualization tool to NGL viewer for faster loading and smoother operation. NGL viewer can fully take advantage of the capabilities of modern web browsers (Rose and Burley, 2018). For example, by using hardware-accelerated graphics through WebGL API, WDSPdb 2.0 can provide more visualization modes (cartoon, balls and sticks, surface, spacefill, et. al.) and coloring schemes (rainbow, by secondary structure, by chain, by b-factor, *etc*.) without reducing performance. We implemented a strategy to view the variants and the featured sites at the same time, which would help the users to understand the impact of the variants in the structural context intuitively. As for the WDSP tool, we added options for tuning two parameters, the homolog-searching data bank and the iterative times, in order to meet different user requirements.

5. References

1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**(7571):68-74.

Bateman, A., *et al.* (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1):D158-D169.

Buchan, D.W., *et al.* (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res*, **41**(Web Server issue):W349-357.

Cerami, E., *et al.* (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data (vol 2, pg 401, 2012). *Cancer Discov*, **2**(10):960-960.

Chang, M.T., *et al.* (2016) Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol*, **34**(2):155-+.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**(9):755-763.

Eddy, S.R. (2011) Accelerated Profile HMM Searches. *Plos Comput Biol*, **7**(10).

Finn, R.D., *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res*, **45**(D1):D190-D199.

Finn, R.D., *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, **44**(D1):D279-D285.

Forbes, S.A., *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, **45**(D1):D777-D783.

Gonzalez-Perez, A., *et al.* (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Meth.*, **10**(11):1081-1082.

Kerrien, S., *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, **40**(D1):D841-D846.

Landrum, M.J., *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, **46**(D1):D1062-D1067.

Lek, M., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans.

Nature, **536**(7616):285-291.

Letunic, I. and Bork, P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research*, **46**(D1):D493-D496.

Liu, X.M., Jian, X.Q. and Boerwinkle, E. (2011) dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. *Human Mutation*, **32**(8):894-899.

Liu, X.M., *et al.* (2016) dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, **37**(3):235-241.

McLaren, W., *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol*, **17**(1):122.

Rose, A.S. and Burley, S.K. (2018) Interactive Exploration of Non-Covalent Interactions with the NGL Viewer. *Biophys J*, **114**(3):343a-343a.

Sigrist, C.J.A., *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Research*, **41**(D1):E344-E347.

Stirnimann, C.U., *et al.* (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci.*, **35**(10):565-574.

Wang, Y., *et al.* (2013) A method for WD40 repeat detection and secondary structure prediction. *PLoS One*, **8**(6):e65705.

Webb, B. and Sali, A. (2016) Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci*, **86**:2 9 1-2 9 37.

Wu, X.H., *et al.* (2010) The effect of Asp-His-Ser/Thr-Trp tetrad on the thermostability of WD40-repeat proteins. *Biochemistry*, **49**(47):10237-10245.

Wu, X.H., *et al.* (2012) Identifying the hotspots on the top faces of WD40-repeat proteins from their primary sequences by beta-bulges and DHSW tetrads. *PLoS One*, **7**(8):e43005.

Xu, C. and Min, J. (2011) Structure and function of WD40 domain proteins. *Protein & cell*, **2**(3):202-214.