

Protein	Residue	Sequence	Residue	Sequence	Residue	Sequence	Residue	Sequence	Residue	Sequence
LEDGF	1	MTRDFKPGDLIFAKMKGYPHWPARVDEVPDGAVKPP-----TNK-LP-FFFGTHE-----TAFLGPKDIFPYS	62							
HDGF2	1	MPHAFKPGDLVFAKMKGYPHWPARVDDIADGAVKPP-----PNK-YPIFFFGTHE-----TAFLGPKDLFPYD	62							
HDGF13	5	RPREYKAGDLVFAKMKGYPHWPARVDELPEGAVKPP-----ANK-YPIFFFGTHE-----TAFLGPKDLFPYK	66							
HDGF	6	RQKEYKCGDLVFAKMKGYPHWPARVDEMPEAAVKST-----ANK-YQVFFFGTHE-----TAFLGPKDLFPYE	67							
HDGF11	1	~MPMYKSGDLVFAKMKGYAHWPARVEHMT-----Q-----PNR-YQVFFFGTHE-----TAFLSPKRLFPYK	55							
MUM1	405	EPRSEVGMVLVWHKHKKYPPFWPAVVKSVRQRDKKA-----S-VLYIEGHMNPMMKGFVSLKSLKHF-	465							
N-PAC	2	AAVSLRLGDLVWGKLGGRYPWPWGKIVNPPKDLKKPR-----GK-KCFVFKFFGTED-----HAWIKVEQLKPYH	64							
DNMT3A	286	DGRGFGIGELVWGKLGGRYSWWPGRIVSWMMTGRSRA-----AEGTRVWVWFGDGKF-----SVVCVGLKMP-LS	348							
DNMT3B	219	DGKEFGIGDLVWGKLGGRYSWWPAMVSVWATSQRQA-----MSGMRVWVWFGDGKF-----SEVSADKLV-LG	281							
MBD5	1379	RPRTFNVGDLVWGQIKGLTSWPKLVREDDVHNSCQ-----S-VLYIEGHMNPMMKGFVSLKSLKHF-	1432							
PWWP2A	649	DGRTICVGDIVWAKLYGFPWWPARTLITITVSRKDNG-----LL--VRQEARISWFGSP-T-----TSFLALSQSPFL	713							
BRPF1	1079	EDSPLDALDLVWAKCRGYPSYPALIIDPKMPREGMFHHGVPIVPPPLEVLKLGEOQTQEAHEHLYLVLFDFDNKRT---	1166							
BRPF3	1070	DRGDLEPLELVWAKCRGYPSYPALIIDPKMPREGLLHNGVPIVPPPLDVLKLGEOQKQAEAGEKLFLVLFDFDNKRT---	1157							
BRPF2	923	AASVLEPLKVVWAKCSGYPSYPALIIDPKMPRVPGHHNGVTIPAPPLDVLKIGEHOQTSKDEKLFLVLFDFDNKRS---	1010							
NSD1_C	1750	AGKKPHYREIVWVKVGRYRWWPAEICHPRVAPSNID-----KMRHDVGEFFVLFVFGSN-D-----YLVWTHQARVFPYM	1816							
NSD2_C	874	AGKKLHFQDIIVWVKLGNRYRWWPAEIVCHPKNVPPNIQ-----KMKHEIGEFPVFFVFGSK-D-----YYWTHQARVFPYM	940							
NSD3_C	954	AGKKLHYKQIVWVKLGNRYRWWPAEICNPRSVPLNIQ-----GLKHDLDGDFVFFVFGSH-D-----YYWVHQGRVFPYV	1020							
NSD1_N	317	TPLKYEVGDLVWAKKRRPWPWCRTICSDPLINTHSK-----MKVSNRRPYRQYYVEAFGDPSE-----RAWVAGKAIVMFE	387							
NSD2_N	216	HLLKYNVGDVWVKVSGYPPWPCMVSSADPLLHSYTK-----LKGQ-KKSARQYHVQFFGDAPE-----RAWIFBKSLVAFE	285							
NSD3_N	264	TGKVFQGDVWVKVGYTPWPCMVSSDPQLVHTK-----IN--TRGAREYHVQFFSNQPE-----RAWVHKKRVREYK	331							
MSH6	86	TSCDFSPGDLVWAKMEGYPPWPCLVYNHPFDGTFIR-----EKGKSVR-VHVQFFDDSPSPT-----RGWVSKRLLKPYT	152							
ZMYND8	266	CEPCSNPHPLVWAKLKGFPFWPAKALRDKDGQV-----DARFFGQH-D-----RAWVPIINCYLM-	319							
ZMYND11	180	CYPCIPNHELWAKMKGFGFWPAKVMQKEDNQV-----DVRFFGHHH-----RAWIPSENIQDI-	234							
ZCWPW1	311	AYASYIPGSIIVAKQYGYPPWPGMTIESDPLGEYFL-----FTSHLDSLPSKYHVTFEGETVS-----RAWIPVNMLKNFQ	381							
ZCWPW2	92	VYSQLPLGSLVVLVLIQNWPSWPGILCPDRFKGKYVT-----YDP--DGNVEEYHHEFLGDPHS-----RSWIKATFVGHYS	160							

Protein	Residue	Sequence	Residue	Sequence	Residue	Sequence	Residue	Sequence	Residue	Sequence
LEDGF	63	ENKEKYGKPN-----KRRG-----FNEGWEIENNPKVKFS-----	93							
HDGF2	63	KCKDKYGKPN-----KRRG-----FNEGWEIQNNPHASYS-----	93							
HDGF13	67	EYKDKFGKSN-----KRRG-----FNEGWEIENNPGVKFT-----	97							
HDGF	68	ESKEKFGKPN-----KRRG-----FSEGLWEIENNPTVKAS-----	98							
HDGF11	56	ECKEKFVKPN-----KRRG-----FSAGLWEIENNPTVQAS-----	86							
MUM1	466	DCKEKQTLNLQ-----AREE-----FNQDIGWCVSLITDYVRLVGLCGSFAAGSFLYYAADISYPVRKSIQQD-----	527							
N-PAC	65	AHKEEMIKIN-----KRRG-----FQQAVDAVEEFLRRRAKGDQTSSSHNSDDKNR-----	110							
DNMT3A	349	SFCSAFHQATYN-----KQPMYRKAI-----Y-EVLQVASSRAGKLPVCHDSDESDTAKAVEVQNKPMIEWALGGFQPSGPKGLEPPE	426							
DNMT3B	282	LFSQHFNLATFN-----KLVSYRKAM-----Y-HALEKARVRAGKTFP-----SSPGDS---LEDQLKPMLEWAHGGFKPTGIEGLKPNN	352							
MBD5	1433	GLEAYSrvRKRN-----RKSGLNNHLEAAIHEAMSELDKMSGTVHQIP-----	1476							
PWWP2A	714	ENFQSRFNK-----KRRGL-----YRKATTEAATAAKQLTPEVRALLTQFET-----	755							
BRPF1	1167	VNQDLDEKMLE-----GRKSNIRKSVQIAYHRAIQHRSKVQGEQS-----	1207							
BRPF3	1158	VEDTVDKLMLE-----GRKTSIRKSVQVAYDRAMIHLSRVRGPHS-----	1198							
BRPF2	1011	IDETIDKLMME-----GRNSSIRKAVRIAEDRAMNHLSRVHGEP-----	1051							
NSD1_C	1817	EGDVSSKDKMGK-----GVDGT-----YKKALQEAARFEELKA-----	1850							
NSD2_C	941	EGDRGSRYQGVR-----GIGRV-----FKNALQEAARFREIKL-----	974							
NSD3_C	1021	EGDKSF-AEGQT-----SINKT-----FKKALEEAARFQELKA-----	1053							
NSD1_N	388	GRHQFEELPVLRRRG-----KQKEKG-----YRHKVPPKILSKWEASVGLAEQ-----	430							
NSD2_N	286	GEGQFEKLCQESAKQAPTKAETK-----LLKPIISGKLRAQWEMGIVQAEAAASMSVEERKAKFTF-----	346							
NSD3_N	332	GHKQYEELLAETKQASNHSEKQK-----IRKPRQRERAQWDIGIAHAEK-----	377							
MSH6	153	GSKSKEAQKG-----GHFYSAKPE-----ILRAVQRADEALNKDKIKRLELAVCDEPSE-----	201							
ZMYND8	320	-SKEIPFSVKK-----TKSI-----FNSAMQEMEYVENIRKRFVFNYS-----	358							
ZMYND11	235	-TVNIHRLHVK-----RSMG-----WKKACDELELHQRFLRE--GRFWKS-----	271							
ZCWPW1	382	ELSLELSVM-----KRRRN-----DCSQKLGVALMMAQEAQIISIQERVNLFQFWSRFNGSNSN-----	435							
ZCWPW2	161	ITLKEPKCK-----NKKK-----WYKSALQEAACLLYGSHEQRLEMCLSKLQDKSETHDKVA-----	213							

Supplementary Note: Structure-guided sequence alignment of all known PWWP domains in human proteins. There are over 20 PWWP domain-containing proteins encoded in the human genome that can be categorized into 6 classes. The sequence alignment was first generated by Clustal Omega and then manually adjusted based on known structures to match the putative DNA interacting basic residues in loop regions at similar positions. Putative DNA interacting basic residues in the patches 1 and 2 are highlighted in green. Residues forming the aromatic cage for methyllysine binding are marked with yellow circles. Other conserved residues (mainly in the conserved hydrophobic core) are colored in grey scale based on the score of conservation.