

Examination of TPD52 Family with respect to its Evolutionary History and Domains

Introduction

A protein family is described as a group of proteins that share a common ancestor and evolutionary similar sequences and, therefore similar functional features. The ancestor at the root of hierarchy, is the common ancestor and as going down at the evolutionary tree, subfamilies of proteins are more closely related. Domains are the smallest functional and structural, independently folded units of the proteins. Domains may have evolved from events such as domain deletion, domain fusion etc. In that sense domain architecture might be useful to investigate evolution of a protein family. Moreover, domains are the conserved regions and since they represent the similar regions in different proteins, they can give an idea about protein homology. In this project, D52 protein family is chosen due to its overexpression in tumour cells, it is related especially with breast carcinoma. D52 protein functions are searched and mentioned, also functions are investigated to clarify which regions are conserved and what these conserved regions effects on disease such as cancer are. Furthermore, each subfamily of this protein family is observed to answer following questions: Which regions are conserved in each subfamily and what is the reason of a protein is conserved in a specific subfamily, but it is not conserved in other families? D52 family is searched on PFAM and InterPro, InterPro contains 2066 different sequences. Since 2066 sequences is too much to align and draw phylogenetic tree, data from EggNOG which contains 230 proteins from 85 different species are gathered. For analysing these regions, conservation scores are calculated. While conservation scores are calculating, sequences are aligned and for each position of consensus sequence's conservation proportion is found, for this step an algorithm that is written by us is used. At last, with using chosen domains phylogenetic tree is formed and information about homology of domains is obtained.

Material Method

Project has started with searching protein families on PFAM database. While determining protein family, functions, expression of protein and domain organisations were examined and D52 protein has chosen. After choosing D52 protein family, in order to find domains of D52 family PFAM database is used. However, since there are lots of information about many species, InterPro database which is provided by EMBL-EBI is used to find predicted domains and important sites, is used. InterPro contains 2066 protein sequences from

different organisms, this database includes homologous protein sequences of TPD52 such as TPD53 and TPD54. However, distinguish proteins out of 2066 proteins is hard, therefore another database EggNOG which based on orthology predictions and phylogenetic data, is used to obtain sequences. Protein sequences are obtained as FASTA format from EggNOG and saved in “ENOG4111M9H.fa” file. These sequences are aligned by using MEGA7 and saved in “ENOG4111M9H-aligned.fa” file. Then for the first step their consensus sequence which represents the most common amino acid in each position is found and for each position conservation value is calculated. Following piece of code explains the algorithm of measurement of conservation values. Conservation scores have range between 0 and 1.

```

1  import re
2
3  def fastareader(filename):
4      seqDict = {}
5      with open(filename, "r") as f:
6          icerik = f.read()
7          myDict = [elem.replace("\n", "").strip("\r") for elem in re.split(r'>.*\n', icerik) if elem != ""]
8
9      keys = [item.lstrip(">") for item in icerik.split("\n") if item.startswith('>')]
10
11     seqDict = dict(zip(keys, myDict))
12     return seqDict
13
14
15 def consensusWithIdentity(sequences):
16     consensus_sequence = ''
17     consensus_vals = {}
18     for index, amino_acid in enumerate(sequences[0]):
19         counter = {}
20         for sequence in sequences:
21             if index < len(sequence):
22                 if sequence[index] in counter.keys():
23                     counter[sequence[index]] += 1
24                 else:
25                     if sequence[index] != "-":
26                         counter[sequence[index]] = 1
27         maxOne = max(counter, key=counter.get)
28         #print(maxOne)
29
30         if counter[maxOne] == 0:
31             consensus_sequence += '-'
32             consensus_vals[index] = 0
33         else:
34             consensus_sequence += maxOne
35             # print counter[maxOne] , sum(counter.values())
36             consensus_vals[index] = round(float(counter[maxOne])/sum(counter.values()), 3)
37             # print (consensus_vals)
38     return consensus_sequence , consensus_vals
39
40
41 def main():
42     myDict = fastareader("TPD52-aligned.fa")
43     consensus_seq, consensus_vals = consensusWithIdentity(list(myDict.values()))
44
45     with open("identity.txt", "w") as f:
46         for key in consensus_vals.keys():
47             f.write("{}\t{}\t{}\n".format(key, consensus_seq[key], consensus_vals[key]))
48
49     print("[*] Done")
50
51
52 if __name__ == "__main__":
53     main()

```

Amino acids on each position and calculation results are written in “identity.txt” file to be used draw conservation plot. By using RStudio, conservation plot is drawn. This plot will be used to analyse which regions are more conserved and to interpret if there is a mutation at well conserve region what the mutation’s effect might be.

For the second step, by using MEGA7 alignment, phylogenetic tree is constructed and Newick format of the phylogenetic tree is tried to be obtained. However, MEGA7 gave an error, therefore Newick format of the tree is obtained from EggNOG. Then, phylogenetic tree is drawn by using Figtree. In order to be consistent about phylogenetic tree, phylogenetic tree from EggNOG's itself is also taken account.

At last, human TPD52 protein is chosen. By using CDART, similar domain architectures under TPD52 protein superfamily are searched. Proteins' that are within this architecture, sequences are gathered and by using sequences their new phylogenetic tree is drawn. By using these phylogenetic trees and domain architecture, interpretations will be done about their homology. In addition, we used SMART and EggNOG tools to verify our results.

Results

In this study we focused on human TPD52 protein. TPD52 is tumour protein that placed in 8th chromosome in human. It has high-level expression in liver, kidney and colon tissues and low-level expression of heart, lung and skeletal muscle. The studies which are related with D52 homologous proteins, indicate that D52 protein play a role on cell proliferation and calcium signalling (Lewis et al, 2007). Other studies reveal D52 proteins regulate their activities with D52-like proteins, work with them, this may conclude that D52 proteins have a potential effect on controlling cell division. In addition, D52 have some effects on B cell differentiation. During differentiation from B cells to plasma cells, TPD52 level is observed at maximum level. D52 protein is also thought to be as target gene that increasing copy number of 8th chromosome. As result TPD52 has significant relationship especially breast carcinoma as well as other type of carcinomas, it may it may contribute to tumour initiation and progression.

After searching expressions and functions of TPD52 firstly, TPD52 proteins from different organism are obtained and aligned. Aligned sequences are used to create phylogenetic tree. Aligned sequences is shown in Figure1.

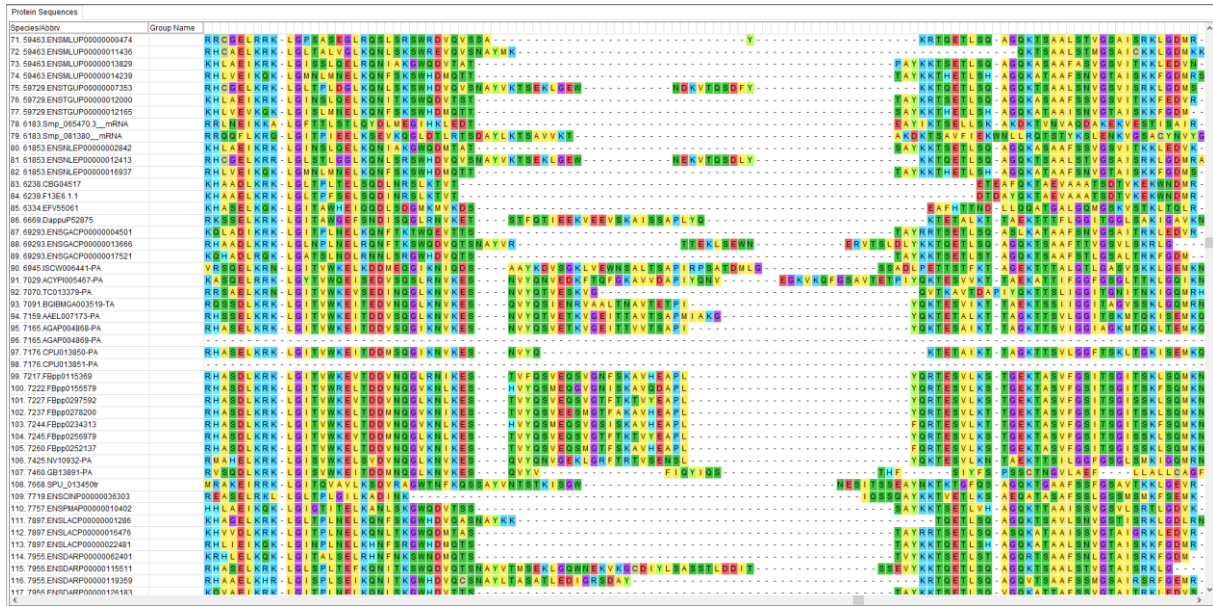


Figure 1: Protein sequence alignment

Full version of alignment data is uploaded to GitHub. According to multiple sequence alignment, TPD52 proteins from different species are mostly conserved for many position. There are some gaps for few organisms however at most region amino acid's properties remain similar even there is an amino acid change at specific position. To be able to interpret this alignment properly data consensus sequence is generated, then by using consensus sequence conservation score is calculated and plot of the conservation score (Figure 2) is drawn.

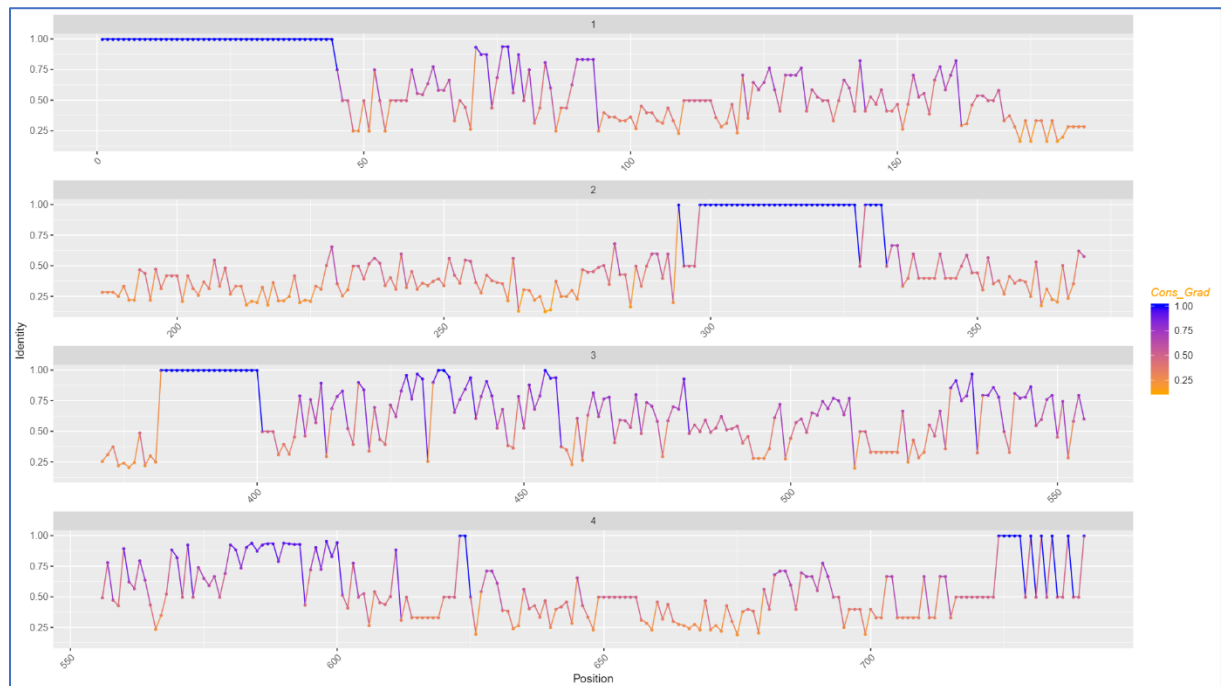


Figure 2: Conservation scores of each position

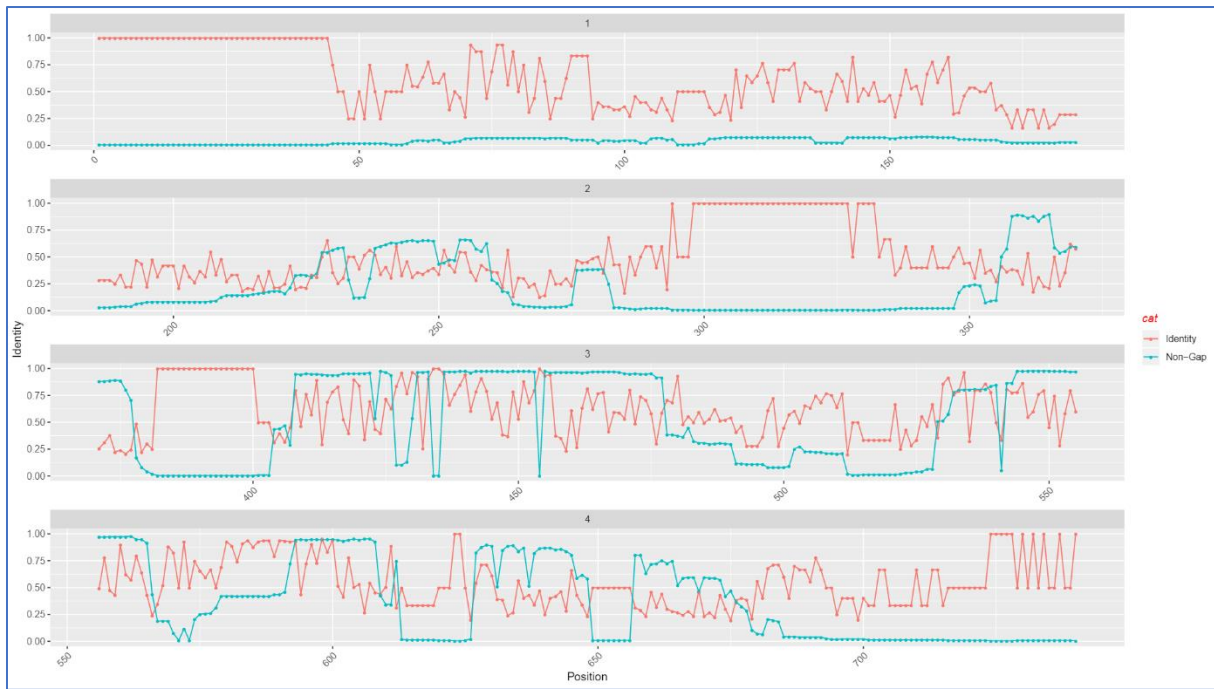


Figure 3: Conservation scores and non_gap region distribution

Figure 2 is also uploaded to GitHub. When the Figure 2 is examined, high conservation scores can be seen for most regions. There are some regions such as between position 300 and 325, conservation scores are 1. As a result, TPD52 protein change at that positions seems never changed in evolutionary scenario. This result occurs because of amino acids in that region are presumably observed in a few organisms and therefore in the alignment there are many gaps. However, this result is improper to interpret graph. That is why Figure 3 is plotted and it represents the number of gaps for each position. When number of gaps of positions that they have as conservation score 1, are examined, they are significantly high. On the other hand, there are also some regions that there are many changes such as amino acid substitution. For very conserved regions with significantly high conservation score, it can be said that if there is a mutation in that regions its effect is more significant than the regions that have lower conservation scores. Because when conservation score is low, many changes occurred evolutionary process without observing disease. As conclusion, if a mutation occurs at very conserved position, occurring a disease related with that mutation is probably observed on the phenotype of the organism.

For the next step, phylogenetic tree of these proteins is generated via both EggNOG and FigTree to pay attention to homogeny of these proteins. Figure 4 illustrates the phylogenetic tree which is created via FigTree. The list of organism names that have proteins shown in phylogenetic tree, is given under the tree and uploaded to GitHub.

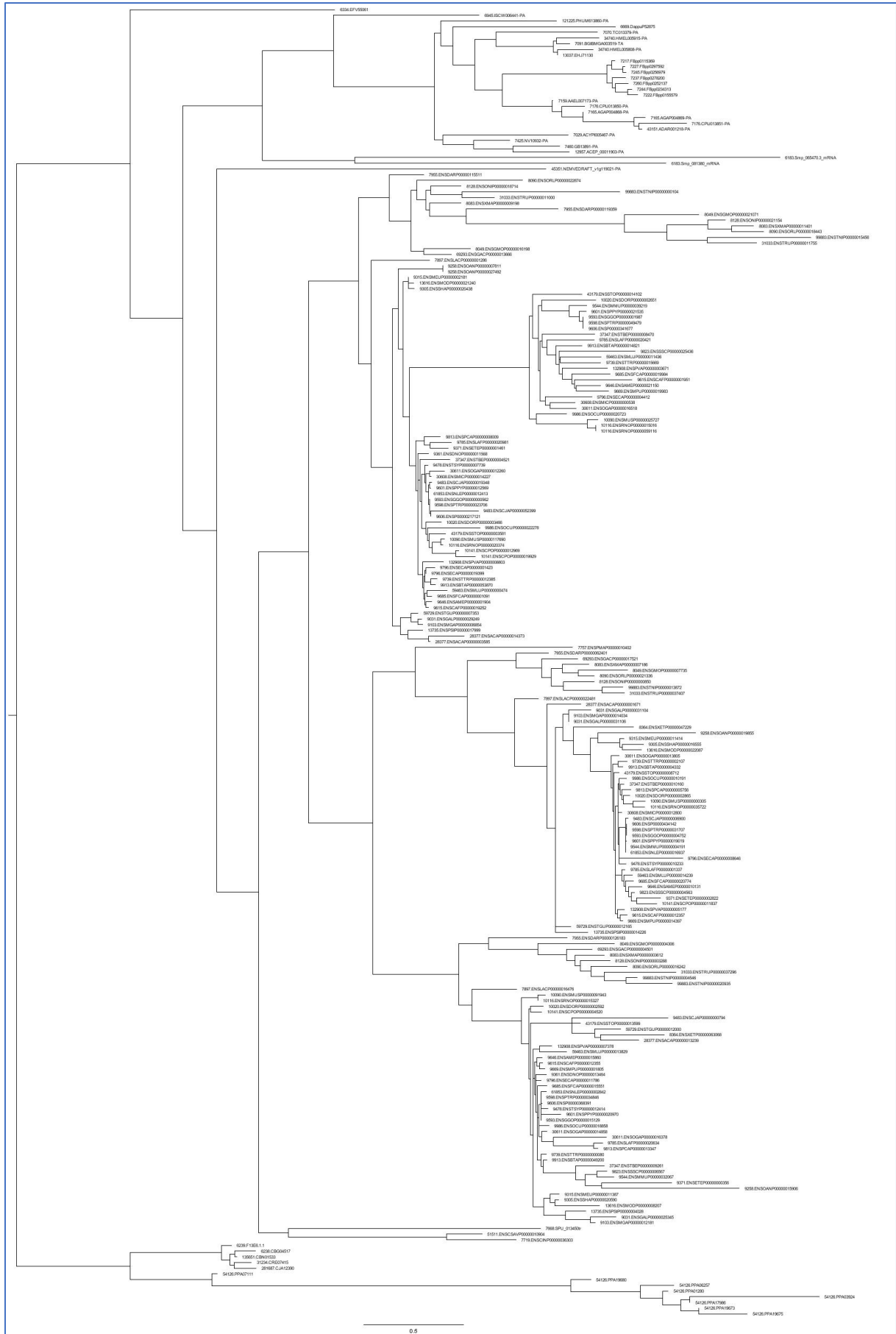


Figure 4: Phylogenetic tree of proteins

1	ACEP_00011903-PA	ACEP_00011903-PA	Acta cephalotes 12957	aliase	57	ENSRORLP0000001944	ENSRORLP0000001916	Oryzias latipes 5090	alia	113	ENSLAF000000020981	ENSLAF000000027960	Loxodonta africana	9785	ali
2	ENSPFF000000019018	TPD52L1	Pongo abelii 9601	aliase:ABM011138	58	ENSRORLP0000001396	TPD52L1	Oryzias latipes 5090	aliase:ENSBFT_0	114	ENSLAF000000011387	ENSLAF000000014601	Loxodonta africana	9785	ali
3	ENSPFF000000021535	TPD52L1	Pongo abelii 9601	aliase:ABM011138	59	ENSBTEFP00000002822	TPD52L1	Echinops telfairi 9371	aliase:D53,	115	ENSLAF000000020241	ENSLAF000000022756	Loxodonta africana	9785	ali
4	ENSPFF000000012569	TPD52L1	Pongo abelii 9601	aliase:ABM011138	60	ENSBTEFP00000003564	TPD52L1	Echinops telfairi 9371	aliase:D53,	116	FBpp0297592	CB5174	Drosophila melanogaster 7227	aliase:AEU13599_A,	
5	ENSPFF000000010181	TPD52L1	Pongo abelii 9601	aliase:ABM011138	61	ENSBTEFP00000001461	TPD52L1	Echinops telfairi 9371	aliase:D54,	117	EPV31041	TSP_08372	Tricholaima spiralis 6134	aliase:ABM011138	
6	ENSCOCFP000000018588	TPD52L1	Oryzias latipes 5090	aliase:AA	62	ENSHLEFP00000002842	ENSHLEFP00000002256	Nomascus leucogenys 41853		118	ENSONIFP000000021134	ENSONIFP000000021788	Oreochromis niloticus 8128		
7	ENSCOCFP000000022278	ENSCOCFP000000020982	Oryzias latipes 5090	aliase:AA	63	ENSHLEFP000000012413	ENSHLEFP000000010178	Nomascus leucogenys 41853		119	ENSONIFP00000003288	ENSONIFP000000020242	Oreochromis niloticus 8128		
8	ENSCOCFP000000011558	Oryzias latipes 5090	aliase:AA	64	ENSHLEFP000000014937	ENSHLEFP000000013932	Nomascus leucogenys 41853		120	ENSONIFP000000018714	ENSONIFP000000014566	Oreochromis niloticus 8128			
9	ENSCOCFP000000020723	ENSCOCFP000000016443	Oryzias latipes 5090	aliase:AA	65	TCU13379-PA	TCU13379-PA	Tribolium castaneum 7070	aliase:TCG052	121	ENSONIFP000000000580	ENSONIFP000000000672	Oreochromis niloticus 8128		
10	ENSBTF000000013599	TPD52L1	Itidomyia trideclinator 43179	aliase	66	FBpp0275200	GA18710	Drosophila pseudoobscura 7237	aliase:CBM0	122	ENSBTF000000020945	ENSBTF000000017763	Tetradodon nigroviridis 99883		
11	ENSBTF000000014102	TPD52L1	Itidomyia trideclinator 43179	aliase	67	Seq_065470.3	MDM	Seq_065470.3	MDM	123	ENSBTF000000014846	ENSBTF000000012486	Tetradodon nigroviridis 99883		
12	ENSBTF000000008712	TPD52L1	Itidomyia trideclinator 43179	aliase	68	Seq_061390	MDM	Seq_061390	MDM	124	ENSBTF000000001014	TPD52L1	Tetradodon nigroviridis 99883	aliase:CA	
13	ENSBTF000000003551	TPD52L1	Itidomyia trideclinator 43179	aliase	69	ISCM04941-PA	ISCM04941-PA	Isodon scapularis 6945	aliase	125	ENSBTF000000013072	TPD52L1	Tetradodon nigroviridis 99883	aliase:CA	
14	ENSBTF000000015016	ENSBTF000000011304	Rattus norvegicus 10116	aliase	70	ENSDORFP00000002845	TPD52L1	Dipodomys ordii 10020	aliase:ENSDOR0	126	ENSBTF000000004846	ENSBTF000000021111	Tetradodon nigroviridis 99883		
15	ENSBTF000000015327	TPD52L1	Rattus norvegicus 10116	aliase:ABM06	71	ENSDORFP000000026251	TPD52L1	Dipodomys ordii 10020	aliase:ENSDOR0	127	ENSCJAP00000002399	ENSCJAP000000007868	Callithrix jacchus 9493	aliase	
16	ENSBTF000000015016	TPD52L1	Rattus norvegicus 10116	aliase:ABM06	72	ENSDORFP00000002846	TPD52L1	Dipodomys ordii 10020	aliase:ENSDOR0	128	ENSCJAP000000019348	TPD52L1	Callithrix jacchus 9493	aliase:ACFV01	
17	ENSBTF000000035722	TPD52L1	Rattus norvegicus 10116	aliase:ABM06	73	ENSDORFP00000002592	TPD52L1	Dipodomys ordii 10020	aliase:ENSDOR0	129	ENSCJAP000000006900	TPD52L1	Callithrix jacchus 9493	aliase:ACFV01	
18	ENSBTF0000000059116	TPD52L1	Rattus norvegicus 10116	aliase:ABM06	74	ENSBTF000000036391	TPD52L1	Homo sapiens 9606	aliase:ACQ0966,ACQ	130	ENSCJAP000000000794	TPD52L1	Callithrix jacchus 9493	aliase:ACFV01	
19	ENSBTF000000010233	TPD52L1	Rattus norvegicus 10116	aliase:ABM06	75	ENSBTF000000014912	TPD52L1	Homo sapiens 9606	aliase:ABM11939,ALL	131	ENSCJAP000000004812	TPD52L1	Equus caballus 9796	aliase:HTD_3923, F	
20	ENSBTF000000012414	TPD52L1	Rattus norvegicus 10116	aliase:ABM06	76	ENSBTF000000014677	TPD52L1	Homo sapiens 9606	aliase:ABM11939,ALL	132	ENSCJAP000000001423	TPD52L1	Equus caballus 9796	aliase:FBM27_HOR	
21	WU1045-PA	WU1045-PA	Neomys viridipennis 7425	aliase:WU1045, WU1045, F	77	ENSBTF000000017121	TPD52L1	Homo sapiens 9606	aliase:AF004429,AF0	133	ENSCJAP000000005646	TPD52L1	Equus caballus 9796	aliase:FBM27_HOR	
22	CF1J013851-PA	CF1J013851-PA	Culex quinquefasciatus 7176	aliase	78	ENSBTF000000017121	TPD52L1	Homo sapiens 9606	aliase:AF004429,AF0	134	ENSCJAP000000001437	TPD52L1	Equus caballus 9796	aliase:FBM27_HOR	
23	CF1J013850-PA	CF1J013850-PA	Culex quinquefasciatus 7176	aliase	79	ENSBTF000000011985	TPD52L1	Ornithorhynchus anatinus 9258	alia	135	ENSCJAP000000019399	TPD52L1	Equus caballus 9796	aliase:FBM27_HOR	
24	ENSBTF000000003568	TPD52L1	Xenopus (Silurana) tropicalis 8364	aliase	80	ENSBTF000000007811	TPD52L1	Ornithorhynchus anatinus 9258	alia	136	ENSBTF000000014387	TPD52L1	Equus caballus 9796	aliase:FBM27_HOR	
25	ENSBTF000000004729	TPD52L1	Xenopus (Silurana) tropicalis 8364	aliase	81	ENSBTF0000000027452	ENSBTF0000000021592	Ornithorhynchus anatinus 9258	alia	137	ENSBTF000000011985	TPD52L1	Equus caballus 9796	aliase:FBM27_HOR	
26	ENSBTF000000002107	TPD52L1	Turaxopoda truncatus 9739	aliase:D53,MD	82	ENSBTF0000000031104	ENSBTF0000000021976	Gallus gallus 9031	alia	138	ENSBTF000000000474	ENSBTF000000000525	Myotis lucifugus 59463	ali	
27	ENSBTF000000013885	TPD52L1	Turaxopoda truncatus 9739	aliase:D54,MD	83	ENSBTF0000000031104	TPD52L1	Gallus gallus 9031	aliase:AAW0201	139	ENSBTF000000014239	ENSBTF000000012568	Myotis lucifugus 59463	ali	
28	ENSBTF000000013669	TPD52L1	Turaxopoda truncatus 9739	aliase:HTD_3923, F	84	ENSBTF0000000029249	TPD52L1	Gallus gallus 9031	aliase:AAW0201	140	ENSBTF000000014239	ENSBTF000000012568	Myotis lucifugus 59463	ali	
29	ENSBTF000000000080	TPD52L1	Turaxopoda truncatus 9739	aliase:D52,MD	85	ENSBTF0000000025945	TPD52L1	Gallus gallus 9031	aliase:AAW0201	141	ENSBTF000000013029	ENSBTF000000010121	Myotis lucifugus 59463	ali	
30	ADAM011218-PA	ADAM011218-PA	Amegilla daeclia 43151	aliase:ADAM01	86	ENSBTF0000000001177	TPD52L1	Pteropus vampyrus 132908	aliase:D53,	142	ENSBTF0000000032067	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
31	Deppu021875	ENSBTF00000002575	Daphnia pulex 6669	aliase:UL732554, D52B3	87	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	143	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
32	ENSBTF000000019252	TPD52L1	Canis lupus familiaris 9615	aliase:CF	88	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	144	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
33	ENSBTF0000000001951	TPD52L1	Canis lupus familiaris 9615	aliase:CF	89	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	145	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
34	ENSBTF000000012355	TPD52L1	Canis lupus familiaris 9615	aliase:CF	90	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	146	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
35	ENSBTF000000012355	TPD52L1	Canis lupus familiaris 9615	aliase:CF	91	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	147	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
36	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	92	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	148	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
37	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	93	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	149	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
38	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	94	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	150	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
39	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	95	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	151	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
40	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	96	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	152	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
41	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	97	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	153	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
42	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	98	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	154	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
43	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	99	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	155	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
44	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	100	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	156	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
45	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	101	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	157	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
46	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	102	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	158	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
47	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	103	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	159	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
48	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	104	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	160	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
49	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	105	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	161	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
50	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	106	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	162	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
51	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	107	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	163	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
52	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	108	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	164	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
53	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	109	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	165	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
54	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	110	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	166	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
55	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	111	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	167	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
56	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	112	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	168	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
57	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	113	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus 132908	aliase:D54,	169	ENSBTF0000000039219	TPD52L1	Macaca mulatta 9544	aliase:CBM01267, F	
58	ENSBTF000000014227	TPD52L1	Microtus murinus 30608	aliase:HTD_3923, F	114	ENSBTF00000000007378	TPD52L1	Pteropus vampyrus							

At last, domain architecture of TPD52 human protein is examined. However, domain architecture itself cannot be explicated (Figure 5). Therefore, by using proteins' that are within domain architecture a new phylogenetic tree is constructed with blast=100 (Figure 5).



Figure 5: Proteins within domain architecture

This phylogenetic tree verifies the result of orthologous organisms in the multiple sequence alignment phylogenetic tree. For instance, as Figure 5 is examined, Homo Sapiens (Human) and Mus musculus (Mouse) are placed as they have orthologous proteins of TPD52, also another example reveals that Ornithorhynchus anatinus (TPD52) share common region that is very similar to Homo Sapiens (TPD52) and they are orthologous proteins. Then tree in Figure 5 is verified by using SMART and EggNOG. Figure 6 illustrates the part of tree which

is generated by EggNOG, in that tree there are more organisms to compare each other and PFAM domain architecture can be seen beside the tree.

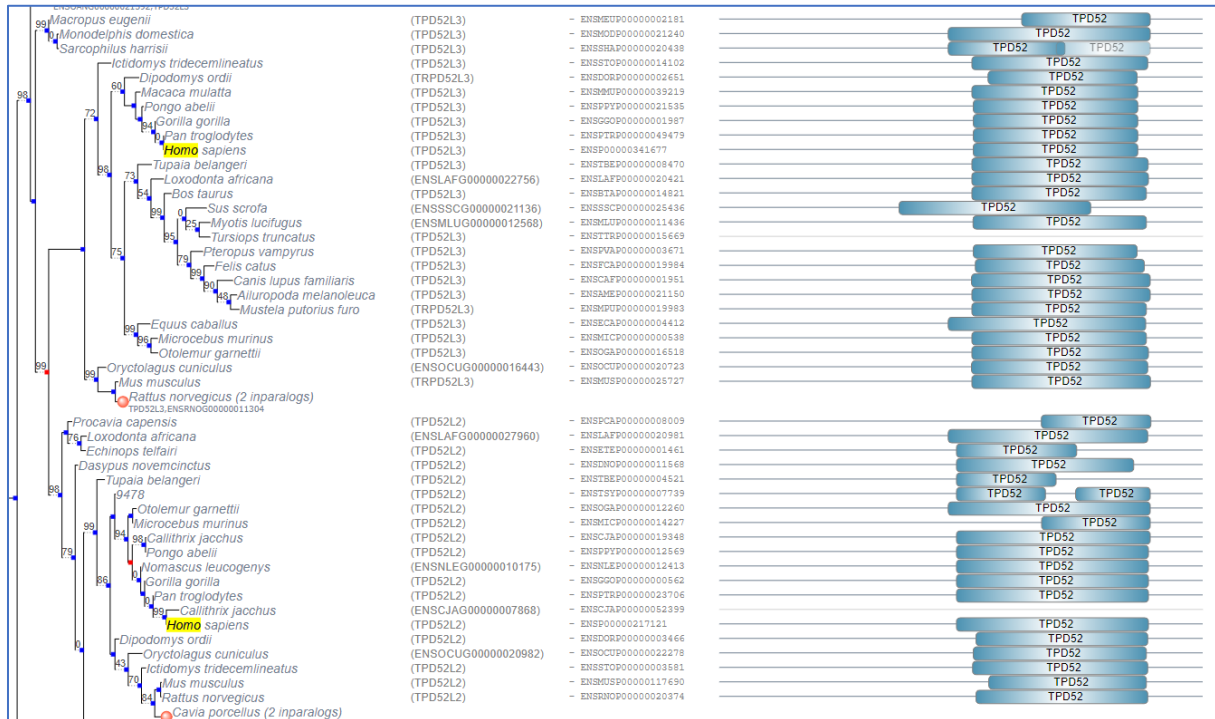


Figure 6: Domain alignment tree by EggNOG

Since there is no specific filtration in these tools number of organisms therefore homologous proteins are much more that we consider. When we look at Homo sapiens protein, phylogenetic tree is the evidence of gene duplication. In addition, while domain architectures are examined, it can be seen that, domains are very conserved, and these organisms have strong relations between each other. Figure 7 represents a closer look to domain architecture of these organisms. Figure 7 is also generated by EggNOG for aligned blocks and shows SMART domains of the species.

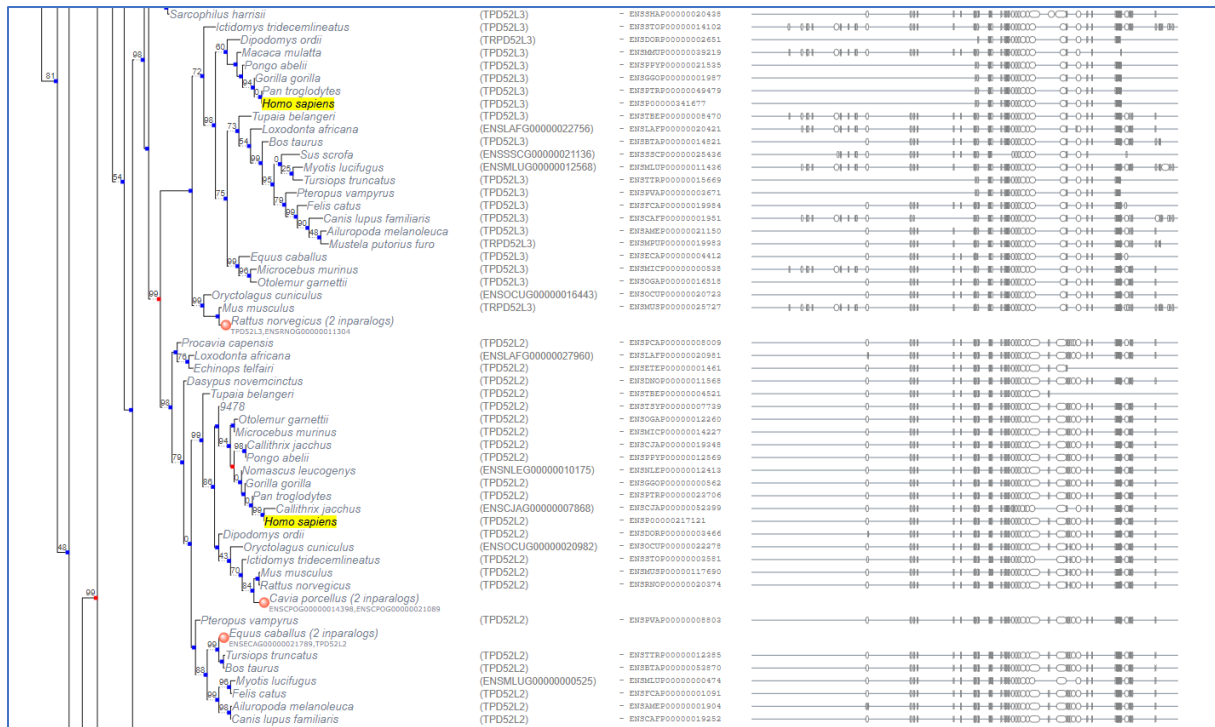


Figure 7: Closer look to aligned blocks

Discussion

This study focused on TPD52 protein in human while its domains were examined and interpreted. Other than domains, conservation scores were calculated and interpreted. Since there are lots of data, and very big phylogenetic tree, small proportion of phylogenetic tree is used while data was construing. In results part, homology of TPD52 protein was explained with few examples. For the future studies, big picture of phylogenetic tree can be studied and homology of the proteins can be proved by many examples.

At first study focused on only TPD52 proteins, then TPD52 family extended to TPD52-like and TPD53, TPD54 proteins. This project is planned to study with superfamily and subfamilies of TPD52 protein. However, we could not be able to distinguish subfamilies. Subfamilies are also can be subject of future studies. Which functions are conserved at which subjects might be good for future studies.

Conservation values can be evaluated by not using consensus sequence but amino acid properties. Since some amino acids have similar properties such as positively charged groups or hydrophobic side chains, when there is an amino acid change, its effect might not be as harmful as amino acid property change.

References

- (n.d.). Retrieved from http://atlasgeneticsoncology.org/Genes/GC_TPD52.html
- A testis-specific and testis developmentally regulated tumor protein D52 (TPD52)-like protein TPD52L3/hD55 interacts with TPD52 family proteins. Cao Q, Chen J, Zhu L, Liu Y, Zhou Z, Sha J, Wang S, Li J. *Biochem Biophys Res Commun*. 2006 Jun 9;344(3):798-806. Epub 2006 Apr 19.
- Carugo, O., & Eisenhaber, F. (2016). *Data Mining Techniques for the Life Sciences*. Totowa: Humana Press.
- EMBL-EBI, I. (n.d.). InterPro. Retrieved from <http://www.ebi.ac.uk/interpro/entry/IPR007327/proteins-matched?start=20>
- European Bioinformatics Institute Protein Information Resource SIB Swiss Institute of Bioinformatics. (2018, December 05). Tumor protein D52. Retrieved from <https://www.uniprot.org/uniprot/P55327#function>
- European Bioinformatics Institute Protein Information Resource SIB Swiss Institute of Bioinformatics. (n.d.). Tumor protein D52 (TPD52): A novel B-cell/plasma-cell molecule with unique expression pattern and Ca (2)-dependent association with annexin VI. Retrieved from <https://www.uniprot.org/citations/15576473>
- European Bioinformatics Institute Protein Information Resource SIB Swiss Institute of Bioinformatics. (n.d.). Identification of homo- and heteromeric interactions between members of the breast carcinoma-associated D52 protein family using the yeast two-hybrid system. Retrieved from <https://www.uniprot.org/citations/9484778>
- Family: TPD52 (PF04201). (n.d.). Retrieved from <https://pfam.xfam.org/family/PF04201#tabview=tab0>
- Gene Tree Image. (n.d.). Retrieved from <http://www.ensembl.org/Multi/GeneTree/Image?gt=ENSGT00940000155294>
- Induction of tumorigenesis and metastasis by the murine orthologue of tumor protein D52. Lewis JD, Payton LA, Whitford JG, Byrne JA, Smith DI, Yang L, Bright RK. *Mol Cancer Res*. 2007 Feb;5(2):133-44.
- Isolation and characterization of a novel gene expressed in multiple cancers. Chen SL, Maroulakou IG, Green JE, Romano-Spica V, Modi W, Lautenberger J, Bhat NK. *Oncogene*. 1996 Feb 15;12(4):741-51.
- Koonin, E. V. (1970, January 01). *Evolutionary Concept in Genetics and Genomics*. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK20255/>
- The tumor protein D52 family: Many pieces, many puzzles. (2004, November 06). Retrieved from <https://www.sciencedirect.com/science/article/pii/S0006291X0402409X>
- What are protein families? (2017, March 27). Retrieved from <https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-families>

