Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supplementary Information

For

The relationship between classification of multi-domain proteins using an alignment-free approach and their functions: a case study with Immunoglobulins.

Ramachandra M. Bhaskara[a], Prachi Mehrotra[a], Ramaswamy Rakshambikai[a], Mutharasu Gnanavel[a], Juliette Martin[b] and Narayanaswamy Srinivasan[a*]

[1]Molecular Biophysics Unit,
[2]Indian Institute of Science Mathematics Initiative,
Indian Institute of Science
Bangalore 560012,
India

[3]Bases Moléculaires et Structurales des Systèmes Infectieux
UMR 5086 CNRS/Université Lyon
7, passage du Vercors
69367 Lyon cedex 7
France

[*]Corresponding Author Prof. N Srinivasan (*ns@mbu.iisc.ernet.in*)

Table of Contents for Supplementary Files

Supplementary File1
This document contains
1. Supplementary Figure(s) S1-S4
2. Supplementary Table(s) S1-S6

Supplementary File 2-4
These files contain the UniProt accession codes:
Protein Kinases accession codes with Family and Sub-family names (**Supplementary File 2;** n=1498).
Ig domain family accession codes with GO terms and SwissProt entries' (**Supplementary File 3;** n=455) and
All Ig domain family accession codes (**Supplementary File 4;** n=7121).

Supplementary File 5-13
These files contain final trees of Pk and Ig proteins clustered using LMS and ClustalW at both domain and full-length sequence levels

Supplementary File 5:
Dendrogram of Ig 455 sequences clustered using ClustalW tau distances using domain level.

Supplementary File 6:
Dendrogram of Ig 455 sequences clustered using ClustalW tau distances using full-length level.

Supplementary File 7:
Dendrogram of Ig 455 sequences clustered using LMS distances using domain level.

Supplementary File 8:
Dendrogram of Ig 455 sequences clustered using LMS distances using full-length level.

Supplementary File 9:
Dendrogram of Ig 7121 sequences clustered using LMS distances using full-length level parsed at cut-off of 0.2. The clusters are coloured.

Supplementary File 10:
Dendrogram of Pk sequences clustered using ClustalW tau distances using domain level.

Supplementary File 11:
Dendrogram of Pk sequences clustered using ClustalW tau distances using full-length level.

Supplementary File 12:
Dendrogram of Pk sequences clustered using LMS distances using domain level.

Supplementary File 13:
Dendrogram of Pk sequences clustered using LMS distances using full-length level.

Supplementary File 14:
This file contains the UniProt accession codes of Antibody heavy chain sequences from mammals (fasta format) and their corresponding sub-type specifications.

Supplementary Information File 1
This document contains

1. Supplementary Figure(s) S1-S4
2. Supplementary Table(s) S1-S6

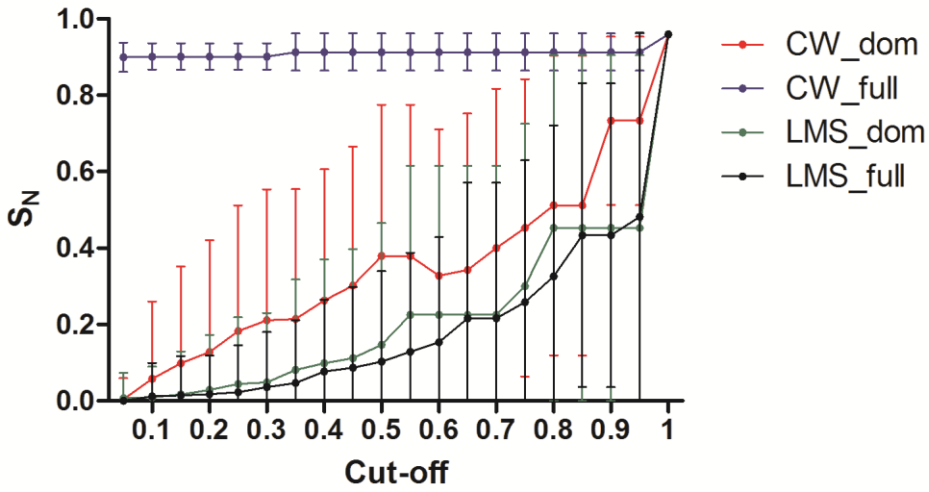


Figure S1: Cluster purity of protein kinases: Relative entropy values averaged over all the clusters as a function of dendrogram parsing value. The entropy computations are done at the kinase group for every cluster, where the number of possible states of a sequence is 7 corresponding to the 7 kinase groups. Here too we observe that the LMS_full performs better than the rest of the methods as the rate of increase of the function is far slower than the alignment based methods.

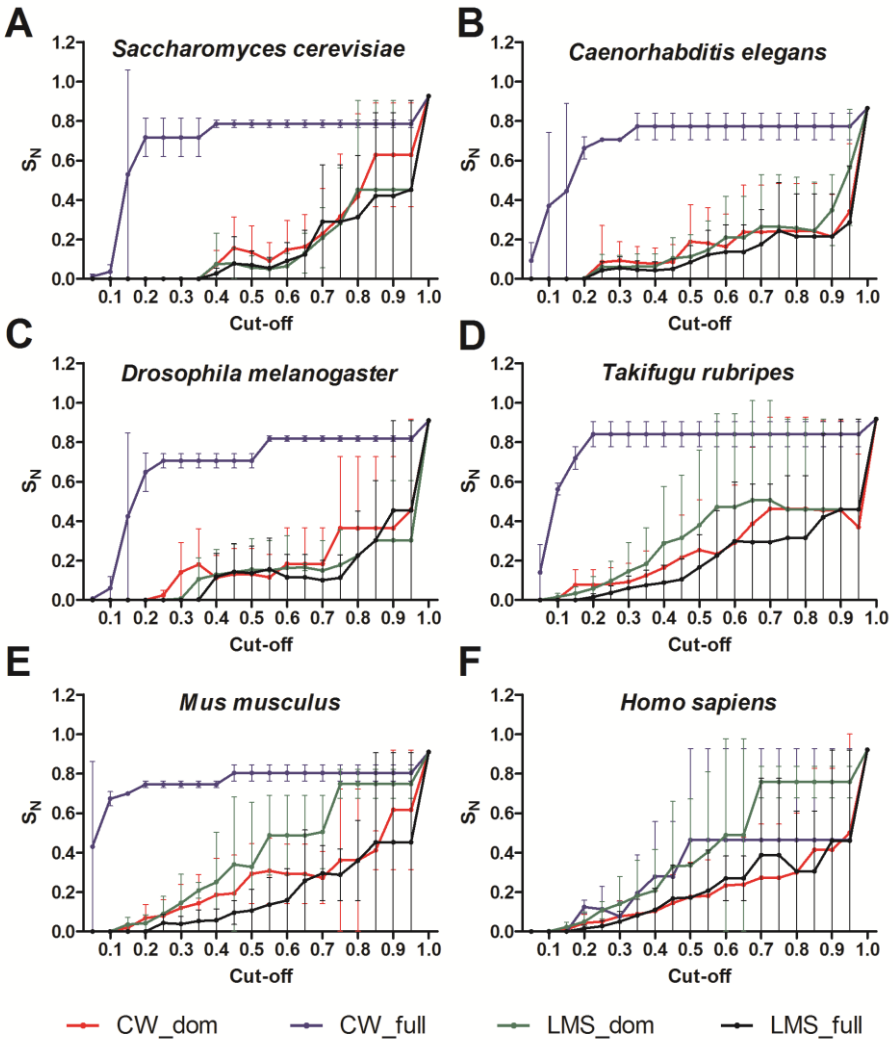


Figure S2: Species specific performance of local matching score and clustal tau distance based clustering of protein kinases. Protein kinases from the six model organisms were clustered using the LMS and CW based distances computed at both at the level of kinase domain alone and using the full length gene products were clustered to obtain trees. Observe that the rate of increase of the relative entropy or the inverse of cluster purity is slow as a function of cut-off value for LMS-full length. In case of yeast and *C. elegans* where multi-domain protein kinases are few, the LMS-full, LMS-dom and CW_dom perform more or less in a similar manner.

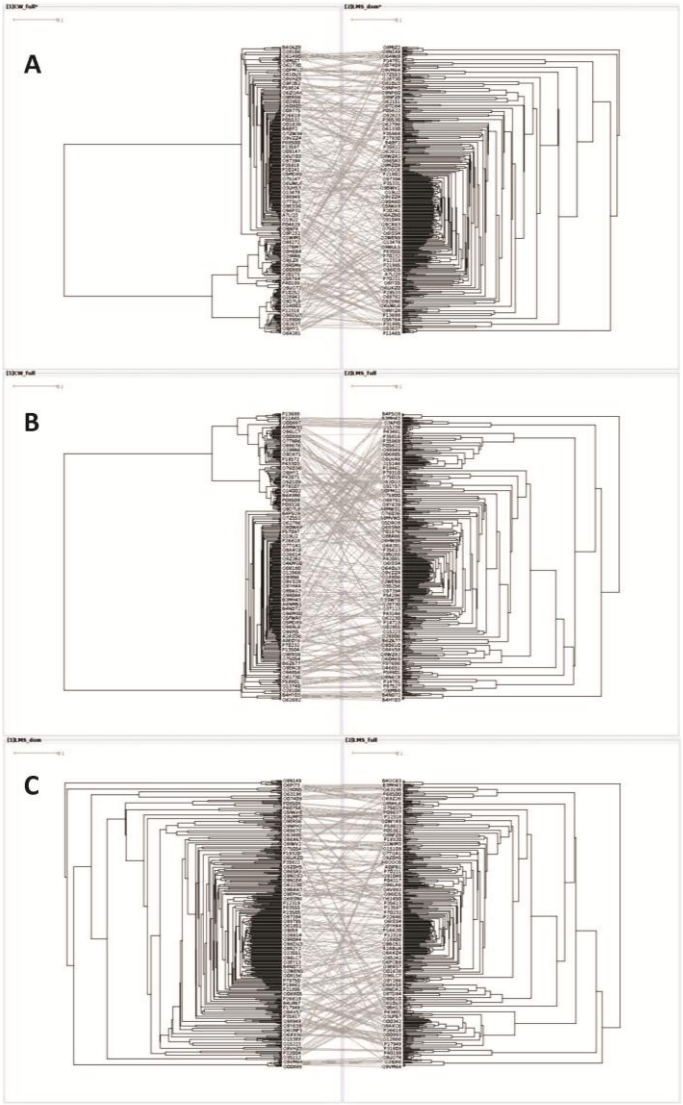


Figure S3: Immunoglobulin tanglegrams: Tanglegrams used for comparing the two different tree topologies having the same set of leaves (*n*=455). The rest of the three comparisons are shown here **A.** CW_full and LMS_dom trees **B.** CW_full and LMS_full and **C.** LMS_dom and LMS_full. The parallel lines join the equivalent leaves from the two trees. The criss-crossing of the lines in between represent the extent of the tangle between the two trees. Note that using LMS based clustering produces disjoint distribution of distances for Immunoglobulins, i.e. two distinct branches which are distant to each other.

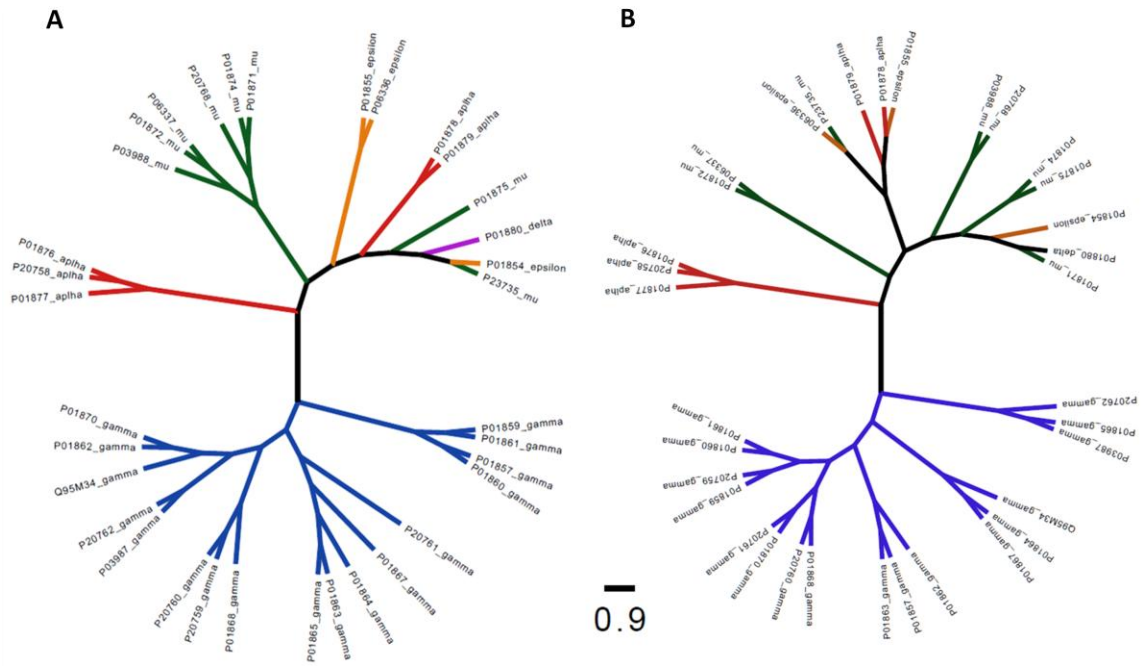


Figure S4: Immunoglobulin subtype clustering: Clustering of Antibody heavy chain mammalian sequences (*n*=34) from Pfam using **A.** LMS based and **B**. CW based distances. The colours represent the different Ig isotypes (*blue*=IgG (17), *red*=IgA (5), *green*=IgM (8), *yellow*=IgE (3) and *purple*=IgD (1)). We can see that the IgG is clustering well with both the methods, but the IgM is only well clustered in LMS based method. We find that the dispersion of IgE and IgD sequences is common to both the clustering but IgA sequences co-cluster with IgE sequences in the CW based tree. The correlation between the two trees is (n=1122; r=0.8; p<0.05; Pearson's correlation).

Table S1: Dataset of protein kinase and immunoglobulin sequences used for the current study. The fasta sequences of all the datasets used are provided as Supplementary files 2-4. The ***n*** value represents the number of sequences in each dataset; ***k*** represents the number of annotated protein kinase subfamilies for each genome and ***d*** represents the number of unique domain architectures observed for each dataset of Ig_2 domain containing proteins in PFam. There is no description of distinct subfamilies definitions for Ig2 containing sequences.

A. Protein Kinases	***N***	***k***
YEAST (*Saccharomyces cerevisiae*)	70	20
CAEEL (*Caenorhabditis elegans*)	214	59
DROME (*Drosophila melanogaster*)	96	44
FUGU (*Takifugu rubripes*)	418	72
MOUSE (*Mus musculus*)	326	76
HUMAN (*Homo sapiens*)	374	81
ALL (*All Protein Kinases*)	1498	92
B. Immunoglobulins	***N***	***d***
Ig2 dataset (SP)	455	200
Ig2 dataset (Full)	7121	1467

Table S2: Classification of protein kinases using LMS and full-length protein sequences

C_i	N_i	$S_{n,i}$	k_i	Functional Similarity (GO)			Domain architectural similarity			Predominant Sub-family	All Sub-families within the cluster	Domain organization	Sub-family defining features	PMID
				BP	CC	MF	J_{PQ}	γ_{PQ}	D_{PQ}					
1	14	0.00	1	2.11	0.62	1.77	0.65	0.36	0.70	RSK	RSK	Has two tandem kinase domains	Has two tandem kinase domains; C-ter kinase domain acts as a regulatory domain; S232, S372, S389, and T581 are crucial for activation	15632195
2	**9**	**0.00**	**1**	**1.54**	**0.44**	**1.62**	**1.00**	**1.00**	**1.00**	**DYRK**	**DYRK**	**Single domain kinase**	**Dual specificity Kinase; activation loop contains Y-X-Y motif and HCD instead of HRD; Autophosphorylation of Tyr residues required for localization to**	**9748265, 9932450**
3	**11**	**0.00**	**1**	**1.76**	**0.40**	**2.64**	**1.00**	**1.00**	**1.00**	**CAMK2**	**CAMK2**	**Single domain kinase**	**Have an auto-inhibitory mechanism analogous to a pseudo substrate; Ca^{++} and phosphorylation of T286 required for activation; forms functional oligomers via self association region to form a flower like structure with 8-10 petals.**	**10473573, 1659571**
4	12	0.00	1	1.38	0.50	1.55	0.76	0.53	0.84	SGK	SGK	Mostly Single domain kinase; few have PX domain at the C-terminus	HM and T loop are required for activation by characteristic associations with HM kinase and PI3; PX is required for targeting to endosomes.	16619268
5	9	0.00	1	2.20	0.81	2.00	0.63	0.44	0.78	PKC	PKC	Multi-domain kinase; Has C1,C2 and Pk domain	C1 is involved in binding to diacyl glycerol; C2 is Calcium sensor	10764742

6	122	0.68	28	1.65	0.90	1.66	0.34	0.09	0.52	Ack	Abl,Ack,ALK,Axl,CCK4,DDR,EGFR,Eph,FAK,Fer,FGFR,JakA,KIN16,KIN6,LISK,Lmr,Met,Musk,PDGFR,Ret,Ror,Ryk,Src,STE11,STKR,Syk,Tie,TK-Sp1	Mostly multi-domain kinases; Belong to Tyrosie kinase group; Diverse domain combinations	Tyrosine kinase like features specific to Kinase domain and the tethered domains; In Ack kinases (representative) SH3, SAM and GTPase domains are tethered required for interaction	8497321
7	260	0.77	55	1.49	0.27	1.42	0.56	0.30	0.66	CAMKL	AKT,ALK,CAMK1,CAMKL,CASK,CCK4,CDK,CLK,DAPK,DCAMKL,DDR,DMPK,Dual,DYRK,FAK,Fer,GRK,GSK,InsR,IRA,LRRK,MAPKAPK,MAST,MLCK,MLK,NDR,PDK1,PIM,PKA,PKD,RAD53,RAF,RCK,RIPK,RSK,RSKL,RSKR,Sev,SRPK,STE11,STE20,STE7,STKR,Syk,Trio,TSSK,TTBK,TTBKL,VRK,Worm10,Worm6,Worm7,Worm8,Worm9,YANK	Mostly multi-domain kinases; Most of them belong to CAMK group. In CAMKL, Pk domain tethered to UBA domain and long C-ter overhangs	Ca^{++} binding regions are at the N-ter or the C-ter of the Kinase domains and in few kinases the Calmodulin binding region is at the C-ter of the Kinase domain; Few kinases have EF had domains and KA1 domains at the C-ter.	10413400
8	**7**	**0.00**	**1**	**1.62**	**0.08**	**1.61**	**1.00**	**1.00**	**1.00**	**CAMKL**	**CAMKL**	**In CAMKL, Pk domain tethered to UBA domain and long C-ter overhangs**	**Few kinases have EF had domains and KA1 domains at the C-ter. Has variable C-ter domain unassigned regions. Functions in proteins involved in Ubiquitination pathway; associates with EF-Tu**	**10413400**
9	78	0.34	8	1.62	0.34	1.57	0.97	0.95	0.98	CDK	CDK,CDKL,DYRK,GSK,MAPK,PSK,RCK,RSK	Predominantly single domain kinases except RSK; All belong to CMGC group	PSTAIRE helix motif in the Kinase domain is important for interaction with cyclin; WP insert at position 227, T160 are required for activation.	8756328, 10559988
10	5	0.00	1	2.98	0.08	2.56	0.87	0.60	0.89	GRK	GRK	Two domain kinase having Pk and RGS domain	S484 and T485 phosphorylation required for binding to arrestin; C-ter extension of Pk domain required for nucleotide binding.	18339619

11	16	0.00	1	3.02	0.21	2.02	1.00	0.48	1.00	GRK	GRK	Two domain kinase having Pk and RGS domain	S484 and T485 phosphorylation required for binding to arrestin; C-ter extension of Pk domain required for nucleotide binding.	18339619
12	**15**	**0.00**	**1**	**2.17**	**0.82**	**1.68**	**1.00**	**1.00**	**1.00**	**CDK**	**CDK**	**Single domain kinases**	**PSTAIRE helix motif in the Kinase domain is important for interaction with cyclin; WP insert at position 227, T160 are required for activation.**	**8756328, 10559988**
13	40	0.49	10	1.52	0.44	1.45	0.76	0.61	0.81	CAMK1 and MLCK	Akt,AKT,CAMK1,DCAMKL,MAPKAPK,MLCK,PDK1,PKA,PKC,PSK	Mostly single domain kinases with both N and C-ter overhangs and domain unassigned regions	Characteristic C-ter autoinhibitory and calmodulin binding region	7641687
14	**7**	**0.00**	**1**	**2.40**	**0.97**	**1.99**	**1.00**	**1.00**	**1.00**	**CAMKL**	**CAMKL**	**In CAMKL, Pk domain tethered to UBA domain and long C-ter overhangs**	**Few kinases have EF had domains and KA1 domains at the C-ter. Has variable C-ter domain unassigned regions. Functions in proteins involved in Ubiquitination pathway; associates with EF-Tu**	**10413400**
15	10	0.00	1	1.79	0.95	1.46	0.81	0.62	0.85	CK2	CK2	Single domain kinases	R198K substitution and G199 is characteristic of CK2 in comparison with CMGC; E180 is important for regulation of activation loop.	15273306
16	12	0.00	1	2.50	0.64	2.04	0.85	0.70	0.88	MAPK	MAPK	Single domain kinases	Characteristic T-X-Y motif in activation loop for regulation of activity; docking groove and ED motif adjacent to it help in recognition of specific interacting proteins.	12639708
17	34	0.19	3	1.58	0.00	1.48	1.00	1.00	1.00	TTBKL	Dual,TTBKL,Worm6	TTBKL and Worm6 have single Pk domain; while Dual kinases have two Pk domains	All belong to CK1 group and specific to worm lineage; TTBKL may function to phosphorylated Tau proteins.	16923168

18	**22**	**0.00**	**1**	**1.77**	**0.80**	**1.43**	**1.00**	**1.00**	**1.00**	**CK1**	**CK1**	**Single domain kinases**	**Characteristic replacement of APE motif in activation loop to SIN**	**7768349**
19	6	0.00	1	3.00	2.21	2.06	0.73	0.47	0.85	PKC	PKC	Multi-domain protein kinase; Has C1,C2 and Pk domain	C1 is involved in binding to diacyl glycerol; C2 is Calcium sensor	10764742
20	12	0.00	1	1.54	0.22	1.87	0.85	0.70	0.88	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases;	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
21	15	0.15	2	1.72	0.57	1.60	0.44	0.17	0.47	Trio	Trio, MLCK	Multi-domain Kinases having Ig like domains associated	GEFdomain association helps in GDP exchange and involved in Cell growth and actin polymerization	8643598
22	**7**	**0.00**	**1**	**1.75**	**1.17**	**1.72**	**1.00**	**1.00**	**1.00**	**DYRK**	**DYRK**	**Single domain kinase**	**Dual specificity Kinase; activation loop contains Y-X-Y motif and HCD instead of HRD; Autophosphorylation of Tyr residues required for localization to**	**9748265, 9932450**
23	11	0.00	1	1.87	0.66	1.67	0.84	0.67	0.84	PKA	PKA	Single domain kinases	Activation regulated by cAMP coupled with dissociation regulatory subunit and phosphorylation at 3 sites 239, 240, 241	20027184
24	18	0.00	1	1.91	0.25	1.54	0.90	0.79	0.92	STE7	STE7	Mainly single kinase domain, sometimes associated with PB1	Dual specific kinase and phosphorylates S/T/Y residues.	1628831
25	9	0.15	2	1.98	0.41	1.66	1.00	1.00	1.00	MAPK(5)	GSK,MAPK	Single domain kinases	CMGC group; In MAPK, characteristic T-X-Y motif in activation loop for regulation of activity; docking groove and ED motif aid in recognition of specific interacting proteins.	12639708
26	14	0.00	1	1.47	0.17	2.15	0.77	0.40	0.72	PKG	PKG	Kinase domain and cNMP binding domain	cNMP binding domain at N terminus of Pkinase	20027184
27	**7**	**0.00**	**1**	**1.65**	**0.09**	**1.53**	**1.00**	**1.00**	**1.00**	**CAMKL**	**CAMKL**	**In CAMKL, Pk**	**Few kinases have EF had**	**10413400**

												domain tethered to UBA domain and long C-ter overhangs	**domains and KA1 domains at the C-ter. Has variable C-ter domain unassigned regions. Functions in proteins involved in Ubiquitination pathway; associates with EF-Tu**	
28	39	0.10	2	1.49	0.08	1.57	0.69	0.31	0.75	Fer(32)	Fer,Src	Protein kinase and SH2 domain	Both are Tyrosine kinases. Fer has FCH domain in addition to SH2 occasionally.	9722593
29	11	0.00	1	1.81	0.79	1.58	0.95	0.63	0.84	PKN	PKN	Multi-domain; Protein kinase, HR1	HR1 binds to Rho GTPases	8647255
30	13	0.00	1	1.98	0.64	1.69	0.92	0.85	0.94	CLK	CLK	Usually single kinase domain; Sometimes PHD and SPK domains are also tethered	The "R" HRD motif in activation loop is replaced by "T" or other hydrophobic residues. This cluster consists of sequences with single kinase domain only.	15273306
31	37	0.13	2	1.53	0.13	1.50	0.92	0.85	0.94	CAMKL	CAMKL,TSSK	TSSK has single kinase domain; In CAMKL, Pk domain tethered to UBA domain and long C-ter overhangs	Both subfamilies belong to CAMK group and are regulated by Ca/Calmodulin binding. Few kinases have EF had domains and KA1 domains at the C-ter. Has variable C-ter domain unassigned regions. Functions in proteins involved in Ubiquitination pathway; associates with EF-Tu	10413400
32	12	0.00	1	1.90	0.13	1.91	0.92	0.41	0.93	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
33	23	0.31	4	1.67	0.72	1.54	0.59	0.27	0.67	CDK,NDR,RAF	CDK(6),DYRK(5),NDR(6),RAF(6)	High diversity in domain organization and group level function. CDK, DYRK are single domain.	NDR is tethered to UBA and Alpha2-MRAP_N domain. RAF is tethered to C1, RBD domains.	
34	6	0.00	1	2.45	0.71	1.87	1.00	1.00	1.00	MAPK	MAPK	Only Pk domain	Characteristic T-X-Y motif in activation loop for regulation of activity; docking groove and ED	12639708

													motif adjacent to it help in recognition of specific interacting proteins.	
35	12	0.00	1	1.57	0.13	1.41	0.73	0.45	0.79	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases;	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
36	**16**	**0.00**	**1**	**2.32**	**0.54**	**1.84**	**1.00**	**1.00**	**1.00**	**MAPK**	**MAPK**	**Only Pk domain**	**Characteristic T-X-Y motif in activation loop for regulation of activity; docking groove and ED motif adjacent to it help in recognition of specific interacting proteins.**	**12639708**
37	17	0.00	1	3.01	1.65	2.59	0.92	0.40	0.93	STKR	STKR	Protein kinase, Activin_recp, TGF_beta_GS	Ser/Thr kinase that is a receptor kinase	8909794
38	7	0.00	1	1.78	0.38	1.67	0.86	0.71	0.86	RSK	RSK	Has two tandem kinase domains	Has two tandem kinase domains; C-ter kinase domain acts as a regulatory domain; S232, S372, S389, and T581 are crucial for activation	15632195
39	9	0.00	1	2.07	0.73	2.44	0.81	0.68	0.88	PKC	PKC	Multi-domain protein kinase; Has C1,C2 and Pk domain	C1 is involved in binding to diacyl glycerol; C2 is Calcium sensor	10764742
40	11	0.00	1	1.82	0.38	1.37	0.94	0.39	0.95	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases;	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
41	**8**	**0.00**	**1**	**1.97**	**0.35**	**1.44**	**1.00**	**1.00**	**1.00**	**MAPKAPK**	**MAPKAPK**	**Single kinase domain**	**Short overhangs on either side of kinase domain**	
42	8	0.00	1	1.59	0.26	1.41	0.71	0.43	0.78	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases;	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
43	**5**	**0.00**	**1**	**1.98**	**1.17**	**1.82**	**1.00**	**1.00**	**1.00**	**CDK**	**CDK**	**Single domain Kinases**	**PSTAIRE helix motif in the Kinase domain is important for interaction with cyclin; WP insert at position 227, T160 are required for activation.**	**8756328, 10559988**
44	**6**	**0.00**	**1**	**2.78**	**1.30**	**2.01**	**1.00**	**1.00**	**1.00**	**STE7**	**STE7**	**Mainly single kinase domain,**	**Dual specific kinase and phosphorylates S/T/Y**	**1628831**

											sometimes associated with PB1	**residues.**	
45	10	0.00	1	1.78	0.44	1.48	0.71	0.43	0.78	CK1	CK1	Single Pk domain	Characteristic replacement of APE motif in activation loop to SIN	7768349
46	14	0.00	1	1.57	0.45	1.74	1.00	1.00	1.00	CAMK1	CAMK1	Single kinase domain	Characteristic autoinhibitory and calmodulin binding region at the C terminusof kinase domain	10413400
47	7	0.00	1	1.67	0.86	1.62	0.71	0.46	0.77	NDR	NDR	Mainly single domain kinase	Auto inhibitory sequence and an N-ter regulatory SMA domain important for binding to MOB proteins	16607288, 20027184
48	16	0.00	1	1.64	0.02	1.60	0.94	0.88	0.95	CDK	CDK	Single domain Kinases	PSTAIRE helix motif in the Kinase domain is important for interaction with cyclin; WP insert at position 227, T160 are required for activation.	8756328, 10559988
49	6	0.00	1	1.91	0.16	1.51	0.83	0.27	0.87	STE11	STE11	Single kinase domain. Also associated with PB1, zf-rbx1 and LOH1CR1	It is a MAP3K; This cluster contains sequences tethered to PB1 domain with a linker region; PB1 mediates heterdimerization with other proteins	17452462
50	6	0.00	1	3.13	0.64	1.86	0.82	0.53	0.82	RSK	RSK	Has two tandem kinase domains	Has two tandem kinase domains; C-ter kinase domain acts as a regulatory domain; S232, S372, S389, and T581 are crucial for activation	15632195
51	9	0.00	1	2.25	2.08	2.22	0.61	0.26	0.65	Trk	Trk	Pkinase, LRR, Lrrct1, Ig		
52	**11**	**0.00**	**1**	**1.83**	**0.49**	**1.86**	**1.00**	**0.77**	**1.00**	**Akt**	**Akt**	**PH, Pkinase**	**Regulated by binding to phosphoinosites which binds at PH domain**	**16540465**
53	6	0.00	1	2.79	0.49	1.61	0.60	0.27	0.69	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases;	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
54	**7**	**0.00**	**1**	**1.86**	**0.00**	**1.57**	**1.00**	**1.00**	**1.00**	**STE11**	**STE11**	**Single kinase domain. Also associated with**	**It is a MAP3K; This cluster contains sequences tethered to PB1 domain with a linker**	**17452462**

												PB1, zf-rbx1 and LOH1CR1	**region; PB1 mediates heterdimerization with other proteins**	
55	27	0.00	1	2.45	1.62	2.09	0.60	0.33	0.67	Eph	Eph	Usually associated with 5 domains apart from Pkinase Ephrin_lbd,, fn3, SAM, GCC2_GCC3, Interfer-bind	Required for binding of Ephrin ligand and activation of the tyrosine kinase domain	12094214
56	11	0.00	1	1.55	0.17	1.84	0.81	0.62	0.79	MAST	MAST	Pkinase, DUF1908 and PDZ	PDZ domain helps in interaction with dystrophin network and microtubule	10404183
57	8	0.00	1	2.35	0.21	1.45	0.88	0.75	0.90	PIM	PIM	Single domain kinases	Characteristic ERPXPX sequence in hinge region, 2 extra B strand regions in N terminal lobe of kinase required for constitutive activation	15525646
58	24	0.00	1	1.87	0.62	1.67	0.76	0.45	0.81	Src	Src	Pkinase,Sh3,SH2	SH# and SH2 required for substrate recruitment, localization, regulation of activity.	15489908
59	**6**	**0.00**	**1**	**4.00**	**1.27**	**3.22**	**1.00**	**0.47**	**1.00**	**STKR**	**STKR**	**Protein kinase, Activin_recp, TGF_beta_GS**	**Ser/Thr kinase that is a receptor kinase**	**8909794**
60	11	0.00	1	2.04	0.85	1.56	0.91	0.82	0.93	DYRK	DYRK	Single domain kinase	Dual specificity Kinase; activation loop contains Y-X-Y motif and HCD instead of HRD; Autophosphorylation of Tyr residues required for localization to	9748265, 9932450
61	8	0.00	1	2.19	1.15	2.03	0.84	0.43	0.86	RAF	RAF	Pkinase, C1 and RBD	Required for binding to Ras and activation	16508002
62	**7**	**0.00**	**1**	**1.68**	**0.07**	**1.53**	**1.00**	**1.00**	**1.00**	**CAMKL**	**CAMKL**	**In CAMKL, Pk domain tethered to UBA domain and long C-ter overhangs**	**Few kinases have EF had domains and KA1 domains at the C-ter. Has variable C-ter domain unassigned regions. Functions in proteins involved in Ubiquitination pathway; associates with EF-Tu**	**10413400**
63	**8**	**0.00**	**1**	**1.91**	**0.32**	**1.52**	**1.00**	**1.00**	**1.00**	**STE20**	**STE20**	**Mostly multi-domain kinases;**	**High diversity of the Kinase domain itself; Mostly function**	**10837245**

												associated with DUF 1241, CNH domains	**in the MAPK pathway by activating MAP3K**	
64	13	0.00	1	2.84	1.61	2.25	0.51	0.19	0.56	FGFR	FGFR	Associated with different Ig family domains like Ig, Vset, Iset, Herpe_GE	Required to interact with growth factors and proteoglycans	14732692
65	10	0.00	1	3.04	0.13	2.06	0.63	0.24	0.71	MLK	MLK	Pkinase, Ank repeats and SH3	The cluster contains SH3 domain containing sequences. Contains leucine zipper which mediates dimerization and activation by autophosphorylation.	15610029
66	11	0.00	1	2.30	1.76	2.04	0.59	0.17	0.59	JakA	JakA	FERM, 2kinase domains, SH2	Requird for binding to cytokine receptors and activation	12039028
67	15	0.00	1	1.72	0.34	1.50	0.78	0.64	0.83	DAPK	DAPK	Single kinase domain. Rarely tethered to death domain and Ank repeats	Activated by Ca^{2+}/Calmodulin binding at Cterminus if kinase domain. Other domains required for interactions.	11579085, 12730201
68	12	0.00	1	2.17	0.77	1.97	0.75	0.36	0.79	Tec	Tec	Pkinase, PH, Sh3, SH2 and BTK.	Required for localization	7503742
69	7	0.00	1	1.47	0.35	1.39	0.86	0.48	0.89	STE20	STE20	Diverse domain combinations; Mostly multi-domain kinases;	High diversity of the Kinase domain itself; Mostly function in the MAPK pathway by activating MAP3K	10837245
70	27	0.15	2	2.68	1.65	2.06	0.49	0.14	0.55	PDGFR	PDGFR,VEGFR	Pkinase, Ig like domains	Specific inserts in kinase domain where phosphorylation takes place required for activation.	2550144
71	**8**	**0.00**	**1**	**3.12**	**0.83**	**2.73**	**1.00**	**1.00**	**1.00**	**Trbl**	**Trbl**	**Single kinase domain**	**High Ser and Pro content N-ter to kinase domain, G-S-P-S-P-P motif 60 residues N-ter of kinase domain, DLKLRKF motif in sub-domain IV B of kinase domain required for function of the kinase and nuclear localization.**	**16963228**

72	7	0.00	1	2.61	0.92	2.06	0.57	0.24	0.73	PKC	PKC	Multi-domain protein kinase; Has C1,C2 and Pk domain	C1 is involved in binding to diacyl glycerol; C2 is Calcium sensor	10764742
20	10	0.00	1	2.56	1.79	2.66	0.95	0.61	0.82	InsR	InsR	Pkinase, fn3, furin-like and Recep-L-domain	Required for binding of ligand	19459609
74	15	0.00	1	2.62	1.77	2.17	0.74	0.46	0.76	Eph	Eph	Usually associated with 5 domains apart from Pkinase Ephrin_lbd,, fn3, SAM, GCC2_GCC3, Interfer-bind	Required for binding of Ephrin ligand and activation of the tyrosine kinase domain	12094214
75	**6**	**0.00**	**1**	**2.00**	**0.39**	**1.45**	**1.00**	**1.00**	**1.00**	**MAPKAPK**	**MAPKAPK**	**Single kinase domain**	**Short overhangs on either side of kinase domain**	
76	6	0.00	1	3.35	1.32	2.57	1.00	0.47	1.00	STKR	STKR	Protein kinase, Activin_recp, TGF_beta_GS	Ser/Thr kinase that is a receptor kinase	8909794
77	12	0.15	2	1.96	0.85	1.81	0.66	0.24	0.70	Csk,Syk	Csk(6), Syk(6)	Pkinase, SH2 in both and additional SH3 only in Csk	Both kinases belogning to TK and both contain SH2 domain required for binding to ligands an localization	15489908
78	5	0.00	1	2.06	0.72	2.13	0.88	0.40	0.89	Abl	Abl	Pkinase, SH3, SH2 and Factin_bind domains	This is the cluster has sequences with with F-actin-bind domain. Required for interaction	9144171
79	12	0.00	1	1.67	0.47	1.58	0.76	0.51	0.81	CAMKL	CAMKL	In CAMKL, Pk domain tethered to UBA domain and long C-ter overhangs	Few kinases have EF had domains and KA1 domains at the C-ter. Has variable C-ter domain unassigned regions. Functions in proteins involved in Ubiquitination pathway; associates with EF-Tu	10413400
80	**8**	**0.00**	**1**	**1.77**	**0.00**	**1.61**	**1.00**	**1.00**	**1.00**	**PIM**	**PIM**	**Single domain kinases**	**Characteristic ERPXPX sequence in hinge region, 2 extra B strand regions in N terminal lobe of kinase required for constitutive activation**	**15525646**

81	17	0.00	1	2.19	0.81	1.79	0.42	0.16	0.58	DMPK	DMPK	Diverse domain combinations with Pk and DMPK	DMPK is required for dimerization and forms a coiled coil structure	12832055
82	7	0.00	1	1.96	0.58	1.46	0.89	0.31	0.77	DCAMKL	DCAMKL	Either single kinase domain or associated with DCX domain	This is cluster associated with DCX domain required for localization.	10441322
83	**9**	**0.00**	**1**	**1.68**	**0.17**	**1.53**	**1.00**	**1.00**	**1.00**	**SRPK**	**SRPK**	**Single domain kinases**	**Deviations from classical CMGC kinases at specific positions like CMGC Glu, DFG motif in activation segment is changed to DLG and this is followed by an Asn**	**15273306**
84	9	0.00	1	2.02	0.81	2.35	0.76	0.41	0.73	PKD	PKD	Pkinase, C1, PH	Required for localization	16540465
85	6	0.00	1	3.35	0.10	2.94	0.83	0.67	0.87	MLK	MLK	Pkinase, SH3, SAM, Ank repeats, IFT57	This cluster consists of singe kinase domain sequences and IFT57 tethered cases. Contains leucine zipper which mediates dimerization and activation by autophosphorylation.	15610029
86	**5**	**0.00**	**1**	**1.52**	**0.15**	**1.57**	**1.00**	**1.00**	**1.00**	**CDKL**	**CDKL**	**Single domain kinases**	**Large C terminal overhangs, T-X-Y motif in activation loop like in MAPK but no reports of its autophosphorylation. It also has a conserved Y in the Gly loop.**	**15273306**
87	6	0.00	1	1.59	0.65	2.40	0.92	0.38	0.86	LISK	LISK	Mostly single pkinase domain, also associated with LIM and PDZ domains	The cluster contains the associated domains required for localization and interaction.	9382826, 10704826
88	7	0.00	1	2.76	1.02	2.12	0.76	0.26	0.80	DDR	DDR	Pkinase, Ig like domains, F5-F8_type C domain	F5_F8_type_C mediates collagen-binding	18697744
89	6	0.00	1	2.33	0.30	2.06	0.83	0.67	0.87	MAPK	MAPK	Single domain kinases	Characteristic T-X-Y motif in activation loop for regulation of activity; docking groove and ED motif adjacent to it help in recognition of specific interacting proteins.	12639708

90	6	0.00	1	1.95	0.80	1.65	0.83	0.67	0.87	CDK	CDK	Single domain kinases	PSTAIRE helix motif in the Kinase domain is important for interaction with cyclin; WP insert at position 227, T160 are required for activation.	8756328, 10559988
91	**6**	**0.00**	**1**	**4.19**	**6.18**	**2.83**	**1.00**	**1.00**	**1.00**	**PHK**	**PHK**	**Single domain kinases**	**Present as hexadecamer of 4 homotetramers where only one is catalytic and others are regulatory in nature. The alpha and beta subunits have a specific phosphorylation sites at 729,735, 1015. The delta subunit has a characteristic Cterminal region for binding calmodulin**	**9362479**
92	6	0.00	1	1.59	0.00	1.52	0.83	0.67	0.87	YANK	YANK	Single domain kinases	Unknown Novel Kinase	---

The bold rows show 100% pure clusters with conserved domain-architecture.

Table S3: Overlap between the LMS based clusters and Hanks and Hunter sub-families of protein kinases

Domain Architecture conserved **(HH)**				
Single domain kinase sub-families				
CAMK1(1)	**CLK(2)**	**PIM(1)**	**YANK(2)**	**Worm9(1)**
CAMK2(1)	**DYRK(5)**	**PKA(3)**	**TTBK(1)**	**Worm10(1)**
CDK(7)	**GSK(3)**	**SGK(1)**	**VRK(1)**	**KIN6(1)**
CDKL(2)	**MAPK(6)**	**STE7(3)**	**Worm6(1)**	**Lmr(1)**
CK1(2)	**MAPKAPK(4)**	**Trbl(3)**	**Worm7(1)**	**PSK(1)**
CK2(1)	**PHK(1)**	**TTBKL(2)**	**Worm8(1)**	**RSKR(1)**
SRPK(1)	**TK-Sp1(1)**			
Multi-domain kinase sub-families (HH)				
AKT(4)	**PKC(5)**	**Syk(3)**	**CASK(1)**	**PDK1(1)**
CAMKL(7)	**PKG(1)**	**Tec(1)**	**Csk(1)**	**PKD(1)**
DCAMKL(2)	**PKN(1)**	**Trk(1)**	**EGFR(1)**	**RAD53(1)**
DDR(3)	**RSK(5)**	**VEGFR(1)**	**FAK(1)**	**Ror(1)**
Dual(2)	**Src(3)**	**Axl(1)**	**IRAK(1)**	**Ryk(1)**
Tie(1)	**CCK4(1)**	**GRK(3)**		
Domain architecture not conserved **(HH)**				
Abl(2)	**MAST(2)**	**STKR(5)**	JAkA(2)	**LRRK(1)**
DAPK(2)	**NDR(3)**	ALK(2)	MLCK(3)	**Met(1)**
Fer(3)	**PDGFR(2)**	CCK(2)	Trio(2)	**Musk(1)**
InsR(2)	**RAF(3)**	DMPK(2)	TSSK(2)	**RET(1)**
LISK(2)	**STE11(4)**	Eph(3)	**Ack(1)**	**RIPK(1)**
MLK(3)	**STE20(9)**	FGFR(2)	**KIN16(1)**	**RSKL(1)**
Sev(1)				

1. Sub-families shown in bold for pure clusters in LMS based clustering
2. The number in the parenthesis shows the number of LMS clusters into which the sub-family related proteins are divided
3. The Shaded boxes represent sub-families which cluster with other sub-families, but all the members belong to the same clusters

Table S4: Classification of Ig2 domain (PF 13895) containing proteins (*n*=455). Clustering of Ig2 containing proteins into 96 clusters using LMS at full-length gene products at a tree parsing cut-off of 0.2. The representative member from the cluster and their domain architecture are presented. Note that the pure clusters (having single domain architecture) also have a high GO similarity values within the cluster at MF level.

Cluster No.	Cluster size	C_u	Rep member	PFam architecture id	GO similarity			Architecture similarity			Representative domain architecture
					CC	BP	MF	JK	GK	DD	
1	4	3	P04217	138208	3.03	Nan	nan	0.75	0.67	0.81	Ig_2~Ig_2~Ig_2~Ig_2~Ig_2
2	64	35	O18906	117061	0.86	1.46	0.89	0.41	0.13	0.51	V-set~Ig_2
3	4	3	P18572	132341	1.92	3.30	4.39	1.00	0.32	0.75	Ig_2~I-set
4	4	2	P50895	170325	1.62	3.88	4.73	1.00	0.92	0.86	Ig_2~C2-set_2~Ig_2~Ig_2~Ig_3
5	2	2	Q6AZB0	197982	2.62	4.81	1.20	1.00	0.71	0.78	Ig_2~I-set~Ig_2~I-set~fn3~fn3~fn3
6	**2**	**1**	**Q8R5M8**	**127544**	**2.54**	**4.17**	**3.68**	**1.00**	**1.00**	**1.00**	**V-set~C2-set_2~Ig_2**
7	**3**	**1**	**Q1WIM3**	**127544**	**1.90**	**3.31**	**3.92**	**1.00**	**1.00**	**1.00**	**V-set~C2-set_2~Ig_2**
8	**3**	**1**	**Q1WIM1**	**127544**	**0.93**	**2.70**	**nan**	**1.00**	**1.00**	**1.00**	**V-set~C2-set_2~Ig_2**
9	7	4	O35112	143191	1.42	2.77	1.20	0.70	0.56	0.77	C2-set_2~Ig_2~Ig_2~Ig_2
10	3	3	Q9N1E6	145890	1.30	2.31	2.75	0.61	0.22	0.54	C2-set_2~Ig_2
11	**3**	**1**	**O35158**	**161167**	**1.09**	**4.35**	**1.20**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_2~I-set~I-set~I-set~fn3~fn3~fn3**
12	6	4	P31997	146398	1.44	0.96	1.20	0.68	0.37	0.69	V-set~Ig_3~Ig_2
13	3	2	P16573	146385	0.99	2.12	0.46	0.78	0.67	0.78	V-set~Ig_2~Ig_2~ig
14	2	2	P70232	220661	2.02	2.37	nan	1.00	1.00	0.90	I-set~I-set~I-set~I-set~Ig_2~fn3~fn3~fn3~Bravo_FIGEY
15	**3**	**1**	**Q8K1G0**	**117061**	**2.17**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**V-set~Ig_2**
16	**5**	**1**	**P12960**	**149331**	**1.95**	**3.18**	**4.52**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_2~I-set~I-set~I-set~Ig_2~fn3~fn3~fn3**
17	4	2	P22063	141064	2.92	4.01	2.86	1.00	0.86	0.91	Ig_2~Ig_3~I-set~I-set~I-set~Ig_2~fn3~fn3~fn3
18	8	3	P97528	220443	1.97	2.28	nan	1.00	0.87	0.86	I-set~I-set~I-set~I-set~I-set~I-set~Ig_2~fn3~fn3
19	5	2	O94779	150331	1.88	2.43	nan	1.00	1.00	0.96	I-set~Ig_2~I-set~I-set~I-set~I-set~fn3~fn3~fn3~fn3
20	16	12	Q08156	165086	1.14	2.25	1.83	0.69	0.39	0.64	Ig_2~Ig_2~I-set~Ig_3~Pkinase_Tyr
21	**5**	**1**	**P78310**	**117061**	**3.52**	**5.10**	**3.82**	**1.00**	**1.00**	**1.00**	**V-set~Ig_2**
22	**2**	**1**	**P43146**	**152881**	**1.07**	**2.11**	**1.19**	**1.00**	**1.00**	**1.00**	**Ig_2~I-set~I-set~I-set~fn3~fn3~fn3~fn3~fn3~fn3~Neogenin_C**
23	5	2	O60469	155022	1.42	4.32	1.65	0.90	0.87	0.95	Ig_3~Ig_2~I-set~I-set~I-set~I-set~I-set~I-set~I-set~fn3~fn3~fn3~fn3~I-set~fn3~fn3
24	**3**	**1**	**O88775**	**132341**	**0.93**	**2.70**	**nan**	**1.00**	**1.00**	**1.00**	**Ig_2~I-set**
25	**4**	**1**	**P12318**	**124672**	**1.07**	**2.61**	**2.91**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_2**
26	**5**	**1**	**A3RFZ7**	**124672**	**1.06**	**3.27**	**3.34**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_2**

27	3	3	P12314	157668	1.94	5.35	3.28	0.67	0.44	0.81	Ig~Ig_2~Ig_2
28	**3**	**1**	**P08508**	**124672**	**1.09**	**nan**	**3.12**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_2**
29	10	8	Q6DN72	140781	0.94	Nan	0.96	0.78	0.45	0.62	Ig_2~Ig_2~Ig_3
30	**4**	**1**	**P18461**	**146464**	**2.18**	**2.89**	**2.18**	**1.00**	**1.00**	**1.00**	**Ig_2~I-set~I-set~Pkinase_Tyr**
31	3	2	Q62786	146934	3.60	6.42	1.20	1.00	1.00	0.81	V-set~Ig_2~V-set~V-set~Ig_2
32	**2**	**1**	**Q61450**	**125874**	**0.84**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_3**
33	3	2	A6QQC6	157765	3.54	Nan	nan	0.78	0.50	0.74	V-set~I-set~Ig_2
34	7	7	Q8NDA2	256492	2.56	1.46	2.06	0.34	0.22	0.51	VWA_2~Ig_2~I-set~I-set~I-set~I-set~I-set~I-set~Ig_2~I-set~Ig_2~I-set~I-set~I-set~I-set~I-set~I-set~I-set~I-set~I-set~I-set~I-set~I-set~G2F~EGF_CA~EGF_CA~EGF_CA~cEGF
35	5	4	P13597	142128	0.91	2.48	nan	0.80	0.58	0.69	ICAM_N~Ig_2~Ig_2~Ig_2
36	4	2	Q28806	142128	1.22	2.76	3.22	1.00	1.00	0.91	ICAM_N~Ig_2~Ig_2~Ig_2
37	**2**	**1**	**Q8TDY8**	**196056**	**1.14**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**Ig_2~I-set~Ig_2~I-set~fn3~fn3~fn3~fn3~fn3**
38	**2**	**1**	**P01876**	**132886**	**Nan**	**6.03**	**6.18**	**1.00**	**1.00**	**1.00**	**C1-set~Ig_2~C1-set**
39	**2**	**1**	**A6NGN9**	**135694**	**3.03**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**V-set~Ig_2~Ig_2**
40	**2**	**1**	**Q7TNR6**	**135694**	**3.03**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**V-set~Ig_2~Ig_2**
41	3	3	Q7TQA1	224370	1.63	Nan	1.88	1.00	1.00	0.85	Ig_2~Ig_2~Ig_2~Ig_3~Ig_2~Ig_2~Ig_2~Ig_3~Ig_2~Ig_3~Ig_2
42	5	4	O75054	137641	0.89	Nan	nan	0.75	0.58	0.75	V-set~V-set~V-set~V-set~Ig_2~V-set~V-set
43	**2**	**1**	**Q5VJ70**	**123060**	**2.31**	**3.50**	**4.99**	**1.00**	**1.00**	**1.00**	**I-set~Ig_2**
44	12	4	B3MKS0	134883	0.93	4.92	4.54	0.92	0.78	0.83	Ig_2~Ig_2~I-set~I-set~fn3~fn3
45	4	2	P59822	133590	1.32	2.77	2.03	1.00	1.00	0.89	Ig_2~Ig_2~TIR
46	3	2	P13504	161174	1.74	5.05	3.32	0.83	0.62	0.78	I-set~Ig_2~Ig_3~TIR
47	3	2	P27931	140781	1.58	Nan	3.49	0.56	0.56	0.74	Ig_2~Ig_2~Ig_3
48	3	2	P14719	158971	2.30	6.02	3.97	0.83	0.62	0.78	Ig_2~Ig_2~Ig_3~TIR
49	3	2	Q62929	133590	1.32	3.29	2.91	0.78	0.67	0.85	Ig_2~Ig_2~TIR
50	6	4	P59823	180968	1.42	2.71	1.87	0.73	0.50	0.71	Ig_2~Ig_3~TIR
51	4	2	Q9JHY1	117061	2.22	3.44	nan	0.67	0.50	0.76	V-set~Ig_2
52	**2**	**1**	**Q68FQ2**	**117061**	**3.74**	**5.70**	**4.29**	**1.00**	**1.00**	**1.00**	**V-set~Ig_2**
53	5	5	A8MWS1	161735	0.67	Nan	1.39	0.55	0.13	0.53	Ig_2~ig~ig
54	**3**	**1**	**Q6X936**	**140845**	**3.31**	**5.21**	**1.20**	**1.00**	**1.00**	**1.00**	**I-set~C2-set_2~Ig_2~Ig_2~Ig_3**
55	2	2	Q6UWL6	197334	0.87	2.12	nan	0.75	0.57	0.82	I-set~C2-set_2~Ig_2~Ig_2
56	**2**	**1**	**Q8BR86**	**161156**	**2.75**	**2.88**	**1.20**	**1.00**	**1.00**	**1.00**	**I-set~C2-set_2~Ig_2~I-set~I-set**

57	6	2	O97799	141521	2.57	5.38	2.84	0.79	0.64	0.82	Ig_2~I-set~Ig_3~Pkinase_Tyr
58	**3**	**1**	**P11627**	**136140**	**1.70**	**2.77**	**5.22**	**1.00**	**1.00**	**1.00**	**Ig_2~Ig_2~I-set~I-set~I-set~I-set~fn3~fn3~fn3~fn3~Bravo_FIGEY**
59	12	4	O75019	159298	0.85	2.72	1.33	1.00	0.77	0.88	Ig_2~Ig_3~Ig_2~Ig_3
60	4	2	Q0PMG2	191424	2.94	3.19	nan	1.00	1.00	0.85	I-set~Ig_2~Ig_2~I-set~I-set~I-set~MAM
61	**3**	**1**	**P60755**	**139092**	**1.95**	**3.47**	**nan**	**1.00**	**1.00**	**1.00**	**I-set~Ig_2~Ig_2~I-set~I-set~Ig_2~MAM**
62	3	2	P57097	133966	3.16	4.61	1.77	0.83	0.62	0.78	I-set~Ig_2~fn3~Pkinase_Tyr
63	**4**	**1**	**O75900**	**142322**	**3.45**	**4.77**	**5.34**	**1.00**	**1.00**	**1.00**	**Peptidase_M10~ShK~Ig_2**
64	2	2	Q9EPF2	222744	1.64	5.70	nan	0.80	0.55	0.72	V-set~C2-set_2~ig~Ig_2~I-set
65	4	2	O76036	125874	1.10	Nan	1.22	0.75	0.50	0.76	Ig_2~Ig_3
66	4	2	P97798	152881	1.12	2.22	nan	1.00	0.94	0.96	Ig_2~I-set~I-set~I-set~fn3~fn3~fn3~fn3~fn3~fn3~Neogenin_C
67	3	2	P97685	136140	3.77	5.80	nan	1.00	1.00	0.95	Ig_2~Ig_2~I-set~I-set~I-set~I-set~fn3~fn3~fn3~fn3~Bravo_FIGEY
68	4	3	P32507	127544	1.96	3.51	1.94	0.75	0.46	0.62	V-set~C2-set_2~Ig_2
69	**3**	**1**	**P97300**	**131527**	**2.54**	**5.36**	**5.61**	**1.00**	**1.00**	**1.00**	**I-set~Ig_2~I-set**
70	4	3	P97686	158237	2.02	4.33	4.62	1.00	0.90	0.88	Ig_2~Ig_3~I-set~Ig_2~I-set~I-set~fn3~fn3~fn3~fn3~Bravo_FIGEY
71	5	4	P35739	137247	2.40	3.71	2.52	0.67	0.42	0.70	LRR_8~Ig_2~Ig_2~Pkinase_Tyr
72	3	3	Q95242	177384	1.29	2.03	-nan	0.67	0.56	0.81	Ig_2~Ig_3~Ig_2~Ig_2~Ig_2~Ig_2
73	4	2	P20786	146464	1.66	3.24	2.11	1.00	0.88	0.89	Ig_2~I-set~I-set~Pkinase_Tyr
74	4	4	Q05030	177664	3.52	4.75	2.51	0.75	0.49	0.74	ig~Ig_2~I-set~Ig_2~Pkinase_Tyr
75	9	2	P11464	136434	3.03	5.01	nan	1.00	1.00	0.89	V-set~Ig_2~Ig_2~Ig_2
76	3	3	Q8BKG3	205979	1.15	3.61	1.79	1.00	0.78	0.87	Ig_2~Ig_2~I-set~I-set~I-set~I-set~I-set~I-set~Pkinase_Tyr
77	10	4	B3MH43	195379	1.09	3.52	2.19	1.00	0.91	0.83	I-set~I-set~Ig_2~I-set~I-set~Pkinase_Tyr
78	3	2	P35822	139347	1.04	3.57	3.82	1.00	0.92	0.90	MAM~Ig_2~fn3~fn3~Y_phosphatase~Y_phosphatase
79	**3**	**1**	**Q15223**	**127544**	**1.88**	**2.81**	**1.81**	**1.00**	**1.00**	**1.00**	**V-set~C2-set_2~Ig_2**
80	**2**	**1**	**Q9JLB9**	**127544**	**2.82**	**4.41**	**3.10**	**1.00**	**1.00**	**1.00**	**V-set~C2-set_2~Ig_2**
81	**4**	**1**	**Q15109**	**156889**	**1.13**	**3.78**	**3.57**	**1.00**	**1.00**	**1.00**	**Ig_2~C2-set_2~Ig_2**
82	6	3	Q58EX2	178129	0.88	2.70	nan	0.84	0.71	0.94	Ig_2~I-set~Ig_2~I-set~I-set~I-set~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3~fn3
83	**2**	**1**	**Q9NTN9**	**155285**	**1.09**	**2.04**	**1.39**	**1.00**	**1.00**	**1.00**	**Sema~PSI~Ig_2**
84	**2**	**1**	**O75326**	**155285**	**2.38**	**4.73**	**2.88**	**1.00**	**1.00**	**1.00**	**Sema~PSI~Ig_2**
85	3	3	Q96RL6	240393	0.80	2.70	4.04	0.72	0.43	0.66	V-set~C2-set_2~I-set~ig~Ig_2
86	10	6	Q08ET2	195203	0.96	2.48	3.45	0.66	0.36	0.65	V-set~ig~Ig_2

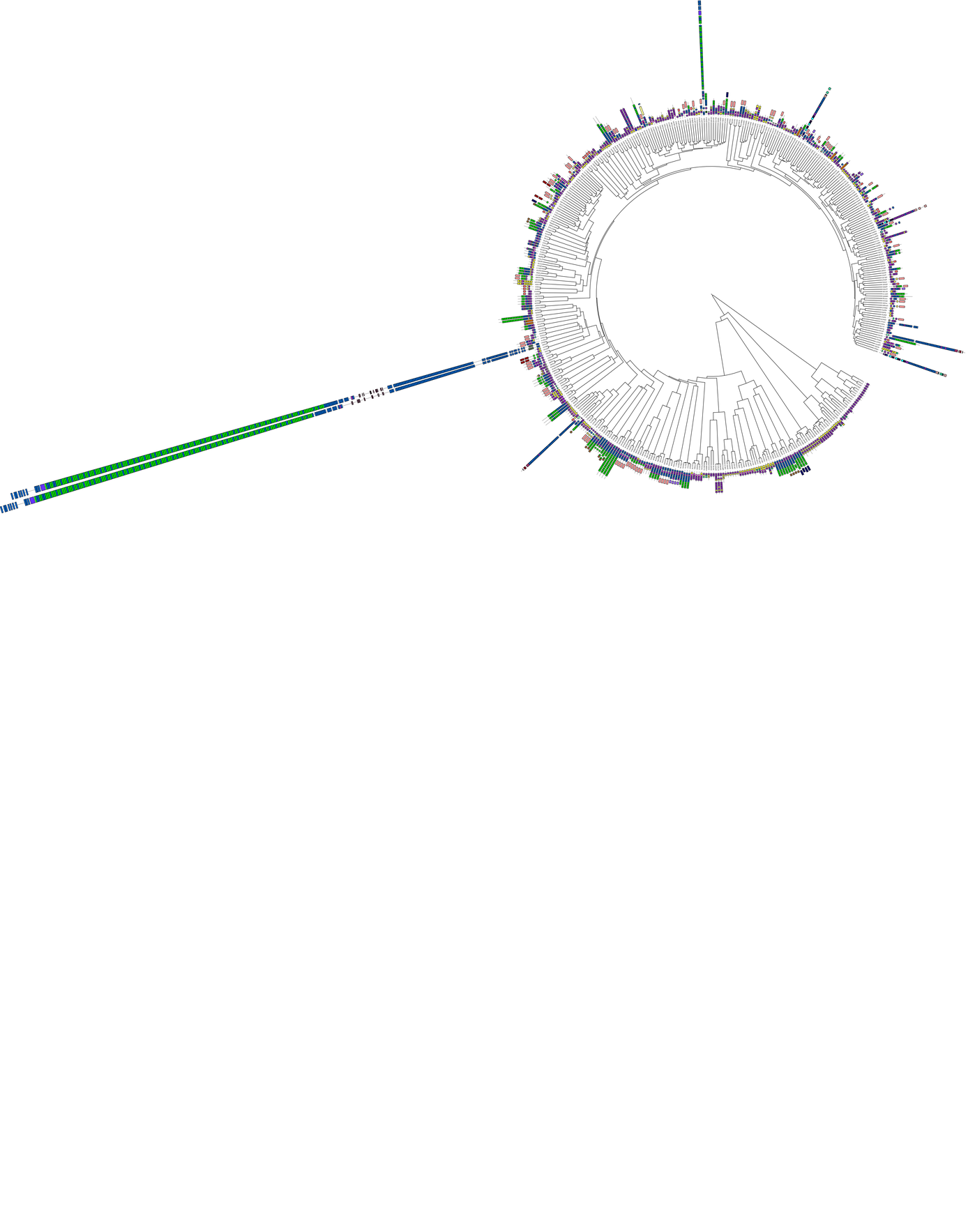
87	**2**	**1**	**Q4V892**	**167299**	**1.31**	**5.91**	**2.02**	**1.00**	**1.00**	**1.00**	**Ig_2~TIR**
88	3	3	A7LCJ3	218008	1.16	2.99	2.83	0.87	0.76	0.89	V-set~C2-set_2~Ig_2~Ig_2~Ig_2~Ig_2~Ig_2~Ig_2~Ig_2~I-set~Ig_2~I-set~Ig_2~Ig_2~Ig_2~Ig_2~Ig_3~Ig_2
89	4	2	P35590	158953	1.46	2.92	1.75	0.73	0.60	0.83	hEGF~Ig_2~fn3~fn3~fn3~Pkinase_Tyr
90	**3**	**1**	**Q3T113**	**132341**	**0.93**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**Ig_2~I-set**
91	5	4	Q98949	133966	1.06	2.12	1.78	0.78	0.55	0.72	I-set~Ig_2~fn3~Pkinase_Tyr
92	4	4	Q28260	212578	2.51	4.64	4.61	0.88	0.67	0.79	I-set~C2-set~I-set~I-set~C2-set~Ig_2~Ig_2
93	4	4	P53767	189175	2.45	3.79	2.09	0.88	0.80	0.85	Ig_2~Ig_2~I-set~I-set~Ig_2~Ig_2~I-set~Pkinase_Tyr
94	3	2	P35918	133851	2.88	4.29	2.11	0.83	0.78	0.83	V-set~I-set~Ig_2~Ig_2~I-set~Pkinase_Tyr
95	3	2	P35916	168294	2.01	4.69	2.40	1.00	1.00	0.91	Ig_2~I-set~Ig_2~Ig_2~Ig_2~I-set~Pkinase_Tyr
96	**3**	**1**	**Q29RR6**	**117061**	**0.93**	**nan**	**nan**	**1.00**	**1.00**	**1.00**	**V-set~Ig_2**

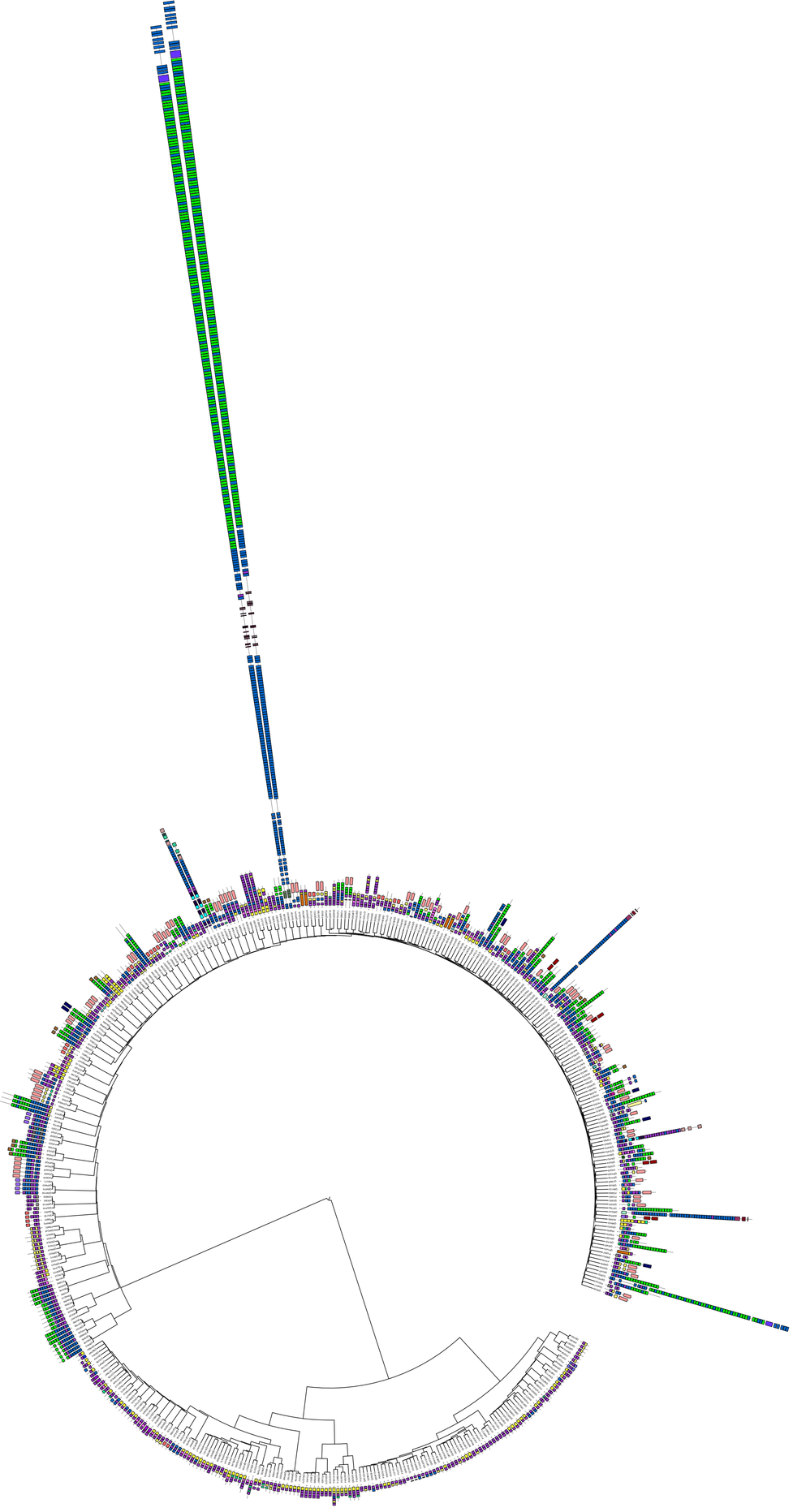
Table S5: Comparison of pair-wise distance scores in protein kinase and immunoglobulin families. All pair-wise values obtained using ClustalW and LMS at the domain (CW_dom and LMS_dom) and full length (CW_full and LMS_full) are compared against each other. A Pearson correlation coefficient value (*r*) depicts the overlap in the pair-wise relationship scores obtained by the two methods at both domain and full length values. Note that the values above 0.50 are represented in bold denoting that there is sufficient correlation. The absolute values of the scores are also compared by comparing the distributions of these scores using paired students-*t* tests.

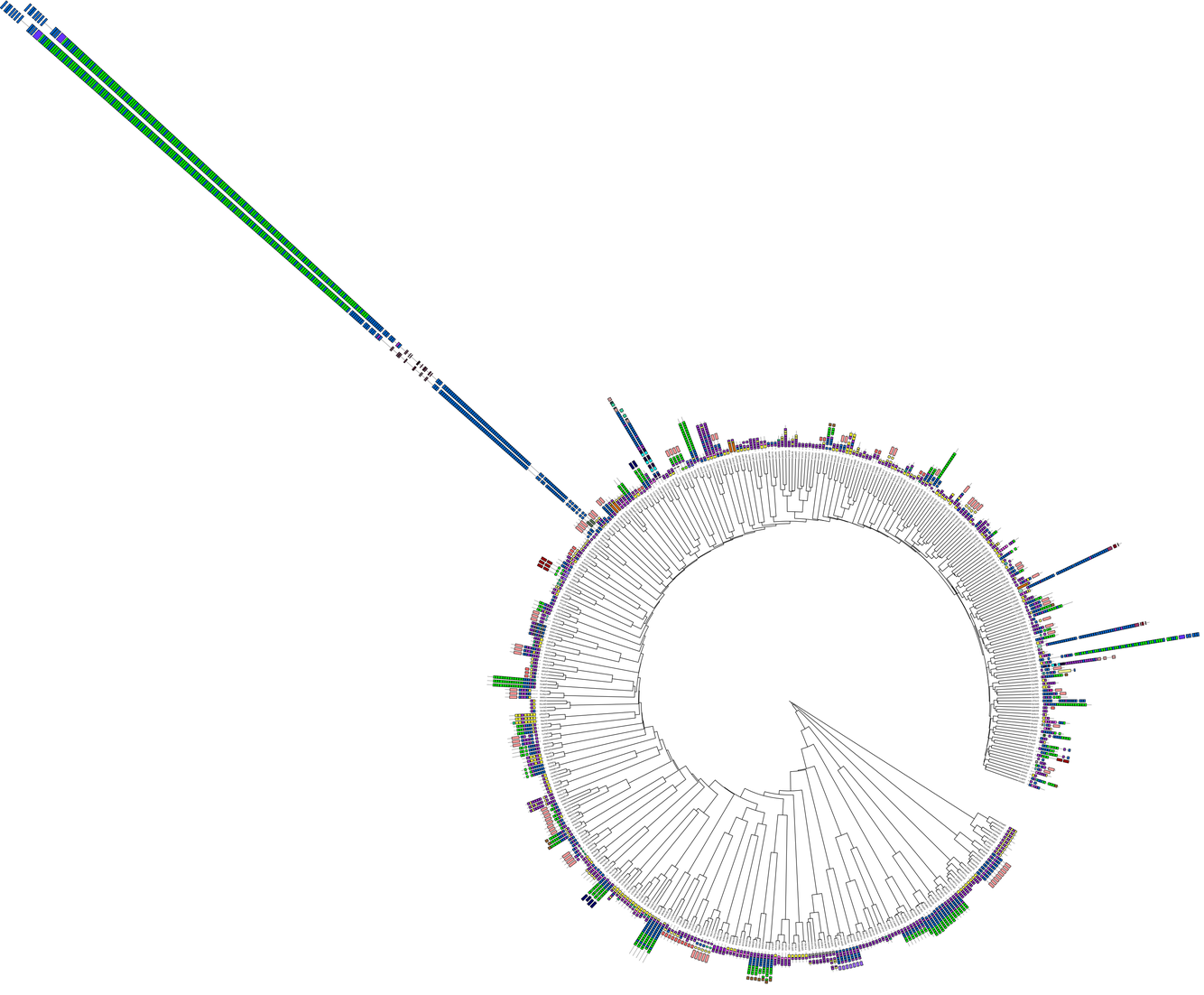
Protein kinases (*n* = 1498)				
Comparisons	**Pearson *r***	***t***	***df***	**p-value**
CW_full vs. CW_dom	0.15	170.22	1121251	< 2.20E-16
CW_full vs. LMS_full	- 0.08	-86.24	1121251	< 2.20E-16
CW_full vs. LMS_dom	- 0.02	-24.86	1121251	< 2.20E-16
CW_dom vs. LMS_full	0.01	14.82	1121251	< 2.20E-16
CW_dom vs. LMS_dom	0.02	21.46	1121251	< 2.20E-16
LMS_full vs. LMS_dom	**0.86**	1821.06	1121251	< 2.20E-16
Ig_2 dataset (*n* = 455)				
Comparisons	**Pearson *r***	***t***	***df***	**p-value**
CW_full vs. CW_dom	**0.59**	271.8	103284	< 2.20E-16
CW_full vs. LMS_full	0.25	80.39	103284	< 2.20E-16
CW_full vs. LMS_dom	**0.50**	118.1	103284	< 2.20E-16
CW_dom vs. LMS_full	0.30*	352.1	103284	< 2.20E-16
CW_dom vs. LMS_dom	**0.76**	153.7	103284	< 2.20E-16
LMS_full vs. LMS_dom	0.23	198.5	103284	< 2.20E-16

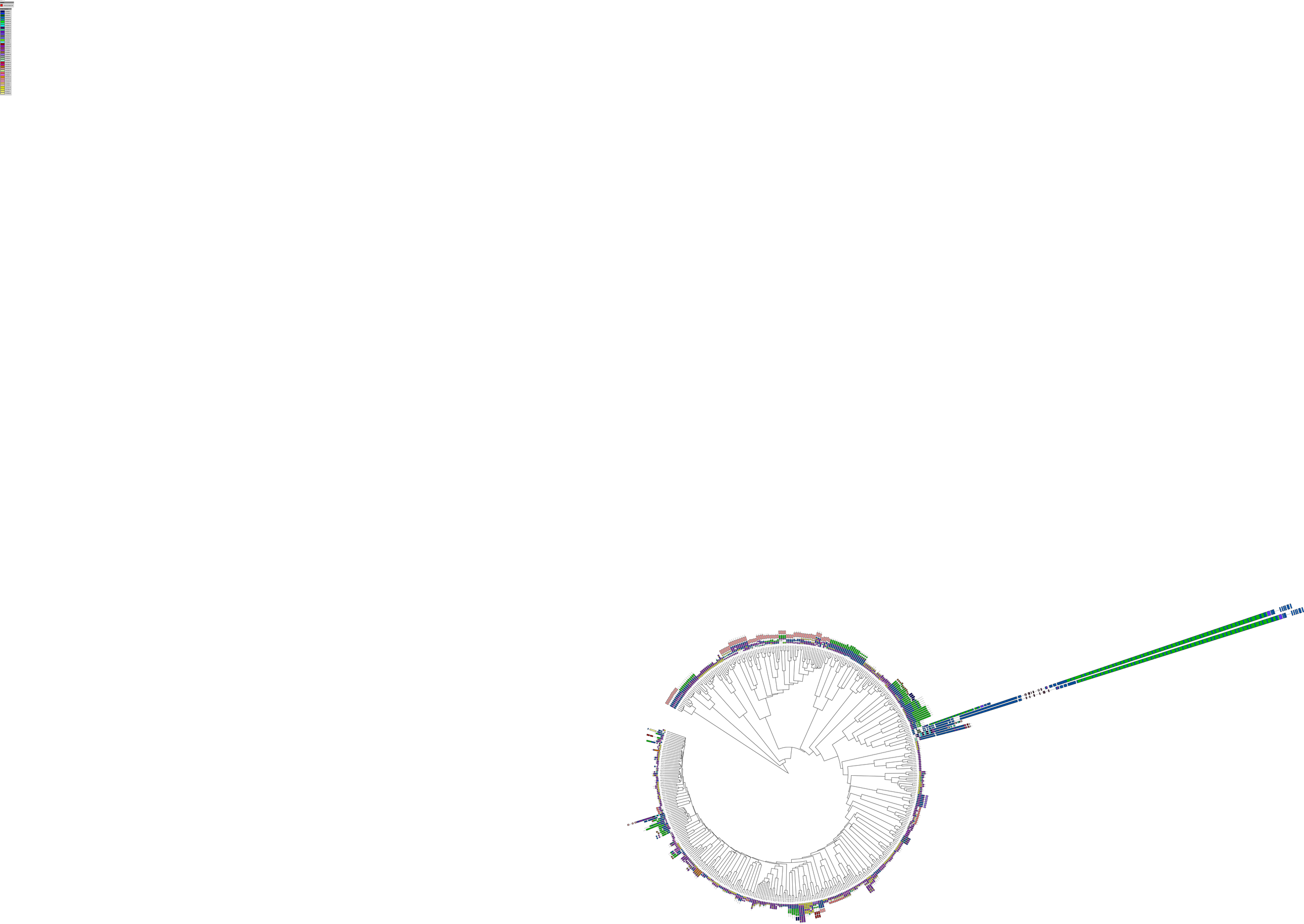
Table S6: Clustering of Protein kinases (*n*=1498) sequences using different methods. The processing time was computed using a personal workstation with a 2.93x8 GHz, Intel i7 processor and 8GB RAM running Ubuntu v12.04. The number of clusters generated at 0.3 threshold and fragment length used in the computations is also shown.

Method	# of Clusters	Threshold	word-length	Time
CW	10	0.3	NA	165m48.269s
LMS	60	0.3	5	23m44.529s
k-tuple	2	0.3	3	0m11.180s
CD-Hit	287	0.3	3	19m3.001s







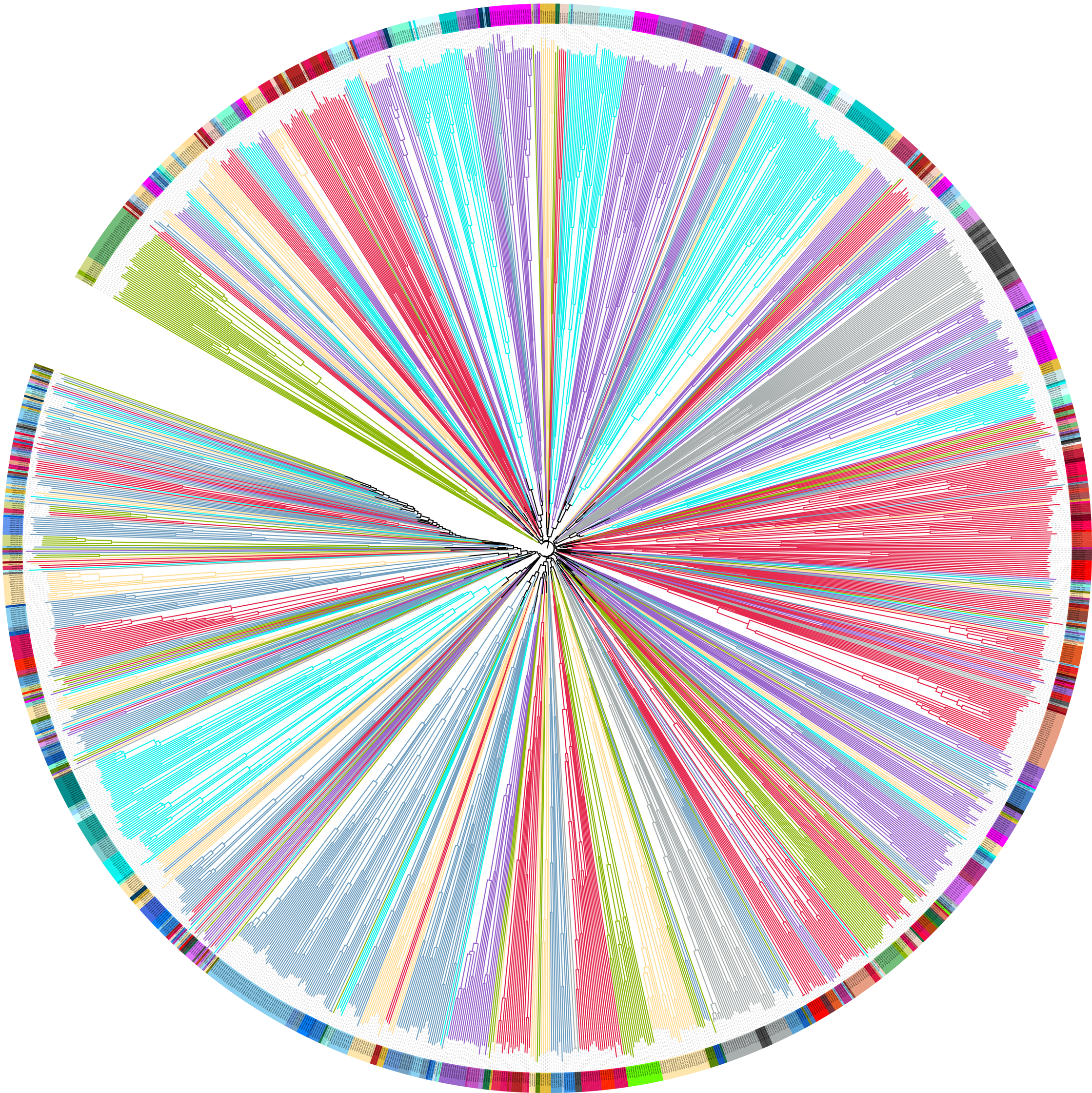


Legend:

- Edges

Color ranges:

- RSK
- PKG
- TSSK
- PKC
- SGK
- Akt
- PKN
- MAPK
- CAMKL
- CDK
- GSK
- DYRK
- RCK
- PKD
- STE11
- Ack
- DMPK
- DAPK
- EGFR
- Ryk
- Syk
- JakA
- MLK
- Worm7
- Fer
- FAK
- Met
- RAD53
- GRK
- STE20
- STE7
- Worm9
- PDK1
- PKA
- NDR
- MAST
- YANK
- CAMK1
- CDKL
- RIPK
- Lmr
- CLK
- Worm6
- AKT
- MAPKAPK
- CK2
- TTBKL
- Worm8
- TTBK
- SRPK
- Eph
- KIN16
- KIN6
- STKR
- Musk
- Ret
- TK-Sp1
- Dual
- VRK
- Axl
- InsR
- ALK
- Tie
- VEGFR
- Ror
- Trk
- PDGFR
- DDR
- FGFR
- PIM
- Worm10
- LRRK
- CK1
- CCK4
- Sev
- CAMK2
- LISK
- MLCK
- Trio
- Tec
- Src
- Csk
- PSK
- Trbl
- IRAK
- RSKR
- RSKL
- DCAMKL
- Abl
- RAF
- CASK
- PHK



Legend:

Color Edges

Color ranges:

- RSK
- PKD
- CCK4
- Axl
- MLCK
- STE11
- NDR
- CDK
- Ack
- PKN
- PKG
- Eph
- RAF
- STE20
- FGFR
- VEGFR
- JakA
- MLK
- Met
- MAST
- Lmr
- Musk
- STKR
- CAMKL
- Trio
- PDGFR
- InsR
- SRPK
- STE7
- KIN16
- DDR
- CASK
- LRRK
- PDK1
- Ror
- FAK
- KIN6
- ALK
- Tie
- DMPK
- Abl
- MAPK
- RSKL
- TTBK
- EGFR
- Sev
- Ret
- DYRK
- Trk
- MAPKAPK
- DAPK
- Fer
- CAMK2
- SGK
- PKC
- Worm6
- Worm9
- PKA
- TTBKL
- CAMK1
- Src
- TSSK
- PHK
- Worm7
- CK1
- CLK
- VRK
- PIM
- Worm8
- Syk
- Trbl
- CK2
- PSK
- CDKL
- GSK
- RSKR
- IRAK
- YANK
- GRK
- TK-Sp1
- RCK
- Csk
- Akt
- Worm10
- Ryk
- RIPK
- RAD53
- Tec
- LISK
- DCAMKL
- AKT
- Dual

Legend:

Color Edges

Color ranges:

- Eph
- Src
- FGFR
- VEGFR
- PDGFR
- InsR
- DDR
- Trk
- Tec
- Abl
- Met
- Axl
- EGFR
- FAK
- JakA
- Ack
- Ror
- Fer
- Syk
- ALK
- Tie
- Csk
- Ret
- KIN16
- TK-Sp1
- Musk
- KIN6
- Sev
- CCK4
- Ryk
- PKA
- MAST
- RSK
- PKC
- Akt
- SGK
- PKN
- STE20
- MAPK
- DYRK
- CDK
- CDKL
- RCK
- GSK
- RIPK
- CLK
- CK1
- CAMKL
- STKR
- NDR
- DMPK
- GRK
- CK2
- STE7
- STE11
- PIM
- SRPK
- Trbl
- TTBK
- Worm6
- TTBKL
- Dual
- YANK
- RAF
- VRK
- CASK
- MLK
- IRAK
- LRRK
- TSSK
- CAMK1
- Worm8
- Worm9
- Trio
- MLCK
- PDK1
- RSKL
- Worm10
- Worm7
- RSKR
- Lmr
- LISK
- PKD
- CAMK2
- MAPKAPK
- PKG
- DCAMKL
- PHK
- PSK
- AKT
- RAD53
- DAPK

Legend:

Color Edges

Color ranges:

- Eph
- PKA
- RSK
- CAMK2
- CAMK1
- SGK
- MAPKAPK
- DCAMKL
- DAPK
- PKG
- PHK
- PSK
- Akt
- AKT
- PKC
- MLCK
- PDK1
- PKN
- CDK
- MAPK
- CK1
- Src
- CAMKL
- CK2
- CDKL
- RCK
- GSK
- DYRK
- PIM
- SRPK
- CLK
- PKD
- GRK
- MAST
- Trbl
- TSSK
- RAF
- YANK
- DMPK
- STE11
- STE20
- Worm6
- TTBKL
- Dual
- CASK
- RIPK
- VRK
- TTBK
- LRRK
- NDR
- Sev
- InsR
- ALK
- Trio
- CCK4
- MLK
- RAD53
- STE7
- Worm10
- Syk
- Worm7
- Worm9
- STKR
- DDR
- Fer
- IRAK
- RSKL
- Worm8
- FAK
- RSKR
- FGFR
- Tec
- Trk
- Csk
- JakA
- LISK
- Abl
- Ror
- Ack
- Lmr
- Axl
- EGFR
- Tie
- Met
- Ryk
- KIN16
- Musk
- KIN6
- PDGFR
- TK-Sp1
- Ret
- VEGFR