Genome-Wide Search for Tyrosine Phosphatases in the Human Genome Through Computational Approaches Leads to the Discovery of Few New Domain Architectures

Evolutionary Bioinformatics
Volume 15: 1–10
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934319840289

Teerna Bhattacharyya and Ramanathan Sowdhamini
National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India.

ABSTRACT: Reversible phosphorylation maintained by protein kinases and phosphatases is an integral part of intracellular signalling, and phosphorylation on tyrosine is extensively utilised in higher eukaryotes. Tyrosine phosphatases are enzymes that not only scavenge phosphotyrosine but are also involved in wide range of signalling pathways. As a result, mutations in these enzymes have been implicated in the pathogenesis of several diseases like cancer, autoimmune disorders, and muscle-related diseases. The genes that harbour phosphatase domain also display diversity in co-existing domains suggesting the recruitment of the catalytic machinery in diverse pathways. We have examined the current draft of the human genome, using a combination of 3 sequence search methods and validations, and identified 101 genes encoding tyrosine phosphatase-containing gene products, agreeing with previous reports. Such gene products adopt 37 unique domain architectures (DAs), including few new ones and harbouring few co-existing domains that have not been reported before. This semi-automated computational approach for detection of gene products belonging to a particular superfamily can now be easily applied at whole genome level on other mammalian genomes and for other protein domains as well.

KEYWORDS: protein superfamily, dephosphorylation, cell signalling, sequence search algorithms, putative domains

RECEIVED: February 27, 2019. **ACCEPTED:** March 4, 2019.

TYPE: Computational Bioinformatics Tools for Evolutionary Genomics - Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors would like to thank RS's JC Bose fellowship (SB/S2/JCB-071/2015) funded by Science & Engineering Research Board, India.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ramanathan Sowdhamini, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bellary Road, Bangalore 560065, Karnataka, India. Email: mini@ncbs.res.in

Introduction

Protein phosphorylation and dephosphorylation are important switches in signalling pathways and are regulated by kinases and phosphatases, respectively.[1] This reversible modification of residues is common across all clades of life, and about 30% of all eukaryotic proteins may be phosphorylated.[2] Tyrosine phosphorylation is essential in controlling several aspects in the life cycle of a cell-growth, differentiation, metabolism, cell cycle, intercellular communications, cell migration, gene transcription and also ion channels, immune response, and survival of the cell. It is easily discernible that abnormal tyrosine phosphorylation would be associated with several human diseases, including cancers, diabetes, rheumatoid arthritis, hypertension, and myopathies.[3] The phosphorylation on tyrosine residue is much more common in multi-cellular eukaryotes, and this post-translational modification serves as a fundamental mechanism for numerous important aspects of eukaryote physiology.[4]

It has been observed that the number of phosphatases in the human genome is much less than that of kinases, yet they are known to play diverse roles as modular enzymes. It is thus interesting to study the diversity in domain architecture (DA) of phosphatases and the evolution of their versatile biochemical functions.[1] Unlike kinases, which are derived from a common ancestor, phosphatases have evolved in structurally and mechanistically distinct families.

A recent report liberally classifies human protein phosphatases into 10 structural folds. Three such folds share a common cysteine-based catalytic motif and include the families protein tyrosine phosphatase (PTP) and low-molecular-weight phosphatase (LMWP). Other than Ser/Thr phosphatase families PPP and PPM, histidine phosphatases and alkaline phosphatases also find a place in this fold classification.[5] The family-specific classification of tyrosine phosphatase superfamily that has been followed in this study is presented in Figure 1A.

Protein tyrosine phosphatases are further divided into 2 subfamilies: the classical PTPs, which include receptor-type and cytosolic or non-receptor-type PTP, and dual-specificity phosphatase (DuSP). LMWP has no sequence homology with PTP and has not been included in our studies. However, both share the signature active-site motif HCXXGXGRS(T), where X is any residue and the conserved cysteine (C) acts as a nucleophile in the catalytic mechanism (Figure 1B), to give a phosphocysteinyl intermediate, which is the rate-determining step in the reaction. Therefore, it is thought to be a case of convergent evolution.[1,6]

Previous studies from the lab on tyrosine phosphatases in the first draft of the human genome, termed as CAPS 2003 GWS henceforth, showed that about 100 gene products containing tyrosine phosphatase domain adopt 26 different DAs and carry out various enzymatic functions.[1] We have re-examined the genome-wide survey using improved computational approaches and the current draft of the human genome. Through our sequence search, we could identify more than 500 gene products of the human genome containing at least 1 phosphatase domain, and these could be mapped

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

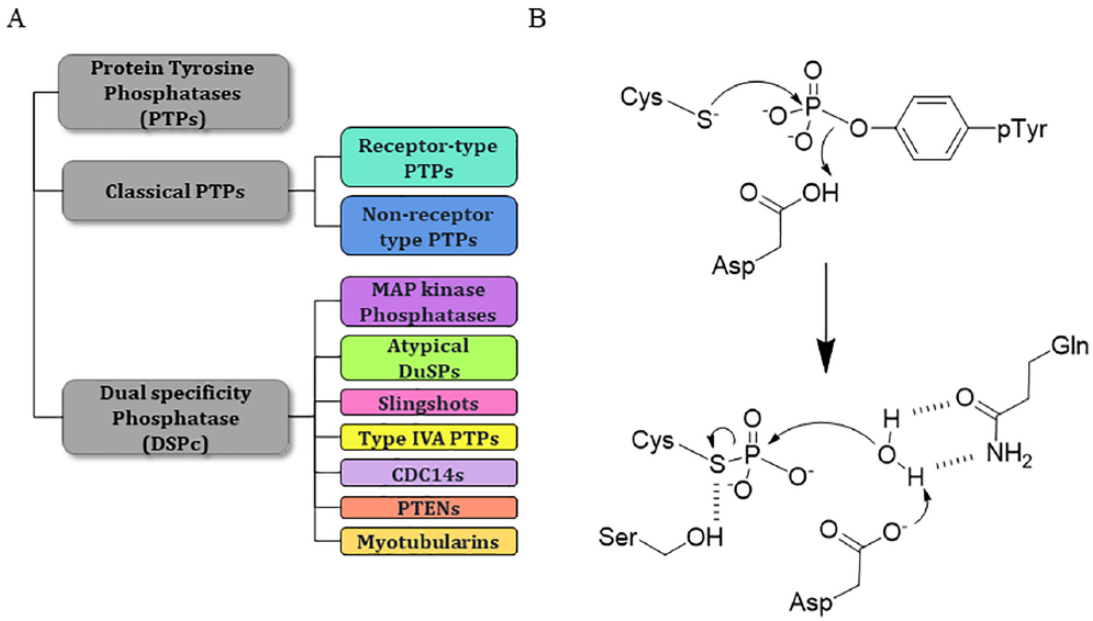

Figure 1. (A) Family-level classification of tyrosine phosphatases showing subfamilies classical PTPs and DuSPs, which act on only phosphorylated tyrosine residues and range of substrates, respectively. (B) Catalytic mechanism of tyrosine phosphatases showing key residues: cysteine (nucleophile), aspartic acid (general acid/base catalyst), serine (stabilises thiolate ion), and glutamine (helps position water in the active site for hydrolysis of the phosphorylated substrate). DuSPs indicate dual-specificity phosphatases; PTPs, protein tyrosine phosphatase.

to 101 genes. Domain annotation of the 500 odd gene products reveals that human tyrosine phosphatases adopt 37 unique DAs and the catalytic domains co-occur with 37 types of domains in these DAs. Such DA analysis further emphasises the modular nature of tyrosine phosphatases. Our findings are in accordance with the previous reports on the tyrosine phosphatase repertoire of the human genome, and this approach for detection of gene products and domain annotation of sequences can be easily extended to other mammalian genomes.

Methods

In all, 36 FASTA sequences for proteins belonging to the superfamily (phosphotyrosine protein) phosphatases II (SCOPe superfamily: 52799) were retrieved from the PASS2.6 database,[7] an in-house database that contains superfamilies with less than 40% sequence identity, in accordance with Structural Classification of Proteins extended (SCOPe; version 2.06). Hidden Markov Models, created from the structure-based sequence alignment for these 36 members, were also downloaded from the PASS2.6 database.[7]

The complete set of sequences of gene products encoded by the human genome (109052 sequences in January 2018) were downloaded from National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens) to form the database. The sequence searches[8-10] were carried out against this database.

Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) builds its own scoring matrix called position-specific scoring matrix (PSSM), which takes into account the conservation of amino acids at alignment positions. It carries out iterative BLAST runs using this PSSM and incorporates new hits obtained from each round of BLAST into the PSSM. Therefore, PSI-BLAST can be used for recognising remote homologues from a set of sequences. Position-Specific Iterative Basic Local Alignment Search Tool version 2.3.0+ was used to carry out sequence to profile-based search on the sequences encoded by the human genome obtained from NCBI, with queries obtained from PASS2.6. A total of 20 iterations were used, the E-value cut-off was kept as 0.00001, and the inclusion threshold was 0.1.

Pattern Hit Initiated Basic Local Alignment Search Tool (PHI-BLAST) permits the user to employ pattern as a seed for alignment. The patterns can be sequence motifs reported in the literature for the particular superfamily. PHI-BLAST version 2.3.0+ was used to search against the sequences encoded by human genome from NCBI, with the (Phosphotyrosine protein) phosphatases II superfamily members as queries. The patterns were obtained for these 36 queries using MOTIFS, an in-house server for prediction of sequence motifs, and included the catalytic motif of tyrosine phosphatases (VHCXXGXGR(S/T); Table 1). The motifs were also compared with published reports in the literature.[11,12] An E-value cut-off of 0.00001 and inclusion threshold of 0.1 were applied and the number of iterations used was retained as 20.

The hmmsearch module of the HMMER suite package (version 3.1b2) was used as a profile-to-sequence search-based approach. PASS2 HMM (generated from PASS2 alignment) for members of superfamily 52799 was used as a query for hmmsearch against sequences encoded by the

Table 1. Example of sequence motifs used for PHI-BLAST search using MOTIFS program. The key residues corresponding to the consensus sequence motif are marked in bold.

MOTIF NUMBER	MOTIF IN PROSITE FORMAT	CORRESPONDING CONSENSUS MOTIF	BIOLOGICAL ROLE OF MOTIF
1	[ACGFKMLPSRV]-[ACDFIHLPTV]-[ACDGIMLTV]-[EDIMLT**V**]-[Y**H**KLG]-[A**C**GKNS]-[AEHKLNQPSRTVY]-[AEDGHMNPR]-[SD**G**V]-[ACFIKLRVY]-[D**G**HNSTY]-[**R**LF]-[A**S**K**T**G]-[ACDGKPRV]-[ACFILQSTV]-[FIMLPRTVY]-[ACEILVY]	VHCXXGXGR(S/T)	Catalytic motif
2	[EFIL**N**V]-[ADFIKLNPRTY]-[ACEDGF**K**LNPT]-[ADGFHKL**N**SRW]-[FIHLQP**R**TVY]-[ACFIMLTV**Y**]	NXKNRY	Substrate recognition motif in classical PTPs
3	[AFIMLQTVY]-[AGKQ**R**Y]-[AEDGIKLNQSRV]-[ACEDGFHKLNQSRVY]-[CEKLQ**R**]	RXXR	Stabilises active-site residues through hydrogen bonds

human genome, as given by NCBI, with an E-value cut-off of 0.1. Hits above the inclusion threshold were taken for further analysis.

The hits obtained from PSI-BLAST, PHI-BLAST, and hmmsearch were pooled and compared with check for redundancy. Venn diagram, representing these hits, was constructed using Venny.[13] Redundant entries were recognised using accession IDs and removed. Furthermore, sequences with 100% identity were identified using CD-HIT[14] and were removed from the final dataset. However, isoforms were retained, as they might be physiologically relevant. Subsequently, the accession IDs of the final set of FASTA sequences were mapped to the NCBI Gene database, using NCBI Batch Entrez.

The final set of FASTA sequences, obtained from all 3 sequence search methods, was validated using hmmscan and RPS-BLAST (using CD-search against Pfam and CDD). The hmmscan module from HMMER suite package (version 3.1b2) employs sequence to HMM-based searches for annotation of domains. HMM, corresponding to the Pfam database (version 31.0), was used for this purpose. RPS-BLAST was carried out using online batch mode of CD-search and against Conserved Domains Database (CDD). The E-value used for both the searches were kept at 0.01. The sequences were annotated with CDD domains and the domain coordinates were obtained. Hits from hmmscan and CD-search with target coverage more than 70% were taken into account and the overlaps were resolved using a Python script, where containment of a domain inside a bigger domain was resolved based on parameters such as i-E-value and domain boundaries. The unique DAs were also extracted using this script. However, through further manual pruning of domain overlaps based on i-E-value and NCBI records of domain boundaries for each sequence, the final list of unique DAs was obtained. The number and type of co-existing domains were also calculated. A comparison with previously reported DAs was also performed.[1,4] TMHMM[15] and SignalP[16] were used for the prediction of transmembrane helical regions and signal peptides for the tyrosine phosphatase sequences obtained, respectively.

Alignment-free DA-similarity search (ADASS) is an algorithm used to compare DAs adopted by proteins, based on a dataset-dependent score called DA Distance (DAD) score.[17] The DAs are divided into triplets (to efficiently compare single-domain members with other multi-domain ones, N and C termini are introduced as domains for each DA) and through matching and scoring all the triplets, domain neighbourhood information is assessed. Using an empirical scoring scheme, the DAD scores are calculated. Domain architecture (DA) pairs that are similar will acquire low DAD scores, and highly dissimilar DA pairs will have a high DAD score. Using the set of 37 unique DAs, ADASS was applied to generate a distance matrix, which was in turn used to construct a neighbour-joining tree in Phylip[18] and visualised in iTOL.[19]

Results

Genome-wide search for tyrosine phosphatase domain containing gene products

A total of 837, 837, and 759 hits were obtained from PSI-BLAST, PHI-BLAST, and hmmsearch runs, respectively. These were pooled to obtain a non-redundant set of hits for which the corresponding FASTA sequences were retrieved. These FASTA sequences were further filtered for 100% sequence identity, as described in the 'Methods' section, giving rise to 575 total sequences. It was found that 202 of them are annotated as natural proteins in RefSeq (eg, NP_000243.1 or myotubularin) and the other 373 sequences correspond to 'Predicted' proteins or sequences, which have transcript evidence and/or protein homology support (eg, XP_005271136.1 or receptor-type tyrosine-protein phosphatase F isoform X1). The predicted sequences primarily consisted of isoforms of the tyrosine phosphatase-containing gene products. These were also included in the following analysis, as they were found to contain the signatures of tyrosine phosphatases and at least 1 catalytic domain.

The 575 sequences were further mapped to the NCBI Gene database, and 101 genes encoding these tyrosine phosphatase domain containing sequences were obtained. Out of these genes, receptor-type and non-receptor-type tyrosine phosphatases are

encoded by 20 and 17 genes, respectively. The remaining 64 genes encode dual-specificity phosphatase-containing gene products.

Domain architectures adopted by tyrosine phosphatase-containing gene products of the human genome

The catalytic domain associates with several other domains, whose function maybe to aid dephosphorylation by the catalytic domain (eg, GRAM domain adjacent to a catalytic domain in Myotubularins bind to phosphoinositides, the primary substrates) or a different function altogether (eg, DNA-J domain in GAK and DNAJ6 proteins, which is associated with a chaperone system).

The 575 sequences (101 genes) corresponding to tyrosine phosphatase-containing gene products were annotated for domains using hmmscan and RPS-BLAST against the Pfam and CDD, respectively. The results of the 2 domain annotation searches were pruned to resolve domain overlaps. These and other steps for validation of domain boundaries from NCBI revealed that they adopt 37 unique DAs.

The presence and location of transmembrane helices and signal peptide regions were detected for these sequences using TMHMM and SignalP, respectively. It was found that 32% and 26% of the sequences contained at least 1 transmembrane helix and signal peptide region. Of the 184 gene products that were annotated by TMHMM to contain at least 1 transmembrane helix or TMH, most were receptor-type tyrosine phosphatases. However, 5 gene products, which were annotated with the presence of 4 transmembrane helices were dual specificity phosphatases – TPTE and TPTE2 (Transmembrane Phosphatase with TEnsin homology and its homologue).[20]

The associated domains had functions ranging from kinase activity (eg, DNAJ6 and GAK) to adhesion or interaction between receptor-type tyrosine phosphatases (MAM domain, in PTPRK, PTPM, PTPRT, and PTPRU). Diversity among domains associated with the catalytic domain was highest in the case of DuSPs, pointing to its diversity of substrates. It was also seen that domain repeats are common in the classical tyrosine phosphatases. The domains found to commonly occur in repeats were the fibronectin type-III (fn3) domain, immunoglobulin or Ig domain, and PDZ domain.

ADASS was applied on the set of 37 unique DAs and the 3 main classes of PTPs, namely, DuSPs, non-receptor or cytosolic (cPTP), and receptor-type (rPTP) tyrosine phosphatases were found to be clustered accordingly (Supplementary Figure 3). The only DuSPs, which clustered with the rPTPs, are TPTE and TPTE2 (marked in yellow star in Supplementary Figure 3), perhaps due to the transmembrane helices, like the rPTPs (further discussion about this DA can be found later). In our previous analysis (CAPS 2003 GWS), we had identified 26 unique DAs, adopted by 96 tyrosine phosphatase domain containing gene products, starting from the first draft of the human genome.[1,21] The then reported DAs were compared with the current GWS results (as elucidated in Figure 2), and the 11 DAs identified in CAPS 2018 GWS are marked by red star symbols. For example, in the case of the MAP kinase phosphatases DUSP16 or MKP-7, it was seen that the sequence limits previously annotated as CH2 domain have been currently annotated as Rhodanese domain.[22] The increase in the number of unique DAs over time demonstrates the requirement for continued annotation efforts of domains of unknown function or unannotated regions of the sequences. These DAs primarily include those adopted by gene products, which had not been considered by the previous GWS effort from the lab. A DA described in the previous GWS that has not been considered in the current analysis corresponds to cdc25 phosphatase, which does not belong to the tyrosine phosphatase superfamily (SCOPe ID: 52799). It also shares poor sequence and structural similarity with classical tyrosine phosphatases and dual specificity phosphatases.[23] Further about the 11 DAs has been discussed in the next section. Apart from these, the DAs reported by CAPS 2003 GWS could be annotated with 15 additional domains, and these are marked by red plus symbols in Figure 2. Three of these new domains are being reported first time by CAPS 2018 GWS (Ricin-B lectin, PTN13_u3 and Endomucin) and these 4 DAs (corresponding to DA numbers 20, 30, 35, and 36 in Figure 2) appear to be new.

Domain architectures adopted by tyrosine phosphatases

The 11 DAs discussed in the previous section are mostly adopted by gene products that had not been included in CAPS 2003 GWS. In Table 2, the Gene Ontology terms associated with gene products that represent these 11 DAs are provided. The gene products have various molecular functions and are localised at different locations in the cell, and the diversity in DA enables the gene products to carry out different functions. Some of these DAs also harbour domains that have not been reported elsewhere as well, for example, DA number 30. A few of the newer DAs as given in Table 2 are discussed in detail in the following.

Slingshot phosphatases (SSH1-3) consist of 3 domains (DA No. 5 in Table 2): an N-terminal domain, which is conserved in the slingshot phosphatases (SSH_N), and a DEK C-terminal domain (DEK_C, known as a chromatin-associated protein that is linked with cancers and autoimmune disease) co-existing with DuSPs domain (DSPc). Slingshot or SSH family of phosphatases regulate actin dynamics by dephosphorylating and activating cofilin, which is an actin-depolymerising factor. Phosphatase activity of SSH1 is mediated by the SSH_N domain.[24] Furthermore, the SSH_N domain plays an important role in the localisation of SSH1 in the lamellipodium in the cell.[24]

TPTE or transmembrane phosphatase with TEnsin homology has been well-studied as a voltage-sensitive phosphatase (VSP; DA 9). Voltage-gated channels possess a characteristic sensor domain – a pore helix. In the case of VSPs (TPTE and

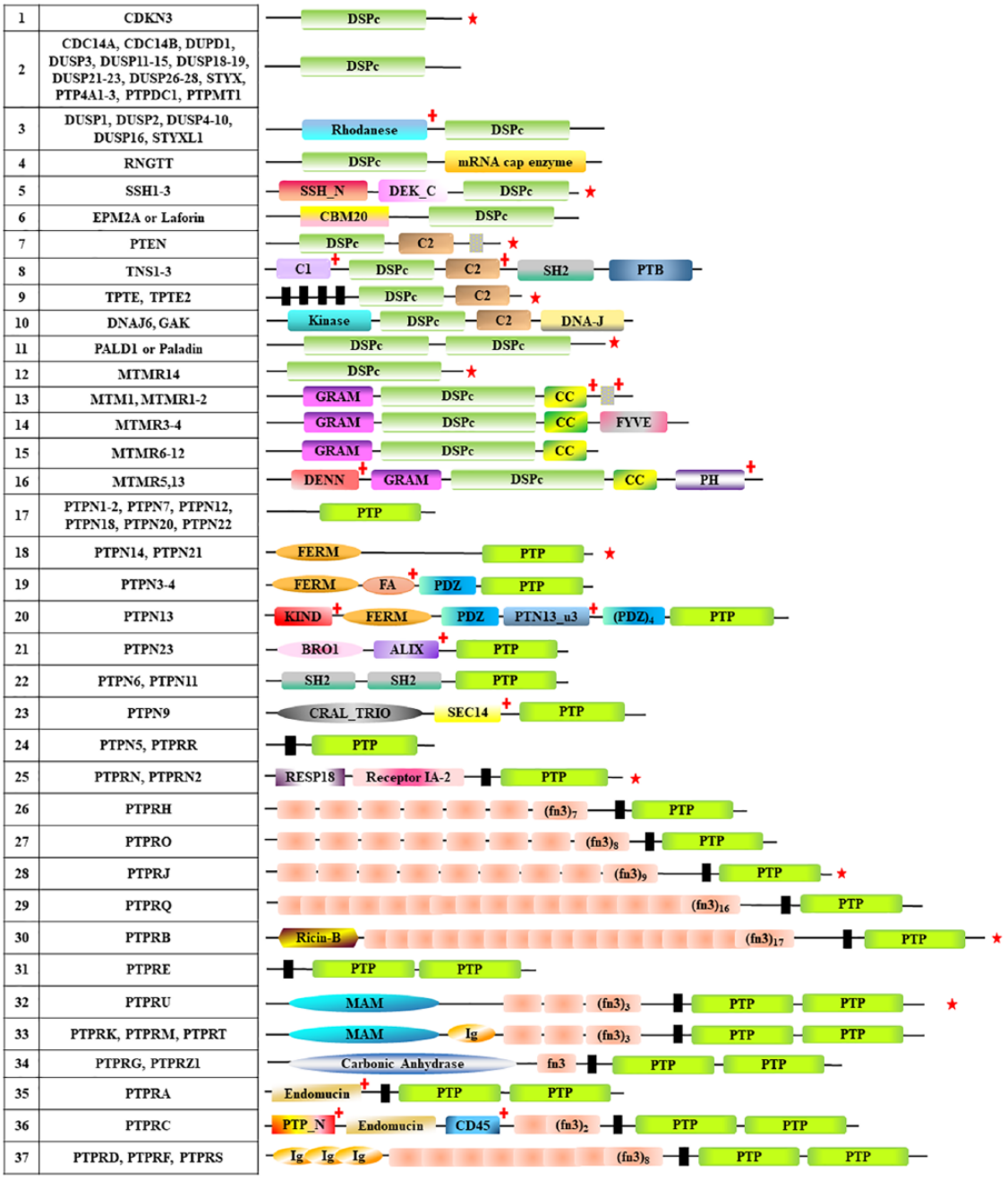

Figure 2. Domain architectures (DAs) adopted by tyrosine phosphatase-containing gene products of the human genome. A total of 37 unique DAs were identified through domain annotation of sequences obtained from the genome-wide search. The gene products corresponding to the DAs are also listed. The tyrosine phosphatase domains are in different shades of green and differentiated based on the class. The black boxes represent transmembrane helices and the grey box in the DAs adopted by PTEN, MTM1, MTMR1, and MTMR2 corresponds to a predicted PDZ binding motif. The individual domains and the DAs that were not reported in the previous GWS from the lab (CAPS 2003 GWS) have been marked with red plus symbols and stars, respectively. Further details on domain names have been provided in Supplementary Table 2.

TPTE2), it was found that the 4 N-terminal transmembrane helices form an unique sensor domain and the pore region, like those of voltage-gated channels, is replaced with a PTEN domain, followed by a C2 domain. Voltage-sensitive phosphatases act on membrane phosphoinositides PI(4,5)P_2 and PI(3,4,5)P_3, and it is possible that the phosphatase activity is regulated by voltage changes.[25,26]

Paladin or PALD1 (DA number 11, Table 2) has 2 PTP domains, which retain consensus active-site motifs, but lack other characteristics of PTPs. Human Paladin is implicated in inhibition of insulin signalling, but there have not been further elucidations of its substrate(s). In chick embryo, it was shown that Paladin plays a role in neural crest formation and migration. Paladin was labelled as an 'antiphosphatase' in the study in chick embryo, as active-site cysteine residues were not essential for Paladin activity in neural crest formation.[27] Paladin is found in all vertebrates and it will be interesting to study DAs in genomes other than human and its associated functions.

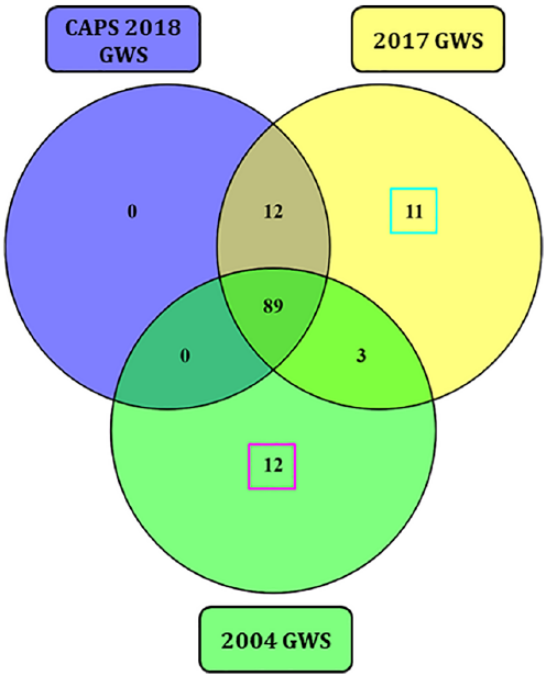

Figure 3. Venn diagram representing genes reported as tyrosine phosphatase encoding by different groups. The 12 genes (marked in pink box in the diagram) that are shown to be reported by only 2004 GWS have either been renamed or declared pseudogenes. The 11 genes (marked in cyan box) reported only by 2017 GWS are either from a different fold or superfamily and have not been considered for CAPS 2018 GWS. Further details have been provided in Supplementary Table 1.

PTPRN and its homologue, PTPRN2, are 2 receptor-type tyrosine phosphatases (DA 25 in Table 2) with an extracellular N-terminal RESP18 domain, a membrane-proximal Receptor IA-2 domain (with ferredoxin-like fold), a single transmembrane helix, and a PTP domain. PTPRN, also kown as IA-2 or ICA512 (islet cell antigen 512), is an autoantigen of Type I diabetes and associates with secretory granules of neuroendocrine cells, including pancreatic beta-cells. On exocytosis, the cytoplasmic domain of PTPRN is cleaved and translocates to the nucleus, where it binds E3-SUMO ligase protein inhibitor of activated signal transducer and activator of transcription-y (PIASy) and upregulates insulin expression.[28]

DA 30 (Table 2) is adopted by a receptor-type tyrosine phosphatase, PTPRB. Apart from a transmembrane helix and a single phosphatase domain, there are several repeats of fn3 domains, which range from 15 to 17 in number among isoforms. There is also an N-terminal domain, which is annotated as a Ricin B-lectin domain by both CD-search and hmmscan against CDD and Pfam databases, respectively, with significant E-values and is being reported for the first time. The Ricin B-lectin domain has been shown to bind simple sugars such as galactose or lactose. The domain has been reportedly found in a variety of proteins serving diverse functions such as enzymatic activity, inhibitory toxicity, and signal transduction.[29]

Genome-wide search efforts in the literature

Over the years, there have been various efforts to understand the repertoire of tyrosine phosphatases in the human genome better. As genome annotations change over time, with improved tools for sequence analysis and growing biochemical and structural studies, it becomes important to continuously update the figures reported for the human phosphatome.

While working on the sequence search studies, we came across 2 studies on human tyrosine phosphatases, one of them recent, and have compared our results with these. It is worth mentioning that we chose to be more stringent in the classification of tyrosine phosphatases and did not include lower molecular weight tyrosine phosphatases, Sac and cdc25 phosphatases. Due to paucity of gene mapping information in the CAPS 2003 GWS, the comparison of reported human tyrosine phosphatase-encoding genes has been performed with only 2017 GWS and 2004 GWS and is elucidated in Figure 3 and Supplementary Table 1. The differences observed on comparison of the 3 genome-wide search efforts (CAPS 2018 GWS, 2017 GWS, and 2004 GWS), such as a change in gene name or declaration of a gene as a pseudogene, have been reported in the last column 'Comments' in Supplementary Table 1. Literature search was carried out for each gene, which was reported by either or both the other GWS but not by CAPS 2018 GWS. In most cases, it was seen that the genes which were not reported belonged to another fold or another superfamily altogether.

We have also looked at the domains associated with catalytic domains in the gene products we have obtained through our sequence search. We found 37 different domains associating with the tyrosine phosphatase domains and compared with the domains reported elsewhere (GWS-CAPS 2003 GWS, 2004 GWS and 2017 GWS; Figure 4). Through our domain annotation protocol, we could identify 3 domains, which are exclusive to CAPS 2018 GWS as seen in Figure 4 (Endomucin, PTN13_u3 and Ricin B-lectin, please see subsequent sections for details). There are few domains that are exclusive to CAPS 2003 GWS and 2004 GWS; these are either no longer found to associate with PTP domains or their nomenclature has changed over the last 15 years. Details about each domain have been provided in Supplementary Table 2.

Three new domains associated with the catalytic domains

Three domains were found to co-occur with phosphatase domain, which are exclusive to our GWS, namely, Ricin B-lectin, Endomucin, and PTN13_u3 domains (Figure 4).

Ricin is a well-known toxic plant lectin with AB_5 quaternary arrangement. The A chain of Ricin retains the toxic part and the B-chain is involved in targeting cells and binding carbohydrates. The Ricin B-chain adopts a β-trefoil fold and is characterised by the presence of 3 (Q-X-W) repeats[30] and

Table 2. Molecular function, biological process, and cellular localisation information for gene products representing the 11 new domain architectures.

DOMAIN ARCHITECTURE NUMBER	REPRESENTATIVE GENE PRODUCT	MOLECULAR FUNCTION GO TERMS	BIOLOGICAL PROCESS GO TERMS	CELLULAR LOCALISATION GO TERMS
1	CDKN3	Protein serine/threonine phosphatase activity (GO:0004721), protein serine/threonine phosphatase activity (GO:0004722) and protein tyrosine phosphatase activity (GO:0004725)	Regulation of cyclin-dependent protein serine/threonine kinase activity (GO:0000079), G1/S transition of mitotic cell cycle (GO:0000082), negative regulation of cell proliferation (GO:0008285) and cell cycle arrest (GO:0007050)	Nucleus (GO:0005634), cytoplasm (GO:0005737), and perinuclear region of cytoplasm (GO:0048471)
5	SSH1	Actin binding (GO:0003779), phosphoprotein phosphatase activity (GO:0004721) and protein binding (GO:0005515)	Cell morphogenesis (GO:0000902), protein dephosphorylation (GO:0006470) and regulation of actin polymerization or depolymerization (GO:0008064)	Cytoskeleton (GO:0005856), lamellipodium (GO:0030027), and plasma membrane (GO:0005886)
7	PTEN	Phosphatidylinositol-3-phosphatase activity (GO:0004438), platelet-derived growth factor receptor binding (GO:0005161) and anaphase-promoting complex binding (GO:0010997)	Regulation of cyclin-dependent protein serine/threonine kinase activity (GO:0000079), angiogenesis (GO:0001525) and regulation of B-cell apoptotic process (GO:0002902)	Extracellular region (GO:0005576), nucleus (GO:0005634), and cytoplasm (GO:0005737)
9	TPTE	Phosphoprotein phosphatase activity (GO:0004721), protein tyrosine phosphatase activity (GO:0004725), and hydrolase activity (GO:0016787)	protein dephosphorylation (GO:0006470), signal transduction (GO:0007165), and peptidyl-tyrosine dephosphorylation (GO:0035335)	Membrane (GO:0016020)
11	PALD1 or paladin	Protein serine/threonine phosphatase activity (GO:0004722), protein tyrosine phosphatase activity (GO:0004725), and protein binding (GO:0005515)	Protein dephosphorylation (GO:0006470) and peptidyl-tyrosine dephosphorylation (GO:0035335)	Cytosol (GO:0005829) and nucleus (GO:0005634)
12	MTMR14	Phosphatidylinositol-3-phosphatase activity (GO:0004438), phosphatidylinositol-3,5-bisphosphate 3-phosphatase activity (GO:0052629), and protein serine/threonine phosphatase activity (GO:0004722)	Phosphatidylinositol biosynthetic process (GO:0006661), macroautophagy (GO:0016236), and protein dephosphorylation (GO:0006470)	Cytosol (GO:0005829), perinuclear region of cytoplasm (GO:0048471), and ruffle (GO:0001726)
18	PTPN14	Protein tyrosine phosphatase activity (GO:0004725), transcription coregulator activity (GO:0003712), and receptor tyrosine kinase binding (GO:0030971)	Lymphangiogenesis (GO:0001946), regulation of transcription, DNA-templated (GO:0006355) and negative regulation of cell proliferation (GO:0008285)	Nucleus (GO:0005634), nucleoplasm (GO:0005654), and cytoplasm (GO:0005737)
25	PTPRN	GTPase binding (GO:0051020), spectrin binding (GO:0030507), and ubiquitin-like protein ligase binding (GO:0044389)	Response to reactive oxygen species (GO:0000302), regulation of transcription, DNA-templated (GO:0006355), and response to glucose (GO:0009749)	Endosome (GO:0005768), golgi apparatus (GO:0005794), and plasma membrane (GO:0005886)

(Continued)

Table 2. (Continued)

DOMAIN ARCHITECTURE NUMBER	REPRESENTATIVE GENE PRODUCT	MOLECULAR FUNCTION GO TERMS	BIOLOGICAL PROCESS GO TERMS	CELLULAR LOCALISATION GO TERMS
28	PTPRJ	Phosphoprotein phosphatase activity (GO:0004721), platelet-derived growth factor receptor binding (GO:0005161), and protein binding (GO:0005515)	Negative regulation of cell proliferation (GO:0008285), negative regulation of platelet-derived growth factor receptor signalling pathway (GO:0010642), and peptidyl-tyrosine dephosphorylation (GO:0035335)	Immunological synapse (GO:0001772), plasma membrane (GO:0005886), and cell-cell junction (GO:0005911)
30	**PTPRB**	Phosphoprotein phosphatase activity (GO:0004721), transmembrane receptor protein tyrosine phosphatase activity (GO:0005001), and protein binding (GO:0005515)	Angiogenesis (GO:0001525), protein dephosphorylation (GO:0006470), and neutrophil degranulation (GO:0043312)	Plasma membrane (GO:0005886), specific granule membrane (GO:0035579), and receptor complex (GO:0043235)
32	PTPRU	Transmembrane receptor protein tyrosine phosphatase activity (GO:0005001), beta-catenin binding (GO:0008013), and hydrolase activity (GO:0016787)	Cell adhesion (GO:0007155), negative regulation of cell proliferation (GO:0008285), and animal organ regeneration (GO:0031100)	Plasma membrane (GO:0005886), integral component of plasma membrane (GO:0005887), and cell-cell junction (GO:0005911)

Abbreviation: GO, gene ontology.

variable disulphide bridges within each repeat. The N-terminal region of PTPRB or Protein Tyrosine Phosphatase Receptor Type B was predicted to retain a Ricin B-lectin domain, with Q-X-W, P-X-W, and H-S-W at the classical Q-X-W motif regions. Sequence comparison with few other lectin families, through PCA, also showed that the PTPRB Ricin B-lectin domain indeed clustered with other domains within the Ricin B-like lectins superfamily (50370) in SCOPe. Details about the Ricin B-lectin domain of PTPRB have been provided in Supplementary Figure 4 and Supplementary Table 3.

Endomucin, also known as sialomucin or mucin-like sialoglycoprotein, is a membrane-bound glycoprotein expressed luminally by endothelial cells.[31,32] The sequences of PTPRA and PTPRC were annotated with the presence of an endomucin domain at their N-terminus. Both receptor-type tyrosine phosphatases and the presence of an endomucin domain may allude to the cell-cell adhesion-related functions of PTPRA and PTPRC. While it has been seen that the extracellular region of PTPRA, and not the catalytic domain, is required for modulation of cell surface expression of a neural adhesion molecule NB-3, role of PTPRC in adhesion in context of T cell adhesion has been well documented till date. Further information has been provided in Supplementary Table 3.

The gene products of PTPN13 or PTP, non-receptor type 13, were annotated with the presence of a linker domain PTN13_u3. This unstructured linker is present between the first and second PDZ domains of the 5 total PDZ domains present on this protein. PTPN13 is known to negatively regulate FAS-induced

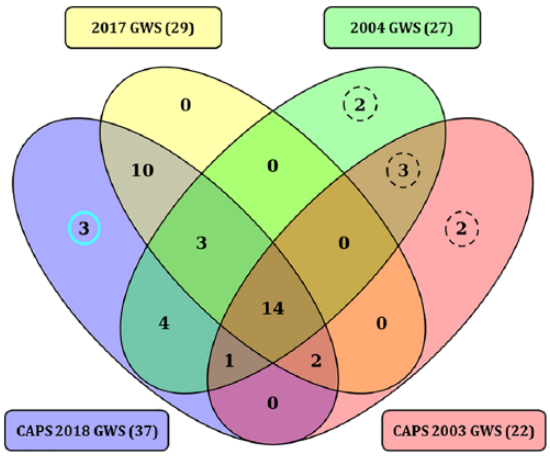

Figure 4. Comparison of the domains found to associate with PTP domains, across different GWS reported in literature. A total of 37 different domains were annotated in the 575 gene products containing tyrosine phosphatase domain(s). The figures adjacent to each GWS name stand for the number of domains that associate with tyrosine phosphatase domains, as reported by the GWS; 14 domains are found to be reported across all the GWS, from 2003 to 2018. The 2 older reports – the CAPS 2003 GWS and 2004 GWS – mention few domains that are exclusive to their findings and are marked in circles with dotted line. However, these domains are no longer annotated in tyrosine phosphatase domain containing gene products or have been renamed. Three domains (marked by cyan circle) are being reported for the first time by CAPS 2018 GWS to associate with PTP domains. These were not reported by even the latest GWS, that is, 2017 GWS. Further details about the associating domains have been provided in Supplementary Table 2. PTP indicates protein tyrosine phosphatase.

apoptosis in different types of tumours[33] and it was found that single-nucleotide polymorphisms (SNPs) mapped on this gene are implicated in several types of cancer. Details about this linker can be found in Supplementary Table 3.

Discussion

Tyrosine phosphatases are known antagonists of kinases and play an equally important role in various physiological processes. It is also known that they adopt diverse DAs and have a plethora of substrates. In 2003, a genome-wide survey for tyrosine phosphatase gene products in the first draft of human genome published from the lab reported 96 such sequences.[1] In 2004, Alonso and co-workers published another report on genes encoding tyrosine phosphatases in the human genome. A total of 107 human PTP genes were reported, but the numbers have changed over the years with better annotation. Many of the then reported genes have been fused or declared as pseudogenes.[4] A fold-level classification and subsequent sequence search were reported in 2017 by Chen and co-workers; 115 genes encoding different families of tyrosine phosphatases were reported. This classification is, however, not strict and includes members from different superfamilies.

We looked at the current draft of the human genome to study the tyrosine phosphatase-encoding genes. Tyrosine phosphatases share distant homologies among each other, owing to their multi-domain nature. Hence, for our search, we used a combination of 3 sequence search methods, 2 of them being profile to sequence-based search methods (PSI-BLAST, PHI-BLAST) and one being a HMM-based sequence search method (hmmsearch) to detect remote relationships. Through these searches, putative tyrosine phosphatase domain containing gene products were identified in the human genome and these were further annotated for co-existing domains. As compared with our previous attempt,[1] even if the computational search algorithms are the same, the methods have gone through constant improvements. Furthermore, the human genome database has undergone substantial improvement, so has the number of recognised domains in Pfam and CDD database. Altogether, the work carried out and reported in this article underscores the importance of updates of secondary data resources. This also demonstrates the application of a near-automatic and computational pipeline, as shown for tyrosine phosphatases.

We report a total of 575 gene products and these retain at least 1 tyrosine phosphatase domain. The 575 gene products also include predicted hit-sequences with transcript and/or protein homology evidence as referenced by NCBI. The gene products could be mapped to 101 tyrosine phosphatase-encoding genes, which is in accordance with previous reports on the human tyrosine phosphatase repertoire. It was found that out of the 100 odd genes identified, 20 genes encode receptor-type tyrosine phosphatases and 17 encode for non-receptor-type tyrosine phosphatases. The remaining 64 genes were found to encode DuSP containing gene products, which included MAP-kinase phosphatases, slingshots, myotubularins, and other tyrosine phosphatases. The 37 unique DAs adopted by these 575 gene products contain 37 different domains, which associate with the catalytic domain(s). It was also observed that repeats of domains such as fn3, PDZ, and Ig are common among receptor-type tyrosine phosphatases. Interestingly, DuSPs, which may also act on substrates other than phosphorylated tyrosine, were found to associate with a large variety of domains such as kinase domain (DNAJ6, GAK) and Carbohydrate Binding Module or CBM20 (Laforin or EPM2A). Three domains which had not been reported previously to associate with the PTP domains, namely, Endomucin, PTN13_u3, and Ricin B-lectin domains, are being reported here for the first time. Endomucin and Ricin B-lectin are extracellular domains annotated on receptor-type tyrosine phosphatases PTPRA&C and PTPRB, respectively. Although specific roles of an extracellular N-terminal domain in both PTPRA and PTPRC have been well studied in context of protein-protein interactions, this is the first time the region is being annotated by an Endomucin domain, which is known to be involved in cell-cell adhesion. The Ricin B-lectin domain has already been documented in other organisms including mouse, and the human enzyme, PTPRB, annotated through our efforts will be a promising candidate for study of role in carbohydrate binding. Apart from these 2, a cytosolic PTP, PTPN13, was annotated with the presence of a linker domain PTN13_u3. This is an unstructured domain and has no roles attributed so far. It is, however, tempting to correlate this linker region with another in PTPN4, which regulates the adjacent PDZ domains.

Such genome-wide searches not only help in understanding the repertoire of tyrosine phosphatases in the human genome, but can aid in genome-scale structural bioinformatics efforts[34] and studies of domain evolution among multi-domain tyrosine phosphatases[11] in mammalian genomes and other functionally versatile protein domains as well. The workflow used in this genome-wide survey is appropriate for the detection of remote homologs and annotation and discovery of new DAs and associated domains and can be extended to other genomes as well as other protein families. The gene products with interesting DAs or domains which our study has identified can be taken forward for biochemical and structural studies in context of new functions or disease implications.

Underlying research material related to this manuscript can be accessed from this URL for download: http://caps.ncbs.res.in/download/hptp

Author Contributions

The work was conceptualised by RS. TB carried out the entire work and analysis. TB wrote the first draft of the manuscript and RS improved it.

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

ORCID iD

Ramanathan Sowdhamini https://orcid.org/0000-0002-6642-2367

REFERENCES

1. Bhaduri A, Sowdhamini R. A genome-wide survey of human tyrosine phosphatases. *Protein Eng*. 2003;16:881–888.
2. Bhaduri A, Sowdhamini R. Domain architectural census of eukaryotic gene products containing O-protein phosphatases. *Gene*. 2006;366:246–255.
3. Zhang Z. Protein tyrosine phosphatases: structure and function, substrate specificity, and inhibitor development. *Annu Rev Pharmacol Toxicol*. 2002;42:209–234.
4. Alonso A, Sasin J, Bottini N, et al. Protein tyrosine phosphatases in the human genome. *Cell*. 2004;117:699–711.
5. Chen MJ, Dixon JE, Manning G. Genomics and evolution of protein phosphatases. *Sci Signal*. 2017;54:1–18.
6. Kolmodin K, Aqvist J. The catalytic mechanism of protein tyrosine phosphatases revisited. *FEBS Lett*. 2001;498:208–213.
7. Ghosh P, Bhattacharyya T, Mathew OK, et al. PASS2 version 6: a database of structure-based sequence alignments of protein domain superfamilies in accordance with SCOPe. *Database*. Epub ahead of print 2019. doi:10.1093/database/baz028.
8. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. Epub ahead of print 1997. doi:10.1093/nar/25.17.3389.
9. Zhang Z. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res*. Epub ahead of print 1998. doi:10.1093/nar/26.17.3986.
10. Eddy S. Profile hidden Markov models. *Bioinformatics*. 1998;14:755–763.
11. Andersen JN, Mortensen OH, Peters GH, et al. Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Mol Cell Biol*. 2001;21:7117–7136.
12. Andersen JN, Vecchio RL Del, Kannan N, et al. Computational analysis of protein tyrosine phosphatases: practical guide to bioinformatics and data resources. *Methods*. 2005;35:90–114.
13. Oliveros JC. VENNY. An interactive tool for comparing lists with Venn Diagrams. *BioinfoGP of CNB-CSIC*, 2007, http://bioinfogp.cnnb.csic.es/tools/venny/index.ht
14. Huang Y, Niu B, Gao Y, et al. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. Epub ahead of print 2010. doi:10.1093/bioinformatics/btq003.
15. Krogh A, Larsson B, Von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol*. Epub ahead of print 2001. doi:10.1006/jmbi.2000.4315.
16. Petersen TN, Brunak S, von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. Epub ahead of print 2011. doi:10.1038/nmeth.1701.
17. Syamaladevi DP, Joshi A, Sowdhamini R. An alignment-free domain architecture similarity search (ADASS) algorithm for inferring homology between multi-domain proteins. *Bioinformation*. Epub ahead of print 2013. doi:10.6026/97320630009491.
18. Felsenstein J. PHYLIP (Phylogeny Inference Package). Distributed by the author Department of Genome Sciences University of Washington Seattle. Epub ahead of print 2005. doi:10.1523/JNEUROSCI.0009-08.2008.
19. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. Epub ahead of print 2016. doi:10.1093/nar/gkw290.
20. Tapparel C, Reymond A, Girardet C, et al. The TPTE gene family: cellular expression, subcellular localization and alternative splicing. *Gene*. Epub ahead of print 2003. doi:10.1016/j.gene.2003.09.038.
21. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. Epub ahead of print 2001. doi:10.1038/35057062.
22. Masuda K, Shima H, Watanabe M, et al. MKP-7, a novel mitogen-activated protein kinase phosphatase, functions as a shuttle protein. *J Biol Chem*. 2001;276:39002–39011.
23. Zhang Z-Y. Mechanistic studies on protein tyrosine phosphatases. *Prog Nucleic Acid Res Mol Biol*. 2003;73:171–220.
24. Kurita S, Watanabe Y, Gunji E, et al. Molecular dissection of the mechanisms of substrate recognition and F-actin-mediated activation of cofilin-phosphatase Slingshot-1. *J Biol Chem*. 2008;283:32542–32552.
25. Sutton KA, Jungnickel MK, Jovine L, et al. Evolution of the voltage sensor domain of the voltage-sensitive phosphoinositide phosphatase VSP/TPTE suggests a role as a proton channel in eutherian mammals. *Mol Biol Evol*. 2012;29:2147–2155.
26. Kumanovics A, Levin G, Blount P. Family ties of gated pores: evolution of the sensor module. *FASEB J*. 2002;16:1623–1629.
27. Roffers-Agarwal J, Hutt KJ, Gammill LS. Paladin is an antiphosphatase that regulates neural crest cell formation and migration. *Dev Biol*. Epub ahead of print 2012. doi:10.1016/j.ydbio.2012.08.007.
28. Primo ME, Klinke S, Sica MP, et al. Structure of the mature ectodomain of the human receptor-type protein-tyrosine phosphatase IA-2. *J Biol Chem*. Epub ahead of print 2008. doi:10.1074/jbc.M708144200.
29. Villafranca JE, Robertus JD. Ricin B chain is a product of gene duplication. *J Biol Chem*. 1981;256:554–556.
30. Hazes B. The (QxW)3 domain: a flexible lectin scaffold. *Protein Sci*. 1996;5:1490–1501.
31. Zahr A, Alcaide P, Yang J, et al. Endomucin prevents leukocyte-endothelial cell adhesion and has a critical role under resting and inflammatory conditions. *Nat Commun*. Epub ahead of print 2016. doi:10.1038/ncomms10363.
32. Matsubara A, Iwama A, Yamazaki S, et al. Endomucin, a CD34-like sialomucin, marks hematopoietic stem cells throughout development. *J Exp Med*. Epub ahead of print 2005. doi:10.1084/jem.20051325.
33. Niu J, Huang Y-J, Wang L-E, et al. Genetic polymorphisms in the PTPN13 gene and risk of squamous cell carcinoma of head and neck. *Carcinogenesis*. 2009;30:2053–2058.
34. Barr AJ, Ugochukwu E, Lee WH, et al. Large-scale structural analysis of the classical human protein tyrosine phosphatome. *Cell*. 2009;136:352–363.