Database, 2022, 1–6
DOI: https://doi.org/10.1093/database/baac088
Original article

ChemBioPort: an online portal to navigate the structure, function and chemical inhibition of the human proteome

Lihua Liu[1], Evianne Rovers[1] and Matthieu Schapira[1,2,*]

[1]Structural Genomics Consortium, University of Toronto, MaRS South Tower, Suite 700, 101 College Street, Toronto, Ontario M5G 1L7, Canada
[2]Department of Pharmacology and Toxicology, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada

[*]Corresponding author: Tel: 1-416-978-3092; Email: matthieu.schapira@utoronto.ca

Citation details: Liu, L., Rovers, E. and Schapira, M. ChemBioPort: an online portal to navigate the structure, function and chemical inhibition of the human proteome. *Database* (2022) Vol. 2022: article ID baac088; DOI: https://doi.org/10.1093/database/baac088

Abstract

Chemical probes are important tools to investigate the function of proteins, evaluate their potential as therapeutic targets and provide chemical starting points for drug discovery. As a result, a growing federation of scientists aims to generate chemical probes for all human druggable proteins. A diverse array of data typically guides target selection and chemical probe discovery: information on protein function can help prioritize targets, domain architecture can provide insight on druggability, structural data enables molecular design and existing chemical ligands can serve as foundation or inspiration for chemical probe development. But these heterogenous data types are dispersed across a variety of public repositories that are difficult to cross-reference by non-experts. We developed ChemBioPort, an online resource that allows users to combine queries related to the ontology, domain architecture or name of human proteins to produce downloadable tables that integrate information on function, disease association, essentiality, tissue enrichment, domain architecture, structure and chemical ligands of proteins. Users can convert these tables into dendrograms reflecting sequence similarity, onto which they can graphically project all data types, linked via a mouse-click to their original repositories or published articles. This interface will support the growing community of chemical biologists, chemists, cell and structural biologists on their perilous journey from genes to medicines.

Database URL: https://chembioport.thesgc.org

Introduction

The exploration and validation of therapeutic targets is a long and hazardous voyage that often starts with genetic studies linking a gene to a disease and ends with the observed efficacy of a drug in the clinic. Chemical probes—drug-like ligands that potently and selectively engage a target protein in cells—are an important type of tools to guide scientists in this journey, as they can be used to link a protein to a phenotype (1–3). This realization catalyzed pilot initiatives to generate chemical probes for entire protein families such as kinases (4), bromodomains (5), methyltransferases (6) or SLC carriers (7), leading to Target 2035, an ambitious mission where an international federation of scientists aims to produce a chemical probe for all druggable human proteins by the year 2035 (8, 9).

To support the prioritization of targets and the development of chemical probes, we previously created Chromohub and Ubihub, online interfaces integrating data on the disease association, structure and chemical inhibition of proteins involved in chromatin- and ubiquitin-mediated signaling, respectively (10, 11). Because Target 2035 expands beyond these specific areas of human biology, here we present ChemBioPort, a web portal on the structure, function and chemical ligands of all human proteins. ChemBioPort not only expands dramatically the scope of our previous interfaces, but also integrates novel data types, such as gene essentiality from the gnomAD database (12), mass spectrometry-derived tissue enrichment of proteins from the GTEX Consortium (13) or chemical inhibitors from Novartis' chemogenetic library (14) or the Chemical Probes portal. Novel features are implemented such as ontology- or domain-based search functions, and a mechanism to project heterogenous data types on circular dendrograms derived from multiple sequence alignments of any ensemble of proteins.

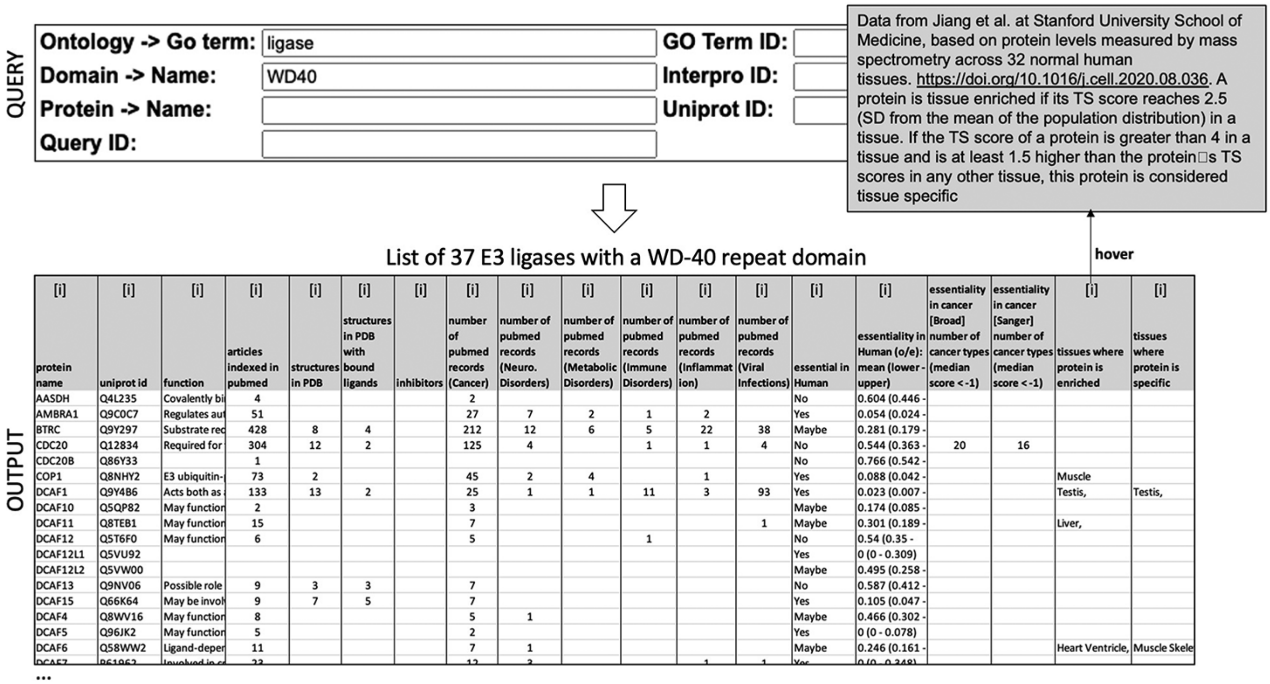
ChemBioPort is an intuitive portal to the human proteome that allows medicinal chemists, structural or cell biologists to navigate and integrate diverse arrays of datatypes relevant to chemical probe development such as disease association, tissue expression profiles, gene function and ontology, protein domain architecture, structural homologs or chemical inhibitors.

Data collection

Proteins can be queried based on the gene ontology associated to their enzyme classification number extracted from

Received 26 July 2022; Revised 14 September 2022; Accepted 20 September 2022

The Gene Ontology (15, 16) or associated to their structural domains extracted from the European Bioinformatics Institute (EBI) (17) ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go. Protein names, Uniprot IDs and domains are extracted from the EBI's Uniprot and Interpro databases (17, 18). Disease association is derived from MeSH terms in PubMed as previously described (10) and PubMed records for any gene are extracted from NCBI's gene2PubMed resource



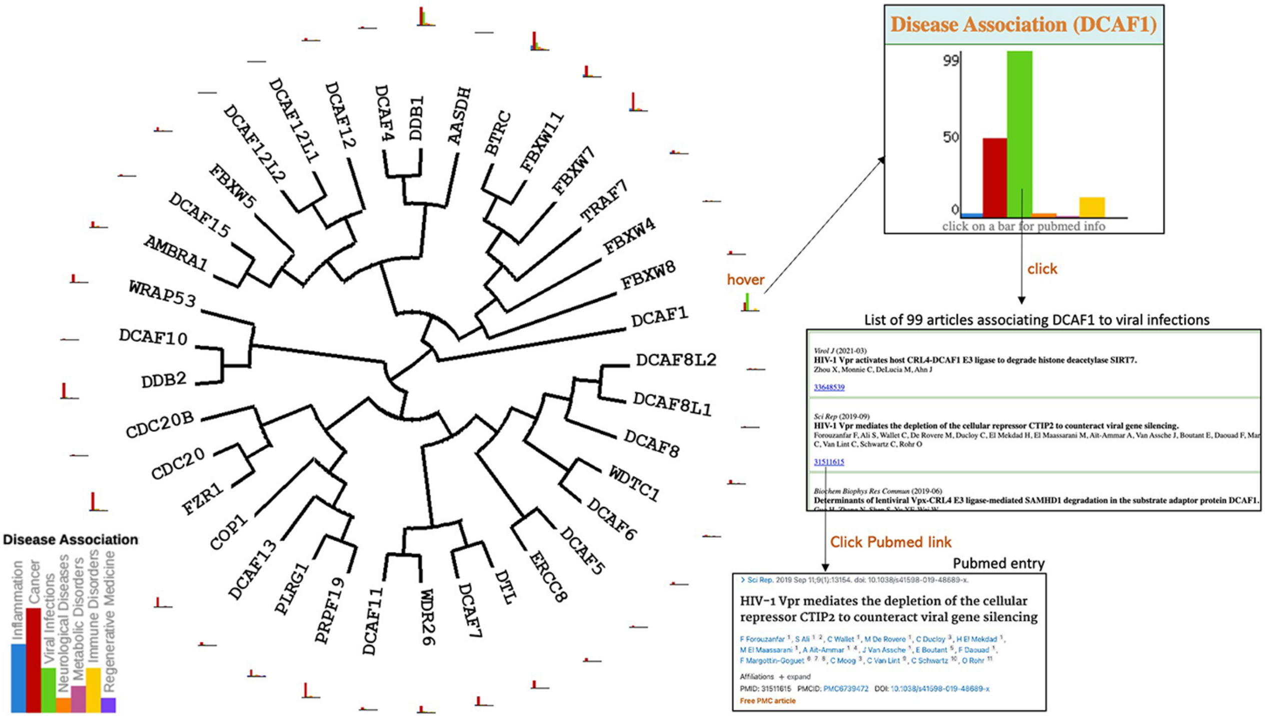
Data from Jiang et al. at Stanford University School of Medicine, based on protein levels measured by mass spectrometry across 32 normal human tissues. https://doi.org/10.1016/j.cell.2020.08.036. A protein is tissue enriched if its TS score reaches 2.5 (SD from the mean of the population distribution) in a tissue. If the TS score of a protein is greater than 4 in a tissue and is at least 1.5 higher than the protein□s TS scores in any other tissue, this protein is considered tissue specific

List of 37 E3 ligases with a WD-40 repeat domain

[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	[i]	essentiality in cancer [Broad]	essentiality in cancer [Sanger]	[i]	[i]
protein name	uniprot id	function	articles indexed in pubmed	structures in PDB	structures in PDB with bound ligands	inhibitors	number of pubmed records (Cancer)	number of pubmed records (Neuro. Disorders)	number of pubmed records (Metabolic Disorders)	number of pubmed records (Immune Disorders)	number of pubmed records (Inflammation)	number of pubmed records (Viral Infections)	essential in Human	essentiality in Human (o/e): mean (lower - upper)	number of cancer types (median score < -1)	number of cancer types (median score < -1)	tissues where protein is enriched	tissues where protein is specific
AASDH	Q4L235	Covalently bi	4				2						No	0.604 (0.446 -				
AMBRA1	Q9C0C7	Regulates aut	51				27	7	2	1	2		Yes	0.054 (0.024 -				
BTRC	Q9Y297	Substrate rec	428	8	4		212	12	6	5	22	38	Maybe	0.281 (0.179 -				
CDC20	Q12834	Required for	304	12	2		125	4		1	1	4	No	0.544 (0.363 -	20	16		
CDC20B	Q86Y33		1										No	0.766 (0.542 -				
COP1	Q8NHY2	E3 ubiquitin-	73	2			45	2	4		1		Yes	0.088 (0.042 -			Muscle	
DCAF1	Q9Y4B6	Acts both as	133	13	2		25	1	1	11	3	93	Yes	0.023 (0.007 -			Testis,	Testis,
DCAF10	Q5QP82	May function	2				3						Maybe	0.174 (0.085 -				
DCAF11	Q8TEB1	May function	15				7					1	Maybe	0.301 (0.189 -			Liver,	
DCAF12	Q5T6F0	May function	6				5			1			No	0.54 (0.35 -				
DCAF12L1	Q5VU92												Yes	0 (0 - 0.309)				
DCAF12L2	Q5VW00												Maybe	0.495 (0.258 -				
DCAF13	Q9NV06	Possible role	9	3	3		7						No	0.587 (0.412 -				
DCAF15	Q66K64	May be invol	9	7	5		7						Yes	0.105 (0.047 -				
DCAF4	Q8WV16	May function	8				5	1					Maybe	0.466 (0.302 -				
DCAF5	Q96JK2	May function	5				2						Yes	0 (0 - 0.078)				
DCAF6	Q58WW2	Ligand-deper	11				7	1					Maybe	0.246 (0.161 -			Heart Ventricle,	Muscle Skele
DCAF7	P61962	Involved in c	23				12	3			1	1	Yes	0 (0 - 0.348)				

...

Figure 1. Querying ChemBioPort. A simple menu (top) is used to produce a downloadable list of proteins (bottom) with annotations extracted from a heterogenous array of data sources. Hovering over 'information' icons [i] provides details on the data source and interpretation (top-right).



List of 99 articles associating DCAF1 to viral infections

Virol J (2021-03)
HIV-1 Vpr activates host CRL4-DCAF1 E3 ligase to degrade histone deacetylase SIRT7.
Zhou X, Monnie C, DeLucia M, Ahn J
33648539

Sci Rep (2019-09)
HIV-1 Vpr mediates the depletion of the cellular repressor CTIP2 to counteract viral gene silencing.
Forouzanfar F, Ali S, Wallet C, De Rovere M, Ducloy C, El Mekdad H, El Maassarani M, Ait-Ammar A, Van Assche J, Boutant E, Daouad F, Mar
C, Van Lint C, Schwartz C, Rohr O
31511615

Biochem Biophys Res Commun (2019-06)
Determinants of lentiviral Vpx-CRL4 E3 ligase-mediated SAMHD1 degradation in the substrate adaptor protein DCAF1.

> Sci Rep. 2019 Sep 11;9(1):13154. doi: 10.1038/s41598-019-48689-x.

HIV-1 Vpr mediates the depletion of the cellular repressor CTIP2 to counteract viral gene silencing

F Forouzanfar [1], S Ali [1 2], C Wallet [1], M De Rovere [1], C Ducloy [3], H El Mekdad [1], M El Maassarani [1], A Ait-Ammar [1 4], J Van Assche [1], E Boutant [5], F Daouad [1], F Margottin-Goguet [6 7 8], C Moog [3], C Van Lint [9], C Schwartz [10], O Rohr [11]

Affiliations + expand
PMID: 31511615 PMCID: PMC6739472 DOI: 10.1038/s41598-019-48689-x
Free PMC article

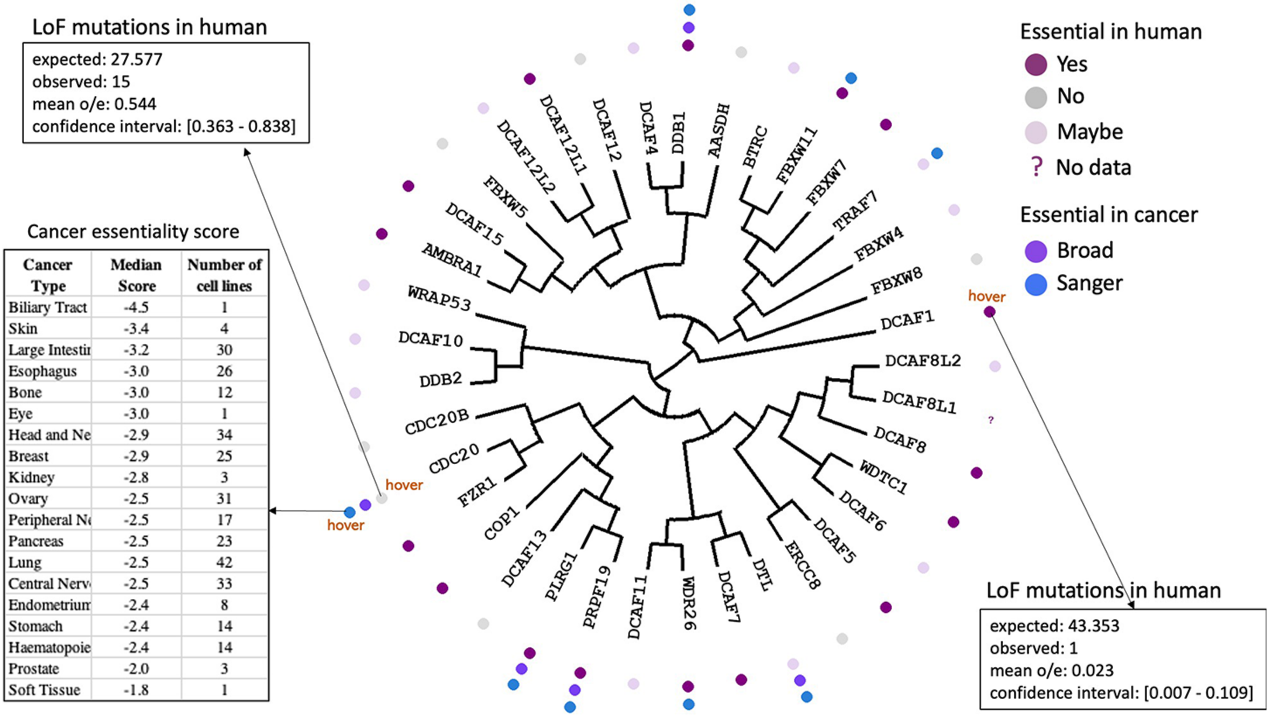
Figure 2. Disease association. Evolutionary dendograms for any table (Figure 1) are automatically generated, onto which data can be projected. Here, histograms representing the number of articles indexed in PubMed linking each gene to specific disease areas are shown. Hovering over histogram icons opens larger bar charts where bars can be clicked to produce the list of articles, each linked to their respective PubMed entry.

ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz. Protein structure coverage and structures of bound chemical ligands (molecules with six or more carbons and a molecular weight lower than 1200 Da) are obtained from the Protein Data Bank (PDB) (19). To avoid artifactual chemical inhibitors often found in the literature, we made the choice to use two curated data sources: the chemogenetics library composed of over 4000 compounds manually selected via expert crowdsourcing at Novartis (14) and the Chemical Probes portal (www.chemicalprobes.org). Gene essentiality in human is downloaded from the Broad Institute's gnomAD database (12). As recommended by gnomAD, a gene is considered essential in human if the upper bound of the oe confidence interval is <0.35. If the lower bound is >0.35, we consider the gene non-essential. If upper bound >0.35 and lower bound <0.35, the gene may be essential which we annotate as 'Maybe'. Gene essentiality in cancer is derived from CRISPR knockout screens at the Broad and Wellcome Sanger Institutes (20, 21). A gene is considered essential in a given cell line if the CRISPR score is lower than −1.0 [Broad] or its corrected log-fold change is lower than −1.0 [Sanger]. Tissue enrichment of proteins derived from mass spectrometry is obtained from the GTEx Consortium (13): a protein is enriched in a given tissue if its tissue specific (TS) score reaches 2.5 (SD from the mean of the population distribution). If the TS score of a protein is >4 in a tissue and is at least 1.5 higher than the protein's TS scores in any other tissue, this protein is considered tissue specific. Disease association derived from PubMed data and structural and chemical coverage derived from the PDB are automatically updated weekly while other data are manually updated annually.

Web interface

Any combination of queries is made on a simple online user interface to retrieve human proteins matching specific names or Uniport IDs, gene ontologies or protein domains. For instance, searching for the GO term 'ligase' and domain name 'WD40' retrieves 37 human ubiquitin ligases containing a WD40-repeat (WDR) domain (Figure 1). Heterogenous lists of protein names or uniport IDs can also be pasted in the query window.

The resulting downloadable table provides the function of each protein (as annotated in Uniprot), the number of articles indexed in PubMed, a good indicator to distinguish highly characterized from underexplored proteins, the number of structures in the PDB (users can opt to only include crystal structures, generally more reliable for structure-based drug design), the number of chemical inhibitors and the number of chemical ligands found with the protein in the PDB. Hovering over or clicking on any of these numbers opens a new window providing further information, such as PDB codes mapped onto the domain architecture of the protein or the chemical structure of ligands, all of which are illustrated in the next section. The next series of data relates to disease association, starting with the number of PubMed entries linking the protein to a specific disease area such as cancer or viral infections,



LoF mutations in human

expected: 27.577
observed: 15
mean o/e: 0.544
confidence interval: [0.363 - 0.838]

Cancer essentiality score

Cancer Type	Median Score	Number of cell lines
Biliary Tract	-4.5	1
Skin	-3.4	4
Large Intestir	-3.2	30
Esophagus	-3.0	26
Bone	-3.0	12
Eye	-3.0	1
Head and Ne	-2.9	34
Breast	-2.9	25
Kidney	-2.8	3
Ovary	-2.5	31
Peripheral N	-2.5	17
Pancreas	-2.5	23
Lung	-2.5	42
Central Nerv	-2.5	33
Endometriun	-2.4	8
Stomach	-2.4	14
Haematopoie	-2.4	14
Prostate	-2.0	3
Soft Tissue	-1.8	1

LoF mutations in human

expected: 43.353
observed: 1
mean o/e: 0.023
confidence interval: [0.007 - 0.109]

Figure 3. Essentiality in cancer and in healthy individuals. Cancer dependency data derived from CRISPR knockout screens at the Broad and Wellcome Sanger Institutes (20, 21) is mapped next to each protein name in the dendogram. Hovering over icons provides details on essentiality score and number of cell lines representing each cancer type. The essentiality of each gene in healthy individuals derived from observed versus expected LoF mutations (12) is also projected on the tree.

the essentiality of the gene in healthy humans and in cancer, as well as the list of healthy tissues where the protein is enriched or specific.

The output table can be converted on the fly into an evolutionary tree via the automated generation of a multiple sequence alignment produced by ICM (Molsoft, San Diego) (22) on our server where the default Uniprot isoform sequence for each protein is used. Alignments are converted to Newick strings with ICM, which are then represented as dendograms onto which data can be projected, as illustrated in the next section. To avoid the rate-limiting step of generating multiple sequence alignments (which can take up to 1 minutes for ensembles of 80 or more proteins), Newick strings are saved and automatically retrieved when the same query is repeated at a later time or a previously saved query ID (automatically generated for each novel query) is entered.

Application for guiding the development of chemical probes targeting E3 ligases associated with cancer and viral infection

Discovering putative therapeutic targets

Once the evolutionary tree of the 37 E3 ligases listed in Figure 1 is generated, users can easily map onto the tree any combination of data extracted from our database. For instance, histograms can be projected next to each protein name to reflect the number of articles linking the protein to specific disease areas (Figure 2). This bird's eye view clearly shows for instance that, based on the published literature, CDC20 and DCAF1 are among the most associated with cancer and viral infections respectively. Hovering over an icon brings an enlarged histogram where any bar can be clicked to produce the list of articles associating the gene to a specific disease, each with a link to the corresponding PubMed page (Figure 2).

The fact that a protein such as CDC20 is associated to cancer based on the literature does not mean that the link is causative. To further investigate this possibility, CRISPR knockout screening data from the Broad and the Wellcome Sanger Institutes (20, 21) reflecting the essentiality of a gene in cancer can be mapped with color-coded icons on the tree (Figure 3). Hovering over an icon next to a gene provides the list of cancer types, the corresponding number of cell lines and essentiality score for the gene of interest for that particular cancer type. Verifying that the data obtained from the two independent institutes is in agreement can add confidence to the result. For instance, CDC20 is one of a handful of WDR-containing E3 ligases that are essential in cancer according to both data sets.

A gene that is essential in cancer may not be a good therapeutic target if it is also essential in healthy tissues. The portal can be used to identify genes essential in the former and not the latter by projecting gene essentiality from the gnomAD database. This data is obtained by comparing the observed and expected numbers of loss-of-function (LoF) mutations in healthy individuals (12): if observed LoF mutations are significantly rarer than expected, it is likely that the gene is essential. This integrative analysis for example reveals that CDC20 (15 observed LoF mutations versus 27 expected) is the only gene in our list that is essential in cancer and not in healthy tissues (Figure 3). On the other hand, DCAF1 is essential

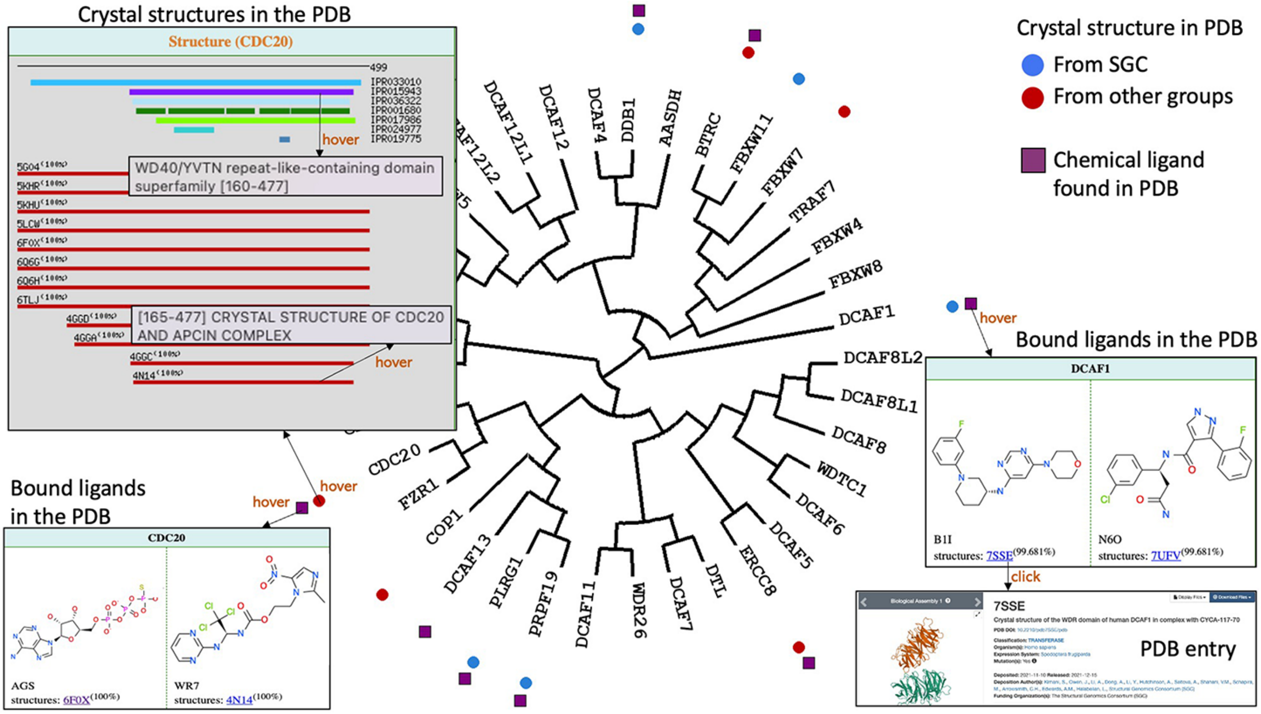


Figure 4. Structural and chemical coverage. Circular icons highlight which proteins have a crystal structure in the PDB. Hovering over an icon provides details on the domain architecture of the protein (including domain name and Interpro ID), the crystallized domains and their clickable PDB codes and titles linked to the associated PDB pages. Data on chemical ligands found in PDB structures were also mapped as square icons. Hovering brings image of the chemical structures and PDB codes.

(one LoF mutation observed versus 43 expected) and stoichiometric doses of drugs targeting this protein have higher chances of being toxic (Figure 3).

Evaluating chemical tractability

Once a functionally interesting target is identified, a natural next step is to assess its chemical tractability: are structures of the protein revealing a well-defined drug binding site? Are drug-like ligands occupying this site? ChemBioPort rapidly provides answers to these questions. First, the structural coverage of all proteins on the tree can be accessed by hovering over icons next to the protein name. A window appears, listing all structures of the protein in the PDB, and highlighting the structural domains of the protein (defined by their name and their Interpo ID) covered by each PDB code. Hovering over the PDB codes provides the PDB title. Second, structures in complex with drug-like ligands can be mapped onto the tree. Hovering over the corresponding icons displays the chemical structures and clickable PDB codes linked to corresponding pages on the PDB website.

For instance, it becomes immediately clear that a drug-like ligand is found in a structure of CDC20 with PDB code 4N14. Details on the structural coverage indicate that this structure covers the WDR domain of CDC20 (Figure 4). The same tools reveal two drug-like ligands found in the structure of DCAF1 (Figure 4).

With just a few clicks, an annotated table of all human WDR-containing E3 ligases was extracted from the database, an evolutionary dendogram was generated, and bird's eye views of disease association, structural and chemical coverage were produced. Details on the domain architecture and chemical tractability of functionally interesting proteins were rapidly accessed.

A similar integrative analysis can be conducted with any collection of human proteins, selected by name, ontology or structural domain.

Conclusion

The development of a chemical probe requires a multidisciplinary team of scientists that often need to access data outside of their area of expertise. To approach an emerging and underexplored target class, a medicinal chemist may need to intersect disease association with the structural and chemical coverages of the entire protein family; to design a target engagement assay, a cell biologist may want to understand the domain architecture of a protein; to develop a binding assay, a biochemist may need to know whether a peptide was co-crystallized with the protein; to engineer ligand selectivity, a medicinal chemist may need to know the structure of protein homologs. By collecting and integrating functional-, structural- and chemical-genomic data, we believe that ChemBioPort is a unique resource to support the rapidly expending community of scientists engaged in the chemical coverage of the human genome (8, 9).

Acknowledgement

The algorithm used to generate phylogenetic trees from Newick strings was originally written by Xi Ting Zhen.

Funding

The Structural Genomics Consortium is a registered charity (no. 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol Myers Squibb, Genentech, Genome Canada through Ontario Genomics Institute (OGI-196), EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medi cines Initiative 2 Joint Undertaking (EUbOPEN grant no. 875510), Janssen, Merck KGaA (aka EMD in Canada and USA), Pfizer and Takeda. MS gratefully acknowledges support from NSERC (grant no. RGPIN-2019-04416), CQDM (Quantum Leap-176) and MITACS accelerate (IT13051).

Conflict of interest

None declared.

Data availability

All data is freely accessible at https://chembioport.thesgc.org/.

References

1. Frye,S.V. (2010) The art of the chemical probe. *Nat. Chem. Biol.*, **6**, 159–161.
2. Bunnage,M.E., Chekler,E.L.P. and Jones,L.H. (2013) Target validation using chemical probes. *Nat. Chem. Biol.*, **9**, 195–199.
3. Arrowsmith,C.H., Audia,J.E., Austin,C. *et al.* (2015) The promise and peril of chemical probes. *Nat. Chem. Biol.*, **11**, 536–541.
4. Fedorov,O., Müller,S. and Knapp,S. (2010) The (un)targeted cancer kinome. *Nat. Chem. Biol.*, **6**, 166–169.
5. Wu,Q., Heidenreich,D., Zhou,S. *et al.* (2019) A chemical toolbox for the study of bromodomains and epigenetic signaling. *Nat. Commun.*, **10**, 1915.
6. Scheer,S., Ackloo,S., Medina,T.S. *et al.* (2019) A chemical biology toolbox to study protein methyltransferases and epigenetic signaling. *Nat. Commun.*, **10**, 19.
7. Casiraghi,A., Bensimon,A. and Superti-Furga,G. (2021) Recent developments in ligands and chemical probes targeting solute carrier transporters. *Curr. Opin. Chem. Biol.*, **62**, 53–63.
8. Carter,A.J., Kraemer,O., Zwick,M. *et al.* (2019) Target 2035: probing the human proteome. *Drug Discov. Today*, **24**, 2111–2115.
9. Müller,S., Ackloo,S., Al Chawaf,A. *et al.* (2022) Target 2035—update on the quest for a probe for every protein. *RSC Med. Chem.*, **13**, 13–21.
10. Liu,L., Zhen,X.T., Denton,E. *et al.* (2012) ChromoHub: a data hub for navigators of chromatin-mediated signalling. *Bioinformatics*, **28**, 2205–2206.
11. Liu,L., Damerell,D.R. and Koukouflis,L. (2019) UbiHub: a data hub for the explorers of ubiquitination pathways. *Bioinformatics*, **35**, 2882–2884.
12. Karczewski,K.J., Francioli,L.C., Tiao,G. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
13. Jiang,L., Wang,M., Lin,S. *et al.* (2020) A quantitative proteome map of the human body. *Cell*, **183**, 269–283.e19.
14. Canham,S.M., Wang,Y., Cornett,A. *et al.* (2020) Systematic chemogenetic library assembly. *Cell Chem. Biol.*, **27**, 1124–1129.
15. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
16. Carbon,S., Douglass,E. and Good,B.M. Gene Ontology Consortium. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.

17. Blum,M., Chang,H.-Y., Chuguransky,S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
18. Bateman,A., Martin,M.-J. and Orchard,S. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
19. Berman,H.M., Westbrook,J., Feng,Z. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
20. Meyers,R.M., Bryan,J.G., McFarland,J.M. *et al.* (2017) Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.*, **49**, 1779–1784.
21. Behan,F.M., Iorio,F., Picco,G. *et al.* (2019) Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature*, **568**, 511–516.
22. Abagyan,R.A. and Batalov,S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.