Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D995–D1002***
doi: 10.1093/nar/gkw1072

Pharos: Collating protein information to shed light on the druggable genome

Dac-Trung Nguyen[1,†], Stephen Mathias[3,†], Cristian Bologa[3], Soren Brunak[8], Nicolas Fernandez[2], Anna Gaulton[9], Anne Hersey[9], Jayme Holmes[3], Lars Juhl Jensen[8], Anneli Karlsson[12], Guixia Liu[3,10], Avi Ma'ayan[2], Geetha Mandava[1], Subramani Mani[3], Saurabh Mehta[5,6], John Overington[12], Juhee Patel[3,11], Andrew D. Rouillard[2], Stephan Schürer[5,7], Timothy Sheils[1], Anton Simeonov[1], Larry A. Sklar[3,4], Noel Southall[1], Oleg Ursu[3], Dusica Vidovic[5,7], Anna Waller[3,4], Jeremy Yang[3], Ajit Jadhav[1], Tudor I. Oprea[3,*] and Rajarshi Guha[1,*]

[1]National Center for Advancing Translational Science, 9800 Medical Center Drive, Rockville, MD 20850, USA, [2]Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA, [3]Translational Informatics Division, School of Medicine, University of New Mexico, Albuquerque, NM 87131, USA, [4]Center for Molecular Discovery, University of New Mexico Cancer Center, University of New Mexico, Albuquerque, NM 87131, USA, [5]Center for Computational Science, University of Miami, Coral Gables, FL 33146, USA, [6]Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India, [7]Department of Molecular and Cellular Pharmacology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA, [8]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark, [9]European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [10]East China University of Science and Technology, Dept. Pharmaceutical Sciences, Shanghai, China, [11]BA/MD Program, School of Medicine, University of New Mexico, Albuquerque, NM 87131, USA and [12]BenevolentAI, 40 Churchway, London NW1 1LW, UK

Received September 15, 2016; Revised October 17, 2016; Editorial Decision October 24, 2016; Accepted October 24, 2016

ABSTRACT

The 'druggable genome' encompasses several protein families, but only a subset of targets within them have attracted significant research attention and thus have information about them publicly available. The Illuminating the Druggable Genome (IDG) program was initiated in 2014, has the goal of developing experimental techniques and a Knowledge Management Center (KMC) that would collect and organize information about protein targets from four families, representing the most common druggable targets with an emphasis on understudied proteins. Here, we describe two resources developed by the KMC: the Target Central Resource Database (TCRD) which collates many heterogeneous gene/protein datasets and Pharos (https://pharos.nih.gov), a multimodal web interface that presents the data from TCRD. We briefly describe the types and sources of data considered by the KMC and then highlight features of the Pharos interface designed to enable intuitive access to the IDG knowledgebase. The aim of Pharos is to encourage 'serendipitous browsing', whereby related, relevant information is made easily discoverable. We conclude by describing two use cases that highlight the utility of Pharos and TCRD.

INTRODUCTION

In 2014, the National Institutes of Health (NIH) initiated the Illuminating the Druggable Genome (IDG) program (https://commonfund.nih.gov/idg/index). The goal of the IDG program is to shed light on poorly characterized proteins that can potentially be modulated using small molecules or biologics. The program comes at a time when genomic information suggests that at least 3000 gene coded proteins can be 'drugged', yet only 10% of these potential

*To whom correspondence should be addressed. Tel: +1 814 404 5449; Fax: +1 301 217 5736; Email: guhar@mail.nih.gov
Correspondence may also be addressed to Tudor Oprea. Tel: +1 505 925 4756; Fax: +1 505 925 7625; Email: toprea@salud.unm.edu
†These authors contributed equally to this work as first authors.

targets have an FDA approved drug (1). From the point of view of funded research, *Edwards et al.* (2) reported a bibliometric analysis indicating that 75% of research is focused on studying only 10% of the known mammalian proteins. Based on data that we accumulated to develop the Target Central Resource Database (TCRD), during the period 2011–2015, the NIH funded 270 491 R01 project grants to study 7934 targets and just 11 of which (0.14%) of the 7934 targets considered during 2011–2015 accounted for 10% of the R01's funded. There are multiple reasons for having understudied, or even unstudied targets and some of which are discussed in Edwards *et al.* (2). We refer to these unstudied proteins as 'dark'.

Clearly, there is a need to be able to access comprehensive, diverse data about protein targets and present such data in a manner that can be used to shed light on potential dark targets. To achieve these goals, the IDG initiated the Knowledge Management Center (KMC) which was initially tasked with collating and disseminating data on approximately 1700 targets from the four families enriched for existing drug targets: ion channels, nuclear receptors, GPCRs and kinases. However, current efforts have gone beyond these four families, to consider all ∼20 000 human protein targets, motivated by the opportunity to expand what is considered druggable (3). These efforts have culminated in the Target Central Resource Database (TCRD), an integrated database of diverse data sources and data types and a multimodal web based platform called Pharos, to disseminate and explore the data within TCRD. These resources allow researchers to explore all data around dark targets in the context of well-studied targets

There currently exist a number of resources that have aggregated data around genes or protein targets. For example, GeneCards (4) and UniProt (5) are comprehensive resources on genes and protein targets respectively, that aggregate a wide variety of information, with the former including extensive links to commercially available tools (e.g. antibodies) to probe targets. While information on antibodies and other tools are collected in TCRD, it goes beyond to include downstream data types such as mouse phenotype information (http://www.mousephenotype.org/) and GWAS (https://www.ebi.ac.uk/gwas/) data. Furthermore, Pharos attempts to present these varied datatypes in a comprehensive, linked fashion, rather than simply displaying individual data types independently. A recently released resource that is somewhat similar in nature to the current work is OpenTargets (https://www.opentargets.org/). However, the scope of OpenTargets is primarily to enable disease specific target validation, as opposed to broadly collating knowledge about all targets. Another resource focusing on the druggable genome is DGIdb (6), a database that collects drug-gene interactions. By definition this resource focuses on well-studied targets and thus does not address the specific challenge of dark targets catalogued *via* the IDG program. DrugBank (7) and the Therapeutic Target Database (8) also aggregate data for protein targets, but their primary focus is the targets of drugs and thus by definition do not contain information on understudied or unstudied targets, for which small molecule probes may not be available.

The current paper describes the Pharos platform that presents the contents of the TCRD. In the following sections, we describe the overall architecture, the data sources considered in the TCRD and the user interface features implemented in the Pharos platform.

MATERIALS AND METHODS

TCRD is the central data repository for the IDG KMC and TCRD is the primary data source for the IDG-KMC project-wide web portal Pharos. The TCRD integrates diverse datasets, using well-defined workflows that employ source APIs, relevant to human genes and proteins and also serves as a platform for data integration and analytics. The Pharos application is the interface to the TCRD data and provides both a HTML user interface along with a REST API. TCRD releases are imported into a local database for pre-processing (which primarily focuses on transforming, indexing and linking different data types to enable rapid retrieval) and then displayed by Pharos. For an example of data transformations for tissue expression data see Supplementary Information. While all TCRD data are available *via* the Pharos application, users wishing to work with the original, unprocessed form of the TCRD database can access it from http://juniper.health.unm.edu/tcrd/. An ER diagram of the TCRD database is available in Supplementary Figure S1 and licensing information for individual data sources contained within the TCRD are available in the Supplementary Table S1). Source code for the Pharos platform is available, under the MIT license, from https://spotlite.nih.gov/ncats/pharos.

Data sources

The datasets in TCRD comprise of a wide array of knowledge and data types about genes, proteins and small molecules collected and processed from numerous resources. It includes text-mined bibliometric associations and statistics from the biomedical and patent literature, mRNA and protein expression data, disease and phenotype associations, bioactivity data, drug target interactions, and processed datasets about the functions of genes and proteins from 66 resources organized into 114 datasets imported from the Harmonizome (9). TCRD also makes use of existing biological ontologies, which we integrated to construct the bespoke Drug Target Ontology (DTO, http://drugtargetontology.org). The full list of data sources is included in Supplementary Table S2.

Target classifications

Based on the data collected for each target, the KMC has constructed a high level classification scheme, termed the Target Development Level (TDL). TDL characterizes the degree to which they are studied or not studied, as evidenced by publications, tool compounds and other features. The TDL scheme serves as the primary grouping of targets, clearly delineating those targets that are unstudied (labeled Tdark) from those that have more information about them (labeled Tclin, if associated with approved drugs with known mechanism of action (10), Tchem, if associated with small molecule activities in ChEMBL or Tbio if not associated with small molecule or drug activities but have a GO

MF or BP leaf term annotated or else have a confirmed OMIM phenotype). DrugCentral (11) aggregates target-disease information, drug target bioactivity data, which are used to categorize Tclin and Tchem, and feeds into TCRD. See http://juniper.health.unm.edu/tcrd/ for a more in depth description of the TDL classification scheme. Along with the TDL scheme, we have employed DTO to support a formal classification and annotation of the IDG protein families, building on top of prior classification schemes for kinases (13), GPCRs (12–14), ion channels (15) and nuclear receptors (13). Though the DTO, being an ontology, allows for sophisticated inferencing and hypothesis generation, Pharos currently employs the DTO primarily as a simple classification scheme to complement the TDL categories.

RESULTS

Presentation and usability features

The Pharos platform is designed to be broadly applicable and of use to both computational and non-computational scientists. The platform focuses on three classes of users: biologists and clinical researchers (with an interest in characterizing and validating novel targets and identifying key small molecules or biologics), funding agencies (with an interest in exploring the research landscape so as to generate new ideas for research funding and direction) and finally computational scientists (with an interest in data mining and supporting target validation projects). Thus Pharos provides a REST API (https://pharos.nih.gov/idg/api/v1) that supports programmatic access to search functionality and all data contained within TCRD. The API is designed to be self-describing and responses are made in JSON format. The API is of primary interest to computational scientists and developers building novel applications on top of it. However, we anticipate that the most common interaction is *via* the web interface. Hence we focus on a description of features implemented in the web interface that enhance usability and exploration of the knowledgebase.

Search functionality

As noted above, Pharos ingests a TCRD release and performs a pre-processing step prior to data display. The pre-processing step primarily focuses on linking or transforming a number of data types to allow for rapid retrieval and visualization. A key pre-processing step is indexing the relevant fields for a given entity (i.e. target, disease and compound) to support free text search, autosuggest and complex filtering functionality. The combination of free-text search and filters allows for easy drill down when faced with large result sets. Text search is enhanced by the availability of autocomplete suggestions, grouped by categories. An example of this behavior is shown in Figure 1A. The autocomplete feature is designed to be the primary entry point for exploring target data and for hypothesis generation. In addition to text search, sequence similarity search allows the user to paste in an amino acid sequence and identify targets with a similarity greater than a user-specified cutoff. Finally, a batch search function is also available, that allows a user to paste in multiple gene symbols or protein accession codes and retrieve their records in one go.

Most searches (including general text searches) will return hits for targets, publications, ligands and diseases. The user interface is designed to support intuitive drill down into the hits within each of these types of entities, with a particular focus on protein targets. This is enabled using faceted filters that support easy construction of complex filtering rules. Figure 1C is a screenshot of the main entry point to the list of targets obtained *via* a search or by browsing all available targets.

The filter panel on the left hand side consists of 5 filters that we consider the most commonly used. Selecting a filter automatically filters the list of targets, and multiple filters are combined using logical AND. The filters also include the count of entities that match a given filter value, and when selected displays the number of matching entities (which may be different from the first number due to the inclusion of other filters). Pharos uses 51 filters that include ontology terms (e.g. GO (16), Disease Ontology (17), DTO and Panther (18)), NIH grant types and counts, tissue expression data, pathway relationships and so on. Combining filters allows one to construct sophisticated queries. For example, identifying multiple targets associated with two or more diseases, could lead to co-morbidity hypotheses (19). The list of filters can be filtered using text search and complex filter combinations. These settings can be saved by simply bookmarking the URL. This enables easy sharing of specific searches between users.

All data viewable in the interface are available for download both for individual targets as well as multiple targets. The data are made available in the form of multiple CSV formatted files contained in a single ZIP archive, with metadata describing columns included as a text file within the archive.

As one of the goals of the IDG KMC was to organize data on unstudied and understudied targets, the notion of a *target dossier* (similar in concept to the OpenPHACTS consortium target dossier, http://td.inab.org/) was developed to allow a user to collect data as they browsed the database. The dossier is analogous to an e-commerce shopping cart and allows a user to collect targets, diseases and publications as they continue browsing. The dossier functionality supports multiple dossiers, allowing the user to collect information for separate purposes, e.g., different projects. Data associated with the entities in any given dossier can be downloaded as on the main interface. Similarly, all visualization tools available on the main interface can be applied to the entities contained within a dossier.

A common task when exploring understudied targets is to compare the data available around them to other targets. In particular, comparison to targets in the same family could be useful in understanding whether more resources should be expended on illuminating the understudied ones. While we expect that an in-depth analysis will be performed using custom tools and data exported from Pharos, the user interface supports visual (side by side) target comparison of two or more targets (e.g. https://pharos.nih.gov/idg/targets/compare?q=Q05586,Q9UBN1)

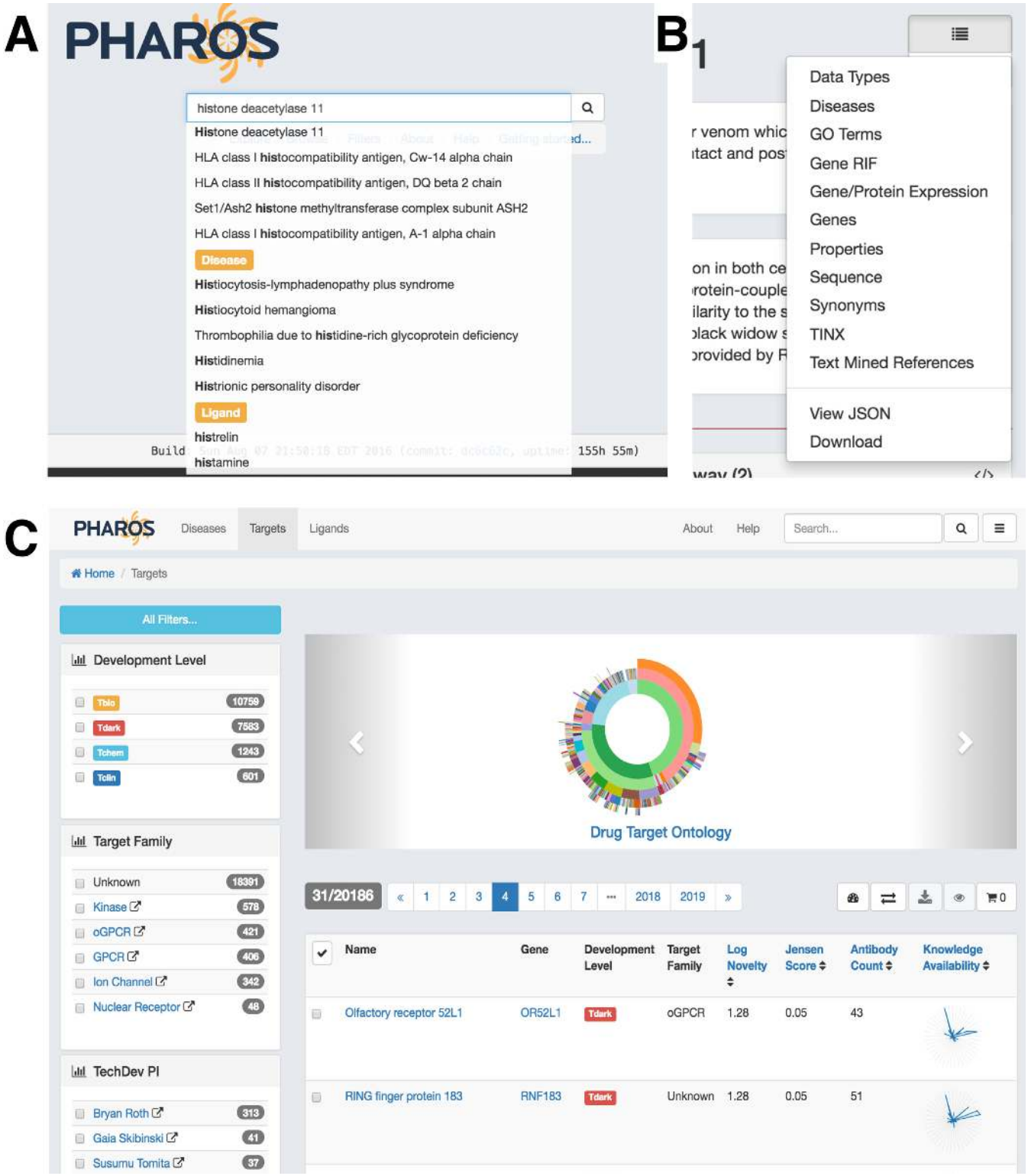

Figure 1. Examples of Pharos UI elements designed to enhance usability and encourage serendipitous discovery. (**A**) categorized autosuggest functionality available in free text searches. (**B**) The Table of Contents widget, on a target detail page, that provides the user with an overview of the data types available for the target being viewed and allows direct navigation to individual data types. In addition, the widget supports of all data for the current target and viewing the JSON representation for this target available from the underlying API. (**C**) A screenshot of the target list view that is obtained either *via* free text search or by browsing the entire set of targets.

Target detail pages

All information about a target is accessible via individual pages. The goal of these pages is to display all data that have been collected by the KMC about the given target. Because of the wide variety of data types that are collected, these pages can be quite large. To enhance usability each page provides a table of contents (Figure 1B) that enables the user to directly jump to data types of interest or download all the available data for the current target. Individual data types are represented as panels, with link-outs or visualizations depending on the nature of the datatypes (e.g., tissue expression data as a color coded homunculus, Supplementary Information, Supplementary Figure S3). In contrast, for publications associated with a target, a list of them is provided in a table, but in addition, a summary using a word cloud generated from the abstracts is presented. For many data types such as grant applications or GO terms, the user interface enables using that data to perform a new search, allowing for easy (even serendipitous) exploration of the IDG target space.

Data visualization components

Depending on the data type, Pharos implements a number of visualizations throughout the interface. For example, he target list view employs visualizations including radial pie charts, word clouds and sunburst (20) diagrams depending on whether the data type is categorical (e.g. target class or target family), textual (GO or Uniprot keywords) or hierarchical (Drug Target Ontology or PANTHER classification). In the target list view these visualizations act as filters—clicking on a given pie segment in the Target Family visual, for example, will filter the current list of targets to match the selected target family.

An important aspect of the work done by the IDG KMC is to collect and process a wide variety of heterogeneous datasets that describe the properties and functions of genes and proteins. A key resource that was developed by the IDG KMC is a simplified uniform representation of knowledge about genes and proteins. This project is called the Harmonizome (9). The Harmonizome datasets provide numeric representation of 72 million associations between all mammalian genes and their attributes collected from 66 open online major resources. Using metadata associated with each data source, we summarized knowledge around a target, by aggregating 114 datasets from the Harmonizome into 41 sub-groups and visualizing this as a radar chart. When displayed in a column in the table of targets, the plots provide a visual summary about the amount and type of knowledge that is available about the target. Importantly, this allows the user to scan the table to examine the shape of the radar plots for each target. In other words, the radar plots that look similar (Figures 1C and 2A) imply that the corresponding targets have similar data types associated with them. Individual radar plots can be expanded, and then explored using different aggregation schemes, as well as overlays combining the radar plots for several targets, or groups of targets (Figure 2B).

An additional visualization of Harmonizome data is by the use of the harmonogram (9). This is essentially a heatmap representation of the cumulative probabilities for the target/dataset associations. The data sources are presented on the Y-axis and the targets on the X-axis. The visualization in Pharos is interactive allowing zooming and selections. Figure 3 displays harmonograms for two sets of targets. Figure 3A corresponds to kinase targets with a Tclin classification (i.e. relatively well studied). This is evidenced by a heatmap that is largely populated, with high cumulative probability values. On the other hand, Figure 3B is the harmonogram for GPCRs with a Tdark classification. It is evident that there are much fewer data associations for this target set (grey representing no data). More specifically, the bands of grey represent *holes* in the knowledge space for this set of targets. The interactive visualization enables grouping of data sources by their type (e.g. genomic data sources or chemical data sources) allowing the user to easily identify targets for which specific types of data may be missing (or poorly populated).

We refer the reader to Supplementary Information for a discussion of other visualization components available in Pharos.

Ranking targets

The interface allows one to rank targets using a variety of parameters including the novelty score (a measure of the extent to which the published literature refers to the target) and the PubMed Score (described at https://pharos.nih.gov/idg/pmscore). We have also implemented a ranking scheme based on Harmonizome data. Specifically, for each target we compute the sum of the cumulative probabilities across all 114 data sources captured in the Harmonizome and defined this as the Data Availability Score (DAS). Clicking on the radar chart column header allows one to sort the targets based on their total knowledge availability as represented by the DAS. It is important to realize that target ranking based on individual parameters is only a first step in target prioritization. While there are examples of target prioritization using individual parameters such as GO or DO terms (21), in general, target prioritization is heavily contextual, where the context could be a disease state or a biological process.

Use cases

The wide variety of data presented via Pharos supports multiple modes of interaction, ranging from guided browsing to direct access to specific target pages. We describe two use cases that highlight the role that Pharos could play in enabling research on understudied targets.

Novel targets that may play a role in obesity. A number of targets play an important role in the regulation of food intake and dysregulation of these targets can lead to a variety of metabolic disorders and play a role in obesity (22). We can start from the Target view (https://pharos.nih.gov/idg/targets) and use the *Disease* filter to search for targets associated with 'Obesity'. This gives 432 targets, which we can further filter down by using the *GWAS Trait* filter (selecting 'Obesity'). This leaves 18 targets of which 15 do not belong to the IDG families. Thus we focus on the GPCR, ion channel and kinase targets. At this stage we can download data on these targets, or else save them to a new dossier titled

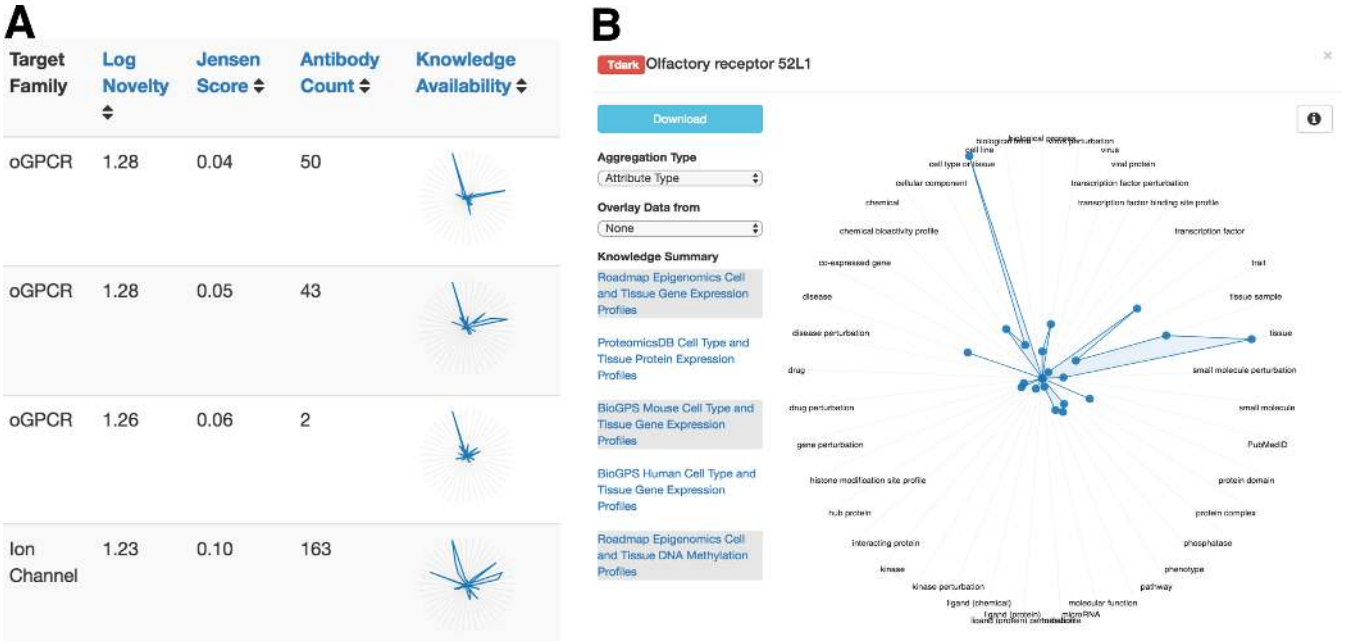

Figure 2. The use of radar charts to summarize data availability for multiple targets schematically (**A**) or in interactive detail for a single target (**B**). The data underlying all radar charts are obtained from the Harmonizome resource (9) and can be downloaded from Pharos for individual targets.

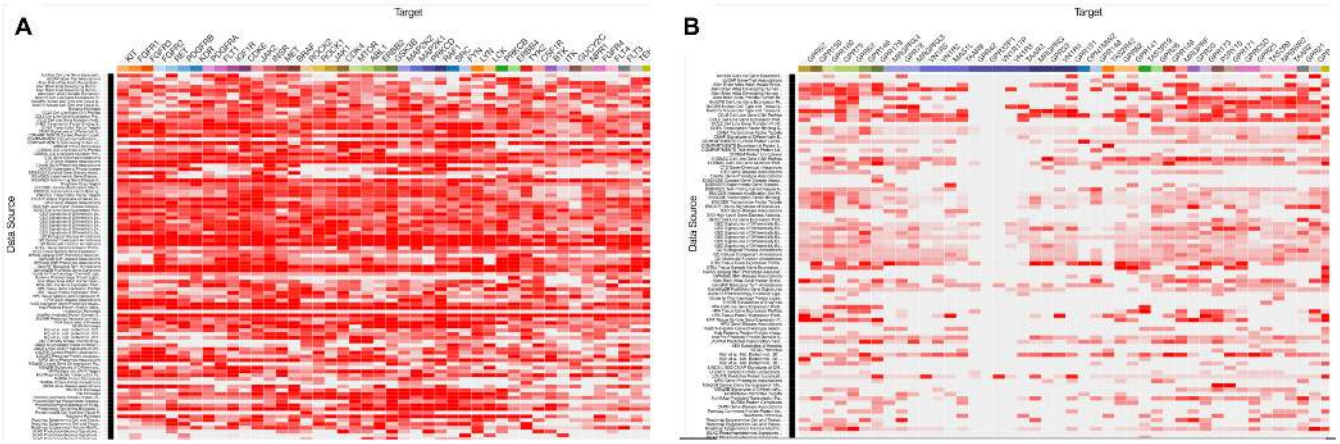

Figure 3. Harmonograms—heatmaps of the cumulative probability of association between a target and a Harmonizome data source, for two target sets. (**A**) Kinase targets with a Tclin classification. (**B**) GPCRs with a Tdark classification. Brighter red indicates more data associated with a target for a given data source and grey represents no data associated with a target in that data source.

'Obesity Targets'. In parallel, we can view the data availability around these targets by generating a harmonogram, which would highlight that KCNMA1 is well studied, experimentally, whereas ALPK1 is somewhat sparser. Focusing on ALPK1, we see from the target detail page that it is part of 15 funded grants. We can then rerun the search using grant 5R01NS044385-12 as the query and identify the targets studied in it, associated with obesity (via the *Disease* filter). As expected this includes ALPK1, but also identifies KIF7, which was not in our initial search results (since it did not belong to the GPCR, kinase or ion channel families). Given that KIF7 is under study, it may be useful for further investigation and thus could be added to the 'Obesity Targets' dossier. At any point it is also possible to identify other diseases that targets are associated with (such as gout for ALPK1) and then explore data associated with those targets, saving items of interest to the dossier for later study.

Identify diseases & researchers that are related directly or indirectly to nociception targets. Targets involved in nociception are spread out amongst multiple protein families. We consider a user who is interested in diseases (and their associated targets) that are related to nociception (or comorbid with diseases related to nociception). The starting point would be a text query for 'nociception', which will generate a result set of 77 targets, 1 disease and 30 publications. The user could simply focus on the identified disease (Neuropathy, hereditary sensory and autonomic) and stop there. However, it is useful to explore what diseases are associated with the 77 targets. To contrast established and

novel targets, the results could be reduced to focus on the 32 Tclin and Tdark targets via the *Development Level* filter. At this point the *Disease* filter will list diseases associated with this subset. These include ones that are clearly related to nociception (such as pain agnosia and neuralgia) but also others that are not obviously related (such as cancers and a number of psychiatric disorders). The user could focus on one or more of these diseases and explore the targets associated with them (possibly saving these in a custom dossier). To identify researchers, the user can employ the *TechDev PI* filter to identify IDG-funded researchers working on any of the targets. The user could drill down to the specific targets being actively studied and from the interface get in touch with the lab conducting experiments. In parallel, using the *R01 Count* filter the research could select targets for which there are multiple grants funded and then explore the targets being studied as part of those grants and jump out to NIH RePORTER (https://projectreporter.nih.gov/) to get further details on who is studying these targets.

DISCUSSION

Given background information from TCRD, Pharos serves as entry point into the druggable genome initially envisaged by the IDG program, but has gone beyond the initial set of ∼1700 targets to incorporate the entire human proteome. As a result, users now have a much richer contextual space within which data on understudied targets may be considered. Given the wide variety of data types collected by TCRD, effective access and presentation via Pharos enables users to find what they want, but also point users in the direction of related, possibly relevant information that they may not have considered initially.

Ongoing work focuses on incorporating more data types into TCRD (in particular epigenomic and metabolomic data) and expanding on the some of the current data types. For example, grant funding data has been very useful to identify research 'hot spots' and inclusion of health economic data would provide a complementary view of which targets are currently of interest versus those in which interest is growing. Other efforts include better highlighting of provenance (why and where did something match a search query), target prioritization (via similarity searches and temporal analysis of appropriate data sources to identify 'rising targets') and using the semantic capabilities of the DTO.

In conclusion, the Pharos platform is designed to allow efficient exploration of the currently defined druggable genome, with the ability to go beyond this pre-defined subset of targets. Together with integration of experimental results from the IDG funded Technology Development groups, this platform will support research scientists wishing to understand the knowledge landscape around the druggable genome, with the hope of shedding light on the dark corners thereby expanding what is considered druggable.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Susumu Tomita and members from his lab for providing valuable feedback on the Pharos interface during an interactive demonstration. We would also like to thank Gaia Skibinski, Michael McManus and Jing-Ruey Joanna Yeh for useful suggestions on improving usability of the interface. We are grateful to Christian Stolte for permission to use the homunculus visualization for tissue expression.

FUNDING

National Institutes of Health (NIH) Common Fund [CA189205 to T.O.]; NIH Common Fund [CA189201 to A.M.]; Novo Nordisk Foundation [NNF14CC0001 to S.B. and L.J.J.]. Intramural Research Program, NCATS [DTN, GM,TS, AS, NS, AJ, RG]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
2. Edwards,A.M., Isserlin,R., Bader,G.D., Frye,S.V., Willson,T.M. and Yu,F.H. (2011) Too many roads not taken. *Nature*, **470**, 163–165.
3. Makley,L.N. and Gestwicki,J.E. (2013) Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. *Chem. Biol. Drug Des.*, **81**, 22–32.
4. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Stein,T.I., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, **54**, doi:10.1002/cpbi.5.
5. Bateman,A., Martin,M.J., O'Donovan,C., Magrane,M., Apweiler,R., Alpi,E., Antunes,R., Ar-Ganiska,J., Bely,B., Bingley,M. *et al.* (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
6. Wagner,A.H., Coffman,A.C., Ainscough,B.J., Spies,N.C., Skidmore,Z.L., Campbell,K.M., Krysiak,K., Pan,D., McMichael,J.F., Eldred,J.M. *et al.* (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.
7. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y.F., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*, **42**, D1091–D1097.
8. Yang,H., Hong,Y., Chu,Q., Li,Y.H., Lin,T., Jin,Z., Yu,C.Y., Feng,X., Zhe,C., Feng,Z. *et al.* (2015) Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.*, **44**, D1069–D1074.
9. Rouillard,A.D., Gundersen,G.W., Fernandez,N.F., Wang,Z., Monteiro,C.D., McDermott,M.G. and Ma'ayan,A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**, doi:10.1093/database/baw100.
10. Santos,R., Ursu,O., Gaulton,A., Bento,A.P., Donadi,R.S., Bologa,C., Karlsson,A., Al-Lazikani,B., Hersey,A., Oprea,T.I. *et al.* (2017) A comprehensive Map of molecular drug targets. *Nat. Rev. Drug Discov.*, doi:10.1038/nrd.2016.230.
11. Ursu,O., Holmes,J., Knockel,J., Bologa,C.G., Yang,J.J., Mathias,S.L., Nelson,S.J. and Oprea,T.I. (2017) DrugCentral: online drug compendium. *Nucleic Acids Res.*, doi:10.1093/nar/gkw993.
12. Munk,C., Isberg,V., Mordalski,S., Harpsoe,K., Rataj,K., Hauser,A.S., Kolb,P., Bojarski,A.J., Vriend,G. and Gloriam,D.E. (2016) GPCRdb: the G protein-coupled receptor database - an introduction. *Br. J. Pharmacol.*, **173**, 2195–2207.
13. Southan,C., Sharman,J.L., Benson,H.E., Faccenda,E., Pawson,A.J., Alexander,S.P.H., Buneman,O.P., Davenport,A.P., McGrath,J.C.,

Peters,J.A. *et al.* (2016) The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.*, **44**, D1054–D1068.
14. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
15. Saier,M.H., Reddy,V.S., Tsu,B.V., Ahmed,M.S., Li,C. and Moreno-Hagelsieb,G. (2016) The transporter classification database (TCDB): recent advances. *Nucleic Acids Res.*, **44**, D372–D379.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
17. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.-W.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
18. Mi,H., Poudel,S., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.
19. Hu,J.X., Thomas,C.E. and Brunak,S. (2016) Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.*, **17**, 615–629.
20. Stasko,J. and Zhang,E. (2000) *InfoVis'00*. IEEE, pp. 57–65.
21. Chen,Y.-A., Tripathi,L.P. and Mizuguchi,K. (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, **6**, e17844.
22. Xu,Y.L., Jackson,V.R. and Civelli,O. (2004) Orphan G protein-coupled receptors and obesity. *Eur. J. Pharmacol.*, **500**, 243–253.