Project SCANCER: An Open-access Drug Repurposing and Data-mining Platform to Enhance Target Validation and Optimize International Research Efforts against Highly Progressive Cancers

David Dora (✉ dora.david@med.semmelweis-univ.hu)

Semmelweis University: Semmelweis Egyetem <https://orcid.org/0000-0002-3138-8816>

Timea Dora

Budapest University of Technology and Economics: Budapesti Muszaki es Gazdasagtudomanyi Egyetem

Gabor Szegvari

National Koranyi Institute for Tuberculosis and Pulmonology: Orszagos Koranyi Tbc es Pulmonologiai Intezet

Csongor Gerdán

National Koranyi Institute for Tuberculosis and Pulmonology: Orszagos Koranyi Tbc es Pulmonologiai Intezet

Zoltan Lohinai

National Koranyi Institute for Tuberculosis and Pulmonology: Orszagos Koranyi Tbc es Pulmonologiai Intezet

Software article

Keywords: drug-repurposing, cancer research, drug database, data mining, lung cancer

Posted Date: January 31st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1257157/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The expanding body of potential therapeutic targets requires easily accessible, structured, and transparent real-time interpretation of molecular data. We invite research teams to join and aim to enhance the cooperative work of more experienced groups to harmonize international efforts to overcome less prevalent, devastating malignancies. We show how we integrated our previous expertise in small-cell lung cancer (SCLC) with data mining approaches to initiate a new platform to overcome highly progressive cancers such as triple-negative breast and pancreatic cancer. Here, we integrate available fundamental data and present a novel, open access, data-mining, drug repurposing platform, deriving our searches from the entries of Clue.io. Through the front end, users can create their drug-target list to select the clinically most relevant targets for further functional validation assays or drug trials. SCANCER integrates searches from publicly available databases, such as PubChem, Drug Bank, PubMed, and EMA; and complementing them with background information on compounds using entities like UniProt, String, and genecards. Cancer drug discovery requires a convergence of complex, often disparate fields. We present a simple, transparent, and user-friendly drug repurposing database.

Introduction

There is an increasing need for open-access drug repurposing databases for researchers in the translational and clinical field due to the emergence of many potential new therapeutic targets every year. This trend will likely continue in the future with exponentially increasing repositories full of data, and data mining requires particular expertise and qualifications that cannot be expected from a single group of researchers. The development of novel pharmaceuticals takes enormous effort and immense financial and human resources, a decade of research and clinical trials until approval with more advancements in prevalent cancers. Research groups usually select their target and related research directions without cross-optimization of efforts across individual groups. Industry AI-based drug developments have limitations and also require preclinical validation. Therefore, *in silico* approaches with solid human resources might assist than fully overtake all steps.

The therapeutic potential of a molecular target, the suitable inhibitor (or agonist), might already be available in another indication, or a small molecule lead compound is available that needs further modification and preclinical testing. Nowadays, a wide selection of databases is available for various goals in drug repurposing (Parvathaneni et al., 2019; Gns et al., 2019; Tanoli et al., 2021). Omics data about molecular targets can be retrieved from *Uniprot* (UniProt Consortium. 2021), *Genecards* (Stelzer et al 2016), *TTD* (Chen et al, 2002), *STITCH* (Kuhn et al, 2008), *BioGRID* (Oughtred et al, 2020), and *STRING* (Szklarczyk et al, 2019); or corresponding pathways from *KEGG* (Kanehisa and Goto 2000), *Pathways Common* (Rodchenkov et al, 2020), or *Reactome* (Griss et al, 2020) with to a certain level of deficient or overlapping information. Information about drugs launched or in development (drug omics data) can be obtained from *Pubchem* (Kim et al, 2019), *Drug Bank* (Wishart et al, 2018), *Drug Map Central* (Fu et al, 2013), or *FDA or EMA Label* repositories. Also, the Clinicaltrials.gov, *SIDER* (Kuhn et al, 2015), or *FDA*

Adverse Event Reporting System (FAERS) platforms provide essential information on stages of drug testing.

Platforms for browsing and visualizing drug-target interactions and drugs in disease-context are already available for many users, such as *Drug Target Profiler* (Tanoli et al, 2018), *Cancer Genome Interpreter* (Tamborero et al, 2018), *SwissTargetPrediction* (Daina et al, 2019), *OpenTargets* (Hecker et al, 2012), and *PharmGKB* (Eichelbaum et al, 2009). These platforms include many entries and a gargantuan amount of unstructured information that is sometimes time-consuming and difficult to handle, especially for those lacking expertise in the field. The data mining techniques require specific processes usually nonexistent in less prevalent cancers and often lack financial incentive for clinical testing (Josephs et al, 2019). Cancer drug discovery requires a convergence of complex, often disparate fields. There is a great need for simple, transparent and user-friendly drug repurposing databases.

One recent endeavor with high impact is Broad Institute's Clue.io, which serves as a drug repurposing hub curated and annotated collection of FDA-approved drugs, clinical trial drugs, and preclinical tool compounds with a companion information resource (Corsello et al., 2017). In this paper, we present a novel, open access, data-mining, drug repurposing platform, deriving our searches from the entries of *Clue.io*. Project Scancer provides researchers and clinicians (especially in cancer research) an easy-to-use platform to retrieve all the necessary information for a freely selected array of potential therapeutic targets with a parallelly-working, data-mining application. In addition, Project Scancer provides detailed biological information on the selected set of target molecules using open-access databases, such as Uniprot, GeneCards, Gene Ontology and STRING. This way, the user receives a concise summary on the biological relevance of every target, that is explicitly important for researchers who are not experts in molecular biology.

Methods

Clue.io and dataPatch R script

Target inclusion/exclusion depends on search results from a clue.io query. Scancer consists of 3 separate R scripts. The first script -clue.R- calls various clue.io REST API endpoints to build up a result table (Fig. 2). If the main API call does not find any component for a target, that target will not be involved in further processing steps since no known drug repurposing approaches are available in Clue.io. R looks up the input target list in two ways. First, it tries to access a private and/or shared Google Sheet file. It requires a unique "key" (a token) given to clue.R via a simple environmental variable. If this secret key is available for the script, it authenticates by *gargle* package to access Google Sheet API services. Next, it reads the sheet and takes the values from its first three columns. An ID string also identifies the Google Sheet, and it is passed via an operating system environment. If there is no API key/Google Sheet identifier, then clue.R tries to load a TSV file from the INPUT directory of the Scancer directory. Clue.R merges the outputs of various clue.io API calls and saves the composing table into an RDS file (R-specific data

format to store and load R objects). At the next stage, data patch.R reads this RDS file and restores the data frame from it.

FDA Label function

dataPatch.R collects additional details from external resources. Most of them are used by clue.io itself, but gaps can occur in its dataset. For example *EMA* data is not included at all, so appending it is an improvement. Another plus is harvesting direct links to drug labels from the search *interface of FDALabel*. fdaLabel function implements this feature that contains several internal sub-functions (some of them are anonymous and vectorized). This function then sends a POST HTTP request to a specific FDALabel link - the same that is used by the human search interface to send a query. fdaLabel function uses pert_iname from clue.io as a search parameter and uses the following parameters to define the search criteria:

- document types of labeling types as documentTypeCodes: Human Rx 34391-3, Human OTC 34390-5, Vaccine 53404-0
- labeling section: selectedLabelingType: 0, sectionTypeCode: Active Ingredient 2-55106-9 fdaLabel uses its internal function - getFDALabelResults - to send the query and interpret the HTTP response from the FDA service.

The result may contain several duplications for a specific product. It is a consequence of FDA strict registration and regulation procedures (products are allowed for different time ranges or can be withdrawn, etc). The function uses a weighting method in order to reduce these duplications and return the most relevant items. fdaLabel function filters out products that are parts of aid kits by verifying presence of simple words first aid|kit|KIT in product names. It receives responses from the FDALabel service in JSON format. It parses, prioritizes and filters the list of returned drug products based on their market date and match-mode with pert_iname (exact, prefix, suffix, inner, not at all). This is a necessary step since the product list contains copies of drugs with only different dates.

PubMed function

This function searches for the compound name received from clue.io and sends a search request to PubMed® service of *National Center for Biotechnology Information* (NCBI). It restricts the result set by including only clinical trials, meta-analyses, randomized controlled trials, reviews and systematic reviews. The results are ordered by the best match algorithm of PubMed®. Our function picks the top 3 of the result set and stores it in the global datatable. If there is no hit at all, dataPatch.R provides the used search URL and this search can be re-initiated and/or refined by users of Scancer. pubMed function uses a simple XPath query to extract PubMed identifiers embedded into the resulting HTML source code.

EMA function

EMA function complements the dataset with product details from the website of *European Medicines Agency* (EMA). It sends search queries to the site through the human-dedicated interface - similar to

fdaLabel and pubMed functions above. It also uses XPath expression to extract the first hit from the result list and to identify the most specific link to a document in PDF format. There is no strict naming/referencing convention for these documents on *EMA*, this function tries to discover a PDF document in the following order (with descending priority):

- summary in English,
- product information,
- refusal public assessment report,
- public statement which is not "non-renewal" and not "authorization"
- Otherwise it selects any PDF document that is encountered on the page of the product first. If EMA finds a PDF reference on the product page, it verifies the existence of the referred PDF file by sending an HTTP HEAD request to the *EMA* web server. This additional check ensures that Scancer will not provide an URL with "Page Not Found"-experience to its users.

XmlUniProt function

This function collects data from the *UniProt* website. *UniProt* provides APIs to access and query its data. Easy to access the human readable contents in machine readable formats (for example XML, RDF, etc.). The usage of the *UniProt website REST API* is straightforward, since the input target list also contains UniProt identifiers. xmlUniProt extracts GO (*Gene Ontology*) molecular function and cellular component terms, *STRING* and *Reactome* references from the received XML data. These specific entries are stored in simple R lists and added to the already collected data in a new column: UniProtData

RenderWebPage function and renderWebPage.R script

RenderWebPage is responsible for rendering human-readable HTML documentation from "patched" data produced by dataPatch.R. This function iterates on rows of the input table created by dataPatch.R. It prepares a compounded, hierarchical data structure from the data table. This data structure helps to simplify data access from the template file which is an important substance of generating out HTML.

MultivaluedCellsToHTML function

A single compound can have multiple related values as elements of various resources. For example a compound can have two *Mechanism of Action* items (MoA); or a *PubChem* reference along with a *DrugBank* reference. These values have to be organized into the same row as the compound and items of the same categories must be displayed in a single table cell. multivaluedCellsToHTML function handles these cases. MultivaluedCellsToHTML function uses these functions to compose corresponding URLs for each identifier from different data sources, including ChEMBL, PubChem and DrugBank.

Presentation layer

A web browser is a "mandatory" software of each end user's computer, so HTML is a clear choice to summarize, visualize and deliver collections of texts, images and hyperlinks. An important part of this rendering is building an HTML source file and populating it with the collected data in an user-friendly way. Scancer follows the popular *Model-View-Controller* design pattern even though it composes only static HTML output (View) from the data (Model) at this stage of the workflow. *(NOTE: However previous actions and functionalities of the workflow can be interpreted as the Controller part of the MVC pattern.)*

Templating

Most web frameworks incorporate a templating system - as a result of their own solution or reusing a 3rd party component. These templating components are not tightly coupled with web services, any software can use their power. Scancer uses the *whisker* package, which implements the *Mustache* template language. This approach excludes any occurrence of program source code from the template code. It provides a strict separation for the *View* layer of Scancer. The structure of the rendered HTML output is based on Bootstrap components and their hierarchy. *(NOTE: The current hierarchical structure and the content of components cannot support a responsible page design. scancer output is tailored for desktop browsing.)*

Dynamic request for STRING network image and a simple cache for HTML/XML contents

Current functionality of Scancer is very similar to a web crawler. This kind of interaction with popular web servers requires respecting their policies controlled by their robots.txt files and described in public pages (FAQ, Usage Guideline, Terms of Service etc.) in order to minimize the load on their resources. Respecting their resources also helps to avoid a possible block or denied access from a well-protected website.

Scancer uses an event-based method, a dynamic DOM modification from JavaScript (see *scancer.js*). If the user opens the STRING accordion of a target content, then the *getString* function (set up as on onclick handler) inserts the image tag. The HTTP request of the STRING network image is initiated by the browser as it modified the *Document Object Model* and loads the missing/uncached image content to complete the rendering of the missing part of the document. It loads the images belonging to only the visited STRING panels of the document.

The simple cache is also purposed to saving resources on websites. scancer send HTTP requests and receive HTTP responses via its *getPageCached* function. This function checks the *cache.tsv* file in the caching folder and returns the content immediately, when it has been already downloaded earlier. If the looked up entry is missing from the cache, the function downloads it and adds it to the cache. This does not just spare remote resources, but it speeds up the subsequent queries to a specific server since Scancer does not need to wait between politely after requests served by the cache-solution.

Description of versions and runtime environment is accessible at:

<https://cycle20.github.io/scancer/Software%20Versions%20and%20Runtime%20Environment>

Results

Construction and Content

Each entry obtained from the search results in the interactive online platform of project Scancer is referenced and has at least one scientific piece of evidence. Figure 1 shows a flowchart on the processing steps and workflow of Scancer. First, users of Scancer can start their workflow by opening the project's starting page on Github (<https://cycle20.github.io/scancer/>). Then, users can upload their target list with three pieces of information into a google spreadsheet (Target INPUT). The first piece (Fig. 2 **Column A**), asks for the HUGO ID. In **column B**, users can give a "Label" for every target for classification and clustering, useful in later work. The third piece (Fig. 2 **column C**) is the UniprotKB ID of the searched gene. Inputs for the HUGO and Label columns are limited to 12 characters. On the right side of the spreadsheet (columns E-K), hyperlinks provide access to the results page on Github, and in the "Results of Update Request" box, users can check the query's status. Hitting the "Start Rendering" button located on columns E-F starts the query. The area within E1-K8 are protected and automatically overwritten if edited (Fig. 2).

By clicking on the "Result page" link on the target spreadsheet, we can access the results of our query within approximately 30 minutes. Clicking on the hyperlink in cell H6 we can follow the progress of the query (Fig. 2, **arrow**). A new query overwrites the earlier one in the web application, but every previous version is saved on Github under the "Result of Update Request" link (<https://github.com/cycle20/scancer/actions/workflows/clue.yml>). A scrollable panel displays all the targets on the left with at least one valid drug compound available. Scancer automatically excludes entries where no drug or small molecule inhibitor/agonist is available according to the Clue.io repurposing hub. The platform creates a table for every target, where different columns indicate the mechanism of action (MoA), clinical status (preclinical, phase 1, phase 2, phase 3, or launched), and the search resources from PubMed EMA and the direct entry from Clue.io. Furthermore, the query table includes hyperlinks with DrugBank, PubChem, and ChEMBL IDs to quickly access the compounds' chemical and pharmacological properties (Fig. 3).

Project Scancer also gives a comprehensive, highly structured overview of the selected targets regarding their molecular biology data, including molecular function (Gene Ontology), their connectome (STRING), participation in pathways, and cellular localization (Reactome) retrieved from various databases. Hyperlinks to GeneCards and DrugBank Target Search are also available but differently structured as for UniProt entries. The "STRING" entry opens a static string map for the target and provides a hyperlink to string-db.org (Fig. 4). The following entry carries the "Molecular Functions / Subcellular Localisations" title, where the two main hyperlinks (source) lead to UniProt's "Function" and "Subcellular Localization" pages. Molecular function entries and target localizations are also provided as text separately, where hyperlinks lead to the QuickGo platform to obtain further information about relevant compartment-specific molecular pathways (Fig. 4). The last entry named "Pathways" provides links to every Reactome database, where the target's participation is visualized in every relevant metabolic pathway (Fig. 4).

Supplementary Video 1 shows a short tutorial about the functionality of the program and the main steps to generate a query.

Current content

This current database includes 97 targets that focus on molecular targets of SCLC (Dora et al, 2021), but can be expanded or replaced with targets concerning other aggressive malignancies, including triple-negative breast cancer, glioblastoma multiforme, and pancreatic cancer. Users and groups can expand the content and assign their own tasks to indicate progress percentages in a different platform. Also, it is possible to give in the "label" column to address targets and classify them into different groups.

Discussion

Open access journals and databases are an essential basis for drug target developments. Endeavours, like the TCGA database and Oncomine (Rhodes et al, 2004) has contributed vastly to accelerate drug research in oncology the latest decade and concurrently multiple other enterprises emerged to assist researchers with valuable genomic, transcriptomic and proteomic data in the pursue for novel cancer biomarkers (Krempel et al, 2018; Wishart et al, 2021; Banck et al, 2021; Pantziarka et al, 2021). However, the information is not well structured for specific diseases, including rare and highly progressive cancers (Gadaleta et al, 2011; Creighton 2018). Moreover, it is challenging and time-consuming to associate the latest biomarkers with drugs to pick the optimal way and contribute to the field. Only a few research groups with diverse expertise can participate, leaving many researchers without an equal opportunity of involvement. Also, the individual interest of these groups might not represent the optimal way to examine diseases. Therefore, we propose a novel, optimal target selection methodology. Notably, after the success of PD-L1-inhibitors in non-small cell lung cancer (NSCLC), where 5-year-survival in extensive-stage disease increased from 2–25%, there has been a keen interest to expand on immunotherapy utilization in small cell lung cancer (SCLC) as well. However, two anti-PD-1 immunotherapies, nivolumab and pembrolizumab, have had their FDA approval (Horn et al., 2018; Pas-Ares et al., 2019), they were withdrawn after the confirmatory phase III trials did not reach statistical significance for overall survival. Nevertheless, PD-L1 expression in SCLC has never been unequivocally correlated with the response.

Open-access data on the latest research requiring further validation is of high interest to the field. In low-prevalence and highly aggressive cancers, scarcity of available tissue samples limit research, so there is an unmet need to share and optimize resources in the field. It has been decades with only modest therapeutic advancements for highly progressive cancers such as pancreatic cancer, triple-negative breast cancer, glioblastoma multiforme, or SCLC, with an unmet need for advances. To enhance drug target development, we believe that project Scancer can serve as an easy-to-use, semi-comprehensive data-mining platform for drug-repurposing and can assist significantly smaller research groups in the fight against malignancies. An outstanding advantage of Project Scancer compared to other drug-repurposing platforms is that it does not require any experience in database handling and provides information on the biological background of our target molecules in a processed way that is easy to

understand. The latter feature can be used for educational purposes as well. We believe that Project Scancer is a useful addition to the field of drug-repurposing in cancer science and oncology and will be particularly useful for smaller research groups with limited expertise in database handling.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Description of versions, software code, runtime environment and all data is accessible at:
<https://cycle20.github.io/scancer/Software%20Versions%20and%20Runtime%20Environment>

Competing interests

The authors declare that they have no competing interests

Funding

Zoltan Lohinai acknowledge funding from the Hungarian National Research, Development and Innovation Office (OTKA #124652, OTKA #129664 and OTKA #128666). David Dora acknowledge funding from Semmelweis University's "Start-up" grant.

Authors' contributions

Conceptualization, DD and ZL; Methodology and Software, CG and GS; Validation and Formal analysis, DD, TD and CG; Investigation, DD and CG; Resources, DD and ZL; Data curation, ZL, GS and CG; Visualisation, DD and TD; Supervision and Project administration, ZL, DD and TD; Funding acquisition, ZL; Writing-original draft, DD and ZL; Writing-reviewing and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors thank Glen J. Weiss, MD, MBA for critical review and feedback on this manuscript.

References

- Banck, H., Dugas, M., Müller-Tidow, C., & Sandmann, S. (2021). Comparison of Open-access Databases for Clinical Variant Interpretation in Cancer: A Case Study of MDS/AML. *Cancer genomics & proteomics*, 18(2), 157–166. <https://doi.org/10.21873/cgp.20250>
- Chen, X., Ji, Z. L., & Chen, Y. Z. (2002). TTD: Therapeutic Target Database. *Nucleic acids research*, 30(1), 412–415. <https://doi.org/10.1093/nar/30.1.412>
- Corsello, S., Bittker, J., Liu, Z. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med* 23, 405–408 (2017). <https://doi.org/10.1038/nm.4306>
- Creighton C. J. (2018). Making Use of Cancer Genomic Databases. *Current protocols in molecular biology*, 121, 19.14.1–19.14.13. <https://doi.org/10.1002/cpmb.49>
- Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res.* 2019 Jul 2;47(W1):W357-W364. doi: 10.1093/nar/gkz382. PMID: 31106366; PMCID: PMC6602486.
- Dora D, Rivard C, Yu H, Pickard SL, Laszlo V, Harko T, Megyesfalvi Z, Dinya E, Gerdan C, Szegvari G, Hirsch FR, Dome B, Lohinai Z. Characterization of Tumor-Associated Macrophages and the Immune Microenvironment in Limited-Stage Neuroendocrine-High and -Low Small Cell Lung Cancer. *Biology*. 2021; 10(6):502. <https://doi.org/10.3390/biology10060502>
- Eichelbaum, M., Altman, R.B., Ratain, M. & Klein, T.E. New feature: pathways and important genes from PharmGKB. *Pharmacogenet. Genomics* 19, 403 (2009).
- Fu C, Jin G, Gao J, Zhu R, Ballesteros-Villagrana E, Wong ST. DrugMap Central: an on-line query and visualization tool to facilitate drug repositioning studies. *Bioinformatics*. 2013 Jul 15;29(14):1834-6. doi: 10.1093/bioinformatics/btt279. Epub 2013 May 15. PMID: 23681121; PMCID: PMC3702253.
- Gadaleta, E., Lemoine, N. R., & Chelala, C. (2011). Online resources of cancer data: barriers, benefits and lessons. *Briefings in bioinformatics*, 12(1), 52–63. <https://doi.org/10.1093/bib/bbq010>
- Gns HS, Gr S, Murahari M, Krishnamurthy M. An update on Drug Repurposing: Re-written saga of the drug's fate. *Biomed Pharmacother.* 2019 Feb;110:700-716. doi: 10.1016/j.biopha.2018.11.127. Epub 2018 Dec 12. PMID: 30553197.
- Griss J, Viteri G, Sidiropoulos K, Nguyen V, Fabregat A, Hermjakob H. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol Cell Proteomics*. 2020 Sep 9. doi: 10.1074/mcp. [PubMed PMID: 32907876](https://pubmed.ncbi.nlm.nih.gov/32907876/).
- Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, Gilson MK, Bourne PE, Preissner R. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.* 2012 Jan;40(Database issue):D1113-7. doi: 10.1093/nar/gkr912. Epub 2011 Nov 8. PMID: 22067455; PMCID: PMC3245174.

Horn L, Mansfield AS, Szczesna A, Havel L, Krzakowski M, Hochmair MJ, Huemer F, Losonczy G, Johnson ML, Nishio M *et al* (2018) First-line atezolizumab plus chemotherapy in extensive-stage small-cell lung cancer. *N Engl J of Med* 379, 2220–2229. [PubMed] [Google Scholar]

Josephs, K. S., Berner, A., George, A., Scott, R. H., Health Education England's Genomic Education Programme, Firth, H. V., & Tatton-Brown, K. (2019). Genomics: the power, potential and pitfalls of the new technologies and how they are transforming healthcare. *Clinical medicine (London, England)*, 19(4), 269–272. <https://doi.org/10.7861/clinmedicine.19-4-269>

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, 49(D1), D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>

Kreppeel, R., Kulkarni, P., Yim, A., Lang, U., Habermann, B., & Frommolt, P. (2018). Integrative analysis and machine learning on cancer genomics data using the Cancer Systems

Biology Database (CancerSysDB). *BMC bioinformatics*, 19(1), 156. <https://doi.org/10.1186/s12859-018-2157-7>

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000 Jan 1;28(1):27-30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409.

Kuhn M, Letunic I, Jensen LJ, Bork P. [The SIDER database of drugs and side effects](#). *Nucleic Acids Res.* 2015 Oct 19. doi: 10.1093/nar/gkv1075

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. *Nucleic acids research*, 36(Database issue), D684–D688. <https://doi.org/10.1093/nar/gkm795>

Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 2020 Oct 18

Pantziarka, P, Capistrano I, R., De Potter, A., Vandeborne, L., & Bouche, G. (2021). An Open Access Database of Licensed Cancer Drugs. *Frontiers in pharmacology*, 12, 627574. <https://doi.org/10.3389/fphar.2021.627574>

Parvathaneni V, Kulkarni NS, Muth A, Gupta V. Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug Discov Today*. 2019 Oct;24(10):2076-2085. doi: 10.1016/j.drudis.2019.06.014. Epub 2019 Jun 22. PMID: 31238113.

Paz-Ares L, Dvorkin M, Chen Y, Reinmuth N, Hotta K, Trukhin D, Statsenko G, Hochmair MJ, Özgüroğlu M, Ji JH *et al* (2019) Durvalumab plus platinum–etoposide versus platinum–etoposide in first-line treatment

of extensive-stage small-cell lung cancer (CASPIAN): a randomised, controlled, open-label, phase 3 trial. *Lancet* 394, 1929–1939. [[PubMed](#)] [[Google Scholar](#)]

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004 Jan-Feb;6(1):1-6. doi: 10.1016/s1476-5586(04)80047-2. PMID: 15068665; PMCID: PMC1635162.

Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res*. 2020 Jan 8;48(D1):D489-D497. doi: 10.1093/nar/gkz946. PMID: 31647099; PMCID: PMC7145667.

Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis*, [Current Protocols in Bioinformatics\(2016\)](#), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019 Jan 8;47(D1):D607-D613. doi: 10.1093/nar/gky1131. PMID: 30476243; PMCID: PMC6323986.

Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, de Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. 2018 Mar 28;10(1):25. doi: 10.1186/s13073-018-0531-8. PMID: 29592813; PMCID: PMC5875005.

Tanoli Z, Alam Z, Ianevski A, Wennerberg K, Vähä-Koskela M, Aittokallio T. Interactive visual analysis of drug-target interaction networks using Drug Target Profiler, with applications to precision medicine and drug repurposing. *Brief Bioinform*. 2018 Dec 18. doi: 10.1093/bib/bby119. Epub ahead of print. PMID: 30566623.

Tanoli Z, Seemab U, Scherer A, Wennerberg K, Tang J, Vähä-Koskela M. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief Bioinform*. 2021 Mar 22;22(2):1656-1678. doi: 10.1093/bib/bbaa003. PMID: 32055842; PMCID: PMC7986597.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D480-D489. doi: 10.1093/nar/gkaa1100. PMID: 33237286; PMCID: PMC7778908.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox

C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2017 Nov 8. doi: 10.1093/nar/gkx1037.

Wishart, D. S., Bartok, B., Oler, E., Liang, K., Budinski, Z., Berjanskii, M., Guo, A., Cao, X., & Wilson, M. (2021). MarkerDB: an online database of molecular biomarkers. *Nucleic acids research*, 49(D1), D1259–D1267. <https://doi.org/10.1093/nar/gkaa1067>

Figures

Figure 1

Flowchart of functionality. Flowchart describes the main steps of Project Scancer's functionality, including data input, Clue.io target search, cross-referencing in databases (Datapatch) and molecular background information on selected targets (Render).

	A	B	C	D	E	F	G	H	I	J	K
1	HUGO	Label	UNIPROT_KB_ID		Baseline page:	https://cycle20.github.io/scancer/baseline/index_target_with_data.html					
2	ALDH2	LOW	P05091		Result page:	https://cycle20.github.io/scancer/index_target_with_data.html					
3	ANXA1	LOW	P04083		Start Rendering ...	<div>Program messages</div> <div>Result of Update Request</div> <div>Workflow started. Please follow its progress on GitHub</div> <div>https://github.com/cycle20/scancer/actions/workflows/clue.yml</div>					
4	ANXA3	LOW	P12429								
5	BCAT1	LOW	P54687								
6	BCL3	LOW	P20749								
7	BIRC3	LOW	Q13489								
8	C1QA	LOW	P02745								
9	CASP14	LOW	P31944								
10	MX1	LOW	P20591								
11	MYC	LOW	P01106								
12	PGC	LOW	P20142								
13	PRAME	LOW	P78395								
14	PRF1	LOW	P14222								
15	PTGS2	LOW	P35354								
16	RAB27B	LOW	Q00194								
17	REL	LOW	Q04864								
18	ROS1	EURO	P08922								
19	TNFSF10	LOW	P50591								
20	TP63	LOW	Q9H3D4								
21	VAMP8	LOW	Q9BV40								
22	VSNL1	LOW	P62760								
23	YBX3	LOW	P16989								
24	ADAM23	HIGH	Q75077								
25	AMH	HIGH	P03971								
26	PAK3	HIGH	Q75914								
27	PAK7	INHIBITOR	Q9P286								
28	PTN	HIGH	P21246								

Gene HUGO name

Designated label / classification

UNIPROT ID

Figure 2

Input table for molecular targets. Users can enter selected targets' HUGO name, label and Uniprot ID in columns A, B and C. Hitting "Start Rendering" will initiate the Clue.io search (arrowhead). Progress can be traced by clicking on hyperlink in cells H6-K6 (arrow). Clicking on the hyperlink in cell F2-I2 reveals the results page.

Drug Targets - Has CLUE.IO Entries

Created at 2021-06-09 17:23:34 UTC

BIRC3 :: Compounds				
Compound	MoA ⁱ	Clinical Status	Resources ⁱ	DrugBank/PubChem/ ChEMBL
AT-406	IAP antagonist	Phase 1	From PubMed: PubMed 33460574 PubMed 25716544 PubMed 27915971 From EMA: Found PDF CLUE.IO ClinicalTrials	CHEMBL2158051 PubChem: 25022340
AZD5582	XIAP inhibitor	Preclinical	From PubMed: PubMed 33007800 From EMA: Found PDF CLUE.IO	CHEMBL3108827 PubChem: 49847690
birinapant	XIAP inhibitor	Phase 2	From PubMed: PubMed 31766284 PubMed 26333381 PubMed 30631429 CLUE.IO ClinicalTrials	CHEMBL3039522 PubChem: 49836020 DB11782
GDC-0152	XIAP inhibitor	Phase 1	From PubMed: PubMed 22623056 From EMA: Found PDF CLUE.IO ClinicalTrials	CHEMBL2063869 PubChem: 46940575 DB12380

Figure 3

List of drug targets. Clicking on the labels of selected targets (column on left side) unveils available compound list (black box) describing also mechanism of action (MoA, dashed box), clinical status (red box), resources of information on PubMed (green box) and DrugBank/PubChem/ChEMBL entries (blue box).

Figure 4

Details on the target's molecular background. Panel A shows “results page” menu for network map from String.io with static string map and hyperlink to String-db entry. Panel B displays hyperlinks to “molecular function” and “subcellular localisation” to browse the UniProt database on molecular background. Panel C shows hyperlinks to visualize “Reactome” pathways of the selected target.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [scancerQuickIntroSVideo.mp4](#)