

# Integration of the Drug-Gene Interaction Database (DGldb) with open crowdsourcing efforts

Sharon Freshour<sup>1,2,†</sup>, Susanna Kiwala<sup>2,†</sup>, Kelsy C. Cotto<sup>1,2,†</sup>, Adam C. Coffman<sup>2</sup>, Joshua F. McMichael<sup>2</sup>, Jonathan Song<sup>1,2</sup>, Malachi Griffith<sup>1,2,3,4,\*</sup>, Obi L. Griffith<sup>1,2,3,4,\*</sup>, Alex H. Wagner<sup>1,2,5,6,\*</sup>

<sup>1</sup> Department of Medicine, Division of Oncology, Washington University School of Medicine, St Louis, MO, 63110, USA

<sup>2</sup> McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO, 63108, USA

<sup>3</sup> Department of Genetics, Washington University School of Medicine, St Louis, MO, 63110, USA

<sup>4</sup> Siteman Cancer Center, Washington University School of Medicine, St Louis, MO, 63110, USA

<sup>5</sup> The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, 43215, USA

<sup>6</sup> Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, 43210, USA

\* To whom correspondence should be addressed. Tel: 1-614-355-1645; Fax: 1-614-355-6833; Email:

[Alex.Wagner@nationwidechildrens.org](mailto:Alex.Wagner@nationwidechildrens.org)

Correspondence may also be addressed to Obi L. Griffith. Tel: 1-314-747-9248; Fax: 1-314-286-1810; Email:

[obigriffith@wustl.edu](mailto:obigriffith@wustl.edu)

Correspondence may also be addressed to Malachi Griffith. Tel: 1-314-286-1274; Fax: 1-314-286-1810; Email:

[mgriffit@wustl.edu](mailto:mgriffit@wustl.edu)

† The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

## ABSTRACT

The Drug-Gene Interaction Database (DGldb, [www.dgldb.org](http://www.dgldb.org)) is a web resource that provides information on drug-gene interactions and druggable genes from various sources including publications, databases, and other web-based sources in one resource. These drug, gene, and interaction claims are normalized and grouped to identify aliases, merge concepts, and reduce redundancy. The information contained in this resource is available to users through a straightforward search interface, an application programming interface (API), and TSV data downloads. DGldb 4.0 is the latest major update of this database. Seven new sources have been added, bringing the total number of sources included to 41. Of the previously aggregated sources, 15 have been updated. DGldb 4.0 also includes improvements to the process of drug normalization and grouping of imported sources. Other notable updates include further development of automatic jobs for routine data updates, more sophisticated query scores for interaction search results, extensive manual curation of interaction source link outs, and the inclusion of interaction directionality. A major focus of this update was integration with crowd-sourced efforts, including leveraging the curation activities of Drug Target Commons, using Wikidata to facilitate term normalization, and integrating into NDEX for producing network representations.

## INTRODUCTION

Originally released in 2013, the Drug-Gene Interaction database (DGldb)(1) serves as a single resource joining information on drug-gene interactions and druggability from multiple diverse sources. Previously, DGldb has released two major updates, DGldb 2.0(2) (in 2016) and 3.0(3) (in 2018) that included improvements to the user interface and search response times, the addition of an API, the introduction and improvement of gene and drug grouping methods, and the expansion of source content through the inclusion of new sources and updates of existing sources. Since the release of DGldb 3.0, many of the existing sources have been substantially updated and new sources have

become available. To help ensure that DGldb offers current information, automatic updates have been further developed for multiple sources and a new background job management system (Sidekiq, [sidekiq.org](https://sidekiq.org)) has been implemented for routine job scheduling in the latest DGldb release (DGldb 4.0), described here. In addition to updates and integration of new sources, DGldb 4.0 focuses on numerous improvements of search results, including an updated query score for interaction results and improved drug normalization. With these improvements, we have also made an effort to integrate crowd-sourced data and sources in several areas, including the addition of the crowd-sourced Drug Target Commons(4) as a drug-gene interaction source, and the use of the open, community Wikidata(5) resource for drug normalization. We also illustrate the value of our integration efforts in downstream community tools, through the incorporation of our data into NDEx(6).

## NEW AND UPDATED CONTENT

In an effort to ensure that DGldb offers diverse and contemporary information, we have both updated many existing sources and added several new sources to DGldb 4.0. For new sources, we have added two new sources of drug-gene interaction records and five new sources of druggable gene category records. The new sources of drug-gene interactions are Drug Target Commons(4) and COSMIC(7) (Supplemental Table 1). Drug Target Commons provides an extensive database of drug-gene interactions that have been curated through crowd-sourcing curation of expert knowledge. COSMIC provides an additional source of curated information on cancer mutations and drug resistance. The new category sources are Tempus xT(8), a panel of actionable cancer therapy target genes, a list of the top priority genes of the Illuminating the Druggable Genome (IDG)(9) Initiative, the Human Protein Atlas(10), Oncomine(11), a clinical cancer biomarker assay, and Pharos(12), which hosts data as part of the IDG program and focuses on understudied targets in three broad classes of druggable genes: G protein-coupled receptors, ion channels, and kinases (Supplemental Table 1). From these new sources, we have added 23,916 new drug-gene interaction records and 8,463 new druggable gene category records (Figure 2).

Additionally, we have also updated multiple sources including large, well-curated sources such as DrugBank(13), Guide to Pharmacology(14) Interactions, Gene Ontology(15), (16), OncoKB(17), PharmGKB(18), and the Therapeutic Target Database(19) (Figure 2, Supplementary Table 1). To facilitate routine updates, the importer for PharmGKB has been updated to an online importer that can be run periodically using DGldb's new automated job scheduling system (see Technology Improvements). Similarly, Pharos(12), one of our new sources, has been implemented as an online updater. Of the 41 sources in DGldb, 12 sources are now imported using the online updater format, including Entrez(20), the core source of gene concepts for gene grouping in DGldb, and Ensembl(21), a key source of gene aliases. In DGldb 4.0, the number of genes imported from Entrez has increased from 41,102 to 42,767 and Ensembl has been updated from version 90\_38 to version 101\_38. We have also migrated several sources from domain-specific language (DSL) importers to the TSV importer style implemented in DGldb 3.0.

In DGldb 4.0, the database structure and presentation model was updated to allow sources to be imported with multiple source types. This change enables merging of sources that were previously duplicated for each independent claim type (drug, gene, interaction, druggable gene category). For example, we currently import both interaction claims and druggable gene category claims from Guide to Pharmacology(14) with two separate importers. Instead of listing two independent sources (GuideToPharmacologyInteractions, GuideToPharmacologyGenes), we can now have a single source (GuideToPharmacology). This is intended to simplify and unify claim sources to aid in downstream interpretation. Additionally, supporting sources that have multiple source types enables easy extension to collect more informative claim type information. For example, some claims from CIVIC(22) can be imported with the additional category of drug resistance. Overall, these changes will increase the efficiency and accuracy of the process of importing and updating sources in DGldb 4.0 compared to previous versions and will make it easier for users to evaluate individual sources.

## **IMPROVED TRANSPARANCY AND DETAILS ON LICENSING OF SOURCES**

In DGldb 4.0, we have made a significant effort to update and improve the information we provide on licensing of sources imported into our database through manual curation of licensing types and details for every source. This information is now readily available on the sources page and sources with special licenses are noted on the downloads page (Figure 1). Since DGldb 3.0, several existing sources have made changes to their licensing, making data from some sources more broadly available and data from other sources more restricted. Notably, PharmGKB(18) has moved to a more permissive Creative Commons Attribution-ShareAlike 4.0 International License and DrugBank(13) has adopted a custom non-commercial license. In contrast, OncoKB(17) has restricted API access to registered/approved non-commercial research use only, and JAX-CKB(23) has restricted API access to negotiated licenses only. Both resources continue to provide access to a portion of their data for free through their respective web clients. When exporting data TSVs for downloading, we partition data with restricted license types separately and note the details of licensing. If data cannot be made available for downloading, users are given directions on how data can be obtained from the sources themselves.

## **MONTHLY DATA RELEASES**

DGldb 3.0 implemented online updaters that imported data from dynamic sources (such as CIViC(22), Guide to Pharmacology(14), OncoKB(17), etc.) periodically, usually monthly. As a result, the static TSV data releases available on our Downloads page would quickly become outdated. For DGldb 4.0, we have implemented monthly data releases of these TSVs to coincide with monthly runs of the online updaters, to ensure that TSVs available for download reflect the most up to date information in our database (Figure 1). The Downloads page now makes available the current Gene, Drug, Interaction, and Category TSVs as well as historic TSVs from past runs. These serve as de-facto snapshots of the data in DGldb over time.

## **INTEGRATION WITH NDEX AND OTHER EXTERNAL RESOURCES**

The Network Data Exchange (NDEX)(6) is a community resource that allows sharing and publishing of biological data in a network-based format. For DGldb, integration with the NDEX platform provides a resource for the visual representation of relationships and interactions between drugs and genes present in our database, allowing users to easily explore a global overview of drugs and genes of interest. NDEX TSVs are generated monthly and automatically uploaded to the NDEX server to keep the DGldb network in NDEX up-to-date (Figure 1). Previously, DGldb has been integrated with other community resources such as GeneCards(24), which uses DGldb as a source for interactions between GeneCards gene entries and drugs, and BioGPS(25), which offers a DGldb plug-in for searching drug-gene interactions. Since the publication of DGldb 3.0, DGldb has been regularly integrated with additional resources such as CancerTracer(26), Gene4Denovo(27), SL-BioDP(28), TargetDB(29), and OncoGemini(30).

## **UPDATED QUERY AND INTERACTION SCORES**

One of the main features added in DGldb 4.0 is the concept of relative query scores for interaction search results. Previously, interaction search results displayed only a static score based on evidence of an interaction (i.e. the number of publications and sources supporting an interaction claim). This score did not take into account whether the gene and drug involved in a given interaction were also part of a large number of other interactions and, thus, had a low specificity that should be penalized. In addition, when searching for a set of genes or drugs, the previous interaction score would also not prioritize results with overlapping interacting drugs or genes, which might be of more interest to the user, particularly in drug discovery and pathway applications.

DGIdb 4.0 now provides a query score that is relative to the search set and considers the overlap of interactions in the result set. For interaction searches using a gene list, the query score depends on the evidence scores (publications and sources), the number of genes from the search set that interact with the given drug, and the ratio of average known gene partners for all drugs to the known partners for the given drug (Figure 3). Similarly, for interaction searches using a drug list, the query score depends on the evidence scores (publications and sources), the number of drugs from the search set that interact with the given gene, and the ratio of average known drug partners for all genes to the known partners for the given gene (Figure 3). In effect, this means that genes and drugs with many overlapping interactions in the search set will rank more highly, with the caveat that drugs or genes involved in many interactions, in general, will have lowered scores (Figure 3).

The static interaction score has also been improved. It now mirrors the query score except that it removes the dependency on queried gene or drug sets and only uses evidence scores, the ratio of average known gene partners for all drugs to the known partners for the given drug, and the ratio of average known drug partners for all genes to the known partners for the given gene (Figure 3). In summary, the introduction of the relative query score will provide users with a score that gives a more intuitive ranking of drugs or genes based on the search set of interest, allowing the prioritization of drugs or genes that have overlapping, specific interactions with the search set. Similarly, the improvements to the static interaction provide a more nuanced scoring system that takes into consideration the number of interactions for a drug-gene pair, in addition to the previous evidence scores, giving a more informative static score.

## OTHER SEARCH RESULT IMPROVEMENTS

Another major focus of the DGIdb 4.0 update is the improvement of features related to search results and information available on the user interface (Figure 1). One such improvement is the addition of link outs to specific interaction evidence, where available. These will allow users to browse to a specific source page for an interaction claim which might provide more information than available on DGIdb.

We have also added information on the directionality of interaction types. This directionality indicates whether the interaction type of a specific interaction result has an inhibitory or activating effect overall (when the directionality is available for an interaction type). While the directionality may be clear for some interaction types, e.g. activators have activating directionality, some interaction types are less intuitive, e.g. chaperones have activating directionality. Inclusion of directionality can make it easier for users to distinguish interactions that are more relevant for their purposes. For example, a user interested in exploring drugs that inhibit a particular gene will readily see that chaperone interactions are activating, without further investigation of the interaction, and might ignore those records.

Finally, while we introduced drug and gene filters to the interaction search view in the last major update, we have had several requests to define how these filters are implemented. To address this lack of transparency on the UI, we have now added a link to the FAQ page where these filters are now defined.

## IMPROVEMENT OF DRUG NORMALIZATION APPROACH

In addition to the improvement of features related to search results, another main focus of the DGIdb 4.0 update is the improvement of drug grouping and normalization. Previously, we grouped drug claims using a rule-based pairwise association approach. This process was cumbersome, requiring a lengthy and complete re-grouping of all claims whenever we updated sources in order to generate consistent groupings. We revised this approach by creating a normalization component independent of the claims aggregated by DGIdb, that could be run on a per-source basis. When redesigning this part of DGIdb, we took steps to enable reuse of this normalizer as a modular component for other resources. To this end, we leveraged and contributed to a drug normalization service from the Variant

Interpretation for Cancer Consortium (VICC) (the “*thera-py*” Python Package; source code online at <https://github.com/cancervariants/therapy-normalization>). Among our contributions to *thera-py* was a normalizer for the Wikidata(5) resource, further enabling community contributions to assist in concept normalization both for DGIdb and other resources reliant upon the VICC normalization services.

Drug claims from DGIdb were normalized using the ChEMBL and Wikidata normalizers from *thera-py*. Rules were written to formalize grouper behavior based upon match characteristics of a query. Briefly, these rules prioritize matches to primary labels over aliases, exact case over case-insensitive, and ChEMBL over other normalizers. An algorithm for constructing a merged drug concept from normalizer results was specified, enabling a standardized set of aliases for a given concept identifier. Pseudocode for this algorithm is provided (see Supplementary Data), and all implemented code is available on our public repository (see Data Availability).

## TECHNOLOGY IMPROVEMENTS AND IMPLEMENTATION

To handle increased web traffic and integration with other tools, we have upgraded DGIdb to Rails 6 (from Rails 5), upgraded to Ruby 2.6.5 (from Ruby 2.3), upgraded to PostgreSQL 12 (from PostgreSQL 9.6), and upgraded the server to the latest Ubuntu LTS release (20.04). In addition to the new features and performance benefits these upgrades bring, they will ensure that we continue to remain on supported software versions that receive regular security updates.

In order to keep DGIdb’s underlying source data current, we had previously implemented an automated job scheduling framework using DelayedJob to schedule monthly runs of online updaters. In this release, we switched to using Sidekiq (Figure 1). In contrast to DelayedJob, Sidekiq offers a convenient user interface which makes identifying job failure reasons and rescheduling of failed jobs easier. Furthermore, the addition of Airbrake, an online tool for exception tracking, gives error reviews and notifies the development team of these errors in real-time (for instance, via email).

To ease future implementation of fixes and new features, we moved testing to a GitHub continuous integration (CI) workflow which allows us to continuously test newly committed code for errors against multiple versions of Ruby and PostgreSQL.

## SUMMARY AND FUTURE DIRECTIONS

With our most recent release, DGIdb has received significant improvements to source content, functionality such as searching and grouping, and underlying application technology. We have significantly expanded the number of records in our database through the addition of new sources and updates of existing sources. Furthermore, we have improved our ability to maintain regular content updates through the implementation of additional online importers for several sources and the use of Sidekiq for automatic job processing. We have revised our process for drug grouping and normalization to be batched by resource and to leverage continual improvement through community contributions to the VICC *thera-py* normalizers and the Wikidata public-domain crowdsourcing platform. Finally, several important changes have been made to search results and the information presented on the UI. We have implemented more sophisticated notations of relative and static interaction scores, improved the relevance of interaction source link outs wherever possible, and included the concept of directionality for interaction results.

Although the updates in DGIdb 4.0 have improved the usability and content of our resource, we expect there will still be a need for future improvements. One technology improvement on our roadmap is converging the public facing API with the internal code that powers the web views. Ultimately, we want the APIs available for general use to be the same ones powering our HTML pages. This would provide an even more fully featured API to end users while reducing our overall maintenance burden by eliminating redundant code. Additionally, one possible feature to add to the user-interface is displaying information on related genes on the gene page for a given gene. Finally,



we also plan to continue updating sources to online updaters, where possible, and migrating TSV-based sources from the legacy DSL importer style to the updated DGldb 3.0 TSV importer style.

## AVAILABILITY

DGldb is an open access database and web interface ([www.dgldb.org](http://www.dgldb.org)) with open source code available on GitHub (<https://github.com/griffithlab/dgi-db>) under the MIT license. We also provide data downloads for drug claims, gene claims, and interaction claims on the website in addition to a SQL data dump (<http://dgldb.org/downloads>). Information about the API and its endpoints can also be found on the website (<http://dgldb.org/api>).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

## ACKNOWLEDGEMENTS

We want to thank the creators and maintainers of the seven new resources added to DGldb and the many previously incorporated resources, as well as our growing community of users for notifying us of minor and major issues and for their suggestions on new features and improvements to DGldb. We would also like to thank the members of the NDEX team, and in particular Dexter Pratt and Rudolf Pillich for their efforts in integrating DGldb data into the NDEX resource.

## FUNDING

This work was supported by the National Human Genome Research Institute [K99HG010157 to A.H.W. and A.C.C., R00HG007940 to M.G.] and the National Cancer Institute [U01CA209936, U01CA248235, U24CA237719 to M.G., O.L.G., S.K., A.C.C., A.H.W. and J.F.M.]. Funding for open access charge: departmental funding

## CONFLICT OF INTEREST

None declared

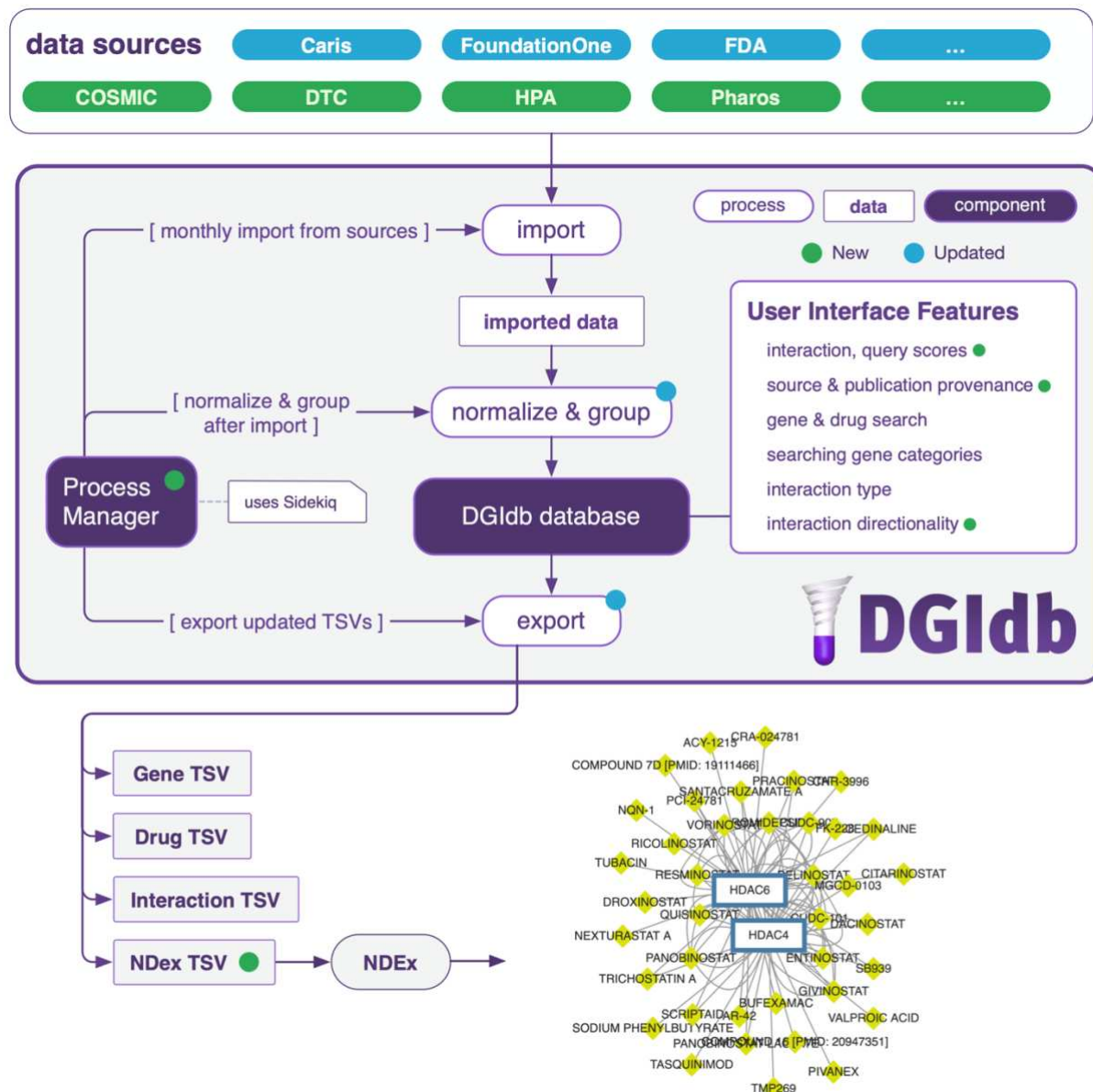
## REFERENCES

1. Griffith,M., Griffith,O.L., Coffman,A.C., Weible,J.V., McMichael,J.F., Spies,N.C., Koval,J., Das,I., Callaway,M.B., Eldred,J.M., et al. (2013) DGldb: mining the druggable genome. *Nat. Methods*, 10, 1209–1210.
2. 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) DGldb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.*, 44, D1036–D1044.
3. Cotto,K.C., Wagner,A.H., Feng,Y.-Y., Kiwala,S., Coffman,A.C., Spies,G., Wollam,A., Spies,N.C., Griffith,O.L. and Griffith,M. (2017) DGldb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res.*, 10.1093/nar/gkx1143.
4. Tanoli,Z., Alam,Z., Vähä-Koskela,M., Ravikumar,B., Malyutina,A., Jaiswal,A., Tang,J., Wennerberg,K. and Aittokallio,T. (2018) Drug Target Commons 2.0: a community platform for systematic analysis of drug-target interaction profiles. *Database*, 2018, 1–13.
5. Vrandečić,D. and Krötzsch,M. (2014) Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57, 78–85.
6. Pratt,D., Chen,J., Pillich,R., Rynkov,V., Gary,A., Demchak,B. and Ideker,T. (2017) NDEX 2.0: A Clearinghouse for Research on Cancer Pathways. *Cancer Res.*, 77, e58–e61.

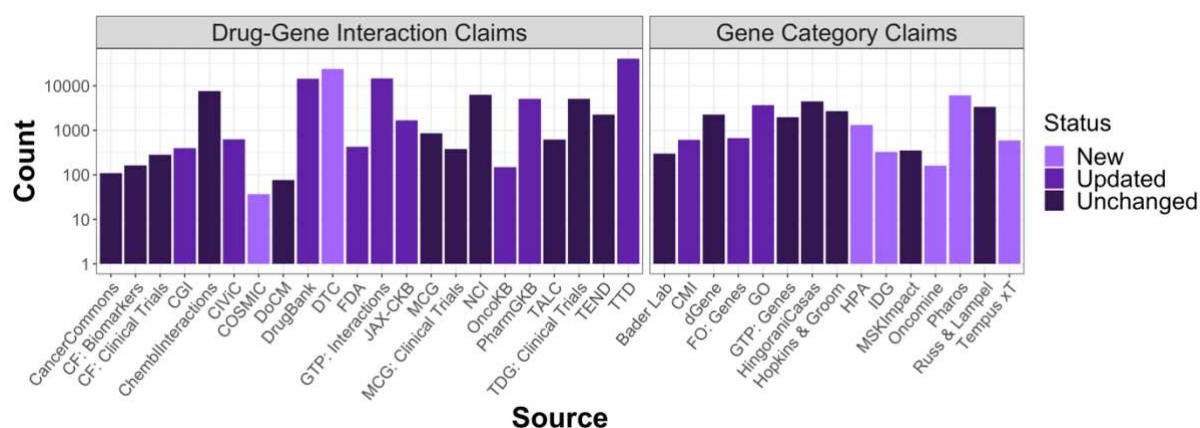
7. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*, 47, D941–D947.
8. Beaubier, N., Tell, R., Lau, D., Parsons, J.R., Bush, S., Perera, J., Sorrells, S., Baker, T., Chang, A., Michuda, J., et al. (2019) Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. *Oncotarget*, 10, 2384–2396.
9. Rodgers, G., Austin, C., Anderson, J., Pawlyk, A., Colvis, C., Margolis, R. and Baker, J. (2018) Glimmers in illuminating the druggable genome. *Nat. Rev. Drug Discov.*, 17, 301–302.
10. Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017) A pathology atlas of the human cancer transcriptome. *Science*, 357.
11. Williams, H.L., Walsh, K., Diamond, A., Oniscu, A. and Deans, Z.C. (2018) Validation of the OncoPrint focus panel for next-generation sequencing of clinical tumour samples. *Virchows Arch.*, 473, 489–503.
12. Nguyen, D.-T., Mathias, S., Bologna, C., Brunak, S., Fernandez, N., Gaulton, A., Hersey, A., Holmes, J., Jensen, L.J., Karlsson, A., et al. (2017) Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, 45, D995–D1002.
13. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46, D1074–D1082.
14. Armstrong, J.F., Faccenda, E., Harding, S.D., Pawson, A.J., Southan, C., Sharman, J.L., Campo, B., Cavanagh, D.R., Alexander, S.P.H., Davenport, A.P., et al. (2020) The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res.*, 48, D1006–D1021.
15. 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.
16. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47, D330–D338.
17. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017) OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*, 2017.
18. Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B. and Klein, T.E. (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, 92, 414–417.
19. Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., Zhang, R., Zhu, J., Ren, Y., Tan, Y., et al. (2020) Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.*, 48, D1031–D1041.
20. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R., et al. (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, 43, D36–42.
21. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020) Ensembl 2020. *Nucleic Acids Res.*, 48, D682–D688.
22. Griffith, M., Spies, N.C., Krysiak, K., McMichael, J.F., Coffman, A.C., Danos, A.M., Ainscough, B.J., Ramirez, C.A., Rieke, D.T., Kujan, L., et al. (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.*, 49, 170–174.
23. Patterson, S.E., Liu, R., Statz, C.M., Durkin, D., Lakshminarayana, A. and Mockus, S.M. (2016) The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum. Genomics*, 10, 4.
24. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazon, Y., et al. (2016) The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics*, 54, 1.30.1–1.30.33.

25. Wu,C., Jin,X., Tsueng,G., Afrasiabi,C. and Su,A.I. (2016) BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic Acids Res.*, 44, D313–6.
26. Wang,C., Yang,J., Luo,H., Wang,K., Wang,Y., Xiao,Z.-X., Tao,X., Jiang,H. and Cai,H. (2020) CancerTracer: a curated database for inpatient tumor heterogeneity. *Nucleic Acids Res.*, 48, D797–D806.
27. Zhao,G., Li,K., Li,B., Wang,Z., Fang,Z., Wang,X., Zhang,Y., Luo,T., Zhou,Q., Wang,L., et al. (2020) Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res.*, 48, D913–D926.
28. Deng,X., Das,S., Valdez,K., Camphausen,K. and Shankavaram,U. (2019) SL-BioDP: Multi-Cancer Interactive Tool for Prediction of Synthetic Lethality and Response to Cancer Treatment. *Cancers* , 11.
29. De Cesco,S.P., Davis,J.B. and Brennan,P.E. (2020) TargetDB: A target information aggregation tool and tractability predictor. 10.1101/2020.04.21.052878.
30. Nicholas,T.J., Cormier,M.J., Huang,X., Qiao,Y., Marth,G.T. and Quinlan,A.R. (2020) OncoGEMINI: Software for Investigating Tumor Variants From Multiple Biopsies With Integrated Cancer Annotations. 10.1101/2020.03.10.979591, pre-print, not peer-reviewed.





**Figure 1:** Overview of main components of DGldb. Data sources are imported from outside resources (over 40 as of DGldb 4.0), normalized and grouped with internal processes to prepare records to be displayed in DGldb, and exported to TSV downloads or integrated with other resources. Process management is handled by Sidekiq for automation of importing, normalization and grouping, and exporting. A subset of new data sources are highlighted in green, a subset of updated pre-existing data sources are highlighted in blue. The updated sources highlighted in this figure are some of the sources that have been updated through manual curation. Information on additional sources and their status in DGldb 4.0 can be found in Figure 2 and Supplemental Table 1. New features and technologies from DGldb 4.0 are indicated with green dots, pre-existing features and technologies that have been updated are indicated with blue dots.



**Figure 2:** DGIdb 4.0 content by source. The number of drug-gene interaction claims (first panel) and druggable gene categories (second panel) are separated into three categories: sources that are new, sources that existed in the DGIdb previously but have been updated for 4.0, or sources that existed previously but have not been updated. Abbreviations: CF = Clarity Foundation, CGI = Cancer Genome Interpreter, CMI = Caris Molecular Intelligence, FO = Foundation One, GO = Gene Ontology, GTP = Guide to Pharmacology, HPA = Human Protein Atlas, IDG = Illuminating the Druggable Genome, JAX-CKB = JAX-Clinical Knowledgebase, MCG = My Cancer Genome, MSK = Memorial Sloan Kettering, OncoKB = Precision Oncology Knowledge Base, TALC = Targeted Agents in Lung Cancer, TDG = The Druggable Genome, TEND = Trends in the Exploration of Novel Drug targets, and TTD = Therapeutic Target Database.

**A**

$$\text{relative drug specificity} = \frac{\text{average known gene partners for all drugs}}{\text{known gene partners for drug } d}$$

$$\text{relative gene specificity} = \frac{\text{average known drug partners for all genes}}{\text{known drug partners for gene } g}$$

$$\text{evidence score} = \text{publication count} + \text{source count}$$

Query Interaction Score (gene search):

$$\text{evidence score} \cdot \text{queried genes interacting with drug } d \cdot \text{relative drug specificity}$$

Query Interaction Score (drug search):

$$\text{evidence score} \cdot \text{queried drugs interacting with gene } g \cdot \text{relative gene specificity}$$

Interaction Score (no query):

$$\text{evidence score} \cdot \text{relative drug specificity} \cdot \text{relative gene specificity}$$

**B**

**Interaction Search Results** drug-gene interactions for your genes

Unique Matches Ambiguous or Unmatched

Search Term: "MEK1" ▶ MAP2K1

Drug	Interaction Type & Directionality	Sources	PMIDs	Query Score	Interaction Score
AZD-8330	inhibitor (inhibitory)	TALC MyCancerGenome TdpClinicalTrial ChemBioInteractions CKB	29733091	12.06	1.2
COBIMETINIB	allosteric modulator, inhibitor (inhibitory)	TALC MyCancerGenome TdpClinicalTrial ChemBioInteractions CancerClinicalTrials ChemBioInteractions CKB DrugBank GuideToPharmacologyInteractions CancerCommons OncoKB MyCancerGenomeClinicalTrial	31227918 30341304 20569975 32123303 26384768 3087262	9.76	0.97
CI-1040	allosteric modulator, inhibitor (inhibitory)	MyCancerGenome ChemBioInteractions CKB GuideToPharmacologyInteractions	12370306 10910144	13.78	0.69

Download as TSV

**Preset Filters**

**Unique Matches**

- ☒ Search Term: MEK1
- ☒ Search Term: MEK2

**Interaction Search Results** drug-gene interactions for your genes

Unique Matches Ambiguous or Unmatched

Search Term: "MEK1" ▶ MAP2K1

Drug	Interaction Type & Directionality	Sources	PMIDs	Query Score	Interaction Score
COBIMETINIB	allosteric modulator, inhibitor (inhibitory)	TALC MyCancerGenome TdpClinicalTrial ChemBioInteractions CancerClinicalTrials ChemBioInteractions CKB DrugBank GuideToPharmacologyInteractions CancerCommons OncoKB MyCancerGenomeClinicalTrial	31227918 30341304 20569975 32123303 26384768 3087262	19.52	0.97
CI-1040	allosteric modulator, inhibitor (inhibitory)	MyCancerGenome ChemBioInteractions CKB GuideToPharmacologyInteractions	12370306 10910144	13.78	0.69

Download as TSV

**Preset Filters**

**Unique Matches**

- ☒ Search Term: MEK1
- ☒ Search Term: MEK2
- ☒ Search Term: BRAF
- ☒ Search Term: KRAS

**C**

**ZEB1 AND SALINOMYCIN** Interaction Record

Summary Claims

**SALINOMYCIN** ▶ **ZEB1**

**Main Info:**

Gene	ZEB1
Drug	SALINOMYCIN
Interaction Types & Directionality	n/a
Interaction Score	11.47

**Publications:**

Sánchez-Tilló et al., 2014, The EMT activator ZEB1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma., Cell Death Differ.

**ZEB1 AND DOXORUBICIN** Interaction Record

Summary Claims

**DOXORUBICIN** ▶ **ZEB1**

**Main Info:**

Gene	ZEB1
Drug	DOXORUBICIN
Interaction Types & Directionality	n/a
Interaction Score	0.13

**Publications:**

Sánchez-Tilló et al., 2014, The EMT activator ZEB1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma., Cell Death Differ.

**Figure 3:** Overview of DGIdb's new query interaction scores and interaction scores. **A)** Schematic of how each of the new scores is calculated within DGIdb. Gene and drug queries both return a query interaction score that is dependent on the search terms. Each interaction has a score that is calculated independently of other search terms. **B)** Example of a query interaction score changing based on the terms searched. In the first panel, only *MEK1* and *MEK2* were searched and the score for the interaction between *MEK1* and Cobimetinib was 9.76. In the second panel, *BRAF* and *KRAS* were added to the search query. These both interact with Cobimetinib and thus raise the query interaction score. **C)** Example of interaction score. The panel on the left shows the interaction between *ZEB1* and Salinomycin. This is only interaction for Salinomycin and thus it has a high interaction score. The panel on the right shows the interaction between *ZEB1* and Doxorubicin. Doxorubicin is involved in 87 interactions within DGIdb and thus has a much lower interaction score.