DGIdb 3.0: a redesign and expansion of the drug–gene interaction database

Kelsy C. Cotto[1,†], Alex H. Wagner[1,*,†], Yang-Yang Feng[1], Susanna Kiwala[1], Adam C. Coffman[1], Gregory Spies[1], Alex Wollam[1], Nicholas C. Spies[1], Obi L. Griffith[1,2,3,4,*] and Malachi Griffith[1,2,3,4,*]

[1]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO 63108, USA, [2]Siteman Cancer Center, Washington University School of Medicine, St Louis, MO 63110, USA, [3]Department of Medicine, Washington University School of Medicine, St Louis, MO 63110, USA and [4]Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

ABSTRACT

The drug–gene interaction database (DGIdb, www.dgidb.org) consolidates, organizes and presents drug–gene interactions and gene druggability information from papers, databases and web resources. DGIdb normalizes content from 30 disparate sources and allows for user-friendly advanced browsing, searching and filtering for ease of access through an intuitive web user interface, application programming interface (API) and public cloud-based server image. DGIdb v3.0 represents a major update of the database. Nine of the previously included 24 sources were updated. Six new resources were added, bringing the total number of sources to 30. These updates and additions of sources have cumulatively resulted in 56 309 interaction claims. This has also substantially expanded the comprehensive catalogue of druggable genes and anti-neoplastic drug–gene interactions included in the DGIdb. Along with these content updates, v3.0 has received a major overhaul of its codebase, including an updated user interface, preset interaction search filters, consolidation of interaction information into interaction groups, greatly improved search response times and upgrading the underlying web application framework. In addition, the expanded API features new endpoints which allow users to extract more detailed information about queried drugs, genes and drug–gene interactions, including listings of PubMed IDs, interaction type and other interaction metadata.

INTRODUCTION

The drug–gene interaction database (DGIdb, www.dgidb.org) was first released in 2013 to consolidate drug–gene interactions and potentially druggable genes into a single resource with a powerful interface to query these data (1). DGIdb 2.0 was released in 2016 and featured substantial content updates, a more intuitive user interface and the inclusion of an application programming interface (API) (2). Despite the success of the DGIdb, 2.0 became outdated in several ways. New sources have become available since 2.0's release and many existing sources within the database required updates. The notion of grouping *drug claims* (an assertion of a drug concept by a constituent resource) to match a canonical drug source (PubChem compounds), and grouping *gene claims* (an assertion of a gene concept by a constituent resource) to a canonical gene source (NCBI Entrez Gene) was described in the 2.0 paper (3,4). In brief, grouping allows the DGIdb to relate disparate representations of a gene or drug concept through a core entity, or *group*. As the database grew, the grouping success rates dropped and new classes of drugs were not fully represented in the search results. Additionally, while the DGIdb grouped drug and gene claims, interactions were still reported at the claim level, with no concept of interaction groups. Furthermore, the initial API that was released alongside 2.0 had limited endpoints that prevented users from extracting additional information about their queried drug–gene interactions, such as PubMed IDs (PMIDs) supporting the interactions and other interaction metadata. Finally, the DGIdb has grown enormously since its initial publication in 2013, and as a result, interaction searches had become significantly slower. This new release addresses these limitations, providing an improved user interface, quicker response times for searches and new meth-

ods for accessing and incorporating the DGIdb into bioinformatics workflows.

EXPANDED CONTENT

For 3.0, there has been a substantial expansion of the content of the database through the addition of new sources and the updating of existing sources. Six new sources have been added, which brings the total number of sources represented by the DGIdb to 30. Three of these new sources provide drug–gene interactions from prominent expert-curated databases of clinically actionable variants similar to the CIViC and Clearity Foundation sources, already included in the DGIdb (Supplementary Table S1) (5). These include the Precision Oncology Knowledge Base (OncoKB), cancer genome interpreter (CGI) and The Jackson Laboratory Clinical Knowledgebase (CKB) (Supplementary Table S1) (6,7). From these three new sources, 2822 new interaction claims were added. These sources were added as anti-neoplastic interaction sources, resulting in an 36.7% increase in anti-neoplastic drugs. Drug–gene interactions were also added from the United States Food and Drug Administration (FDA) Pharmacogenomics website and the National Cancer Institute (NCI) Cancer Gene Index, resulting in an additional 276 and 6231 interaction claims, respectively (Supplementary Table S1). In addition to these sources, the existing druggable gene category sources were expanded with data extracted from a new druggable genome paper that used computational approaches to identify druggable genes from genome-wide association studies (Supplementary Table S2) (8). The inclusion of this source into the DGIdb results in an additional 2300 potentially druggable genes, a 58% increase from 2.0 (Figure 1). Notably, this independent definition of the druggable genome almost completely encapsulates many other gene categories (e.g. kinases, G-protein coupled receptors) that are expected to be good drug targets (Supplementary Figure S1).

As the DGIdb has matured, maintaining current representations of existing sources has become increasingly important relative to identifying new interaction sources. Curated updates of drug–gene interaction data from My Cancer Genome and the targeted and biologic therapies for non-small-cell lung cancer (TALC) have occurred, resulting in moderate increases in drug–gene interactions from these sources (9,10). To help prevent regularly updated sources from becoming outdated in the DGIdb, we have rewritten and expanded *online updaters* to keep frequently updated sources current within the database. With these more frequent, incremental updates to the DGIdb, there has been a substantial increase in drug–gene interaction content from IUPHAR's Guide to Pharmacology and CIViC, a 23 and 200% increase, respectively (5,11). Additionally, a major update of DrugBank from version 4.3 to 5.0 was performed, resulting in a 30% increase in interaction claims from this source (12). Gene Ontology was also updated which provided a moderate increase for the druggable gene categories (13). For this release, 18 493 new interaction claims were added, of which 50% were from updated sources and 50% from new sources. In total, there are now 6106 druggable genes and 29 783 drug–gene interactions, which cover 41 100 genes and 9495 drugs, within the DGIdb.

The enhancements to the online updaters have also been applied to Entrez Gene, from which 99% of all gene claims made by the DGIdb constituent sources were grouped (Supplementary Figure S2) (3). Another major change from 2.0 to 3.0 is that the canonical drug source for the DGIdb has switched from using PubChem compounds to ChEMBL molecules (14). This switch has added 1.7 million ChEMBL molecules to the database for potential matching to drug claims. Importantly, switching to ChEMBL has added 195 antibody drugs (e.g. trastuzumab, cetuximab), a drug class that is absent from the PubChem database and frequently requested by users. These antibody drugs matched to 539 distinct drug claims from the constituent sources of the DGIdb. With ChEMBL as the canonical drug source and the improvements to the grouping strategy below, 80.2% of all drug claims now group. Many of the resources we pull from strive to be as comprehensive as possible, and sometimes include broad classes of drug or therapy (e.g. 'hormone therapy', 'mtor inhibitors', 'chemotherapy', 'radiation', 'antibiotics', etc.), which account for a large percentage of the remaining drug claims.

NEW FEATURES AND ENHANCEMENTS

A major feature added in 3.0 is a user-selectable series of preset filters. These allow users to select drug–gene interactions that align with some of the most common search use cases. While it was possible for a user to filter on some of the following attributes in 2.0, these concepts have now been structured as stand-alone filters that can help focus a user's search results with a single click. The filters currently include FDA-approved drugs, anti-neoplastic drugs, immunotherapy drugs, clinically actionable genes, genes included in the druggable genome definition and drug resistant genes. *FDA approval status* is extracted from ChEMBL v23 (current as of this writing) (14). *Anti-neoplastic drugs* are defined by inclusion in an anti-neoplastic drug–gene interaction source (e.g. My Cancer Genome), or as a drug with an anti-neoplastic attribute from its constituent source. Immunotherapy drugs are defined as any drug with an attribute of 'immunosuppressive agent', 'immunomodulatory agent' or 'immunostimulant'. *Clinically actionable* genes are genes that constitute the DGIdb 'clinically actionable' gene category, and by definition is used to inform clinical action (e.g. the Foundation One diagnostic gene panels). Similarly, *druggable genome* genes are genes listed in the DGIdb 'druggable genome' gene category (8,11,13,15–19). *Drug resistance* genes are defined by the Gene Ontology as genes that confer drug resistance or susceptibility (GO identifier 0042493), and are maintained in the DGIdb through the 'drug resistance' gene category (13). To incorporate these new filters, we have redesigned the data model, expanded the definition of the druggable genome and restructured the user interface (UI) to include a redesigned search form and results interface. The UI also now features drug, gene and interaction views to leverage the new grouping strategies and preset filters.

Since the last release, several changes have been made to the drug grouping strategy to improve overall grouping percentages. We have added support for fuzzy searching when grouping if direct matches to drug groups are not found, en-
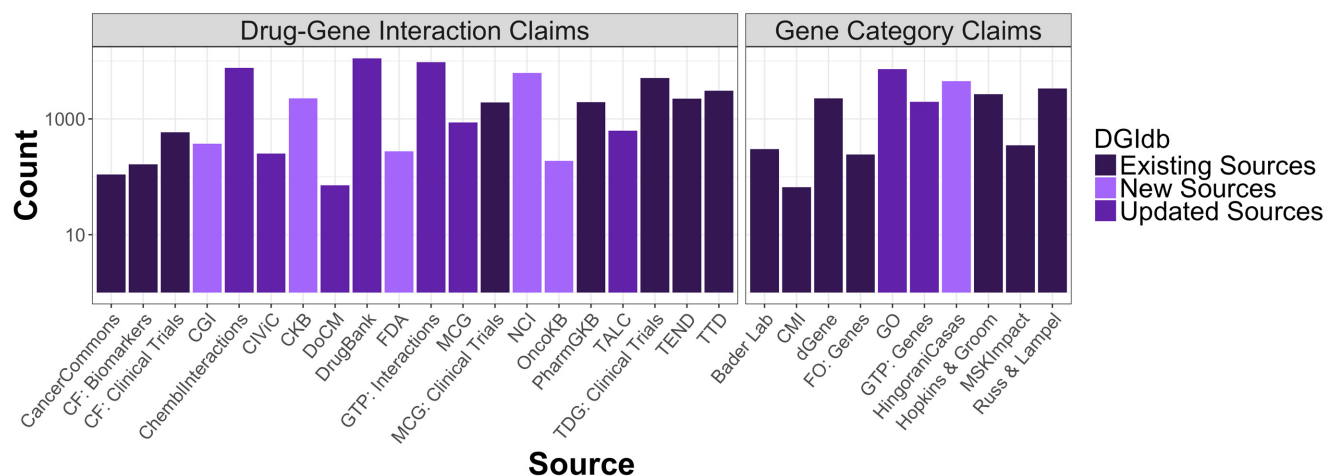
Figure 1. DGIdb 3.0 content by source. The number of drug–gene interaction claims (first panel) and druggable gene categories (second panel) are separated into three categories: sources that existed in the DGIdb previously, sources that existed in the DGIdb previously but have been updated for 3.0 or sources that are entirely new to the DGIdb. Abbreviations: CF: Clearity Foundation; CGI = Cancer genome interpreter, CKB = JAX-Clinical Knowledgebase, CMI = Caris Molecular Intelligence, FO = Foundation One, GTP = Guide to pharmacology, MCG = My cancer genome, OncoKB = Precision Oncology Knowledge Base, TALC = Targeted agents in lung cancer, TTD = Therapeutic Target Database, TEND = Trends in the exploration of novel drug targets, GO = Gene Ontology and MSK = Memorial Sloan Kettering.

abling grouping of drug claims with slightly different means of joining multi-word terms (whitespace, dashes and underscores). Code has also been added to identify and remove aliases that are highly ambiguous to improve exact matching of drug claims.

In addition to improved drug grouping, we have added interaction grouping, linking together interaction claims from multiple sources that describe the exact same drug–gene interaction. The grouping efficiency for interactions is 75.2%. The successful grouping of an interaction claim is dependent on the associated gene claims and drug claims successfully grouping. As a result, interaction grouping percentages are closely related to the grouping percentages for drugs and genes (Supplementary Figure S2). The introduction of interaction groups has led to a noticeable improvement in response times on the UI for interaction searches, even when query sizes and result sizes grow larger (Figure 2). Before interaction groups were created, searches in 2.0 had to query almost twice as many database tables as in 3.0. Interaction groups allow for more efficient queries, which leads to a 14-fold reduction in response times when searching the DGIdb.

We have expanded and added additional API endpoints that now allow users to extract more information on the drug–gene interactions provided by the DGIdb. These include endpoints to list all the interaction groups, gene groups and drug groups in the DGIdb, as well as endpoints to view detailed information about an individual interaction, gene or drug group. Drug/gene/interaction group endpoints include various metadata about the group including its constituent claims. The information presented in the interaction search results endpoint has been expanded to include the Entrez and Chembl ID of the gene and drug involved in an interaction, as well as a list of publications that support each interaction. Users can also apply the preset filters to their interaction search through the API. This brings the interaction search endpoint in sync with the user interface. Moreover, the information presented in the interaction search results and the new individual endpoints for interactions, genes and drugs more closely mirrors the views available via the user interface. These new endpoints allow users to more efficiently export all of the data available in the DGIdb.

USER INTERFACE (UI) UPDATES

The 3.0 release of the DGIdb features a dramatic overhaul of the UI that reflects many of the backend changes detailed above. The search interactions page now allows the user to apply the new preset filters in addition to the existing advanced filters (e.g. source databases, gene categories and interaction types) directly on the search form (Figure 3). On performing a search, the user is redirected to the redesigned search interactions results page, enabled by a new concept of interaction groups instead of individual interaction claims. This new view displays search results as lists of visually distinct panels for each search term mapped to gene or drug groups. A list of uniquely matched terms is shown in one tab whereas a second tab summarizes the list of ambiguously matched search terms and unmatched terms. Each panel displays a table of interaction groups, along with summaries of the respective interaction types, sources, supporting publication PMIDs and a ranking. Users interested in a particular interaction can navigate from the interaction search results to the corresponding interaction group view. Similarly, detailed information about an interacting gene or drug can be found in the associated gene or drug group view (Supplementary Figure S3). These new group views feature (i) a *summary tab*, which details all relevant information collected from the various source claims (e.g. aliases, FDA approval status, supporting publications); (ii) an *interactions tab* (for gene and drug group views), which lists summary panels of each interaction in the DGIdb for this gene or
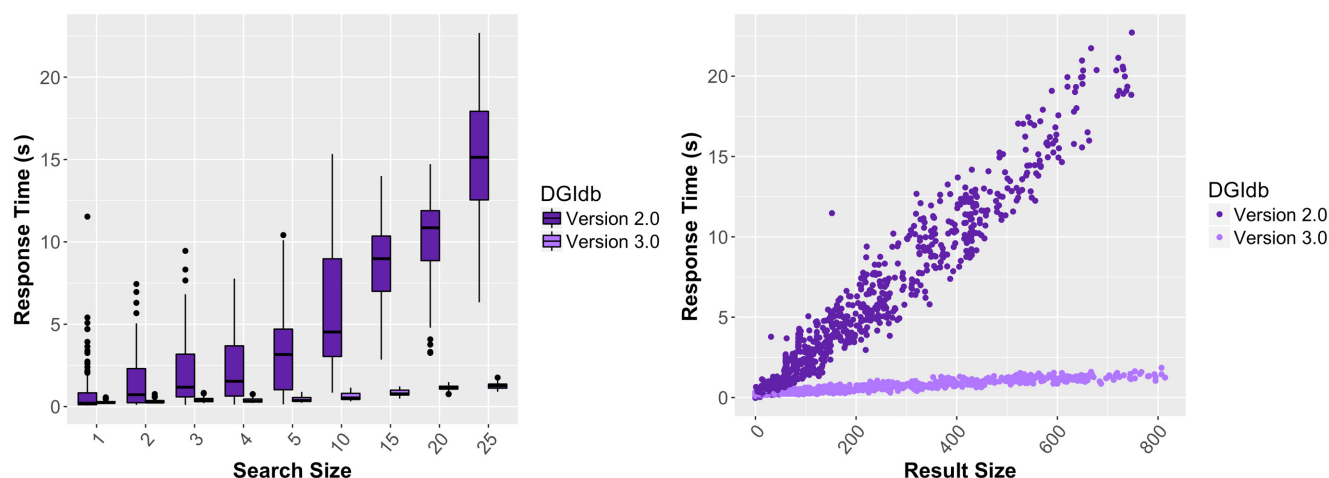
Figure 2. Comparison of the DGIdb 2.0 and 3.0 response time by search and result size. The left panel summarizes observed response times ($n = 100$) for searches of various sizes (1, 2, 3, 4, 5, 10, 15, 20 and 25) of randomly selected genes. The right panel summarizes observed response times for the corresponding interaction search result sizes.

drug; and (iii) a *claims tab* with detailed information about each claim supporting the group.

USAGE AND ACCESSIBILITY

The utility of the DGIdb as a resource is reflected in its substantial web and especially API traffic. The website receives ∼1700 unique users and ∼2800 sessions a month, an increase of 42 and 39%, respectively, since the initial release of the DGIdb. Additionally, 42% of unique visitors within the past year were returning users. Within the last year, the DGIdb API has been used by 3689 unique IP addresses for a total of 939 429 requests, as seen in Supplementary Figure S4. At the time of publication, 2.0 was used by a number of bioinformatics tools in the development of their platforms. DGIdb has since been integrated into four additional tools, bringing the total number of known platforms that utilize the DGIdb to 10. These platforms are GeneCards (http://www.genecards.org), CVE (20), rDGIdb (21), PANDA (22), iCAGES (icages.usc.edu), BioGPS (23), Omics Pipe (24), GEMINI (25), StationX (www.stationxinc.com) and IHLDB.rf (www.lungcancerdatabase.com). These integrations highlight the accessibility and utility of the DGIdb API, which has been expanded to include new endpoints in 3.0.

To handle the increasing usage of the DGIdb, the database backend has undergone significant updates. One of the most notable is the upgrade from Rails 3 to the Rails 5 framework. We also added 22 new tables to the PostgreSQL database schema consisting of 2 326 676 records supporting the described changes and added 4065 new lines of code to the repository, an expansion of the codebase by 18.9%. This activity can be seen through the commit history to the GitHub repository in Supplementary Figure S5. For access to this code and data, please see the availability section below.

The open source DGIdb software was previously available under the GNU General Public License 3.0 (GPL3) and has been re-released under the more permissive MIT

license. The data contained within the DGIdb are available under the licenses assigned by their host sources which makes it possible to integrate the DGIdb into any workflow. Since 2.0, the database has moved from being hosted on a local server to being hosted on Amazon Web Services. With this change in server host, the availability of the DGIdb has been expanded through the release of two Amazon Machine Images, a development environment and production environment. By providing this production environment for users, we now support a quick-start solution for private, cloud-based applications.

CONCLUSION AND FUTURE DIRECTIONS

DGIdb 3.0 has undergone significant changes. The number of drug–gene interactions and druggable genome definitions has been substantially expanded through the updating of existing sources and inclusion of new sources. To improve searches, we have included preset filters that allow searches for commonly requested gene or drug classes. To utilize these filters, a more intuitive UI was created and new fields to the database schema were added. Additionally, there are new ways to both access the data within the DGIdb and to deploy local or cloud instances of the database.

While the DGIdb remains a powerful resource for querying drug–gene interactions, we anticipate future changes that will improve the user experience and the content within it. One change is that claim information will be linked to the licensing terms from constituent sources through the various methods of accessing the DGIdb including the web interface, API endpoints and data downloads. Second, while grouping statistics for claims have improved in the DGIdb 3.0, there are still a significant percentage of claims that remain ungrouped. To address this, the drug grouper will need to be optimized to handle the exceptions that currently prevent successful grouping. Third, much of the user interface is a series of static renders by the server. A potential future enhancement is a more reactive, client-side web application that would allow for more dynamic visualization and exploration of the data. Finally, other databases (e.g. CIViC)

Figure 3. The DGIdb interaction search interface. (**A**) A search field accepts drug or gene identifiers, depending on which tab is selected and provides auto-completion suggestions for search terms. (**B**) Preset options are provided to filter search results based on the attributes of mapped interactors. (**C**) Advanced filters allow the user to further filter based on source database, gene category or interaction type. (**D**) The interaction search results (Unique Matches) view shows results for search terms that were matched within the DGIdb. (**E**) The interaction search results (Ambiguous or Unmatched) view shows search terms that were either ambiguously matched or unmatched within the DGIdb. (**F**) Additional side panels provide a brief summary of unmatched and ambiguously matched terms.

have had success utilizing a community curation model. In an effort to enable community feedback to address the complexity of grouping and representing drug–gene interaction data, the DGIdb will be moved to a similar model. Expected changes include users creating login accounts, suggesting sources or claims for inclusion and verification that a source or claim has enough support to be included within the DGIdb. These changes will require significant code updates but will greatly increase the utility of the DGIdb.

AVAILABILITY

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

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

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) DGIdb—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) DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.
3. 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–D42.
4. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.
5. 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.
6. Chakravarty,D., Gao,J., Phillips,S., 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.*, **1**, doi:10.1200/PO.17.00011.
7. 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.
8. Finan,C., Gaulton,A., Kruger,F.A., Lumbers,R.T., Shah,T., Engmann,J., Galver,L., Kelley,R., Karlsson,A., Santos,R. *et al.* (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.*, **9**, eaag1166.
9. Yeh,P., Chen,H., Andrews,J., Naser,R., Pao,W. and Horn,L. (2013) DNA-Mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): a catalog of clinically relevant cancer mutations to enable genome-directed anticancer therapy. *Clin. Cancer Res.*, **19**, 1894–1901.
10. Simon,G.R. and Somaiah,N. (2014) A tabulated summary of targeted and biologic therapies for non-small-cell lung cancer. *Clin. Lung Cancer*, **15**, 21–51.
11. 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.
12. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., 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.
13. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
14. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Krüger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
15. 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.
16. Kumar,R.D., Chang,L.-W., Ellis,M.J. and Bose,R. (2013) Prioritizing potentially druggable mutations with dGene: an annotation tool for cancer genome sequencing data. *PLoS One*, **8**, e67980.
17. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
18. Cheng,D.T., Mitchell,T.N., Zehir,A., Shah,R.H., Benayed,R., Syed,A., Chandramohan,R., Liu,Z.Y., Won,H.H., Scott,S.N. *et al.* (2015) Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.*, **17**, 251–264.
19. Russ,A.P. and Lampel,S. (2005) The druggable genome: an update. *Drug Discov. Today*, **10**, 1607–1610.
20. Mock,A., Murphy,S., Morris,J., Marass,F., Rosenfeld,N. and Massie,C. (2017) CVE: an R package for interactive variant prioritisation in precision oncology. *BMC Med. Genomics*, **10**, 37.
21. Thurnherr,T., Singer,F., Stekhoven,D.J. and Beerenwinkel,N. (2016) Genomic variant annotation workflow for clinical applications [version 2; referees: 2 approved]. *F1000Research*, **5**, 1963.
22. Hart,S.N., Moore,R.M., Zimmermann,M.T., Oliver,G.R., Egan,J.B., Bryce,A.H. and Kocher,J.-P.A. (2015) PANDA: pathway and annotation explorer for visualizing and interpreting gene-centric data. *PeerJ*, **3**, e970.
23. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
24. Fisch,K.M., Meißner,T., Gioia,L., Ducom,J.-C., Carland,T.M., Loguercio,S. and Su,A.I. (2015) Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics*, **31**, 1724–1728.
25. Paila,U., Chapman,B.A., Kirchner,R. and Quinlan,A.R. (2013) GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.*, **9**, e1003153.