OXFORD

*Briefings in Bioinformatics,* 22(2), 2021, 1656–1678

**doi: 10.1093/bib/bbaa003**
Advance Access Publication Date: 14 February 2020
Review Article

# Exploration of databases and methods supporting drug repurposing: a comprehensive survey

Ziaurrehman Tanoli, Umair Seemab, Andreas Scherer, Krister Wennerberg, Jing Tang and Markus Vähä-Koskela

Corresponding authors: Ziaurrehman Tanoli, Senior researcher, Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Biomedicum 1, Haartmaninkatu 8, 00290, Helsinki, Finland. Tel: +358 50 318 5639; E-mail: zia.rehman@helsinki.fi; Jing Tang, Assistant Professor, Academy of Finland Research Fellow, Network Pharmacology for Precision Medicine Group, Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Finland, E-mail: jing.tang@helsinki.fi; Markus Vähä-Koskela, senior researcher Institute for Molecular Medicine Finland (FIMM), Biomedicum 2U, P.O.Box 20, Tukholmankatu 8,FI-00014 University of Helsinki, Finland, E-mail: markus.vaha-koskela@helsinki.fi

## Abstract

Drug development involves a deep understanding of the mechanisms of action and possible side effects of each drug, and sometimes results in the identification of new and unexpected uses for drugs, termed as drug repurposing. Both in case of serendipitous observations and systematic mechanistic explorations, confirmation of new indications for a drug requires hypothesis building around relevant drug-related data, such as molecular targets involved, and patient and cellular responses. These datasets are available in public repositories, but apart from sifting through the sheer amount of data imposing computational bottleneck, a major challenge is the difficulty in selecting which databases to use from an increasingly large number of available databases. The database selection is made harder by the lack of an overview of the types of data offered in each database. In order to alleviate these problems and to guide the end user through the drug repurposing efforts, we provide here a survey of 102 of the most promising and drug-relevant databases reported to date. We summarize the target coverage and types of data available in each database and provide several examples of how multi-database exploration can facilitate drug repurposing.

**Key words:** drug repositioning; biomolecular databases; drug databases; disease databases; drug–target interaction databases

## Introduction

Drug repurposing/repositioning is the process of assigning indications for drugs other than the one(s) that they were originally developed for. This definition is somewhat subjective; the cases where drugs are assigned uses in different forms of a class of indications, such as, in different cancer types, which could have been expected, particularly if the mechanism of action is the same, and thus, may not be considered repurposing. Drug repositioning implies the involvement of an unexpected element and is usually distinguished from the utility extension, in which a drug is launched for different forms or stages of the same indication, sharing the same mode of action. Examples of utility extension include the approval of dasatinib for newly diagnosed chronic myeloid leukaemia (CML), having first been approved only for imatinib-relapsed CML [1], and antiangiogenic antibody bevacizumab gaining approval first in colon cancer and later in other solid cancers. Utility extension is also known as market

**Ziaurrehman Tanoli** is a senior researcher at Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland.
**Umair Seemab** is a PhD researcher at Haartman Institute, University of Helsinki, Finland.
**Andreas Scherer** is Research Coordinator at Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland.
**Krister Wennerberg** is a professor and group leader at Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Denmark.
**Jing Tang** is assistant professor at Faculty of medicine, University of Helsinki, Finland.
**Markus Vähä-Koskela** is a senior researcher at Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland.
**Submitted:** 23 September 2019; **Received (in revised form):** 9 December 2019

expansion and is a central drug development strategy for the pharmaceutical industry to ensure revenue as the average cost of developing a drug from scratch ranges from 1 to 2 billion dollars [2, 3], while drugs granted extended uses mainly incur only the regulatory and administrative costs. Discovery of entirely new indications for drugs is also highly lucrative as several development steps can be minimized. For example, clinical trial data existing on adult individuals for the 'morning sickness' drug thalidomide contributed to its approval for multiple myeloma in 2012 at an estimated cost of only $40–80 million [4]. Another example of a profitable genuine drug repositioning is Minoxidil, which was originally developed for hypertension but was later repositioned to treat male hair loss [5]. Similarly, kinase inhibitor imatinib was developed and first approved to target BCR-ABL in CML but was later was found to be effective in targeting KIT in gastrointestinal stromal tumors (GIST) [6].

Apart from drug repositioning and utility extension, in many cases, a distinction can be made for drug development repositioning (sometimes called drug rescue), where a drug candidate developed for a certain indication fails but ends up being used for a different indication [7]. An example of this type of repositioning is Sildenafil (Viagra), which failed to meet its primary endpoints in angina pectoris and hypertension, but instead is very successful as a medicine for erectile dysfunction [8]. Other examples of such drug repositioning include cancer drugs crizotinib, sorafenib, azacitidine and decitabine, all of which failed to reach the markets in their original indications, yet now are important tools in the treatment of other types of cancers [9].

Underlying both traditional drug development and drug repositioning are mechanistic explanations, which depend on sufficient drug target and phenotypic annotations. Since the 1990s and onwards, rapid development in the high throughput screening technologies has created an environment for expediting the discovery process by enabling huge libraries of compounds and molecular targets to be interrogated in a short amount of time. At the same time, advances in the computational methods and availability of public databases have vastly increased the possibility to create novel models and hypotheses for drug mechanisms and to narrow down the top hits by *in silico* validations [10–15].

This has created opportunities to assess the potential for new drug uses even before the experimental testing, which has proven particularly attractive for orphan diseases, in which traditional drug development is limited [16, 17]. In the United States, drug development and clinical research for rare diseases is encouraged by fast track FDA approval and marketing protection and tax alleviations, creating a niche for drug repositioning efforts that can offset the smaller revenue expectations arising from the limited number of patients.

Public data repositories are a considerable asset to drug development, but one of the biggest challenges of elevating drug repositioning to an informed and consistent parallel alternative to primary indication-oriented drug development has been mapping of the mechanisms of action and downstream interactions of the agents [13, 18]. Several reviews have been published highlighting the tools and methodologies leading to drug repositioning. For instance, Dudley *et al.* [19] have published a review on computational methods for drug repositioning and classified the methodologies either as drug or disease-based repositioning. Jin *et al.* [20], have linked existing drug-repositioning methods with their integrated biological and pharmaceutical knowledge and have discussed how to customize a new drug-repositioning pipeline for specific studies. Sam *et al.* [21] have presented web-based tools that can aid in repositioning of the drugs. Li *et al.* [22] have summarized recent progress in computational drug repositioning into four parts: repositioning strategies, computational approaches, validation methods and application areas. Song *et al.* [23] have discussed major tools and resources that have been developed for repositioning the drugs, and Yang *et al.* [24] have recently provided an extensive review on the use of artificial intelligence for drug repurposing. Several databases are being developed every year to support the drug repositioning and are published in the database-related issues of the journals. For instance, Nucleic Acids Research Database Issue contains information on more than 1700 unique databases and 64 new databases [25]. However, another major challenge in drug repositioning is that the mechanistic and phenotypic data necessary to distil new purposes for the 'old' drugs are spread out over a vast and increasing number of data repositories, with data that may vary significantly in quality and reliability. In this review, we have provided a comprehensive review of such publicly available databases by placing them into four categories: chemical, biomolecular, drug-target interaction and disease databases and then further dividing each into subcategories. Furthermore, we have compared the databases using various parameters, such as the number of chemicals, genes, diseases, etc., in order to facilitate the researchers in selecting appropriate databases for specific purposes. We also highlight several new databases that have not been previously covered in any review (e.g. [10, 23]), including: DepMap [26], cBioPortal [27], Probes & Drugs portal [28], DrugComb [29], DrugTargetCommons(DTC) [30], DrugTargetProfiler (DTP) [31], IDAAPM [32], PharmacoDB [33] and DisGeNet [34]. Some of these databases can provide additional datasets and create news ways to support drug repurposing. For example, DrugComb [29] provides combination responses in terms of sensitivity and synergy measures; PharmacoDB [33] has integrated and standardized several drug sensitivity resources and DTC [30] provides a crowd sourcing platform to integrate drug target interactions. The present survey will provide readers a useful comparison on drug repurposing databases and help them to select right database for their analysis.

The rest of the manuscript is structured into three sections, starting in Section 2 by providing an overview about all the databases, including data statistics and subcategories. Section 3 further highlights the recommended databases for each subcategory based on data quality and comprehensiveness. In Section 4, we have provided representative applications of the databases and shown how these databases can be used for drug repositioning. Finally, Section 5 concludes the manuscript with key points.

## Databases providing drug repositioning information

The databases covered in this research are divided into four main categories i.e. chemical, biomolecular, drug–target interaction and disease databases. The main category names are defined after the compilation of database list and assessing their scope. These categories are further divided into subcategories and explained in subsequent sections. Each main section contains a table, providing short introduction and summary statistics for the databases in terms of various parameters (e.g. number of citations, application programming interface (API), number of proteins, chemicals, diseases and protein to disease associations) falling in each of the subcategories for that section.

Several databases fall into more than one subcategory because of the multiple data types. In Figure 1, we have highlighted only those databases that belong to more than

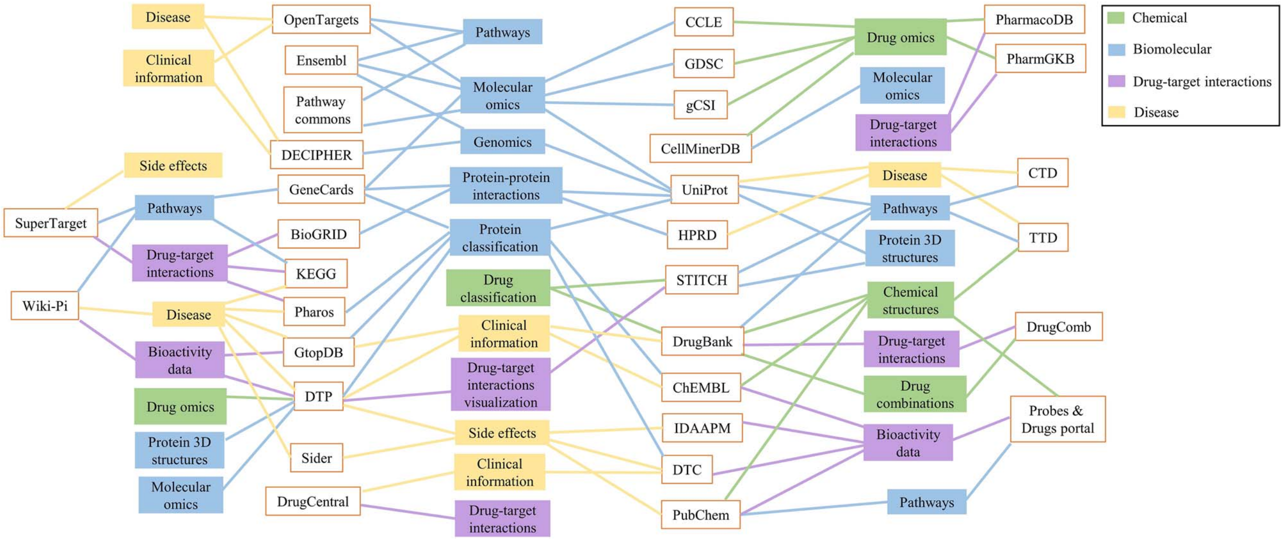

**Figure 1**. Drug repositioning databases categorized into more than one subcategory. Some subcategories are shown more than once in order to facilitate the interpretation of database relationships.

one subcategory in order to highlight heterogeneity in the data types. The subcategories are color coded according to four main categories and the links are also color coded. Drug repositioning can be a complicated process that involves various steps and may require various types of data analysis and experimental validation. Figure 1 can help the researchers to select the most relevant databases, as required by their repositioning applications. In the Tables 1–4, we have listed the databases as a single main category, with the explicit acknowledgement that the databases may also belong to other categories using color-coding scheme as shown in Figure 1. A single main or subcategory for the databases was assigned after thoroughly reading the database descriptions and analyzing data types to determine which datatype is the primary focus and main strength of a database. For example, main strength of the KEGG [35] database is the pathway information, but it also contains additional datatypes, such as drug target interactions, as shown in Figure 1. Hence, we assigned 'Pathways' as primary subcategory for KEGG, which is listed under the main category 'biomolecular', as shown in Table 2. Similarly, main strength of DrugBank [36] is drug target interactions data, which is clearly visible by reading its description and checking the data statistics, but it also provides clinical, drug classification, chemical structures, pathways and drug combination information. Figure 1 can be especially helpful for researchers who are interested in various datatypes or cross-database comparative analysis.

## Chemical databases

There are 12 chemical databases, which have been further divided into four subcategories: drug combination, drug classification, chemical structures and drug omics databases. The subcategory 'Drug combination' contains databases with screening data for combinatorial therapies. Drug classification databases provide classification for the drugs based on the mechanism of action, structure similarity or other parameters. The subcategory 'Chemical Structures' contains databases with data on chemical structures. Drug omics databases contain drug response data for various cancer cell lines. The detailed information about individual databases is shown in Table 1. Most of the databases that fall into the chemical category do not provide application programming interface (API) i.e. no tick mark under API column in the Table 1. The last column (Cit) shows the sum of the citations for all publications on a database and was computed on 25 May 2019, 18:00 CET. This number might increase as more and more researchers will be citing these databases, but it can give an idea about which of the databases are being mostly used. Reference information for some of the databases is missing as we could not find associated publications.

## Biomolecular databases

There are 52 biomolecular databases that we have covered in this review. For a better understanding, we have divided the databases into seven subcategories: genomics, proteomics, protein classification, protein 3D structures, molecular omics, protein-protein interactions and pathways databases. Genomics databases comprise of gene visualizations, whereas proteomics databases contain protein sequences and visualizations. Protein 3D structure databases have comprehensive data on protein structures. Molecular omics databases comprise RNA or protein expression-related information. The subcategory 'protein–protein interactions' contains databases on protein-protein interactions and protein complexes. The subcategory 'pathways' lists databases with signaling pathways and 'protein classification' comprises databases that can provide some sort of classification system for proteins or genes. More details on individual databases are shown in Table 2, and in case the information is not available, those sections in the table are left blank and the column headers are abbreviated at the bottom of the table.

## Drug–target interactions databases

We have divided this category into three subcategories: bioactivity data, drug–target binary interactions and drug–target interaction visualizations. We covered 17 databases, which contain data on either bioactivity or drug–target interactions. The subcategory 'bioactivity data' lists only those databases that report compound-target interactions in terms of dose response

**Table 1.** List of chemical databases; some databases may also fall into more than one (main) category (+chemical, −biomolecular, ∗drug target interactions, •disease)

| Category | Database name | Link | Key features | API | Ref | Cit ≥ |
|---|---|---|---|---|---|---|
| Drug combinations | DrugComb+,∗ | https://drugcomb.fimm.fi/ | Drug combinations tested on a variety of cancer cell lines (466 K drug combinations tested on 2204 cancer cell lines) | ✓ | [29] | 0 |
| | DrugCombDB+ | http://drugcombdb.denglab.org/s | Contains high-throughput screening assays of drug combinations, 561 drugs and 104 cancer cell lines | | [37] | 0 |
| | Drug Combination Database (DCDB)+ | http://www.cls.zju.edu.cn/dcdb/ | Provides combined activity/indications for each drug combination. 1363 drug combinations, 904 individual drugs and 805 targets | | [38] | 114 |
| Drug classification | Medsafe+ | https://medsafe.govt.nz/profs/class/classintro.asp | Provides drug classification into 16 classes | | | |
| | Anatomical Therapeutic Chemical (ATC)+ | https://www.whocc.no/atc/structure_and_principles/ | Provide drug classification into five levels | | | |
| Chemical structures | ChemSpider+ | http://www.chemspider.com/ | Contains >67 million chemical structures and physiochemical properties | ✓ | [39] | 0 |
| | ChemDB+ | http://cdb.ics.uci.edu/ | Contains ∼65 million chemical structures and molecular properties. It also predicts the 3D structures of molecules | | [40] | 101 |
| Drug omics | Connectivity Map (CMAP)+,− | https://clue.io/cmap | A genome-scale library of cellular signatures that catalogs transcriptional responses to chemical and genetic perturbation. It contains one million profiles resulting from perturbations of multiple cell types | | [41] | 3122 |
| | Genomics of drug sensitivity in cancer (GDSC)+,− | http://www.cancerrxgene.org/ | Cancer-driven alterations identified in 11 289 tumors from 29 tissues (integrating somatic mutations, copy number alterations, DNA methylation, and gene expression), mapped onto 1001 human cancer cell lines and correlated with sensitivity to 265 compounds | ✓ | [42] | 388 |
| | Cancer Therapeutics Response Portal (CTRP)+ | https://portals.broadinstitute.org/ctrp/ | Multidimensional enrichment analysis to explore the associations between groups of small molecules and groups of cancer cell-lines in a new quantitative sensitivity dataset | | [43] | 559 |
| | Profiling Relative Inhibition Simultaneously in Mixtures (PRISM)+,− | https://depmap.org/portal/prism/ | PRISM is a powerful approach to rapidly screen thousands of drugs across hundreds of human cancer models on an unprecedented scale | | [44] | 63 |
| | PharmacoDB+,∗ | https://pharmacodb.pmgenomics.ca/ | A web-application assembling the largest *in vitro* drug screens in a single database and allowing users to easily query the union of studies released to date. Drugs: 759, targets: 621, cell lines: 1691, tissues: 41 | | [33] | 70 |

Ref = reference article, Cit = total number of citations for each individual article for the database (computed on 25 May 2019, 18:00 CET).

**Table 2.** Databases on genomics, protein 3D structures, protein classification, protein-protein interaction, pathways and molecular omics.; Some databases may also fall in more than one (main) category (+chemical, −biomolecular, ∗drug target interactions, •disease)

| Category | Database name | Link | Key features | V | S | NC | API | Ref | Cit |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | ≥ | | | | ≥ |
| Genomics | Vega genome browser− | http://vega.archive.ensembl.org/index.html | High-quality gene models produced by the manual annotation of vertebrate genomes | ✓ | 10 | ✓ | ✓ | [45] | 534 |
| | UCSC genome browser− | https://genome.ucsc.edu/ | Displays assembly contigs and gaps, mRNA, multiple gene predictions, cross-species homologies, repeats | ✓ | 46 | ✓ | ✓ | [46] | 15 712 |
| | Ensembl− | https://www.ensembl.org/index.html | Display gene annotation and predicted gene locations for analysis | ✓ | 5K | ✓ | ✓ | [47] | 36 484 |
| | GenBank− | https://www.ncbi.nlm.nih.gov/genbank/ | Contains annotated collection of all publicly available DNA sequences | | 0.4M | | | [48] | 3761 |
| | DECIPHER−,• | https://decipher.sanger.ac.uk | Online repository of genetic variation with associated phenotypes that facilitate the identification and interpretation of pathogenic genetic variation in patients with rare disorders | ✓ | | | | [49] | 952 |
| | PharmVar−,∗ | https://www.pharmvar.org/ | PharmVar is a central repository for pharmacogene (PGx) variation that focuses on haplotype structure and allelic variation | ✓ | | | | [50] | 59 |
| Proteomics | UniProt−,• | https://www.uniprot.org/ | Contains comprehensive data on proteins, gene ontology, pathways, gene taxonomy and associated diseases | ✓ | 904K | | ✓ | [51] | 34 803 |
| | Swiss-Prot− | https://www.uniprot.org/statistics/Swiss-Prot | Swiss-Prot is the manually annotated and reviewed section of the UniProt | ✓ | 9473 | | | [52] | 2668 |
| | GENCODE− | https://www.gencodegenes.org | Classifies all gene features in human and mouse genomes | ✓ | 2 | ✓ | ✓ | [53] | 2645 |
| | Encyclopedia of DNA elements (ENCODE)− | https://www.genome.gov/10005107/ | Two levels of annotations: (i) integrative-level annotations, including a registry of candidate cis-regulatory elements and (ii) ground-level annotations derived directly from experimental data | ✓ | 2 | ✓ | ✓ | [54] | 8906 |
| | Consensus CDS (CCDS)− | https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi | Identifies a core set of human and mouse protein coding regions that are consistently annotated and of high quality | | 2 | | | [55] | 633 |
| | GeneCards− | https://www.genecards.org/ | Automatically integrates gene-centric data from ~150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information | ✓ | | | | [56] | 616 |

Continued

**Table 2.** Continues

| Category | Database name | Link | Key features | V | S<br>≥ | NC | API | Ref | Cit<br>≥ |
|---|---|---|---|---|---|---|---|---|---|
| | GenomeRNAi– | http://www.genomernai.org/Index | Contains phenotypes from RNA interference (RNAi) screens in Drosophila and *Homo sapiens* | | 2 | | | [57] | 130 |
| | HUGO Gene Nomenclature Committee (HGNC)– | https://www.genenames.org/ | Approves unique symbols and names for human loci, including protein-coding, ncRNA and pseudogenes, to allow unambiguous scientific communication | | | ✓ | ✓ | [58] | 17 828 |
| | OrthoDB– | https://www.orthodb.org/ | The hierarchical catalogue of orthologs mapping genomics to functional data (37M genes) | | 1367 | | ✓ | [59] | 133 |
| | UniGene– | https://www.ncbi.nlm.nih.gov/unigene/ | UniGene is a database that provides automatically generated nonredundant sets (clusters) of transcript sequences, each cluster representing a distinct transcription locus | | 140 | | ✓ | [60] | 61 |
| | UniRef– | https://www.uniprot.org/uniref/ | UniRef provides clustered sets of sequences from UniProt, hides redundant sequences and obtains complete coverage of the sequence space at three resolutions | ✓ | 904K | | | [61] | 1747 |
| | neXtProt– | https://www.nextprot.org/about/nextprot | A comprehensive human-centric discovery platform, offering its users a seamless integration of and navigation through protein-related data | ✓ | | | ✓ | [62] | 412 |
| | RefSeq– | https://www.ncbi.nlm.nih.gov/refseq/ | RefSeq provides a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies and comparative analyses | ✓ | 55K | ✓ | | [63] | 12 096 |
| | Expression Atlas – | https://www.ebi.ac.uk/gxa/home | Provides gene expression results on >3000 experiments from 40 different organisms, including metazoans and plants | ✓ | 60 | | ✓ | [64] | 779 |

*Continued*

**Table 2.** Continues

| Category | Database name | Link | Key features | V | S | NC | API | Ref | Cit |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | ≥ | | | | ≥ |
| Protein–protein interactions | Human Protein Reference Database (HPRD)–,● | http://www.hprd.org/ | Centralized platform to visualize and integrate information on post-translational modifications, interaction networks and disease association for each protein in the human proteome | | 1 | | | [65] | 3749 |
| | Biological General Repository for Interaction Datasets (BioGRID)–,* | http://thebiogrid.org/ | An interaction repository with data compiled through comprehensive curation efforts. Publications: 69 031; proteins and genetic interactions: 1 676 780; chemical associations: 28 093 and 726 378 post translational modifications | | 71 | | ✓ | [66] | 8254 |
| | Molecular INTeraction database (MINT)– | https://mint.bio.uniroma2.it/ | Molecular interaction (MINT) database focuses on experimentally verified protein-protein interactions (130 733) mined from the scientific literature by expert curators | | 646 | | | [67] | 665 |
| | GPS-Prot– | http://www.gpsprot.org | Integration of different HIV interaction data as well as interactions between human proteins derived from public databases, including MINT, BioGRID and HPRD | | | | | [68] | 25 |
| | Wiki-Pi– | http://severus.dbmi.pitt.edu/wiki-pi/ | A wiki resource on human protein-protein interactions. Unique interactions: 48 419; proteins: 10 492 | | 1 | | | [69] | 30 |
| | Protein Interaction Network Analysis (PINA)– | http://omics.bjcancer.org/pina/ | An integrated platform for protein interaction network construction, filtration, analysis, visualization and management | | 6 | | | [70] | 500 |
| | MPIDB– | http://www.jcvi.org/mpidb/ | Manually curated microbial interactions from literature and databases (IntAct, DIP, BIND, MINT). Experimental interactions: 22 530, bacterial species/strains: 191 | | 191 | | | [71] | 124 |
| | Search tool for the retrieval of interacting genes (STRING)– | http://string-db.org/ | Aims to collect score and integrate all publicly available sources of protein-protein interaction information. Organisms: 5090; proteins: 24.6M, interactions: >2000M | | 5090 | | ✓ | [72] | 16 117 |

*Continued*

**Table 2.** Continues

| Category | Database name | Link | Key features | V | S | NC | API | Ref | Cit |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | ≥ | | | | ≥ |
| | Mammalian protein-protein interaction (MIPS)+,– | http://mips.helmholtz-muenchen.de/proj/ppi/ | Manually curated high-quality data on mammalian protein-protein interactions by expert curators (proteins >900) | | 10 | | | [73] | 516 |
| | IntAct– | http://www.ebi.ac.uk/intact/ | Open source database system and analysis tools for molecular interaction data. Total interactions: 582 671 | ✓ | | | | [74] | 3953 |
| | Database of Interacting Proteins (DIP)– | http://dip.doe-mbi.ucla.edu/dip/Main.cgi | Contains experimentally determined interactions between proteins. Proteins: 28 850; and interactions: 81 923 | | 834 | | | [75] | 3949 |
| Molecular omics | Cancer Cell Line Encyclopedia (CCLE)+,– | https://portals.broadinstitute.org/ccle | Large cancer cell line collections broadly capture the genomic diversity of human cancers and provide valuable insight into anti-cancer drug response. Cell lines: 1457; genes: 84 434; mutations: 1 159 663; distribution scores: 118 661 636; methylation scores: 411 948 577 | ✓ | | | | [76] | 3455 |
| | gCSI– | | RNA sequencing and single-nucleotide polymorphism (SNP) array analysis of 675 human cancer cell lines. Comprehensive analyses of transcriptome features including gene expression, mutations, 2200 gene fusions and expression of nonhuman sequences | | | | | [77] | 267 |
| | Dependency Map (DepMap)– | https://depmap.org/portal/ | Systematically identify biomarkers of genetic vulnerabilities and drug sensitivities in hundreds of cancer models and tumors, to accelerate the development of precision treatments (4686 compounds screened against 578 cell lines) | | | | ✓ | [26] | 235 |

*Continued*

**Table 2.** Continues

| Category | Database name | Link | Key features | V | S<br>≥ | NC | API | Ref | Cit<br>≥ |
|---|---|---|---|---|---|---|---|---|---|
| | ICGC– | https://dcc.icgc.org/ | The ICGC Data Portal provides many tools for visualizing, querying, and downloading cancer data, which is released on a quarterly schedule | | | | ✓ | [78] | 308 |
| | cBioPortal– | https://www.cbioportal.org/ | A web resource for exploring, visualizing, and analyzing multidimensional cancer genomics data. The portal reduces molecular profiling data from cancer tissues and cell lines into readily understandable genetic, epigenetic, gene expression, and proteomic events | | | | ✓ | [27] | 4463 |
| | NCBI-GEO– | http://www.ncbi.nlm.nih.gov/geo/ | A genomics data repository supporting array and sequence-based data. Users may query and download experiments and curated gene expression profiles | | 1600 | | | [79] | 14 532 |
| | ArrayExpress– | http://www.ebi.ac.uk/arrayexpress/ | Public database of microarray experiments and gene expression profiles. It contains data from >7000 public sequencing and 42 000 array-based studies (>1.5 million assays) | | 200 | | | [80] | 2365 |
| | Princeton University MicroArray database (PUMAdb)– | http://puma.princeton.edu | Stores raw and normalized data from microarray experiments and provides interfaces for data retrieval, analysis and visualization | | 50 | | | [81] | 1163 |
| | CellMiner– | http://discover.nci.nih.gov/cellminer/ | Allows rapid access to transcript expression levels of 22 217 genes, 360 microRNAs and 18 549 compounds | | | | | [82] | 465 |

Continued

**Table 2.** Continues

| Category | Database name | Link | Key features | V | S<br>≥ | NC | API | Ref | Cit<br>≥ |
|---|---|---|---|---|---|---|---|---|---|
| | CellMinerCDB+,− | https://discover.nci.nih.gov/cellminercdb/ | CellMinerCDB is an interactive web application that simplifies access and exploration of cancer cell line pharmacogenomic data across different sources. | | | | | [83] | 2 |
| Pathway information | PathwayCommon− | http://www.pathwaycommons.org/ | Pathways including biochemical reactions, complex assembly and physical interactions involving proteins, DNA, RNA, small molecules and complexes. Interactions: 442 182; pathways: 1668 | | 414 | | ✓ | [84] | 768 |
| | Kyoto Encyclopedia of Genes and Genomes (KEGG)−,*,• | http://www.genome.jp/kegg/ | A reference knowledge base that Integrates genomic, chemical and systemic functional information. Pathway maps: 627 677; functional hierarchies: 224 555; KEGG orthology: 22 710; genes: 29 041 034; drugs: 10 957; biochemical reactions: 11 152; disease-related network elements: 813; human gene variants: 255 | ✓ | 6221 | | ✓ | [85] | 34 643 |
| | Reactome− | http://www.reactome.org | Manually curated and peer-reviewed pathway database. Proteins: 92 000; complexes: 93 684; pathways: 20 760 | ✓ | 16 | | ✓ | [86] | 757 |
| | MetaCyc− | http://metacyc.org | Curated database of experimentally elucidated metabolic pathways from all domains of life. Pathways: 2698; enzymatic reactions: 15418; transport reactions: 637; protein complexes: 4117; transporters: 363; compounds: 15 263; GO\ignorespacesterms: 6686 | | 20 | | | [87] | 2852 |
| Protein 3D structures | Protein Data Bank (PDB)− | http://www.rcsb.org | Contains 3D structures of proteins, nucleic acids and complex assemblies from protein synthesis to health and disease. Structures: 151079, proteins: 140135 | ✓ | | ✓ | ✓ | [88] | 65 699 |

*Continued*

**Table 2.** Continues

| Category | Database name | Link | Key features | V | S<br>≥ | NC | API | Ref | Cit<br>≥ |
|---|---|---|---|---|---|---|---|---|---|
| | Orientations of Proteins in Membranes (OPM)– | http://opm.phar.umich.edu | Unique experimental structures of transmembrane proteins and some peripheral proteins and membrane-active peptides | | 350 | | | [89] | 1287 |
| | Proteopedia– | http://proteopedia.org | Helps to bridge the gap between 3D structure & function of biomacromolecules | | | | | [90] | 126 |
| | TOPSAN– | http://www.topsan.org | Web-based platform for exploring and annotating structures determined by structural genomics efforts. Structures >7250 | | | | | [91] | 71 |
| | GPCRdb–,* | https://gpcrdb.org/ | Contains data, diagrams and web tools for G protein-coupled receptors (GPCRs). Users can browse all GPCR structures and the largest collections of receptor mutants (proteins: 15 147; ligands: 144 889; ligand interactions: 15 720) | ✓ | 3547 | | ✓ | [92] | 2177 |
| Protein classification | Protein families (PFAM)– | https://pfam.xfam.org/ | The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs) | | | | | [93] | 26 311 |
| | Protein ANalysis THrough Evolutionary Relationships (PANTHER)– | http://pantherdb.org/ | PANTHER classification system was designed to classify proteins in order to facilitate high-throughput analysis. Proteins have been classified according to molecular function, biological process and pathways | | 132 | | | [94] | 5674 |

V = database containing variants data are tick-marked, S = number of species, NC = tick-marked if contains noncoding RNA data, Ref = reference article, Cit = total number of citations for each individual article for the database (computed on 25 May 2019, 18:00 CET).

**Table 3.** Drug target interaction databases; some databases may also fall in more than one (main) category (+chemical, −biomolecular, ∗drug target interactions, •disease)

| Category | Database name | Link | Key features | C | T | I | API | Ref | Cit |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ≥ | ≥ | ≥ | | | ≥ |
| Bioactivity data | DrugTargetCommons (DTC)−,∗,• | http://drugtargetcommons.fimm.fi/ | Bioactivity data with manually curated assays, protein classification into super families, clinical phase and adverse effects, disease indications, disease gene associations for ∼3000 proteins | 1.6M | 13K | 14.8M | ✓ | [30] | 16 |
| | ChEMBL+,−,∗,• | https://www.ebi.ac.uk/chembl/ | Contains bioactivity data, compound structures and properties, clinical study references and disease indications | 1.9M | 12K | 15.5M | ✓ | [95] | 2490 |
| | PubChem+,−,∗,• | https://pubchem.ncbi.nlm.nih.gov/ | Provides chemical structures and physical properties, biological activities, patents, health, safety, toxicity data and many others | 95M | 58K | 264.8M | ✓ | [96] | 1909 |
| | BindingDB∗ | https://www.bindingdb.org/bind/index.jsp | Database of measured binding affinities. | 0.7M | 7.2K | 1.2M | ✓ | [97] | 1510 |
| | DrugCentral∗,• | http://drugcentral.org/ | Provides information on active ingredient chemical entities, pharmaceutical products, drug mode of action, indications, pharmacologic action | 4.5K | 11K | 15K | ✓ | [98] | 61 |
| | Guide to PHARMACOLOGY (GtopDB)∗ | http://www.guidetopharmacology.org/ | Information on approved targets and experimental drugs | 9.7K | 2.9K | 31.2K | ✓ | [99] | 1184 |
| | PDSP Ki∗ | https://pdspdb.unc.edu/pdspWeb/ | Contains internally derived Ki, or affinity, values for a large number of drugs at an expanding number of G-protein coupled receptors, ion channels, transporters and enzymes | | | | | [100] | 257 |
| | Probes & Drugs portal+,−,∗ | https://www.probes-drugs.org/home/ | A public resource joining together focused libraries of bioactive compounds (probes, drugs, specific inhibitor sets, etc.) with commercially available screening libraries | 67K | 8.7K | 0.96M | | [28] | 13 |
| Drug-target interactions | DrugBank+,−,∗,• | https://www.drugbank.ca/ | Combines drug (i.e. chemical, pharmacological and pharmaceutical) information with drug target (i.e. sequence, structure and pathway) information | 12K | 5K | 18.9K | ✓ | [36] | 7126 |
| | Drug gene interaction database (DGIdb)∗ | http://www.dgidb.org/ | Searching, and filtering of information on drug-gene interactions and the druggable genome, mined from over thirty trusted sources | 9501 | 41K | 29K | ✓ | [101] | 210 |
| | PharmGKB+,• | https://www.pharmgkb.org/ | Contains comprehensive data on genetic variation on drug response for clinicians and researchers | 694 | 900 | | ✓ | [102] | 695 |

Continued

**Table 3.** Continues

| Category | Database name | Link | Key features | C | T | I | API | Ref | Cit |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ≥ | ≥ | ≥ | | | ≥ |
| | SuperTarget−,∗,• | http://insilico.charite.de/supertarget/ | A resource for analyzing drug-target interactions and drug side effects | 0.2M | 6.2K | 0.33M | | [103] | 503 |
| | GLIDA∗ | http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/ | A GPCR database focusing on GPCRs and their ligands | 24K | 3.7K | 39.1K | | [104] | 215 |
| | SwissTargetPrediction∗ | http://www.swisstargetprediction.ch/ | An on-line tool to predict targets for drugs using 2D and 3D structures | 0.38M | 3K | 0.58M | | [105] | 197 |
| | Cancer Genome Interpreter (CGI)∗ | https://www.cancergenomeinterpreter.org/ | Supports the identification of tumor alterations that drive the disease and detect those that may be therapeutically actionable | 310 | 837 | | | [106] | 82 |
| Drug target interactions visualization | DrugTargetProfiler (DTP)+,−,∗,• | http://drugtargetprofiler.fimm.fi/ | Users can search database as well as upload their own data. Interaction score is computed based on curated bioactivity data, protein family and assays. Figures as well as tabular data can be exported. Contains data on >200 mutant targets and also integrated gene expression and drug sensitivity data on various cancer cell lines | 0.9M | 6K | 4.4M | | [31] | 2 |
| | Search Tool for Interactions of Chemicals (STITCH)+,−,∗ | http://stitch.embl.de/ | Stores known and predicted interactions of chemicals and proteins. Also contains pathways and drug-drug interaction data | 0.43M | 9.6M | | ✓ | [107] | 1160 |

C = compounds, T = targets, I = interactions, Cit = total number of citations for each individual article on the database, Ref = reference article.

**Table 4.** Disease databases; some databases may contain heterogeneous datatypes and hence placed in more than one (main) category (+chemical, −biomolecular, ∗drug target interactions, •disease)

| Category | Database name | Link | Key features | C | P | D | A | API | Ref | Cit |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ≥ | ≥ | ≥ | ≥ | | | ≥ |
| Disease | OpenTargets−,• | https://www.opentargets.org | A comprehensive and robust data integration for access to and visualization of potential drug targets associated with disease | | 28K | 10K | 3M | ✓ | [108] | 106 |
| | Catalogue Of Somatic Mutations In Cancer (COSMIC)• | https://cancer.sanger.ac.uk/cosmic | A comprehensive source of information on manually curated somatic mutations and their frequencies in human cancers | | 709 | | | | [109] | 6294 |
| | Pharos−,∗,• | https://pharos.nih.gov/idg/index | Focusses on druggable genome (DG) to illuminate the uncharacterized and/or poorly annotated portion of the DG. Four drug-targeted protein families: G-protein-coupled receptors (GPCRs); nuclear receptors (NRs); ion channels (ICs); and kinases | 0.2M | 10K | 12K | 0.2M | ✓ | [110] | 118 |
| | DisGeNET• | http://www.disgenet.org/ | A discovery platform containing publicly available collections of genes and variants associated to human diseases (210 498 variant-disease associations) | | 17K | 24K | 0.6M | ✓ | [34] | 955 |
| | BioMuta• | https://hive.biochemistry.gwu.edu/biomuta | A single-nucleotide variation (3 732 175 SNVs) associated with 42 cancer types | | 19K | 42 | 3.7M | ✓ | [111] | 92 |
| | Therapeutic target database (TTD)−,• | http://bidd.nus.edu.sg/group/cjttd/ | Information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs directed at each of these targets | 34K | 3.1K | | | | [112] | 466 |
| | Online Mendelian Inheritance in Man (OMIM)• | http://omim.org | OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily | | 16K | 5K | | ✓ | [113] | 4729 |
| | Comparative Toxicogenomics Database (CTD)−,∗,• | http://ctdbase.org/ | It provides manually curated information about chemical–gene/protein interactions, chemical–disease and gene–disease relationships | 15.6K | 46.7K | 7K | 37.7K | | | 954 |
| Clinical information | ClinicalTrials• | https://clinicaltrials.gov/ | Contains clinical studies, adverse effects and disease indications. | | | | | | | |
| | Drugs@FDA• | https://www.accessdata.fda.gov/scripts/cder/daf/ | FDA approved drugs and their dosage information | | | | | | | |

*Continued*

**Table 4.** Continues

| Category | Database name | Link | Key features | C<br>≥ | P<br>≥ | D<br>≥ | A<br>≥ | API | Ref | Cit<br>≥ |
|---|---|---|---|---|---|---|---|---|---|---|
| | FDA Orange Book• | https://www.accessdata.fda.gov/scripts/cder/ob/ | Approved drugs with therapeutic equivalence evaluations | | | | | | | |
| | EU Clinical Trials register• | https://www.clinicaltrialsregister.eu/ctr-search/ | Interventional clinical trials that are conducted in the EU | | | | | | | |
| | Japan PMDA• | https://www.pmda.go.jp/english/review-services/reviews/approved-information/drugs/0002.html | Scientific reviews of marketing authorization application of pharmaceuticals and medical devices, monitoring of their post-marketing safety | | | | | | | |
| | NIH DailyMed• | https://dailymed.nlm.nih.gov/dailymed/index.cfm | Trustworthy information about marketed drugs in US | | | | | ✓ | | |
| | REPURPOSED DRUG DATABASE• | http://drugrepositioningportal.com/repurposed-drug-database.php | Contains original and new indications for 240 repurposed drugs | | | | | | | |
| | repoDB• | http://apps.chiragjpgroup.org/repoDB/ | Contains a standard set of drug repositioning successes and failures that can be used to fairly and reproducibly benchmark computational repositioning methods. repoDB data were extracted from DrugCentral and ClinicalTrials.gov | | | | | | [114] | 59 |
| | RepurposeDB• | http://repurposedb.dudleylab.org/ | RepurposeDB help to develop predictive models of drug repositioning and can have major influence in personalized drug repositioning | | | | | | [115] | 84 |
| Drug side effects | Side Effect Resource (Sider)• | http://sideeffects.embl.de/ | Contains side effects and indications (5868 side effects for 1430 drugs or therapies) | 1.4K | | | | | [116] | 970 |
| | IDAAPM*,• | http://idaapm.helsinki.fi | Contains ADMET, adverse effects, molecular descriptors and bioactivity data. There are 2.5 million drug-adverse effect pairs, 3382 targets and 36 963 binding affinity data | | 3.3K | | | | [32] | 10 |
| | Everyday health• | https://www.everydayhealth.com/drugs/ | Contains side effects and dosage information | | | | | | | |
| | VigiAccess• | http://www.vigiaccess.org/ | Contains adverse reactions and ADR reports | | | | | | [117] | 2 |

C = compounds, P = proteins, A = protein-disease associations, Cit = total number of citations for each individual article on the database, Ref = reference article.

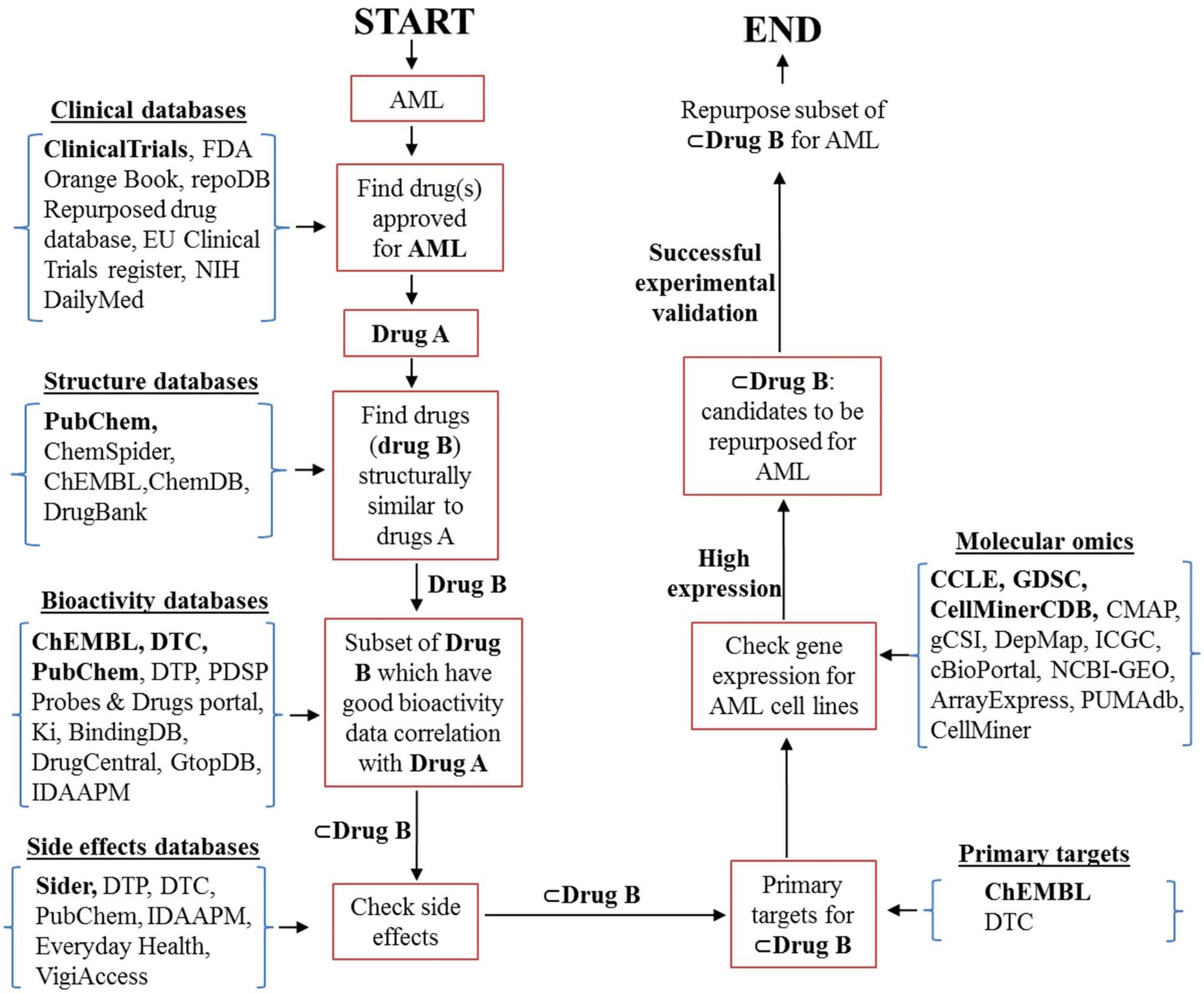
START
AML
Find drug(s) approved for AML
Drug A
Find drugs (drug B) structurally similar to drugs A
Drug B
Subset of Drug B which have good bioactivity data correlation with Drug A
⊂Drug B
Check side effects
⊂Drug B
Primary targets for ⊂Drug B
Check gene expression for AML cell lines
High expression
⊂Drug B: candidates to be repurposed for AML
Successful experimental validation
Repurpose subset of ⊂Drug B for AML
END
Clinical databases
ClinicalTrials, FDA Orange Book, repoDB Repurposed drug database, EU Clinical Trials register, NIH DailyMed
Structure databases
PubChem, ChemSpider, ChEMBL,ChemDB, DrugBank
Bioactivity databases
ChEMBL, DTC, PubChem, DTP, PDSP Probes & Drugs portal, Ki, BindingDB, DrugCentral, GtopDB, IDAAPM
Side effects databases
Sider, DTP, DTC, PubChem, IDAAPM, Everyday Health, VigiAccess
Molecular omics
CCLE, GDSC, CellMinerCDB, CMAP, gCSI, DepMap, ICGC, cBioPortal, NCBI-GEO, ArrayExpress, PUMAdb, CellMiner
Primary targets
ChEMBL
DTC

**Figure 2**. Disease-based drug repurposing workflow using databases listed in Tables 1–4. Databases are mentioned inside brackets. Dark font shows recommended databases and normal font shows alternative databases. The workflow describes steps for disease-based computational drug repurposing for AML; however, the same sequence of steps and the listed databases can be used for other diseases.

measurements (e.g. IC50, Kd, Ki) whereas drug–target interaction databases report quantitative data (binary interactions). The drug-target interaction visualization databases in addition to providing information on bioactivity or drug-target binary interactions also provide visualizations for interaction networks, which can be exported as figures. Databases are assessed using parameters, such as the number of compounds (C), number of targets (T), number of compound-target interactions (I), number of citations (Cit) and API, as shown in Table 3. Citation information was computed on 25 May 2019, 18:00 CET. Table 3 contains further details and statistics on individual databases.

### Disease databases

There are 19 databases in this category as shown in Table 4. Databases under disease subcategory mostly contain disease-gene associations. The subcategory 'clinician information' contains databases that report clinical studies and disease indications while the subcategory 'side effects' represents those databases that show possible side effects of the drugs in diseased patients. The databases in disease category are represented using parameters, such as the number of compounds, proteins, diseases, protein-disease associations, API (tick-marked if data access is provided by API) and number of citations (citation information was computed on 25 May 2019, 18:00 CET).

## Recommended databases in each subcategory

Tables 1–4 provide short descriptions, statistics, datatypes for subcategories and other characteristics for individual databases, but it is still difficult to select best databases from a comprehensive list of 102 databases using 4 main categories and 17 subcategories. Hence, in this section, we have tried to suggest the best database(s) for each subcategory based on statistics, quality, availability, data redundancy, features, variety of data types and database usage (based on citations).

### Chemical databases

Among drug combination databases, DrugComb [29] and DrugCombDB [37] provide qualitative drug combination responses in terms of sensitivity and synergy measures. Additionally, DrugComb also provides visualization of drug combination data for >4500 drugs; hence, we recommend DrugComb for drug combination data mining. For Drug omics, we propose GDSC [42] as it is one of the most comprehensive drug omics database with >1000 cancer cell lines and with data from >75 000 experiments. GDSC also provides API to programmatically access data and linked drug omics with molecular omics, as shown in Table 1.

### Biomolecular databases

Among genomics databases, we suggest Ensembl [47] as it is very comprehensive, focusing on >5000 species with deep curation and API access. Furthermore, it contains data for protein variants, genomic visualizations, linked pathways and molecular omics information. Among the proteomics databases, UniProt [118] could be the one choice as it integrates a wide variety of data types, such as pathways, protein-protein interactions (PPI), molecular omics, protein sequences, structures, classification and protein disease associations, as shown in Figure 1. UniProt is quite comprehensive, focusing on ∼1 million species and also contains data on protein variants, which can be accessible using APIs (Table 2). Among PPI databases, BioGRID [66] could be used as it contains 1.3 million nonredundant PPIs. Additionally, BioGRID has associated PPI data with drug target interactions with few minimized steps for drug repurposing applications. Among molecular omics databases, we suggest CCLE [76] as it links drug omics with molecular omics and contain comprehensive data on 1457 cell lines, 84 434 genes, ∼1 million mutations and 400 million methylation scores. If someone is interested in interactive visualization of molecular and drug responses across cancer cell lines, then CellMiner CDB [83] could be a good choice. Among pathway databases, KEGG [85] is one of the most comprehensive and nonredundant database (0.6 million pathways). It contains pathways for >6000 species with this information linked to diseases and drug target interaction information, as shown in Table 2. It is also the most cited among the pathway databases and contains some variant data. PDB [119] is one of the most comprehensive databases when it comes to protein 3D structures, with >140 K protein structures accessible via API.

### Drug target interaction databases

Among bioactivity or drug target interactions databases, PubChem [120] and ChEMBL [95] could be the databases of choice as these contain enormous nonredundant bioactivity datasets with more than 15 million curated bioactivities. Additionally, both have linked bioactivity data with several other datatypes, such as chemical structures and physiochemical properties, clinical data and disease indications. STITCH [107] and DTP [31] are good if someone is interested in drug target interaction visualization, as can be seen in Table 3. DTP is based on curated bioactivates whereas STITCH combines bioactivities with predicted compound target interactions using a weighted mechanism.

### Disease databases

OpenTargets [108] and DisGeNET [34] are the most comprehensive databases when it comes to protein and disease associations, both having information on >17K proteins and >10K diseases. DisGeNET, additionally, has associated protein variants with human diseases. If someone is specifically interested in cancer biomarkers, COSMIC [109] could be a good choice as it contains comprehensive and curated list of somatic mutations associated with different cancers, as shown in Table 4. Clinical trials (https://clinicaltrials.gov/) is one of the most comprehensive clinical DBs, containing information on both failed and successful trials, as well as drug combinatorial studies for 0.3 million clinical studies. The problem with most of the clinical databases is that the data are not in structured format and are not downloadable in case of some databases (require manual copy/paste or web scrapping). Among the side effects databases, we suggest Sider [116] as it has a structured format that is easily downloadable and quite comprehensive (5868 side effects for 1430 drugs).

## Examples of drug repositioning, database use, methods and limitations

In this section, to illustrate how databases can be used for drug repositioning, we provide examples of the prerequisites and methods of drug activity prediction and the databases that have been used. We also illustrate a few important issues that affect several types of data and influence both the primary drug development and repositioning efforts. For specific examples of drug repositioning other than the ones provided below, the reader is referred to the Drug Repurposing Portal (http://www.drugrepurposingportal.com/).

### Drug repositioning is facilitated by molecular target annotation

Approval of small molecule compounds for new indications requires a sufficient degree of mechanistic insight, which in turn is dependent on identifying the drug binding partners at molecular level. Several endeavors are underway to catalogue comprehensive binding profiles of existing small molecules and cellular macromolecules [121–124]. To facilitate drug-target annotation efforts, an open annotation and query platform called Drug Target Commons (DTC) [30], which provides a summary of currently known molecular targets for thousands of drugs, has recently become available. Of all the molecular targets, kinase inhibitors are among the most studied as they function both as therapeutics and as molecular tools to examine cellular processes. Because of the conserved nature of the ATP binding pocket of the kinases, kinase inhibitors are more than likely to bind and inhibit several targets. This implies that repurposing of such inhibitors is a task requiring cataloguing affinities and assigning functional ranking to the targets. For example, Abl kinase-inhibitor imatinib has found use outside chronic myeloid leukaemia, as it also targets at least wild type c-Kit and PDGFR receptor tyrosine kinases within a well-tolerated clinical concentration range [125]. Another example of repurposing effort that relied on drug annotation information is the repurposing of SRC kinase inhibitor saracatinib. It was originally developed for several types of cancers but was abandoned due to poor efficacy. Later on, it was known to possess affinity toward FYN kinase and was proceeded to Phase II testing for Alzheimer's disease, in which FYN plays an important pathogenic role [126, 127]. While that 159 Alzheimer patient trial unfortunately did not show clinical benefits for saracatinib, brain imaging indicated that the drug may have influenced some pathological processes [126], which possibly can support generation of new combination regimens for this devastating disease in the future. Other examples include anti-inflammatory asthma medication montelukast, which was found to bind dipeptidyl peptidase IV [128], a promising drug target in type II diabetes. Diclofenac, simvastatin, ketoconazole and itraconazole, on the other hand, were found to bind to estrogen receptors, whereby they potentially could find uses in targeting estrogen-dependent cancers [128].

In addition to knowing which targets small molecules can bind, it is important to know the mode of action (inhibitory/activating). For example, anti-helminthic agent pyrvinium binds to and activates CK1$\alpha$ kinase, which has a net inhibitory effect on the Sonic Hedgehog pathway by virtue of CK1$\alpha$ acting as an endogenous antagonist for this pathway [129]. However, even drugs with broad or only partially defined molecular target ranges can be repurposed. One such example is the famous drug disulfiram, also known as Antabuse, that has been used to treat alcohol abuse since 1948 because one of its metabolites inhibits aldehyde dehydrogenase (ALDH), generating discomfort upon alcohol consumption. However, it displays anti-cancer effects which are still incompletely understood and are at least independent of the ALDH-inhibitory function and associated with protein accumulation of nuclear protein NPL4 in a copper-dependent fashion [130].

## Drug-induced cellular responses and conditional cell states can be matched to facilitate drug repositioning

Drug repositioning often begins from observations in large phenotypic screens, which can then be matched to underlying mutations and gene expression patterns. The National Cancer Institute cell line panel (NCI-60), the Cancer Cell Line Encyclopedia (CCLE) and the collaborative Genomics of Drug Sensitivity in Cancer (GDSC) databases constitute the largest collections of genomic and drug response data from different cancer types to date. The data in these databases can be accessed, explored and visualized using the highly useful CellminerCDB portal [83]. The databases offer drug response measurements, transcription levels and mutation data, where CCLE features a greater number of mutations in at least KRAS, PTEN, BRAF, NRAS and MSH6 than in the GDSC or NCI-60 databases due to greater sequencing depth. GDA [131] is similar to CellminerCDB in many aspects: to simplify the user experience, expression data from both NCI-60 and CCLE have been combined into a single data point for each gene as the data between these databases are highly concordant. Transcriptomic and drug response data in databases, such as GDA, CCLE and GDSC, can be correlated to reveal drug mechanisms. In one study, high gene expression levels of fatty acid desaturase 2 (FADS2) were identified as a common denominator among ~19 000 basal transcript levels across 823 different human cancer cell lines responding to the poorly characterized anti-cancer agent ML239. Subsequently, the anti-cancer potency of ML239 was demonstrated to be reduced in the cells knocked out for FADS2 or simultaneously treated with FADS2 inhibitor [132, 133].

Cell-intrinsic changes, including epigenetic modulation and mutations, and external factors, such as cell-cell contacts; stroma and matrix; soluble signalling mediators; pressure and oxygen tension; may influence cell, tissue and body-wide transcriptional landscapes. Some of the molecular alterations are conserved across patients, mutations, and specific conditions, and can be categorized as disease- or state-associated/enriched (Table 2) [134, 135], which is the basis for drug connectivity mapping. As an example, the gene expression alterations triggered by three candidate compounds, disulfiram (Antabuse), metropolol, a beta1-receptor blocker used to treat high blood pressure, and nonsteroidal anti-inflammatory agent sulindac, were observed to match the gene expression patterns enriched in the neurocognitive disorder Fragile X, as reported in NCBI GEO and in unpublished sources [136]. These agents improved cognitive performance as measured in Fmr1 KO2 mice, which serve as an animal model of Fragile X. However, in most cases, rapidly fluctuating and heterogeneous transcriptomes necessitate feature selection. This means filtering of only the most relevant genes for predictive modelling, or, conversely, eliminating irrelevant genes that might confound correlative drug response analyses [137]. A form of feature selection useful for bulk transcriptomic data is deconvolution, whereby diverse cellular background signals can be separated from the gene expression changes associated with and predicting drug responses. As an example, the gene expression signatures of coding mutations across a large number of different cell types, called core transcriptomes, were filtered out from the gene expression changes triggered by the drugs to reveal the drug-specific, tissue-independent effects [137]. In this study, after core transcriptome filtering, several novel AKT inhibitors or FOXO and AMPK activators were found, which were able to extend *C. elegans* nematode lifespan, similar to the known agents affecting these pathways. The approach offered a promising strategy to complement other deconvolution tools, such as CIBERSORT, ImmuCC and DeMixT, which will likely prove useful in repurposing drugs based on mixed transcriptomes [138–140]. Tissue-specific transcriptomes can be accessed via Genotype-Tissue Project (GTEx) [141].

## Computational tools and machine learning facilitate drug structure-guided target prediction and repositioning

Concomitant with experimental drug-target discovery efforts that provide quantitative molecular affinity data, computational tools are being built to predict/infer new molecular targets using orthogonal drug-target space deconvolution, where the molecular structures of both the drugs and targets help guide predictions [105, 142]. As an example of drug-target prediction based on molecular docking, antipsychotic agent thioridazine was found among 1500 FDA-approved compounds to possess anti-inflammatory activity by binding and inhibiting IKK, critical for the NF-$\kappa$B pathway, which was also validated in experimental assays [143]. Similarly, virtual docking predicted inhibitory activity for five compounds from a collection of more than 1400 FDA-approved drugs against *Pseudomonas aeruginosa* quorum-sensing (population-wide virulence) mechanisms, with antipsychotic agent pimozide displaying *in vitro* activity in inhibiting bacterial virulence gene expression [144]. A recent study introduced CDRscan, an algorithm that incorporates drug response assay data from GDSC, genomic data from CCLE and virtual docking based on structural fingerprints, including quantitative structure activity relationships (QSAR) information from DrugBank [145]. CDRscan predicted anti-cancer activity for 176 of 1385 approved nononcology drugs, with 27 compounds showing strong predicted efficacy for at least one of the 25 cancer types in at least 10% of the cell lines evaluated, and four agents being assigned predicted anticancer activity against more than 90% of all cancer types.

## Incorporating more omics variables creates new *in silico* drug repositioning opportunities awaiting clinical translation

New algorithms display ever increasing accuracy to match observed drug response patterns when additional variables are incorporated, such as drug-protein-interactions and protein-protein interactions (PPI). Addition of PPI data from STRING [72] generated a new method, termed HNMDRP, that outperformed several state-of-the-art machine learning-based methods in correctly correlating responses of the drugs with the underlying combined weighted scores, calculated using drug structural, target cell mutational, drug-protein and PPI data [147]. Cross-database deep learning models have been used to predict drug responses and drug synergies based on drug-induced transcriptional changes [148, 149] and KEGG-defined signaling pathways [150]. Yet another promising computational repurposing tool is Dr Insight, which uses correlations between various 'omics' and clinical data and drug responses to enrich for drug-disease and drug-target connections without the need to extract or limit feature sets, thus, outperforming other well established drug repurposing methods, such as CMap [41], sscMap [151] and NFFinder [152, 153]. No experimental validations were conducted in these studies, which limits the possibilities to evaluate their translational potential, and while response predictions for several compounds outside the training sets matched published data, it is difficult to know how these new algorithms would perform on previously unpublished, poorly annotated molecules.

Machine learning can even be used for drug repositioning by identifying the drug-disease links from word combinations in published texts. In a recent example, drug candidates for psoriasis and other inflammatory conditions were extracted from Pubmed abstracts using several different machine learning algorithms, of which partial least squares discriminant analysis outperformed other well-known approaches, such as random forest and LASSO regression [154]. The gene expression signatures associated with cellular responses to the drug candidates matched transcriptomic signatures associated with psoriasis, and at least five of the 20 top candidate drugs proposed by the machine learning approach already had been observed to have ameliorating activity in psoriasis, as listed in CTD [155], DrugBank [36] and Pharmacogenomics Knowledgebase. However, all hits were relatively promiscuous immunomodulators and could be expected to have at least some activity in psoriasis and in other immune-mediated diseases. This implies that text mining might work best to extend drug utility rather than offer entirely new or unexpected uses for existing compounds.

### Incorporating clinical databases for drug repurposing

An important consideration for any new repositioning study is whether the repositioned drug(s) would display the expected phenotypic effects both in preclinical models and in real patients/subjects. In this regard, databases in which clinical drug responses and side effects have been collected (Table 4) may be particularly useful. Cheng *et al.* [156] provided an example of pharmacoepidemiologic multi-database drug repurposing by inferring to potential cardiovascular side effects for a set of compounds based on the cardiovascular side effects of other drugs annotated in the patient databases IBM Watson/Truven Health Analytics MarketScan and Optum Clinformatics, and sharing close protein-protein interaction networks with the drugs of interest. By this approach, hydroxychloroquine was observed to be associated with significantly reduced risk for coronary arterial disease, and experimental testing corroborated previously published data that the drug inhibited several risk-associated cellular processes, notably TLR7/9- and TNF-$\alpha$ signaling. As another example, terazosin, a drug approved for the treatment of benign prostatic hyperplasia, and in rare cases hypertension, was found to enhance the activity of phosphoglycerate kinase 1 (PGK1), which promotes glycolysis and cellular ATP production, which in turn was shown to increase neuronal survival in animal models of Parkinson's disease [157]. This novel hypothetical use was validated by association in two clinical databases, Parkinson's Progression Markers Initiative database and Truven Health Analytics MarketScan, where terazosin was associated with reduced risk to develop Parkinson's disease, reduced disease progression and amelioration of select clinical symptoms in the diagnosed patients [157].

### Drug repurposing workflow

The previous examples showcase how multidimensional data stored in the databases can be mined for drug repositioning. In this final section, in order to help readers further understand how various databases could be combined to generate drug repurposing hypotheses, we outline one possible integrative workflow for drug repurposing. As an example, we have used Acute Myeloid Leukaemia (AML) as the disease of interest, but this can be applied to any disease.

We started by searching clinical databases for drugs approved for AML. A list of appropriate databases is given in Figure 2; however, we recommend ClinicialTrials.gov (shown in bold-face) for this purpose as it is the most comprehensive and accessible database for 0.3 million clinical studies. We name this list of approved drugs for AML as Drug A. The next step could be to find drugs that are significantly similar (e.g. Tanimoto similarity >0.3) to Drug A in terms of 2D chemical structure. We recommend PubChem [96] here because it not only contains ∼9 million structures, but also provides computed fingerprints through API. We name this list Drug B. The next step is to correlate bioactivity data for Drug A with Drug B. A subset of Drug B that has significant correlation with Drug A is proceeded to the next step where we need to check side effects for ⊂Drug B using Sider database [116]. The next step is to search for primary targets for ⊂Drug B, which can be obtained from ChEMBL [95] or DTC [30]. After that, we need to see whether these protein targets are highly expressed in AML cell lines. This can be done by processing data from CCLE [76], GDSC [44] or CellMinerCDB [83]. The subset of Drug B, whose primary targets were highly expressed in AML cell lines, can be the possible repurposing candidate for AML. The subset of Drug B can be further validated by experimentation.

## Conclusions

In this review, we analysed 102 databases, which can directly or indirectly affect the designing of the computational pipelines for drug repositioning. We have provided some characteristics and statistics for the databases and assigned those into primarily 4 main and 17 subcategories based on the main strengths of the databases. We have also provided information on how some databases have integrated multiple datasets and suggested preferable databases for each subcategory (or data type). We have also highlighted some of the machine learning methods, which are using databases to train models in order to find new indications for approved drugs, and provided a computational workflow using listed databases, which can help researchers to setup novel drug repurposing pipelines.

In general, the main limitations of the machine learning-based drug repositioning efforts stem from data heterogeneity and overabundance of variables as compared to the number of samples. Fortunately, both database hosts and the research community as a whole are aware of the main hurdles, and several solutions are being developed to address these concerns [158].

Taken together, the early examples of drug repositioning using database-driven *in silico* exploration herald an expanding wave of discoveries of increasing accuracy that can be matched by experimental evidence. Combining several technical and model-specific improvement scans facilitates the identification of new uses for existing drugs, thus accelerating MoA establishment for new chemicals.

### Key Points

- This manuscript provides a survey on available databases, which could directly or indirectly support drug repurposing. For the ease of understanding, the databases are divided into 4 main and 17 subcategories.
- In total, 102 databases are summarized, and the best databases are recommended for each subcategory based on data quality, comprehensiveness, data types and usage.
- The manuscript also sheds light on how databases can be used for drug repositioning and provides examples of the methods for drug activity prediction by using data from public databases.
- Finally, a systematic workflow for disease-based drug repurposing is provided showing the usage of given databases that can lead to assigning new applications for drug repurposing.

## Funding

The work was financially supported by European Research Council (ERC) starting grant agreement DrugComb (No. 716063), Academy of Finland grant (No. 317680), European Commission H2020 EOSC-life (No. 824087) and Medicinska Understödsföreningen Liv och Hälsa.

## References

1. Shah NP, Tran C, Lee FY, *et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* 2004;**305**:399–401.
2. Paul SM, Mytelka DS, Dunwiddie CT, *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**:203–14.
3. Dickson M, Gagnon JP. The cost of new drug discovery and development. *Discov Med* 2009;**4**:172–9.
4. Landgren O, Iskander K. Modern multiple myeloma therapy: deep, sustained treatment response and good clinical outcomes. *J Intern Med* 2017;**281**:365–82.
5. Zappacosta AR. Reversal of baldness in patient receiving minoxidil for hypertension. *N Engl J Med* 1980;**303**:1480–1.
6. Joensuu H. Treatment of inoperable gastrointestinal stromal tumor (GIST) with Imatinib (Glivec, Gleevec). *Med Klin (Munich)* 2002;**97**(Suppl 1):28–30.
7. Langedijk J, Mantel-Teeuwisse AK, Slijkerman DS, *et al.* Drug repositioning and repurposing: terminology and definitions in literature. *Drug Discov Today* 2015;**20**:1027–34.
8. Boolell M, Gepi-Attee S, Gingell JC, *et al.* Sildenafil, a novel effective oral therapy for male erectile dysfunction. *Br J Urol* 1996;**78**:257–61.
9. Gupta SC, Sung B, Prasad S, *et al.* Cancer drug discovery by repurposing: teaching new tricks to old dogs. *Trends Pharmacol Sci* 2013;**34**:508–17.
10. GNS HS, GR S, Murahari M, *et al.* An update on drug repurposing: re-written saga of the drug's fate. Biomed. *Pharmacotherapy* 2019;**110**:700–16.
11. Verbaanderd C, Meheus L, Huys I, *et al.* Repurposing drugs in oncology: next steps. *Trends Cancer* 2017;**3**:543–6.
12. Lotfi Shahreza M, Ghadiri N, Mousavi SR, *et al.* A review of network-based approaches to drug repositioning. *Brief Bioinform* 2017;**19**:878–92.
13. Ji X, Freudenberg JM, Agarwal P. Integrating biological networks for drug target prediction and prioritization. *Methods Mol Biol* 2019;**1903**:203–18.
14. Mullard A. 2018 FDA drug approvals. *Nat Rev Drug Discov* 2019;**18**:85–9.
15. Lin A, Giuliano CJ, Palladino A, *et al.* Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Sci Transl Med* 2019;**11**:eaaw8412.
16. Shaughnessy AF. Old drugs, new tricks. *BMJ* 2011;**342**:d741.
17. Pushpakom S, Iorio F, Eyers PA, *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**:41–58.
18. Lotfi Shahreza M, Ghadiri N, Mousavi SR, *et al.* A review of network-based approaches to drug repositioning. *Brief Bioinform* 2018;**19**:878–92.
19. Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. *Brief Bioinform* 2011;**12**:303–11.
20. Jin G, Wong STC. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today* 2014;**19**:637–44.
21. Sam E, Athri P. Web-based drug repurposing tools: a survey. *Brief Bioinform* 2017;**20**(1):299–316.
22. Li J, Zheng S, Chen B, *et al.* A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;**17**: 2–12.
23. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. *Brief Bioinform* 2009;**10**:579–91.
24. Yang X, Wang Y, Byrne R, *et al.* Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev* 2019;**119**:10520–94.
25. Rigden DJ, Fernández XM. The 26th annual nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Res* 2019;**47**:D1–7.
26. Tsherniak A, Vazquez F, Montgomery PG, *et al.* Defining a cancer dependency map. *Cell* 2017;**170**:564–576.e16.
27. Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1.
28. Skuta C, Popr M, Muller T, *et al.* Probes & drugs portal: an interactive, open data resource for chemical biology. *Nat Methods* 2017;**14**:759–60.
29. Zagidullin B, Aldahdooh J, Zheng S, *et al.* DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res* 2019;**47**(W1):W43–51.
30. Tanoli Z, Alam Z, Vähä-Koskela M, *et al.* Drug target commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. *Database* 2018;**2018**:1–13.
31. Tanoli Z, Alam Z, Ianevski A, *et al.* Interactive visual analysis of drug–target interaction networks using drug target profiler, with applications to precision medicine and drug repurposing. *Brief Bioinform* 2018.
32. Legehar A, Xhaard H, Ghemtio L. IDAAPM: integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data. *J Chem* 2016; **8**:33.

33. Smirnov P, Kofia V, Maru A, *et al*. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res* 2018;**46**:D994–1002.
34. Piñero J, Bravo À, Queralt-Rosinach N, *et al*. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–9.
35. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
36. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
37. Liu H, Zhang W, Zou B, *et al*. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res* 2019;**48**:D871–D881.
38. Liu Y, Hu B, Fu C, *et al*. DCDB: drug combination database. *Bioinformatics* 2010;**26**:587–8.
39. Editorial: ChemSpider-a tool for Natural Products research. 2015
40. Chen JH, Linstead E, Swamidass SJ, *et al*. ChemDB update full-text search and virtual chemical space. *Bioinformatics* 2007;**23**:2348–51.
41. Subramanian A, Narayan R, Corsello SM, *et al*. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**:1437–1452.e17.
42. Iorio F, Knijnenburg TA, Vis DJ, *et al*. A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**:740–54.
43. Seashore-Ludlow B, Rees MG, Cheah JH, *et al*. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015;**5**:1210–23.
44. Yu C, Mannan AM, Yvone GM, *et al*. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol* 2016;**34**:419–23.
45. Ashurst JL, Chen C-K, Gilbert JGR, *et al*. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* 2005;**33**:D459–65.
46. Kent WJ, Sugnet CW, Furey TS, *et al*. The human genome browser at UCSC. *Genome Res* 2002;**12**:996–1006.
47. Flicek P, Amode MR, Barrell D, *et al*. Ensembl 2011. *Nucleic Acids Res* 2011;**39**:D800–6.
48. Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* 2018;**46**:D41.
49. Bragin E, Chatzimichali EA, Wright CF, *et al*. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res* 2014;**42**:D993–1000.
50. Gaedigk A, Ingelman-Sundberg M, Miller NA, *et al*. The pharmacogene variation (PharmVar) consortium: incorporation of the human cytochrome P450 (*CYP*) allele nomenclature database. *Clin Pharmacol Ther* 2018;**103**:399–401.
51. Consortium U. UniProt: a hub for protein information. *Nucleic Acids Res* 2014;**43**:D204–12.
52. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res* 1997;**25**:31–6.
53. Harrow J, Frankish A, Gonzalez JM, *et al*. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res* 2012;**22**:1760–74.
54. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
55. Farrell CM, O'Leary NA, Harte RA, *et al*. Current status and new features of the consensus coding sequence database. *Nucleic Acids Res* 2014;**42**:D865–72.
56. Stelzer G, Rosen N, Plaschkes I, *et al*. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics* 2016;**54**:1.30.1–33.
57. Schmidt EE, Pelz O, Buhlmann S, *et al*. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Res* 2013;**41**:D1021–6.
58. Gray KA, Seal RL, Tweedie S, *et al*. A review of the new HGNC gene family resource. *Hum Genomics* 2016;**10**(6).
59. Waterhouse RM, Tegenfeldt F, Li J, *et al*. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* 2013;**41**:D358–65.
60. Zhuo D, Zhao WD, Wright FA, *et al*. Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. *Genome Res* 2001;**11**:904.
61. Suzek BE, Huang H, McGarvey P, *et al*. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;**23**:1282–8.
62. Gaudet P, Michel P-A, Zahn-Zabal M, *et al*. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res* 2017;**45**:D177–82.
63. O'Leary NA, Wright MW, Brister JR, *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45.
64. Petryszak R, Keays M, Tang YA, *et al*. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* 2016;**44**:D746–52.
65. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database–2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
66. Oughtred R, Stark C, Breitkreutz B-J, *et al*. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;**47**:D529–41.
67. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;**40**:D857–61.
68. Fahey ME, Bennett MJ, Mahon C, *et al*. GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. *BMC Bioinformatics* 2011;**12**:298.
69. Orii N, Ganapathiraju MK. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. *PLoS One* 2012;**7**:e49029.
70. Cowley MJ, Pinese M, Kassahn KS, *et al*. PINA v2.0: mining interactome modules. *Nucleic Acids Res* 2012;**40**:D862–5.
71. Goll J, Rajagopala SV, Shiau SC, *et al*. MPIDB: the microbial protein interaction database. *Bioinformatics* 2008;**24**:1743–4.
72. Szklarczyk D, Franceschini A, Wyder S, *et al*. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;**43**:D447–52.
73. Pagel P, Kovac S, Oesterheld M, *et al*. The MIPS mammalian protein-protein interaction database. *Bioinformatics* 2005;**21**:832–4.
74. Orchard S, Ammari M, Aranda B, *et al*. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 2014;**42**:D358–63.
75. Salwinski L, Miller CS, Smith AJ, *et al*. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004;**32**:449D–51.

76. Barretina J, Caponigro G, Stransky N, *et al*. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603.
77. Klijn C, Durinck S, Stawiski EW, *et al*. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol* 2015;**33**:306–12.
78. Zhang J, Baran J, Cros A, *et al*. International cancer genome consortium data portal–a one-stop shop for cancer genomics data. *Database (Oxford)* 2011;**2011**:bar026.
79. NCBI GEO: archive for functional genomics data sets—update. Google Search.
80. Kolesnikov N, Hastings E, Keays M, *et al*. Array express update—simplifying data submissions. *Nucleic Acids Res* 2015;**43**:D1113–6.
81. Gollub J, Ball CA, Binkley G, *et al*. The Stanford microarray database: data access and quality assessment tools. *Nucleic Acids Res* 2003;**31**:94–6.
82. Reinhold WC, Sunshine M, Liu H, *et al*. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* 2012;**72**:3499–511.
83. Rajapakse VN, Luna A, Yamade M, *et al*. CellMiner-CDB for integrative cross-database genomics and pharmacogenomics analyses of cancer cell lines. *iScience* 2018;**10**:247–64.
84. Cerami EG, Gross BE, Demir E, *et al*. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;**39**:D685–90.
85. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
86. Fabregat A, Jupe S, Matthews L, *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55.
87. Caspi R, Billington R, Fulcher CA, *et al*. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 2018;**46**:D633–9.
88. Gutmanas A, Alhroub Y, Battle GM, *et al*. PDBe: protein data bank in Europe. *Nucleic Acids Res* 2014;**42**:D285–91.
89. Lomize MA, Lomize AL, Pogozheva ID, *et al*. OPM: orientations of proteins in membranes database. *Bioinformatics* 2006;**22**:623–5.
90. Hodis E, Prilusky J, Martz E, *et al*. Proteopedia—a scientific ‘wiki’ bridging the rift between 3D structure and function of biomacromolecules. *Genome Biol* 2008;**9**:R121.
91. Weekes D, Krishna SS, Bakolitsa C, *et al*. TOPSAN: a collaborative annotation environment for structural genomics. *BMC Bioinformatics* 2010;**11**:426.
92. Pándy-Szekeres G, Munk C, Tsonkov TM, *et al*. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res* 2018;**46**:D440–6.
93. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**:D427–32.
94. Mi H, Muruganujan A, Ebert D, *et al*. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 2019;**47**:D419–26.
95. Gaulton A, Hersey A, Nowotka M, *et al*. The ChEMBL database in 2017. *Nucleic Acids Res* 2016;**45**:D945–54.
96. Wang Y, Bryant SH, Cheng T, *et al*. PubChem BioAssay: 2017 update. *Nucleic Acids Res* 2016;**45**:D955–63.
97. Gilson MK, Liu T, Baitaluk M, *et al*. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;**44**:D1045–53.
98. Ursu O, Holmes J, Bologa CG, *et al*. DrugCentral 2018: an update. *Nucleic Acids Res* 2019;**47**:D963–70.
99. Alexander SPH, Kelly E, Marrion NV, *et al*. The concise guide to PHARMACOLOGY 2017/18: overview. *Br J Pharmacol* 2017;**174**.
100. Roth BL, Lopez E, Patel S, *et al*. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscience* 2000;**6**:252–62.
101. Wagner AH, Coffman AC, Ainscough BJ, *et al*. DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res* 2015;gkv1165.
102. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002;**30**:163–5.
103. Hecker N, Ahmed J, von Eichborn J, *et al*. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res* 2012;**40**:D1113–7.
104. Okuno Y, Tamon A, Yabuuchi H, *et al*. GLIDA: GPCR–ligand database for chemical genomics drug discovery–database and tools update. *Nucleic Acids Res* 2008;**36**:D907–12.
105. Daina A, Michielin O, Zoete V. Swiss Target Prediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res* 2019;**47**:W357–64.
106. Tamborero D, Rubio Pérez C, Déu Pons J, *et al*. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* 2018;**10**(25): 25.
107. Szklarczyk D, Santos A, von Mering C, *et al*. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;**44**:D380–4.
108. Carvalho-Silva D, Pierleoni A, Pignatelli M, *et al*. Open targets platform: new developments and updates two years on. *Nucleic Acids Res* 2019;**47**:D1056–65.
109. Forbes SA, Beare D, Boutselakis H, *et al*. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;**45**:D777–83.
110. Nguyen D-T, Mathias S, Bologa C, *et al*. Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res* 2017;**45**:D995–1002.
111. Dingerdissen HM, Torcivia-Rodriguez J, Hu Y, *et al*. BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Res* 2018;**46**:D1128–36.
112. Yang H, Qin C, Li YH, *et al*. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;**44**:D1069–74.
113. Amberger JS, Bocchini CA, Schiettecatte F, *et al*. OMIM.org: online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015;**43**:D789–98.
114. Brown AS, Patel CJ. A standard database for drug repositioning. *Sci Data* 2017;**4**.
115. Hodos RA, Kidd BA, Shameer K, *et al*. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip Rev Syst Biol Med* 2016;**8**:186–210.
116. Kuhn M, Letunic I, Jensen LJ, *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–9.
117. VigiAccess SPR. Promoting public access to VigiBase. *Indian J Pharm* 2016;**48**:606–7.
118. Bateman A, Martin MJ, O’Donovan C, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017; **45**:D158–69.

119. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
120. Wang, Yanli, *et al*. PubChem bioassay: 2014 update. *Nucleic acids research* 2013;**42**:D1075–D1082.
121. Lomenick B, Olsen RW, Huang J. Identification of direct protein targets of small molecules. *ACS Chem Biol* 2011;**6**: 34–46.
122. Elkins JM, Fedele V, Szklarz M, *et al*. Comprehensive characterization of the published kinase inhibitor set. *Nat Biotechnol* 2016;**34**(95).
123. Santos R, Ursu O, Gaulton A, *et al*. A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 2017;**16**:19–34.
124. Bamborough P, Drewry D, Harper G, *et al*. Assessment of chemical coverage of kinome space and its implications for kinase drug discovery. *J Med Chem* 2008;**51**:7898–914.
125. Pardanani A, Tefferi A. Imatinib targets other than bcr/abl and their clinical relevance in myeloid disorders. *Blood* 2004;**104**:1931–9.
126. van Dyck CH, Nygaard HB, Chen K, *et al*. Effect of AZD0530 on cerebral metabolic decline in Alzheimer disease. *JAMA Neurol* 2019.
127. Kaufman AC, Salazar SV, Haas LT, *et al*. Fyn inhibition rescues established memory and synapse loss in Alzheimer mice. *Ann Neurol* 2015;**77**:953–71.
128. Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;**8**:e1002503.
129. Li B, Fei DL, Flaveny CA, *et al*. Pyrvinium attenuates hedgehog signaling downstream of smoothened. *Cancer Res* 2014;**74**:4811–21.
130. Skrott Z, Mistrik M, Andersen KK, *et al*. Alcohol-abuse drug disulfiram targets cancer via p97 segregase adaptor NPL4. *Nature* 2017;**552**:194–9.
131. Caroli J, Sorrentino G, Forcato M, *et al*. GDA, a web-based tool for genomics and drugs integrated analysis. *Nucleic Acids Res* 2018;**46**:W148–56.
132. Carmody LC, Germain AR, VerPlank L, *et al*. Phenotypic high-throughput screening elucidates target pathway in breast cancer stem cell–like cells. *J Biomol Screen* 2012; **17**:1204–10.
133. Rees MG, Seashore-Ludlow B, Cheah JH, *et al*. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2016;**12**:109.
134. Levy H, Wang X, Kaldunski M, *et al*. Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. *Genes Immun* 2012;**13**: 593–604.
135. Wang Y, Yella J, Jegga AG. Transcriptomic data mining and repurposing for computational drug discovery. *Methods Mol Biol* 1903;**2019**:73–95.
136. Tranfaglia MR, Thibodeaux C, Mason DJ, *et al*. Repurposing available drugs for neurodevelopmental disorders: the fragile X experience. *Neuropharmacology* 2019;**147**:74–86.
137. Xu C, Ai D, Suo S, *et al*. Accurate drug repositioning through non-tissue-specific Core signatures from cancer transcriptomes. *Cell Rep* 2018;**25**:523–535.e5.
138. Newman AM, Liu CL, Green MR, *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;**12**:453–7.
139. Wang Z, Cao S. Morris JS, *et al*, Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *iScience* 2018;**9**:451–60.
140. Chen Z, Huang A, Sun J, *et al*. Inference of immune cell composition on the expression profiles of mouse tissue. *Sci Rep* 2017;**7**:40508.
141. Chen Z, Huang A, Sun J, *et al*. Inference of immune cell composition on the expression profiles of mouse tissue. *Sci Rep* 2017;**7**.
142. Mervin LH, Bulusu KC, Kalash L, *et al*. Orthologue chemical space and its influence on target prediction. *Bioinformatics* 2018;**34**(1):72–9.
143. Baig MS, Roy A, Saqib U, *et al*. Repurposing thioridazine (TDZ) as an anti-inflammatory agent. *Sci Rep* 2018;**8**:12471.
144. Mellini M, Di Muzio E, D'Angelo F, *et al*. In silico selection and experimental validation of FDA-approved drugs as anti-quorum sensing agents. *Front Microbiol* 2355;**2019**:10.
145. Chang Y, Park H, Yang H-J, *et al*. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep* 2018;**8**:8857.
146. Wei D, Liu C, Zheng X, *et al*. Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. *BMC Bioinformatics* 2019;**20**:44.
147. Zhang F, Wang M, Xi J, *et al*. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**:3355.
148. Tan M, Özgül OF, Bardak B, *et al*. Drug response prediction by ensemble learning and drug-induced gene expression signatures. 2018.
149. Zhang T, Zhang L, Payne PRO, *et al*. *Synergistic drug combination prediction by integrating multi-omics data in deep learning models*, 2018
150. Haider S, Yao CQ, Sabine VS, *et al*. Pathway-based subnetworks enable cross-disease biomarker discovery. *Nat Commun* 2018;**9**:4746.
151. Zhang S-D, Gant TW. sscMap: an extensible java application for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 2009;**10**:236.
152. Setoain J, Franch M, Martínez M, *et al*. NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic Acids Res* 2015;**43**:W193–9.
153. Chan J, Wang X, Turner JA, *et al*. Breaking the paradigm: Dr insight empowers signature-free, enhanced drug repurposing. *Bioinformatics* 2019;**35**:2818–26.
154. Patrick MT, Raja K, Miller K, *et al*. Drug repurposing prediction for immune-mediated cutaneous diseases using a word-embedding–based machine learning approach. *J Invest Dermatol* 2019;**139**:683–91.
155. Davis AP, Grondin CJ, Johnson RJ, *et al*. The comparative toxicogenomics database: update 2019. *Nucleic Acids Res* 2019;**47**:D948–54.
156. Cheng F, Desai RJ, Handy DE, *et al*. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 2018;**9**:2691.
157. Cai R, Zhang Y, Simmering JE, *et al*. Enhancing glycolysis attenuates Parkinson's disease progression in models and clinical databases. *J Clin Invest* 2019;**129**:4539–49.
158. Mirza B, Wang W, Wang J, *et al*. Machine learning and integrative analysis of biomedical big data. *Genes (Basel)* 2019;**10**:87.
159. Tyner JW, Tognon CE, Bottomly D, *et al*. Functional genomic landscape of acute myeloid leukaemia. *Nature* 2018;**562**:526–31.