Arch. Pharm. Res. (2019) 42:879–889
https://doi.org/10.1007/s12272-019-01176-3

Online ISSN 1976-3786
Print ISSN 0253-6269

Archives of Pharmacal Research
www.springer.com/12272

REVIEW

In silico drug repositioning: from large-scale transcriptome data to therapeutics

Ok-Seon Kwon[1,3] · Wankyu Kim[2] · Hyuk-Jin Cha[1,3] · Haeseung Lee[2]

Received: 30 January 2019 / Accepted: 26 July 2019 / Published online: 3 September 2019

Abstract Drug repositioning is an attractive alternative to conventional drug development when new beneficial effects of old drugs are clinically validated because pharmacokinetic and safety profiles are generally already available. Since ~ 30% of drugs newly approved by the US food and drug administration (FDA) are developed through drug repositioning, identifying novel usage for existing drugs is an emerging strategy for developing disease treatments. With advances in next-generation sequencing technologies, available transcriptome data related to diseases have expanded rapidly. Harnessing these resources enables a better understanding of disease mechanisms and drug mode of action (MOA), and moves toward personalized pharmacotherapy. In this review, we briefly outline publicly available large-scale transcriptome databases and tools for drug repositioning. We also highlight recent approaches leading to the discovery of novel drug targets, drug response biomarkers, drug indications, and drug MOA.

Keywords Drug repositioning · In silico drug repositioning · Transcriptome · Pharmacogenomics · Big data

✉ Haeseung Lee
haeseung@ewha.ac.kr

[1] Research Institute of Pharmaceutical Sciences, Seoul National University, Seoul 08826, Republic of Korea

[2] Ewha Research Center for Systems Biology, Department of Life Sciences, Ewha Womans University, Seoul 03760, Republic of Korea

[3] College of Pharmacy, Seoul National University, Seoul 08826, Republic of Korea

Introduction

Despite major scientific and technological advances in not only basic research but also drug discovery and development, the number of new drugs approved by the US Food and Drug Administration (FDA) has steadily declined. This trend is called Eroom's Law, and is the reverse of the more familiar Moore's Law that refers to the exponential increase in the number of transistors in a dense integrated circuit (Scannell et al. 2012). To improve the productivity and success rate of drug discovery, many novel strategies have been developed, including target structure-based drug design (Chen and Butte 2016), disease modeling with stem cell technology (Xia and Wong 2012; Kondo et al. 2017), and drug repositioning (or repurposing) (Hernandez et al. 2017).

Among these strategies, drug repositioning could potentially overcome the challenges of drug discovery (Hernandez et al. 2017; Novac 2013). In this approach, toxicity, pharmacokinetics, and pharmacodynamics profiles of a given drug are fully characterized during clinical trials. Once the efficacy of old drugs or drug candidates on new indications is established, successful therapeutic intervention can be anticipated with less risk from failure during clinical trials. There are many examples of serendipitous success for drug repositioning, as exemplified by thalidomide (originally developed for sedation and nausea, and now prescribed for multiple myeloma), sildenafil (originally developed for angina but now used to treat male erectile dysfunction), and raloxifene (originally developed for breast cancer and now an established treatment for osteoporosis) (Pushpakom et al. 2018). Currently, ~ 30% of drugs newly approved by the FDA are derived from drug repositioning. Encouraged by these successes, more systematic approaches are being applied to identify new

indications for old drugs, drug candidates, and drugs withdrawn from the market via drug repositioning and repurposing (Kim et al. 2016).

Recent advances in next-generation sequencing and high-throughput technologies have rapidly expanded available biological and chemical datasets, ushering in the era of big data (Costa 2014). Furthermore, computational methods taking advantage of these datasets have been actively developed to couple diseases with novel therapeutics (Jin and Wong 2014). In particular, data-driven approaches based on large-scale transcriptome data have accelerated the discovery of candidate drugs across a wide range of diseases (Chen and Butte 2016), some of which are enrolled in clinical trials (Jahchan et al. 2013). Transcriptional profiling, especially of mRNA, provides a comprehensive view of biological changes that reflect the overall consequences of multiple genetic variations. Thus, disease mechanisms and drug mode of action (MOA) can often be elucidated based on altered transcriptome profiles.

In this review, we describe publicly available resources widely used for transcriptome-based drug repositioning, and focus on recently developed computational approaches to identify new drug targets, drug response biomarkers, drug indication, and drug MOA that utilize these databases.

Public resources for in silico drug repositioning

A number of notable reference datasets widely used in transcriptome-based drug repositioning have recently been published and updated (Table 1) (Chen and Butte 2016; Kannan et al. 2016). These are divided into (i) disease-based, (ii) drug-based, and (iii) knowledge-based datasets depending on the biological perspective that the data describe. Disease-based datasets such as The Cancer Genome Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE) include gene expression profiles and detailed information on clinical or preclinical samples. Drug-based datasets such as CMap, LINCS, and CTRP include gene expression profiles of drug perturbation, drug efficacy, known targets, and other drug-related features. Knowledge-based datasets such as Gene Ontology, MSigDB, and KEGG include gene or protein functional annotations that are used to understand or interpret mechanisms of disease or drug action based on a set of genes. Datasets listed herein have been thoroughly investigated for hypothesis testing and discovery in the pharmacogenomics field, but we selectively describe their applications in identifying drug targets, drug response biomarkers, drug indication, and drug MOA (Fig. 1).

Clinical and preclinical transcriptome data

With rapid advances in systemic approaches, integrative analysis of multiple-layer omics data from various sources is widely used not only to identify drug–disease relationships, but also to discover optimal drug targets and/or drugs for diseases (Kannan et al. 2016). Thus, it is of the utmost importance to develop better integrative platforms and databases for disease-related information that are publically accessible.

Patient-derived transcriptome data

The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov) project, a joint collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), is one of the largest public resources for multi-layer cancer genomics, with over 11,000 patient profiles representing 36 cancer types, and 15 genomic assays per tumor type. TCGA data contain information on tumors such as gene expression, copy number variation, somatic mutations, single-nucleotide polymorphisms (SNPs), and clinical outcomes with pathological annotation. Even though TCGA database contains comprehensive information of cancer, TCGA has not been complete in the aspect of missing information such as transcriptome from normal tissue or drug treatment history. In this line, considering the difference of analytical breadth such as incomplete information, and emergent themes across cancer type and organ of origins, TCGA launched the Pan-Cancer analysis project to provide comprehensive information about cancer. Through TCGA Pan-Cancer Atlas project, comprehensive database of 12 different tumor type including a total of 5074 tumor sample has been assessed for clinical, genomic, epigenomic, transcriptional and proteomic data on at least one platform each (Cancer Genome Atlas Research et al. 2013). Moreover, Pan-Cancer Atlase reclassifies human tumor into three major categories based on molecular similarities: cell-of-origin pattern, oncogenic processes and signaling pathways (Sanchez-Vega et al. 2018). The integrative data from TCGA Pan-Cancer Atlas, this is a powerful emerging resource as we enter a new era of cancer treatment.

Cell line-derived transcriptome data

The Cancer Cell Line Encyclopedia (CCLE; https://portals.broadinstitute.org/ccle) database is a large-scale genomic dataset of gene expression, copy number, and DNA sequencing data from 1457 human cancer cell lines, encompassing 36 tumor types. Similarly, transcriptome data from in vitro cancer cell lines are provided in the Genomics of Drug Sensitivity in Cancer (GDSC) database,

Table 1 Publically accessible databases widely used in transcriptomic-based in silico drug repositioning

Category	Name	Description (as of December 2018)	URL
Omics data repositories	GEO	Raw and processed transcriptome data from multiple platforms	https://www.ncbi.nim.nih.gov/geo/
	SRA	Sequencing data from multiple platforms	https://www.ncbi.nlm.nih.gov/sra
	ArrayExpress	Raw and processed transcriptome data from multiple platforms	https://www.ebi.ac.uk/arrayexpress/
Disease-based	ICGC	Genomic, transcriptomic, epigenomic and clinical data from > 24,000 tumours (22 different tumour types)	https://icgc.org/
	TCGA	Genomic, transcriptomic, epigenomic and clinical data from > 11,000 tumours (33 different tumour types)	http://tcga-data.nci.nih.gov/
	CCLE	Genomic, transcriptomic and epigenomic data from > 1000 cancer cell lines	http://www.broadinstitute.org/ccle
	GDSE	Genomic, transcriptomic and epigenomic data from > 1000 cancer cell lines	https://wwwcancerrxgene.org/
Durg-based	CMap	Gene expression profiles for 1309 chemical compounds in 5 cancer cell lines	https://portals.broadinstitute.org/cmap/
	LINCS	Gene expression profiles for perturbagens (20,413 chemicals and 2119 genetic knockdown/overexpression) across 77 cell lines	https://clue.io/
	NCI60	Drug response data (GI50, LC50 values) of 60 cancer cell lines for 45,449 compounds	https://dtp.cancer.gov/discovety_development/nci-60/
	CTRP	Drug response data (AUC, EC50 values) of 860 cancer cell lines for 481 compounds	https://portals.broadinstitute.org/ctrp/
	CCLE	Drug response data (AUC, IC50 values) of 504 cancer cell lines for 24 compounds	http://www.broadinstitute.org/ccle
	GDSE	Drug response data (AUC, IC50 values) of 714 cancer cell lines for 142 compounds	https://www.cancerrxgene.org/
	NCl-ALMANAC	Therapeutic activity for pairwise combinations (> 5000 pairs) of 104 FDA-approved anticancer drugs against NCI-60 cell lines	https://dtp.cancer.govincialmanac
	PubChem Bioassay	Chemical compound screening data, including > 3.4 M unique chemical compounds, > 12 K protein targets, and > 1 M assays.	https://pubchem.ncbi.nlm.nih.gov/
	ChEMBL	Chemical compound screening data, including > 2.2 M unique chemical compounds, > 12 K protein targets, and > 1 M assays	https://www.ebi.ac.uk/chembl/
Knowledge-based	Gene ontology	Database collection of over 15,000 genes with gene-ontology, including 13,212 biological process, 1547 cellular components and 4162 molecular functions	http://www.geneontology.org/
	MsigDB	Data repository, contained 17,810 genes sets, with 8 major collection Database collection of over 2300 biological pathways for 25 different species	http://software.broadinstitute.org/gsea/msigdb http://www.wikipathways.org
	KEGG	Database collection for genomes, pathways, disease and compounds information, including 3947 genes, 200 pathway and 9324 gene-pathway association	http://www.genome.jp/kegg
	BioCarta	Database collection of 1396 genes with 254 pathway and 4417 gene-pathway association	http://www.biocarta.com
	Reactome	Database collection of 7535 genes with 1638 pathway and 83,680 gene-pathway association	http://www.reactome.org

which includes over 1000 cancer cell lines, and NCI-60, which covers 60 cancer cell lines. Moreover, by compiling cell line and compound sensitivity data using CCLE, GDSC, and NCI-60, a profound understanding of the connections between pharmacological vulnerability and molecular signature of a responsive cancer cell line can be gleaned (Barretina et al. 2012; Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer 2015). However, standardization with additional curation and processing to combine information of cell line and drug

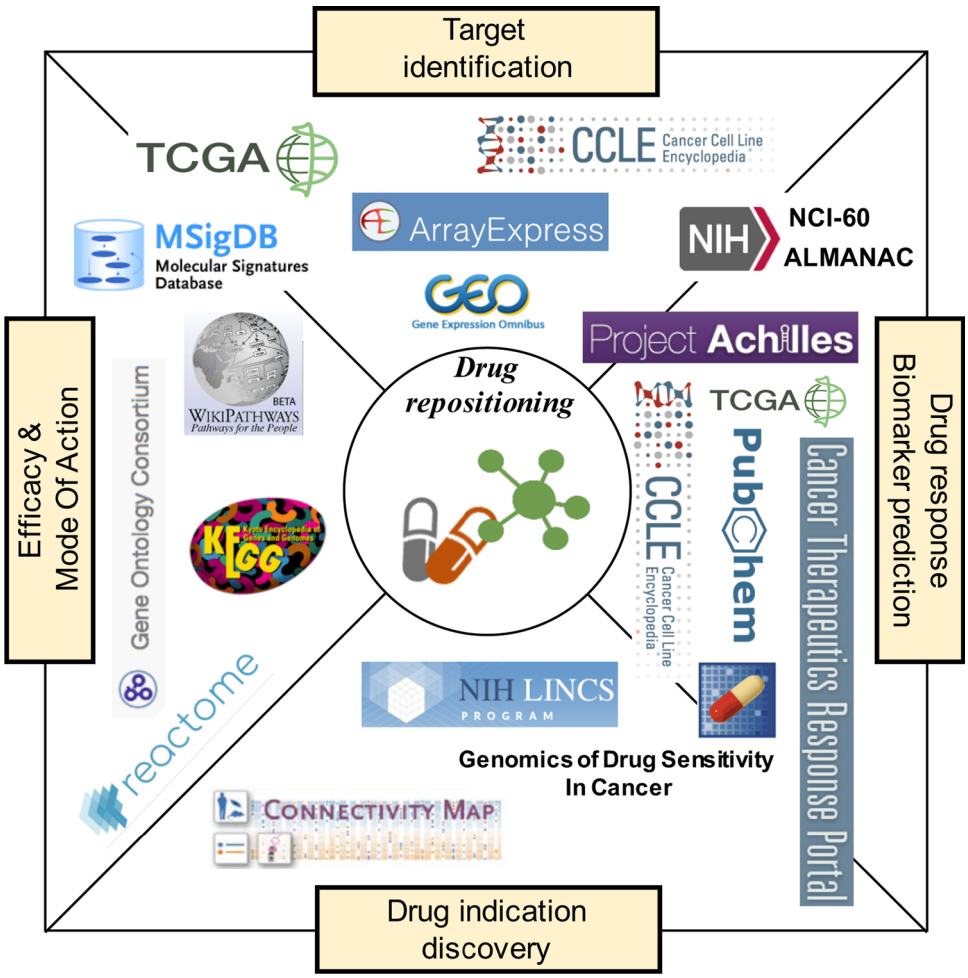

Fig. 1 Public databases utilized in drug repositioning pipelines

treatment is a challenge that is required to enable integrative analysis.

Transcriptome data following treatment

The connectivity map (build 02)

The connectivity map (CMap) is a collection of genome-wide gene expression data from five human cancer cell lines treated with 1309 compounds obtained using the Affymetrix microarray platform (Lamb et al. 2006). The concept of CMap is to establish a comprehensive reference database of drug-induced gene expression profiles to compare with a set of genes representing the biological state of interest, and to discover functional connections between them. It provides a web-based tool that performs simple pattern matching analysis with CMap reference data based on a user-submitted gene list, but is no longer updated or modified (https://portals.broadinstitute.org/cmap/).

Library of integrated network-based cellular signatures (LINCS) L1000

LINCS L1000, also referred to as LINCS, or an extended version of CMap, is a resource containing 1.3 million gene expression profiles associated with 20,413 chemical perturbagens (e.g., small molecules or drugs) and ~ 5000 genetic perturbagens (e.g., single-gene knockdown or overexpression) (Subramanian et al. 2017). Data were acquired using the L1000 assay developed by the Broad Institute CMap team to facilitate rapid high-throughput gene expression profiling at low cost. The L1000 assay measures the expression of 978 landmark genes, and expression values for remaining genes are estimated by a linear model using a diverse collection of transcriptome data from Affymetrix microarray data in Gene Expression Omnibus (GEO). LINCS L1000 datasets are fully downloadable from GEO (accession: GSE92742) and are easily accessible via the cloud-based software platform CLUE (https://clue.io/).

Knowledge-based gene annotations

In drug development pipelines, the knowledge base, which includes information on drugs, biological implications of drugs, and clinical outcomes, can reveal associations and thereby provide integrative implications (Fotis et al. 2018). To provide biological insight relevant to drug development, utilizing molecular interaction data gathered from various knowledge bases is a potentially powerful method. Below, gene annotation databases that illuminate the biological background by exploring molecular mechanisms and molecular interactions are briefly described.

Gene annotation databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and the Molecular Signatures Database (MSigDB) provide diverse types of interaction models, including signaling pathways, metabolic networks, and regulatory interactions, based on transcriptome data. The KEGG database collection integrates genomic and chemical information. In terms of systemic information, the KEGG database includes KEGG Pathway containing pathway maps, KEGG Disease comprising disease entries, and KEGG Drug that includes comprehensive information on drugs, approved in Japan, the USA, and Europe. In particular, KEGG Pathway, which contains manually drawn pathway maps, provides intuitive information on interactions between genes and proteins (Kanehisa et al. 2018). By contrast, the GO project aims to provide ontologies of genes defined with their own properties. GO provides ontologies and annotation information for three domains: cellular component (CC), biological process (BP), and molecular function (MF) (Zhang et al. 2014; Rhee et al. 2008). Meanwhile, MSigDB, developed for gene set enrichment analysis (GSEA), covers a large number of gene sets with annotations and links from external resources including KEGG, GO, GEO, and ArrayExpress (Liberzon et al. 2011). Together, these knowledge-based databases provide a foundation for computational drug repositioning based on transcriptome analysis, and collate valuable information such as target identification and MOA of drugs.

Web-based drug repositioning tools

Exploring complex large data sets described above often requires high-performance computing resources but access is difficult without proficient computer skills. A number of user-friendly interface-based web tools that assist research in drug repositioning have lowered this barrier for all scientists regardless of their computational backgrounds (Sam and Athri 2019). Since most transcriptome-based studies initiate hypothesis testing on the sets of differentially expressed genes (DEGs) that represent the biological state of interest, various web-based analytic tools have been developed to associate these DEGs with drugs.

CLUE (https://clue.io/l1000-query) provides a cloud-based query tool to find positive or negative connections between a user-submitted gene set and all the signatures in LINCS L1000 (Subramanian et al. 2017). The term signature here refers to a vector of differential gene expression values (Z score) induced by individual perturbagen in LINCS L1000. CLUE returns a list of approximately 50,000 unique perturbagens, including small molecules, single-gene knockdown and overexpression, with a score based on the amount of inducing expressional changes of the input genes.

L1000CDS^2 (http://amp.pharm.mssm.edu/L1000CDS2/) is another LINCS L1000 signature search engine (Duan et al. 2016). It processed LINCS L1000 data to define the signatures using the characteristic direction method (Clark et al. 2014). Predictive performance of *L1000CDS*^{2} was tested on expression signatures from human cells infected with Ebola virus. Based on these signatures, kenpaullone, a GSK3B/CDK2 inhibitor was predicted and its dose-dependent efficacy in inhibiting Ebola infection was validated in vitro.

DeSigN (http://design-v2.cancerresearch.my/query) associates drug efficacy with a user-submitted gene set by comparing it against drug response-related gene expression signatures for 140 drugs (Lee et al. 2017). The individual expression signature of a drug was defined as a differential gene expression profile derived by using its drug sensitivity (IC50) and baseline gene expression against cancer cell lines in GDSC data. DeSigN was validated using four different drug sensitivity studies deposited in the GEO database. In addition, bosutinib, a src tyrosine kinase inhibitor, was predicted as a sensitive drug for oral squamous cell carcinoma (OSCC) and its efficacy was demonstrated by in vitro viability assay.

Prediction of novel drug–target interactions

Drug polypharmacology (Hopkins 2008), in which a single drug acts on multiple targets, implies the therapeutic potential of a drug for new indications, and thus facilitates innovative and successful drug repositioning (Reddy and Zhang 2013). Drug-induced transcriptome data reflect the combined effects of multiple targets of a drug, providing insight into its MOA or unintended off-targets. CMap and LINCS are the most comprehensive resources for exploring novel drug–target interactions (DTIs). From CMap data, high correlations among gene expression changes caused by drugs sharing the same target have been systematically shown (Wang et al. 2013). Several methods have been developed to expand known drug–target relationships

based on drug similarity at the gene expression level (Hizukuri et al. 2015; Iwata et al. 2017). On the other hand, drug-induced differentially expressed genes (DEGs) comprise only a small proportion of known target genes, but are distributed close to targets in the functional protein–protein interaction (PPI) network (Isik et al. 2015). Based on these observations, a target prediction model was developed that integrates drug-induced DEGs and the network topology of PPIs.

Genetically perturbed transcriptome data can also be utilized to seek novel DTIs using drug-induced transcriptome data. Importantly, novel connections between a drug and its target gene can be inferred from common expression signatures shared by both drug treatments and loss of gene function in yeast systems (Hughes et al. 2000). This idea was applied in human cancer cells using LINCS L1000 data, which led to the discovery of compound BRD-1868 that targets Casein Kinase 1A1, which is related to drug resistance in lung cancer (Lantermann et al. 2015; Subramanian et al. 2017). Another similar approach comprehensively predicted novel DTIs between 1124 drugs and 829 target proteins by correlating gene expression patterns caused by chemical and genetic perturbations (Sawada et al. 2018). Notably, this approach distinguished predicted DTIs by inhibitory and activatory interactions, depending on whether a genetic perturbation directly compared with a drug is knockdown or overexpression.

Identification of drug response biomarkers

In drug development pipelines, most drugs are developed based on the molecular features of a given disease. In terms of drug repositioning, identification of indicators or biomarkers of repurposed drugs is critical to match the appropriate drug with the right patient based on predicted drug responses (Kelloff and Sigman 2012). With great advances in sequencing technologies, large-scale transcriptome data and pharmacogenomics-based disease models have emerged that aid the identification of biomarkers and the prediction of drug responses. In the Cancer Therapeutics Response Portal (CTRP) database, transcriptome-based biomarkers of drug sensitivity have been identified by integrating drug response profiles for 481 anticancer drugs across 860 cancer cell lines (Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer 2015). Drug response profiles from CTRP can be utilized to predict drug responses in cell lines, which have particular disease features or defined gene signatures, suggesting that drugs may sensitize certain disease features. For example, sensitivity patterns of 481 chemical compounds were correlated with $\sim$ 19,000 basal transcript levels across 823 different human cancer cell lines, and this demonstrated that analyzing the basal gene expression profile of cell lines can predict drug responses and illuminate the mechanisms of small molecules (Rees et al. 2016). Furthermore, based on validation with previously annotated targets and drugs, as exemplified by BCL2 and ABT-199 (Rees et al. 2016) and SLC35F2 and YM-155 (Winter et al. 2014; Rees et al. 2016), ML239 was newly identified, after being originally identified by phenotypic screening to selectively eliminate epithelial breast cancer cells, and found to activate fatty acid desaturase 2 (FADS2) (Rees et al. 2016). Furthermore, chemoresistance score was defined, which is strongly correlated with mesenchymal cancer traits, by leveraging integrative transcriptome data from both CTRP and CCLE (Hong et al. 2018). Furthermore, analyzing the association between drug response profiles and genome-wide RNAi screening data in the Achilles project (Tsherniak et al. 2017) identified *ITGB3*, highly expressed in mesenchymal-type lung cancer cell lines (Bae et al. 2016; Hong et al. 2016), as an Achilles' heel for chemoresistant cancer cells with mesenchymal traits. In conclusion, dependency on ITGB3 was considered to be one of the major factors determining the responses of most chemotherapeutic drugs (Hong et al. 2018). Thus, leveraging publicly available pharmacogenomics data linked to diseases offers a promising approach for identifying drug biomarkers with statistical reliability.

Discovery of novel drug indications

Systems biology approaches have utilized large-scale pharmacogenomics data to identify previously unrecognized relationships between diseases and drugs. These approaches generally begin by defining a gene expression signature (e.g., a collection of genes representing a disease state) and comparing it directly against compound signatures in reference databases such as CMap or CTRP. This query signature can be derived from a disease, drug perturbation, or genetic perturbation, and used to perform (i) drug–disease, (ii) drug–drug, or (iii) drug–gene comparisons (Fig. 2). Below, several studies that have discovered new drug indications through such comparisons are described.

The first case, the most prevalent approach, typically defines a disease signature as a set of DEGs obtained by comparing disease and corresponding control (healthy) states, and seeks a drug whose perturbation reverses the disease signature. For example, disease signatures were generated for 100 diseases using microarray data from GEO, and each disease signature was mapped to 164 drug signatures in CMap (build 01) (Dudley et al. 2011; Sirota et al. 2011). Among the highly anti-correlated disease–drug pairs, many known disease–drug relationships were

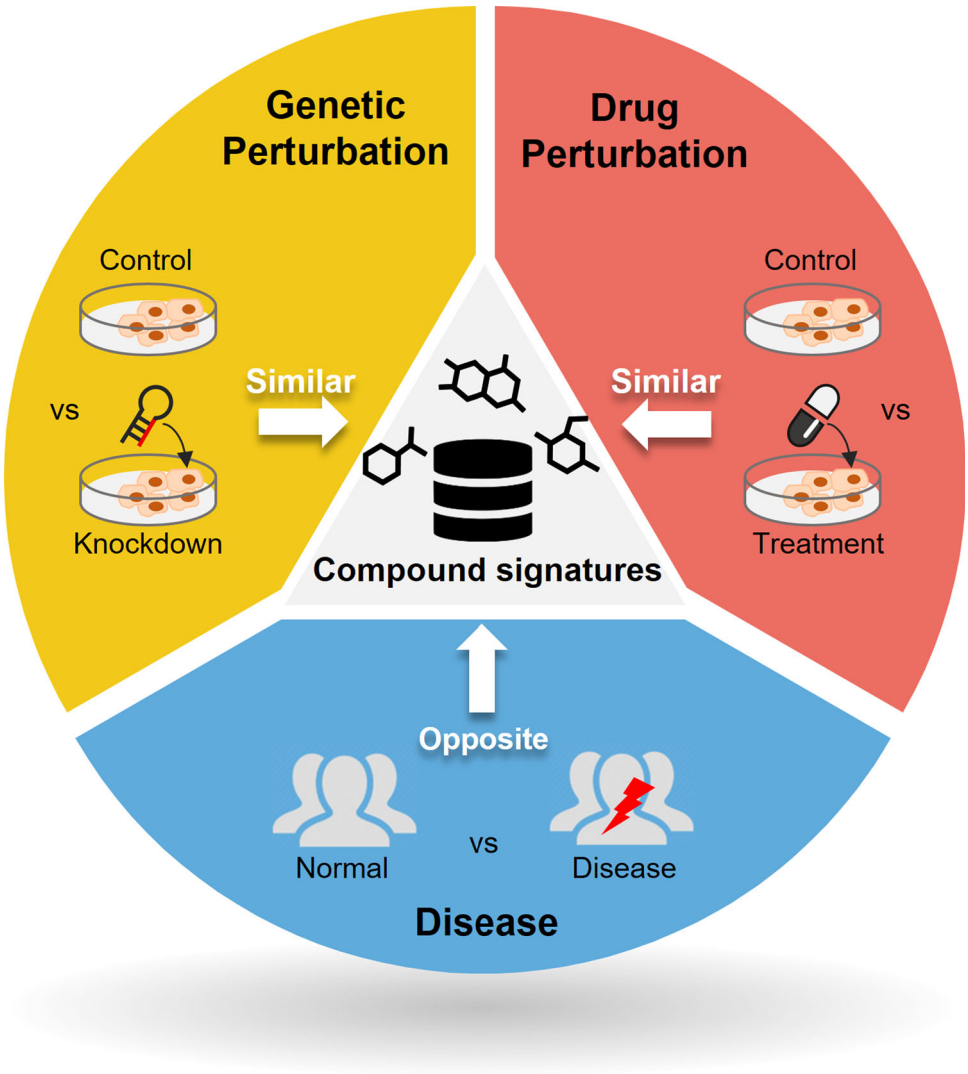

Fig. 2 Three signature types (disease, drug perturbation, and genetic perturbation) used to compare compound signatures in pharmacogenomics databases for identifying novel relationships between diseases and drugs

recovered, along with new associations including cimetidine (a histamine H_2 receptor antagonist for antiulcer treatment) for the treatment of lung adenocarcinoma, and topiramate (a voltage-gated sodium and calcium channel blocker as an anticonvulsant) for the treatment of inflammatory bowel disease. A similar systematic approach using a small cell lung cancer (SCLC) expression signature found that antidepressant drugs (imipramine, a tricyclic antidepressant; promethazine, a histamine H_1 receptor antagonist for allergies; and bepridil, an amine calcium channel blocker) are potent inducers of apoptosis in SCLC (Jahchan et al. 2013). These findings led to the enrolment into clinical trials of a related molecule, the tricyclic antidepressant desipramine, for the treatment of SCLC (NCT01719861, phase IIa clinical trials). In another example, comparison of a metastatic colon signature against compound signatures in CMap (build 02) resulted in the identification of citalopram (a selective serotonin reuptake inhibitor and antidepressant), troglitazone (a ligand mimetic of PPARγ and antihyperglycemic agent), and enilconazole (a fungicide) drugs for the treatment of colorectal cancer metastasis (van Noort et al. 2014). A common assumption of these studies is that a strong anti-correlation between a disease and drug signatures indicates that the drug may potentially have a therapeutic effect on the disease.

Disease signatures can also be used to characterize disease states for other biological systems (e.g., cancer cells or organoids), and may be associated with drug activity such as IC50, EC50, and AUC values. For example, a mesenchymal score was calculated using a mesenchymal signature for each cancer cell line available in CTRP, and correlated with cell line sensitivity against 481 compounds (Viswanathan et al. 2017). The authors found that ferroptosis inducers (e.g., RSL3, ML210, and ML162) were selectively potent against mesenchymal cancer cells via inhibition of a lipid peroxidase pathway. A similar approach using a YM155-resistant signature led to the

discovery of BCL2 homology 3 mimetics (ABT-263, ABT-737, and WEHI-539) that selectively ablate abnormal human embryonic stem cells (hESCs) resistant to YM155, which specifically eliminates undifferentiated hESCs (Lee et al. 2013; Cho et al. 2018).

Drug–drug comparisons can be used to extrapolate knowledge on a given drug to other drugs based on similarity, assuming that drugs whose perturbations cause similar gene expression changes may have similar therapeutic effects. Indeed, drugs with similar MOAs were significantly enriched in the sub-modules of a large-scale drug association network constructed based on drug-induced transcriptional similarity from CMap data (Iorio et al. 2010). In this network, fasudil, a Rho-kinase inhibitor and vasodilator, was clustered with well-known autophagy inducers, and its effect on autophagy enhancement was validated.

LINCS contains drug-induced transcriptome data for additional perturbagens causing perturbations 15-fold beyond the range included in CMap, providing an excellent opportunity for exploring candidate compounds. One approach retrieved LINCS data to identify drugs whose signatures (i.e., DEGs from comparisons before and after drug treatment) are similar to those of known glioblastoma (GBM) drugs (Lee et al. 2016). By integrating this signature similarity with other features such as drug targets and chemical structures, 14 drugs were predicted for the treatment of GBM, and more than half displayed anti-proliferative activity against patient-derived GBM cells.

In the final case, a disease is linked to a drug based on similarity between transcriptomic signatures generated from a genetic perturbation (e.g., knockdown or overexpression of a disease biomarker) and a drug. This concept was first applied as an alternative to targeting the poorly druggable gene encoding integrin beta 3 (*ITGB3*), responsible for chemoresistance in mesenchymal lung cancer (Hong et al. 2018). From LINCS data, atorvastatin (a HMG-CoA reductase inhibitor for anti-dyslipidemia) mimicked expression changes caused by knockdown of *ITGB3* and was identified as a chemosensitizer. A similar approach was performed for the *N*-acetylgalactosaminyl-transferase 14 (*GALNT14)* protein, the expression of which is strongly correlated with lung cancer recurrence and metastasis (Lee et al. 2008; Kwon et al. 2015). Due to a lack of feasible drugs that directly inhibit the GALNT14 protein, the authors generated the gene expression signature of shRNA-mediated *GALNT14* depletion in metastatic lung cancer, and identified bortezomib (the first-in-class proteasome inhibitor for multiple myeloma) (Argyriou et al. 2008), which likely reverses *GALNT14*-dependent gene expression (Kwon et al. 2018).

All the above studies successfully identified drugs by matching transcriptomic signatures of drugs and therapeutic targets, rather than attempting to inhibit these protein targets directly. Given that many potential molecular targets identified from cancer genomic profiling are undruggable (Lazo and Sharlow 2016), this approach could prove to be a viable strategy in cancer pharmacology.

Identification of drug mode of action

Identification of the molecular pathways and adverse effects of a compound are crucial for drug repositioning. Traditionally, the MOA of a drug has been predicted based on analysis of chemical structure, gene expression profiles following drug treatment (Lamb 2007), and side effect similarity (Campillos et al. 2008). Furthermore, most of these approaches are only applied to drugs that are well-characterized based on the available structure and documented side effect (Iorio et al. 2010). Thus, when prior information on drugs is lacking, gene signature-based methods are the most cost-effective approach for elucidating the MOA (Lamb et al. 2006; Lamb 2007). In this regard, a subset of differential gene expression data following treatment can provide profound information on connections between drugs, pathways, and diseases through pharmacogenomics (e.g., CMap and LINCS) and pathway (e.g., KEGG and GO) databases. As an example of the power of identifying the MOA of a drug, fasudil was newly identified to induce cellular autophagy through network analysis, and could therefore be applicable for neurodegenerative disorders (Iorio et al. 2010). Moreover, they provide their approach for discovering MOA with publically accessible tool, named as Mode of Action by NeTwoRk Analysis (MANTRA, http://mantra.tigem.it). Similarly, another study analyzed 16,268 compound and 68 human cell lines, gathered from LINCS, and performed pathway enrichment analysis to reveal active pathways (Iwata et al. 2017). By mapping onto KEGG biological pathways using genes up- or down-regulated by the compound, the authors proposed a computational approach to identify not only active pathways, but also target proteins and therapeutic indications. In another example, bortezomib was found to interrupt tumor metastasis in lung cancer (Kwon et al. 2018), caused by an unexpected off-target effect within cells, independent of proteasome inhibition. The authors subsequently discovered the MOA of bortezomib on metastasis by conducting KEGG and GSEA analyses with combined transcriptome data from bortezomib-treated cells and GALNT14 expressing lung cancer patients from TCGA. It was concluded that bortezomib suppresses the TGFβ-dependent gene signature, and thereby inhibits tumor metastasis in lung cancer (Kwon et al. 2018).

Conclusions and future directions

With rapid advances in technology, complex transcriptome datasets that reflect disease systems, ranging from single cells to patients, are emerging and expanding. Diverse transcriptome datasets enable researchers to achieve efficient drug repositioning by prediction of drug side effects, drug indications, and drug MOA. However, transcriptome data has inherent challenging limitations. First, integrating gene expression data from different platforms (e.g., microarray, RNA-seq, or L1000 assay) is intriguing as the range of genes measured varies depending on the platform. Second, since changes in mRNA levels are relatively more sensitive than changes in DNA molecules, the effects of various environmental factors that induce gene expression changes are comprehensively reflected in transcriptome data. In particular, large-scale data such as CMap generated over a long period of time contain considerable experimental variations or noise due to batch effects. Additional pre-processing steps are therefore required to yield reliable results. Third, in contrast to detecting genetic variations compared to a reference genome, there is no representative reference to define whether gene expression levels are high or low. To determine if gene expression has changed under a certain condition, there should be comparable gene expression data under a control condition. The former two limitations can be overcome by applying sophisticated normalization methods, but the last one is still challenging.

Despite these limitations, systems biology approaches leveraging these datasets can reveal previously unrecognized relationships between drugs and diseases, and provide alternative strategies for associating drugs with newly identified targets, regardless of their druggability. Since drugging undruggable molecular targets represents a major hurdle for traditional drug discovery methods, new therapeutic indications for existing drugs can prove crucial for certain diseases. In particular, for personalized pharmacotherapy, an emerging approach for disease treatment that takes into account individual characteristics of each patient, discovering therapeutic indications from matched drugs may be of great benefit to drug development and drug repositioning. In this article, we reviewed publicly available large-scale transcriptome databases and multidisciplinary methods utilizing the data within them. These big data approaches hold great promise for overcoming the limitations of traditional drug discovery pipelines and supporting the field of precision pharmacology.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) via grants funded by the Korea government (MSIT; NRF-2017R1A6A3A11030794, NRF-2017M3C9A5028690, and NRF-2019R1C1C1008710) to HSL, WKK, and OSK.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Argyriou AA, Iconomou G, Kalofonos HP (2008) Bortezomib-induced peripheral neuropathy in multiple myeloma: a comprehensive review of the literature. Blood 112:1593–1599

Bae GY, Hong SK, Park JR, Kwon OS, Kim KT, Koo J, Oh E, Cha HJ (2016) Chronic TGFbeta stimulation promotes the metastatic potential of lung cancer cells by Snail protein stabilization through integrin beta3-Akt-GSK3beta signaling. Oncotarget 7:25366–25376

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, De Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, Macconaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483:603–607

Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321:263–266

Cancer Cell Line Encyclopedia, Genomics of Drug Sensitivity in Cancer (2015) Pharmacogenomic agreement between two cancer cell line data sets. Nature 528:84–87

Cancer Genome Atlas Research, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45:1113–1120

Chen B, Butte AJ (2016) Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther 99:285–297

Cho SJ, Kim KT, Jeong HC, Park JC, Kwon OS, Song YH, Shin JG, Kang S, Kim W, Shin HD, Lee MO, Moon SH, Cha HJ (2018) Selective elimination of culture-adapted human embryonic stem cells with bh3 mimetics. Stem Cell Rep 11:1244–1256

Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, Ma'ayan A (2014) The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinform 15:79

Costa FF (2014) Big data in biomedicine. Drug Discov Today 19:433–440

Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, Readhead B, Tritsch SR, Hodos R, Hafner M, Niepel M, Sorger PK, Dudley JT, Bavari S, Panchal RG, Ma'ayan A (2016) L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl 2:16015

Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, Morgan AA, Sarwal MM, Pasricha PJ, Butte AJ (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 3:96ra76

Fotis C, Antoranz A, Hatziavramidis D, Sakellaropoulos T, Alexopoulos LG (2018) Network-based technologies for early drug discovery. Drug Discov Today 23:626–635

Hernandez JJ, Pryszlak M, Smith L, Yanchus C, Kurji N, Shahani VM, Molinski SV (2017) Giving drugs a second chance:

overcoming regulatory and financial hurdles in repurposing approved drugs as cancer therapeutics. Front Oncol 7:273
Hizukuri Y, Sawada R, Yamanishi Y (2015) Predicting target proteins for drug candidate compounds based on drug-induced gene expression data in a chemical structure-independent manner. BMC Med Genom 8:82
Hong SK, Park JR, Kwon OS, Kim KT, Bae GY, Cha HJ (2016) Induction of integrin beta3 by sustained ERK activity promotes the invasiveness of TGFbeta-induced mesenchymal tumor cells. Cancer Lett 376:339–346
Hong SK, Lee H, Kwon OS, Song NY, Lee HJ, Kang S, Kim JH, Kim M, Kim W, Cha HJ (2018) Large-scale pharmacogenomics based drug discovery for ITGB3 dependent chemoresistance in mesenchymal lung cancer. Mol Cancer 17:175
Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690
Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell 102:109–126
Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, Di Bernardo D (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci USA 107:14621–14626
Isik Z, Baldow C, Cannistraci CV, Schroeder M (2015) Drug target prioritization by perturbed gene expression and network information. Sci Rep 5:17417
Iwata M, Sawada R, Iwata H, Kotera M, Yamanishi Y (2017) Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. Sci Rep 7:40164
Jahchan NS, Dudley JT, Mazur PK, Flores N, Yang D, Palmerton A, Zmoos AF, Vaka D, Tran KQ, Zhou M, Krasinska K, Riess JW, Neal JW, Khatri P, Park KS, Butte AJ, Sage J (2013) A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. Cancer Discov 3:1364–1377
Jin G, Wong ST (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today 19:637–644
Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M (2018) New approach for understanding genome variations in KEGG. Nucleic Acids Res 47:D590–D595
Kannan L, Ramos M, Re A, El-Hachem N, Safikhani Z, Gendoo DM, Davis S, Gomez-Cabrero D, Castelo R, Hansen KD, Carey VJ, Morgan M, Culhane AC, Haibe-Kains B, Waldron L (2016) Public data and open source tools for multi-assay genomic investigation of disease. Brief Bioinform 17:603–615
Kelloff GJ, Sigman CC (2012) Cancer biomarkers: selecting the right drug for the right patient. Nat Rev Drug Discov 11:201–214
Kim RS, Goossens N, Hoshida Y (2016) Use of big data in drug development for precision medicine. Expert Rev Precis Med Drug Dev 1:245–253
Kondo T, Imamura K, Funayama M, Tsukita K, Miyake M, Ohta A, Woltjen K, Nakagawa M, Asada T, Arai T, Kawakatsu S, Izumi Y, Kaji R, Iwata N, Inoue H (2017) iPSC-based compound screening and in vitro trials identify a synergistic anti-amyloid beta combination for Alzheimer's disease. Cell Rep 21:2304–2312
Kwon OS, Oh E, Park JR, Lee JS, Bae GY, Koo JH, Kim H, Choi YL, Choi YS, Kim J, Cha HJ (2015) GalNAc-T14 promotes metastasis through Wnt dependent HOXB9 expression in lung adenocarcinoma. Oncotarget 6:41916–41928
Kwon O-S, Lee H, Kong H-J, Park JE, Lee W, Kang S, Kim M, Kim W, Cha H-J (2018) In silico Drug Repositioning of bortezomib to reverse metastatic effect of GALNT14 in lung cancer. https://doi.org/10.1101/394163
Lamb J (2007) The connectivity map: a new tool for biomedical research. Nat Rev Cancer 7:54–60
Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313:1929–1935
Lantermann AB, Chen D, Mccutcheon K, Hoffman G, Frias E, Ruddy D, Rakiec D, Korn J, Mcallister G, Stegmeier F, Meyer MJ, Sharma SV (2015) Inhibition of casein kinase 1 alpha prevents acquired drug resistance to erlotinib in EGFR-mutant non-small cell lung cancer. Cancer Res 75:4937–4948
Lazo JS, Sharlow ER (2016) Drugging undruggable molecular cancer targets. Annu Rev Pharmacol Toxicol 56:23–40
Lee ES, Son DS, Kim SH, Lee J, Jo J, Han J, Kim H, Lee HJ, Choi HY, Jung Y, Park M, Lim YS, Kim K, Shim Y, Kim BC, Lee K, Huh N, Ko C, Park K, Lee JW, Choi YS, Kim J (2008) Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res 14:7397–7404
Lee MO, Moon SH, Jeong HC, Yi JY, Lee TH, Shim SH, Rhee YH, Lee SH, Oh SJ, Lee MY, Han MJ, Cho YS, Chung HM, Kim KS, Cha HJ (2013) Inhibition of pluripotent stem cell-derived teratoma formation by small molecules. Proc Natl Acad Sci USA 110:E3281–E3290
Lee H, Kang S, Kim W (2016) Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. PLoS ONE 11:e0150460
Lee BK, Tiong KH, Chang JK, Liew CS, Abdul Rahman ZA, Tan AC, Khang TF, Cheong SC (2017) DeSigN: connecting gene expression with therapeutics for drug repurposing and development. BMC Genom 18:934
Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27:1739–1740
Novac N (2013) Challenges and opportunities of drug repositioning. Trends Pharmacol Sci 34:267–272
Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, Mcnamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M (2018) Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18:41
Reddy AS, Zhang S (2013) Polypharmacology: drug discovery for the future. Expert Rev Clin Pharmacol 6:41–47
Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE, Soule CK, Alexander B, Li A, Montgomery P, Kotz JD, Hon CS, Munoz B, Liefeld T, Dancik V, Haber DA, Clish CB, Bittker JA, Palmer M, Wagner BK, Clemons PA, Shamji AF, Schreiber SL (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 12:109–116
Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9:509–515
Sam E, Athri P (2019) Web-based drug repurposing tools: a survey. Brief Bioinform 20:299–316
Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D, Daian F, Gao Q, Bailey MH, Liang WW, Foltz SM, Shmulevich I, Ding L, Heins Z, Ochoa A, Gross B, Gao J, Zhang H, Kundra R, Kandoth C, Bahceci I, Dervishi L, Dogrusoz U, Zhou W, Shen H, Laird PW, Way GP, Greene CS, Liang H, Xiao Y, Wang C, Iavarone A, Berger AH, Bivona TG, Lazar AJ,

Hammer GD, Giordano T, Kwong LN, Mcarthur G, Huang C, Tward AD, Frederick MJ, Mccormick F, Meyerson M, Cancer Genome Atlas Research, Van Allen EM, Cherniack AD, Ciriello G, Sander C, Schultz N (2018) Oncogenic signaling pathways in the cancer genome atlas. Cell 173(321–337):e10

Sawada R, Iwata M, Tabei Y, Yamato H, Yamanishi Y (2018) Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures. Sci Rep 8:156

Scannell JW, Blanckley A, Boldon H, Warrington B (2012) Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 11:191–200

Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 3:96ra77

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside P, Gray NS, Clemons PA, Silver S, Wu X, Zhao WN, Read-Button W, Wu X, Haggarty SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, Golub TR (2017) A next generation connectivity map: L1000 platform and the FIRST 1,000,000 profiles. Cell 171(1437–1452):e17

Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC (2017) Defining a cancer dependency map. Cell 170(564–576):e16

Van Noort V, Scholch S, Iskar M, Zeller G, Ostertag K, Schweitzer C, Werner K, Weitz J, Koch M, Bork P (2014) Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. Cancer Res 74:5690–5699

Viswanathan VS, Ryan MJ, Dhruv HD, Gill S, Eichhoff OM, Seashore-Ludlow B, Kaffenberger SD, Eaton JK, Shimada K, Aguirre AJ, Viswanathan SR, Chattopadhyay S, Tamayo P, Yang WS, Rees MG, Chen S, Boskovic ZV, Javaid S, Huang C, Wu X, Tseng YY, Roider EM, Gao D, Cleary JM, Wolpin BM, Mesirov JP, Haber DA, Engelman JA, Boehm JS, Kotz JD, Hon CS, Chen Y, Hahn WC, Levesque MP, Doench JG, Berens ME, Shamji AF, Clemons PA, Stockwell BR, Schreiber SL (2017) Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. Nature 547:453–457

Wang K, Sun J, Zhou S, Wan C, Qin S, Li C, He L, Yang L (2013) Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS Comput Biol 9:e1003315

Winter GE, Radic B, Mayor-Ruiz C, Blomen VA, Trefzer C, Kandasamy RK, Huber KVM, Gridling M, Chen D, Klampfl T, Kralovics R, Kubicek S, Fernandez-Capetillo O, Brummelkamp TR, Superti-Furga G (2014) The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. Nat Chem Biol 10:768–773

Xia X, Wong ST (2012) Concise review: a high-content screening approach to stem cell research and drug discovery. Stem Cells 30:1800–1807

Zhang J, Xing Z, Ma M, Wang N, Cai YD, Chen L, Xu X (2014) Gene ontology and KEGG enrichment analyses of genes related to age-related macular degeneration. Biomed Res Int 2014:450386