Briefings in Bioinformatics, 19(3), 2018, 506–523

doi: 10.1093/bib/bbw112
Advance Access Publication Date: 9 January 2017
Paper

A review of connectivity map and computational approaches in pharmacogenomics

Aliyu Musa, Laleh Soltan Ghoraie, Shu-Dong Zhang, Galina Glazko, Olli Yli-Harja, Matthias Dehmer, Benjamin Haibe-Kains and Frank Emmert-Streib

Corresponding author: Frank Emmert-Streib, Department of Signal Processing, Predictive Medicine and Data Analytics Laboratory, Tampere University of Technology, Korkeakoulunkatu 1, FI-33720 Tampere, Finland. Tel.: 00358 50301 5353; E-mail: frank.emmert-streib@tut.fi

Abstract

Large-scale perturbation databases, such as Connectivity Map (CMap) or Library of Integrated Network-based Cellular Signatures (LINCS), provide enormous opportunities for computational pharmacogenomics and drug design. A reason for this is that in contrast to classical pharmacology focusing at one target at a time, the transcriptomics profiles provided by CMap and LINCS open the door for systems biology approaches on the pathway and network level. In this article, we provide a review of recent developments in computational pharmacogenomics with respect to CMap and LINCS and related applications.

Key words: *pharmacogenomics*; *drug discovery*; *bioinformatics*; *drug repurposing*; *drug repositioning*; *big data*

Introduction

Recently, there is an increasing interest in the computational analysis of drug perturbation data sets. Such data types are now routinely used to aid our understanding in drug discovery and disease therapeutics [1, 2]. With the rapid accumulation of genomics and chemical informatics data in the past decade, several new systematic approaches to drug discovery have been proposed. For example, some study the drug–target structural relationships for specific drugs to discover new targets implicated in diseases, whereas others predict biochemical interactions of small molecules with their respective targets using, e.g. the Connectivity Map (CMap) approach [3–5]. However, for either type of investigations, machine learning [6] and biomedical text mining [7] approaches have been vital to uncover hidden relationships between drugs and potential new indications. Overall, applying these methods on drug perturbation data sets

Aliyu Musa is a PhD Student at Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology. His research focuses on 'Big Data' analysis for drug discovery and cancer therapeutics.
Laleh Soltan Ghoraie is a Postdoctoral Research Fellow at Princess Margaret Cancer Centre, University Health Network. She is interested in applications of Machine Learning in Bioinformatics.
Shu-Dong Zhang is a Senior Lecturer in Stratified Medicine (Statistics/Bioinformatics) at Northern Ireland Centre for Stratified Medicine, University of Ulster. His research focuses on the analysis of large-scale gene expression profiling data for drugs and diseases, and their applications in biomarker discovery for stratified medicine and drug repurposing.
Galina Glazko is assistant professor of Biostatistics and Computational Biology at Department of Biomedical Informatics, University of Arkansas for Medical Sciences. Her research focuses mainly on computational biology and biostatistics and their application in gene regulatory networks.
Olli Yli-Harja is Professor at Tampere University of Technology Department of Signal Processing. He has been involved in development of computational tools and software for systems biology using advanced methods of signal processing and statistics.
Matthias Dehmer is Professor at UMIT, Department for Biomedical Computer Science and Mechatronics. He is interested in Graph Theory, Data Science, Data Analysis, Big Data, Complex Networks and Machine Learning.
Benjamin Haibe-Kains is Scientist at the Princess Margaret Cancer Centre, University Health Network and Assistant Professor at the University of Toronto. His research focuses on the development and application of machine learning algorithms to analyze high-throughput genomic data in biomedicine, mostly in cancer studies.
Frank Emmert-Streib is Associate Professor in the Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology. His research interests lie in computational biology, predictive analytics and data science.

has proven to be beneficial in enhancing the understanding of the connection between genes, drugs and diseases [8–10] because such methodologies can lead to generation of novel hypotheses beyond classical pharmacology by translating new knowledge from genomic *in vitro* screens and cell-based assays to the patients.

Computational screening of drugs has been greatly facilitated by the advent of connectivity mapping methods, specifically CMap and the Library of Integrated Network-based Cellular Signatures (LINCS) [3, 11]. CMap and LINCS are comprehensive, large-scale drug perturbation databases containing transcriptomic profiles of dozens of cultivated cell lines treated with thousands of chemical compounds serving as reference databases. That means, these 'big data' resources provide simple yet important platforms to characterize 'signatures' of gene expression changes induced by small molecules. Such drug perturbation signatures have been used to determine connections, similarities or dissimilarities among diseases, drugs, genes and pathways, but we are far from fully understanding their capabilities.

The purpose of this article is to provide a state-of-the art survey of recent advances in CMap studies and related methods used in drug discovery, as well as reviewing computational tools that have been applied in the field. Furthermore, we discuss examples of applications of these methodologies being currently used both in drug repurposing/repositioning and in drug discovery process. An earlier review of connectivity mapping has been provided by Qu *et al.* [12], neglecting, however, methodological developments. A complementary presentation has been given in [13] focusing on publicly available resources and databases that can be used for generic genomic investigations of disorders.

Put simply, the goal of the CMap in genomic drug discovery studies is to identify disease or drug-associated gene signatures that correlate with perturbations on the transcriptomics level as a response to administrated small molecules or drugs [14]. It is a common approach used to identify inverse drug–disease relationships by comparing disease molecular features and drug molecular features, such as gene expression. This approach starts by generating a disease gene expression signature by comparing disease samples and normal tissue samples, followed by querying drug–gene expression reference databases. This makes the CMap technique effective and widely popular in drug discovery, posing a primary advantage, as it does not require a detailed mechanism of action (MoA) or prior knowledge of drug targets to work [15]. However, CMap comes with some limitations, such as limited drug perturbation data, a limited drug coverage, dosage-dependent conditions and the uncertainty of applying cell lines or animal model expression patterns to human systems. Also, the methodology can be expensive and time-consuming before it can generate a significant portion of all safe dosage conditions for a limited number of cell lines for CMap [12].

The connectivity mapping methods

CMap: the connectivity map

The connectivity map was introduced by Lamb *et al.* [3] in 2006. The basic concept of CMap is to use a reference database containing drug-specific gene expression profiles and compare it with a disease-specific gene signature. The CMap method is performed by simply submitting a list of genes thought to be relevant to a particular disease. A researcher is returned a list of drugs having either presumptive efficacy for the disease or, more realistically, whole mechanisms of action that are well known, thereby enhancing biological understanding of the disease. This allows identifying connections between drugs, genes and diseases. The overall goal of CMap is to predict potentially therapeutic drug candidates.

The principal workflow of CMap is shown in Figure 1. A phenotype of interest such as a disease or biological condition is described by a gene expression signature, i.e. a set of genes that uniquely represents the underlying phenotype. In [3], the gene signature corresponds to a list of differentially expressed genes (DEG), named h, that contains up- and downregulated genes as shown Figure 1A.

The gene signature set is then used to query the CMap catalog of gene expression profiles. The CMap database is a collection of paired gene expression profiles representing a series of structured microarray experiments. All experiments were conducted using a microarray platform (Affymatrix HT_HG_U133A array with 22 283 probesets in addition to HG_U133A with 22 277 probesets) and standardized preprocessing (MAS 5.0). The experiments were carried out in various cell lines to perturbagens (drugs and bioactive small molecules) at varying concentrations and time points against vehicle controls. The initial database (Build 1) contained 455 instances, i.e. treatment-control pairs, where treatment constitutes a selection of 165 drugs, 42 different concentrations, 2 time points and 5 cell lines. The updated version (Build 2) contains 6100 instances with more drugs (1309) and concentration (156) but the same cell lines, for a parallel series of analysis. The instance is the basic unit of data and metadata in CMap. Each instance is uniquely identified by an instance identifier. After preprocessing, the resulting probe-level summaries are subject to further analysis (scaling treatment values to corresponding vehicle controls, thresholding, etc.). The fold change of treatment to control values was calculated for each probeset, sorted into decreasing order and converted to a rank vector, separately for each instance. Thus, the probeset that is most upregulated will receive Rank 1 and the most downregulated will receive 22 283. So, for Build 2, the CMap database is $n = 22,283 \times p = 6100$ matrix. The instance rankings are used to compare query lists. It is important to note that while these rankings may be perceived as a crude form of summarization, the absence or sparsity of treatment replication precludes usage of summaries incorporating variation. Hence, for every drug, there is an instance representation in the reference database, corresponding to the treatment and the control condition.

The gene signature, h, is compared with the ranked probesets of the treatment versus control gene expression profiles that are ranked in descending order according to the fold changes of the probesets. By splitting the gene signature, h, into two lists containing only upregulated genes, $h\uparrow$, and downregulated genes, $h\downarrow$, a so-called connectivity score is estimated via several auxiliary variables using a nonparametric rank-ordered Kolmogorov–Smirnov (KS) test, similar to the method introduced in [16].

The resultant 'connectivity score' is normalized using random permutation described in [3] by Lamb *et. al.*, assuming values from -1 to $+1$ to reflect the closeness or connection between the expression profiles. A positive connectivity score is obtained for having most of the downregulated genes at the top of the reference profile and most of the upregulated genes at the bottom (Figure 1B). In contrast, a negative connectivity score is obtained for a reversed mapping, meaning that most of the upregulated genes are at the bottom of the reference profile and most of the downregulated genes are at the top [17]. A positive

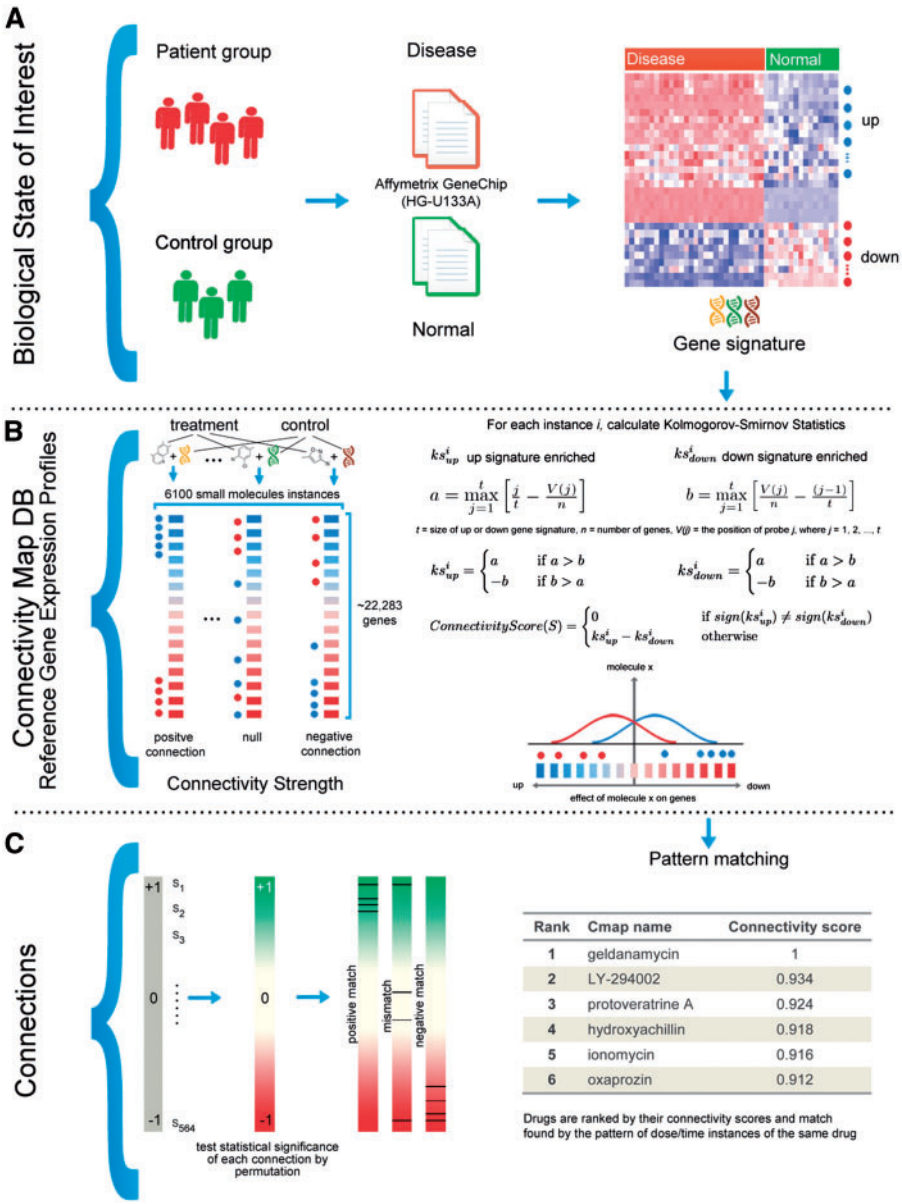

Figure 1. Mechanistic overview of the working principle of the CMap method and the CMap database for drug discovery.

correlation denotes the degree of similarity and a negative correlation emphasizes an inverse similarity between a query signature and a reference profile derived from an individual chemical perturbation; thus, implicating the exposure to a particular chemical can mimic or reverse the expression pattern of the biological state of interest. A null connectivity score occurs when the up- and downregulated genes are randomly distributed over the reference profile. See Figure 1B for a visualization of the different cases. Overall, the results are obtained as a list of connectivity scores for all small molecules in the reference database, one connectivity score for each small molecule. Finally, the top-scoring drugs are selected by sorting all connectivity scores in descending order and identifying a relevance threshold (Figure 1C). Unfortunately, in [3], no measure of statistical significance, via a statistical hypothesis test, has been used formally. In contrast, only a basic approach has been suggested involving a resampling procedure.

Since the first introduction of the CMap principle and methodology, there have been numerous applications of this approach by many research groups with a particular focus in drug discovery and development. Therefore, the CMap approach can be used as a method of screening chemicals by matching the gene signature of a novel pertubagen against the reference profile [18, 19]. The chemicals sharing similar gene expression pattern, similar activities or mechanisms can be retrieved. A highly representative phenotype-specific gene signature set of a given biological state; pathological, genomic perturbations or induced by chemicals is seen as the first step of implementing CMap technique. The signature can be generated through a computational analysis using the genome-wide gene expression profiles. Although there is no precise way of creating optimal gene signatures, the conventional approach is to identify and use the DEG that are statistically significant displaying an association with a given phenotype.

Reference drug perturbation databases and data sets

There are a few valuable databases and data sets containing gene expression response profiles effected by chemical compounds that are publicly available. Hence, these data provide information about the perturbation effects that drugs have on the transcriptomics level of a cell. In Table 1, we provide an overview of the most important generic resources. However, we would like to note that there are additional disease-specific resources available, e.g. for cancer [20], that provide also disease-relevant relationships with drug compounds and targets. Henceforth, we focus on the two largest general purpose drug perturbation data sets CMap and LINCS L1000.

Table 1. An overview of generic drug perturbation databases and data sets

Database/ data set	Description	URL link
CMap [3]	A database of genome-wide gene expression profiles produced on treatment of 564 gene expression profiles generated for five cancer cell lines (Build 1). The current version consists of 1309 compounds and $\sim 7,000$ gene expression profiles (Build 2).	https://www.broadinstitute.org/CMap/
LINCS L1000 [11]	The Library of Integrated Cellular Signatures (LINCS) is an NIH program, which funds the generation of perturbation profiles across multiple cell and perturbation types, as well as readouts, at a massive scale. The data consist of $\sim$20000 perturbagens, $\sim$15 cell lines, $\sim 1,400,000$ gene expression profiles and 25 assays.	http://www.lincsproject.org/
DP14 and DP92 [21]	The DP14 data set contains GEPs of OCI-LY3 cell line (a human diffuse large B-cell lymphoma cell line) treated with 14 distinct individual compounds and profiled at 6, 12 and 24 h following compound treatment, all in triplicate. For treatment, two different concentrations of the compounds corresponding to IC20 at 24 h and IC20 at 48 h were used. GEP of DMSO-treated samples and profiled at the three different time points, all in octuplicate were used as control, resulting in 276 GEPs from this data set. DP92 data set contains GEPs of 92 distinct FDA-approved, late-stage experimental and tool compounds in three different B-cell lymphoma cell lines (OCI-LY3, OCI-LY7 and U-2932), profiled at 6, 12 and 24 h following compound treatment. All compounds were treated using IC20 at 24 h concentration. DMSO was used as control media at each of the three time points, resulting in 857 GEPs.	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60408
GEODB [21]	This data set contains GEP of 13 different compounds, obtained from nine independent expression sets obtained from the Gene Expression Omnibus (GEO). Each expression set had at least six DMSO controls and six samples for compound treatment. Three of the expression sets were profiled on MCF7 breast cancer cell lines (GSE9936—three compounds, GSE5149 and GSE28662—two compounds), and two on MDA-MB-231 metastatic breast cancer lines (GSE33552—two compounds). The rest of the expression sets were profiled in a B-cell lymphoma cell lines, which are chronic lymphocytic leukemia patient-derived cell lines(GSE14973), K422 non-Hodgkin's lymphoma cell lines (GSE7292), lytic-permissive lymphoblastoid cell lines (GSE31447), diffuse large B-cell lymphoma patient-derived cell lines (GSE40003) and mantle cell lymphoma cell lines (GSE34602).	http://www.ncbi.nlm.nih.gov/geo/
Follicular lymphoma [22]	CB33, SUDHL4 and SUDHL6 cells provided by R. Dalla-Favera (Columbia University, NY) were maintained in IMDM (Life Technology), supplemented with 10% FBS (Gemini) and antibiotics. The HF1 follicular cell line provided by R. Levy (Stanford University, CA) was maintained in DMEM (Life Technology), supplemented with 10% FBS and antibiotics. Cells were tested negative for mycoplasma. Cells were not further authenticated. Antibodies: rabbit anti-MYC (XP) (Cell Signaling Technology); rabbit anti-FOXM1 and mouse anti-GAPDH (SantaCruz); rabbit anti-HMGA1, anti-ATF5, anti-NFYB, mouse anti-TFDP1 (Abcam), Alprostadil, Clemastine, Cytarabine and Troglitazone (Tocris), Econazole nitrate and Promazine hydrochloride (Sigma) were reconstituted in DMSO (Sigma).	http://cancerres.aacrjournals.org/content/early/2015/11/20/0008-5472.CAN-15-0828.abstract
RAF-inhibitor resistant [23]	The data set consists of 143 proteomic/phenotypic entities under 89 perturbation conditions. In perturbation experiments, the drugs are applied to cell cultures after SkMel-133 cells are grown to about 40% confluence in complete RPMI-1640 medium (10% heat-inactivated fetal bovine serum, 100 units/ml each of penicillin and streptomycin and incubated at 37°C in 5% CO_2) in six-well plates. After 24 h drug administration, the perturbed cells are harvested. In control experiments (i.e. no drug condition), cells are treated with the DMSO drug vehicle for 24 h.	http://elifesciences.org/content/4/e04640v1

CMap

The CMap database consists of genome-wide transcriptional expression profiles of bioactive compounds from cultured cell lines. In the original CMap study [3], the reference database consisted of 564 gene expression profiles generated from exposing five different human cell lines (MCF7, PC3, SKMEL5, HL60 and ssMCF7) with 164 small molecules [3] (Build 1). In Build 2, this has been significantly extended to 1309 approved small molecules applied to the same five human cell lines leading to over 7000 gene expression profiles. Build 1 and Build 2 use an Affymetrix platform for generating the gene expression data. So far, several methods have been developed using the CMap database (either Build 1 or Build 2), either for new drug repositioning/repurposing approaches or for improving the performance of the original CMap method, also in comparison with other data sets [24–27]. Notably, Cheng *et al.* [28] presented a systematic approach to quantitatively assess the performance of such methods. Hence, this study can be seen as a benchmark approach to assess any new methodology in the future.

LINCS L1000

The LINCS supported by the NIH, comprises ∼5000 genetic perturbagens (e.g. single-gene knockdowns or overexpressions) and ∼15000 perturbagens induced by chemical compounds (e.g. drugs) [29]. To date, over one million gene expressions have been profiled and collected for this project using the L1000 technology [29]. The L1000 platform has been developed at the Broad Institute by the CMap team to facilitate rapid, flexible and high-throughput gene expression profiling at lower costs. Specifically, the L1000 technology measures the expression of only 978 so-called landmark genes, and the expression values for the remaining genes are estimated by a computational model using additional data from the Gene Expression Omnibus (GEO) [30]. A user-friendly access to the database is provided by the LINCS cloud Web page (http://www.lincscloud.org/l1000/), which is a Web-based application allowing users to browse and query the LINCS database.

In a simplified view, the L1000 data can be considered as a 'big matrix' where the rows correspond to 22 268 genes and the columns are the millions of perturbations induced by the small molecules. It is clear that such a large data set presents new challenges to computational systems biologists who aim to analyze and visualize Big Data. In Table 2, we provide a brief overview of tools and software developed so far to explore and understand the L1000 database.

CMap variations and extensions

ssCMap: statistically significant connectivity map

New methods of pattern matching algorithm and data normalization were applied using CMap approach to help reduce noise effects, results interpretation and strengthen the methods reliability in generating unproven hypotheses [26]. For example, an important method has been introduced by Zhang *et al.* [33],

Table 2. Tools and softwares developed for browsing, visualizing and querying the LINCS database

Name	Description	Features	URL link
Enrichr [31]	Enrichr is an easy-to-use intuitive enrichment analysis Web-based tool providing various types of visualization summaries of collective functions of gene lists.	Access, Search, Navigation, Integration, Visualization and Signature Enrichment	http://amp.pharm.mssm.edu/Enrichr
LINCS Data Portal	The current version of the portal has features for searching and exploring LINCS database.	Access, Search, Browse and Navigation	http://lincsportal.ccs.miami.edu/dcic-portal
Slicr	Slicr (LINCS L1000 Slicer GSE70138 data only) is a metadata search engine that searches for LINCS L1000 gene expression profiles and signatures matching users input parameters.	Access, Search, Navigation, Integration, Visualization and Signature Enrichment	http://amp.pharm.mssm.edu/Slicr
L1000CDS^2 [32]	L1000CDS^2 queries gene expression signatures against the LINCS L1000 to identify and prioritize small molecules that can reverse or mimic the observed input expression pattern.	Access, Search, Navigation, Integration, Visualization and Signature Enrichment	http://amp.pharm.mssm.edu/L1000CDS2
LIFE	A semantically enhanced Web-based application that enables access, navigation and exploration of a knowledge base built by integrating and indexing all the LINCS data types. LIFE allows access, navigation and exploration of LINCS assays, biomolecules, related concepts and LINCS screening results via a variety of views such as proteins, genes, cell lines, small molecules. LIFE provides flexible navigation of the LINCS assay and data landscape via list functionality covering important assay biomolecules and concepts; this enables a variety of use cases.	Access, Query, Search, Browse, Navigation and Download	http://life.ccs.miami.edu/life
iLINCS	iLINCS is a portal that handles LINCS L1000 and KinomeScan data. It facilitates integration of LINCS data-derived signatures with other genome-scale signatures.	Access, Search, Navigation, Leverage Ontology, Visualization and Download	http://life.ccs.miami.edu/life
LINCS Canvas Browser [29]	Compact visualization of thousands of L1000 experiments; clustering of perturbations based on signature similarity; interactive gene list enrichment analysis using 32 gene set libraries; query up- and downregulated gene lists against over 140 000 L1000 conditions.	Access, Search, Navigation, Integration, Visualization and Signature Enrichment	http://www.maayanlab.net/LINCS/LCB

called statistically significant connectivity map (ssCMap). The approach uses connectivity score computation with permutation tests at both treatment instance level and treatment set level that offers a statistical means to control over the possible false connections between the gene signature and the reference profiles. Because the CMap concept uses the entire genomic information of the patients and of the drug, one may view this approach as an attempt at systems treatment. However, it suffers from having many draw backs as mentioned in [33]. In particular, it has no specific reference to the biological functions altered by the disease in question. A top-ranked drug could be misleading for having strong effects on a subset of functions at the expense of altering other functions that are not associated with the disease [34].

The ssCMap method introduces a new ranking score using the following steps. First, treatment and control instances are treated similarly, making the effect of the treatment instances to be determined by DEG. Second, the genes that are affected by the treatment instance, that is, genes that are highly differentially expressed, are given more weight. Finally, the up- and downregulated genes are handled equally, in such a way that 2-fold of the up- or downregulation of a gene has the same relevance in constructing the reference profile. The genes are ordered using the absolute value of their log expression ratios (fold change), as the up- and downregulated genes are considered the same. Moreover, the most significant gene will be at the top of the list, while most of the insignificant gene will be at the bottom. This ensures that the genes are ranked by their importance in the reference profile [33]. Assuming there are in total N genes, the first gene in the list will be assigned a rank N if it is upregulated, or a rank $-N$ if it is downregulated. In general, the ith gene in the list will be ranked with $(N - \mathrm{i} + 1)$ for upregulation or $-(N - \mathrm{i} + 1)$ for downregulation. The ssCMap uses new scoring scheme for representing a query gene signature either with ordered or unordered gene list. The important gene expressed will be assigned a rank m or $-m$ depending on whether it is up- or downregulated, where m is the number of genes in the gene signature. The connection strength [33] is calculated between reference profile R and gene signature s to measure a connection between reference profile and gene signature.

$$C(R, \mathrm{s}) = \sum_{\mathrm{i}=1}^{m} R(g_{\mathrm{i}})\mathrm{s}(g_{\mathrm{i}}). \tag{1}$$

Where g_{i} represents the ith gene in the signature, $s(g_{\mathrm{i}})$ is its signed rank in the signature and $R(g_{\mathrm{i}})$ is this gene's signed rank in the reference profile (Equation 1). To have maximum connection between reference profile and gene signature, Zhang *et al.* achieved it by matching m genes and their regulation status in the reference profile and the gene signature in the correct order (for ordered gene signature) as shown in Equation 2. For an unordered gene signature, all the genes in the list have equal weight because there is no particular ordering; therefore, maximum connection strength for unordered is calculated using Equation 3.

$$C^{0}_{max}(N, m) = \sum_{\mathrm{i}=1}^{m} (N - \mathrm{i} + 1)(m - \mathrm{i} + 1). \tag{2}$$

$$C^{u}_{max}(N, m) = \sum_{\mathrm{i}=1}^{m} (N - \mathrm{i} + 1). \tag{3}$$

The overall connectivity score (c) is calculated by dividing the connection strength with the maximum connection strength of a given gene signature and reference profile Equation 4. The connectivity score ranges from -1 to 1, where 1 indicates a maximum positive connection of gene signature with the reference profile, while -1 indicates a negative connection. To test the connection score, ssCMap uses a simple procedure to test the null hypothesis between the gene signature and the reference profile that is achieved by generating a random gene signature of ordered/unordered list using random selection without replacement with equal probability of either up- or downregulation. After generating the signature, ssCMap calculates the connectivity score (c) of each signature as well as the P-value associated with the connectivity score denoted by P. Here, $\tilde{c}$ is the connectivity score between a random gene signature and a reference profile. The same procedure is repeated to estimate the sampling distribution of the random signatures. Zhang *et al.* provide a user-friendly software application for the ssCMap algorithm [35].

$$ConnectivityScore(c) = C(R, \mathrm{s})/C_{max}(N, m). \tag{4}$$

CMapBatch: a meta-analysis of drug response

Fortney *et al.* [27] have recently adapted a parallel CMap approach across multiple gene signatures of a disease, and named the method 'CMapBatch'. Specifically, instead of applying CMap to one individual gene signature, the authors apply it to multiple gene signatures for the same disease and then combine the resulting outcomes. Therefore, their approach is similar to a meta-analysis. It is common for a complex disease to have more than one signature available, and this justifies the application of CMap to multiple gene signatures of a disease. Previously, other groups [36, 37] addressed this issue by combining those different gene signatures before applying CMap [35]. However, Fortney *et al.* emphasize that combining gene signatures is problematic for strongly nonoverlapping gene sets. This problem has been addressed by CMapBatch.

Formally, for each disease signature, CMapBatch obtains a list of connectivity scores corresponding to all the small molecules (1309 in CMap Build 2) and combines them by using the Rank Product method [38] to assign a consensus ranking on each drug for all the tested gene signatures. The Rank Product method was originally developed to identify DEG for replicated experiments based on the ranking of the individual experiments. Fortney *et al.* analyzed 21 signatures ($s = 21$) for lung cancer obtained from Oncomine [27, 39]. The results reveal that CMapBatch produces indeed a more stable list of drugs when compared with the individual gene signatures. Specifically, the median overlap of the top 50 drugs for 21 individual gene signatures was 22, but for CMapBatch, the overlap was 39 drugs. Furthermore, for a FDR threshold value of 0.01, 247 small molecules have been identified that significantly reverse the gene expression changes of the tested signatures.

The method was used to further highlight more effective drug candidates inhibiting cancer growth and the results compare favorably with the results of the original CMap. Thus, scaling up transcriptional knowledge increases the hit percentage significantly from 44 to 78% of the top-ranked drugs. Moreover, the resultant drug hits were characterized *in silico* and showed slow growth significantly in nine lung cancer cell lines from the NCI-60 collection [27]. In total, 247 candidate therapeutics were identified for which two genes, CALM1 and PLA2G4A, are found to be markers for drug targets in lung cancer [40].

Despite the fact that CMapBatch was only tested for lung cancer, the proposed meta-analysis can be used for any disease phenotype to prioritize therapeutics.

Extensions of the CMap similarity metric

The CMap ability of finding connections and similarities between genes, diseases and drugs makes it useful in many applications but has a few draw backs. One of these is failure to apply a comprehensive measure to validate the significance of a gene signature when queried against reference profiles [33]. Several studies have focused on improving the original KS statistics used as the 'similarity metric' by CMap. We highlight some of these methodologies in Table 3.

High-performance computing platforms in CMap

As a computational and bioinformatics framework, connectivity mapping has been underpinned by the powers of modern computers. Throughout the development of connectivity mapping, particularly CMap and its extensions, intensive permutation tests are required to provide statistical rigor, and the ever-growing expansion of the reference database has required faster processing and/or better software architectures to fulfill such requirements.

To address these issues related to the computational demands, Zhang and his group developed high-performance computing (HPC) models of connectivity mapping, called cudaMap [45], which uses the computing power offered by the graphics processing units (GPUs) of modern computers; a recent extension is QUADrATiC [46], which is a scalable gene expression connectivity mapping framework for repurposing Food and Drug Administration (FDA)-approved drugs. The framework uses multiple processor cores to achieve high-speed connectivity mapping. Furthermore, concerted efforts have also been made to formulate and standardize the procedures for creating quality gene signatures across multiple data sets [47] and determining the optimal lengths of query gene signatures [48].

Computational evaluation of CMap methods

Transcriptional expression profiles are widely used to find drug–disease or drug–drug relationships that could lead to new methods in drug discovery [28]. However, a remaining challenge is to evaluate methods based on such data sets. Despite the success of various CMap approaches, there are few ways to quantitatively evaluate the performance of the connectivity score for the association between drugs and diseases by computational means. There are two ways to computationally evaluate CMap: first, evaluate drug–drug relations [18, 42] and second, evaluate disease–drug relations [28].

In evaluating drug–drug relationship, a drug signature is used to query CMap to retrieve related drugs that have the same ATC codes or chemical structures that are similar as studied in [18, 42]. However, in evaluating disease–drug relations, a disease signature is used to query CMap to retrieve known drugs notably in [28].

Iskar *et al.* [18] were among the first to study a quantitative evaluation of CMap methods to identify similar compounds using an ATC classification. They created labeled benchmark sets using compound chemical similarities and ATC codes. They focused on early retrieval performance where the false-positive rate (FPR) is <0.1. At these FPRs, their calculated AUCs were significantly different from random.

Cheng *et al.* [42] also used the ATC codes to benchmark the similarity metrics using two different methods: the batch DMSO control and mean-centering normalization. Focusing on early retrieval performance (FPR = 0.1), eXtreme cosine (XCos) method outperforms the original CMap similarity metric based on KS test. It is also robust in terms of drug–drug relationship prediction with compounds that have higher treatment effect on treated cell lines. Therefore, the authors further extended the method for evaluating various CMap similarity metrics with compound profiles that have higher treatment effect.

However, not all performance evaluations tend to work as pointed out by [49] because of the following reasons: First, a lack of high-quality disease signatures, as many diseases may not be represented accurately by the reference profiles in the gene signature. Second, the benchmark sets used to represent the drug–disease association might not be comprehensive enough to capture all drug–disease linkages. Finally, the drug cellular profiles are limited to only treating fewer cell lines, which explains why some of the neoplastic disease signatures perform better than nonneoplastic disease signatures [28].

Applications of CMap in pharmacogenomics

Since the introduction of Build 1 in 2006, the CMap database and the CMap method have been applied in a large number of pharmacogenomics studies. These studies can be categorized with respect to their application purpose. Specifically, CMap has been used to identify novel phenotypic relations for disease treatment, for drug repurposing/repositioning and for studying drug combinations [50].

Discovering novel phenotypic relations

The most fundamental but also the most difficult task for which the CMap database can be used is to identify a novel therapeutic treatment for a disease [5]. This is also called a lead discovery. It aims at establishing an advantageous connection between the administration of a drug and a phenotypic response of the patient. Several studies used a CMap analysis to improve the understanding of disease/phenotype associations by combining some of the therapeutic agents identified in cancer [51–53]. These studies have shown the full potential of the application of CMap in drug discovery and in identifying cancer disease therapeutic targets. Table 4 provides a list of applications in finding drug targets or pathways and their associations with a disease.

As an example, McArt *et al.* [60] used the ssCMap to find connections for small molecule candidates that can be used for a phenotypic analysis in the laboratory [35]. Specifically, their study used a DNA microarray and RNA sequencing platform, and they identified the same gene signature for which the resulting drug (cotinine) suppressed androgen-driven cell proliferation [61]. Furthermore, they experimentally validated cotinine, which inhibits proliferation in LNCaP cells [60].

Recently, a study conducted by Lim *et al.* [53] used a gastric cancer gene signature to query CMap. The results of their analysis showed that histone deacetylase inhibitors (HDAC), which include vorinostat and trichostatin A, were potential drug candidates for treating gastric cancer [53]. These findings were experimentally validated *in vitro* using gastric cancer cell lines, where vorinostat significantly inhibited cell viability in a dose-dependent manner [53].

Spijkers-Hagelstein *et al.* used CMap to demonstrate a reverse effect of PI3K inhibitors in infants with MLL-rearranged acute lymphoblastic leukemia (ALL). The study found the PI3K inhibitor LY294002 to be significantly effective in reversing prednisolone-resistance profile and induce sensitivity [51, 62]. Moreover, the prednisolone-sensitizing effects of LY294002 on two cell lines studied consist of five downregulated genes, namely PARVB,

Table 3. List of methodologies that extend the CMap similarity metric

Method name	Description	Advantage	Disadvantage
ProbCMap: Probabilistic drug connectivity mapping [41]	A probabilistic connectivity mapping by [41] was introduced as a model-based alternative to the original CMap. The method uses a probabilistic model that focuses on the relevant gene expression effects of a drug as a probabilistic latent factor derived from the data on cell lines.	• Finding functionally and chemically similar drugs based on transcriptional response profiles. • It has been shown that gene expression response factors between cell lines can be promising when a multisource probabilistic model is used. • The method allows retrieval of a combination of drugs. • It also shows how drug combination retrieval provides complementary information when compared with a single-drug retrieval.	• It is more sensitive to platform differences. • The method intentionally ignores possible cell line-specific effects of the drugs. • The approach relies on the assumption that it is suitably chosen based on the probabilistic model.
Connectivity score based on partial-rank metrics [26]	This extension of the connectivity score was introduced by Segal *et al.* [26]. They apply partial-rank metrics and empirical null distributions for scoring CMap queries by accommodating a query order, in contrast to the KS scoring, which uses a rank ordering of gene expression profiles in the target instance to generate an ordering of the query.	• More effective methods than KS by computing a per experiment score that measures 'closeness' between the signature and the reference profiles. • New approaches measuring closeness for the common scenario wherein the query constitutes an ordered list. • Advance an alternate inferential approach based on generating empirical null distributions that characterize the scope, and capture dependencies, embodied by the database.	• Hard to develop effective fitting algorithms for large instances. • Number of inferential problems surrounding use of metrics extended to partial rankings, such as reconciling asymptotic distributions.
XCos: Cosine-based similarity [42]	The xCosine is introduced as alternative method used to computationally evaluate the similarity between reference profile and gene signature. In this novel CMap approach, Cheng *et al.* used the Anatomical Therapeutic Chemical (ATC) classification as the benchmark to measure differences and similarities of XCos method to other CMap scoring methods, data processing methods and signature sizes [42].	• XCos outperforms CMap when used with a larger number of features (top 500). • Help find the analytical approaches that are more accurate in evaluating the CMap data. • Finds good transcriptional response to drug treatment that appears to have sufficient consistency in MoA. • The method is used to determine the compound classes, which have robust expression profiles in the CMap data. • It emphasizes early retrieval, which is important because in repositioning the aim is to sacrifice some true positives to keep false positives low.	• Multiple ATC codes per compound can lead to errors, and redundant ATC codes may inflate results. • Many ATC codes do not properly characterize MoAs. • Averaging over multiple cell lines averages biological variation for compounds that may have differential responses in the multiple cell lines.
XSum: Systematic evaluation of connectivity map [28]	This method uses a similarity metric that systematically evaluates multiple CMap methodologies by assessing their performance on many drug profiles across a curated data set consisting of multiple disease gene signatures [28].	• Using XSum, CMap can significantly enrich true positive drug-indication pairs by a novel matching algorithm. • It can be used as an effective similarity measure to enhance the KS statistics as well as filtering drug-induced data. • The algorithm has a relative early retrieval performance. • It can help tremendously in experimental validation using small number of hypotheses. • The overall retrieval performance is weak. • The drug–disease benchmark standard was not able to capture all known drug–disease association.	

(continued)

Table 3. Continued

Method name	Description	Advantage	Disadvantage
		• As the CMap performance is not optimized, that process is prone to be overfitting and bias.	
Module-based chemical function similarity search [43]	This approach evaluates CMap (Build 1) data set using expression pattern comparison-based chemical function similarity search, seen as an improvement of CMap that can provide more biological information of the chemicals.	• Module-based expression pattern comparison provides a possibility to identify functional modules or pathways with two similar profiles. • It can help in finding chemicals that are functionally alike because they affect similar pathways or biological processes. • Uses GO [44] modules to reduce feature selection.	• It is limited to GO system to define gene set. • When searching for related profiles for a given chemical, both module based and CMap give similar rankings, especially when two target chemicals have close ranks.

D123, FCGR1B, PSTPIP2 and S100A2. Interestingly, the mentioned genes appear to be expressed in children with ALL samples with prednisolone-resistant, but not in ALL samples with prednisolone-sensitive samples.

Another interesting study from Engerud *et al.* [25] found by applying CMap that HSF1 and HSF1-related gene signatures are correlated with a high-risk disease state in endometrial cancer, and they also shed light on the underlying biological mechanisms. The results showed how HSF1 levels can predict a response to drugs targeting HSP90 or any possible protein synthesis. Furthermore, their results also justified that the HSF1 level and HSF1-related signatures impact on carcinogenesis during disease progression and found that HSF1 can be used for developing new therapeutic targets [17]. Therefore, HSP90 inhibitors are seen as novel targeted therapeutics for patients with high HSF1 levels in tumors [25, 63].

In addition, a similar approach of CMap application has been used to investigate relationships between drugs and microRNAs (miRNAs) [64]. Jiang *et al.* proposed a novel high-throughput approach to identify the biological links between drugs and miRNAs in 23 different cancers and constructed the Small Molecule-MiRNA Network for each cancer to systematically analyze the properties of their associations. They concluded that most of the miRNA modules comprised miRNAs that had similar target genes and functions or were members of the same miRNA family. The majority of the drug modules involved compounds with similar chemical structures, modes of action or drug interactions. Another common approach is to identify drug–miRNA relationships by comparing disease molecular features and drug molecular features, such as gene expression. Wang *et al.* [65] proposed a novel computational approach to identify associations between small molecules and miRNAs based on functional similarity of DEG. The results show 2265 associations between FDA-approved drugs and diseases, where 35% of the associations have been reported in the literature. Also, 19 potential drugs were identified for breast cancer, in which 12 drugs were reported by previous studies. Their studies provide a valuable perspective for repurposing drugs and predicting novel drug targets, which may provide new way for miRNA-targeted therapy [65].

Duan *et al.* introduced an improved computational method that potentially shows the importance of using the newly generated publicly available LINCS L1000 data set to rapidly prioritize small molecules that could reverse or mimic expression in disease and other biological states. The DEG of these profiles were calculated using the characteristic direction method [66]. The L1000CDS^2 uses the users' input of either a gene-set method or cosine distance method to compare the input signatures with the L1000 signatures to perform the search via a state-of-the-art Web interface. The L1000CDS2 method provides prioritization of thousands of small-molecule signatures, and their pairwise combinations, predicted to either mimic or reverse an input signature. It also predicts drug targets for all the small molecules profiled using L1000 assay. To further showcase the usefulness of the approach, they collected expression signatures from human cells infected with Ebola virus at 30, 60 and 120 time points. Querying these signatures against L1000CDS^2, kenpaullone compound was identified. A GSK3B/CDK2 inhibitor has shown a dose-dependent efficacy in inhibiting Ebola infection *in vitro* without causing cellular toxicity in human cell lines [67].

Using the CMap approach, Zhu *et al.* found vorinostat as a possible candidate therapeutic drug in gastric cancer. The HDAC inhibitor (e.g. vorinostat and trichostatin A) has an inverse correlation with a gastric gene signature, which shows an interesting therapeutic target. Studies have already revealed the efficacy of vorinostat as therapeutic drug that suppresses growth of various cancer cell lines [68]. Moreover, many analysis of cancer-related cell lines and gastric cancer patients showed vorinostat to be effective in altering expression levels, hence making it effective for the upregulation of autophagy-specific genes [69, 70].

Siu *et al.* [71] highlighted the potential benefits of polyphyllin D as a therapeutic drug for non-small cell lung cancer (NSCLC). Interestingly, the extracts of the *Paris polyphylla* plant, containing polyphyllin D, have been long used in traditional Chinese medicine for cancer treatment [72]. Their CMap analysis indicated that polyphyllin D is a trigger for estrogen receptor-induced apoptosis and mitochondria-mediated apoptotic pathways [73].

CMap-based elucidation of drug MoA

In pharmacology, understanding the exact effect of an active compound on a system represented, e.g. by a gene signature, is the central focus. Specifically, it is important to identify possible new compounds that are performing activities based on particular targets [12]. Given a compound phenotypic gene signature, the CMap method [3] can be applied to identify such novel active compounds. Thus, it provides a new hypothesis-generating tool to identify signaling pathways affected by a compound, connecting a biological state to the discovery of

Table 4. An overview of the application of CMap for a number of different diseases

Disease	Method	Data set	Result	Drug	Reference
CNS injuries	CMap tool	Human MCF7 breast adenocarcinoma (GSE34331)	The findings show the hypothesis that inhibition of calmodulin signaling might allow neurons to alleviate substrate derived neurite growth restriction and CNS regeneration.	Calmodulin and piperazine phenothiazine (repurposed)	[54]
GBM	Pathway analysis and CMap tool	GBM data sets (GSE4290, GSE7696, GSE14805, GSE15824 and GSE16011)	Investigated antitumor drugs in GBM cell lines and identify novel drugs that can suppress GBM tumors.	Thioridazine	[55]
Gaucher disease (GD1)	Pathway analysis and CMap tool	GD1 mouse (GSE2308)	Predicted highly enriched anti-helminthic compounds for new drug action on GD1 and repurposing.	Albendazole and oxamniquine	[52]
Ovarian cancer	CMap tool	MCF7 and PC3 cell lines (GSE5258)	Found a compound as PI3K/AKT pathway inhibitor that shows the mechanism of cancer therapeutics.	Thioridazine	[56]
Stem cell leukemia (SCL)	GSEA and CMap tool	hESCs cell lines (GSE54508)	Found two HDAC inhibitors as potential inducers that can be used in treating SCL and acute megakaryoblastic leukemias.	Trichostatin A and suberoylanilide hydroxamic acid	[57]
T-cell acute lymphoblastic leukemia (T-ALL)	GSEA and CMap tool	Human and mouse T-ALL cell lines (GSE12948, GSE8416 and GSE14618)	Identified interconnecting regulatory pathways as therapeutic targets for T-ALL.	HDAC, PI3K and HSP90 inhibitors	[51]
Prostate cancer	CMap tool	Celastrol- and gedunin-treated cell lines (GSE5505 and GSE5508)	Identified target pathways of androgen receptor (AR) signaling and modulation of HSP90 MoA.	Celastrol and gedunin	[17]
Gastric cancer	Hierarchical clustering and CMap tool	Yonsei gastric cancer (GSE13861)	Predicted two possible drug candidates for gastric cancer therapy.	Vorinostat and trichostatin A	[53]
Myelomatosis	CMap tool	Human myeloma cell lines (GSE14011)	Found a drug with potential to induce suppression of cyclin D2 promoter regulation.	Pristimerin	[58]
AML	CMap tool	AML data (GSE7538)	Predicted novel treatment of human primary AML with parthenolide and transcriptional response of cells.	Celastrol	[59]

disease–gene–drug connections, depending on the level of observed changes, i.e. the molecular or functional (anatomical) level.

Availability of computational approaches has sparked usability of network models and system biology approaches to obtain a deeper understanding of the basic biological drug–disease relations [57]. Specifically, methods have been developed to aid in finding drugable targets and drug compounds based on a basic understanding of biological processes in the pathway level. These include methods such as integrating a functional protein association network to form a new model, finding information on a known target and enriched pathways, small molecules with high connectivity score, investigating side-effect scores based on ranked gene signatures and the use of novel methods from machine learning to evaluate CMap data set [74–77].

There are also many other functional phenotype-based approaches that use the CMap resource to understand MoA [7, 78–80]. It is widely known that many drugs with therapeutic targets in cancer prognosis and diagnosis have been identified using CMap. For example, CMap designated the mTOR inhibitor rapamycin as a potential therapy for dexamethasone-resistant ALL in children. A clinical trial is currently underway for assessing this possible new indication [81]. A similar approach by Li *et al.* has shown its power in discovering chemicals sharing similar biological mechanisms and chemicals reversing disease states. They used CMap and gene ontology (GO) [44] modules to partition genes into small biological categories and performed expression pattern comparison within each category [43]. The method shows robustness in finding chemicals sharing similar biological effects by using a reduced similarity matrix to measure the biological distances between query and reference profiles. This will pave in reducing experimental noises and marginal effects and directly correlates chemical molecules with gene functions.

Iorio *et al.* [4] generated a drug network (DN) from the CMap database using a novel distance metric that is able to score the similarity between gene expression profiles and drug treatment. The authors partitioned the DN using graph theory tools to identify groups of drugs (communities) that are densely interconnected [63]; the same method was also applied by [82, 83]. Their results revealed that these groups were significantly enriched with drugs of a similar MoA and therapeutic purpose and, hence, can be used for such predictive purposes. Their analysis exemplified their method studying HSP90 and CDK2 inhibitors and showed that the predicted MoAs correspond to results known in the literature [25, 63, 84]. Interestingly, their method revealed a previously unknown MoA link between fasudil, a Rho-kinase inhibitor, and autophagy. An experimental validation indeed confirmed this connection suggesting a repositioning of this drug because so far fasudil is approved in Japan for the treatment of cerebral vasospasm characterized by blood vessel obstruction.

Kibble *et al.* uses CMap approach to show, via the case study of the natural product pinosylvin, how the combination of two complementary network-based methods can provide novel mechanistic insights. They illustrate that elucidating the MoA of multi-targeted natural products through transcriptional response-based approaches can lead to unbiased hypotheses that might not have been otherwise conceived and, hence, to truly novel and even surprising findings [85].

Dudley *et al.* have shown that CMap data contain sufficient information about the dynamic activities of human genes for reconstructing gene–gene interactions in drug-perturbed cancer cells. They successfully applied a Gaussian Bayesian network framework [86] to reconstruct a subnetwork containing validated interactions between genes with known roles in the apoptosis pathway. In addition, their network successfully predicted key players and interactions in drug-induced apoptosis, including both intrinsic and extrinsic apoptosis pathways [87].

Choi *et al.* [5] proposed another computational optimization method using CMap to find drug MoA. Their study used gene expression signatures of disease states or physiological processes with gene expression signatures of small-molecule drugs to predict novel functional associations between small molecules sharing the same MoA. The heat-shock protein 90 inhibitors (HSP90i) were identified in the study as a candidate that suppresses homologous recombination (HR) in epithelial ovarian cancer (EOC) patients [5]. They further showed that sublethal concentrations of HSP90i 17-AAG suppresses HR sensitivity observed in ovarian cancer cells [5, 88]. Hence, the authors concluded that the combination of 17-AAG and PARP inhibitors (PARPis) olaparib or carboplatin in EOCs that inhibit HR will be effective when developing PARPi resistance [5].

Shigemizu *et al.* [15] introduced a novel methodology similar to the partial-rank metric, by using gene expression profile to apply the CMap concept to identify candidate therapeutics for MoA, targeting possible functions that are beyond drug repositioning [89]. The method uses drug candidates in a pool of compounds that downregulate the overexpressed genes, or upregulate the underexpressed genes, for a given abnormal phenotypic condition and demonstrate the utility of their approach for drug repositioning. The authors pointed out that the improved functionality of their method will help in identifying a drug or a group of drugs with potential heterogeneous properties. On the other hand, the method can be used to find genes that can be targeted by a set of identified compounds. For instance, the genes RPL35, LAMB1 and CAV1 have been found to be breast cancer targets [15, 90]. Finally, the result of their functional analysis indicated that the MoA of tamoxifen is given by downregulating TGF-β signaling [15].

Drug repurposing

Generally, drug repurposing refers to investigating drugs that are already used for treating a particular disease to see if they can be safely and effectively used for treating other diseases. The terms repurposing and repositioning are used interchangeably. Owing to the fact that the repurposing of a drug builds on previous research and development efforts, new candidate therapies could be ready for clinical usage more quickly and at reduced costs. Over the past years, many approaches have been developed for the generic drug repurposing; however, in the following, we will focus on investigations that have been using CMap to repurpose drugs and to identify novel targets.

For instance, Kunkel and his group [37] used CMap to determine ursolic acid, a natural compound that is e.g. present in apples, as a lead compound for reducing fasting-induced muscle atrophy. They used rodents for an *in vivo* validation of the therapeutic concept, demonstrating that ursolic acid is a potentially interesting therapy candidate for muscle atrophy and perhaps other metabolic diseases.

Applying the connectivity mapping approach to acute myeloid leukemia (AML), Ramsey *et al.* integrated gene signatures from a mouse model of AML and a cohort of AML patients to query the ssCMap. They identified entinostat as a candidate drug able to alter the AML condition toward the normal state. This prediction was followed up experimentally in cell line as well as mouse models, and the authors were able to validate the

predicted effects of entinostat on the signature genes, and showed that *in vivo* treatment with this compound resulted in prolonged survival of leukemic mice [91].

Johnstone *et al.* used a comparative microarray analysis of compound-induced changes in gene expression for a possible drug repurposing, and they discovered a novel compound. This finding suggests a possible mechanism of calmodulin signaling using piperazine as promoters of central nervous system (CNS) neurite growth [54]. This study suggests that calmodulin can be seen as a novel target enhancing neuron regeneration. Furthermore, their analysis showed that a previously unrecognized potential for piperazine phenothiazine antipsychotics can be repurposed for neuron regeneration [54].

Jin *et al.* [92] presented a novel computational drug-repurposing method to screen a combined set of drugs together for treating type 2 diabetes [93]. Interestingly, they found that a combination of Trolox C and Cytisine is effective for the treatment of type 2 diabetes, but if used separately, neither of the drugs are effective. Similarly, Sirota *et al.* [94] integrated a new gene expression database from 100 diseases and 164 drug compounds, yielding predictions for all drug compounds that show a high consistency with already known therapeutics. As a demonstration for a novel prediction, an experimental validation for the antiulcer drug cimetidine was provided as a candidate therapeutics in the treatment of lung adenocarcinoma (LA).

Malcomson *et al.* [95] has recently applied computational drug repurposing successfully, as well, by using sscMap to identify candidate drugs that could be used to induce A20 and to normalize the inflammatory response in cystic fibrosis. A20 (TNFAIP3) is a known nuclear factor-kB regulator, which is reduced in airway cells. The authors used a co-expression-based analysis to create a gene signature consisting of A20 showing high correlation. Then, Malcomson *et al.* performed a connectivity mapping analysis using the sscMap framework. The identified candidate drugs were subsequently validated in airway epithelial cells, confirming that ikarugamycin and quercetin have anti-inflammatory effects mediated by induction of A20. They used small interfering RNA experiments to illustrate that the anti-inflammatory effect of these two drugs is mainly because of A20 induction.

Drug combinations

Rather than using single drugs in treating diseases, combinations of multiple drugs are gaining more and more interest. Such drug combinations are motivated by studies indicating higher efficacy, fewer side effects and less toxicity compared with single-drug treatments [36, 96, 97]. This seems to be particularly appropriate for complex disorders such as cancer, as cancer cells possess compensatory mechanisms to overcome perturbations occurring at the individual signaling pathway level by means of, e.g. mutations of key receptors or cross-talk between pathways [98].

For instance, Lee *et al.* [98] developed the Combinatorial Drug Assembler as a genomic and bioinformatics system by using gene expression profiling to target multiple signaling pathways for a combinatorial drug discovery. The method performs an expression search against a signaling pathway to compare gene expression profiles of patient samples (or cell lines) as input signature, with the expression patterns of the sample treated with different small molecules. The method then finds the best pattern that matches the combination of two drugs across the input signature related to signaling pathways to detect and predict those drugs that could be used in a combination therapy. Furthermore, they performed *in vitro* validations for NSCLC and triple-negative breast cancer (TNBC) cells and found that alsterpaullone and scriptaid as well as irinotecan and semustin for NSCLC, halofantrine and vinblastine for TNBC, showed synergistic effects.

Huang *et al.* [99] proposed a novel systematic computational approach called DrugComboRanker to find synergistic drug combinations and to uncover their MoA. The drug functional framework was built based on genetic profiles and network partitions of various DN clusters using a Bayesian nonnegative matrix factorization. By building disease-specific signaling networks based on disease profiles, drug combinations can be identified by searching drugs whose targets are enriched in the reference signaling module of the disease signaling network. An evaluation of the method was performed for LA and endocrine receptor-positive breast cancer.

Wang and his group [36] performed a meta-analysis to obtain a list of 343 DEG of LA and used this signature to query CMap to identify a combination of compounds whose treatment reverse the expression direction. Compounds in categories such as HSP90 inhibitor, HDAC inhibitor, PPAR agonist and PI3K inhibitor were identified as top candidates. An *in vitro* validation demonstrated that either 17-AAG (HSP90 inhibitor) alone or in combination with cisplatin can significantly inhibit LA cell growth by inducing cell cycle arrest and apoptosis.

Parkkinen *et al.* [41] showed their proposed probabilistic connectivity mapping method is capable of identifying drug combinations. Specifically, they define a combined drug profile consisting of drug pairs by assessing the correlation of their individual profiles. Overall, this leads to a ranking of drug pairs rather than individual drugs. A computational assessment of the proposed method was conducted considering ATC codes and chemical similarity as ground truth. Their hypothesis was that single drugs with ATC codes having minor response effects will not result in a high relevance score, as other drugs with stronger effects will dominate. However, their statistical analysis demonstrated that a combinatorial matching improves the results for many polypharmacologic drugs [41]. The authors highlight how LINCS data set [11] could be used to extend benefits of the group factor analysis-based probabilistic connectivity mapping in drug combination. As it identifies both single or shared responses across a large number of cell types, making it valuable for drug discovery and development would be even possible to impose more structure on the group factor analysis model, by similarly inferring response of a specific cell line to a drug, enabling high relevant information for personalized medicine studies.

Experimental validations

Using a computational biology approach in combination with CMap can help in finding new forms of drugs, predicting drug candidates, pharmacological and toxicological properties in chemicals [19, 100–102]. However, these predictions need to be evaluated experimentally, either by using cell viability after drug treatment *in vitro* or tumor growth after drug treatment *in vivo* and, in some cases, using survival analysis of drug treatment in the clinic. Moreover, disease samples collected from patients are used to investigate the dynamics of disease progression; apart from that, diverse preclinical models, such as cell lines and animal models, could be used in experiments to interpret CMap results, understand disease and validate hypothesis. In this section, we discuss studies that provided such experimental validations.

Notably, Ishimatsu-Tsuji *et al.* identified fluphenazine compound as a novel inducer in hair-growth cycle using CMap.

Moreover, the results showed the additive effect of two compounds that are being ranked by the CMap analysis [100]. Caiment *et al.* studied the reliability of the CMap method for classifying and predicting a drug in different forms. The study was performed on hepatocellular carcinoma and liver cell model exposed to a wide range of different compounds using ssCMap application. The results of the analysis revealed significant positive connections [103]. Moreover, the method showed how the CMap approach is robust in predicting a drug's carcinogenicity based on data from representative *in vitro* models by adding more relevance for predicting human disease state and may be considered as a classification way of discovering new compounds [103]. Also, Wang *et al.* established prediction models for various adverse drug reactions, including severe myocardial and infectious events. Also, they were able to identify drugs with FDA boxed warnings for safety liability effectively. Therefore, it illustrates that a combination of effective computational methods and drug-induced gene expression change can be proven as new cutting edge to have a systematic drug safety evaluation [104].

Public data sets can be leveraged to validate drug hits and understand drug mechanisms, e.g. drug efficacy and toxicity. Using *in silico* drug screening via CMap followed by empirical validations, Cheng *et al.* discovered that thioridazine can reduce the viability of glioblastoma (GBM) cells and GBM stem cells, induce autophagy and affect the expressions of related proteins in GBM cells. Thus, thioridazine has the potential to treat GBM [55]. In addition, thioridazine induces autophagy and apoptosis at a high concentration, functioning through G protein-coupled receptors.

Although drugs in these previous examples were validated in preclinical models, the question of whether the disease gene expression was really reversed in disease models remains unknown. A recent study in a mouse model of dyslipidemia found that treatments that restore gene expression patterns to their norm are associated with the successful restoration of physiological markers to their baselines, providing a sound basis to this computational approach.

PharmacoGx: a computational pharmacogenomics platform

The availability of large-scale perturbation data sets, such as CMap and LINCS L1000, opened new avenues for research in pharmacogenomics. Nonetheless, issues such as lack of standards for annotation, storage, access and analysis challenge the full exploitation of the pharmacogenomics data sets. Hence, unifying platforms are required to integrate the currently existing data sets and the corresponding mining tools. For data integration purposes, such platforms should remove biases of different sources such as batch effects, difference between profiling platforms and cell-specific differences to best characterize drug-induced effects. Furthermore, the unifying platforms should be easy to use so that users can develop new methods and functions for easy data manipulation and mining within the platform [105–107]. To address these issues, PharmacoGx, an open source package, has been recently developed [108]. To the best of our knowledge, PharmacoGx is currently the only integrative platform developed for this purpose.

The PharmacoGx platform comprises two fundamental components: first, efficient data structures to store pharmacological and molecular data and experimental metadata (e.g. molecular profiles of cell lines before and after treatment by compounds) provided by the pharmacogenomics data sets. The storage scheme of PharmacoGx provides a common interface for multiple data sets, standardizes cell line and drug identifiers, and provides easy access to the data. Furthermore, it facilitates easy and side-by-side comparison of the pharmacogemonics data sets that are usually scattered and independently collected.

The second component of PharmacoGx is its set of functions for data manipulation and mining tasks, such as, removing the biases of data and creating signatures representing drug-induced changes in the gene expression of cell lines, implementation of the connectivity mapping analysis and computing the connectivity score to infer links between the drug-induced signatures and phenotypes. Furthermore, it should be noted that such functions are not data set specific. For instance, connectivity mapping analysis can be performed on not only the CMap data set but also the LINCS L1000 and any other drug perturbation data set that will be curated and published in the future. This provides an opportunity to compare the query results from several data sets alongside one another. These features contribute to the uniqueness of the PharmacoGx package.

Connectivity mapping via PharmacoGx: a case study

We designed an experiment to show that PharmacoGx package enables users to easily query the two state-of-the-art perturbation data sets (i.e. CMap and L1000), and facilitates comparison of the results along each other. For this purpose, we illustrate a case study similar to the phenothiazines example by Lamb *et al.* in the original CMap publication. L1000 and CMap both contain profiles of five members of phenothiazine antipsychotics (i.e. chlorpromazine, fluphenazine, prochlorperazine, thioridazine and trifluoperazine). We first generated a small L1000 signature set (Supplementary Materials) consisting of 10 unique instances of the family members and 990 randomly selected perturbation signatures from the L1000 data set. The goal of this experiment is to retrieve phenothiazine family members, from the L1000 and CMap data sets, using a query signature generated from the profile of only one of the family members (e.g. trifluoperazine). We used trifluoperazine's signature to generate a query signature by selecting only genes whose expression values are highly affected by the drug (—t-stat—¿1). This led to a signature of length 458. Query results have been shown in Table 5 as two ranked lists. PharmacoGx matched trifluoperazine signature as the most similar to the query signature in both data sets. The other family members have also been retrieved as top hits in both lists.

Discussion

The CMap methodology has been used in numerous applications by many research groups with a particular focus in drug

Table 5. Results of retrieving phenothiazines using a query signature generated from trifluoperazine profile

L1000 rank	Drug name	CMap rank	Drug name
1	Trifluoperazine	1	Trifluoperazine
2	Fluphenazine	2	Thioridazine
3	Thioridazine	3	Fluphenazine
4	Triflupromazine	4	Prochlorperazine
74	Fluphenazine	20	Chlorpromazine
201	Prochlorperazine		
253	Chlorpromazine		
271	Chlorpromazine		
284	Chlorpromazine		
402	Chlorpromazine		
438	Chlorpromazine		

discovery and development as pointed out in this review. These efforts have been aimed at identifying new therapeutic targets, drug repurposing/repositioning opportunities, finding new MoA for new or existing small molecules, predicting side effects and improving biological understanding. Most of the potentials of CMap mentioned are undoubtedly beneficial in pharmacogenomics research and useful in drug industries, as this approach has been found to be extremely valuable in multiple biomedical research scenarios.

The CMap method uses a simplistic model of pattern matching techniques based on an unproven hypothesis to understand the concept of cell biology in drug discovery. However, there is no account for dynamics associated with the disease or the drug under investigation, multi-organ effects and genetic variations. Therefore, incorporating additional models and data sources will help in understanding the effect of candidate drugs in specific disease settings and appropriate cellular tissue and environmental factors that are more effective in drug discovery/repurposing applications. Applications are not limited to such disease-oriented querying with, for example, illustrations of CMap generating hypotheses concerning MoA being showcased. While CMap has achieved some notable successes [37, 75], pathways and network-based models provide a more realistic system-level insights into the molecular targets of the drug candidates, which is an essential step in drug repurposing/repositioning process and phenotypic-based discovery [84].

Moreover, some limitations of the CMap approach can be highlighted, for example, experimental replicates, a potential issue with the CMap data (Build 1), as most small molecules have only one replicate per cell line for each experiment. This will present some challenges on statistical analysis, such as finding DEG for small molecules compounds. Another limitation is cell line coverage (the experiment was conducted only using five human cancer cell lines and not all small molecules were tested on all cell lines), the limited dosages and time points (several small molecules were tested using 10 mM concentration with 6 h perturbation time point). Another possible limitation in CMap is the presence of potential batch effect, the similarity of gene expression profiles observed for unrelated stimuli in grown or processed cells at the same time. Batch effects have been identified as a significant source of systematic error that can be corrected [82]. Attempts to solve the problem of batch effects have been made in the methods proposed. For example, Iskar *et al.* [18] performed a quantitative evaluation of CMap methods by applying a centered mean approach to normalize the gene expression intensity values in CMap to reduce batch-specific effects. Also, Iorio *et al.* uses the pairwise drug-induced gene expression profile similarity (DIPS) scores between drug pairs in CMap to calculate total enrichment score [4]. They used drug compounds with shared ATC classification, and high chemical similarities to discretize true positives in their approach. This is relevant in willingness to sacrifice true positives to keep false positives low. Notably, Cheng *et al.* used the ATC classification as a benchmark to address batch effects using XCos. The novel XCos approach is used to determine which drug compounds contain robust expression profiles in CMap data, and which analytical approaches are more accurate to use when evaluating CMap data set. Although some of these limitations are derived from the practicality and resource constraints at the time of designing the approach, the caveats associated with such systems abstraction methodology need to be addressed during study design, for example, a proper biological context, relevance of transcriptional changes to disease states, representation of gene signatures to the global expression profile and the overall reliability of the approach. Now with the availability of the LINCS L1000 data set, covering cellular responses upon the treatment of chemical/genetic perturbagen, including over 1.4 million gene expression profiles representing $\sim$15,000 small molecules compounds and $\sim$5000 genes (small hairpin RNA and overexpression) in $\sim$15 cell lines. Researchers can leverage the publicly available data to overcome some of the CMap shortcomings.

The LINCS L1000 still lacks quality needed for comprehensive drug discovery/repurposing, which makes it challenging for understanding the data-processing pipeline and lead inferences, mostly because it uses a noisy platform [109]. The current imputation of the computational inferred genes used by the L1000 in generating the data is also lagging. What is certain is that, the recent methods developed using CMap/LINCS L1000 data have already shown great promises and constantly becoming more appealing to researchers in pharmacogenomics. For more comprehensive understanding of drug MoA, some methodologies incorporating other omics than transcriptomics would be beneficial, including, for instance, methylation array for epigenetic compound such as HDAC inhibitors or 5-AZA-CdR, metabolomics and proteomics, as well as dynamic or longitudinal data, would widen the limited view captured by the single time point of transcriptomic responses. This will give the opportunity to shift drug discovery toward personalized and precision medicine treatment approach to enhance disease therapies.

Conclusion

In this article, we reviewed the connectivity mapping methodology and applications. Perturbation databases, such as CMap or LINCS, offer a wealth of opportunities for computational drug discovery approaches by enabling pharmacogenomics that extends beyond classical pharmacology. A reason for this is that these transcriptomic perturbation databases allow network (nonsingle gene centered) approaches, e.g. at the pathway or network level. So far, the majority of applications are focused on different cancer types. However, the principal ideas can be translated to any other type of complex disease opening in this way the door into a new era of drug discoveries. Research in extending connectivity mapping concept and methodology is ongoing, and there are still aspects such as the application of different similarity metrics that need further investigations. Although few variations and improvements over the original CMap have been proposed, the field lacks systematic evaluations of the new approaches. Therefore, advantages and disadvantages of different methods are so far not precisely measurable.

Key Points

- Comprehensive review of perturbation databases, e.g. CMap and LINCS L1000, that can be used for drug discovery and drug repurposing.
- Surveying applications of CMap and LINCS L1000 for novel pharmacogenomics approaches.
- Presentation of benchmarking approaches for evaluating computational drug discovery approaches.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgment

For professional proof reading of the manuscript we would like to thank Bárbara Macías Solís.

Funding

AM was supported by a fellowship from CIMO (Finland) and FE-S was supported by TUT (Finland). S-DZ was supported by a grant of £11.5M (PI Professor Tony Bjourson) from European Union Regional Development Fund (ERDF) EU Sustainable Competitiveness Programme for N. Ireland, Northern Ireland Public Health Agency (HSC R&D) & Ulster University, and supported by the UK BBSRC/M- RC/EPSRC co-funded grant BB/I009051/1. MD thanks the Austrian Science Funds for supporting this work (project P26142). LSG was supported by the Canadian Cancer Society Research Institute (grant #703886). BH-K was supported by the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre, the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Ministry of Economic Development and Innovation/Ministry of Research & Innovation of Ontario (Canada). GVG was supported in part by the Arkansas INBRE program, with grants from the National Center for Research Resources (P20RR016460) and the National Institute of General Medical Sciences (P20 GM103429) from the National Institutes of Health.

References

1. Schenone M, Dancik V, Wagner BK, *et al*. Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* 2013;**9**(4):232–40.
2. Wang H, Gu Q, Wei J, *et al*. Mining drug–disease relationships as a complement to medical genetics-based drug repositioning: where a recommendation system meets genome-wide association studies. *Clin Pharmacol Ther* 2015;**97**(5):451–4.
3. Lamb J, Crawford ED, Peck D, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**(5795):1929–35.
4. Iorio F, Bosotti R, Scacheri E, *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;**107**(33):14621–6.
5. Choi YE, Battelli C, Watson J, *et al*. Sublethal concentrations of 17-aag suppress homologous recombination dna repair and enhance sensitivity to carboplatin and olaparib in hr proficient ovarian cancer cells. *Oncotarget* 2014;**5**(9):2678–87.
6. Rasmussen CE. *Gaussian Processes for Machine Learning*. Citeseer, New York, 2006.
7. Napolitano F, Zhao Y, Moreira VM, *et al*. Drug repositioning: a machine-learning approach through data integration. *J Cheminform* 2013;**5**:30.
8. Pacini C, Iorio F, Gonçalves E, *et al*. Dvd: an r/cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* 2013;**29**(1):132–4.
9. Kim J, Yoo M, Kang J, *et al*. K-map: connecting kinases with therapeutics for drug repurposing and development. *Hum Genomics* 2013;**7**(1):20.
10. Alaimo S, Bonnici V, Cancemi D, *et al*. Dt-web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC Syst Biol* 2015;**9(Suppl 3)**:S4.
11. Vidovic D, Koleti A, Schurer SC. Large-scale integration of small molecule-induced genome-wide transcriptional responses, kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. *Front Genet* 2014;**5**:342.
12. Qu XA, Rajpal DK. Applications of connectivity map in drug discovery and development. *Drug Discov Today* 2012;**17**(23):1289–98.
13. Kannan L, Ramos M, Re A, *et al*. Public data and open source tools for multi-assay genomic investigation of disease. *Brief Bioinform* 2016;**17**(4):603–15.
14. Dudley JT, Sirota M, Shenoy M, *et al*. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med* 2011;**3**(96):96ra76.
15. Shigemizu D, Hu Z, Hung JH, *et al*. Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer. *PLoS Comput Biol* 2012;**8**(2):e1002347.
16. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**(43):15545–50.
17. Hieronymus H, Lamb J, Ross KN, *et al*. Gene expression signature-based chemical genomic prediction identifies a novel class of {HSP90} pathway modulators. *Cancer Cell* 2006;**10**(4):321–30.
18. Iskar M, Campillos M, Kuhn M, *et al*. Drug-induced regulation of target expression. *PLoS Comput Biol* 2010;**6**(9):e1000925.
19. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 2011;**12**(4):303–11.
20. Ahmed J, Meinel T, Dunkel M, *et al*. Cancerresource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res* 2011;**39(Suppl 1)**:D960–7.
21. Woo JH, Shimoni Y, Yang WS, *et al*. Elucidating compound mechanism of action by network perturbation analysis. *Cell* 2015;**162**(2):441–51.
22. Bisikirska B, Bansal M, Shen Y, *et al*. Elucidation and pharmacological targeting of novel molecular drivers of follicular lymphoma progression. *Cancer Res* 2016;**76**(3):664–74.
23. Korkut A, Wang W, Demir E, *et al*. Perturbation biology nominates upstream–downstream drug combinations in raf inhibitor resistant melanoma cells. *eLife* 2015;**4**:e04640.
24. Tabares-Seisdedos R, Rubenstein JL. Inverse cancer comorbidity: a serendipitous opportunity to gain insight into cns disorders. *Nat Rev Neurosci* 2013;**14**(4):293–304.
25. Engerud H, Tangen IL, Berg A, *et al*. High level of hsf1 associates with aggressive endometrial carcinoma and suggests potential for HSP90 inhibitors. *Br J Cancer* 2014;**111**(1):78–84.
26. Segal MR, Xiong H, Bengtsson H, *et al*. Querying genomic databases: refining the connectivity map. *Stat Appl Genet Mol Biol* 2012;**11**(2).
27. Fortney K, Griesman J, Kotlyar M, *et al*. Prioritizing therapeutics for lung cancer: an integrative meta-analysis of cancer gene signatures and chemogenomic data. *PLoS Comput Biol* 2015;**11**(3):e1004068–03.
28. Cheng J, Yang L, Kumar V, *et al*. Systematic evaluation of connectivity map for disease indications. *Genome Med* 2014;**6**(12):540.
29. Duan Q, Flynn C, Niepel M, *et al*. Lincs canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures. *Nucleic Acids Res* 2014;**42**(W1):W449–60.

30. Barrett T, Wilhite SE, Ledoux P, *et al*. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5.
31. Chen EY, Tan CM, Kou Y, *et al*. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128.
32. Duan Q, Reid SP, Clark NR, *et al*. L1000cds2: lincs l1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl* 2016;**2**:16015.
33. Zhang SD, Gant T. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 2008;**9**(1):258.
34. Chung F, Chiang Y, Tseng A, *et al*. Functional module connectivity map (fmcm): a framework for searching repurposed drug compounds for systems treatment of cancer and an application to colorectal adenocarcinoma. *PloS One* 2014;**9**(1):e86299.
35. Zhang SD, Gant T. sscmap: an extensible JAVA application for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 2009;**10**:236.
36. Wang G, Ye Y, Yang X, *et al*. Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma. *PloS One* 2011;**6**(1):e14573.
37. Kunkel SD, Suneja M, Ebert SM, *et al*. mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metab* 2011;**13**(6):627–38.
38. Breitling R, Armengaud P, Amtmann A, *et al*. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* 2004;**573**(1–3):83–92.
39. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, *et al*. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 2007;**9**(2):166–80.
40. Yeh CT, Wu ATH, Chang PMH, *et al*. Trifluoperazine, an antipsychotic agent, inhibits cancer stem cell growth and overcomes drug resistance of lung cancer. *Am J Respir Crit Care Med* 2012;**186**(11):1180–8. 2015/06/08
41. Parkkinen J, Kaski S. Probabilistic drug connectivity mapping. *BMC Bioinformatics* 2014;**15**:113.
42. Cheng J, Xie Q, Kumar V, *et al*. Evaluation of analytical methods for connectivity map data. In: *Pacific Symposium on Biocomputing 2013*, Kohala Coast, Hawaii, USA, 2013, 5.
43. Li Y, Hao P, Zheng S, *et al*. Gene expression module-based chemical function similarity search. *Nucleic Acids Res* 2008;**36**(20):e137.
44. Harris MA, Clark J, Gene Ontology Consortium, *et al*. The gene ontology (go) database and informatics resource. *Nucleic Acids Res* 2004;**32(Suppl 1)**:D258–61.
45. McArt DG, Bankhead P, Dunne PD, *et al*. cudaMap: a GPU accelerated program for gene expression connectivity mapping. *BMC Bioinformatics* 2013;**14**:305.
46. O'Reilly PG, Wen Q, Bankhead P, *et al*. Quadratic: scalable gene expression connectivity mapping for repurposing fda-approved therapeutics. *BMC Bioinformatics* 2016;**17**(1):1–15.
47. Wen Q, Philip D, O'Reilly PD, *et al*. Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. *BMC Syst Biol* 2015;**9**(5):1–11.
48. Wen Q, Kim C, Hamilton P, *et al*. A gene-signature progression approach to identifying candidate small-molecule cancer therapeutics with connectivity mapping. *BMC Syst Biol* 2016;**17**:211.
49. Cheng J, Yang L. Comparing gene expression similarity metrics for connectivity map. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2013, pp. 165–70.
50. Madani TSA, Ghoraie LS, Manem VSK, *et al*. Predictive approaches for drug combination discovery in cancer. *Brief Bioinform* 2016, doi: 10.1093/bib/bbw104.
51. Sanda T, Li X, Gutierrez A, *et al*. Interconnecting molecular pathways in the pathogenesis and drug sensitivity of T-cell acute lymphoblastic leukemia. *Blood* 2009;**115**(9):1735–45.
52. Yuen T, Iqbal J, Zhu LL, *et al*. Disease-drug pairs revealed by computational genomic connectivity mapping on gba1 deficient, gaucher disease mice. *Biochem Biophys Res Commun* 2012;**422**:573–7.
53. Lim SM, Lim JY, Cho JY. Targeted therapy in gastric cancer: personalizing cancer treatment based on patient genome. *World J Gastroenterol* 2014;**20**(8):2042–50,
54. Johnstone AL, Reierson GW, Smith RP, *et al*. A chemical genetic approach identifies piperazine antipsychotics as promoters of cns neurite growth on inhibitory substrates. *Mol Cell Neurosci* 2012;**50**(2):125–35.
55. Cheng HW, Liang YH, Kuo YL, *et al*. Identification of thioridazine, an antipsychotic drug, as an antiglioblastoma and anticancer stem cell agent using public gene expression data. *Cell Death Dis* 2015;**6**:e1753–05.
56. Kang S, Rho SB, Kim B. A gene signature-based approach identifies thioridazine as an inhibitor of phosphatidylinositol-3-kinase (pi3k)/akt pathway in ovarian cancer cells. *Gynecol Oncol* 2011;**120**(1):121–7.
57. Toscano MG, Navarro-Montero O, Ayllon V, *et al*. SCL/tal1-mediated transcriptional network enhances megakaryocytic specification of human embryonic stem cells. *Mol Ther* 2015;**23**(1):158–70.
58. Tiedemann RE, Schmidt J, Keats JJ, *et al*. Identification of a potent natural triterpenoid inhibitor of proteosome chymotrypsin-like activity and NF-b with antimyeloma activity *in vitro* and *in vivo*. *Blood* 2009;**113**(17):4027–37.
59. Hassane DC, Guzman ML, Corbett C, *et al*. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood* 2008;**111**(12):5654–62.
60. McArt DG, Dunne PD, Blayney JK, *et al*. Connectivity mapping for candidate therapeutics identification using next generation sequencing RNA-seq data. *PLoS One* 2013;**8**(6):e66902–6.
61. Li H, Lovci MT, Kwon YS, *et al*. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci USA* 2008;**105**(51):20179–84.
62. Spijkers-Hagelstein JAP, Pinhancos SS, Schneider P, *et al*. Chemical genomic screening identifies ly294002 as a modulator of glucocorticoid resistance in mll-rearranged infant all. *Leukemia* 2014;**28**(4):761–9.
63. Iorio F, Saez-Rodriguez J, Bernardo DD. Network based elucidation of drug response: from modulators to targets. *BMC Syst Biol* 2013;**7**:139.
64. Jiang W, Chen X, Liao M, *et al*. Identification of links between small molecules and mirnas in human cancers based on transcriptional responses. *Sci Rep* 2012;**2**:282.
65. Wang J, Meng F, Dai E, *et al*. Identification of associations between small molecule drugs and mirnas based on functional similarity. *Oncotarget* 2016;**7**(25):38658–69.
66. Clark NR, Hu KS, Feldmann AS, *et al*. The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics* 2014;**15**:79.

67. McLauchlan H, Elliott M, cohen P. The specificities of protein kinase inhibitors: an update. *Biochem J* 2003;**371**(1):199–204.
68. Claerhout S, Lim JY, Choi W, *et al*. Gene expression signature analysis identifies vorinostat as a candidate therapy for gastric cancer. *PLoS One* 2011;**6**(9):e24662.
69. Khan SA, Virtanen S, Kallioniemi OP, *et al*. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics* 2014;**30**(17):i497–504.
70. Zhu Y, Das K, Wu J, *et al*. Rnh1 regulation of reactive oxygen species contributes to histone deacetylase inhibitor resistance in gastric cancer cells. *Oncogene* 2014;**33**(12):1527–37.
71. Siu FM, Ma DL, Cheung YW, *et al*. Proteomic and transcriptomic study on the action of a cytotoxic saponin (polyphyllin d): induction of endoplasmic reticulum stress and mitochondria-mediated apoptotic pathways. *Proteomics* 2008;**8**(15):3105–17.
72. Wen Z, Wang Z, Wang S, *et al*. Discovery of molecular mechanisms of traditional chinese medicinal formula Si-Wu-Tang using gene expression microarray and connectivity map. *PLoS One* 2011;**6**(3):e18278–03.
73. Lee MS, Chan JY, Kong S, *et al*. Effects of Polyphyllin d, a steroidal saponin in paris polyphylla, in growth inhibition of human breast cancer cells and in xenograft. *Cancer Biol Ther* 2005;**4**(11):1248–54.
74. Laenen G, Thorrez L, Bornigen D, *et al*. Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol Biosyst* 2013;**9**:1676–85.
75. Jahchan NS, Dudley JT, Mazur PK, *et al*. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov* 2013;**3**(12):1364–77.
76. Lee S, Lee K, Song M, *et al*. Building the process-drug-side effect network to discover the relationship between biological processes and side effects. *BMC Bioinformatics* 2011;**12(Suppl 2)**:S2.
77. Pritchard JR, Bruno PM, Hemann MT, *et al*. Predicting cancer drug mechanisms of action using molecular network signatures. *Mol Biosyst* 2013;**9**(7):1604–19.
78. Kumar N, Hendriks BS, Janes KA, *et al*. Applying computational modeling to drug discovery and development. *Drug Discov Today* 2006;**11**(17):806–11.
79. Huang H, Liu CC, Zhou XJ. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc Natl Acad Sci USA* 2010;**107**(15):6823–8.
80. Gu Q, Chen XT, Xiao YB, *et al*. Identification of differently expressed genes and small molecule drugs for tetralogy of fallot by bioinformatics strategy. *Pediatr Cardiol* 2014;**35**(5):863–9.
81. Issa NT, Kruger J, Byers SW, *et al*. Drug repurposing a reality: from computers to the clinic. *Expert Rev Clin Pharmacol* 2013;**6**(2):95–7.
82. Kibble M, Saarinen N, Tang J, *et al*. Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products. *Nat Prod Rep* 2015;**32**(8):1249–66.
83. Jensen K, Ni Y, Panagiotou G, *et al*. Developing a molecular roadmap of drug-food interactions. *PLoS Comput Biol* 2015;**11**(2):e1004048–02.
84. Iorio F, Rittman T, Ge H, *et al*. Transcriptional data: a new gateway to drug repositioning? *Drug Discov Today* 2013;**18**(7):350–7.
85. Kibble M, Khan SA, Saarinen N, *et al*. Transcriptional response networks for elucidating mechanisms of action of multitargeted agents. *Drug Discov Today* 2016;**21**(7):1063–75.
86. Dudley JT, Schadt E, Sirota M, *et al*. Drug discovery in a multidimensional world: systems, patterns, and networks. *J Cardiovasc Transl Res* 2010;**3**(5):438–47.
87. Yu J, Putcha P, Silva JM. Recovering drug-induced apoptosis subnetwork from connectivity map data. *Biomed Res Int* 2015;**2015**:708563.
88. Gao L, Zhao G, Fang JS, *et al*. Discovery of the neuroprotective effects of alvespimycin by computational prioritization of potential anti-parkinson agents. *FEBS J* 2014;**281**(4):1110–22.
89. Ravindranath AC, Perualila-Tan N, Kasim A, *et al*. Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism-of-action analysis. *Mol Biosyst* 2015;**11**(1):86–96.
90. Ma C, Chen HH, Flores M, *et al*. Brca-monet: a breast cancer specific drug treatment mode-of-action network for treatment effective prediction using large scale microarray database. *BMC Syst Biol* 2013;**7(Suppl 5)**:S5.
91. Ramsey JM, Kettyle LMJ, Sharpe DJ, *et al*. Entinostat prevents leukemia maintenance in a collaborating oncogene-dependent model of cytogenetically normal acute myeloid leukemia. *Stem Cells* 2013;**31**(7):1434–45.
92. Jin L, Tu J, Jia J, *et al*. Drug-repurposing identified the combination of trolox c and cytisine for the treatment of type 2 diabetes. *J Transl Med* 2014;**12**:153.
93. Lucas FAS, Fowler J, Kopetz S, *et al*. Abstract 5371: drug repositioning with a bioinformatics platform that integrates the TCGA, CMAP and CCLE. *Cancer Res* 2014;**74(Suppl 19)**:5371.
94. Sirota M, Dudley JT, Kim J, *et al*. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* 2011;**3**(96):96ra77.
95. Malcomson B, Wilson H, Veglia E, *et al*. Connectivity mapping (sscmap) to predict a20 inducing drugs anti-inflammatory action in cystic fibrosis. *Proc Natl Acad Sci USA* 2016;**113**(26):E3725–34.
96. Gupta EK, Ito MK. Lovastatin and extended-release niacin combination product: the first drug combination for the management of hyperlipidemia. *Heart Dis* 2002;**4**(2):124–37.
97. Sun X, Vilar S, Tatonetti NP. High-throughput methods for combinatorial drug discovery. *Sci Transl Med* 2013; **5**(205):205rv1.
98. Lee J, Kim DG, Bae TJ, *et al*. Cda: combinatorial drug discovery using transcriptional response modules. *PloS One* 2012; **7**(8):e42573.
99. Huang L, Li F, Sheng J, *et al*. Drugcomboranker: drug combination discovery based on target network analysis. *Bioinformatics* 2014;**30**(12):i228–36.
100. Ishimatsu-Tsuji Y, Soma T, Kishimoto J. Identification of novel hair-growth inducers by means of connectivity mapping. *FASEB J* 2010;**24**(5):1489–96.
101. Gottlieb A, Stein GY, Ruppin E, *et al*. Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**(1):496.
102. Bao H, Wang J, Zhou D, *et al*. Protein-protein interaction network analysis in chronic obstructive pulmonary disease. *Lung* 2014;**192**(1):87–93.
103. Caiment F, Tsamou M, Jennen D, *et al*. Assessing compound carcinogenicity *in vitro* using connectivity mapping. *Carcinogenesis* 2014;**35**(1):201–7.
104. Wang K, Weng Z, Sun L, *et al*. Systematic drug safety evaluation based on public genomic expression (connectivity map) data: myocardial and infectious adverse reactions as application cases. *Biochem Biophys Res Commun* 2015;**457**(3):249–55.
105. Safikhani Z, El-Hachem N, Quevedo R, *et al*. Assessment of pharmacogenomic agreement. *F1000Res* 2016;**5**:825.

106. Safikhani Z, Freeman M, Smirnov P, *et al*. Revisiting inconsistency in large pharmacogenomic studies. *bioRxiv* 2015;026153.
107. El-Hachem N, Gendoo DM, Ghoraie LS, *et al*. Integrative pharmacogenomics to infer large-scale drug taxonomy. *bioRxiv* 2016;046219.
108. Smirnov P, Safikhani Z, El-Hachem N, *et al*. Pharmacogx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* 2016;**32**(8):1244–6.
109. Young WC, Yeung KY, Raftery AE. Model-based clustering with data correction for removing artifacts in gene expression data. *arXiv*, 2016.