Tampere University

ALIYU MUSA

Network-Based Systems Pharmacogenomics

Methods and Applications

Tampere University Dissertations 302

Tampere University Dissertations 302

ALIYU MUSA

Network-Based Systems Pharmacogenomics

Methods and Applications

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion at Tampere University,
on Friday 2 October 2020, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Assoc. Prof. Frank Emmert-Streib Tampere University Finland	
Pre-examiners	Professor Des Higgins University College Dublin Ireland	Professor Sol Efroni Bar-Ilan University Israel
Opponent	Professor Marc-Thorsten Hütt Jacobs University Germany	

The originality of this thesis has been checked using the Turnitin Originality Check service.

Cover design: Roihu Inc.

ISBN 978-952-03-1682-2 (print)
ISBN 978-952-03-1683-9 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1683-9

PunaMusta Oy – Yliopistopaino
Vantaa 2020

Dedicated to my family and loved ones.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my PhD advisor, Prof. Frank Emmert-Streib for his continous support, patience, guidance and encouragement along the years of my study. I have learnt alot and have enjoyed every single moment working under his supervision. I would also like to thank a member of my thesis committee, Prof. Olli Yli-Harja for his valuable support, guidance and inspiration during my PhD work. Furthermore, I would like to thank my external reviewers, Prof. Des Higgins and Prof. Sol Efroni for spending time and effort to review my dissertation.

During the course of my PhD, I worked with Dr. Shailesh Tripathi, Prof. Matthias Dehmer, Prof. Stuart A Kauffman and Prof. Benjamin Haibe-Kains, I wish to thank them for their thoughts and insights, they have been truly appreciated and definitely boosted the quality of my research work. Prior to coming to the Tampere University, Tampere, Finland, I was fortunate to conduct research under Dr. Ricardo De Matos Simoes and Dr. Shu-Dong Zhang at Queen's University Belfast, Northern Ireland, UK, I want to thank them for endless guideline and technical assistance toward my early research career. To these individuals and to others who have mentored me during my educational journey, I offer my deepest gratitude and appreciation.

Finally, I would like to thank my parents for raising me to this level, my wife for unconditional love and support and my friends for contributing to a joyful experience.

ABSTRACT

Large-scale molecular perturbational data provide signatures that represent changes on a cellular state from a systematic exposition to drugs or other forms of perturbations. Resources like the *Library of Integrative Network-Based Cellular Signatures* (LINCS) enable the identification of signature profiles important, e.g., for drug repositioning or target discovery based on automatic similarity searches across a vast reference profile space. In this thesis, we investigated the LINCS L1000 data repository consisting of nearly 2 million gene expression profiles and additional meta-data. As main results, we obtained: (I) an overview of the characteristics of all available data sets, their interrelations and experimental conditions including specific drugs, cell lines, time points and dosages. (II) a web interface called L1000 Viewer for accessing selected subsets of data from LINCS needed for the experimental design of studies addressing particular questions. (III) drug association networks (DANs) representing relationships for all drugs and small molecules in LINCS. The DANs are very informative in gaining a genomic-scale overview of the relationships among all drugs (including FDA approved drugs) and small molecules and, hence, provide a systems pharmacogenomic drug landscape. Importantly, we assessed the structural connectivity of the DANs by using information from the Anatomical Therapeutic Chemical (ATC) classification of drugs. This allowed us to identify the DAN modules' as therapeutic attractors of ATC drug classes, extending the classic idea of cancer attractors in gene regulatory networks introduced by S. Kauffman to the compound space. In order to utilize our results, all DANs are available via an interactive web site allowing also the exploration of their structural complexity.

CONTENTS

ABBREVIATIONS

API	Application Programming Interface
ATC	Anatomical Therapeutic Chemical
CMap	Connectivity Map
DAN	Drug Association Network
DEGs	Differentially Expressed Genes
GEO	Gene Expression Omnibus
GO	Gene Ontology
JI	Jaccard Index
LINCS	Library of Integrated Network-based Cellular Signatures
MoA	Mechanism of Action
MODZ	Moderated $z-score$
ODM	Object-Data Modeling
XCos	eXtreme cosine method

ORIGINAL PUBLICATIONS

Publication I A. Musa, L. S. Ghoraie, S.-D. Zhang, G. Glazko, O. Yli-Harja, M. Dehmer, B. Haibe-Kains and F. Emmert-Streib. A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics* 19.3 (Jan. 2017), 506–523. ISSN: 1477-4054. DOI: `10.1093/bib/bbw112`.

Publication II A. Musa, M. Dehmer, O. Yli-Harja and F. Emmert-Streib. Exploiting Genomic Relations in Big Data Repositories by Graph-Based Search Methods. *Machine Learning and Knowledge Extraction* 1.1 (2018), 205–210. ISSN: 2504-4990. DOI: `10.3390/make1010012`.

Publication III A. Musa, S. Tripathi, M. Dehmer and F. Emmert-Streib. L1000 Viewer: A search engine and web interface for the LINCS data repository. *Frontiers in Genetics* 3.1 (2019), 7849. DOI: `10.1038/s41598-000`.

Publication IV A. Musa, S. Tripathi, M. Kandhavelu, M. Dehmer and F. Emmert-Streib. Harnessing the biological complexity of Big Data from LINCS gene expression signatures. *PLOS ONE* 13.8 (Aug. 2018), 1–16. DOI: `10.1371/journal.pone.0201937`.

Publication V A. Musa, S. Tripathi, M. Dehmer, O. Yli-Harja, S. A. Kauffman and F. Emmert-Streib. Systems Pharmacogenomic Landscape of Drug Similarities from LINCS data: Drug Association Networks. *Scientific Reports* 9.1 (2019), 7849. DOI: `10.1038/s41598-019-44291-3`.

Author's contribution

Publication I: This publication reviewed advanced computational methods that are used for drug repurposing, as well as for identifying the mechanism of action (MoA) and discovering the phenotypic relationships in gene-expression datasets. Moreover, we used a LINCS dataset for a case-study. The author of this dissertation was responsible for writing the manuscript and for carrying out the case-study analysis.

Publication II: This publication discusses the problem of accessing/querying selected subsets of pharmacogenomic-data from repositories, such as the LINCS, and describes how the lack of querying capabilities in current realizations could be compensated. Moreover, creating a smart interface in the form of a web application was discussed with respect to how it can efficiently eliminate difficulties in accessing "big data". Importantly, this idea is not limited to the LINCS database; rather, the LINCS database was selected due to its popularity, thereby emphasizing the need for the aforementioned search capabilities in related similar databases. Indeed, pharmacogenomic data require new data-accessing strategies; these can be provided by smart interfaces that allow for the sophisticated querying of relevant data sets. The candidate was responsible for presenting and writing the manuscript.

Publication III: This publication describes how we created the L1000 Viewer, a user-friendly web application platform. The L1000 Viewer is a search engine and web application catered towards the LINCS data repository. The web interface is an interactive platform for users to select various types of perturbation profiles; for example, specific cell lines, drugs, dosages, time points and their combinations. In its core, the method is based on the intricate dependence graph structure (network) created from metadata information. Conceptually, this method implements the framework introduced in Publication II. It is accessible via (`http://L1000viewer.bio-complexity.com`). The author was responsible for developing the entire methodology, including the web interface and writing the manuscript.

Publication IV: In this publication, a summary for the distributional characteristics of gene-expression signature profiles for cell lines treated with drugs and small-molecule compounds was presented. Specifically, the LINCS L1000 data was analyzed according to two different layers. The first layer focused on the signature profiles themselves whereas the second layer focused on differentially expressed genes derived from the signature profiles. A linear-regression method revealed changes in the differential expression of genes, thereby demonstrating the biological complexity of the drugs tested. This analysis will help dealing with the overwhelming complexity of large datasets, by providing guidance for both experimental design and follow-up studies. The author was responsible for developing the entire research method, including implementations, performing the analysis and writing the manuscripts.

Publication V: In this publication, a novel approach to network-based systems pharmacogenomics was developed utilizing the connection between genes, drugs, and diseases provided by the LINCS repository. As a result, a Drug Association Networks (DANs) have been created. DANs systematically summarize the therapeutic effects of hundreds of known drugs and thousands of small molecule compounds. Importantly, we showed that the modular structure of the DANs reflect an enrichment of ATC classes, thus integrating the drug-induced changes with the genotype state of the cells meaningfully.

In addition, a web interface was developed and made available on the following website (`http://dan.bio-complexity.com`). The web application can be used to identify drug–drug interactions and can also connect to external links for a drug-target validation. Moreover, the features of the DAN user-interface allow users to browse and download relevant information. The candidate, as the first author, was responsible for implementing the method, performing the analysis and writing the manuscript.

1 INTRODUCTION

1.1 Overview

The accessibility of large-scale perturbation datasets has opened up new possibilities for pharmacogenomics research. However, these datasets are not without their challenges, such as a lack of annotation, storage, access and analysis standards. Unifying platforms are therefore necessary to integrate such datasets with relevant data analysis tools. For data integration purposes, such platforms should eliminate biases from various sources, such as batch effects, profiling platform differences, and cell-specific differences that characterize drug-induced effects. In addition, the platforms should be simple-to-use, which means that users should be able to create new methods for data mining and data manipulation. Despite the continuous generation of large-scale pharmacogenomics datasets, such as LINCS [1], they remain largely underused with respect to their analysis potential. Recently, many computational approaches have motivated researchers to develop network-based models and systems-biology approaches to obtain an in-depth understanding of the basic biological relationships between drugs and diseases [2]. Specifically, various methods have been developed with respect to identifying both druggable targets and drug compounds based on a basic understanding of biological processes at the pathway level. These include the following methods: (i) integrating functional protein-association networks to form a new model, (ii) finding information on a known target and enriched pathways with small molecules with high connectivity scores, (iii) investigating side-effect scores based on ranked gene signatures, and (iv) the use of novel methods from network-based studies to evaluate perturbation datasets [3, 4, 5, 6].

Indeed, a systematic and unbiased approach towards drug prediction is imperative with respect to efficiently classifying new compounds and inferring their potential reuse. The rapid accumulation of data in genomics prompts us to utilize the gene expression data available from public databases with respect to predicting and repurposing drugs. Indeed, integrative computational methods that mine said data sets are relatively fast and cheap; moreover, they can complement traditional methods of drug discovery by using the complementary information available in distinct resources to develop novel therapies. Several promising attempts have been made with respect to drug predictions on different cancer types; these studies use a large sample of gene expression data generated by the connectivity map (CMap) [7, 8]. Transcriptional profiles induced by drugs can be used to characterize biological effects, thus offering fresh techniques for detecting annotations of compounds as well as drug-drug similarities based exclusively on gene expression profiles [9]. Since drug-induced transcriptional profiles can be presented as gene signatures (a set of differentially expressed genes can be obtained by comparing gene expression levels in samples from two distinct disease states or at two distinct times or conditions), they can be used to discover novel drug associations, disease therapies and pathways [7, 10]. While these efforts have certainly enabled researchers to make great strides in characterizing drug categories, yet, determining the consistency of such efforts in predicting new/uncharacterized small molecules still remains a persistent challenge.

Resources such as the CMap and the LINCS catalogue the transcriptional responses to drug treatments in human cell lines for thousands of small molecules that can act as rational drugs. For instance, the CMap introduced in 2006 by the Broad Institute was adapted by several research groups for the purposes of analyses and for developing novel applications with respect to drug discovery and understanding disease [11]. The success of the CMap has motivated its use in identifying new therapeutics by finding drug targets, as well as in identifying possible connections between diseases, genes and drugs [12]. Indeed, the CMap aids researchers in finding new chemical forms of drugs, as well as in predicting possible drug candidates and the pharmacological and toxicological properties of chemicals. The CMap approach has been effectively used to define novel therapeutics for a variety of factors, including multiple cancers and, most recently, inflammatory bowel disease [13] and muscle atrophy [14]. Moreover, new data has been made available in the LINCS according to the same underlying concept as the CMap; however, the size of this data base is much larger than that of the CMap. Using these novel databases, researchers can develop and apply high-dimensional machine learning and statistical methods that can be used to investigate the high-throughput genomics of said datasets (that is, “big data”) with respect to studying drug-related problems for personalized medicine [15].

Over the previous 20 years, technological advancement in high-performance measurement assays, for instance, next-generation sequencing, has led in an enormous increase in genomics data generation capacities. As a result, there are many databases available that provide millions of RNA sequencing datasets, metabolic data, gene expression, protein interaction, and protein structure [16, 17]. These datasets, however, require extensive analysis to find the relevant information, which is problematic since accessing selected subsets is not easy due to the sheer vloume and complexity of connections between different data sets. Unfortunately, most current databases do not provide efficient organizational structures nor interfaces that enable direct access and/or provide relevant summaries. This problem is discussed by focusing on the LINCS, a pharmacogenomic database [7, 18, 19, 20, 21].

In order to rectify this problem, large-scale databases require smart interfaces. By “database”, we mean both the data set together with the database management system; by “smart interface”, we mean that, in addition to a graphical-user interface (GUI), an analytics component is required to analyze the data. In our case, these smart interfaces exploit the connectivity structure between individual data files (see Fig. 1.1 A) by generating a representative network between them (see Fig. 1.1 B). In the instance of LINCS, the connections are established by combining cell lines, drugs and dosages (among others) using gene-expression profile. These connections generally correspond to the various attributes of the information files (metadata information). Once such a network is created, one can rapidly extract selected files using a query function. For example, the drugs D1 and D3 and the dosage Do3 correspond to the cell line C2 (see Fig. 1.1 B and its link back to the yellow data files), because each search combination is a hash with a list of data files connected with it. In this way, a smart interface forms a connection between data and data analysis (see Fig. 1.1 C); it also provides a GUI and data structure function that can efficiently access selected data files.

Indeed, new technology can be used to generate high-throughput genomics in large-

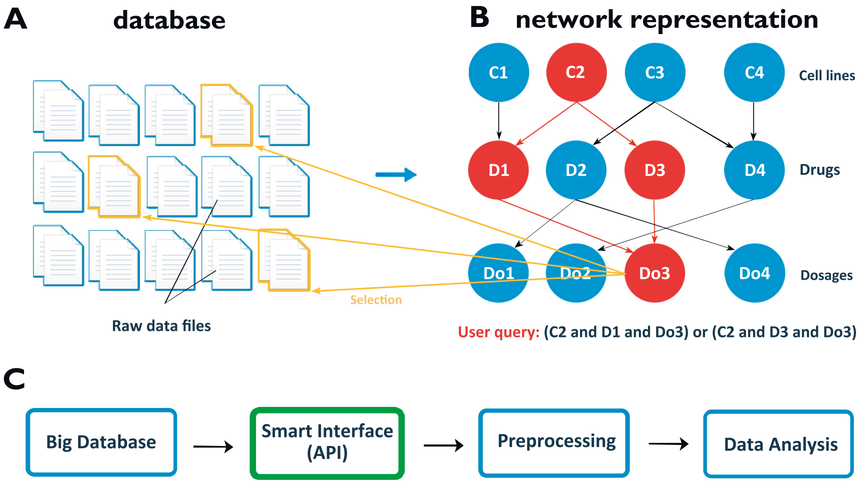


Figure 1.1 (A) A collection of available individual (raw) data files and metadata information from the original LINC data repository (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742`). Yellow indicated the selected files; (B) Network representation of connections between the raw data files (metadata information for either cell line, drug, dosage or time point can be selected). A user query (in red) corresponds to one particular combination of attributes of the data files, which leads to an efficient selection of files of interest from user quary (in yellow); (C) Conceptual integration of the smart interface, with an application programming interface (API), into a conventional data analysis pipeline.

scale databases; however, it is essential to develop methods that can be used for the curation setup by using computational approaches to implement a web interface for data visualization. Considering the problem of high dimensionality and data complexity, it is also necessary to build a statistical procedure to transform the set of large features of possibly correlated variables in the data into a set of values that can be used for network visualization. Obviously, developing a new technique for defining and categorizing drug classes is essential considering the current availability of relevant data; this can be accomplished through computational analysis and biological hypothesis. Therefore, we think a network-based systems representation is the way to go.

1.2 Objectives of the thesis

In recent decades, the ability to process large data sets has increased dramatically due to the development of both publicly available databases and "big data" analysis technologies. For example, the CMap and the LINCs were designed to characterize the mechanism of action (MoA) behind various drugs. However, these databases are severely underutilized due to the complexity of the provided data structures and the gene-disease relations. This is especially true when using perturbational data for different drug applications

In order to address this problem and to enhance the usability of LINCS we developed various methods. First, we obtained a general overview of LINCS by analyzing the general characteristics of the contained information, including different cell lines, dosages and perturbations (see Publication VI). Then we developed a web application, called L1000 viewer, for selectively accessing LINCS data (see Publication II and III). This was necessary because LINCS does not provide a selective access to raw data files. Finally, we developed a network-based method to summarize all information contained in LINCS. This leads to a representation of the pharacogenomic landscape in the compount space in the form of a drug-association network (DAN) (see Publication V).

The objective of this thesis is to harness the biological complexity of "big data" to potentially connect disease-therapeutic drugs by making use of the large high-throughput expression datasets provided by the newly created LINCS L1000 project [15]. The main result is a drug association network (DAN) that summarizes the entire information contained in LINCS whereas structural models of the network consist in enriched ATC classes. We also develop a web-application tool for visualizing the results; it can be used to interactively explore the interactions of different drugs and obtain pharmacological information from the linked data resources.

1.3 Thesis outline

The thesis is outlined as follows; Chapter 1 introduces the wider research area and provides a general overview. Chapter 2 discusses the details of the used datasets and explains external data resources. Chapter 3 presents a general review of the CMap methodology and applications. This chapter summarizes also recent developments of the CMap that have been used to identify drug-target interactions and disease states. Chapter 4 discusses the main results obtained in the thesis, specifically the DAN network, the meta-analysis and the development of new web-based applications. Finally, Chapter 5 presents conclusions.

2 DATA

The L1000 transcriptomics assay developed by the LINCS Center for Transcriptomics explicitly measures 978 landmark transcripts from pure cell lysates using a ligation-mediated amplification technique combined with Luminex detection technology on numerous well plates [15]. Because of the extremely correlated gene expression composition, the unmeasured transcriptome is computationally inferred from the chosen landmark genes to produce a robust reconstruction using an algorithm created in accordance with a big collection of full transcriptomes. In addition, for scaling and normalization, 80 invariant genes are evaluated. This technique is demonstrably similar to RNA sequencing (although much cheaper) and has generated more than one million cellular perturbation profiles that are accessible to the public via `https://clue.io`. Please refer to Publication III–V for in-depth details about the variables and the subsets used in the analysis.

2.1 Metadata information

The LINCS L1000 provides extended metadata specifications that describe reagents, assays and experiments. These includes annotations for the perturbagens (small molecules, siRNA, growth factors and other ligands) and cells, as well as for some of the elements of the experimental metadata. The LINCS data API provides a programmatic pipeline for the annotations and perturbational signatures in the L1000 dataset via a collection of HTTP-based RESTful web services; for example, “cell service” describes the cell line meta-information.

2.2 Gene expression data: Intensity values

The LINCS data is generated by using the LINCS Data Signature Generation Centers (DSGC) and is made accessible for download by the LINCS Data Coordination and Integration Center (DCIC) via the data portal [22]. Using R packages, we created a data pre-processing pipeline composed of functional scripts for analyzing relevant data and directly accessing files on Secure File Transfer Protocol (SFTP) servers. For all "gct" and "gctx" files (provided by the LINCS), an additional pre-processing step was performed using the "parse.gctx" function in the "cmapR" package [23] in order to extract the metadata and the intensity values.

Level-four plate-normalized data (for the L1000 datasets) was downloaded from the LINCS data portal [15]. The processed dataset packages are accessible through the LINCS data portal, including both original expression data and metadata [22].

2.3 Transcriptional signature profiles of perturbation

The LINCS L1000 profiled small-molecule and genetic-interference perturbation from transcription response. Gene expression was assessed only for the landmark genes to increase throughput, all of which were chosen for their ability to impute the expression of the remaining genes. Under a variety of conditions, a single disturbance was often tested, including cell types, dosages, time and concentrations [24]. Each condition produces a single up- or down-regulated z-scores signature. These signatures have been further processed to fit our approach.

Aggregation was performed by counting the amount of samples with *z-scores* greater than 1 for a specific gene. For samples using the same small molecule and the same cell line, gene expression profiles were aggregated (technical / biological replicates and/or distinct times of the small molecule). LINCS L1000 experiments are typically conducted with three or more biological replicates. A consensus of the replicated signature is derived by applying the moderated *z-score* (MODZ) procedure as follows [25]: First, a pairwise Spearman correlation matrix is computed between replicate signatures in the using 978 genes (landmark), not taking into account the trivial self-correlations (set to zero); The weights for each replicate are then calculated as the sum of the correlations between the replicated signature profiles, which are then normalized so that all the weights are equal to 1. Lastly, the consensus signature profile is indicated by the linear combination of the replicate signatures with the coefficients set to the weight values. If this count was more than 20% of the total number of samples for a particular gene, then said gene was included in the aggregated-expression profile.

Next, the above-aggregated gene-expression profiles are collapsed at the small molecule level: gene-expression profiles that correspond to the same small molecule across all cell lines are aggregated to produce the transcriptional signature profiles. Here, a gene is included in the final profile if it is up- or down-regulated by a $z - score$ larger than 1 in more than 30% of the cell lines treated by the same small molecule.

This procedure serves to mitigate the effects of uncorrelated or outlier replicates; it can be thought of as a “de-noised” representation of the given experiment’s transcriptional consequences.

2.4 External data sources

The L1000 small molecules were tested in several cell lines, experimental replicates, doses and time points. For this reason, we mapped DrugBank compounds and direct-measured genes to calculate a single transcriptional profile for each L1000 small molecule (concensus signature), across multiple signatures. We have also linked L1000 small molecules to various database sources in the UniChem database. This was achieved by querying UniChem with each L1000 small molecules’s InChIKey via UniChem API. It enabled us not only to map the small L1000 molecules to DrugBank, but also to PubChem, ChEMBL and KEGG Ligand databases (see Table 2.1). The pipeline also allows us to identify and map FDA-approved drugs to the small molecule identifiers of L1000 data.

Table 2.1 List of five small molecule/compound databases available in UniChem data source.

Source name	Description	% of drugs mapped
PubChem [26]	A database of normalized PubChem compounds (CIDs) from the PubChem Database.	89.34
ChEMBL [27]	A database of bioactive drug-like small molecules and bioactivities abstracted from the scientific literature.	87.47
KEGG Ligand [28]	KEGG LIGAND is a composite database consisting of COMPOUND, GLYCAN, REACTION, RPAIR, RCLASS, and ENZYME databases, whose entries are identified by C, G, R, RP, RC, and EC numbers, respectively.	61.21
LINCS [15]	The LINCS DCIC facilitates and standardized the information relevant to LINCS assays as described in `http://lincsportal.ccs.miami.edu/SmallMolecules/`	93.45
DrugBank [29]	A database that combines drug (i.e. chemical, pharmacological and pharmaceutical) data with drug target (i.e. sequence, structure, and pathway) information.	99.62

3 REVIEW OF METHODS

In this chapter, a review of methods and applications that use the LINCS L1000 datasets are introduced and described. Furthermore, the CMap and alternative methodologies are also described. A detailed review of the overall computational approache and the design of the CMap is presented in Publication I [20].

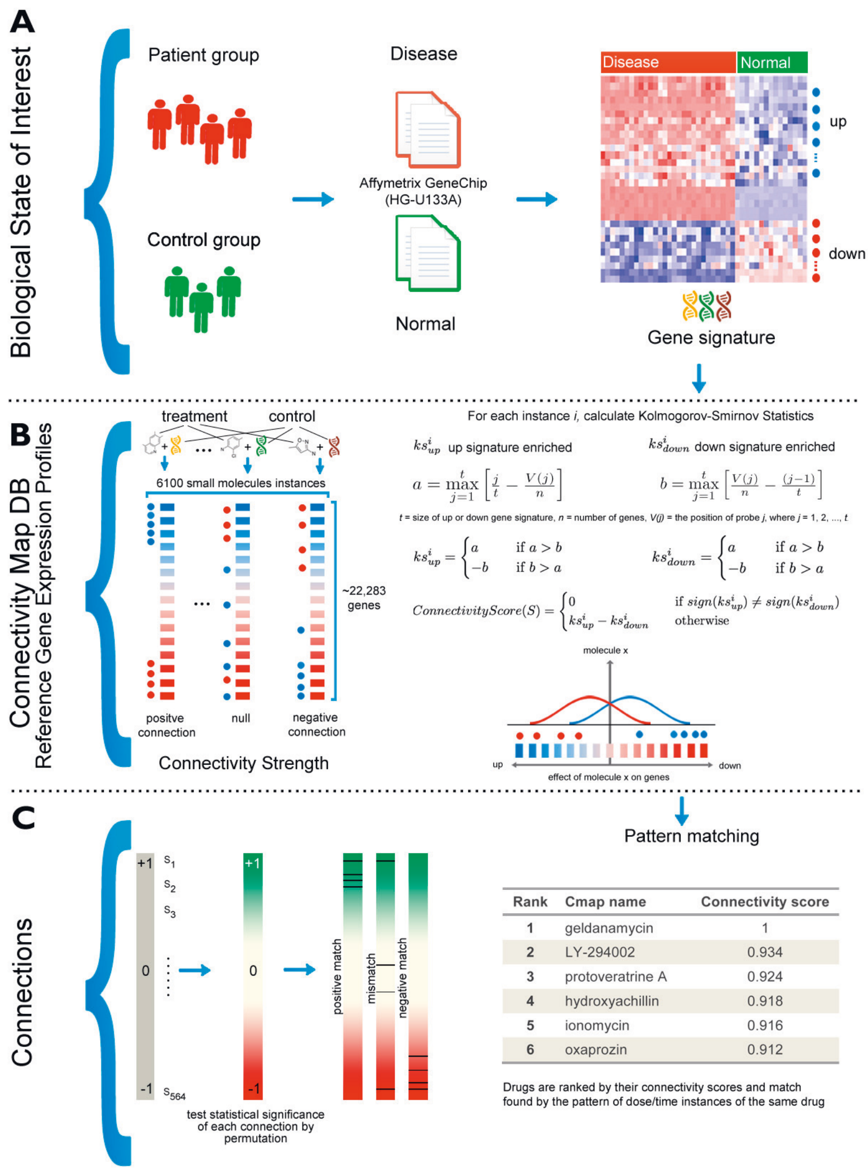
Gene-expression profiling offers insight into the overall measured transcript levels of any particular cell, tissue or organism at a specific point of time under the experimental conditions [30, 31]. Typically, these kinds of research strive to create an understanding of the regulatory networks of differentially expressed genes that may involve mediating biological mechanisms or pathogenesis of diseases. They also strive to define genes that show patterns of expression that correlate with a specific phenotypic trait or reaction to a specific disturbance, identifying molecular markers in the process [31]. These studies are indeed essential because they can be used to diagnose disease and predict clinical outcomes, as well as to identify new potential targets for therapeutic treatment and drug discovery [32]. While genome-wide gene expression studies are influenced by different biological complexities, experimental designs, technological and analytical difficulties, these methods are extremely common with respect to examining the biological processes [33]. In this chapter, we will provide example of methods and applications used to influenced the biological perturbations datasets at a transcript level to find similarities in drugs for further validation.

3.1 Methods and applications

Currently, chemotherapy is the standard treatment for the majority of disseminated malignancies. Recently, however, academics are becoming more and more interested in the prospect of molecular interventions in identifying drug targets for disease treatment [34, 35]. Indeed, applying computational modeling based on biological information is useful, since it can extend our knowledge of the connection between genes, drugs and diseases to improve the precision of our predictions [36, 37, 38]. Models used to simulate cellular or biological processes can provide accurate data and novel hypotheses; moreover, they can translate information between *in vitro* screening, cell-based assays and, eventually to patients. The introduction of the CMap in 2006 by the Broad Institute made this type of modelling popular among researchers with respect to drug discovery and understanding disease, as well as with respect to developing new therapeutics by identifying drug targets and the possible connections between diseases, genes and drugs.

Recently, the computational screening of drugs has been facilitated by the advent of the CMap [7]. The CMap is a comprehensive (and regularly updated) database containing the transcriptomics profiles of many existing small-molecule compounds. The CMap provides a simple (yet important) platform by employing a pattern-matching strategy to determine similarities between the connections of gene signatures among diseases, drugs and groups of genes. It has been used in many studies with respect to discovering treatments for common diseases, such as treating solid tumors, including those associated with various types of cancer; e.g., colon cancer [39, 40], breast cancer [41], and lung adenocarcinoma [42].

The concept of the CMap in drug-discovery studies is based on the identification of disease-associated gene signatures that indirectly correlate with perturbations in transcriptomic signatures associated with the administrated molecules or drugs [43]. In several studies, the procedure of using the CMap approach to identify drugs for treating disease is relatively straightforward [2]. Firstly, find a set of differentially expressed genes obtained by comparing the expression levels of genes in samples drawn from two different tissues, or at two different time points or under two different conditions. Secondly, score the match between the DEG set and the genomic profiles



For each instance *i*, calculate Kolmogorov-Smirnov Statistics

ks^i_{up} up signature enriched

ks^i_{down} down signature enriched

$$a = \max_{j=1}^{t} \left[\frac{j}{t} - \frac{V(j)}{n} \right] \qquad b = \max_{j=1}^{t} \left[\frac{V(j)}{n} - \frac{(j-1)}{t} \right]$$

t = size of up or down gene signature, *n* = number of genes, *V(j)* = the position of probe *j*, where *j* = 1, 2, ..., *t*

$$ks^i_{up} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } b > a \end{cases} \qquad ks^i_{down} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } b > a \end{cases}$$

$$ConnectivityScore(S) = \begin{cases} 0 & \text{if } sign(ks^i_{up}) \neq sign(ks^i_{down}) \\ ks^i_{up} - ks^i_{down} & \text{otherwise} \end{cases}$$

Rank	Cmap name	Connectivity score
1	geldanamycin	1
2	LY-294002	0.934
3	protoveratrine A	0.924
4	hydroxyachillin	0.918
5	ionomycin	0.916
6	oxaprozin	0.912

Drugs are ranked by their connectivity scores and match found by the pattern of dose/time instances of the same drug

Figure 3.1 Mechanistic overview of the working principle of the CMap method and the CMap database for drug discovery. (A) Gene-expression profiles derived from the treatment of cultured human cells with a large number of perturbagens populate a reference database or any given biological state of interest to obtain a query signature. Gene-expression signatures represent any induced or organic cell state of interest. (B) Shows Kolmogorov-Smirnov test algorithms for connectivity score of each reference profile for the direction and strength of enrichment with the query signature. (C) Perturbagens are ranked by this "connectivity score"; those at the top ("positive") and bottom ("negative") are functionally connected with the query state through the transitory feature of common gene-expression changes.

of the drugs calculated by the CMap, ranking the drugs according to their scores. Finally, choose the candidate drugs by selecting those with the highest scores. The CMap is beneficial since it exploits the entire genomic information of the biological state as well as that of the drug in question. However, it also has many drawbacks, as mentioned in the existing literature [44]. Specifically, it ignores the modified states of biological functions that share links with the disease being studied. Moreover, it ignores individual biological functions; for instance, it would ignore a high scoring drug that has beneficial effects on a subset of functions at the expense of harming other functions [45].

3.2 The connectivity mapping

3.2.1 CMap: The connectivity map

The original connectivity map was introduced by Lamb et al. [7] in 2006. The basic concept of CMap is to utilize a reference database containing drug-specific gene expression profiles and compare it to a disease-specific gene signature. 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 Fig. 3.1. A phenotype of interest such as a disease or biological condition in a form of gene expression signature is generated, normally a set of genes that are representative and unique with the underlying phenotype. In [7] the gene signature corresponds to a list of differentially expressed genes, named h, that contain up- and down-regulated genes; see Fig. 3.1 A. The gene signature, h, is compared to the ranked probe sets of the treatment vs. control gene expression profiles that are ranked in descending order according to the fold changes of the probe sets. By splitting the gene signature, h, into two lists containing only up-regulated genes, $h \uparrow$, and down-regulated genes, $h \downarrow$, a so-called *connectivity score* is estimated via several auxiliary variables using a non-parametric rank-ordered Kolmogorov-Smirnov (KS) test; see Fig. 3.1 B, similar to the method introduced in [46]. The *connectivity score* is represented in Equation 3.1 - 3.5.

For each instance i, calculate KS statistics:

ks_{up}^{i} up signature enriched

$$a = \max_{j=1}^{t}\left[\frac{j}{t} - \frac{V(j)}{n}\right] \tag{3.1}$$

ks_{down}^{i} down signature enriched

$$b = \max_{j=1}^{t}\left[\frac{V(j)}{n} - \frac{(j-1)}{t}\right] \tag{3.2}$$

Here t = size of the up or down gene signature; n = number of genes, $V(j)$ = the position of probe j, where $j = 1, 2, ..., t$

$$ks_{up}^{i} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } b > a \end{cases} \tag{3.3}$$

$$ks_{down}^{i} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } b > a \end{cases} \tag{3.4}$$

$$ConnectivityScore(S) = \begin{cases} 0 & \text{if } sign(ks_{up}^{i}) \neq sign(ks_{down}^{i}) \\ ks_{up}^{i} - ks_{down}^{i} & \text{otherwise} \end{cases} \tag{3.5}$$

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, CMap approach can be used as a method of screening chemicals by matching the gene signature of a novel pertubagen against the reference profile [47, 48]. The chemicals sharing similar gene expression pattern, same activities or mechanisms can be easily seen. Conceivably, a highly representative phenotype-specific gene signature set, of pathological, genomic perturbations or induced with chemical is define as the key component of implementing CMap methods. These can be generated through 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 differentially expressed genes that are statistically significant and display an association with a given phenotype.

3.2.2 CMapBatch: A meta-analysis of drug response

Fortney et al. have recently adapted a parallel CMap approach across multiple gene signatures of a disease, called CMapBatch [49]. Specifically, instead of applying CMap to one individual gene signature, they apply it to multiple gene signatures for the same disease and then combine the resulting outcomes. For this reason 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[50, 51] addressed this issue by combining those different gene signatures *before* applying CMap [52]. However, Fortney et al. emphasize that combining gene signatures is problematic for strongly non-overlapping gene sets. This problem is avoided by CMapBatch.

Formally, for each gene 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 [53] assigning a consensus ranking of drugs for all the tested gene signatures. The Rank Product method was originally developed to identify differentially expressed genes for replicated experiments based on the ranking of the individual experiments. Basically, for each drug d in gene signature i for a total of s gene signatures, a rank product values of $RP_d = \Pi_{i=1}^{s}\big(\text{rank}(d_i)\big)^s$ is estimated. Here 'rank' gives the ranking (in increasing order) of drug d in gene signature i. This allows to obtain a ranking of all drugs based on their corresponding

RP_d values. By randomizing the ranks of the drugs in the individual gene signatures, p-values are obtained.

In [49], 21 signatures ($s = 21$) for lung cancer obtained from Oncomine have been studied. 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 drug for 21 individual gene signatures was 22, but for CMapBatch the overlap was 39 drugs. Furthermore, for 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 in inhibiting cancer growth better than the results of the original CMap. It shows promising results in scaling up transcriptional knowledge and significantly increases the hit percentage from 44% to 78% of the top ranked drugs. Moreover, the resulting drug hits were characterised *in silico* and showed slow growth significance in 9 lung cancer cell lines from the NCI-60 collection. 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 [54].

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

3.2.3 Connectivity score based on partial-rank metrics

The CMap accomplished a decent level of success in its applications but has a few set-backs. One of these is the failure to apply a comprehensive measure to validate the significance of a gene signature when queried against reference profiles [55]. To address this issue, an extension of the connectivity score was introduced by Segal et al. [56], where they employ partial-rank metrics for scoring CMap queries by accommodating a query order, in contrast to the Kolmogorov-Smirnov scoring, which uses a rank ordering of gene expression profiles in the target instance to induce an ordering of the query. The paper also provides an alternative inferential approach based on generating empirical null distributions that exploit the scope, and capture dependencies, embodied by the database for refinements of the scoring measure that proved to be more efficient [56].

Shigemizu et al. introduced a novel methodology similar to partial-rank metric, by using gene expression profile from applying CMap concept to identify candidate therepeutics for MoA, targeting possible functions that are beyond drug repositioning [57, 58]. The method uses drug candidates in a pool of compounds that down-regulate the over expressed genes, or up-regulates the under-expressed genes, for a given abnormal phenotypic condition and demonstrate the utility of their approach for drug repositioning. The authors also stressed 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 [57, 59]. Finally, the result of their functional analysis indicated the MoA of tamoxifen is given by down-regulating TGF-β signaling [57].

3.2.4 ProbCMap: Probabilistic drug connectivity mapping

A probabilistic connectivity mapping by [60] 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 the drug as a probabilistic latent factor derived from the data on cell lines. The benefits of the method compared to other approaches are demonstrated for finding functional and chemical similarities of drugs based on transcriptional response profiles. It has also been showed how gene expression response factors between cell lines are the promising when a multi-source probabilistic model is used.

Furthermore, the method outline how probabilistic model-base approach can be extended to allow retrieval of combination of drugs. It showed how set of drugs retrieval provides complementary information when compered with single-drug, which is important in pharmacology and drug discovery, since the drugs have multiple mechanisms of action [61]. Considering the drug similarity validation with the CMap data, the probabilistic connectivity mapping provides a promising alternative. Moreover, the method could be applied to matching known drugs and drug combinations to disease samples, providing in this way novel hypotheses about therapies.

The LINCS dataset [12] comprises of generated data over tens of cell lines, the authors expect other benefits of the Group Factor Analysis-based probabilistic connectivity mapping used to become even more valueble. Being able to identify both shared responses across a large number of cell types, and on the other hand responses specific only to few cell lines, will be highly useful in drug development and discovery. It would be even possible to impose more structure on the Group Factor Analysis model, inferring which cell lines response similarly to the drugs, providing potentially highly relevant information for personalized medicine approaches.

3.2.5 New cosine-based similarity method compared with Kolmogorov-Smirnov statistic used in CMap

In this novel CMap approach, Cheng et. al. uses the Anatomical Therapeutic Chemical (ATC) classification as the benchmark to measure the differences and similarities of eXtreme cosine method (XCos) to other CMap scoring methods, data processing methods, and signature sizes [62]. They used the comparisons to clarify parameter choices, that can be used as new methods for drug repositioning where the gold standard benchmarking datasets are more complicated. The performance score for each method was measured using the AUC (FPR=0.1 and FPR=0.01) in the early discovery phase. The AUC score was used to determine the compound classes which have robust expression profiles in CMap data, and also help to find the analytical approaches that are more accurate in evaluating the data.

Overall, the XCos similarity score, which simply measures the cosine similarity between two signatures, yielded better results when compared with the Kolmogorov-Smirnov (KS)-based CMap method. Moreover, the similarity measure of closely related signatures tends to rely on the genes that changes between the treatment and control samples. Cheng and his group also notice both XCos and KS predicted the same ATC codes when used with a low number of features, for instance, the top 100, however, XCos outperformes when used with a larger number of features (top 500).

3.3 Connectivity Mapping Transcriptomics Datasets

There are a number of valuable data sets and databases containing gene expression response profiles effected by chemical compounds that are publicly available. Hence, these provide information about the perturbation effects that drugs have on the transcriptomics level of a cell. In Table 3.1 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 [63], that provide also disease-relevant relationships with drug compounds and targets. In the following, we discuss the two largest general purpose drug perturbation databases CMap and LINCS L1000 in more detail.

3.3.1 CMap dataset

The CMap database consists of genome-wide transcriptional expression profiles of bioactive compounds from cultured cell lines. In the original CMap study [7], 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 [7] (Build 1). In Build 2 this has been significantly extended to 1,309 approved small-molecules applied to the same five human cell lines leading to over 7,000 gene expression profiles. Build 1 and 2 use an Affymetrix platform for generating the gene expression data. So far, several methods have been developed utilizing 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 the other datasets [49, 56, 64, 65]. Notably, Cheng et al. presented a systematic approach to quantitatively assess the performance of such methods [66]. Hence, this study can be seen as a benchmark approach to assess any new methodology in the future.

3.3.2 LINCS L1000 dataset

The Library of Integrated Network-based Cellular Signatures (LINCS) supported by the NIH, comprises 5806 genetic perturbations (e.g. single gene knockdowns or over-expressions) and 16,425 perturbations induced by chemical compounds (e.g.

drugs) [67]. To date, over one million gene expressions have been profiled and collected for this project using the L1000 technology [67]. The L1000 platform has been developed at the Broad Institute by the CMap team in order to facilitate rapid, flexible and high-throughput gene expression profiling at lower costs. Specifically, the L1000 technology measures the expression of only 1000 so called *landmark* genes, and the expression values for the remaining transcriptome is estimated by a computational model utilizing additional data from the Gene Expression Omnibus (GEO) [16]. A user-friendly access to the database is provided by the LINCS project webpage (`http://www.clue.io/`), which is a web-based application allowing users to browse and query the LINCS database.

In a very 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 dataset presents new challenges to computational systems biologists who aim to analyze and visualize Big Data.

Table 3.1 List of perturbation transcriptomics resources

Resource	Type	Readout	Perturbagens	Cells/Tissues	Signatures	Species	Access
L1000 [15]	Small molecules, biologics, shRNA, cDNA	L1000	42,080	77 (9 core)	473,647	Human	GEO - GSE92742
CMap [7]	Small molecules	Microarray	1,384	4	3,773	Human	GEO - GSE5258
Carcinogenome Project (CRCGN) [8]	Small molecules	L1000	500	4	5,996	Human	`https://clue.io`
DrugMatrix [68, 69]	Small molecules	Microarray	657	9	3,938	Rat	GEO - GSE59927
Fish CMap [70]	Small molecules	Microarray	51	24	55	Fathead minnow and zebra fish	GEO - GSE38070, GSE60202, GSE70807, GSE70936
Gene Perturbation Atlas (GPA) [71]	Genes	Mined from GEO	1,585	1,170	3,072	Human, mouse	`http://biocc.hrbmu.edu.cn/GPA`
DrugSig [72]	Small molecules	Microarray	1,309	NA	5,997	Human	`http://biotechlab.fudan.edu.cn/database/drugsig`
Open TG-GATEs [73]	Small molecules	Microarray	170	2	1,483	Human, rat	`https://dbarchive.biosciencedbc.jp`
DRUG-seq [74]	Small molecules	RNA-seq	433	1	3,464	Human	GEO - GSE120222
ENCODE [75]	shRNA	RNA-seq	421	5	668	Human, fly	`https://www.encodeproject.org`
CREEDS [76]	Small molecules, genes, diseases	Microarray	1,475	2,513	3,879	Human, mouse, rat	`http://amp.pharm.mssm.edu/CREEDS`

4 RESULTS SUMMARY

In this chapter, the findings of Publication III–V are summarized. First, we describe the web application developed in this study (L1000 Viewer) [21], including the database, and the back-end and front-end implementations. Secondly, we present the results of a collection of gene-expression profiles and metadata, which includes many experimental samples covering more than 70 human cell lines [77], all of which were derived from the signature profiles. Lastly, we show the results for the systematic organization of the drugs and small compounds available from the LINCS repository by constructing the DAN [78].

4.1 L1000 Viewer: A Web Interface for the LINCS Data metadata information

To facilitate access to raw data subsets from the LINCS data repository, the L1000 Viewer was developed, an interactive web application that does not require the installation of dedicated software, but can be used with any web browser on any operating system. Furthermore, for each LINCS repository data file, our web application provides a web interface with access to a dedicated database that we created using the graph dependency framework. The dependency structure is specifically organized according to the experimental conditions of the expression profile; it can be depicted as a graph or network [79]. In the network, nodes represent information files; two data files are connected if they share experimental conditions. In order to query the data, the application provides a user-friendly and easy-to-use platform for choosing subsets of raw data files from specific types of perturbation profiles; e.g., for specific cell lines, drugs, dosages and time points. This technique of profile retrieval is efficient and fast because pre-computed graph structures (in data records) already exist. This provides valuable information for the user regarding the experimental design [80] of follow-up computational pharmacogenomics studies based on these data.

4.1.1 Development of the web application

L1000 Viewer, the web application we have developed, consists of three main parts namely; the database, back-end, and front-end implementations. First, in order to store the data in the back-end, we use a MongoDB database [81, 82]. We convert and store all the raw data into a json object structure to enable identifier reference to each profile sample in the database. This enables the data to be stored as a document-oriented structure that allows fast user queries. The document-oriented model maps to the data objects in the application code in the back-end, making the data easy to work with. The MongoDB is a distributed database at its core, therefore, it enables a horizontal scaling, high availability and faster access. Second, for the back-end component, we decided to use Node.js [83] for the server side architecture. A Node.js server environment was utilized to interact with the database through custom object-data modeling (ODM) calls adopted from pseudo relational database representation in Mongoose API [82]. Third, for the work-flow designer on the front-end we used javascript. Specifically, we use Vue.js [84] to created the front-end representation. Vue.js is a widely used javascript framework and the L1000 Viewer uses it for handling all client side user interactions. The connections between the components of the interface are implemented using Vue.js plugins. It provides a mechanism to display and render the structural components from HTML tags. To interactively display the large collection of drug-induced profiles, the HTML5 elements were used to layout the profiles systematically.

4.1.2 Graphical summary and visualization

In addition, we provide a functionality for an interactive visualization for viewing the selected profiles on the web. A user can click on the visualization button from the search results to visualize the selected profiles in different plots (e.g., boxplot representation of the profiles etc.). The metadata information of the selected profiles are also displayed. We provide R scripts for further metadata visualizations.

Specifically, we provide scripts that allow the user to generate graphical summary statistics of their metadata query results. From the download function, the user can immediately download the profiles and use the R scripts on the subset of the data that was retrieved.

4.1.3 L1000 Viewer accessibility

Access to the data indexed by the L1000 Viewer is provided through our web interface via `http://L1000viewer.bio-complexity.com/`. It enhances the biomedical data repository by providing a simple and fast access to LINCS raw data and allows to easily generate subsets of data. In this way, users of the web interface can extract knowledge more efficiently when interfacing with LINCS data.

4.2 Characterization of biological complexity from LINCS gene expression signatures

The LINCS L1000 data is a vast collection of gene expression profiles and meta information that includes many experimental samples covering more than seventy human cell lines. These cell lines are populations of cells descended from an original source cell and having the same genetic make-up, kept alive by growing them in a culture separate from their original source [85, 86]. In the following, we analyze the LINCS L1000 data for two different steps. The first step focuses on the signature profiles themselves and the second step on the differentially expression of genes derived from the signature profiles. This means we are moving from overview distributions on a basic level to characterizations of the biological activity of the cell lines in dependence on multivariate conditions, as given by, e.g., the number of replicates or the duration of applied drug perturbations. Hence, this provides an understanding of the biological functions effected by the perturbations.

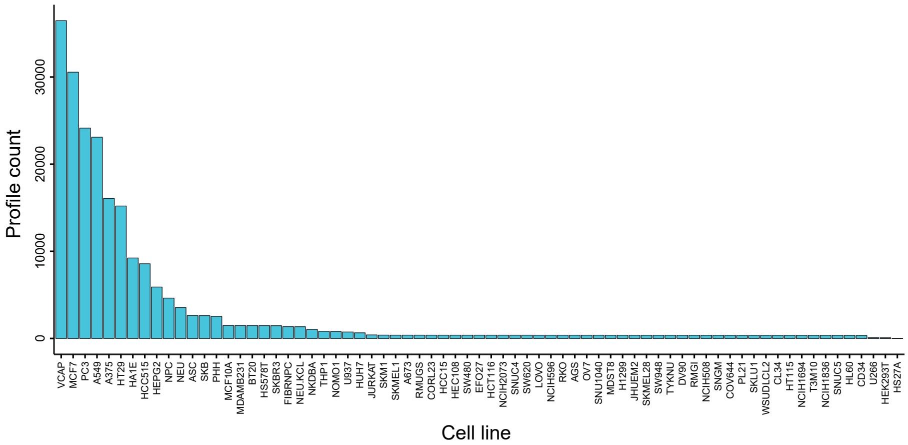


Figure 4.1 Cell line signature profile counts. The drug signature profile count distribution is shown for all 71 cell lines across all experiments in the LINCS L1000 dataset. Each bar gives the number of available signature profiles per cell line.

4.2.1 Transcriptional signature profiles

4.2.1.1 Cell line annotations

Various cancer cell lines and non-transformed primary cultures were used to represent disease models in the LINCS L1000 data [87, 88]. To enable an integration and analysis of large cell-based screening profiles, the cell lines were annotated with labeled terms to identify the associated organs and diseases. In Fig. 4.1 we show the overall distribution of profiled samples for 71 cell lines across all experiments. These counts include all the corresponding cell line profiles. Table 4.1 shows profile count and tissue origin for top 9 cell lines. For obtaining this information, we used the metadata annotations that are available via the Cell Service API [15, 23]. By summation over all cell lines in Fig 4.1 we find that, currently, the total number of signature profiles (excluding the profiles treated with knockdown and overexpression genes) is $215,224$. This number is much smaller than the 1.3 million raw gene expression samples because the replicated raw sample have been summarized for obtaining the signature profiles resulting from a comparison of treatment with control conditions.

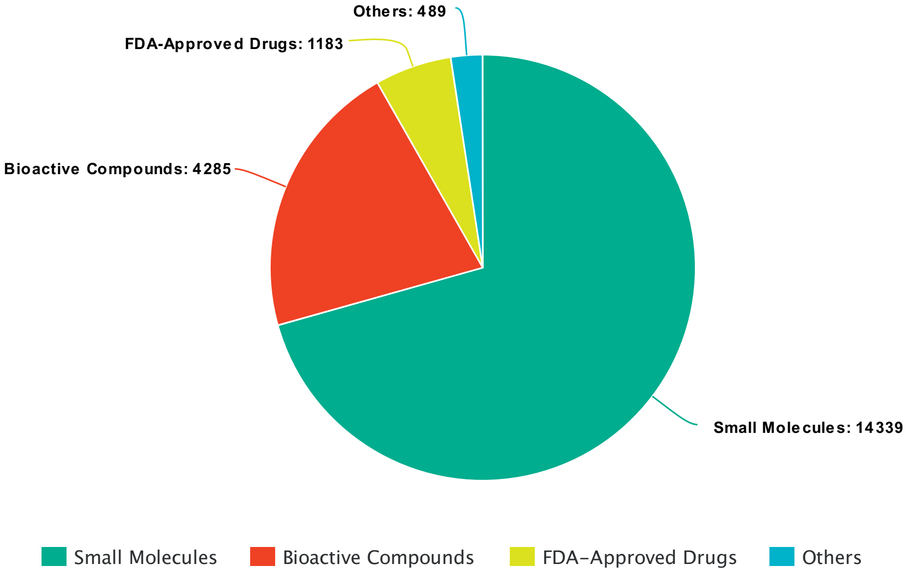


Figure 4.2 Distribution different small molecule used for all the signature profiles.

4.2.1.2 Small molecule annotations

The LINCS L1000 data include experiments for more than 20, 000 small molecule perturbations. The perturbations are applied to the cell culture to induce changes in the gene expression profiles. Furthermore, there are genetic perturbation experiments targeting single genes to control their expression levels, by either suppressing or enhancing them [89]. Detailed information for small molecule perturbations can be retrieved using the Pert Service API that identifies unique and common drugs used in the L1000 dataset. In Fig. 4.2 we show the distribution different small molecule used for all the signature profiles. The 6 experimental conditions considered are: controls, ligands, poscons, compounds, overexpression and shRNAs. The number of controls and compounds is always highest for all cell lines followed by the number of overexpressed profiles.

4.2.1.3 Experimental replicates

Experimental replicates have been investigated and found to be useful in simulation and in boosting analysis [90, 91] and decreasing the number of replicates will adversely affect the power of experiments [92, 93]. For this reason we studied the

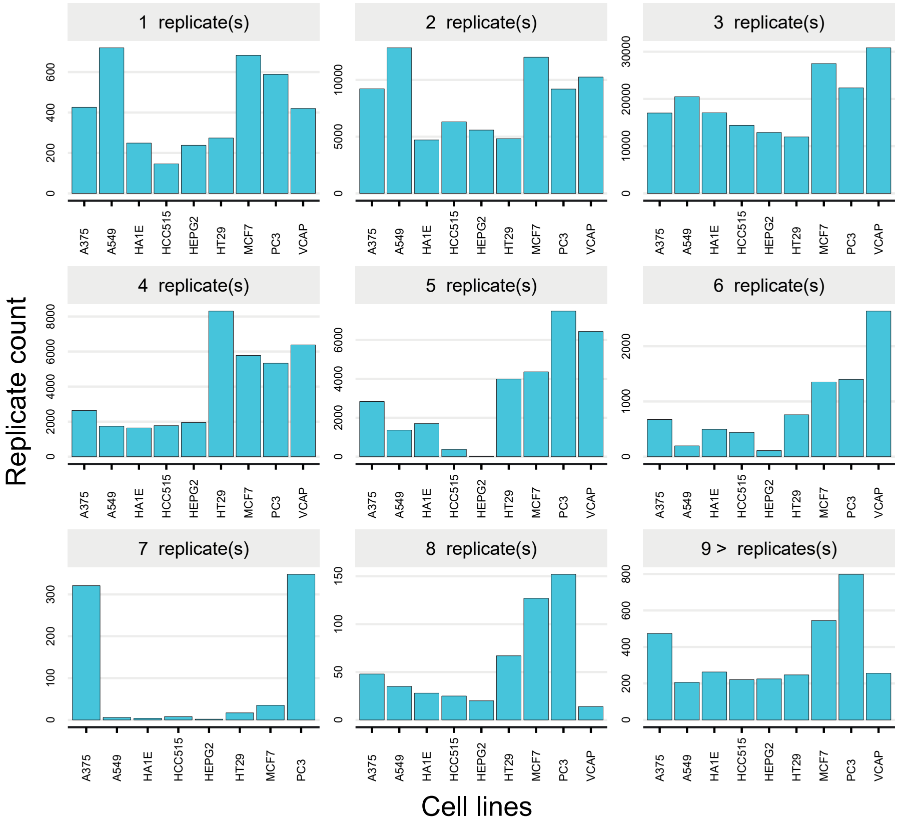


Figure 4.3 Distributions of experimental replicates for the signature profiles. The number of available replicates is shown for small molecule treatments in the LINCS L1000 data for 9 highly profiled cell lines.

distribution of replicate experiments of the LINCS L1000 data. From this we find that the plate variation is ranging mostly between 1 to 8 replicates with the majority of samples having 3 replicates. There are also conditions for which more than 9 replicates have been generated, however, these are rare covering only 1% of all profiles, whereas 1 to 8 replicates cover 99%. The largest number of replicates observed is 27, e.g., found for cell line VCAP, drug Vorinostat, a dosage of 10um and a time duration of 24h. In Fig. 4.3 we show the number of replicated experiments cross the 9 selected cell lines. The figure includes also information about 9 or more replicates and shows that the availability various greatly between the cell lines.

Table 4.1 List of 9 selected cell lines with the highest number of profiles

Cell line	Profile count	Tissue
A375	33,656	Skin
A549	37,577	Lung
HCC515	23,714	Lung
HA1E	26,164	Kidney
HEPG2	21,032	Liver
HT29	30,449	Colon
MCF7	52,373	Breast
PC3	21,032	Prostate
VCAP	21,032	Prostate

4.2.1.4 Dosage and time point annotations

Next, we show in Fig. A.2 results for the number of different dosages (concentrations) applied to the 9 highly profiled cell lines. The figure shows distributions for 8 different concentrations and 9 or more concentrations. However, almost 99% of the treated samples are measured for 1 to 8 different concentrations. From the available 49,400 perturbations, most of them were tested for a duration of 6, 24, 48, 96 and 120 hours. Overall, the number of cell lines per compound represented in the treatments ranged from 1 to 8 different time duration points (see Fig. A.3). Around 99% of the perturbations affected at least one gene significantly in a single cell line after treatment with the varying number of time points.

4.2.2 Differentially expression of genes

4.2.2.1 Differentially expression of genes and small molecule diversity

Our next analysis focuses on the activity level of the gene expression data as quantified by differentially expressed genes. For this analysis we utilized the L1000 raw z-scores from the GEO repository and pre-processed these by using the R L1000 tools [23]. We utilized the signature meta-information in Signature Service API for selecting the same subset of 9 cell lines as in Table 4.1 (with highest signature counts across all cell lines). Here a signature for a small molecule is defined as a vector of z-score values, each representing differential expression of genes profile between small

Table 4.2 Summary of z-score signature profiles for DEGS between treatments and controls on the cell line subset

	Signature profile	**Small molecules**
No significant gene	24	19
At least 1 significant gene	158,030	19,957
At least 50 significant genes	58,739	15,714
At least 100 significant genes	23,867	8,211
Total	**158,054**	**20,009**

molecule treated samples and control samples. In total there are 169, 239 z-score signature profiles for the 9 cell lines that satisfied the well- and plate-based quality control. This signature profile subset comprises 20, 009 small molecules (out of 49, 400 perturbations) that were repeatedly measured between 1 to 8 times. To further simplify the data and the quality of the analysis, we selected 6, 24 and 48h time points. In total this leaves us 158, 054 signature profiles (i.e., any combination of the small molecule, time, and cell line) for our analysis. These signature profiles come from experiments that were carried out on 391 multi-wells, where 362 wells were used for treatment and 29 DSMO wells were for control vehicles.

In order to obtain the number of differentially expressed genes between treatment and control samples for each of the 384 plates we used the z-score signature vectors obtained from the Signature Service setting the z-score threshold to > 2.0 and < -2.0 for up- and down-regulated genes respectively. For measuring the signature type effects that have been shown to be robust in biological interpretations, we use the assigned z-score thresholds to measure the biological effects encoded in the gene expression data. We found that 19, 957 small molecules from 20, 009 that are used in 158, 054 signature profiles yielded at least one gene that is significantly differentially expressed when compared with the corresponding control samples. We further found that 15, 714 small molecules reveal significant differences for at least 50 genes, and 8, 211 small molecules are differentially expressed for at least 100 or more genes. Table 4.2 summarizes these results.

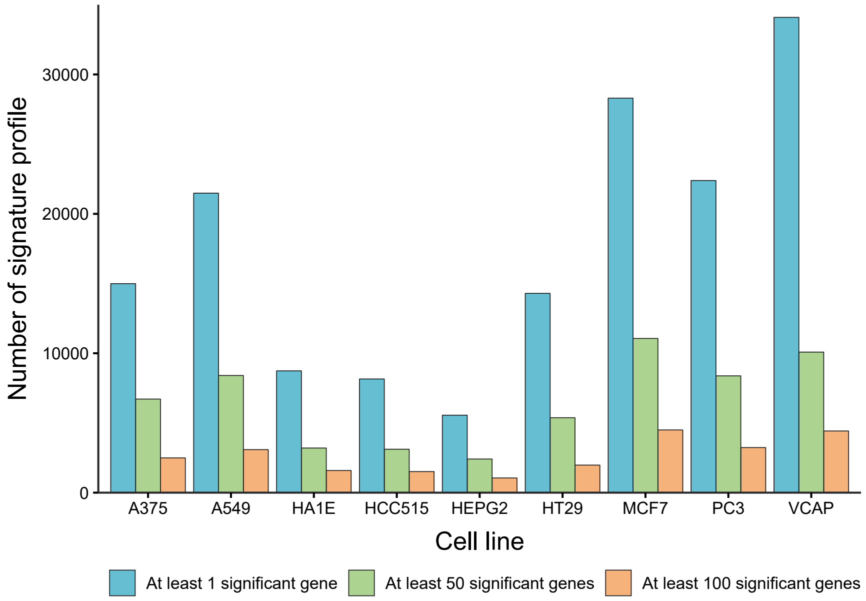


Figure 4.4 Number of significant profiles found when comparing signature profiles of treatment and control samples. The cell lines are categorized according to the number of DEGs and the DEG have been estimated based on the z-score signatures profiles.

4.2.2.2 Cell type specific differentially gene expression

Since not all cell lines measure the transcription effects of small molecules for the same time points, we subset the treatments according to cell lines and evaluate the number of significant genes for the 9 cell lines separately. In Fig. 4.4 we show our results giving the number of signature profiles for each cell line for three categories. The three categories correspond to (I) at least one significant gene, (II) at least 50 significant genes, and (III) at least 100 significant genes when compared with vehicle controls. Since there were only 24 profiles with no significant genes in total, this category is not shown in the figure.

4.2.2.3 Dosage specific differentially gene expression

For studying the effect of drug dosages we repeated a similar analysis as above. Specifically, we systematically classified the small molecule dosages into two categories for ‘low’ and ‘high’ concentrations. The ‘low’ concentration group contains all measurements in nanomolar (nM) and doses less than or equal to 5 micromolar (μ M) while the ‘high’ concentration group includes all measurements greater than 5 μM. In total, we find $63,113$ and $94,941$ signature profiles for low and high dosages respectively. In Fig. A.4, the number of differentially expressed genes is shown for the 9 cell lines and the two dosage categories. From this we observe two different behaviors. First, the number of differentially expressed genes increases with time, e.g., cell line A375 or A549. Second, the number of differentially expressed genes decreases with time. This behavior is only observed for cell line VCAP. The first type of behavior is expected because higher dosages of drugs should result in more severe changes in the expression of genes. The reverse of this effect for cell line VCAP, a prostate cancer cell line, averaged over all drugs is counter intuitive and points to follow-up investigations.

4.2.2.4 Drug Perturbation Specific differentially gene expression

Next, we analyze the number of differentially expressed genes according to the time duration of the treatment with small molecules. In Fig. A.3 we show results for 6 and 24 hours. From this we again observe two different behaviors. First, the number of differentially expressed genes increases with time, e.g., cell line A375 or A549. Second, the number of differentially expressed genes decreases with time, e.g., cell line HA1E or HCC515.

4.3 Drug similarities, prediction and network associations

Traditionally, pharmacology approaches focus on single drugs at a time to study their action, effects or safety [94]. This is similar to traditional molecular biology approaches that focused on single genes or proteins [95]. However, due to modern genomic high-throuhgput technologies, nowadays, it is possible to study many genes or proteins simultenously [96]. Pharmacogenomics and Systems Pharmacogenomics aim to utilize such genomic profiles to expand beyond single drugs [97]. For instance, in [98] drug-target and drug-drug networks have been constructed based on the DrugBank database utilizing information about FDA approved and non-approved drugs and their corresponding targets. However, their analysis focused exclusively on drugs and compounds with known targets and did not take into consideration dynamic activity profiles as represented, e.g., by transcriptomics data. In [99] some disadvantages were avoided by using gene expression profiles for which Pearson correlation-based networks were constructed. A problem is that the used data were generated from many independent, uncoordinated laboratories using varying platforms and samle preprations. Another drawback of this study is the small number of used profiles ($< 7,000$) and the very limited number of studied drugs (200). Similar data were used in [9, 100] but the construction of the drug network differed. Also, their analysis focused on drugs with known MoA. A different approach has been taken in [101] where a drug-drug network has been constructed only based on known side effects of FDA approved drugs. A drawback is the sole focus on negative clinical parameters, limitation to FDA approved drugs and the neglection of dynamical aspects of drug effects. In [102] in addition to gene expressin data also information about chemical structures and drug responses have been used. Unfortuantely, the number of drugs for which all three sources of data are available is very limited. A common shortcoming of all these studies is a lack of conceptual explanations of the drug networks.

The ultimate goal in pharmacology is to understand all properties, effects and actions of all drugs and componds [103]. Hypothetically, this information could be obtained from clinical trials testing each compound for every existing disease including subtypes and stages. From this information one could measure the similarity between different compounds, e.g., based on clinically relevant parameters. This would give the network structure of an ideal compound-space giving all relationships among all compounds corresponding to an ideal drug association network. Due to the practical impossibility of such an approach the question is, is it possible by using genomics data to approximate such an ideal drug association network (DAN)?

The goal of this thesis is to introduce a computational method that provides such an approximation leading to a systematic organization for the thousands of drugs and small compounds that are available from the LINCS repository. Specifically, we introduce a method for constructing DANs based on almost two million gene expression profiles for over $20,000$ chemical perturbagens and seventy-two human cell lines; see Fig. 4.6A. In these networks, nodes correspond to drugs and two drugs are connected if their profile responses are similar, as measured by the statistical sig-

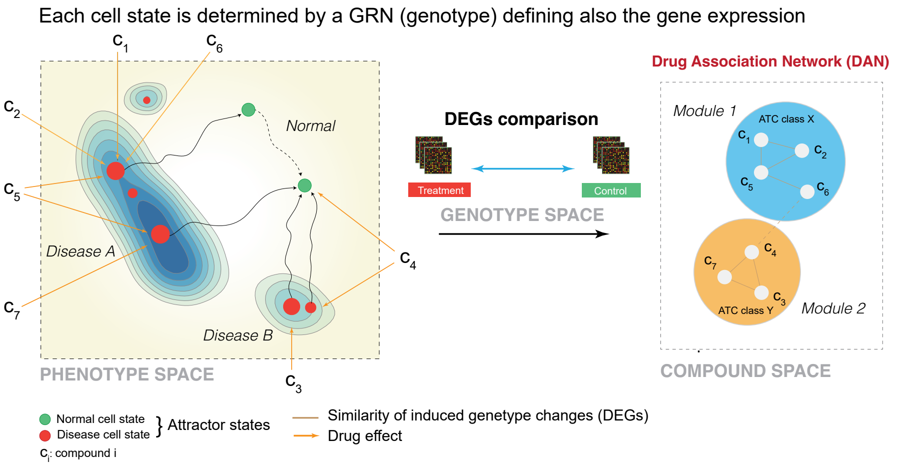


Figure 4.5 Conceptual connection between genotype space, phenotype space and compound space containing DANs. The phnotype space represent the compound C_i targeting disease/normal state i.e. drug effect (left). DEGs comparisons between treatment and control to obtain signature gene changes (center). DAN grouping of similar compounds that share similarities i.e. module representation (right).

nificance of the Jaccard Index (JI); see Fig. 4.7. The profile responses for each drug correspond to estimates of "consensus" signature profiles summarizing the transcriptional effect of drugs across multiple treatments on different cell lines and/or different dosages and time points. Overall, the DANs provide a systematic summary of the entire LINCS data repository and the complex pharmacogenomic landscape of drug similarities. For a conceptual overview see Fig. 4.5.

For obtaining pharmacogenomically meaningful networks, we construct different DANs based on data from different conditions. Specifically, we construct for each cell line a DAN using only the corresponding drug signature profiles. Furthermore, we construct one DAN limited to FDA approved drugs and one DAN for all drugs and small compounds (comprising FDA approved and non-approved drugs). This leads to condition-specific DANs (see Fig. 4.6B for their dependencies). In total, we are inferring 74 different DANs.

In order to analyze and interpret the DANs, we investigate the DANs on three different levels. First, we study the structure of the DANs by identifying network modules, also called communities [104, 105, 106]. This will allow us to gain insights

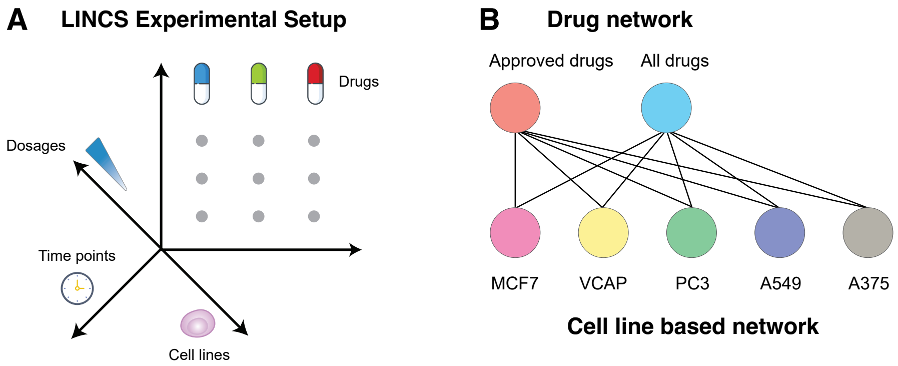


Figure 4.6 (A) Multifactural experimental space of the LINCS data, In total, nearly 2 million L1000 profiles from $42,080$ perturbagens are generated, corresponding to $25,200$ biological entities ($19,811$ compounds, shRNA and/orcDNA against $5,075$ genes, and 314 biologics) for a total of 473,647 signatures (consolidating replicates) taking at different doses and time points. (B) For our analysis we study 7 different DANs, the upper 2 networks are drug based networks and 5 networks are cell line based.

into the structural properties of the networks. Second, we study drugs pairwise by identifying the presence of significant ATC classes in the entire network. This analysis step will show that drugs with similar ATC classes are actually identified in compound space. Third, we study the enrichment of the network modules with respect to ATC classes. By using the ATC classification of drugs, we will demonstrate that the DANs represent a pharmacogenomic landscape of drugs summarizing the entire LINCS repository on a genomic scale.

As a general results, we will show that the ATC code enriched modules in the DANs can be seen as therapeutic attractors of drug classes. We will argue that this allows a conceptual extension of the idea of cancer attractors [107] introduced for gene regulatory networks to represent cell states [108, 109] to DANs representing pharmacological states.

4.3.1 Jaccard Index

Let D_k and D_l be two drugs with regulation profiles R_i and R_j. R_i and R_j are two vectors of length n, whereas n is the number of genes. Their components correspond

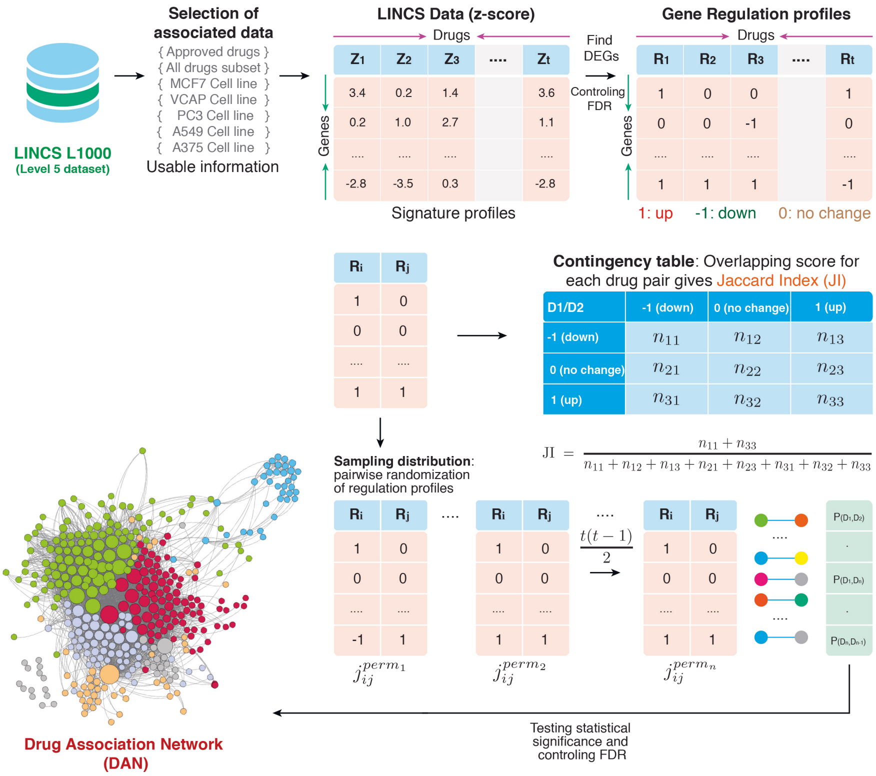


Figure 4.7 Overview of the construction of a DAN. The figure shows the gene expression profile signature of drugs and small molecule compounds from LINCS L1000 subset. Representation of the use of drug-feature matrices of different types to calculate drug connections using Jaccard Index (JI).

to (I) down-regulation (-1), (II) no-change (0) or (III) up-reguation (1). The JI can be estimated from the contingency table (see Table 4.3 and Fig. 4.7) giving the overlap between the two regulation profiles representing the effect of the drugs D_k and D_l:

$$J_{ij} \quad = \quad J(R_i,R_j) = \frac{\left\|G_i \cap G_j\right\|_{/||0,0||}}{\left\|G_i \cup G_j\right\|_{/||0,0||}} = \frac{n_{11}+n_{33}}{n_t} \qquad (4.1)$$

Here $n_t = n_{11}+n_{12}+n_{13}+n_{21}+n_{23}+n_{31}+n_{32}+n_{33}$ is the number of genes showing differential expression.

Table 4.3 Contingency table summarizing the gene regulation profiles R_i and R_j treated by drug D_k and D_l. Here n_{kl} are integer numbers giving the common genes in the categories $k, l \in \{\text{up}, \text{no change}, \text{down}\}$.

$D_i \downarrow / D_j \rightarrow$	-1 (down)	0 (no change)	1 (up)
-1 (down)	n_{11}	n_{12}	n_{13}
0 (no change)	n_{21}	n_{22}	n_{23}
1 (up)	n_{31}	n_{32}	n_{33}

4.3.2 Construction of drug association networks

The first network, we construct for FDA-approved drugs with assigned annotations in DrugBank [110, 111]. For this reason we call this network N_{approved}. In total, there are 1139 approved drugs in LINCS, however, only 381 have an ATC annotation. The drugs with DrugBank IDs are repeated in multiple experiments; therefore, the landmark genes have multiple z-scores from different experiments. We first average the z-scores for each drug from different experiments and use the consensus of the z-scores to construct the DAN, as described in the method section. From this analysis, we obtain a network with 381 nodes and 4251 significant interactions. From this network, we extract the giant connected component (GCC) having 367 drugs (nodes) and 4244 interactions (edges). In Fig. 4.8A, we show the distribution of JI of all significant interactions for this network from profiles having between 100 to 150 DEGs.

The second network we construct, we call N_{all}, is for all available drugs. In LINCS data there are in total 2505 different drugs applied in the different experiments (cell line, dosage and time point). For these, we construct a network with 2505 drugs and 86,585 significant interactions. From this network, we extract the GCC having 2451 nodes and 22636 interactions. In Fig. 4.8B, we show the distribution of JI of all significant interactions for this network from profiles having between 700 to 800 DEGs. The higher the value of the JI the more genes are commonly up- or down-regulated between two drugs.

Next, we construct 72 networks that are specific for the 72 cell lines. For our further analysis, we select from these 72 networks the five networks having the highest num-

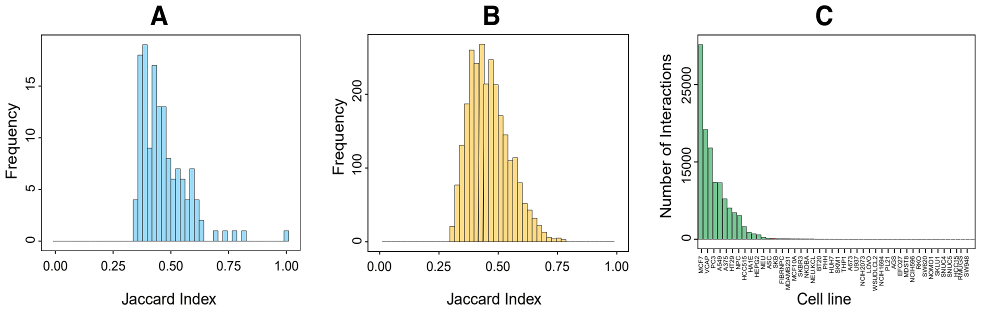


Figure 4.8 (A) Distribution of JI of all significant interactions for N_{approved} from profiles having between 100–150 DEGs. (B) Distribution of JI of all significant interactions for N_{all} from signature profiles having DEGs between 700–800. (C) Number of significant interactions between drugs for different cell lines.

ber of interactions between the drugs; see Fig. 4.8C for the frequency distribution of interactions for all cell lines. These cell lines are MCF7, VCAP, PC3, A549, A375. These 5 networks contain the most information, assuming interactions provide informative knowledge. The high number of interactions in each of these networks (more than $10,000$) ensures also that a sensible identification of modules is feasible.

In Table 4.4, we show a summary of these seven networks and their number of nodes and edges. All of these networks correspond to the GCC of the corresponding network. In the following, we will limit our analysis to these seven networks.

DAN	Used information	Drugs	Edges	Modularity	No. of Modules
N_{approved}	Approved drugs	367	4244	0.318	13
N_{all}	All drugs	2451	22636	0.554	20
N_{MCF7}	MCF7 cell line	750	7144	0.623	11
N_{VCAP}	VCAP cell line	520	2727	0.749	25
N_{PC3}	PC3 cell line	612	4314	0.644	17
N_{A549}	A549 cell line	380	2122	0.561	22
N_{A375}	A375 cell line	635	4286	0.636	14

Table 4.4 Summary of seven DANs constructed from different information. Shown is the information of the giant connected component. Column two describes the used information that characterizes the underlying data for each network.

4.3.3 Network modules in DANs

Our first analysis consists in the identification of the modules in the seven different DAN networks; see Fig. 4.6B. For this, we are using a multilevel community module detection algorithm [112] to find the modules in the networks. The modularity and the number of modules for each network are summarized in Table 4.4. We would like to remark that the number of the modules correspond to labels, i.e., the same label for different networks does not mean it should contain the same drugs. In general, we find the modularity to be similar among the different networks except for N_{approved} and N_{all} which is smaller. This is understandable considering the used data for these networks is different to the others. For the number of modules we observe similar values ranging from 11 to 25 modules. In Fig. 4.9, we show the networks for N_{approved} and N_{all} and the distribution of the number of drugs in the modules.

From the barcharts of boths networks one can see that there are a few modules containing a large number of drugs and the remaining modules contain only a few drugs. These large modules are also clearly visible in the network representation of the DANs on the left-hand-side in Fig. 4.9. In general, the modules in N_{all} are larger than in N_{approved} which is understandable because the former DAN contains 2451 nodes whereas the latter has only 367 (see Table 4.4).

4.3.4 Enrichment analysis of network modules

We performed an enrichment analysis of drugs with ATC codes for the modules detected in each network [113]. In order to test the statistical significance of ATC classes, we use Fisher's Exact Test [114, 115, 116]. Since we are testing multiple hypothesis tests for each module, we apply a Benjamini Hochberg correction to control the FDR. In the enrichment analysis we first find the total number of drugs in a module which are labelled with ATC codes and then we performed Fisher's Exact test to determine which ATC labels are overrepresented in a particular module. The results of this enrichment analysis are shown in Fig. A.1.

The summary of the enrichment analysis of the ATC groups for the modules of

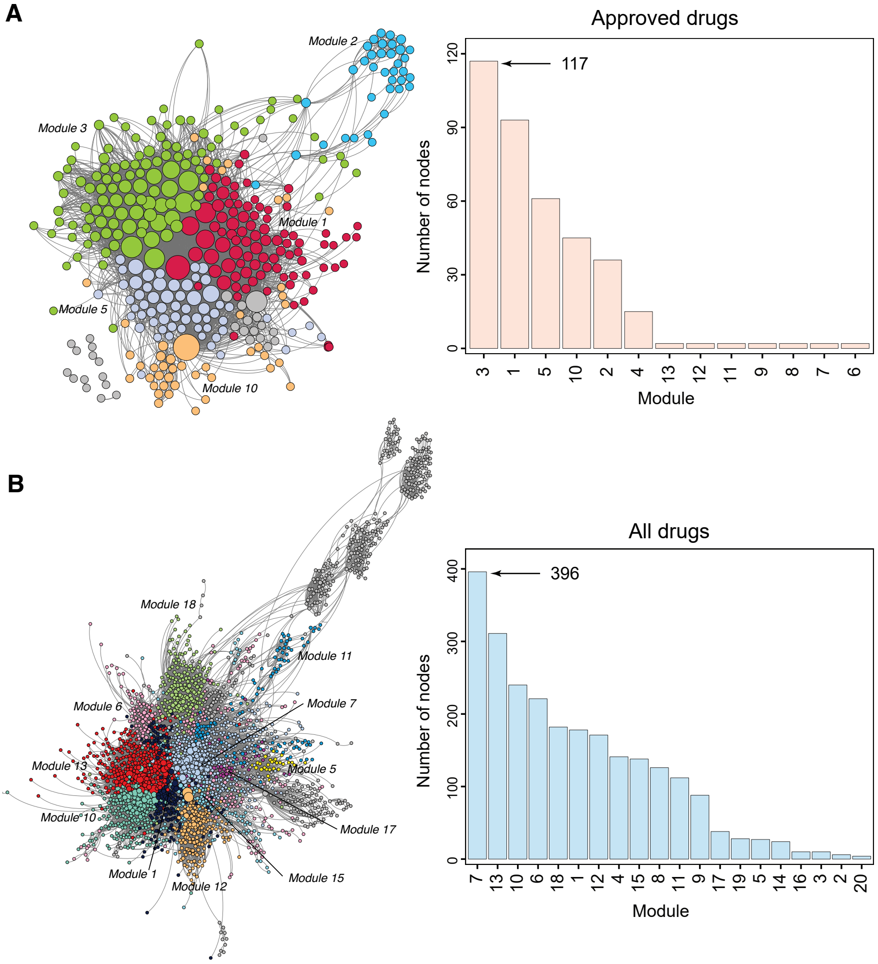


Figure 4.9 (A) Shows the network of FDA-Approved drugs with their corresponding module annotations (Left), and the number of nodes in each module of N_{approved} (Right) (B) The network show All Drugs including approved and non-approved drugs colored based on grouped module (Left), and the number of drugs in each cluster for N_{all} (Right).

the different networks is shown in Table 4.5. In this table, we highlighted the ATC groups which are enriched in at least one module in different networks. We also include those ATC groups which are not significant but holds low q-values between $0.05 < \alpha < 0.15$.

DAN / ATC code	C	D	G	H	J	L	M	N	P	R	S	**SC**	**SM**
Approved drugs		1	1		1			1	1	1	1	7	5
All drugs		1		1		1						3	2
MCF7 cell line						1				1		2	2
VCAP cell line												0	0
PC3 cell line	1	1	2			1			1		1	6	5
A549 cell line	1	1				1	1					4	4
A375 cell line	1					3						2	4
SM (all networks)	3	4	3	1	1	7	1	1	2	2	2		

Table 4.5 Summary of module enrichments shown in Fig. A.1 for all DANs. The columns show ATC classes highlighting if ATC codes are enriched in at least one module in the entire network (see Fig. A.1). SC gives the number of significant ATC classes and SM gives the number of significant modules per network. SM (all networks) gives the number of significant modules in all DANs.

4.3.5 Web interface for DAN of drugs

Furthermore, in order to communicate the wealth of our obtained results efficiently, we developed a web interface accessible at (`http://dan.bio-complexity.com`). Our web application allows to access the drug-drug interactions inferred by our method, and connecting to external links. The features of our DAN user interface enable searching, browsing, exploration and downloading of the network visualizations.

5 CONCLUSION

In this chapter, the contributions of the thesis to the existing literature are discussed, as presented in the Publications I–V.

The LINCS data repository [15], including the L1000 dataset, has been introduced to extend CMap [7] and improve its limitations. However, LINCs has a very intrictate structure of dependencies and is not straight forward to exploit. The general goal of this thesis was not only to show the utiliy of LINCS but to develop resources (methods and web applications) that can be further used by the community. This should allow to overcome at least some of the obstacles to enhance systems pharacogenomics resarch.

In Publication I [20], we reviewed and studied the CMap methodology and its applications. This showed that the CMap methodology provides an interesting view on how to predict disease-effects of a compound based on data from representative in vitro models. Hence, such techniques could be used for the computational analysis of new compounds. In general, the identification of drug targets plays an essential role in understanding the MoAs of various drugs and in designing new drugs.

In Publication II [79], we discussed the general problem of accessing/querying selected subsets of pharmacogenomic-data from repositories, such as LINCS, and described how the lack of querying capabilities in current realizations could be compensated. Furthermore, we discussed smart interfaces utilizing a graph-based file organization of the underlying data. Here it is a key insight that LINCS is not hierarchically flat but can be represented as a graph.

In Publication III [21], We introduced the L1000 Viewer - a search engine and graphical web interface for the LINCS data repository. At the core of our L1000 Viewer is a database that utilizes the intricate dependency structure among the files in the LINCS dataset, allowing these files to be reorganized as a graph and ensuring efficient search capabilities based on graph-oriented operations. Overall, the L1000 Viewer is a useful tool with respect to efficiently accessing selective information from the LINCS data repository for computational-pharmacogenomics studies [117, 118, 119], e.g., for drug repurposing and cancer therapeutics, as well as for understanding the composition and relationships between genes, drugs and diseases. Conceptually, the L1000 Viewer implements our ideas introduced in Publication II.

This thesis also demonstrated how "big data" from the LINCS project can be used to explore different experimental settings, such as cell-line coverage, time points and dosages, using a data pipeline to assess the compound-induced transcriptional effects. This is the content of Publication IV [77]. As a result, we provided summary statistics for the distributional characteristics of gene-expression signature profiles for all cell lines and their perturbagens. In doing so, we identified changes in the differential expression of genes, thereby demonstrating the biological complexity of the perturbagens. In this way, our analysis could help future studies for guiding their experimental design and for harnessing the overwhelming complexity of the LINCS data.

Finally, in Publication V [78], we developed a systems-pharmacogenomics approach and applied it to data from the LINCS repository. As a result, we constructed DANs that summarize hundreds of drugs and thousands of compounds with respect to their therapeutic effects. We demonstrated that the modular structure of the DANs represent enriched ATC classes, thus integrate the drug-induced changes on the genotype states of the cells. Interestingly, our results extend conceptual work conducted by S. Kauffman about *cancer attractors* in the epigenetic landscape of cell states to the compound space representing therapeutic interventions of which a DAN is a particular representative.

REFERENCES

[1] A. B. Keenan, S. L. Jenkins, K. M. Jagodnik, S. Koplev, E. He, D. Torre, Z. Wang, A. B. Dohlman, M. C. Silverstein, A. Lachmann et al. The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. *Cell systems* 6.1 (2018), 13–24.

[2] F. Iorio, J. Saez-Rodriguez and D. d. Bernardo. Network based elucidation of drug response: from modulators to targets. *BMC Systems Biology* 7.1 (2013), 139. ISSN: 1752-0509. DOI: `10.1186/1752-0509-7-139`. URL: `http://www.biomedcentral.com/1752-0509/7/139`.

[3] G. Laenen, L. Thorrez, D. Börnigen and Y. Moreau. Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Molecular BioSystems* 9.7 (2013), 1676–1685.

[4] N. S. Jahchan, J. T. Dudley, P. K. Mazur, N. Flores, D. Yang, A. Palmerton, A.-F. Zmoos, D. Vaka, K. Q. Tran, M. Zhou et al. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer discovery* 3.12 (2013), 1364–1377.

[5] S. Lee, K. H. Lee, M. Song and D. Lee. Building the process-drug–side effect network to discover the relationship between biological Processes and side effects. *BMC bioinformatics*. Vol. 12. 2. BioMed Central. 2011, S2.

[6] J. R. Pritchard, P. M. Bruno, M. T. Hemann and D. A. Lauffenburger. Predicting cancer drug mechanisms of action using molecular network signatures. *Molecular bioSystems* 9.7 (2013), 1604–1619.

[7] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross et al. The Connectivity

Map: using gene-expression signatures to connect small molecules, genes, and disease. *science* 313.5795 (2006), 1929–1935.

[8] A. B. Keenan, M. L. Wojciechowicz, Z. Wang, K. M. Jagodnik, S. L. Jenkins, A. Lachmann and A. Ma'ayan. Connectivity Mapping: Methods and Applications. *Annual Review of Biomedical Data Science* 2 (2019), 69–92.

[9] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences* 107.33 (2010), 14621–14626.

[10] J. Wang, F. Meng, E. Dai, F. Yang, S. Wang, X. Chen, L. Yang, Y. Wang and W. Jiang. Identification of associations between small molecule drugs and miRNAs based on functional similarity. *Oncotarget* 7.25 (2016), 38658.

[11] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine* 3.96 (2011), 96ra77–96ra77.

[12] D. Vidović, A. Koleti and S. C. Schürer. 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. *Frontiers in genetics* 5 (2014), 342.

[13] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha and A. J. Butte. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine* 3.96 (2011), 96ra76–96ra76.

[14] A. M. Brum, J. van de Peppel, L. Nguyen, A. Aliev, M. Schreuders-Koedam, T. Gajadien, C. S. van der Leije, A. van Kerkwijk, M. Eijken, J. P. van Leeuwen et al. Using the connectivity map to discover compounds influencing human osteoblast differentiation. *Journal of cellular physiology* 233.6 (2018), 4895–4906.

[15] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171.6 (2017), 1437–1452.

[16] R. Edgar, M. Domrachev and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 30.1 (2002), 207–210.

[17] S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* 47.D1 (2018), D464–D474.

[18] A. Ma'ayan, A. D. Rouillard, N. R. Clark, Z. Wang, Q. Duan and Y. Kou. Lean Big Data integration in systems biology and systems pharmacology. *Trends in pharmacological sciences* 35.9 (2014), 450–460.

[19] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork. Drug target identification using side-effect similarity. *Science* 321.5886 (2008), 263–266.

[20] A. Musa, L. S. Ghoraie, S.-D. Zhang, G. Glazko, O. Yli-Harja, M. Dehmer, B. Haibe-Kains and F. Emmert-Streib. A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics* 19.3 (Jan. 2017), 506–523. ISSN: 1477-4054. DOI: 10.1093/bib/bbw112.

[21] A. Musa, S. Tripathi, M. Dehmer and F. Emmert-Streib. L1000 Viewer: A search engine and web interface for the LINCS data repository. *Frontiers in Genetics* 3.1 (2019), 7849. DOI: 10.1038/s41598-000.

[22] A. Koleti, R. Terryn, V. Stathias, C. Chung, D. J. Cooper, J. P. Turner, D. Vidović, M. Forlin, T. T. Kelley, A. D'Urso et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic acids research* 46.D1 (2017), D558–D566.

[23] O. M. Enache, D. L. Lahr, T. E. Natoli, L. Litichevskiy, D. Wadden, C. Flynn, J. Gould, J. K. Asiedu, R. Narayan and A. Subramanian. The GCTx format and cmap {Py, R, M} packages: resources for the optimized storage and integrated traversal of dense matrices of data and annotations. *BioRxiv* (2018), 227041.

[24] M. Niepel, M. Hafner, Q. Duan, Z. Wang, E. O. Paull, M. Chung, X. Lu, J. M. Stuart, T. R. Golub, A. Subramanian et al. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nature communications* 8.1 (2017), 1186.

[25] Q. Duan, S. P. Reid, N. R. Clark, Z. Wang, N. F. Fernandez, A. D. Rouillard, B. Readhead, S. R. Tritsch, R. Hodos, M. Hafner et al. L1000CDS 2: LINCS L1000 characteristic direction signatures search engine. *NPJ systems biology and applications* 2 (2016), 16015.

[26] E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant. PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*. Vol. 4. Elsevier, 2008, 217–241.

[27] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40.D1 (2011), D1100–D1107.

[28] S. Goto, Y. Okuno, M. Hattori, T. Nishioka and M. Kanehisa. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic acids research* 30.1 (2002), 402–404.

[29] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 34.suppl_1 (2006), D668–D672.

[30] H. Hieronymus, J. Lamb, K. N. Ross, X. P. Peng, C. Clement, A. Rodina, M. N. and Jinyan Du, K. Stegmaier, S. M. Raj, K. N. Maloney, J. Clardy, W. C. Hahn, G. Chiosis and T. R. Golub. Gene expression signature-based chemical genomic prediction identifies a novel class of {HSP90} pathway modulators. *Cancer Cell* 10.4 (2006), 321–330. ISSN: 1535-6108. DOI: `http://dx.doi.org/10.1016/j.ccr.2006.09.005`. URL: `http://www.sciencedirect.com/science/article/pii/S1535610806002820`.

[31] J. R. Nevins and A. Potti. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat Rev Genet* 8.8 (Aug. 2007), 601–609. URL: `http://dx.doi.org/10.1038/nrg2137`.

[32] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine* 3.96 (2011), 96ra77–96ra77.

[33] T. A. Manolio and F. S. Collins. The HapMap and genome-wide association studies in diagnosis and therapy. *Annual review of medicine* 60 (2009), 443–456.

[34] M. Schenone, V. Dancik, B. K. Wagner and P. A. Clemons. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology* 9.4 (2013), 232.

[35] H. Wang, Q. Gu, J. Wei, Z. Cao and Q. Liu. Mining drug–disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. *Clinical Pharmacology & Therapeutics* 97.5 (2015), 451–454.

[36] C. Pacini, F. Iorio, E. Gonçalves, M. Iskar, T. Klabunde, P. Bork and J. Saez-Rodriguez. DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* 29.1 (2012), 132–134.

[37] J. Kim, M. Yoo, J. Kang and A. C. Tan. K-Map: connecting kinases with therapeutics for drug repurposing and development. *Hum Genomics* 7.1 (2013), 20.

[38] S. Alaimo, V. Bonnici, D. Cancemi, A. Ferro, R. Giugno and A. Pulvirenti. DT-Web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC systems biology* 9.3 (2015), S4.

[39] Y. Hong, T. Downey, K. W. Eu, P. K. Koh and P. Y. Cheah. A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clinical & experimental metastasis* 27.2 (2010), 83–90.

[40] P. Nygren, M. Fryknas, B. Agerup and R. Larsson. Repositioning of the anthelmintic drug mebendazole for the treatment for colon cancer. *Journal of cancer research and clinical oncology* 139.12 (2013), 2133–2140.

[41] T. W. Miller, B. T. Hennessy, A. M. Gonzalez-Angulo, E. M. Fox, G. B. Mills, H. Chen, C. Higham, C. Garcia-Echeverria, Y. Shyr and C. L. Arteaga. Hyperactivation of phosphatidylinositol-3 kinase promotes escape from hormone dependence in estrogen receptor–positive human breast cancer. *The Journal of clinical investigation* 120.7 (2010), 2406–2413.

[42] I. N. Melas, T. Sakellaropoulos, F. Iorio, L. G. Alexopoulos, W.-Y. Loh, D. A. Lauffenburger, J. Saez-Rodriguez and J. P. Bai. Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integrative Biology* (2015).

[43] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. A. Chiang Annie P. and Morgan, M. M. Sarwal, P. J. Pasricha and A. J. Butte. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Science Translational Medicine* 3.96 (2011), 96ra76. DOI: 10.1126/scitranslmed.3002648. eprint: http://stm.sciencemag.org/content/3/96/96ra76.full.pdf. URL: http://stm.sciencemag.org/content/3/96/96ra76.abstract.

[44] S.-D. Zhang and T. Gant. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 9.1 (2008), 258. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-258. URL: http://www.biomedcentral.com/1471-2105/9/258.

[45] J. Li, X. Zhu and J. Y. Chen. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS computational biology* 5.7 (2009), e1000450.

[46] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102.43 (2005), 15545–15550.

[47] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. Van Noort and P. Bork. Drug-induced regulation of target expression. *PLoS computational biology* 6.9 (2010), e1000925.

[48] J. T. Dudley, T. Deshpande and A. J. Butte. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics* 12.4 (2011), 303–311.

[49] K. Fortney, J. Griesman, M. Kotlyar, C. Pastrello, M. Angeli, M. Sound-Tsao and I. Jurisica. Prioritizing therapeutics for lung cancer: an integrative meta-analysis of cancer gene signatures and chemogenomic data. *PLoS computational biology* 11.3 (2015), e1004068.

[50] G. Wang, Y. Ye, X. Yang, H. Liao, C. Zhao and S. Liang. Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma. *PloS one* 6.1 (2011), e14573.

[51] S. D. Kunkel, M. Suneja, S. M. Ebert, K. S. Bongers, D. K. Fox, S. E. Malmberg, F. Alipour, R. K. Shields and C. M. Adams. mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell metabolism* 13.6 (2011), 627–638.

[52] S.-D. Zhang and T. W. Gant. sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. *BMC bioinformatics* 10.1 (2009), 236.

[53] R. Breitling, P. Armengaud, A. Amtmann and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters* 573.1-3 (2004), 83–92.

[54] C.-T. Yeh, A. T. Wu, P. M.-H. Chang, K.-Y. Chen, C.-N. Yang, S.-C. Yang, C.-C. Ho, C.-C. Chen, Y.-L. Kuo, P.-Y. Lee et al. Trifluoperazine, an antipsychotic agent, inhibits cancer stem cell growth and overcomes drug resistance of lung cancer. *American journal of respiratory and critical care medicine* 186.11 (2012), 1180–1188.

[55] S.-D. Zhang and T. W. Gant. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC bioinformatics* 9.1 (2008), 258.

[56] M. R. Segal, H. Xiong, H. Bengtsson, R. Bourgon and R. Gentleman. Querying genomic databases: refining the connectivity map. *Statistical applications in genetics and molecular biology* 11.2 (2012).

[57] D. Shigemizu, Z. Hu, J.-H. Hung, C.-L. Huang, Y. Wang and C. DeLisi. Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer. *PLoS computational biology* 8.2 (2012), e1002347.

[58] A. C. Ravindranath, N. Perualila-Tan, A. Kasim, G. Drakakis, S. Liggi, S. C. Brewerton, D. Mason, M. J. Bodkin, D. A. Evans, A. Bhagwat et al. Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism-of-action analysis. *Molecular BioSystems* 11.1 (2015), 86–96.

[59] C. Ma, H.-I. H. Chen, M. Flores, Y. Huang and Y. Chen. BRCA-Monet: a breast cancer specific drug treatment mode-of-action network for treatment effective prediction using large scale microarray database. *BMC systems biology* 7.5 (2013), S5.

[60] J. A. Parkkinen and S. Kaski. Probabilistic drug connectivity mapping. *BMC bioinformatics* 15.1 (2014), 113.

[61] J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li and Y. Z. Chen. Mechanisms of drug combinations: interaction and network perspectives. *Nature reviews Drug discovery* 8.2 (2009), 111.

[62] J. Cheng, Q. Xie, V. Kumar, M. Hurle, J. M. Freudenberg, L. Yang and P. Agarwal. Evaluation of analytical methods for connectivity map data. *Biocomputing 2013*. World Scientific, 2013, 5–16.

[63] J. Ahmed, T. Meinel, M. Dunkel, M. S. Murgueitio, R. Adams, C. Blasse, A. Eckert, S. Preissner and R. Preissner. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic acids research* 39.suppl_1 (2010), D960–D967.

[64] R. Tabarés-Seisdedos and J. L. Rubenstein. Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders. *Nature Reviews Neuroscience* 14.4 (2013), 293.

[65] H. Engerud, I. Tangen, A. Berg, K. Kusonmano, M. Halle, A. Øyan, K. Kalland, I. Stefansson, J. Trovik, H. Salvesen et al. High level of HSF1 associates with aggressive endometrial carcinoma and suggests potential for HSP90 inhibitors. *British journal of cancer* 111.1 (2014), 78.

[66] J. Cheng, L. Yang, V. Kumar and P. Agarwal. Systematic evaluation of connectivity map for disease indications. *Genome medicine* 6.12 (2014), 95.

[67] Q. Duan, C. Flynn, M. Niepel, M. Hafner, J. L. Muhlich, N. F. Fernandez, A. D. Rouillard, C. M. Tan, E. Y. Chen, T. R. Golub et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic acids research* 42.W1 (2014), W449–W460.

[68] A. Engelberg. Iconix Pharmaceuticals, Inc.–removing barriers to efficient drug discovery through chemogenomics. *Pharmacogenomics* 5.6 (2004), 741–744.

[69] B. Ganter, S. Tugendreich, C. I. Pearson, E. Ayanoglu, S. Baumhueter, K. A. Bostian, L. Brady, L. J. Browne, J. T. Calvin, G.-J. Day et al. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of biotechnology* 119.3 (2005), 219–244.

[70] R.-L. Wang, A. D. Biales, N. Garcia-Reyero, E. J. Perkins, D. L. Villeneuve, G. T. Ankley and D. C. Bencic. Fish connectivity mapping: linking chemical stressors by their mechanisms of action-driven transcriptomic profiles. *BMC genomics* 17.1 (2016), 84.

[71] Y. Xiao, Y. Gong, Y. Lv, Y. Lan, J. Hu, F. Li, J. Xu, J. Bai, Y. Deng, L. Liu et al. Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Scientific reports* 5 (2015), 10889.

[72] H. Wu, J. Huang, Y. Zhong and Q. Huang. DrugSig: A resource for computational drug repositioning utilizing gene expression signatures. *PloS one* 12.5 (2017), e0177743.

[73] Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani and H. Yamada. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic acids research* 43.D1 (2014), D921–D927.

[74] C. Ye, D. J. Ho, M. Neri, C. Yang, T. Kulkarni, R. Randhawa, M. Henault, N. Mostacci, P. Farmer, S. Renner et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nature communications* 9.1 (2018), 4307.

[75] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research* 46.D1 (2017), D794–D801.

[76] Z. Wang, C. D. Monteiro, K. M. Jagodnik, N. F. Fernandez, G. W. Gundersen, A. D. Rouillard, S. L. Jenkins, A. S. Feldmann, K. S. Hu, M. G. McDermott et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature communications* 7 (2016), 12846.

[77] A. Musa, S. Tripathi, M. Kandhavelu, M. Dehmer and F. Emmert-Streib. Harnessing the biological complexity of Big Data from LINCS gene expression signatures. *PLOS ONE* 13.8 (Aug. 2018), 1–16. DOI: `10.1371/journal.pone.0201937`.

[78] A. Musa, S. Tripathi, M. Dehmer, O. Yli-Harja, S. A. Kauffman and F. Emmert-Streib. Systems Pharmacogenomic Landscape of Drug Similarities from LINCS data: Drug Association Networks. *Scientific Reports* 9.1 (2019), 7849. DOI: `10.1038/s41598-019-44291-3`.

[79] A. Musa, M. Dehmer, O. Yli-Harja and F. Emmert-Streib. Exploiting Genomic Relations in Big Data Repositories by Graph-Based Search Methods. *Machine Learning and Knowledge Extraction* 1.1 (2018), 205–210. ISSN: 2504-4990. DOI: `10.3390/make1010012`.

[80] K. Hinkelmann and O. Kempthorne. *Design and analysis of experiments. V. 1. Introduction to experimental design*. Tech. rep. John Wiley and Sons, 1994.

[81] N. Pornputtapong, K. Wanichthanarak, A. Nilsson, I. Nookaew and J. Nielsen. A dedicated database system for handling multi-level data in systems biology. *Source code for biology and medicine* 9.1 (2014), 17.

[82] G. A. Pavlopoulos, D. Malliarakis, N. Papanikolaou, T. Theodosiou, A. J. Enright and I. Iliopoulos. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *Gigascience* 4.1 (2015), 38.

[83] S. Tilkov and S. Vinoski. Node. js: Using JavaScript to build high-performance network programs. *IEEE Internet Computing* 14.6 (2010), 80–83.

[84] B. Nelson. *Getting to Know Vue. js*. 2018.

[85] E. Ong, J. Xie, Z. Ni, Q. Liu, S. Sarntivijai, Y. Lin, D. Cooper, R. Terryn, V. Stathias, C. Chung et al. Ontological representation, integration, and analysis of LINCS cell line cells and their cellular responses. *BMC bioinformatics* 18.17 (2017), 556.

[86] R. B. Stoughton and S. H. Friend. How molecular profiling could revolutionize drug discovery. *Nature Reviews Drug Discovery* 4.4 (2005), 345.

[87] R. G. Lim, C. Quan, A. M. Reyes-Ortiz, S. E. Lutz, A. J. Kedaigle, T. A. Gipson, J. Wu, G. D. Vatine, J. Stocksdale, M. S. Casale et al. Huntington's disease iPSC-derived brain microvascular endothelial cells reveal WNT-mediated angiogenic and blood-brain barrier deficits. *Cell reports* 19.7 (2017), 1365–1377.

[88] L. F. Zerbini, M. K. Bhasin, J. F. de Vasconcellos, J. D. Paccez, X. Gu, A. L. Kung and T. A. Libermann. Computational repositioning and preclinical validation of pentamidine for renal cell cancer. *Molecular cancer therapeutics* 13.7 (2014), 1929–1941.

[89] C. Liu, J. Su, F. Yang, K. Wei, J. Ma and X. Zhou. Compound signature detection on LINCS L1000 big data. *Molecular BioSystems* 11.3 (2015), 714–722.

[90] W. Pan, J. Lin and C. T. Le. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome biology* 3.5 (2002), research0022–1.

[91] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* 11.10 (2010), 733.

[92] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology* 11.10 (2010), R106.

[93] W. E. Berndtson. A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments. *Journal of animal science* 69.1 (1991), 67–76.

[94] S. Piening, F. M. Haaijer-Ruskamp, J. T. de Vries, M. E. van der Elst, P. A. de Graeff, S. M. Straus and P. G. Mol. Impact of safety-related regulatory action on clinical practice. *Drug safety* 35.5 (2012), 373–385.

[95] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America* 27.11 (1941), 499.

[96] M. Vidal. A unifying view of 21st century systems biology. *FEBS letters* 583.24 (2009), 3891–3894.

[97] L. Wang. Pharmacogenomics: a systems approach. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2.1 (2010), 3–22.

[98] M. AY, K.-I. Goh, M. E. Cusick, A.-L. Barabasi, M. Vidal et al. Drug–target network. *Nature biotechnology* 25.10 (2007), 1119–1127.

[99] G. Hu and P. Agarwal. Human disease-drug network based on genomic expression profiles. *PloS one* 4.8 (2009), e6536.

[100] F. Sirci, F. Napolitano, S. Pisonero-Vaquero, D. Carrella, D. L. Medina and D. di Bernardo. Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses. *NPJ Systems Biology and Applications* 3.1 (2017), 23.

[101] H. Ye, Q. Liu and J. Wei. Construction of drug network based on side effects and its application for drug repositioning. *PloS one* 9.2 (2014), e87864.

[102] N. El-Hachem, D. M. Gendoo, L. S. Ghoraie, Z. Safikhani, P. Smirnov, C. Chung, K. Deng, A. Fang, E. Birkwood, C. Ho et al. Integrative cancer pharmacogenomics to infer large-scale drug taxonomy. *Cancer research* 77.11 (2017), 3057–3069.

[103] P. K. Sorger, S. R. Allerheiligen, D. R. Abernethy, R. B. Altman, K. L. Brouwer, A. Califano, D. Z. D'Argenio, R. Iyengar, W. J. Jusko, R. Lalonde et al. Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. *An NIH white paper by the QSP workshop group*. Vol. 48. NIH Bethesda, MD. 2011.

[104] L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* 2005.09 (2005), P09008.

[105] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99.12 (2002), 7821–7826.

[106] S. Tripathi, S. Moutari, M. Dehmer and F. Emmert-Streib. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC bioinformatics* 17.1 (2016), 129.

[107] S. Kauffman. Differentiation of malignant to benign cells. *Journal of Theoretical Biology* 31.3 (1971), 429–451.

[108] S. Huang, I. Ernberg and S. Kauffman. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in cell & developmental biology*. Vol. 20. 7. Elsevier. 2009, 869–876.

[109] J. C. Mar and J. Quackenbush. Decomposition of gene expression state space trajectories. *PLoS computational biology* 5.12 (2009), e1000626.

[110] W. Jiang, T. Miyamoto, T. Kakizawa, T. Sakuma, S.-i. Nishio, T. Takeda, S. Suzuki and K. Hashizume. Expression of thyroid hormone receptor alpha in 3T3-L1 adipocytes; triiodothyronine increases the expression of lipogenic enzyme and triglyceride accumulation. *Journal of Endocrinology* 182.2 (2004), 295–302.

[111] W. Mai, M. F. Janier, N. Allioli, L. Quignodon, T. Chuzel, F. Flamant and J. Samarut. Thyroid hormone receptor α is a molecular switch of cardiac function between fetal and postnatal life. *Proceedings of the National Academy of Sciences* 101.28 (2004), 10332–10337.

[112] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[113] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng and K.-C. Chou. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS one* 7.4 (2012), e35254.

[114] M. Raymond and F. Rousset. An exact test for population differentiation. *Evolution* 49.6 (1995), 1280–1283.

[115] M. P. Fay. Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics* 11.2 (2009), 373–374.

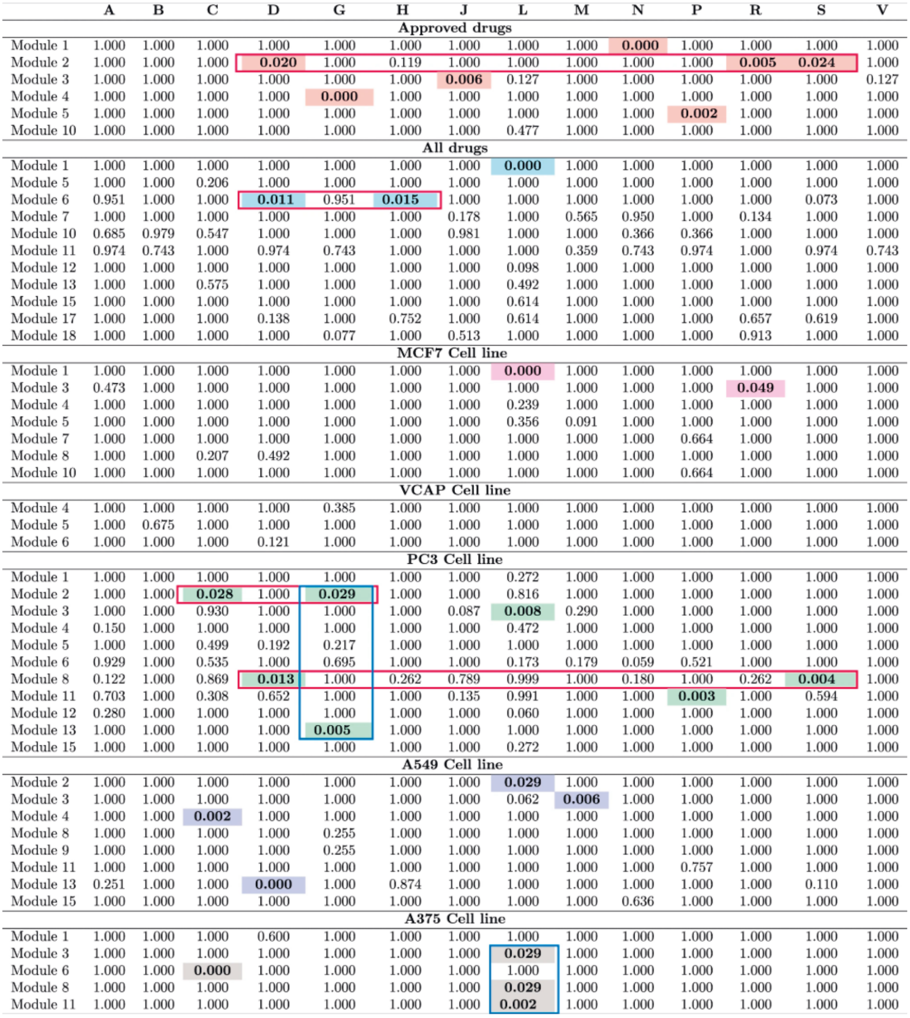
[116] H.-Y. Kim. Statistical notes for clinical researchers: chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics* 42.2 (2017), 152–155.

[117] D. A. Davis and N. V. Chawla. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PloS one* 6.7 (2011), e22670.

[118] F. Emmert-Streib, S. Tripathi, R. d. M. Simoes, A. F. Hawwa and M. Dehmer. The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes. *Systems Biomedicine* 1.1 (2013), 20–28.

[119] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 6 (2017), e26726.

A SUPPLEMENTARY INFORMATION



	A	B	C	D	G	H	J	L	M	N	P	R	S	V
Approved drugs														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000
Module 2	1.000	1.000	1.000	**0.020**	1.000	0.119	1.000	1.000	1.000	1.000	1.000	**0.005**	**0.024**	1.000
Module 3	1.000	1.000	1.000	1.000	1.000	1.000	**0.006**	0.127	1.000	1.000	1.000	1.000	1.000	0.127
Module 4	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.002**	1.000	1.000	1.000
Module 10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.477	1.000	1.000	1.000	1.000	1.000	1.000
All drugs														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	0.206	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	0.951	1.000	1.000	**0.011**	0.951	**0.015**	1.000	1.000	1.000	1.000	1.000	1.000	0.073	1.000
Module 7	1.000	1.000	1.000	1.000	1.000	1.000	0.178	1.000	0.565	0.950	1.000	0.134	1.000	1.000
Module 10	0.685	0.979	0.547	1.000	1.000	1.000	0.981	1.000	1.000	0.366	0.366	1.000	1.000	1.000
Module 11	0.974	0.743	1.000	0.974	0.743	1.000	1.000	1.000	0.359	0.743	0.974	1.000	0.974	0.743
Module 12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.098	1.000	1.000	1.000	1.000	1.000	1.000
Module 13	1.000	1.000	0.575	1.000	1.000	1.000	1.000	0.492	1.000	1.000	1.000	1.000	1.000	1.000
Module 15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.614	1.000	1.000	1.000	1.000	1.000	1.000
Module 17	1.000	1.000	1.000	0.138	1.000	0.752	1.000	0.614	1.000	1.000	1.000	0.657	0.619	1.000
Module 18	1.000	1.000	1.000	1.000	0.077	1.000	0.513	1.000	1.000	1.000	1.000	0.913	1.000	1.000
MCF7 Cell line														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	0.473	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.049**	1.000	1.000
Module 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.239	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.356	0.091	1.000	1.000	1.000	1.000	1.000
Module 7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.664	1.000	1.000	1.000
Module 8	1.000	1.000	0.207	0.492	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.664	1.000	1.000	1.000
VCAP Cell line														
Module 4	1.000	1.000	1.000	1.000	0.385	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	0.675	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	1.000	1.000	1.000	0.121	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PC3 Cell line														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.272	1.000	1.000	1.000	1.000	1.000	1.000
Module 2	1.000	1.000	**0.028**	1.000	**0.029**	1.000	1.000	0.816	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	1.000	1.000	0.930	1.000	1.000	1.000	0.087	**0.008**	0.290	1.000	1.000	1.000	1.000	1.000
Module 4	0.150	1.000	1.000	1.000	1.000	1.000	1.000	0.472	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	0.499	0.192	0.217	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	0.929	1.000	0.535	1.000	0.695	1.000	1.000	0.173	0.179	0.059	0.521	1.000	1.000	1.000
Module 8	0.122	1.000	0.869	**0.013**	1.000	0.262	0.789	0.999	1.000	0.180	1.000	0.262	**0.004**	1.000
Module 11	0.703	1.000	0.308	0.652	1.000	1.000	0.135	0.991	1.000	1.000	**0.003**	1.000	0.594	1.000
Module 12	0.280	1.000	1.000	1.000	1.000	1.000	1.000	0.060	1.000	1.000	1.000	1.000	1.000	1.000
Module 13	1.000	1.000	1.000	1.000	**0.005**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.272	1.000	1.000	1.000	1.000	1.000	1.000
A549 Cell line														
Module 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.029**	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.062	**0.006**	1.000	1.000	1.000	1.000	1.000
Module 4	1.000	1.000	**0.002**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 8	1.000	1.000	1.000	1.000	0.255	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 9	1.000	1.000	1.000	1.000	0.255	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.757	1.000	1.000	1.000
Module 13	0.251	1.000	1.000	**0.000**	1.000	0.874	1.000	1.000	1.000	1.000	1.000	1.000	0.110	1.000
Module 15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.636	1.000	1.000	1.000	1.000
A375 Cell line														
Module 1	1.000	1.000	1.000	0.600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.029**	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.029**	1.000	1.000	1.000	1.000	1.000	1.000
Module 11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.002**	1.000	1.000	1.000	1.000	1.000	1.000

Figure A.1 Enrichment of individual modules in the DANs. Shown are the BH corrected q-values of Fisher's exact tests for the enrichment of ATC codes in each of the modules of the DANs. Modules not shown, do not contain any enriched ATC code. The highlighted cells are statistically significant. The horizontal and vertical boxes highlight the multiple occurance of ATC classes in modules and multiple enriched modules for an ATC class respectively.

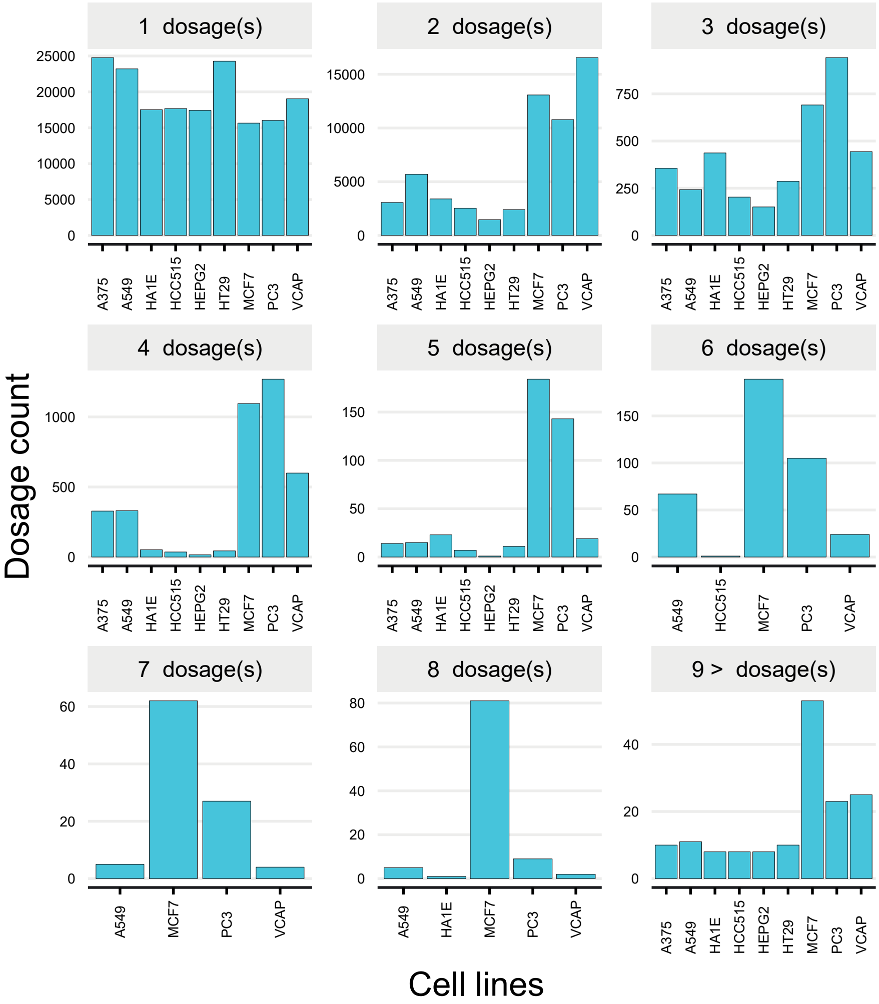


Figure A.2 Distributions of unique dosages for the signature profiles. The number of available profiles is shown for different dosages (concentrations) of small molecules for 9 highly profiled cell lines.

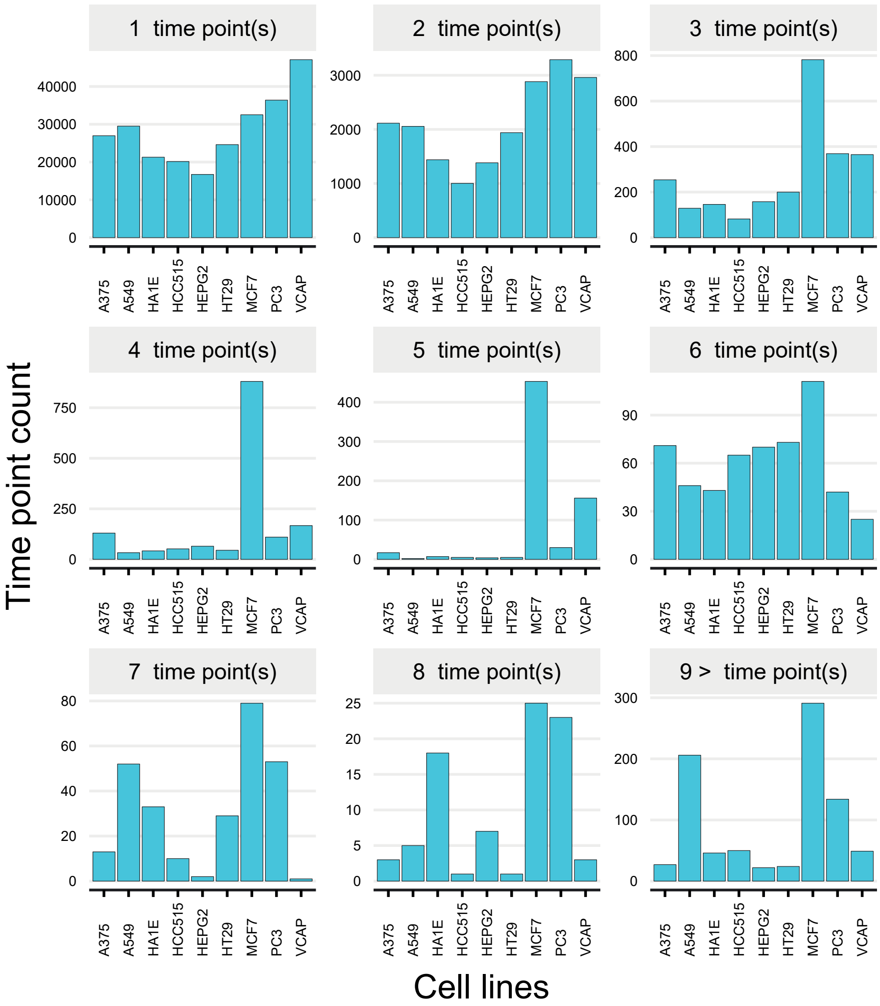


Figure A.3 Number of different time points measured on the 9 cell line subset.The number of cell lines per compound represented in the treatments ranged from 1 to 8 different time points count out of 14. Around 99% of the perturbagens affected at least one gene significantly in a single cell line after treatment with the different number of time points.

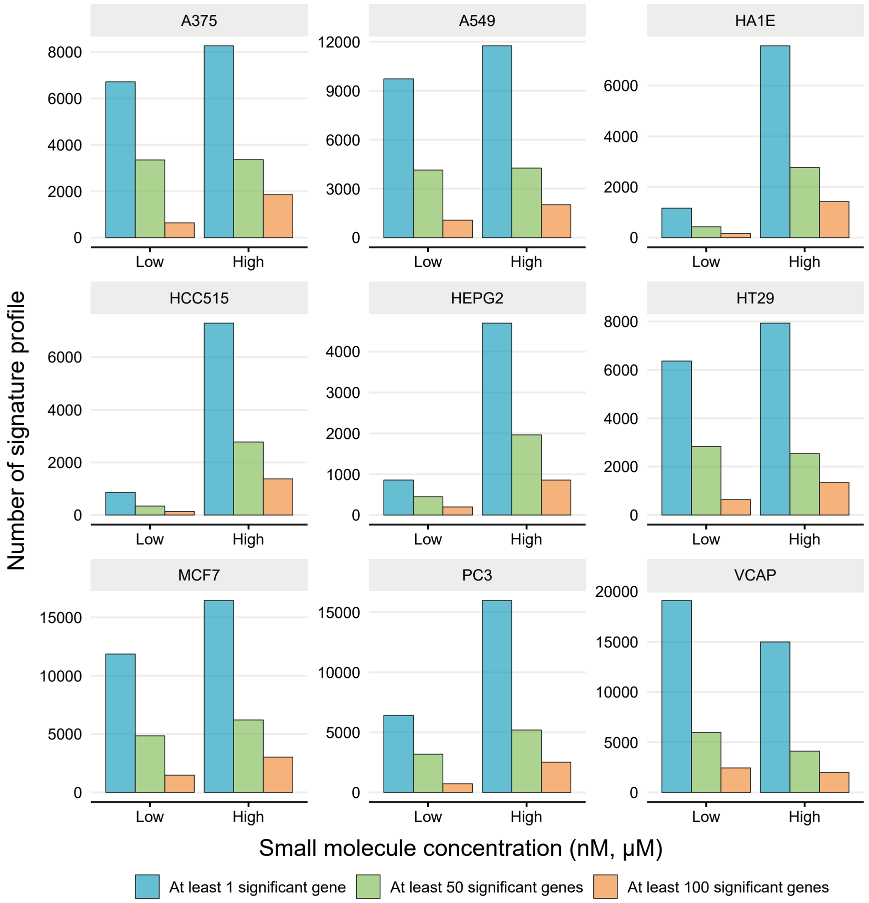


Figure A.4 Dosage specific differentially gene expression. The differential expression of genes for 9 cell lines is shown categorized in Low and High dosages of small molecules.

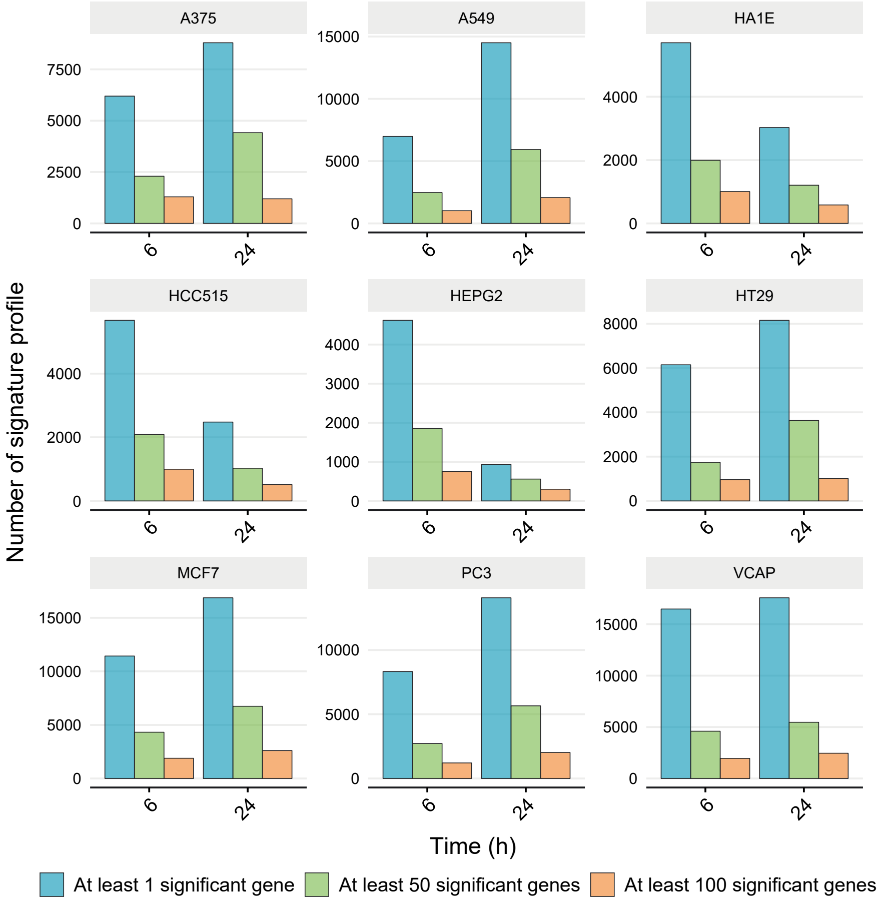


Figure A.5 Time point specific differentially gene expression. The differential expression of genes for 9 cell lines is shown categorized in the time durations (6 and 24h) of drug perturbations.

PUBLICATIONS

PUBLICATION
I

A review of connectivity map and computational approaches in pharmacogenomics

A. Musa, L. S. Ghoraie, S.-D. Zhang, G. Glazko, O. Yli-Harja, M. Dehmer,
B. Haibe-Kains and F. Emmert-Streib

Briefings in Bioinformatics 19.3 (2017), 506–523
DOI: 10.1093/bib/bbw112

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

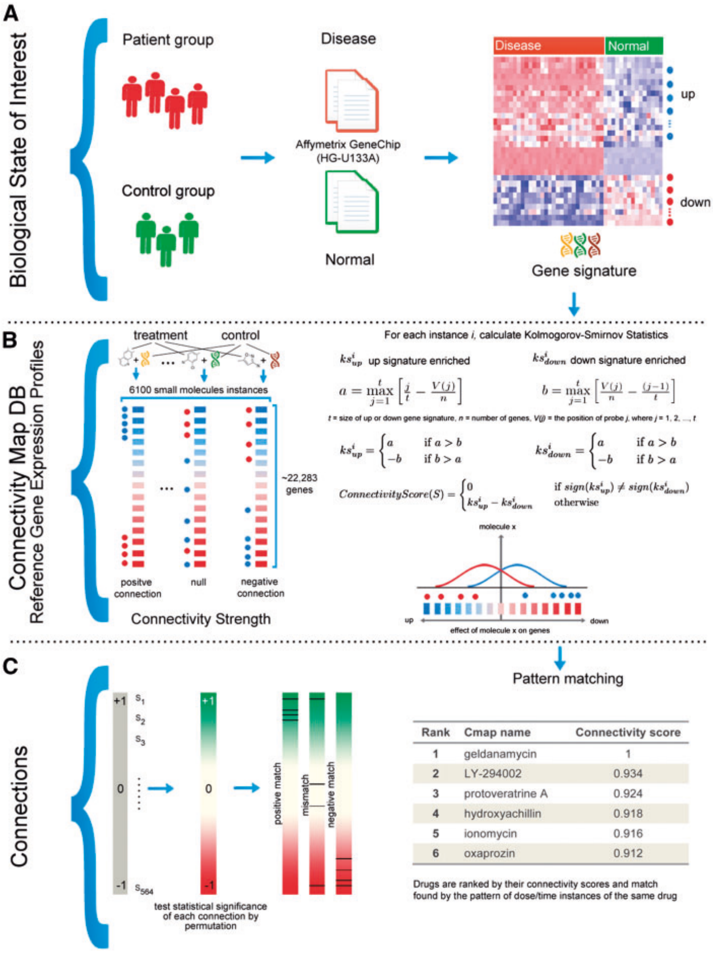

A
Biological State of Interest
Patient group
Control group
Disease
Affymetrix GeneChip (HG-U133A)
Normal
Disease
Normal
up
down
Gene signature
B
Connectivity Map DB
Reference Gene Expression Profiles
treatment
control
6100 small molecules instances
~22,283 genes
positve connection
null
negative connection
Connectivity Strength
For each instance *i*, calculate Kolmogorov-Smirnov Statistics
ks^i_{up} up signature enriched
ks^i_{down} down signature enriched
$a = \max_{j=1}^{t}\left[\frac{j}{t} - \frac{V(j)}{n}\right]$
$b = \max_{j=1}^{t}\left[\frac{V(j)}{n} - \frac{(j-1)}{t}\right]$
t = size of up or down gene signature, n = number of genes, $V(j)$ = the position of probe j, where $j = 1, 2, ..., t$
$ks^i_{up} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } b > a \end{cases}$
$ks^i_{down} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } b > a \end{cases}$
$ConnectivityScore(S) = \begin{cases} 0 & \text{if } sign(ks^i_{up}) \neq sign(ks^i_{down}) \\ ks^i_{up} - ks^i_{down} & \text{otherwise} \end{cases}$
molecule x
up
down
effect of molecule x on genes
C
Connections
+1
S_1
S_2
S_3
0
-1
S_{564}
+1
0
-1
test statistical significance of each connection by permutation
positive match
mismatch
negative match
Pattern matching
Rank | Cmap name | Connectivity score
1 | geldanamycin | 1
2 | LY-294002 | 0.934
3 | protoveratrine A | 0.924
4 | hydroxyachillin | 0.918
5 | ionomycin | 0.916
6 | oxaprozin | 0.912
Drugs are ranked by their connectivity scores and match found by the pattern of dose/time instances of the same drug

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 $(N-i+1)$ for upregulation or $-(N-i+1)$ 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.

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

Where g_i represents the ith gene in the signature, $s(g_i)$ is its signed rank in the signature and $R(g_i)$ 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.

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

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

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].

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

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 ~15, 000 small molecules compounds and ~5000 genes (small hairpin RNA and overexpression) in ~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.

PUBLICATION

II

Exploiting Genomic Relations in Big Data Repositories by Graph-Based Search Methods

A. Musa, M. Dehmer, O. Yli-Harja and F. Emmert-Streib

Machine Learning and Knowledge Extraction 1.1 (2018), 205–210
DOI: 10.3390/make1010012

machine learning &
knowledge extraction

Opinion

Exploiting Genomic Relations in Big Data Repositories by Graph-Based Search Methods

Aliyu Musa [1,2], Matthias Dehmer [3,4,5], Olli Yli-Harja [2,6,7] and Frank Emmert-Streib [1,2,*]

[1] Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland; aliyu.musa@tut.fi
[2] Institute of Biosciences and Medical Technology, 33520 Tampere, Finland; olli.yli-harja@tut.fi
[3] Department of Mechatronics and Biomedical Computer Science, UMIT, 6060 Hall in Tyrol, Austria; matthias.dehmer@umit.at
[4] College of Computer and Control Engineering, Nankai University, Tianjin 300071, China
[5] Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, 4400 Steyr Campus, Austria
[6] Computational Systems Biology Lab, Tampere University of Technology, 33720 Tampere, Finland
[7] Institute for Systems Biology, Seattle, WA 98109, USA
* Correspondence: v@bio-complexity.com; Tel.: +358-50-301-5353

Received: 26 September 2018; Accepted: 21 November 2018; Published: 22 November 2018

Abstract: We are living at a time that allows the generation of mass data in almost any field of science. For instance, in pharmacogenomics, there exist a number of big data repositories, e.g., the Library of Integrated Network-based Cellular Signatures (LINCS) that provide millions of measurements on the genomics level. However, to translate these data into meaningful information, the data need to be analyzable. The first step for such an analysis is the deliberate selection of subsets of raw data for studying dedicated research questions. Unfortunately, this is a non-trivial problem when millions of individual data files are available with an intricate connection structure induced by experimental dependencies. In this paper, we argue for the need to introduce such search capabilities for big genomics data repositories with a specific discussion about LINCS. Specifically, we suggest the introduction of *smart interfaces* allowing the exploitation of the connections among individual raw data files, giving raise to a network structure, by graph-based searches.

Keywords: knowledge extraction; computational pharmacogenomics; systems pharmacogenomics; network science; computational biology; genomics; big data; databases

1. Introduction

In the last 20 years, technological progress in high-throughput assays, e.g., next-generation sequencing, led to a tremendous increase of our data generation capabilities in genomics. As a result, there are many data collections available providing millions of data points about DNA sequence, gene expression, metabolic, protein structure or protein interaction data [1]. However, to reveal the information buried within these data collections, such data need to be analyzable [2]. The problem is that accessing *selected subsets* of these "big data" for performing a dedicated analysis is non-trivial due to the sheer number of data and, more importantly, the complexity of the connections between different data points. Unfortunately, most data collections do not provide efficient interfaces enabling a direct access to subsets of *raw data*, thus hampering downstream analysis.

For reasons of clarity, we would like to highlight that, here, we are concerned with accessing and selecting *raw data*, not knowledge that has been derived by processing and analyzing raw data and stored in *knowledge databases*. Instead, the data repositories we are concerned with in our paper, store raw data files (see Figure 1 for a brief overview). In the following, we discuss this problem by focusing on the pharmacogenomic data repository LINCS (Library of Integrated Network-based Cellular Signatures) [3–8] and describe how this lack in querying capability could be compensated.

2. Preliminaries

Before we discuss the problem under consideration, we would like to clarify a couple of terms used throughout the paper. We use the term *data repository* for a very general collection and storage of individual data files without providing any dedicated accessing or searching capabilities. Sometimes, this may also be referred to as a data library. Here, by lack of *dedicated* accessing and searching capabilities, we mean that information about data files can in principle be searched but in an inefficient way, which may be as simple as a manual browsing of the data.

In contrast, we use the term database to refer to an *organized* collection and storage of data for which a database management system (DBMS) is available that allows querying the data from the database. The term database system refers to the combination of a database with a DBMS. Here, the term "organized" refers to a specific type of a database, e.g., a relational database or object-oriented database.

It is important to note that each type of data organization (data repository, database system, etc.) comes with its own characteristics. Interestingly, the conceptual idea discussed in the following does not fit nicely into any of these well-known, existing categories, but is situated between them, extending and modifying characteristics thereof.

3. The Pharmacogenomics Data Repository LINCS

The LINCS data repository is supported by the NIH (National Institute of Health), comprising 5000 genetic perturbagens (e.g., single-gene knockdowns or overexpressions) and 15,000 perturbagens induced by chemical compounds (e.g., drugs) [9]. To date, almost two million gene expressions have been profiled using the L1000 technology [9]. 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) [10]. Access to the raw data is provided by GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742) but, unfortunately, there is no search functionality provided other than to select all raw files for download. For this reason, LINCS is merely a collection of files usually called a data repository.

This particular example of LINCS described above is typical for the current situation of many big *data repositories*. Here, we want to emphasize that a data repository is not a database itself. Instead, it stands for a more generalized term that indicates the lack of basic functionality usually present within a database yet providing data storage capabilities. In our context, the crucial lack of functionality is the limited capability to provide efficient ways to query the data within the data repository for selecting and downloading subsets of the data (files).

To rectify this problem, in our opinion, big data repositories need functionality we summarize by the term *smart interfaces*. We envision a smart interface as a web interface that enables extensive selection capabilities, providing many features for querying, exploration, downloading and analyzing data and related meta information. It would also support programmatic access via API as a search functionality to all the attributes contained within the data repository. Using the API, computational scientists and developers can access the data and build flexible research pipelines. Given the genomic context of LINCS and related data repositories, the data queries can utilize the dependency structure between individual data files as implied by, for instance, experimental or biological conditions (see example below). Hence, queries perform network or graph-based searches within the data repositories exploiting in this way the existing dependency structure between the individual data files.

> Smart interfaces: A web interface enabling extensive selection capabilities of raw data, providing features for graph-based querying, exploration, downloading and analyzing data and related meta information.

In Figure 1, we show a visualization of our idea. A smart interface exploits the connectivity structure among the raw data files (see Figure 1A), which can also include metadata if available, by generating a network representation among the individual data files (see Figure 1B). In the example of the LINCS database, these connections are given by the combination of cell lines, drugs, dosages of drugs, etc. for which gene expression profiles have been generated. In general, these correspond to the attributes of the data files. Importantly, these attributes remain constant and do not change if more data points are added to a database. Once such a network representation among the data files is generated, a user query extracts quickly the desired data files, e.g., that correspond to cell line C2, the drugs D1 and D3 and the dosage Do3 (see Figure 1B and its connection back shown in yellow to the data files), because each search combination connects to a list of associated data files. In this way, a smart interface forms a connection between the data repository and the preprocessing and analysis of the data (see Figure 1C) and its purpose is to provide a graphical-user-interface and query function for an efficient access to selected data files. We want to emphasize that the network representation should be part of the smart interface because, in this way, it would be easily applicable for practitioners such as biologists or clinicians.

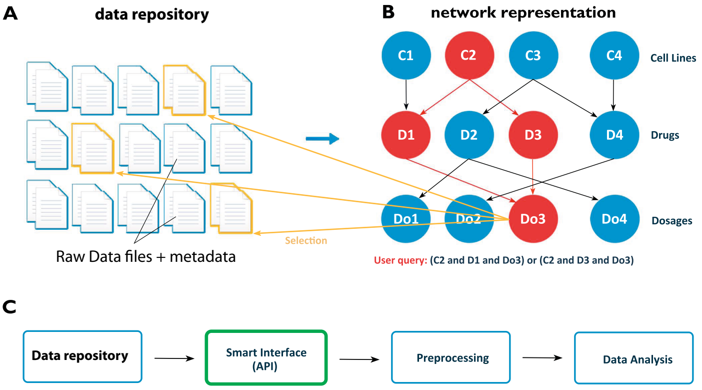


Figure 1. (**A**) A collection of available individual (raw) data files and metadata; (**B**) Network representation of connections between the raw data files. A user query (in red) corresponds to one particular combination of attributes of the data files, which leads to an efficient selection of these (in yellow); (**C**) Conceptual integration of the smart interface, which is an application programming interface (API), into a conventional data analysis pipeline.

4. Technical Considerations

On a technical note, we would like to point out that, here, we focus on a data representation that would allow users to immediately interact with the data. Through the smart interface, users can perform highly specialized queries using attributes that naturally connect the individual data files. The queries can be executed in the web browser or programmatically from the interface. This could be achieved by using modern generalizations of relational databases [11], e.g., NoSQL [12] or graph [13] databases, to efficiently store the data for quick access. Unfortunately, non-relational databases have been naturally fragmented by usage and have drawbacks in scaling, resulting in relatively slow progress in integrating large datasets [14]. However, for genomics problems with a constant number of attributes, e.g., cell lines, drugs, dosages, etc, as is the case for the LINCS data (see below), the known scaling problems of graph databases do not hamper their usage because new data points do not lead to a change in the number of attributes and, hence, the database can grow efficiently in the number of stored data points. An example of this was given by Himmelstein et al. [15] using a graph database for integrating information from 29 public resources to connect compounds, diseases, genes, pharmacologic classes, side effects, etc., which helped to identified network patterns that distinguish treatments from non-treatments drugs [15]. We would like to point out that the result of [15], and similar approaches [16–18], is a knowledge database that operates on processed and analyzed data, not on the raw data files as is our major concern in this paper.

For the LINCS data, one can start from a set of files and select certain attributes to create a network representation by using graph algorithms [19]. This is similar to classical contributions focusing on data and retrieval based on graph theoretical considerations [20,21], which do technically not fall within the strict definition of databases because they are lacking the consideration of database management systems as the most important building block when one refers to the term database. Hence, more research is required to identify if a database structure, an information retrieval system [22,23] or a graph-based file organization system [20,21] provides the most appropriate technical realization for graph-based searches of data repositories in genomics.

5. Conceptual Idea

The general idea of a smart interface is similar to the idea behind Google. If one considers "web sites" as "data files" and realizes that the "connections between web sites" are implicitly provided by the "attributes of data files" (see the example above), then the analogy is apparent. Specifically, Google identifies the connections between web sites by searching the links from and to sites by crawling the web. This establishes a graph structure between the web sites corresponding to a very large network upon which graph-based searches that take user queries into account can be executed.

In the case the world-wide-web (WWW) would consist of only a dozen web sites, there would be no need for a search engine such as Google because a user could quickly go through the list of these web sites manually. However, for billions of web sites, this is no longer feasible (even if such a list would exist) because a linear search would lead to exponential searching times. Interestingly, this is exactly the situation we are facing for data repositories such as LINCS. While the current raw organization of LINCS or similar data repositories is sufficient for certain tasks, it does not favor the selection of complexly determined subsets, such as those required for more advanced or specialized studies. Any additional tool, such as a smart interface, that can be added to facilitate such complex queries, would make these repositories more useful, efficient and popular. This would not only benefit users, but also the repositories themselves by reducing work, reducing download costs and increasing their impact, usage and user satisfaction. This implies also that the true potential of LINCS is currently not yet unlocked due to this limitation. For technical completeness we would like to note that Google uses a NoSQL database of columnar type called BigTable [24].

6. Further Applications

We would like to note that our idea extends beyond the LINCS data repository. Other examples of raw data repositories that would benefit from a similar approach are:

- Gene Expression Omnibus (GEO) [1]
- NCI60 human tumour cell line anticancer drug screen [25]
- ArrayExpress [26]
- Cancer Cell Line Encyclopedia [27]

However, the largest benefit for the community would result from the integration of some (or all) such data repositories to address the problems by a systems biology approach taking holistically all aspects into account. We expect that the smart interface needs to be adapted to the specific characteristics of the data types in the corresponding data repositories but the conceptual core idea would be generic to all these different repositories.

In our opinion, the implementation costs would be rather limited because it only requires a software solution. However, the intellectual costs are considerable because the creation of graph-based relations among the individual data files requires familiarity with basic graph-theoretical concepts and graph-search methods [19,28].

7. Conclusions

The transition from simple data repositories to big pharmacological warehouses requires new forms of data accessing strategies and we think that smart interfaces, enabling graph-based querying capabilities, provide the needed functionalities. While current repositories offer the possibility to mirror data to access it in a local implementation, it would carry unduly efforts and costs for most users, many of whom would not be able to do it (and hence to benefit from the data), and this cost would be best addressed if data repositories offering advanced search technologies were available, whether at the primary curation site, or at a separate publicly accessible resource, or both. Otherwise, the opportunities offered by these big data cannot be translated into new knowledge by means of modern data science [29]. We discussed our idea for the LINCS data repository and provided a specific outline of the graph structure induced by the available data files. However, our idea is not limited to LINCS, but we selected this data repository because of its popularity to emphasize the need for such search capabilities.

Finally, we would like to note that, in our presentation, we focused on the network-based search capabilities and neglected many other data aspects of practical relevance, e.g., data privacy, data quality, etc, to convey a clear message. However, we do not want to miss emphasizing that these aspects are also part of the data analysis pipeline (see Figure 1C) that needs to be integrated into our framework to obtain a functional implementation.

Author Contributions: F.E.S. and A.M. conceived the study. All authors wrote the paper and approved the final version.

Funding: AM thanks the CIMO foundation of Finland for a scholarship. MD thanks the Austrian Science Funds for supporting this work (project P30031).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef] [PubMed]
2. Holzinger, A.; Jurisica, I. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*; Springer: Berlin, Germany, 2014; pp. 1–18.
3. Lamb, J.; Crawford, E.D.; Peck, D.; Modell, J.W.; Blat, I.C.; Wrobel, M.J.; Lerner, J.; Brunet, J.P.; Subramanian, A.; Ross, K.N.; et al. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929–1935. [CrossRef] [PubMed]
4. Ma'ayan, A.; Rouillard, A.; Clark, N.; Wang, Z.; Duan, Q.; Kou, Y. Lean Big Data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.* **2014**, *35*, 450–460. [CrossRef] [PubMed]
5. Campillos, M.; Kuhn, M.; Gavin, A.C.; Jensen, L.J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266. [CrossRef] [PubMed]
6. Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; et al. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. *BioRxiv* **2017**. [CrossRef] [PubMed]
7. Musa, A.; Ghoraie, L.; Zhang, S.D.; Glazko, G.; Yli-Harja, O.; Dehmer, M.; Haibe-Kains, B.; Emmert-Streib, F. A Review of Connectivity Mapping and Computational Approaches in Pharmacogenomics. *Brief. Bioinform.* **2017**, *19*, 506–523.

8. Musa, A.; Tripathi, S.; Kandhavelu, M.; Dehmer, M.; Emmert-Streib, F. Harnessing the biological complexity of Big Data from LINCS gene expression signatures. *PLoS ONE* **2018**, *13*, e0201937. [CrossRef] [PubMed]
9. Vidovic, D.; A, K.; Schurer, S. 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. [PubMed]
10. Barrett, T.; Troup, D.B.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; et al. NCBI GEO: Archive for functional genomics data sets -10 years on. *Nucleic Acids Res.* **2011**, *39*, D1005–D1010. [CrossRef] [PubMed]
11. Codd, E.F. A Relational Model of Data for Large Shared Data Banks. *Commun. ACM* **1970**, *13*, 377–387. [CrossRef]
12. Wiese, L. *Advanced Data Management: For SQL, NoSQL, Cloud and Distributed Databases*; De Gruyter: Berlin, Germany, 2015.
13. Angles, R.; Gutierrez, C. Survey of Graph Database Models. *ACM Comput. Surv.* **2008**, *40*, 1–39. [CrossRef]
14. Zou, L.; Chen, L.; Özsu, M.T. Distance-join: Pattern match query in a large graph database. *Proc. VLDB Endowment* **2009**, *2*, 886–897. [CrossRef]
15. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **2017**, *6*, e26726. [CrossRef] [PubMed]
16. Matthews, L.; Gopinath, G.; Gillespie, M.; Caudy, M.; Croft, D.; de Bono, B.; Garapati, P.; Hemish, J.; Hermjakob, H.; Jassal, B.; et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **2009**, *37*, D619–D622. [CrossRef] [PubMed]
17. Swainston, N.; Batista-Navarro, R.; Carbonell, P.; Dobson, P.D.; Dunstan, M.; Jervis, A.J.; Vinaixa, M.; Williams, A.R.; Ananiadou, S.; Faulon, J.L.; et al. biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PLoS ONE* **2017**, *12*, 1–14. [CrossRef] [PubMed]
18. Touré, V.; Mazein, A.; Waltemath, D.; Balaur, I.; Saqi, M.; Henkel, R.; Pellet, J.; Auffray, C. STON: Exploring biological pathways using the SBGN standard and graph databases. *BMC Bioinform.* **2016**, *17*, 494. [CrossRef] [PubMed]
19. Cormen, T.; Leiserson, C.; Rivest, R.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, USA, 2001.
20. Lipski, W.; Marek, W., File organization, an application of graph theory. In *Automata, Languages and Programming: 2nd Colloquium, University of Saarbrücken 29 July– 2 August 1974*; Loeckx, J., Ed.; Springer: Berlin/Heidelberg, Germany, 1974; pp. 270–279.
21. Lipski, W. Information storage and retrieval ? mathematical foundations II (combinatorial problems). *Theor. Comput. Sci.* **1976**, *3*, 183 – 211. [CrossRef]
22. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*; ACM Press: New York, NU, USA, 1999; Volume 463.
23. Chowdhury, G.G. *Introduction to Modern Information Retrieval*; Facet Publishing: London, UK, 2010.
24. Chang, F.; Dean, J.; Ghemawat, S.; Hsieh, W.C.; Wallach, D.A.; Burrows, M.; Chandra, T.; Fikes, A.; Gruber, R.E. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.* **2008**, *26*, 4. [CrossRef]
25. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823. [CrossRef] [PubMed]
26. Brazma, A.; Parkinson, H.; Sarkans, U.; Shojatalab, M.; Vilo, J.; Abeygunawardena, N.; Holloway, E.; Kapushesky, M.; Kemmeren, P.; Lara, G.G.; et al.. ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **2003**, *31*, 68–71. [CrossRef] [PubMed]
27. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [CrossRef] [PubMed]
28. Dehmer, M.; Emmert-Streib, F., Eds. *Analysis of Complex Networks: From Biology to Linguistics*; Wiley-VCH: Weinheim, Germany, 2009.
29. Emmert-Streib, F.; Moutari, S.; Dehmer, M. The process of analyzing data is the emergent feature of data science. *Front. Genet.* **2016**, *7*, 12. [CrossRef] [PubMed]

PUBLICATION

III

L1000 Viewer: A search engine and web interface for the LINCS data repository

A. Musa, S. Tripathi, M. Dehmer and F. Emmert-Streib

Frontiers in Genetics 3.1 (2019), 7849
DOI: 10.1038/s41598-000

frontiers in Genetics

TECHNOLOGY REPORT
published: 14 June 2019
doi: 10.3389/fgene.2019.00557

L1000 Viewer: A Search Engine and Web Interface for the LINCS Data Repository

Aliyu Musa[1,2], *Shailesh Tripathi*[1,3], *Matthias Dehmer*[3,4,5] *and Frank Emmert-Streib*[1,2]*

[1] *Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland,* [2] *Institute of Biosciences and Medical Technology, Tampere, Finland,* [3] *Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Linz, Austria,* [4] *Department of Mechatronics and Biomedical Computer Science, UMIT, Hall in Tyrol, Austria,* [5] *College of Computer and Control Engineering, Nankai University, Tianjin, China*

The LINCS L1000 data repository contains almost two million gene expression profiles for thousands of small molecules and drugs. However, due to the complexity and the size of the data repository and a lack of an interoperable interface, the creation of pharmacologically meaningful workflows utilizing these data is severely hampered. In order to overcome this limitation, we developed the L1000 Viewer, a search engine and graphical web interface for the LINCS data repository. The web interface serves as an interactive platform allowing the user to select different forms of perturbation profiles, e.g., for specific cell lines, drugs, dosages, time points and combinations thereof. At its core, our method has a database we created from inferring and utilizing the intricate dependency graph structure among the data files. The L1000 Viewer is accessible via http://L1000viewer.bio-complexity.com/.

Keywords: gene expression, big data, pharmacogenomics, web application, visualization, data science

OPEN ACCESS

Edited by:
Mehdi Pirooznia,
National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Zichen Wang,
Icahn School of Medicine at Mount Sinai, United States
Marco Brandizi,
Rothamsted Research (BBSRC), United Kingdom

****Correspondence:***
Frank Emmert-Streib
v@bio-complexity.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics

Received: *13 April 2019*
Accepted: *28 May 2019*
Published: *14 June 2019*

Citation:
Musa A, Tripathi S, Dehmer M and Emmert-Streib F (2019) L1000 Viewer: A Search Engine and Web Interface for the LINCS Data Repository.
Front. Genet. 10:557.
doi: 10.3389/fgene.2019.00557

1. INTRODUCTION

We are living in the era of big data that sparked the establishment of the field data science (Smith, 2006; Ma'ayan et al., 2014; Jin et al., 2015; Emmert-Streib and Dehmer, 2019). For genomics, the recent growth of high-throughput biomedical and pharmacogenomic data (Edgar et al., 2002; Barrett et al., 2013; Woo et al., 2015; Musa et al., 2017) presents opportunities and at the same time challenges for their analysis. Paramount to these problems is ensuring that comparative genomics tools keep pace with the rate at which the data are produced (Tripathi et al., 2014; Smirnov et al., 2016; Stupnikov et al., 2016). A major challenge researchers are facing practically when interacting with "big data" is that most of the relevant information requires a considerable amount of time to subset, preprocess and obtain. Therefore, novel approaches for finding, selecting and downloading specific subdata from large data repositories are required. This is particularly a problem for obtaining raw data (Musa et al., 2018).

One example for such a big data repository is the Library of Integrated Network-based Cellular Signatures (LINCS) (Subramanian et al., 2017). The LINCS L1000 data repository consists of almost two million individual files containing information about the gene expression and metadata of cell lines perturbed by chemicals of certain dosages and durations (Vempati et al., 2014). While there are several desktop or command line software tools available that are capable of processing and manually extracting subsets of large data, these tools require software installation, which can be

TABLE 1 | List of available LINCS L1000 metadata APIs.

Service (API)	Description	URL link
Cell	The cell information service provides cell line meta-information for used in the experiments.	https://clue.io/api#cells
Gene	The gene information service returns meta-information for measured and inferred genes in the LINCS dataset.	https://clue.io/api#genes
Profile	The profile information service returns meta-information for instances in the LINCS dataset.	https://clue.io/api#profiles
Pert	The pert information service returns meta-information for perturbagens in the LINCS dataset.	https://clue.io/api#perts
Plate	The plateInfo service returns plate information.	https://clue.io/api#plates

difficult and time consuming, and are only capable of processing the data locally (Duan et al., 2016; Enache et al., 2017; Fallahi-Sichani et al., 2017). Therefore, the datasets in the repository can only be analyzed if the end-user has specialized software installed. Improvements in software development but also web-based application technologies such as the Node.js and Vue.js JavaScript libraries, have led to the development of advanced web-based applications with animated and interactive features. While there are several interactive web-based tools that can access data via an application programming interface (API) (Subramanian et al., 2017), most of these tools have limited interactivity and sharing capabilities, e.g., by embedding them within web applications such as CMAP (Lamb et al., 2006). Furthermore, they are lacking an integration with biology specific analysis methods, e.g., for performing an enrichment analysis (Rahmatallah et al., 2017). Importantly, all of these tools operate on the signature level of the LINCS data, not the raw data. That means, if a user wants to select a specific subset of raw data for a dedicated analysis, there is no help available.

In order to facilitate the access and subset of raw data from the LINCS data repository we developed the L1000 Viewer. Our software is an interactive web application that does not require the user to install dedicated software, but it operates via any web browser on any operating system. Hence, it is operating system independent. Our web application provides a web interface with access to a dedicated database we created. This database utilizes the graph dependency structure between the individual data files of LINCS because *individual* does not mean *independent*. Specifically, the dependency structure is induced by the experimental conditions of the expression profiles and can be represented as a graph or network (Musa et al., 2018). In this graph, nodes correspond to data files and two data files are connected if they share experimental conditions. Our web application provides an easy-to-use interactive platform allowing the user to select subsets of raw data files that belong to specific forms of perturbation profiles, e.g., for specific cell lines, drugs, dosages and time points. This retrieval of data files is efficient and fast because of the utilization of the precomputed graph structure of the data files. In addition, we are providing software for a graphical summarization of the selected data showing various distributions of experimental parameters, e.g., sample sizes per cell line, sample sizes per concentration and sample sizes per time point. This provides valuable information for the user regarding the experimental design (Hinkelmann and Kempthorne, 2008) of follow-up computational pharmacogenomics studies based on these data.

Our paper is organized as follows. In the next section, we discuss all methods and data we use for our analysis. In the results section, we present our findings and provide results for an example application of our software. In the following sections, we discuss our results in detail. The paper finish with conclusions and an outlook.

2. METHODS

2.1. LINCS Data

The LINCS data is a vast collection of gene expression profiles that includes many experimental samples covering more than seventy human cell lines. These cell lines are populations of cells that descended from an original source cell and having the same genetic make-up. These cells have been kept alive by growing them in a culture separate from their original source (Ong et al., 2017).

Specifically, LINCS contains about $1,328,098$ gene expression profiles as a result from applying $42,553$ perturbagens ($19,811$ small molecule compounds, $18,493$ shRNAs, $3,627$ cDNAs, and 622 biologics) for a total of $476,251$ signatures (consolidating replicates) (Subramanian et al., 2017).

2.2. Metadata and Data Standards

LINCS provides an API to annotations and perturbational signatures in the L1000 data repository via a collection of HTTP-based RESTful web services. An example of such a services is the Cell Service which is a service that describes meta-information for cell lines. **Table 1** lists all the API services provided by LINCS for querying the L1000 metadata. These services support complex queries via simple HTTP GET requests that can be executed in a web browser or within most programming languages.

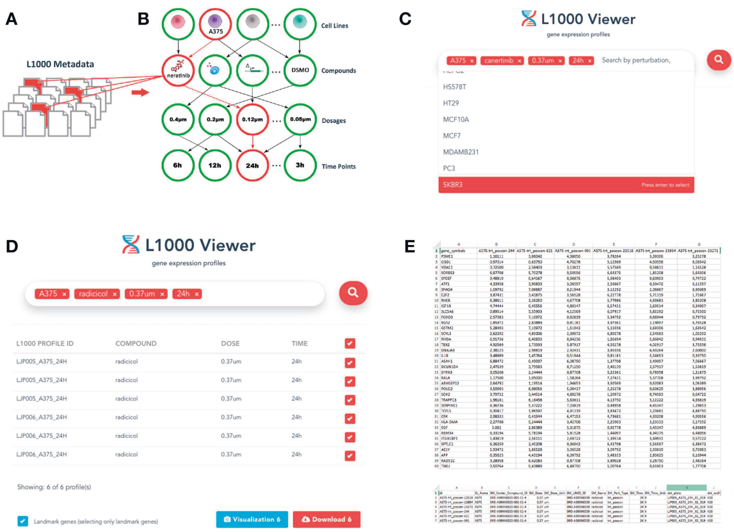
2.3. Development of the Web Application

L1000 Viewer, the web application we have developed, consists of three main parts namely; (I) the database, (II) back-end, and (III) front-end implementations.

First, in order to store the data in the back-end, we use a MongoDB database. We convert and store all the raw data into a json object structure to enable identifier reference to each profile sample in the database. This enables the data to be stored as a document-oriented structure that allows fast user queries. The document-oriented model maps to the data objects in the application code in the back-end, making the data easy to work with. The MongoDB is a distributed database at its core, therefore, it enables a horizontal scaling, high availability and faster access.

The specific document structure is constructed from the experimental conditions of the individual data profiles within the LINCS data repository. As a result we obtained a relational representation of the documents using Mongoose schema. Mongoose provides a straight-forward, schema-based solution to model json object data into relationships. It includes built-in type casting, validation, query building, and logic hooks environment that wraps the Node.js native driver. This is visualized in **Figures 1A,B**. By (I) identifying and (II) utilizing this structure, our L1000 Viewer is able to efficiently provide a list of result profiles corresponding to an user-defined query. For instance, querying for the cell line A375, the drug neratinib, a dosage of 0.12μ and a duration of 24 h (see **Figure 1C**) results in 5565 files (see **Figure 1D**) that match the query list. That means the L1000 Viewer is a search interface that represent a relational structure from the underlying individual profiles corresponding to the instances in the database collection and allows by this an efficient querying of these profiles.

Second, for the back-end component, we decided to use Node.js for the server side architecture. A Node.js server



L1000 PROFILE ID	COMPOUND	DOSE	TIME	☑
LJP005_A375_24H	radicicol	0.37um	24h	☑
LJP005_A375_24H	radicicol	0.37um	24h	☑
LJP005_A375_24H	radicicol	0.37um	24h	☑
LJP006_A375_24H	radicicol	0.37um	24h	☑
LJP006_A375_24H	radicicol	0.37um	24h	☑
LJP006_A375_24H	radicicol	0.37um	24h	☑

Showing: 6 of 6 profile(s)

☑ Landmark genes (selecting only landmark genes)

Visualization 6 Download 6

	A	B	C	D	E	F	G
1	gene_symbols	A375-trt_poscon-244	A375-trt_poscon-621	A375-trt_poscon-991	A375-trt_poscon-22518	A375-trt_poscon-22894	A375-trt_poscon-23271
2	PSME1	1,16111	3,06042	4,38056	3,78264	3,39306	3,25278
3	CISD1	0,97014	0,63750	4,70278	5,12569	4,50556	6,03542
4	VDAC1	0,72500	2,58403	3,13611	5,57569	0,58611	1,16528
5	SORBS3	6,97708	5,70278	0,50556	6,64375	1,85208	5,64306
6	SPDEF	0,48819	0,64167	0,36875	5,98403	0,60903	0,74722
7	ATF1	4,33958	3,90833	3,06597	2,56667	0,33472	6,11597
8	SPAG4	1,09792	7,06667	0,21944	3,12292	2,06667	6,95069
9	E2F2	3,87431	3,41875	3,56528	6,22778	5,71319	1,71667
10	RHEB	6,38611	3,26250	4,67708	5,77986	4,93681	3,85208
11	IGF1R	4,74444	0,45556	4,80347	6,57431	2,69514	0,34097
12	SLC5A6	0,89514	5,55903	4,12569	6,07917	5,82292	0,72500
13	FOXO3	2,57083	7,10972	0,62639	5,34792	0,66944	0,79792
14	RGS2	1,85972	2,83889	0,81181	3,97361	2,19097	6,76528
15	GSTM2	5,28403	7,10972	1,51042	5,15556	3,69306	1,63542
16	SCYL3	2,62292	4,89306	2,30972	6,80278	2,54583	1,02292
17	RHOA	6,91736	6,40833	0,84236	1,85694	5,86042	0,94931
18	TBX2	4,92569	1,73333	5,87917	4,55278	4,02917	3,75556
19	DNAJA3	2,38125	2,98819	2,92431	3,85556	6,43264	2,60000
20	IL1B	3,48889	1,45764	0,51944	0,81181	5,54653	0,33750
21	ASAH1	6,88472	0,49097	6,08750	1,37708	3,49097	7,06667
22	DCUN1D4	2,47639	3,79583	0,71250	2,40139	3,57917	1,53819
23	DYRK3	3,05208	2,24444	0,87708	5,22361	0,78958	1,21875
24	RALA	1,17500	1,05000	1,58264	7,27431	5,57708	3,99792
25	ARHGEF12	2,64792	1,19514	1,94653	3,92569	0,62083	3,06389
26	POLG2	3,50903	6,08056	5,00417	2,25278	0,60625	3,88056
27	SOX2	3,79722	3,44514	4,80278	1,20972	0,74583	0,64722
28	TRAPPC3	1,96181	6,16458	5,63611	4,13750	5,12222	4,93819
29	SERPINE1	6,36736	5,37222	7,03819	0,48958	4,45347	1,29653
30	TCFL5	6,30417	1,96597	4,91319	3,83472	2,23681	4,88750
31	CRK	2,98333	0,41944	6,47153	4,73681	4,30208	4,90556
32	HLA-DMA	2,27708	5,24444	3,42708	2,25903	3,23333	5,17292
33	EGF	5.062	2,86389	5,31875	0,92778	3,45347	4,83889
34	RBM34	6,33194	5,78194	5,81528	1,44097	6,94375	6,48056
35	ITGB1BP1	1,43819	2,56111	2,69722	1,39514	3,49931	0,57222
36	SPTLC2	6,26250	2,40208	0,36042	6,42708	5,56597	6,38472
37	ACLY	5,93472	1,86528	3,06528	3,09792	2,93819	0,72083
38	APP	6,35625	4,43194	6,09792	5,48333	2,85625	6,16944
39	RAD51C	3,28958	6,42083	0,87708	5,89028	3,28750	2,48264
40	TSKU	0,50764	0,43889	6,48750	5,00764	0,65903	1,77708

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	CL_Name	SM_Center_Compound_ID	SM_Dose	SM_Dose_Unit	SM_LINCS_ID	SM_Name	SM_Pert_Type	SM_Time	SM_Time_Unit	det_plate	det_well
2	A375-trt_poscon-22518	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP006_A375_24H_X1_B19	K10
3	A375-trt_poscon-22894	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP006_A375_24H_X2_B19	K10
4	A375-trt_poscon-23271	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP006_A375_24H_X3_B19	K10
5	A375-trt_poscon-244	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP005_A375_24H_X1_B19	K10
6	A375-trt_poscon-621	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP005_A375_24H_X2_B19	K10
7	A375-trt_poscon-991	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP005_A375_24H_X3_B19	K10

FIGURE 1 | Overview of the L1000 Viewer. **(A)** Individual files forming the LINCS data. **(B)** Our graph database inferred from the relations among the individual profiles. **(C)** The interface allowing to enter queries to search the graph database. Shown is an example query discussed in the text. **(D)** The result of a query leads to a downloadable list. **(E)** Gene expression and metadata csv files.

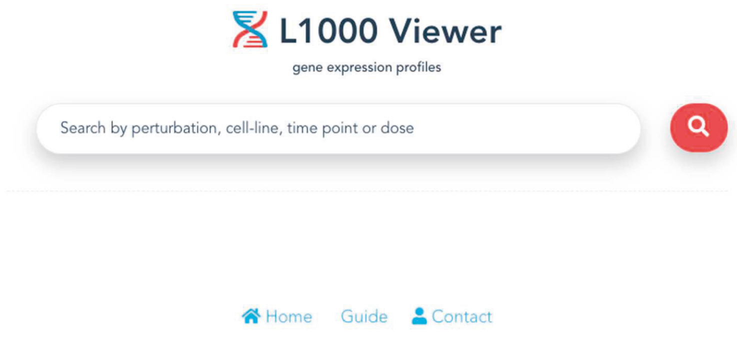


FIGURE 2 | The entry point of the L1000 Viewer application.

environment was utilized to interact with the database through custom object-data modeling (ODM) calls adopted from pseudo relational database representation in Mongoose API. The main benefit of using this model is that you can define schemas for your collections which are then enforced at the ODM layer by architecture. It also has utilities for simplifying Node's callback patterns that make it easier to work with than the standard MongoDB driver alone. In general, this approach makes it even easier to use MongoDB with Node.js. Node.js is a web application development framework that uses convention over configuration. This means it can be efficiently used to spin a back-end development environment and also allows users to quickly understand the source code and contribute to development. It also supports a rich database of user-contributed libraries called packages that ease many complicated tasks, e.g., in handling downloading and archiving requests on the server side. We use packages such as backbone.js, archiver.js, underscore.js etc. to build the back-end. The L1000 Viewer was deployed on a Linux operating system supported by the Node.js runtime library. It is deployed on an Nginx server using Linode node.

Third, for the work-flow designer on the front-end we used javascript. Specifically, we use Vue.js to created the front-end representation. Vue.js is a widely used javascript framework and the L1000 Viewer uses it for handling all client side user interactions. The connections between the components of the interface are implemented using Vue.js plugins. It provides a mechanism to display and render the structural components from HTML tags. To interactively display the large collection of drug-induced profiles, the HTML5 elements were used to layout the profiles systematically.

Overall, the model-view-controller (MVC) software architecture was used to integrate the front-end, back-end and the database. The MVC pattern of design describes the behavior of the application's data, logic, rules, and generates an output based on changes to the application. The advantage of this is it helps in focusing on a specific part of the application name, the ways information is presented to and accepted from, the user.

2.4. Graphical Summary

In addition, we provide a functionality for an interactive visualization for viewing the selected profiles on the web. A user can click on the visualization button from the search results to visualize the selected profiles in different plots (e.g., boxplot representation of the profiles etc.). The metadata information of the selected profiles are also displayed. We provide R scripts for further metadata visualizations. Specifically, we provide scripts that allow the user to generate graphical summary statistics of their metadata query results. From the download function, the user can immediately download the profiles and use the R scripts on the subset of the data that was retrieved.

3. RESULTS

We start this section by describing the basic functionality of the L1000 viewer web application we developed. Then we discuss its specific components in detail and provide an example.

3.1. General Overview of the L1000 Viewer

The L1000 Viewer has an interface allowing the user to enter queries in a disjunctive normal form (DNF), i.e., one can search for the simultaneous presence of search terms in the form,

$$\text{term}_1 \text{ AND } \text{term}_2 \text{ AND } \cdots \text{ AND } \text{term}_n \tag{1}$$

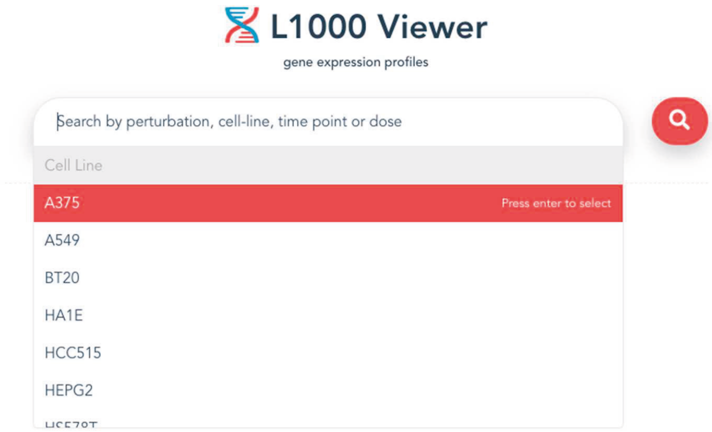


FIGURE 3 | An example of a search term that can be used to search for gene expression profiles.

For instance, in **Figure 1C** we entered the cell line A375, the drug neratinib, a dosage of 0.12μ and a duration of 24 h resulting in all profiles that are simultaneously indexed by cell line A375 AND drug neratinib AND a dosage of 0.12μ AND a duration of 24 h. The user can obtain a comprehensive list of available options directly from the L1000 interface by selecting the search field with a mouse click. This will open a pull-down menu that lists all available options that can be used as a search term in the query. Overall, the major categories for a query are cell lines, drugs and small compounds, dosages and time points.

A query finds any entity that exists among the treatment and control profiles. All queries will return a table of profiles listing unique ID numbers (e.g., LINCS profile ID, Compound), and if selected, a listing of metadata associated with the experiment will also be included in the download link. The interface is the data access point into the L1000 data repository.

The result from a query may be downloaded as a matrix of gene expression profiles. The array contains, for every gene, a binary vector representing the probe signal from the gene expression experiments (Subramanian et al., 2017). We converted the probe IDs to gene symbols for global representation. The L1000 Viewer allows the user to download complete matrices in either .csv or .csv.gz format, conferring flexibility to choose among alternative software analysis packages with optimal criteria and easy matrix subseting.

3.2. Constituting Components

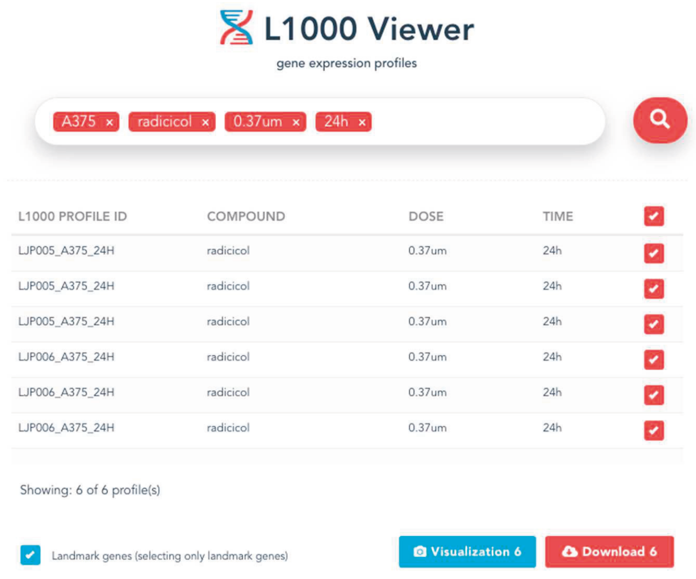
3.2.1. Data Available for Download

A large collection of almost two million L1000 gene expression profile data can be downloaded from the web interface, including the aforementioned GSE70138 from the LJP, CPC and CPD data repositories. Our application provides an easy-to-use and user-friendly interface to query the data repository, simply by searching for the desired experimental conditions.

From the L1000 Viewer web interface, metadata attributes can be used as input keyword to query the data repository. Any metadata associated with the input search can be entered in the search box. By default, the section provides four input fields for metadata: Cell, Perturbation, Dosage, and Time Point (**Figure 1**). Users can add new search terms for specific types of metadata by typing in the search box or remove one by clicking the close (x) sign on the right hand side of each keyword. The tag field is used to enter the keywords which are most descriptive of the input metadata.

3.3. Search Input

The entry point for our L1000 Viewer is to input a search term or a list of metadata query terms in the search box (see **Figure 2**) or paste a symbol (see **Figure 3**) into the search box. In order to provide guidance for setting search parameters, a query term is a list of cell lines, drug compounds, dosages or time points. The search button will only become enabled when the text box



L1000 Viewer

gene expression profiles

A375 × radicicol × 0.37um × 24h ×

L1000 PROFILE ID	COMPOUND	DOSE	TIME	☑
LJP005_A375_24H	radicicol	0.37um	24h	☑
LJP005_A375_24H	radicicol	0.37um	24h	☑
LJP005_A375_24H	radicicol	0.37um	24h	☑
LJP006_A375_24H	radicicol	0.37um	24h	☑
LJP006_A375_24H	radicicol	0.37um	24h	☑
LJP006_A375_24H	radicicol	0.37um	24h	☑

Showing: 6 of 6 profile(s)

☑ Landmark genes (selecting only landmark genes)

FIGURE 4 | An example of the displayed search results using all four metadata information.

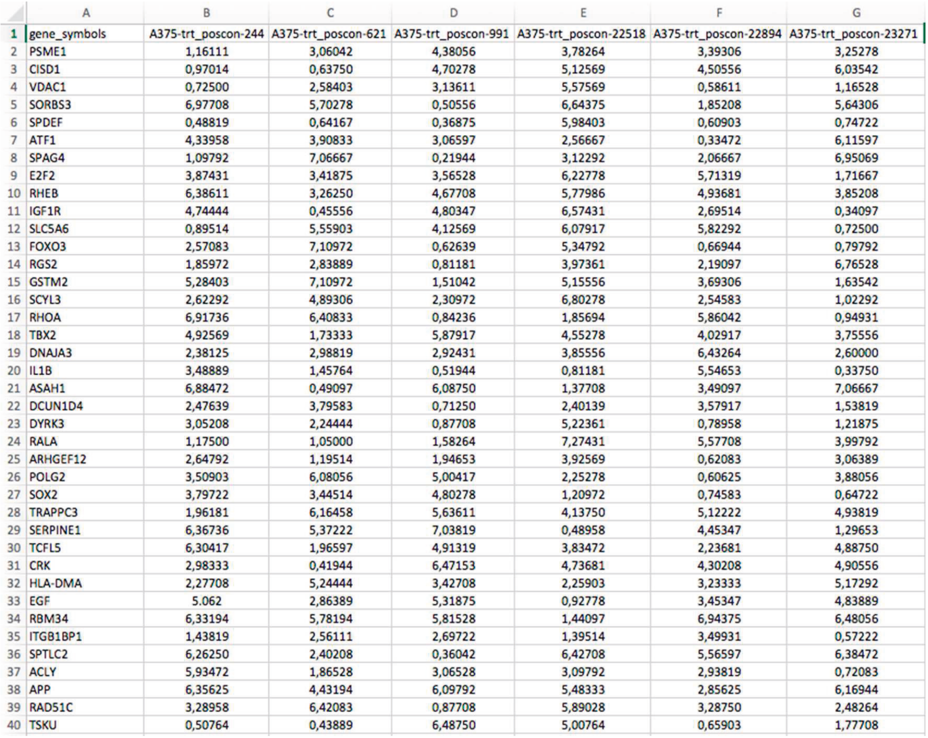
is filled with a search term, or when the text box is filled with a selection from the drop down list. By clicking the search button, the information for the top 50 samples will be displayed in a table below the search box. The interface provides the user with a user-friendly scrolling functionality for displaying more than 50 results.

3.4. Search Results

When a user successfully submits a query, the application will search and retrieve the corresponding profiles that match the user's input term and display the results. The performance of the search results will depend on the user-defined input terms. However, for any given query the application will guarantee fast results within milliseconds. In contrast, when the data is manually processed and retrieved directly from the LINCS data repositories a similar process can consume up to one day.

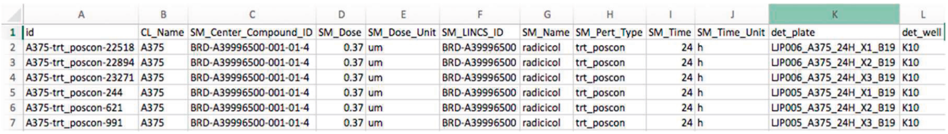
3.5. Download View

After the search results are displayed, the user can select individual profiles in the search results using the check box to fine-tune the results or decide to download all results by checking the first selection box. Then a download button will appear at the bottom of the page in the right corner. Clicking on this button will bring up the download view. The download button will generate and download gene expression profiles and signatures



	A	B	C	D	E	F	G
1	gene_symbols	A375-trt_poscon-244	A375-trt_poscon-621	A375-trt_poscon-991	A375-trt_poscon-22518	A375-trt_poscon-22894	A375-trt_poscon-23271
2	PSME1	1,16111	3,06042	4,38056	3,78264	3,39306	3,25278
3	CISD1	0,97014	0,63750	4,70278	5,12569	4,50556	6,03542
4	VDAC1	0,72500	2,58403	3,13611	5,57569	0,58611	1,16528
5	SORBS3	6,97708	5,70278	0,50556	6,64375	1,85208	5,64306
6	SPDEF	0,48819	0,64167	0,36875	5,98403	0,60903	0,74722
7	ATF1	4,33958	3,90833	3,06597	2,56667	0,33472	6,11597
8	SPAG4	1,09792	7,06667	0,21944	3,12292	2,06667	6,95069
9	E2F2	3,87431	3,41875	3,56528	6,22778	5,71319	1,71667
10	RHEB	6,38611	3,26250	4,67708	5,77986	4,93681	3,85208
11	IGF1R	4,74444	0,45556	4,80347	6,57431	2,69514	0,34097
12	SLC5A6	0,89514	5,55903	4,12569	6,07917	5,82292	0,72500
13	FOXO3	2,57083	7,10972	0,62639	5,34792	0,66944	0,79792
14	RGS2	1,85972	2,83889	0,81181	3,97361	2,19097	6,76528
15	GSTM2	5,28403	7,10972	1,51042	5,15556	3,69306	1,63542
16	SCYL3	2,62292	4,89306	2,30972	6,80278	2,54583	1,02292
17	RHOA	6,91736	6,40833	0,84236	1,85694	5,86042	0,94931
18	TBX2	4,92569	1,73333	5,87917	4,55278	4,02917	3,75556
19	DNAJA3	2,38125	2,98819	2,92431	3,85556	6,43264	2,60000
20	IL1B	3,48889	1,45764	0,51944	0,81181	5,54653	0,33750
21	ASAH1	6,88472	0,49097	6,08750	1,37708	3,49097	7,06667
22	DCUN1D4	2,47639	3,79583	0,71250	2,40139	3,57917	1,53819
23	DYRK3	3,05208	2,24444	0,87708	5,22361	0,78958	1,21875
24	RALA	1,17500	1,05000	1,58264	7,27431	5,57708	3,99792
25	ARHGEF12	2,64792	1,19514	1,94653	3,92569	0,62083	3,06389
26	POLG2	3,50903	6,08056	5,00417	2,25278	0,60625	3,88056
27	SOX2	3,79722	3,44514	4,80278	1,20972	0,74583	0,64722
28	TRAPPC3	1,96181	6,16458	5,63611	4,13750	5,12222	4,93819
29	SERPINE1	6,36736	5,37222	7,03819	0,48958	4,45347	1,29653
30	TCFL5	6,30417	1,96597	4,91319	3,83472	2,23681	4,88750
31	CRK	2,98333	0,41944	6,47153	4,73681	4,30208	4,90556
32	HLA-DMA	2,27708	5,24444	3,42708	2,25903	3,23333	5,17292
33	EGF	5.062	2,86389	5,31875	0,92778	3,45347	4,83889
34	RBM34	6,33194	5,78194	5,81528	1,44097	6,94375	6,48056
35	ITGB1BP1	1,43819	2,56111	2,69722	1,39514	3,49931	0,57222
36	SPTLC2	6,26250	2,40208	0,36042	6,42708	5,56597	6,38472
37	ACLY	5,93472	1,86528	3,06528	3,09792	2,93819	0,72083
38	APP	6,35625	4,43194	6,09792	5,48333	2,85625	6,16944
39	RAD51C	3,28958	6,42083	0,87708	5,89028	3,28750	2,48264
40	TSKU	0,50764	0,43889	6,48750	5,00764	0,65903	1,77708

FIGURE 5 | Example for a gene expression profile matrix (matrix.csv) available in the download zip folder.



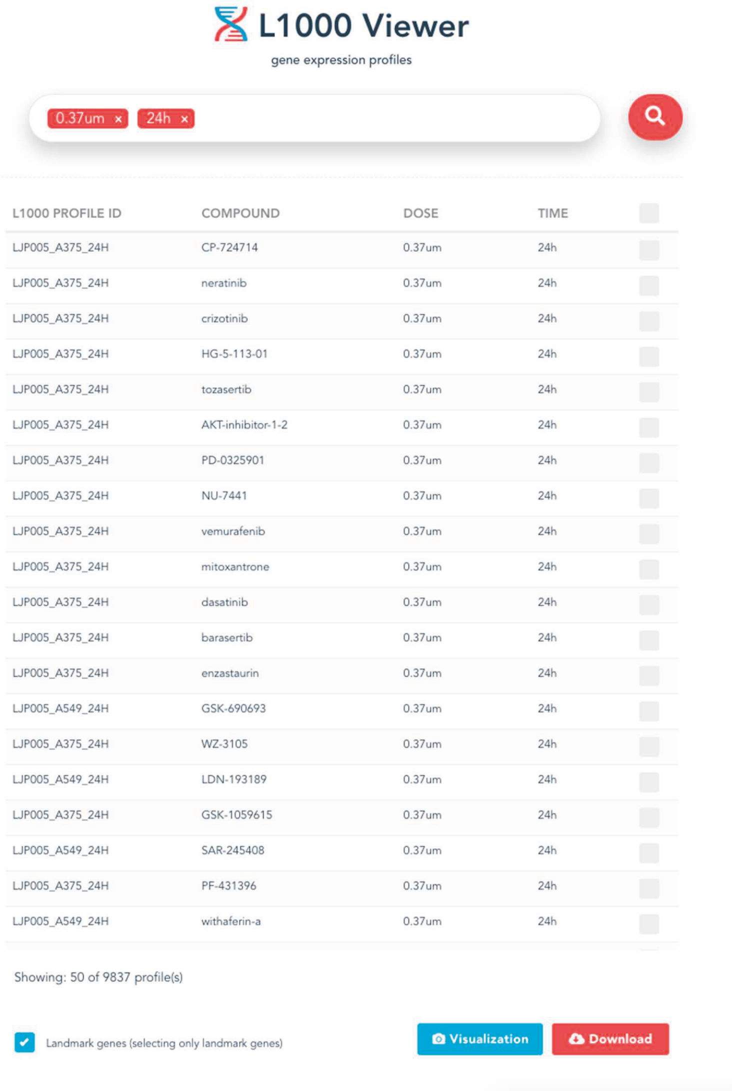
	A	B	C	D	E	F	G	H	I	J	K	L
1	id	CL_Name	SM_Center_Compound_ID	SM_Dose	SM_Dose_Unit	SM_LINCS_ID	SM_Name	SM_Pert_Type	SM_Time	SM_Time_Unit	det_plate	det_well
2	A375-trt_poscon-22518	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP006_A375_24H_X1_B19	K10
3	A375-trt_poscon-22894	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP006_A375_24H_X2_B19	K10
4	A375-trt_poscon-23271	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP006_A375_24H_X3_B19	K10
5	A375-trt_poscon-244	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP005_A375_24H_X1_B19	K10
6	A375-trt_poscon-621	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP005_A375_24H_X2_B19	K10
7	A375-trt_poscon-991	A375	BRD-A39996500-001-01-4	0.37	um	BRD-A39996500	radicicol	trt_poscon	24	h	LJP005_A375_24H_X3_B19	K10

FIGURE 6 | Example for a metadata matrix (metadata.csv) available in the download zip folder.

selected within the search results as .csv files, and will also include the metadata information associated with the profiles in a zipped folder. an example is shown in **Figure 4**.

3.6. Files

There are two files generated that are available in the zipped folder. The first is a comma-separated data matrix file (.csv) named "matrix.csv." It contains the gene expression profiles of the downloaded dataset as shown in **Figure 5**. The rows in the file correspond to all the gene symbol annotations for each profile and the columns correspond to the samples. A second file contains the meta description of the profiles. It is also a comma-separated file named "metadata.csv." This file contains the meta-information of the experiment of each profile, such as



L1000 PROFILE ID	COMPOUND	DOSE	TIME	□
LJP005_A375_24H	CP-724714	0.37um	24h	□
LJP005_A375_24H	neratinib	0.37um	24h	□
LJP005_A375_24H	crizotinib	0.37um	24h	□
LJP005_A375_24H	HG-5-113-01	0.37um	24h	□
LJP005_A375_24H	tozasertib	0.37um	24h	□
LJP005_A375_24H	AKT-inhibitor-1-2	0.37um	24h	□
LJP005_A375_24H	PD-0325901	0.37um	24h	□
LJP005_A375_24H	NU-7441	0.37um	24h	□
LJP005_A375_24H	vemurafenib	0.37um	24h	□
LJP005_A375_24H	mitoxantrone	0.37um	24h	□
LJP005_A375_24H	dasatinib	0.37um	24h	□
LJP005_A375_24H	barasertib	0.37um	24h	□
LJP005_A375_24H	enzastaurin	0.37um	24h	□
LJP005_A549_24H	GSK-690693	0.37um	24h	□
LJP005_A375_24H	WZ-3105	0.37um	24h	□
LJP005_A549_24H	LDN-193189	0.37um	24h	□
LJP005_A375_24H	GSK-1059615	0.37um	24h	□
LJP005_A549_24H	SAR-245408	0.37um	24h	□
LJP005_A375_24H	PF-431396	0.37um	24h	□
LJP005_A549_24H	withaferin-a	0.37um	24h	□

Showing: 50 of 9837 profile(s)

☑ Landmark genes (selecting only landmark genes)

FIGURE 7 | Sample query displaying drug profiles that are treated on different cell lines with 0.37 um concentration and 24 h time point. In total, 9,837 profiles have been retrieved.

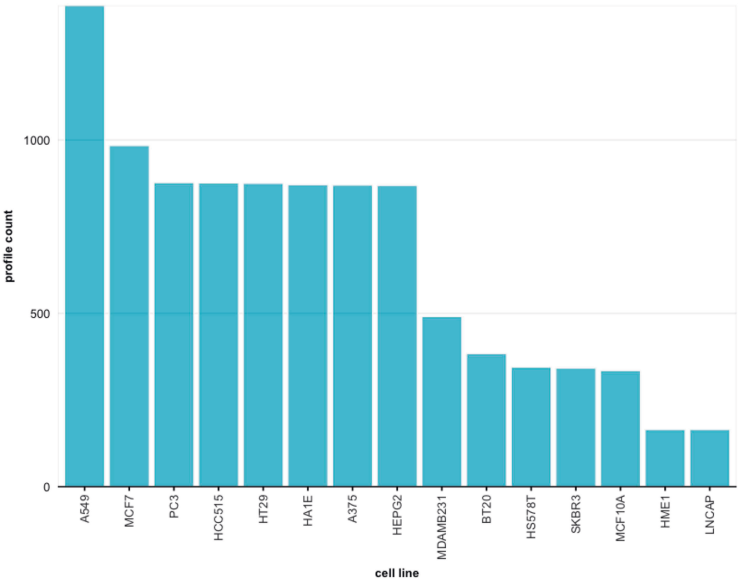


FIGURE 8 | Frequency distribution for cell lines across all experiments retrieved from the query in **Figure 7**.

time points, dosages, profile IDs, etc. The content of the file is shown in **Figure 6**.

3.7. Data Visualization: An Example

In addition to the above search and downloading capability of our L1000 viewer, described above, we provide a graphical summarization of the selected files. Specifically, we provide code that can be used to plot (in an R environment) the statistical distributions of cell lines, dosage concentrations or time points. A user can make use of the scripts to visualize the data obtained directly from a specified query.

For instance, from the query shown in **Figure 7**, setting the concentration to 0.37um and the time points to 24h, 9, 837 profiles are obtained. In **Figure 8** we show the distribution of these 9, 837 profiles over 15 cell lines. Here we leverage the metadata annotations downloaded along with the expression profiles obtained from the Cell Service API to show the distribution of each cell line.

eFor the same query we obtain the distribution of different concentrations of small molecule perturbagens, shown in **Figure 9A**. One can see that there are more than 9 different concentrations available in this data set. The compound information for small molecule perturbagens was retrieved using the Pert service API to identify unique and common compounds used in the L1000 data.

Finally, in **Figure 9B** we show the distribution of available time points in the data set. The R code and guidelines are provided from the web interface in order to subset and visualize the L1000 dataset using user specific query.

3.8. L1000 Viewer Accessibility

Access to the data indexed by the L1000 Viewer is provided through our web interface via http://L1000viewer.bio-complexity.com/. It enhances the biomedical data repository by providing a simple and fast access to LINCS raw data and allows to easily generate subsets of data. In this way, users of the web interface can extract knowledge more efficiently when interfacing with LINCS data.

3.9. Code Availability

All code associated with the L1000 Viewer project is open source. The code is available from the BitBuket repository (https://bitbucket.org/aliocee/devcrew/src/master/). The L1000 Viewer libraries are versioned according to the Semantic Versioning 2.0.0 guidelines (http://semver.org/).

4. DISCUSSION

Advances in experimental and computational methods in biomedical research are now producing large volumes of digital data objects that are rapidly accumulating. At the same time, a variety of bioinformatics tools to handle the analysis of all this data are promptly being developed and published. However, systematic linking of digital data entities for easy access are currently lacking most especially for the LINCS L1000 raw data.

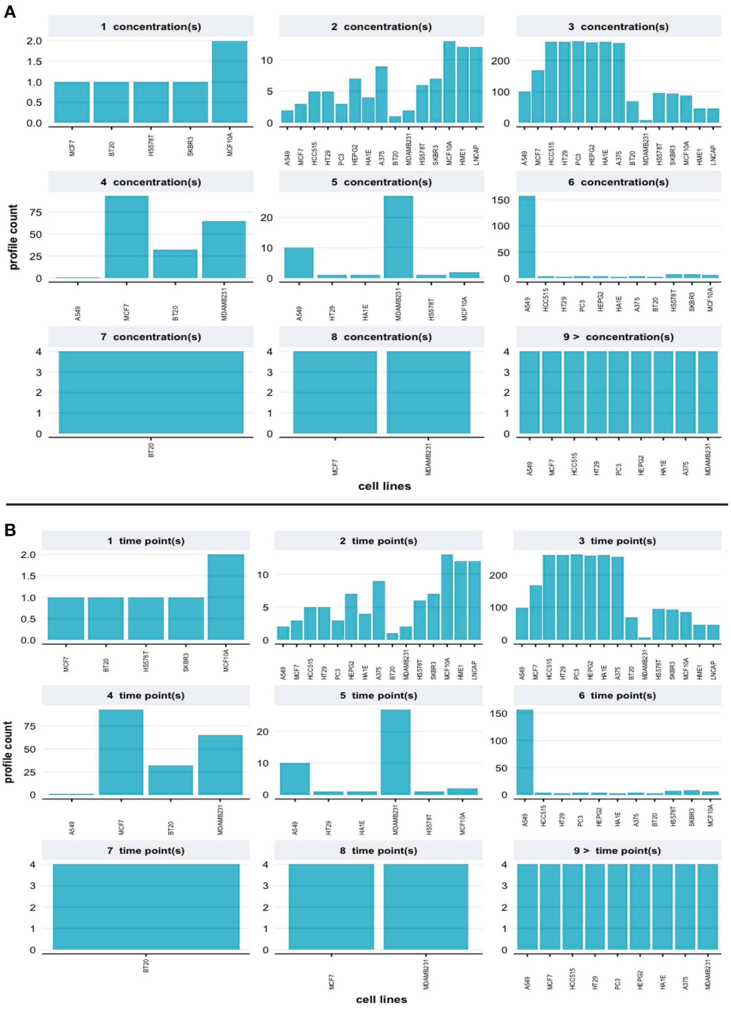


FIGURE 9 | (A) Distributions of different dosages (concentrations) of small molecules for the query in **Figure 7**. **(B)** Distributions of different time points for the query in **Figure 7**.

That means there is a gap between the data availability and how much of it can be employed in applications for extracting useful knowledge.

Previous attempts to build gene expression content-based databases have provided new support for perturbational data accessibility (Subramanian et al., 2017; Wang et al., 2018). The data within these databases is structured, and thus suitable for data access; however, most attempts to represent such data only succeeded in accomplishing this in a complex representation. For example, web-based platforms such as the CLUE Platform (Li et al., 2019), LINCS Data Portal (Koleti et al., 2017), L1000FWD (Wang et al., 2018) or iLINCS (Keenan et al., 2018) provide information about signature profiles and metadata, but there are no easy-to-use resources that enable the user to access selected raw data. Specifically, the CLUE Platform is one of the most comprehensive resources for collective knowledge about the LINCS project and L1000 data, aggregating information from over 20, 000 perturbagens and 400, 000 signature profiles. However, the CLUE Platform is very complex and does not provide direct access to the raw data.

Instead, it provides an open and free API for accessing metadata. Moreover, most of these platforms operate on metadata like annotated cell lines, proteins, and small molecules. Still they lack the simplicity and interactivity for users to access the data (Vempati et al., 2014). In comparison, our L1000 Viewer provides an easy-to-use interface for searching and downloading raw data.

The L1000 Viewer web application will enable the user to easily search the LINCS L1000 raw data via an interactive web interface. The L1000 Viewer is built using Javascript libraries, and is deployed as a Node.js application (Tilkov and Vinoski, 2010) in order to provide quick access. Its front end interface utilizes the core Vue.js libraries (You, 2017) and all gene expression and metadata are stored in a MongoDB database. Furthermore, we developed and integrated an API in our application that enables users to search the LINCS data repository and to automatically generate data for download.

In contrast to stand-alone software that needs to be installed locally on a computer, our L1000 Viewer is a web application that can be accessed via any web browser without the need of installing software on a computer locally. This makes it not only easy to access but ensures also an operating system independent functioning.

5. CONCLUSION

In this paper, we introduced the L1000 Viewer (http://L1000viewer.bio-complexity.com/), a search engine and graphical web interface for the LINCS data repository. The core of our L1000 Viewer is a database that utilizes the intricate dependency structure among the files in the LINCS data. This resulted in a reorganization of the files and enables efficient search capabilities based on graph-oriented operations.

Overall, the L1000 Viewer provides a useful tool for efficiently accessing exclusive information from the LINCS data repository that can be utilized for computational pharmacogenomics studies (Hopkins, 2008; Davis and Chawla, 2011; Emmert-Streib et al., 2013; Himmelstein et al., 2017), e.g., for drug repurposing and cancer therapeutics, as well as for understanding the composition and relationships between genes, drugs and diseases.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138.

AUTHOR CONTRIBUTIONS

FE-S conceived this study. AM and ST performed the analysis. AM, ST, MD, and FE-S wrote the paper and approved the final version of the manuscript.

FUNDING

MD thanks the Austrian Science Funds for supporting this work (project P 30031).

ACKNOWLEDGMENTS

We would like to thank Ricardo de Matos Simoes for fruitful discussions.

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Davis, D. A., and Chawla, N. V. (2011). Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PLoS ONE* 6:e22670. doi: 10.1371/journal.pone.0022670

Duan, Q., Reid, S. P., Clark, N. R., Wang, Z., Fernandez, N. F., Rouillard, A. D., et al. (2016). L1000cds2: lincs l1000 characteristic direction signatures search engine. *npj Syst. Biol. Appl.* 2:16015. doi: 10.1038/npjsba.2016.15

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Emmert-Streib, F., and Dehmer, M. (2019). Defining data science by a data-driven quantification of the community. *Mach. Learn. Knowledge Extract.* 1, 235–251. doi: 10.3390/make1010015

Emmert-Streib, F., Tripathi, S., de Matos Simoes, R., Hawwa, A., and Dehmer, M. (2013). The human disease network: opportunities for classification, diagnosis and prediction of disorders and disease genes. *Syst. Biomed.* 1, 20–28. doi: 10.4161/sysb.22816

Enache, O. M., Lahr, D. L., Natoli, T. E., Litichevskiy, L., Wadden, D., Flynn, C., et al. (2017). The gctx format and cmapPy, R, M packages: resources for the optimized storage and integrated traversal of dense matrices of data and annotations. *Bioinformatics* 35, 1427–1429 doi: 10.1093/bioinformatics/bty784

Fallahi-Sichani, M., Becker, V., Izar, B., Baker, G. J., Lin, J. R., Boswell, S. A., et al. (2017). Adaptive resistance of melanoma cells to raf inhibition via reversible induction of a slowly dividing de-differentiated state. *Mol. Syst. Biol.* 13:905. doi: 10.15252/msb.20166796

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., et al. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 6:e26726. doi: 10.7554/eLife.26726

Hinkelmann, K., and Kempthorne, O. (2008). *Design and Analysis of Experiments: Introduction to Experimental Design*. Chichester: Wiley-Interscience.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690. doi: 10.1038/nchembio.118

Jin, X., Wah, B. W., Cheng, X., and Wang, Y. (2015). Significance and challenges of big data research. *Big Data Res.* 2, 59–64. doi: 10.1016/j.bdr.2015.01.006

Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., Koplev, S., He, E., Torre, D., et al. (2018). The library of integrated network-based cellular signatures nih program: system-level cataloging of human cells response to perturbations. *Cell Syst.* 6, 13–24. doi: 10.1016/j.cels.2017.11.001

Koleti, A., Terryn, R., Stathias, V., Chung, C., Cooper, D. J., Turner, J. P., et al. (2017). Data portal for the library of integrated network-based cellular signatures (lincs) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res.* 46, D558–D566. doi: 10.1093/nar/gkx1063

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Li, A., Lu, X., Natoli, T., Bittker, J., Sipes, N. S., Subramanian, A., et al. (2019). The carcinogenome project: *in vitro* gene expression profiling of chemical perturbations to predict long-term carcinogenicity. *Environ. Health Perspect.* 127:047002. doi: 10.1289/EHP3986

Ma'ayan, A., Rouillard, A. D., Clark, N. R., Wang, Z., Duan, Q., and Kou, Y. (2014). Lean big data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.* 35, 450–60. doi: 10.1016/j.tips.2014.07.001

Musa, A., Dehmer, M., Yli-Harja, O., and Emmert-Streib, F. (2018). Exploiting genomic relations in big data repositories by graph-based search methods. *Mach. Learn. Knowl. Extr.* 1, 205–210. doi: 10.3390/make1010012

Musa, A., Ghoraie, L. S., Zhang, S.-D., Glazko, G., Yli-Harja, O., Dehmer, M., et al. (2017). A review of connectivity map and computational approaches in pharmacogenomics. *Brief. Bioinform.* 18:903. doi: 10.1093/bib/bbx023

Ong, E., Xie, J., Ni, Z., Liu, Q., Sarntivijai, S., Lin, Y., et al. (2017). Ontological representation, integration, and analysis of lincs cell line cells and their cellular responses. *BMC Bioinformatics* 18:556. doi: 10.1186/s12859-017-1981-5

Rahmatallah, Y., Zybailov, B., Emmert-Streib, F., and Glazko, G. (2017). GSAR: Bioconductor package for gene set analysis in R. *BMC Bioinform.* 18:61. doi: 10.1186/s12859-017-1482-6

Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., and Olsen, C. (2016). PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* 32, 1244–1246. doi: 10.1093/bioinformatics/btv723

Smith, F. J. (2006). Data science as an academic discipline. *Data Sci. J.* 5, 163–164. Available online at: https://www.jstage.jst.go.jp/article/dsj/5/0/5_0_163/_article/-char/ja/

Stupnikov, A., Tripathi, S., de Matos Simoes, R., McArt, D., Salto-Tellez, M., Glazko, G., et al. (2016). samExploreR: exploring reproducibility and robustness of RNA-seq results based on SAM files. *Bioinformatics* 32, 3345–3347. doi: 10.1093/bioinformatics/btw475

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452.e17. doi: 10.1016/j.cell.2017.10.049

Tilkov, S., and Vinoski, S. (2010). Node.js: using javascript to build high-performance network programs. *IEEE Int. Comput.* 14, 80–83. doi: 10.1109/MIC.2010.145

Tripathi, S., Dehmer, M., and Emmert-Streib, F. (2014). NetBioV: an R package for visualizing large-scale data in network biology. *Bioinformatics* 30, 2834–2836. doi: 10.1093/bioinformatics/btu384

Vempati, U. D., Chung, C., Mader, C., Koleti, A., Datar, N., Vidović, D., et al. (2014). Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the library of integrated network-based cellular signatures (lincs). *J. Biomol. Screen.* 19, 803–816. doi: 10.1177/1087057114522514

Wang, Z., Lachmann, A., Keenan, A. B., and Ma'ayan, A. (2018). L1000fwd: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* 34, 2150–2152

Woo, J. H., Shimoni, Y., Yang, W. S., Subramaniam, P., Iyer, A., Nicoletti, P., et al. (2015). Elucidating compound mechanism of action by network perturbation analysis. *Cell* 162, 441–451. doi: 10.1016/j.cell.2015.05.056

You, E. (2017). *Vuejs Javascript Framework*. Available online at: https://vuejs.org/

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLICATION IV

Harnessing the biological complexity of Big Data from LINCS gene expression signatures

A. Musa, S. Tripathi, M. Kandhavelu, M. Dehmer and F. Emmert-Streib

PLOS ONE 13.8 (2018), 1–16
DOI: 10.1371/journal.pone.0201937

Publication reprinted with the permission of the copyright holders

RESEARCH ARTICLE

Harnessing the biological complexity of Big Data from LINCS gene expression signatures

Aliyu Musa[1,2], Shailesh Tripathi[1,3], Meenakshisundaram Kandhavelu[2,4], Matthias Dehmer[3,5], Frank Emmert-Streib[1,2]*

1 Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology, Tampere, Finland, **2** Molecular Signaling Lab, Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, Tampere, Finland, **3** University of Applied Sciences Upper Austria, Steyr, Austria, **4** BioMediTech Institute, Tampere University of Technology, Tampere, Finland, **5** Institute for Bioinformatics and Translational Research, UMIT- The Health and Life Sciences University, Hall in Tyrol, Austria

* frank.emmert-streib@tut.fi

OPEN ACCESS

Citation: Musa A, Tripathi S, Kandhavelu M, Dehmer M, Emmert-Streib F (2018) Harnessing the biological complexity of Big Data from LINCS gene expression signatures. PLoS ONE 13(8): e0201937. https://doi.org/10.1371/journal.pone.0201937

Editor: Hong Wanjin, Institute of Molecular and Cell Biology, SINGAPORE

Received: May 15, 2018

Accepted: July 24, 2018

Published: August 29, 2018

Data Availability Statement: The LINCS L1000 data is publicly available for download from (https://clue.io/data) and from the Gene Expression Omnibus (GEO) database with accession number GSE92742 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742).

Funding: Matthias Dehmer thanks the Austrian Science Funds for supporting this work (project P26142).

Competing interests: The authors have declared that no competing interests exist.

Abstract

Gene expression profiling using transcriptional drug perturbations are useful for many biomedical discovery studies including drug repurposing and elucidation of drug mechanisms (MoA) and many other pharmacogenomic applications. However, limited data availability across cell types has severely hindered our capacity to progress in these areas. To fill this gap, recently, the LINCS program generated almost 1.3 million profiles for over 40,000 drug and genetic perturbations for over 70 different human cell types, including meta information about the experimental conditions and cell lines. Unfortunately, Big Data like the ones generated from the ongoing LINCS program do not enable easy insights from the data but possess considerable challenges toward their analysis. In this paper, we address some of these challenges. Specifically, first, we study the gene expression signature profiles from all cell lines and their perturbagents in order to obtain insights in the distributional characteristics of available conditions. Second, we investigate the differential expression of genes for all cell lines obtaining an understanding of condition dependent differential expression manifesting the biological complexity of perturbagents. As a result, our analysis helps the experimental design of follow-up studies, e.g., by selecting appropriate cell lines.

Introduction

Despite continuous progress in our understanding of the genetic origin of diseases our ability of treating and curing such diseases lacks far behind [1–5]. For this reason, it has been proposed to utilize genomic information for the development of drugs to directly translate results from basic research to clinical applications [6, 7]. A particular example of such a genome-scale project is the Library of Integrated Network-based Cellular Signatures (LINCS) program [8].

The LINCS program [8] (https://clue.io), generated genetic and molecular signatures of human cell lines in response to a variety of perturbations. Specifically, a vast library of gene expression profiles that includes over one million experiments covering more than seventy human cell lines has been generated by measuring the expression values for 978 landmark genes, hence, called the LINCS L1000 data. These data include experiments using over 20,000 chemical perturbagens (small drug molecules), namely drug compounds added to the cell culture to induce changes in the gene expression profile. In addition, there are genetic perturbation experiments targeting a single gene to control its expression level, either suppressing it (knockdown) or enhancing it (overexpression). The LINCS L1000 data is publicly available for download from (https://clue.io/data) and from the Gene Expression Omnibus (GEO) database with accession number GSE92742 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742).

The LINCS L1000 data provide an unprecedented compendium of both structural and transcriptomic drug data. However, the availability of such *Big Data* [9, 10] like LINCS L1000, provide also major challenges for their analysis requiring the development of novel approaches and methods. Examples of such approaches for exploring the LINCS L1000 data can be found in [11]. This study focused on finding structural similarities of drugs with a combination of 3D molecular structure to show significant associations of drugs with similar transcriptional changes, supporting the usage of drug-related data [11]. Another study showed that perturbational data can be used for finding common and cell-type specific responses to anti-cancer drug [12]. One major challenge in drug discovery is identifying biochemical interactions of small drug molecules [13]. For this reason, vast effort has been put into discovering the drug MoA and understanding the genetic interactions within cells that will lead to a much fuller understanding of how organisms develop interactions at a cellular level, as well as how diseases such as cancer affect cells and how they can be treated [14, 15]. Several methods such as high-throughput screen is used in identifying interactions of small drug molecules showing activity in biological assays (cellular assays, enzyme activity assays, binding assay) for a single therapeutic target or pathway of interest [16–18]. These examples show the vast use of such data in drug discovery applications.

One problem of the LINCS program is that it constitutes an ongoing endeavor. That means at present there is no foreseeable end when the last samples are deposited. This feature is shared with other genomic data repositories, e.g., Gene Expression Omnibus (GEO) [19], Protein Data Bank (PDB) [20] or Reactome (database of reactions, pathways and biological processes) [21]. All of these data repositories have in common that the data have not been generated from one laboratory sponsored by one funding agency, but multiple independently funded laboratories generated and are still generating data to date. As a consequence, the information contained in such repositories and also in LINCS is a function of time. A problem resulting from this and the fact that multiple laboratories contribute to these data is the lack of global overview statistics that characterize the content of the data. This lack of overview statistics hampers the downstream usage of the LINCS L1000 data for any data analytics application, as outlined above, severely because essentially any statistical data analysis requires knowledge of available sample sizes and available experimental conditions in order to design an analysis properly [22, 23]. For instance, one would like to know how many experiments have three or more replicates for cell line HA1E? Or how many samples are available for cell line A375 having been exposed to four different drug dosages? These and similar questions are currently unanswered and there is no simple way for obtaining such information. For this reason regular updates of the content of such data repositories need to be provided in order to inform the community.

In this paper, we address this problem by exploring and summarizing the LINCS L1000 data as provided by the signature profiles. Specifically, we analyze the LINCS L1000 data for two different layers. In the first layer we focus on the signature profiles themselves and in the second layer we focus on the differentially expression of genes derived from the signature profiles. This means we are moving from overview distributions on a basic level to characterizations of the biological activity of the cell lines in dependence on multivariate conditions, as given by, e.g., the number of replicates or the duration of applied drug perturbations. This will allow to gain insights into the distributions of cell types, time points and small drug molecule dosages across multiple compounds and all experiments conducted so far.

Methods

LINCS L1000 dataset

The LINCS L1000 dataset comprises 5806 genetic perturbations (e.g., single gene knockdown and over-expression) and 16,425 perturbations induced by chemical compounds (e.g., drugs) [24]. So far about 1.3 million gene expression profiles have been generated and collected for this project using the L1000 technology [25]. The L1000 platform has been developed at the Broad Institute by the connectivity map (CMap) team to facilitate rapid, flexible and high-throughput gene expression profiling at lower costs. Specifically, this means the L1000 technology measures expression for 978 *landmark* genes and expression values for the remaining transcriptome is estimated using a computational model based on data from the Gene Expression Omnibus (GEO) [26].

Metadata pipeline

The LINCS data API provides a programmatic pipeline to annotations and perturbational signatures in the L1000 dataset via a collection of HTTP-based RESTful web services. An example for such a service is ‘Cell Service’, which is a service that describes the cell line meta-information. Table 1 lists all the API services provided by the LINCS API for querying the L1000 metadata. These services support complex queries via simple HTTP GET requests that can be executed in a web browser or with most programming languages.

Results

The LINCS L1000 data is a vast collection of gene expression profiles and meta information that includes many experimental samples covering more than seventy human cell lines. These cell lines are populations of cells descended from an original source cell and having the same genetic make-up, kept alive by growing them in a culture separate from their original source [27]. In the following, we analyze the LINCS L1000 data for two different layers. The first layer focuses on the signature profiles themselves and the second layer on the differentially expression of genes derived from the signature profiles. This means we are moving from overview

Table 1. List of LINCS L1000 metadata APIs.

Service	Description	URL link
Cell Service	The Cell information service returns cell line information.	https://clue.io/api#cells
Gene Service	The Gene information service returns meta-information for measured and inferred genes in the LINCS dataset.	https://clue.io/api#genes
Profile Service	The Profile information service returns meta-information for instances in the LINCS dataset.	https://clue.io/api#profiles
Pert Service	The Pert information service returns meta-information for perturbations in the LINCS dataset.	https://clue.io/api#perts
Plate service	The PlateInfo service returns plate information.	https://clue.io/api#plates
Signatures	The Signature information service returns meta-information for signatures in the LINCS dataset.	https://clue.io/api#signatures

https://doi.org/10.1371/journal.pone.0201937.t001 distributions on a basic level to characterizations of the biological activity of the cell lines in dependence on multivariate conditions, as given by, e.g., the number of replicates or the duration of applied drug perturbations. Hence, this provides an understanding of the biological functions effected by the perturbations.

A. Signature profiles

Cell line and small molecule annotations. Various cancer cell lines and non-transformed primary cultures were used to represent disease models in the LINCS L1000 data [28]. To enable an integration and analysis of large cell-based screening profiles in the LINCS project, the cell lines were annotated with labeled terms to identify the associated organs and diseases. In Fig 1 we show the overall distribution of profiled samples for 71 cell lines across all experiments. These counts include all the corresponding cell line profiles. For obtaining this information, we used the metadata annotations that are available via the Cell Service API. By summation over all cell lines in Fig 1 we find that, currently, the total number of signature profiles (excluding the profiles treated with knockdown and overexpression genes) is 215,224. This number is much smaller than the 1.3 million raw gene expression samples because the replicated raw sample have been summarized for obtaining the signature profiles resulting from a comparison of treatment with control conditions.

From Fig 1 it is clear to see that there are many cell lines that are not highly profiled and therefore have low profile counts. For this reason, in the following we focus on the 9 cell lines

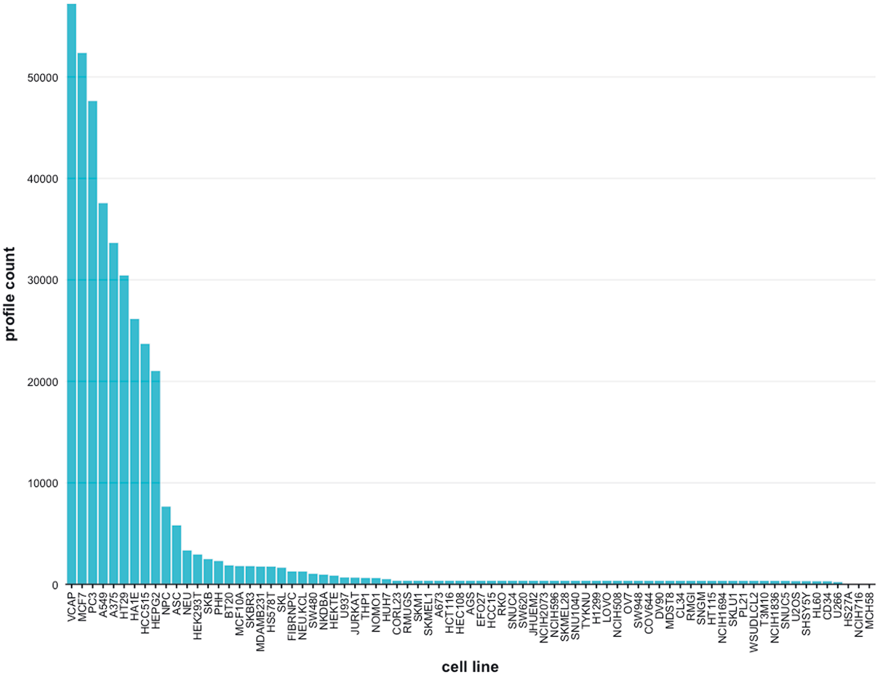


Fig 1. Cell line signature profile counts. The drug signature profile count distribution is shown for all 71 cell lines across all experiments in the LINCS L1000 dataset. Each bar gives the number of available signature profiles per cell line.

https://doi.org/10.1371/journal.pone.0201937.g001

Table 2. Cell lines with the highest number of available signature profiles in the LINCS L1000 data and their corresponding annotation according to the Cell Service API.

Cell line	Profile count	Tissue
A375	33,656	Skin
A549	37,577	Lung
HCC515	23,714	Lung
HA1E	26,164	Kidney
HEPG2	21,032	Liver
HT29	30,449	Colon
MCF7	52,373	Breast
PC3	21,032	Prostate
VCAP	21,032	Prostate

https://doi.org/10.1371/journal.pone.0201937.t002

with the highest profile counts. In Table 2 we show the count distribution of these 9 cell lines, each containing more than 20,000 profiles.

The LINCS L1000 data include experiments for more than 20,000 small molecule perturbations. The perturbations are applied to the cell culture to induce changes in the gene expression profiles. Furthermore, there are genetic perturbation experiments targeting single genes to control their expression levels, by either suppressing or enhancing them [29]. Detailed information for small molecule perturbations can be retrieved using the Pert Service API that identifies unique and common drugs used in the L1000 dataset. In Fig 2 we show the count distribution of 6 different treatment and control samples including genetic and small molecule perturbations. The count distributions shown correspond to the same 9 cell lines as in Table 2.

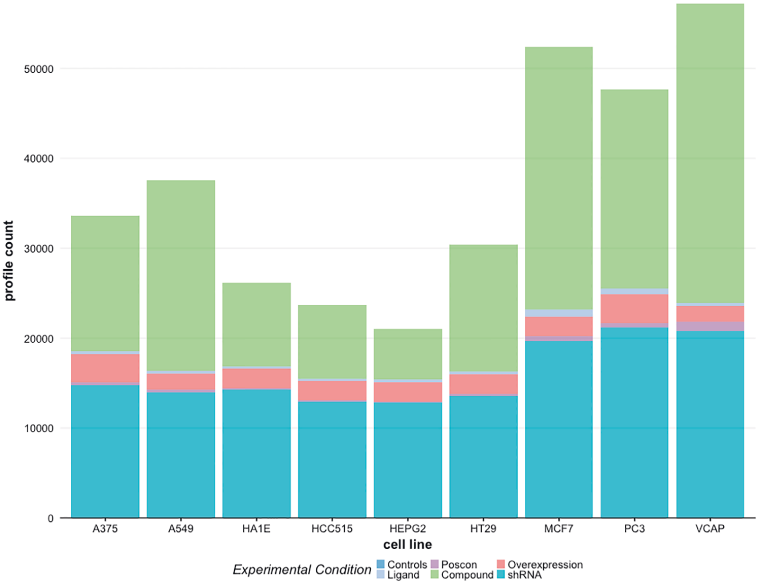


Fig 2. Distribution of experimental conditions for 9 highly profiled cell lines. Each stack bar shows the proportion of available profiles for different small molecules and controls used for the experimental condition.

https://doi.org/10.1371/journal.pone.0201937.g002

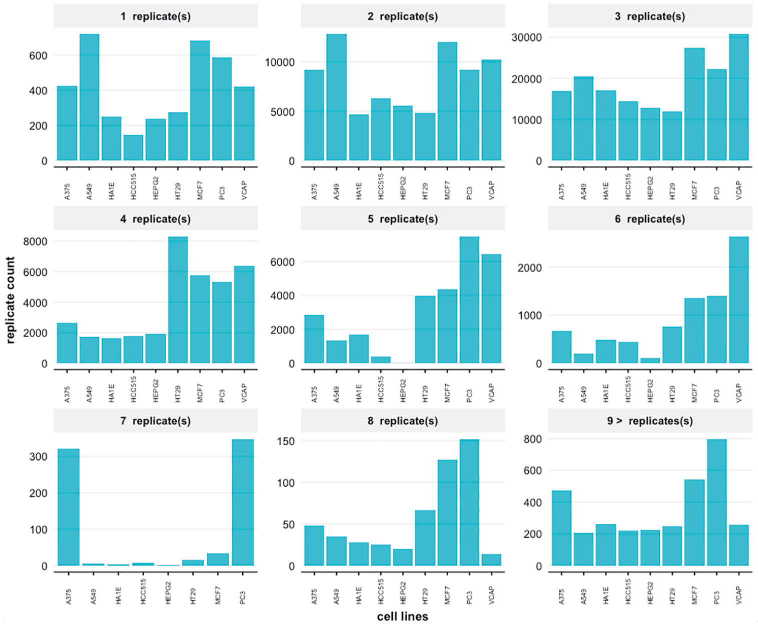


Fig 3. Distributions of experimental replicates for the signature profiles. The number of available replicates is shown for small molecule treatments in the LINCS L1000 data for 9 highly profiled cell lines.

https://doi.org/10.1371/journal.pone.0201937.g003

The 6 experimental conditions considered are: controls, ligands, poscons, compounds, overexpression and shRNAs. As one can see the number of controls and compounds is always highest for all cell lines followed by the number of overexpressed profiles.

Experimental replicates have been investigated and found to be useful in simulation and in boosting analysis [30] and decreasing the number of replicates will adversely affect the power of experiments [30, 31]. For this reason we studied the distribution of replicate experiments of the LINCS L1000 data. From this we find that the plate variation is ranging mostly between 1 to 8 replicates with the majority of samples having 3 replicates. There are also conditions for which more than 9 replicates have been generated, however, these are rare covering only 1% of all profiles, whereas 1 to 8 replicates cover 99%. The largest number of replicates observed is 27, e.g., found for cell line VCAP, drug Vorinostat, a dosage of 10um and a time duration of 24h. In Fig 3 we show the number of replicated experiments cross the 9 selected cell lines. The figure includes also information about 9 or more replicates and shows that the availability various greatly between the cell lines.

Next, we show in Fig 4 results for the number of different dosages (concentrations) applied to the 9 highly profiled cell lines. The figure shows distributions for 8 different concentrations and 9 or more concentrations. However, almost 99% of the treated samples are measured for 1 to 8 different concentrations. From the available 49,400 perturbations, most of them were tested for a duration of 6, 24, 48, 96 and 120 hours. Overall, the number of cell lines per compound represented in the treatments ranged from 1 to 8 different time duration points (see Fig A in S1 File). Around 99% of the perturbations affected at least one gene significantly in a single cell line after treatment with the varying number of time points.

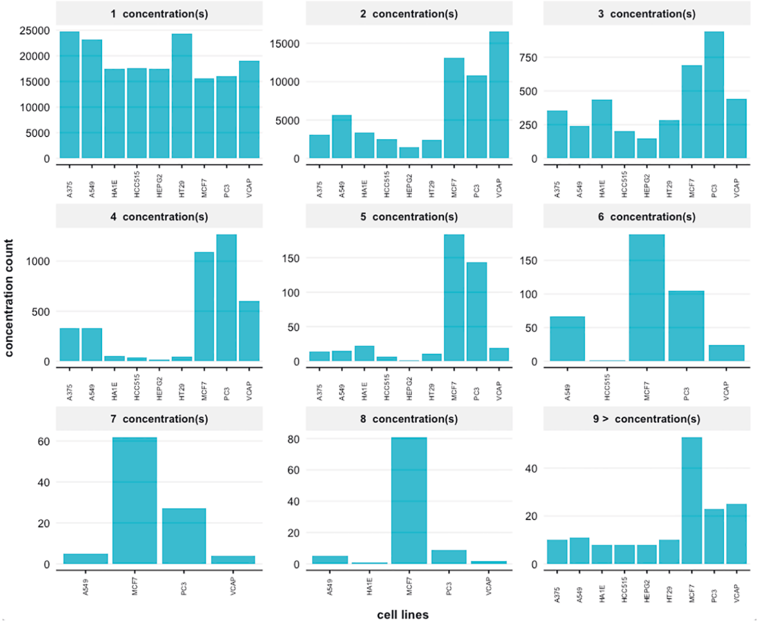


Fig 4. Distributions of unique dosages for the signature profiles. The number of available profiles is shown for different dosages (concentrations) of small molecules for 9 highly profiled cell lines.

https://doi.org/10.1371/journal.pone.0201937.g004

B. Differentially expression of genes

Differentially expression of genes and small molecule diversity. Our next analysis focuses on the activity level of the gene expression data as quantified by differentially expressed genes. For this analysis we utilized the L1000 raw z-scores from the GEO repository and pre-processed these by using the R L1000 tools [32]. We utilized the signature meta-information in Signature Service API for selecting the same subset of 9 cell lines as in Table 2 (with highest signature counts across all cell lines). Here a signature for a small molecule is defined as a vector of z-score values, each representing differential expression of genes profile between small molecule treated samples and control samples. In total there are 169,239 z-score signature profiles for the 9 cell lines that satisfied the well- and plate-based quality control. This signature profile subset comprises 20,009 small molecules (out of 49,400 perturbations) that were repeatedly measured between 1 to 8 times. To further simplify the data and the quality of the analysis, we selected 6, 24 and 48*h* time points. In total this leaves us 158,054 signature profiles (i.e., any combination of the small molecule, time, and cell line) for our analysis. These signature profiles come from experiments that were carried out on 391 multi-wells, where 362 wells were used for treatment and 29 DSMO wells were for control vehicles.

In order to obtain the number of differentially expressed genes between treatment and control samples for each of the 384 plates we used the z-score signature vectors obtained from the Signature Service setting the z-score threshold to $>$ 2.0 and $<$ -2.0 for up- and down-regulated genes respectively. For measuring the signature type effects that have been shown to be robust in biological interpretations, we use the assigned z-score thresholds to measure the biological

Table 3. Summary of z-score signature profiles resulting in differentially expressed genes (DEG) between treatment and control samples for the 9 cell lines in Table 2.

Differentially expressed genes	Signature profiles	Small molecules
No significant gene	24	19
At least 1 significant gene	158,030	19,957
At least 50 significant genes	58,739	15,714
At least 100 significant genes	23,867	8,211
Total	**158,054**	**20,009**

https://doi.org/10.1371/journal.pone.0201937.t003

effects encoded in the gene expression data. We found that 19,957 small molecules from 20,009 that are used in 158,054 signature profiles yielded at least one gene that is significantly differentially expressed when compared with the corresponding control samples. We further found that 15,714 small molecules reveal significant differences for at least 50 genes, and 8, 211 small molecules are differentially expressed for at least 100 or more genes. Table 3 summarizes these results.

Cell type specific differentially gene expression. Since not all cell lines measure the transcription effects of small molecules for the same time points, we subset the treatments according to cell lines and evaluate the number of significant genes for the 9 cell lines separately. In Fig 5 we show our results giving the number of signature profiles for each cell line for three categories. The three categories correspond to (I) at least one significant gene, (II) at least 50 significant genes, and (III) at least 100 significant genes when compared with vehicle controls.

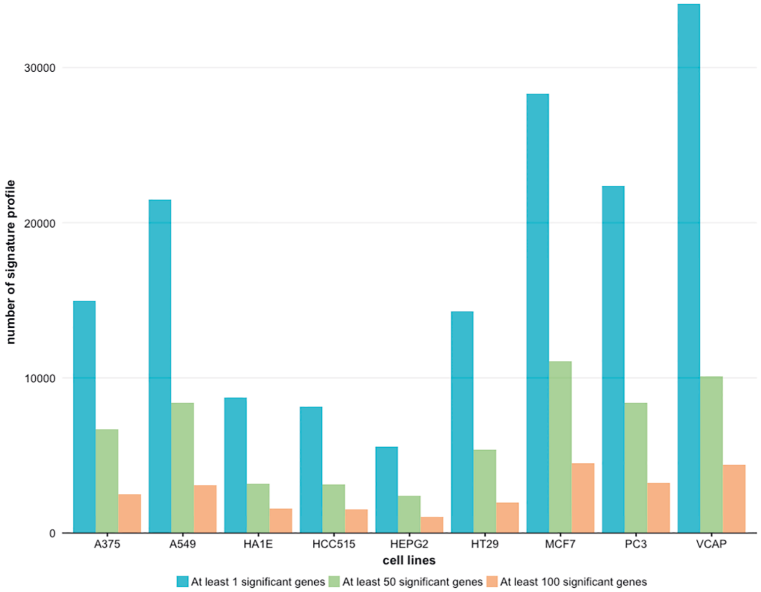


Fig 5. Number of significant profiles found when comparing signature profiles of treatment and control samples. The cell lines are categorized according to the number of DEGs and the DEG have been estimated based on the z-score signatures profiles.

https://doi.org/10.1371/journal.pone.0201937.g005

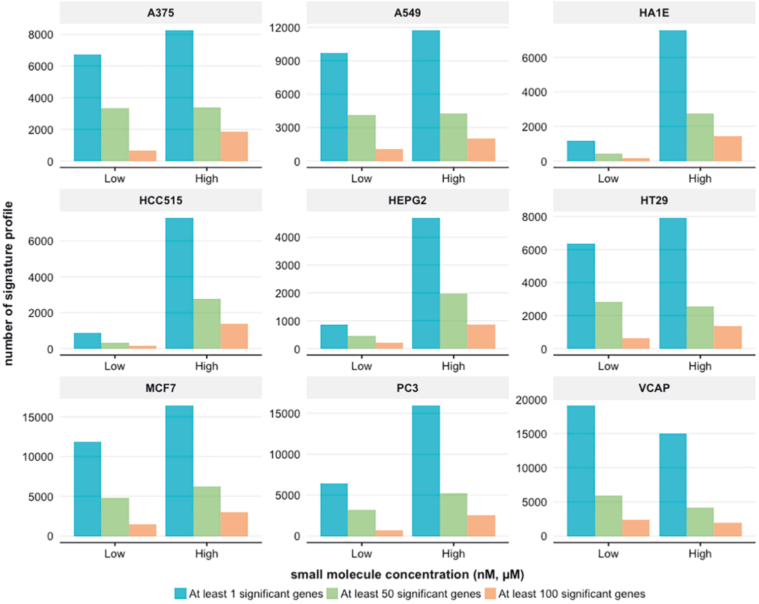


Fig 6. Dosage specific differentially gene expression. The differential expression of genes for 9 cell lines is shown categorized in Low and High dosages of small molecules.

https://doi.org/10.1371/journal.pone.0201937.g006

Since there were only 24 profiles with no significant genes in total, this category is not shown in the figure.

Dosage specific differentially gene expression. For studying the effect of drug dosages we repeated a similar analysis as above. Specifically, we systematically classified the small molecule dosages into two categories for 'low' and 'high' concentrations. The 'low' concentration group contains all measurements in nanomolar (nM) and doses less than or equal to 5 micromolar (μM) while the 'high' concentration group includes all measurements greater than 5 μM. In total, we find 63,113 and 94,941 signature profiles for low and high dosages respectively. In Fig 6, the number of differentially expressed genes is shown for the 9 cell lines and the two dosage categories. From this we observe two different behaviors. First, the number of differentially expressed genes increases with time, e.g., cell line A375 or A549. Second, the number of differentially expressed genes decreases with time. This behavior is only observed for cell line VCAP. The first type of behavior is expected because higher dosages of drugs should result in more severe changes in the expression of genes. The reverse of this effect for cell line VCAP, a prostate cancer cell line, averaged over all drugs is counter intuitive and points to follow-up investigations.

Drug perturbation specific differentially gene expression. Next, we analyze the number of differentially expressed genes according to the time duration of the treatment with small molecules. In Fig 7 we show results for 6 and 24 hours. From this we again observe two different behaviors. First, the number of differentially expressed genes increases with time, e.g., cell line A375 or A549. Second, the number of differentially expressed genes decreases with time, e.g., cell line HA1E or HCC515.

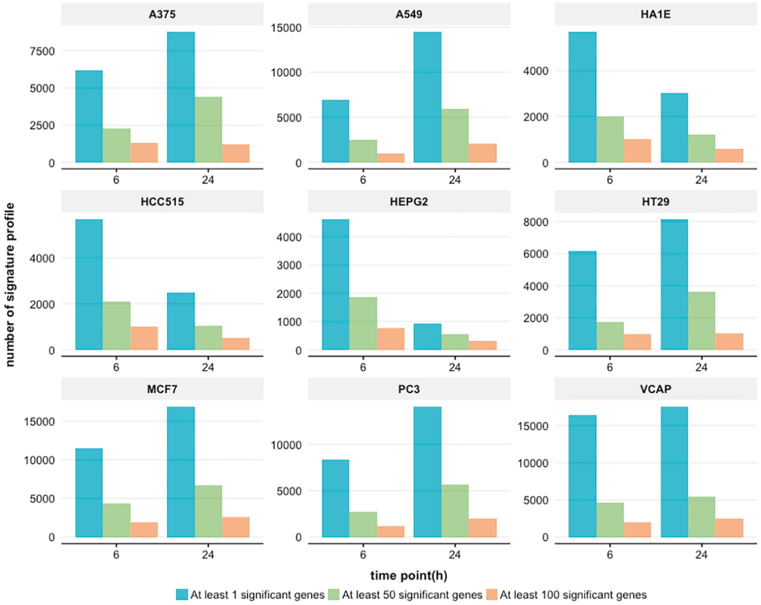


Fig 7. Drug perturbation specific differentially gene expression. The differential expression of genes for 9 cell lines is shown categorized in the time durations (6 and 24h) of drug perturbations.

https://doi.org/10.1371/journal.pone.0201937.g007

Changes in biological activity. Finally, we compare the findings shown in Figs 6 and 7 to reveal changes in the biological activity of the corresponding cell lines. In order to do this, we estimate the fraction of change for each of the two categories ‘at least 50 significant genes’ and ‘at least 100 significant genes’ with respect to the category ‘at least 1 significant gene’. That means we are estimating

$$f_A^{50} = \frac{\#\text{profiles}(\text{at least 50 significant genes}|A)}{\#\text{profiles}(\text{at least 1 significant gene}|A)} \tag{1}$$

$$f_A^{100} = \frac{\#\text{profiles}(\text{at least 100 significant genes}|A)}{\#\text{profiles}(\text{at least 1 significant gene}|A)} \tag{2}$$

wheres A corresponds either to Low dosage or 6 hours and

$$f_B^{50} = \frac{\#\text{profiles}(\text{at least 50 significant genes}|B)}{\#\text{profiles}(\text{at least 1 significant gene}|A)} \tag{3}$$

$$f_B^{100} = \frac{\#\text{profiles}(\text{at least 100 significant genes}|B)}{\#\text{profiles}(\text{at least 1 significant gene}|A)} \tag{4}$$

whereas B corresponds either to high dosage and 24 hours. This results in 8 percentage values for each cell line, 4 values from Fig 6 ($f_{\text{Low}}^{50}, f_{\text{Low}}^{100}, f_{\text{High}}^{50}, f_{\text{High}}^{100}$) and 4 values from Fig 7 ($f_{6\text{ hours}}^{50}, f_{6\text{ hours}}^{100}, f_{24\text{ hours}}^{50}, f_{24\text{ hours}}^{100}$). From these we obtain four straight lines per cell line defined by the pairs ($f_{\text{Low}}^{50}, f_{\text{High}}^{50}$) (green line in Fig 8) and ($f_{\text{Low}}^{100}, f_{\text{High}}^{100}$) (blue line in Fig 8) for dosages and

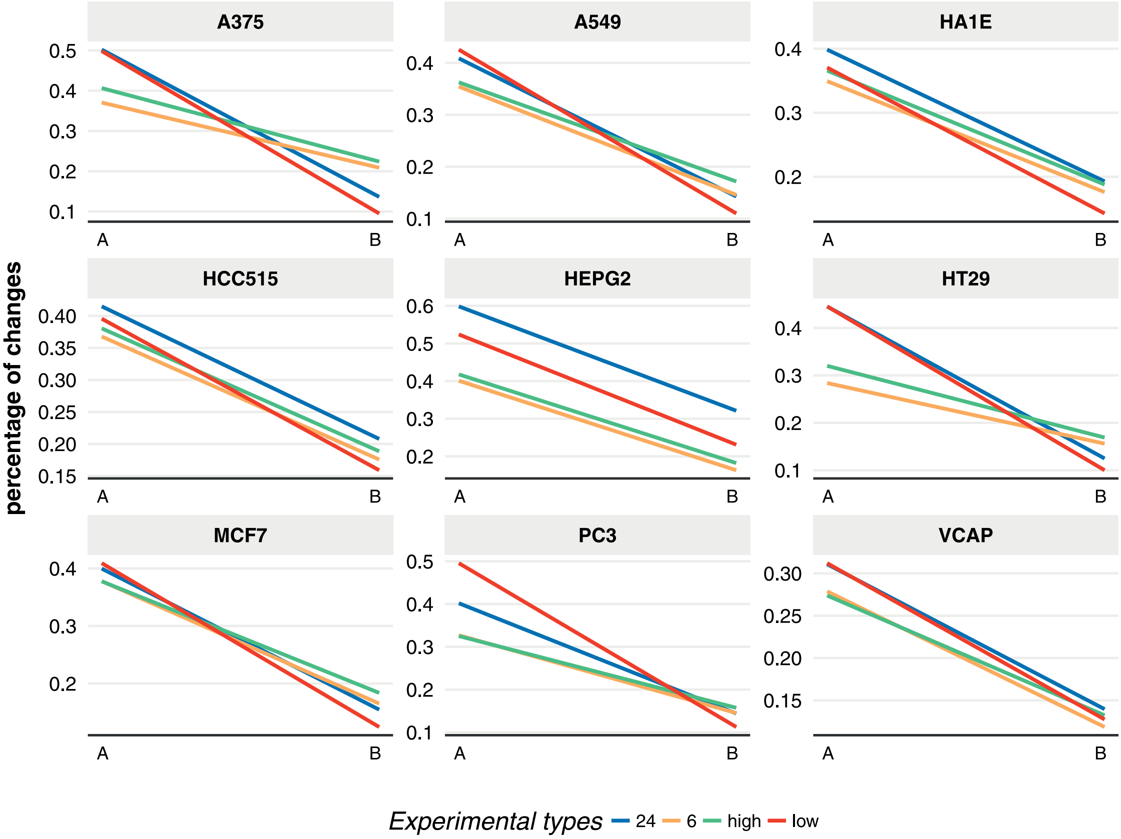


Fig 8. Changes of biological activity. Percentage changes in the number of significant profiles for the cell lines in dependence on the dosages and time points obtained from Figs 6 and 7. A. corresponds either to Low dosage or 6 hours and B. corresponds either to High dosage and 24 hours.

https://doi.org/10.1371/journal.pone.0201937.g008

$(f^{50}_{6\ \text{hours}}, f^{50}_{24\ \text{hours}})$ (red line in Fig 8) and $(f^{100}_{6\ \text{hours}}, f^{100}_{24\ \text{hours}})$ (orange line in Fig 8) for time points. Overall this means Fig 8 shows a summary of the fraction (percentage) of changes in the biological activity in dependence on different experimental conditions.

From Fig 8 we obtain two major observations. First, regarding the slope of the four straight lines, we observe that either these are parallel or they intersect each other. A parallel behavior is observed for cell line HEPG2 or VCAP, whereas an intersection is observed for HT29 or A375. This means that changes in the drug dosages has a nonlinear effect for cell line HT29 or A375 compared to, e.g., cell line HEPG2 or VCAP, if contrasted with changes in the time points. The second major observation from Fig 8 is the change of the top y-scale. For instance for cell line HEPG2 we find the highest percentage change of 60% for 24 hours, whereas for cell line VCAP this is only slightly over 30%. The difference is almost a factor of two in the activity changes.

Discussion

In this study, we analyzed the LINCS L1000 dataset by characterizing different experimental variables including cell types, time points, and dosages. We performed our analysis for two different layers. In layer one we focused on distributional characteristics of signature profiles whereas in layer two we focused on biological activity changes as measured by the number of differentially expressed genes.

Despite the fact that the LINCS L1000 dataset contains information for 71 cell lines, the vast majority of data is available for 9 cell lines only, namely A375, A549, HCC515, HA1E, HEPG2, HT29, MCF7, PC3 and VCAP, as can be seen from Fig 2 and Table 2. Each of these cell lines contains more than 20,000 signature profiles which enables excellent analysis opportunities. In contrast, for 46 cell lines less than 500 signature profiles are available. This means the utility of these 46 cell lines for any pharmacogenomic application is severely limited. Overall this means, that only 12% of all cell lines enable comprehensive large-scale data-driven pharmacogenomic applications.

For the number of replicates, we found that 2, 3 and 4 replicates are the majority for the 9 highly profiled cell lines, see Fig 3. However, also the number of replicates vary greatly between the cell lines. For instance, for HT29 there are over 8000 profiles with four replicates available whereas for A549 there are less than 2000 profiles, which means the difference is a factor of four. For studies requiring a very large number of replicates the cell lines MCF7 and PC3 are preferable because these cell lines provide experimental condition with over 9 replicates. To a lesser extend this is also true for A375. This information is important for planning an analysis in order to prevent an underpowered analysis [33] and ensure accurate estimations in a downstream analysis [34].

From the distribution of dosages (concentrations of drugs or small molecules) we found that most of these are used only with one or two concentrations, see Fig 4. However, for cell line MCF7 small molecules have been applied for even more than 9 different concentrations. Overall, the screening character of the LINCS project is well reflected by the distributions for different concentrations across the 9 cell lines in Fig 4 because of the high variability in the resulting number of signature profiles.

The second part of our analysis focused on the differentially expression of genes. As an overall results we find 24 profiles without any significant gene, 158,030 profiles with at least 1 significant gene, 58,739 with at least 50 significant gene and 23,867 profiles with at least 100 significant genes, see Table 3. For these numbers we averaged over all cell lines and experimental conditions. From this analysis we can conclude that 99.99% of all signature profiles contain at least some activity changes induced by the applied perturbations. Interestingly, the induced activity changes in the expression of genes seem to be moderate because 62% of all signature profiles contain between 1 and 49 significant genes.

It has been pointed out by Iorio et al. [35] that a compound can show inconsistent transcriptional effects when applied across different cell lines, its biological effect may be differentiated when merging gene expression values from different cell lines. Therefore, the compounds that were used to assess the effect on the cell lines may hold a bias towards a particular biological effect, since a cell line might react differently to certain treatment [36–38].

By zooming into the individual cell lines, see Fig 5, these overall observations are confirmed, although, there are certainly noticeable variations in the level of activity changes. For instance, for cell line A375 we find a decrease of around 40% from the number of significant profiles in category one to category two, whereas for cell line VCAP this decrease is only about 30%. This is actually a desirable observation because it means the LINCS data reflect that natural variability and sensitivity of the different human cell lines.

Next, we performed a detailed analysis studying the influence of the dosage and the time points on the individual cell lines. For the dosages we observed two different behaviors, see Fig 6. Behavior one corresponds to an increase in the number of significant profiles when going from low to high dosages, across the three gene categories, e.g., for cell line A375 or HA1E. In contrast, behavior two corresponds to a decrease. Interestingly, this behavior is only observed for cell line VCAP.

For the time points we obtain similar results, see Fig 7. For the first behavior the number of differentially expressed genes increases with time, e.g., cell line A375 or A549. For the second behavior the number of differentially expressed genes decreases with time, e.g., cell line HA1E or HCC515.

An explanation for this is that either lower or higher concentration treatments do not kill cells rapidly. Due to this reason, they should be tested for a longer period of time/days. In experimental setup of the L1000 data it is possible that a higher concentration might not killed the entire population rather induced a resistance population in which cell cycle is not be arrested. Furthermore, it should be also noted that PC3 (high metastasis) and VCAP (moderate) are not in the same state.

Finally, we compared the influence of dosage changes (Fig 6) with the influence of time point changes (Fig 7) in order to reveal changes in the biological activity of the corresponding cell lines and summarized these findings in Fig 8. From this we obtained two major observations. First, either the slope of the four experimental types occurs in parallel or they intersect with each other. Second, the y-scale is not the same for all cell lines. These results demonstrate the nonlinearity of the biological activity of the cell lines as a function of the different experimental conditions (types) and, hence, show the biological complexity of the transcription regulation.

All these results allow to gain insights that go beyond the mere features of gene expression data, e.g., providing information about the number of samples or number of drugs used for perturbing the cell lines. Instead, the second part of our study provides information for selecting cell lines with respect to their activity profiles. This information is important for the design of any pharmacogenomic study regardless of their particular goals because it is the biological activity of genes that decides about the effect of drugs.

Interestingly, in a previous study it has been shown that using additional cell lines provides more information about the compound-induced biological effects when different time points are used in the experimental design [39]. We found two time points (6h and 24h) yielded the most number of significant genes (see Fig 7) in the L1000 data. Therefore, the time point coverage can provide an understanding of how the L1000 data is represented at the gene level. Moreover, the combination of MCF7, VCAP, A549, HT29, and PC3 cell lines covers the majority of the transcriptional effects.

Overall, the LINCS L1000 data provides a rich and valuable source of compound-induced data that addresses some of the problems as mentioned for the CMap data [14, 16]. For example, a limited number of replicates, batch effect sizes and small number of profiles, now are all increased and improved in the L1000 assay. However, there are still shortcomings: First, most of the compounds are profiled at a high single dose only, causing different variability in dosage measurements. Second, the dataset does not explicitly follow the conventional settings of using experimental variables which are needed in a genome-wide transcriptional profiling study [40], but measure only 978 gene transcripts while the rest of the transcriptome was estimated by a model. Finally, the compounds are neither from primary-screening libraries such as FDA-approved nor the molecularly targeted and not highly selective agents that would be of particular interest for researchers [41].

Conclusion

In this paper, we used the Big Data from the LINCS project to explore different experimental settings, such as cell line coverage, time points and dosages using a data pipeline to assess compound-induced transcriptional effects. As a result, first, we provided summary statistics for distributional characteristics of gene expression signature profiles from all cell lines and their perturbagents. Second, we revealed changes in the differential expression of genes manifesting the biological complexity of the perturbagents. As a result, our analysis hopefully helps in harnessing the overwhelming complexity of the LINCS data providing guidance for the experimental design of follow-up studies, e.g., by selecting appropriate cell lines.

Given the limitations of previous datasets such as the CMap [14, 16], our analysis suggests that the L1000 data provide a good opportunity for the characterization of the compound-induced transcriptional effects. Given the volume and complexity of this dataset for drug discovery, it is necessary to understand the potential of the L1000 dataset and how it can be used in a drug research setting where every step is driven by data and rigorous data models. For example, the selection of appropriate tools to access, analyze and create models using the dataset to validate hypotheses. More efficient ways are expected to quickly transform Big Data discoveries into clinical applications.

Supporting information

S1 File. This file is provided as a zip file containing figures and R code related to the manuscript.
(ZIP)

Acknowledgments

Matthias Dehmer thanks the Austrian Science Funds for supporting this work (project P26142).

Author Contributions

Formal analysis: Aliyu Musa.

Methodology: Aliyu Musa.

Supervision: Frank Emmert-Streib.

Visualization: Aliyu Musa, Frank Emmert-Streib.

Writing – original draft: Aliyu Musa, Shailesh Tripathi, Meenakshisundaram Kandhavelu, Matthias Dehmer, Frank Emmert-Streib.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009; 106(23):9362±9367. https://doi.org/10.1073/pnas.0903103106
2. Emmert-Streib F, de Matos Simoes R, Mullan P, Haibe-Kains B, Dehmer M. The gene regulatory network for breast cancer: Integrated regulatory landscape of cancer hallmarks. Front Genet. 2014; 5:15. https://doi.org/10.3389/fgene.2014.00015 PMID: 24550935
3. Lewis SN, Nsoesie E, Weeks C, Qiao D, Zhang L. Prediction of Disease and Phenotype Associations from Genome-Wide Association Studies. PLoS ONE. 2011; 6(11):e27175. https://doi.org/10.1371/journal.pone.0027175 PMID: 22076134

4. de Matos Simoes R, Dehmer M, Emmert-Streib F. Interfacing cellular networks of *S. cerevisiae* and *E. coli*: Connecting dynamic and genetic information. BMC Genomics. 2013; 14:324. https://doi.org/10.1186/1471-2164-14-324 PMID: 23663484

5. Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Molecular Systems Biology. 2007; 3(124):124. https://doi.org/10.1038/msb4100163 PMID: 17625512

6. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature. 2009; 462(7270):175±181. https://doi.org/10.1038/nature08506 PMID: 19881490

7. Dunkel M, GuÈnther S, Ahmed J, Wittig B, Preissner R. SuperPred: drug classification and target prediction. Nucleic acids research. 2008; 36(suppl 2):W55±W59. https://doi.org/10.1093/nar/gkn307 PMID: 18499712

8. Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, et al. The Library of Integrated Network-Based Cellular Signatures {NIH} Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Systems. 2017; p. ±. https://doi.org/10.1016/j.cels.2017.11.001 PMID: 29199020

9. Fan J, Han F, Liu H. Challenges of big data analysis. National science review. 2014; 1(2):293±314. https://doi.org/10.1093/nsr/nwt032 PMID: 25419469

10. Marx V. Biology: The big challenges of big data. Nature. 2013; 498(7453):255±260. https://doi.org/10.1038/498255a PMID: 23765498

11. Hafner M, Heiser LM, Williams EH, Niepel M, Wang NJ, Korkola JE, et al. Quantification of sensitivity and resistance of breast cancer cell lines to anti-cancer drugs using GR metrics. Scientific Data. 2017; 4:170166. https://doi.org/10.1038/sdata.2017.166 PMID: 29112189

12. Niepel M, Hafner M, Duan Q, Wang Z, Paull EO, Chung M, et al. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nature Communications. 2017; 8(1):1186. https://doi.org/10.1038/s41467-017-01383-w PMID: 29084964

13. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics. 2017; p. bbx044. https://doi.org/10.1093/bib/bbx044

14. Musa A, Ghoraie LS, Zhang SD, Glazko G, Yli-Harja O, Dehmer M, et al. A review of connectivity map and computational approaches in pharmacogenomics. Briefings in Bioinformatics. 2017; p. bbw112±bbw112. https://doi.org/10.1093/bib/bbw112

15. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nature Reviews Cancer. 2010; 10(1):59±64. https://doi.org/10.1038/nrc2771 PMID: 20029424

16. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science. 2006; 313 (5795):1929±1935. https://doi.org/10.1126/science.1132939 PMID: 17008526

17. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? Drug discovery today. 2013; 18(7):350±357. https://doi.org/10.1016/j.drudis.2012.07.014 PMID: 22897878

18. Finley SD, Chu LH, Popel AS. Computational systems biology approaches to anti-angiogenic cancer therapeutics. Drug discovery today. 2015; 20(2):187±197. https://doi.org/10.1016/j.drudis.2014.09.026 PMID: 25286370

19. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002; 30:207±210. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295

20. Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research. 2017; 45(D1): D271±D281. https://doi.org/10.1093/nar/gkw1000 PMID: 27794042

21. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research. 2009; 37(suppl 1):D619±D622. https://doi.org/10.1093/nar/gkn863 PMID: 18981052

22. Hinkelmann K, Kempthorne O. Design and Analysis of Experiments: Introduction to experimental design. Chichester: Wiley-Interscience; 2008.

23. Emmert-Streib F. Influence of the experimental design of gene expression studies on the inference of gene regulatory networks: Environmental factors. PeerJ. 2013; 1:e10. https://doi.org/10.7717/peerj.10 PMID: 23638344

24. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Research. 2014; 42(W1):W449±W460. https://doi.org/10.1093/nar/gku476 PMID: 24906883

25. Vidović D, Koleti A, SchuÈrer 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. Frontiers in Genetics. 2014; 5:342. https://doi.org/10.3389/fgene.2014.00342 PMID: 25324859

26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets?update. Nucleic Acids Research. 2013; 41(D1):D991±D995. https://doi.org/10.1093/nar/gks1193 PMID: 23193258

27. Ong E, Xie J, Ni Z, Liu Q, Sarntivijai S, Lin Y, et al. Ontological representation, integration, and analysis of LINCS cell line cells and their cellular responses. BMC Bioinformatics. 2017; 18(17):556. https://doi.org/10.1186/s12859-017-1981-5 PMID: 29322930

28. Lim RG, Quan C, Reyes-Ortiz AM, Lutz SE, Kedaigle AJ, Gipson TA, et al. Huntington's Disease iPSC-Derived Brain Microvascular Endothelial Cells Reveal WNT-Mediated Angiogenic and Blood-Brain Barrier Deficits. Cell Reports. 2017; 19(7):1365±1377. https://doi.org/10.1016/j.celrep.2017.04.021 PMID: 28514657

29. Liu C, Su J, Yang F, Wei K, Ma J, Zhou X. Compound signature detection on LINCS L1000 big data. Molecular BioSystems. 2015; 11(3):714±722. https://doi.org/10.1039/c4mb00677a PMID: 25609570

30. Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome biology. 2002; 3(5):1. https://doi.org/10.1186/gb-2002-3-5-research0022

31. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010; 11(10):1. https://doi.org/10.1186/gb-2010-11-10-r106

32. Lincscloud. LINCS L1000 R tools; 2014. http://support.lincscloud.org/hc/en-us/articles/202062163-L1000-Code-via-GitHub-.

33. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016; 22(6):839±851. https://doi.org/10.1261/rna.053959.115 PMID: 27022035

34. Willems E, Leyns L, Vandesompele J. Standardization of real-time PCR gene expression data from independent biological replicates. Analytical biochemistry. 2008; 379(1):127±129. https://doi.org/10.1016/j.ab.2008.04.036 PMID: 18485881

35. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences. 2010; 107(33):14621±14626. https://doi.org/10.1073/pnas.1000138107

36. Fingar DC, Blenis J. Target of rapamycin (TOR): an integrator of nutrient and growth factor signals and coordinator of cell growth and cell cycle progression. Oncogene. 2004; 23(18):3151±3171. https://doi.org/10.1038/sj.onc.1207542 PMID: 15094765

37. Nigsch F, Hutz J, Cornett B, Selinger DW, McAllister G, Bandyopadhyay S, et al. Determination of minimal transcriptional signatures of compounds for target prediction. EURASIP Journal on Bioinformatics and Systems Biology. 2012; 2012(1):1±10. https://doi.org/10.1186/1687-4153-2012-2

38. Cayrefourcq L, Mazard T, Joosse S, Solassol J, Ramos J, Assenat E, et al. Establishment and Characterization of a Cell Line from Human Circulating Colon Cancer Cells. Cancer Research. 2015; 75(5):892±901. https://doi.org/10.1158/0008-5472.CAN-14-2613 PMID: 25592149

39. Kibble M, Saarinen N, Tang J, Wennerberg K, MaÈkelaÈS, Aittokallio T. Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products. Natural product reports. 2015; 32(8):1249±1266. https://doi.org/10.1039/c5np00005j PMID: 26030402

40. Butcher RA, Schreiber SL. Using genome-wide transcriptional profiling to elucidate small-molecule mechanism. Current Opinion in Chemical Biology. 2005; 9(1):25±30. https://doi.org/10.1016/j.cbpa.2004.10.009 PMID: 15701449

41. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: {L1000} Platform and the First 1,000,000 Profiles. Cell. 2017; 171(6):1437±1452.e17. https://doi.org/10.1016/j.cell.2017.10.049 PMID: 29195078

PUBLICATION

V

Systems Pharmacogenomic Landscape of Drug Similarities from LINCS data: Drug Association Networks

A. Musa, S. Tripathi, M. Dehmer, O. Yli-Harja, S. A. Kauffman and F. Emmert-Streib

Scientific Reports 9.1 (2019), 7849
DOI: 10.1038/s41598-019-44291-3

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 12 October 2018
Accepted: 8 May 2019
Published online: 24 May 2019

Systems Pharmacogenomic Landscape of Drug Similarities from LINCS data: Drug Association Networks

Aliyu Musa[1,2], Shailesh Tripathi[1,6], Matthias Dehmer[3,4,6], Olli Yli-Harja[2,5,7], Stuart A. Kauffman[7] & Frank Emmert-Streib[1,2]

Modern research in the biomedical sciences is data-driven utilizing high-throughput technologies to generate big genomic data. The Library of Integrated Network-based Cellular Signatures (LINCS) is an example for a large-scale genomic data repository providing hundred thousands of high-dimensional gene expression measurements for thousands of drugs and dozens of cell lines. However, the remaining challenge is how to use these data effectively for pharmacogenomics. In this paper, we use LINCS data to construct drug association networks (DANs) representing the relationships between drugs. By using the Anatomical Therapeutic Chemical (ATC) classification of drugs we demonstrate that the DANs represent a systems pharmacogenomic landscape of drugs summarizing the entire LINCS repository on a genomic scale meaningfully. Here we identify the modules of the DANs as therapeutic attractors of the ATC drug classes.

Recent availability of large-scale pharmacogenomic data have presented new opportunities but also challanges for tailored patient treatment, drug design and drug safety[1,2]. Vast efforts have been placed into discovering the drug mode-of-action (MoA) and understanding the genetic interactions within cells for disease treatment[3]. Importantly, it has been found that drug-induced transcriptional profiles from cell lines can be used to characterize therapeutic effects, enabling new computational ways for pharmacogenomics for identifying small drug molecules, compounds and drug-drug similarities solely based on gene expression profiles[4–7].

The Library of Integrated Network-based Cellular Signatures (LINCS) program[8], (https://clue.io/), funded by the Big Data to Knowledge (BD2K) Initiative at the National Institutes of Health (NIH), generated genetic and molecular signatures of human cell lines in response to various perturbations. The LINCS data repository is a vast library of gene expression profiles covering seventy-two human cell lines and include experiments for thousands of chemical perturbagens (small drug molecules), and drugs added to the cell cultures to induce changes in the gene expression profiles. The LINCS data are publicly available from the Gene Expression Omnibus (GEO) database. Based on these data, several advanced computational methods have been proposed for drug repurposing, identification of mode-of-action (MoA) and discovering phenotypic relations[9–11]; for an overview see[12]. The reason why gene expression data can be utilized as surrogates for the structure of chemical compounds to study mechanism of action and phenotypic impact between compounds[13–17] it that in[18] it has been shown that structurally similar compounds have similar gene expression profiles, furthermore compounds with similar gene expression signatures tend to interact with similar protein targets[19].

[1]Predictive Society and Data Analytics Lab, Tampere University, Tampere, Korkeakoulunkatu 10, 33720, Tampere, Finland. [2]Institute of Biosciences and Medical Technology, Tampere University, Tampere, Korkeakoulunkatu 10, 33720, Tampere, Finland. [3]Department for Biomedical Computer Science and Mechatronics, UMIT - The Health and Lifesciences University, Eduard Wallnoefer Zentrum 1, 6060, Hall in Tyrol, Austria. [4]College of Computer and Control Engineering, Nankai University, Tianjin, 300350, P.R. China. [5]Computational Systems Biology Lab, Tampere University of Technology, Korkeakoulunkatu 10, 33720, Tampere, Finland. [6]Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Wehrgrabengasse 1-3, 4400, Steyr, Austria. [7]Institute for Systems Biology, Seattle, WA, 98109, USA. Correspondence and requests for materials should be addressed to F.E.-S. (email: frank.emmert-streib@tuni.fi)

Traditionally, pharmacology approaches focus on single drugs at a time to study their action, effects or safety[20]. This is similar to traditional molecular biology approaches that focused on single genes or proteins[21]. However, due to modern genomic high-throuhgput technologies, nowadays, it is possible to study many genes or proteins simultenously[22]. Pharmacogenomics and Systems Pharmacogenomics aim to utilize such genomic profiles to expand beyond single drugs[23]. For instance, in[24] drug-target and drug-drug networks have been constructed based on the DrugBank database utilizing information about FDA approved and non-approved drugs and their corresponding targets. However, their analysis focused exclusively on drugs and compounds with known targets and did not take into consideration dynamic activity profiles as represented, e.g., by transcriptomics data. In[25] some disadvantages were avoided by using gene expression profiles for which Pearson correlation-based networks were constructed. A problem is that the used data were generated from many independent, uncoordinated laboratories using varying platforms and samle preprations. Another drawback of this study is the small number of used profiles (<7,000) and the very limited number of studied drugs (~200). Similar data were used in[4,17] but the construction of the drug network differed. Also, their analysis focused on drugs with known MoA. A different approach has been taken in[26] where a drug-drug network has been constructed only based on known side effects of FDA approved drugs. A drawback is the sole focus on negative clinical parameters, limitation to FDA approved drugs and the neglection of dynamical aspects of drug effects. In[27] in addition to gene expressin data also information about chemical structures and drug responses have been used. Unfortuantely, the number of drugs for which all three sources of data are available is very limited. A common shortcoming of all these studies is a lack of conceptual explanations of the drug networks.

The ultimate goal in pharmacology is to know all properties, effects and actions of all drugs and componds[28]. Hypothetically, this information could be obtained from clinical trials testing each compound for every existing disease including subtypes and stages. From this information one could measure the similarity between different compounds, e.g., based on clinically relevant parameters. This would give the network structure of an ideal compound-space giving all relationships among all compounds corresponding to an ideal drug association network (iDAN). Due to the practical impossibility of such an approach the question is, is it possible by using genomics data to approximate such an iDAN?

The main purpose of our paper is to introduce a computational method that provides such an approximation leading to a systematic organization for the thousands of drugs and small compounds that are available from the LINCS repository. Specifically, we introduce a method for constructing Drug Association Networks (DAN) based on almost two million gene expression profiles for over 20,000 chemical perturbagens and seventy-two human cell lines. In these networks nodes correspond to drugs and two drugs are connected if their profile responses are similar, as measured by the statistical significance of the Jaccard Index (JI). The profile responses for each drug correspond to estimates of "consensus" signature profiles summarizing the transcriptional effect of drugs across multiple treatments on different cell lines and/or different dosages and time points. Overall, the DANs provide a systematic summary of the entire LINCS data repository and the complex pharmacogenomic landscape of drug similarities. For a conceptual overview see Fig. 1A.

For obtaining pharmacogenomically meaningful networks, we construct different DANs based on data from different conditions. Specifically, we construct for each cell line a DAN using only the corresponding drug signature profiles. Furthermore, we construct one DAN limited to FDA approved drugs and one DAN for all drugs and small compounds (comprising FDA approved and non-approved drugs). This leads to condition-specific DANs (see Fig. 1C for their dependencies). In total, we are inferring 74 different DANs.

In order to analyze and interpret the DANs, we investigate the DANs on three different levels. First, we study the structure of the DANs by identifying network modules, also called communities[29–31]. This will allow us to gain insights into the structural properties of the networks. Second, we study drugs pairwise by identifying the presence of significant Anatomical Therapeutic Chemical (ATC) classes in the entire network. This analysis step will show that drugs with similar ATC classes are actually identified in compound space. Third, we study the enrichment of the network modules with respect to ATC classes. By using the ATC classification of drugs, we will demonstrate that the DANs represent a pharmacogenomic landscape of drugs summarizing the entire LINCS repository on a genomic scale.

As a general results, we will show that the ATC code enriched modules in the DANs can be seen as therapeutic attractors of drug classes. We will see that this allows a conceptual extension of the idea of *cancer attractors*[32] introduced for gene regulatory networks to represent cell states[33,34] to DANs representing pharmacological states (need name).

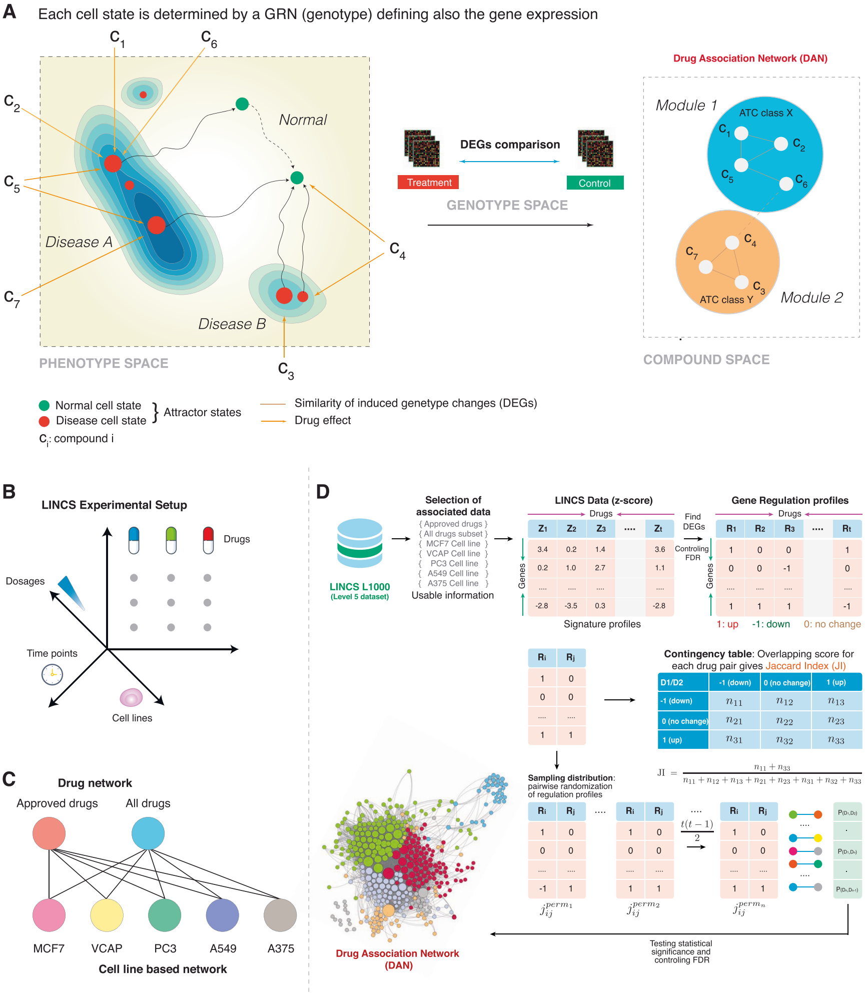
Furthermore, in order to communicate the wealth of our obtained results efficiently, we developed a web interface accessible at (http://dan-network.herokuapp.com). Our web application allows to access the drug-drug interactions inferred by our method, and connecting to external links. The features of our DAN user interface enable searching, browsing, exploration and downloading of the network visualizations.

The paper is organized as follows. In the next section we present the Materials and Methods used for our analysis. Then we present our Results and a Discussion. This paper finishes with Conclusions.

Results

In the following, we first construct DANs from different information corresponding to different characteristics of the LINCS data. This results in DANs having a context specific meaning. Then we will analyze the DANs on three different levels. First, we focus on the structure of the DANs identifying modules in the networks. Second, we study drugs pairwise by identifying the presence of significant ATC classes in the entire network. Third, we study the enrichment of the network modules with respect to ATC classes.

Construction of drug association networks. The first network, we construct for FDA-approved drugs with assigned annotations in DrugBank[35,36]. For this reason we call this network N_{approved}. In total, there are



Z1	Z2	Z3	....	Zt
3.4	0.2	1.4		3.6
0.2	1.0	2.7		1.1
....	....	....		....
-2.8	-3.5	0.3		-2.8

R1	R2	R3	....	Rt
1	0	0		1
0	0	-1		0
....	....	....		....
1	1	1		-1

Ri	Rj
1	0
0	0
....	....
1	1

D1/D2	-1 (down)	0 (no change)	1 (up)
-1 (down)	n_{11}	n_{12}	n_{13}
0 (no change)	n_{21}	n_{22}	n_{23}
1 (up)	n_{31}	n_{32}	n_{33}

$$\mathrm{JI} = \frac{n_{11}+n_{33}}{n_{11}+n_{12}+n_{13}+n_{21}+n_{23}+n_{31}+n_{32}+n_{33}}$$

Ri	Rj
1	0
0	0
....	....
-1	1

$j_{ij}^{perm_1}$

Ri	Rj
1	0
0	0
....	....
1	1

$j_{ij}^{perm_2}$

$\frac{t(t-1)}{2}$

Ri	Rj
1	0
0	0
....	....
1	1

$j_{ij}^{perm_n}$

Figure 1. (**A**) Conceptual connection between genotype space, phenotype space and compound space containing DANs. (**B**) Multifactural experimental space of the LINCS data. (**C**) For our analysis we study 7 different DANs. (**D**) Overview of the connstruction of a DAN. The figure shows the gene expression profile signature of drugs and small molecule compounds from LINCS L1000 subset. Representation of the use of drug-feature matrices of different types to calculate drug connections using Jaccard Index (JI).

1139 approved drugs in LINCS, however, only 381 have an ATC annotation. The drugs with DrugBank IDs are repeated in multiple experiments; therefore, the *landmark* genes have multiple z-scores from different experiments. We first average the z-scores for each drug from different experiments and use the consensus of the z-scores to construct the DAN, as described in the method section. From this analysis, we obtain a network with 381 nodes and 4251 significant interactions. From this network, we extract the giant connected component (GCC) having 367 drugs (nodes) and 4244 interactions (edges). In Fig. 2A, we show the distribution of JI of all significant interactions for this network from profiles having between 100 to 150 DEGs.

The second network we construct, we call N_{all}, is for all available drugs. In LINCS data there are in total 2505 different drugs applied in the different experiments (cell line, dosage and time point). For these, we construct a network with 2505 drugs and 86,585 significant interactions. From this network, we extract the GCC having 2451 nodes and 22636 interactions. In Fig. 2B, we show the distribution of JI of all significant interactions for this network from profiles having between 700 to 800 DEGs. The higher the value of the JI the more genes are commonly up- or down-regulated between two drugs.

Next, we construct 72 networks that are specific for the 72 cell lines. All of these networks are sub-graphs of N_{all}, i.e., $N_{\text{all}}^{CL_l} \subset N_{\text{all}}$, with $CL = \{\textit{list of cell lines in LINCS}\}$, due to the way we summarize all configurations, see

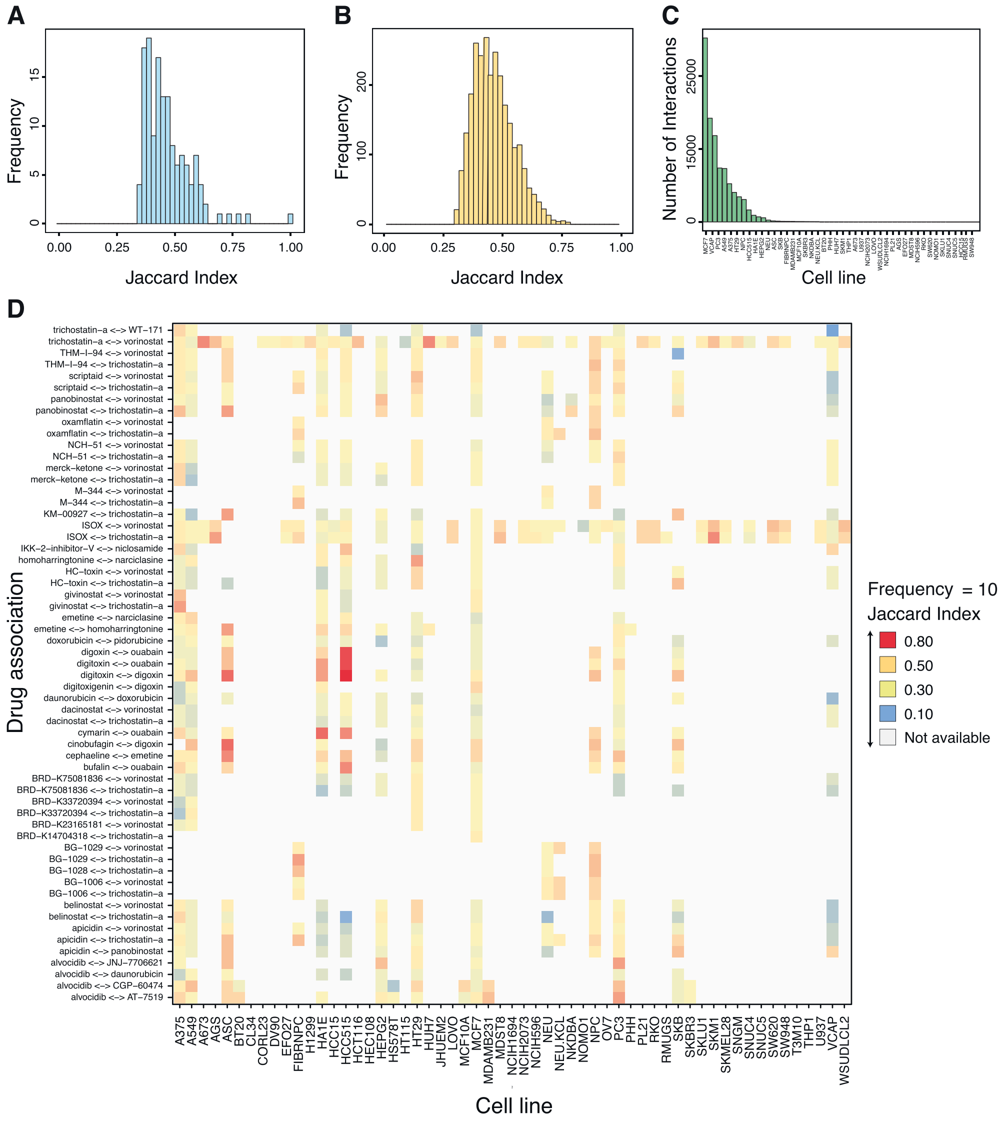


Figure 2. Similarity distribution of drugs over different experiments. (**A**) Distribution of JI of all significant interactions for N_{approved} from profiles having between 100–150 DEGs. (**B**) Distribution of JI of all significant interactions for N_{all} from signature profiles having DEGs between 700–800. (**C**) Number of significant interactions between drugs for different cell lines. (**D**) Heat map showing drug similarities using JI for selected drug-pairs (y-axis) in dependence on cell lines (x-axis) having a JI larger than 0.5 and appearing in ten or more experiments. The color indicates the value of the JI for drug-pairs. The grey color shows drug-pair not available in a given cell line.

Eqn. 5. In addition, it holds $N_{\text{all}} = \cup_{CL_l \in CL} N_{\text{all}}^{CL_l}$. That means, N_{all} contains all significant interactions identified for any cell line.

For our further analysis, we select from these 72 networks the five networks having the highest number of interactions between the drugs; see Fig. 2C for the frequency distribution of interactions for all cell lines. These cell lines are {*MCF7*, *VCAP*, *PC*3, *A*549, *A*375}. These 5 networks contain the most information, assuming

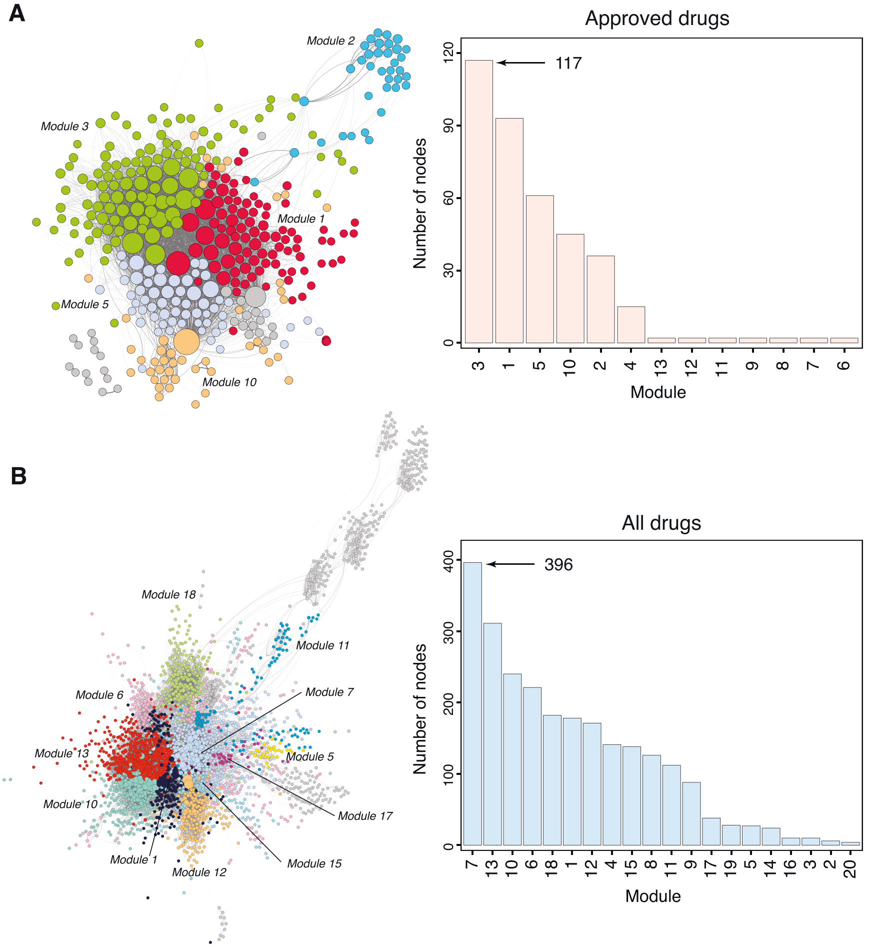


Figure 3. Drug network connecting the most associative drugs using JI and module annotation from LINCS L1000 dataset. The network representation displays drugs as circles (nodes) connected with edges. The colour of drug corresponds to their associative grouped module. (**A**) Shows the network of FDA-Approved drugs with their corresponding module annotations (Left), and the number of nodes in each module of N_{approved} (Right) (**B**) The network show All Drugs including approved and non-approved drugs colored based on grouped module (Left), and the number of drugs in each cluster for N_{all} (Right).

interactions provide informative knowledge. The high number of interactions in each of these networks (more than 10,000) ensures also that a sensible identification of modules is feasible.

In Table 4, we show a summary of these seven networks and their number of nodes and edges. All of these networks correspond to the GCC of the corresponding network. In the following, we will limit our analysis to these seven networks.

Modules in Dans. Our first analysis consists in the identification of the modules in the seven different DAN networks. For this, we are using a multilevel community module detection algorithm[37] to find the modules in the networks. The modularity and the number of modules for each network are summarized in Table 4. We would like to remark that the number of the modules correspond to labels, i.e., the same label for different networks does not mean it should contain the same drugs. In general, we find the modularity to be similar among the different networks except for N_{approved} and N_{all} which is smaller. This is understandable considering the used data for these networks is different to the others. For the number of modules we observe similar values ranging from 11 to 25 modules.

In Fig. 3, we show the networks for N_{approved} and N_{all} and the distribution of the number of drugs in the modules. The networks for the 5 cell lines are shown in Fig. 1–5 in the Supplementary File. From the barcharts of boths

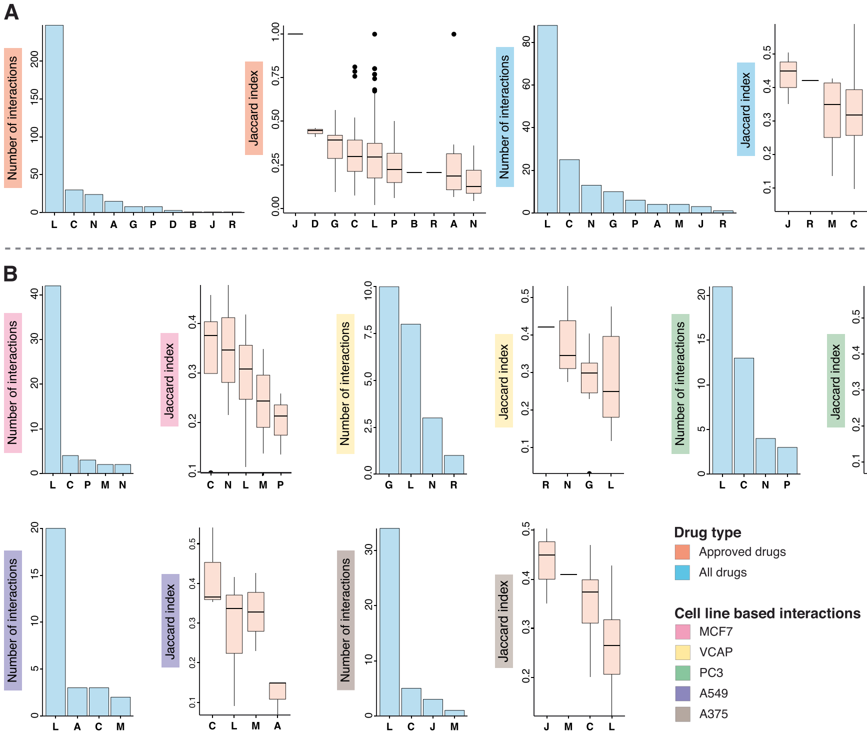


Figure 4. Significant interactions between drugs with the same ATC classes. Here the notation, e.g., *L* means *L* − *L* (x-axis) (similar for other ATC codes) and their corresponding Jaccard Indices. (**A**) Number of significant interactions between the same *ATC* codes (i.e two drugs with the same ATC class) for the networks N_{approved} (right) and N_{all} (left). The boxplots show the distribution of JI of all significant interactions of drugs which are annotated with the same ATC codes of N_{approved} (right) and N_{all} (left). (**B**) Results for the 5 networks N_{MCF7}, N_{VCAP}, N_{PC3}, N_{A549} and N_{A375}. Shown results are similar as for (**A**). The colored y-axis label indicate the type of network analysed.

networks one can see that there are a few modules containing a large number of drugs and the remaining modules contain only a few drugs. These large modules are also clearly visible in the network representation of the DANs on the left-hand-side in Fig. 3. In general, the modules in N_{all} are larger than in N_{approved} which is understandable because the former DAN contains 2451 nodes whereas the latter has only 367 (see Table 4).

Significance of ATC interactions in the entire network.

Next, we analyze pairwise interactions between drugs in terms of their corresponding ATC classes. For this analysis, we use all the significant interactions which are annotated with *ATC* codes in the 7 DANs. The number of interactions and the distribution of their JI values are shown in Fig. 4. In this figure, we show only drug pairs belonging to the same ATC class corresponding to homogene interactions, i.e., the label L refers to the interaction of two drugs, both from ATC class L.

For the network N_{approved} the number of interactions and their JI values are shown in Fig. 4A (left with red label). One can see that interactions between drugs from the ATC class L occur far more often than for any other ATC class. Interestingly, the differences in the values of the JI for these interactions (shown in the boxplot in Fig. 4A) are not that different for different ATC classes. The results are similar for N_{all}.

For the other five networks of the cell lines, the frequency of drug annotations and the distribution of JI values are shown in Fig. 4B. From comparing these five networks we make five observations. First, the number of ATC classes is much smaller than for the two networks N_{approved} and N_{all}. Second, the ATC class L is present in all networks for the cell lines. Third, the overlap between the five cell line networks with respect to the ATC classes is smaller than for the two generic networks. Fourth, the network N_{VCAP} is the only one having more interactions for the ATC class G. Also the difference between the top 4 ATC classes is smaller than for the other networks, except N_{PC3}. Fifth, all of the networks share that the ATC class of the larges JI values do not correspond to the ATC class for the largest number of interactions.

In order to reveal robust interaction patterns, we randomize the ATC class labels of the drugs and determine statistically significant ATC interactions classes. For this analysis, we study homogeneous as well as

D_i↓/D_j→	**−1 (down)**	**0 (no change)**	**1 (up)**
−1 (down)	n_{11}	n_{12}	n_{13}
(no change)	n_{21}	n_{22}	n_{23}
(up)	n_{31}	n_{32}	n_{33}

Table 1. Contingency table summarizing the gene regulation profiles R_i and R_j treated by drug D_k and D_l. Here n_{kl} are integer numbers giving the common genes in the categories $k, l \in \{up, nochange, down\}$.

Code	Description
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito urinary system and sex hormones
H	Systemic hormonal preparations, excl. sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides, and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table 2. Description of ATC annotations. The first level of the ATC classification represents the organ or system in the body on which the therapeutic effect.

	Signature profile	Small molecule
No significant gene	24	19
At least 1 significant gene	158,030	19,957
At least 50 significant genes	58,739	15,714
At least 100 significant genes	23,867	8,211
Total	**158,054**	**20,009**

Table 3. Summary of z-score signature profiles for DEGs between treatments and controls on the cell line subset.

heterogeneous interactions (between drugs from different ATC classes) corresponding to the inter-class effect of drugs. Specifically, we obtain the counts of ATC code combinations from each network (i.e. $A-A$, $A-C$, $B-L$ etc.) by counting their occurancy in each DAN. Then we randomise each network 10,000 times to obtain the null distribution for each ATC class combination using the counts of ATC classes as test statistic for each ATC class. From comparing the null distributions with the test statistics we obtaine p-values to which we apply a Bonferroni multiple testing correction to get the adjusted p-values.

These results demonstrate that the inferred network structure of all DANs capturing meaningful drug-specific information that could be revealed by the significance of selected ATC classes.

Enrichment analysis of network modules. Finally, in order to obtain a pharmacogenomically meaningful interpretation of the DANs, we perform an enrichment analysis of the modules identified in the previous section.

The constructed DANs have nodes corresponding to known and unknown drugs and some of the nodes (drugs) in these networks have Anatomical Therapeutic Chemical (ATC) annotations[38]. We categorized these drugs/nodes with ATC annotations into 14 classes, summarized in Table 2. In addition, we use the label 'X' to indicate drugs for which no drug annotation is known.

We performed an enrichment analysis of drugs with ATC codes for the modules detected in each network. In order to test the statistical significance of ATC classes, we use Fisher's Exact Test[39]. Since we are testing multiple hypothesis tests for each module, we apply a Benjamini Hochberg correction to control the FDR. In the enrichment analysis we first find the total number of drugs in a module which are labelled with ATC codes and then we performed Fisher's Exact test to determine which ATC labels are overrepresented in a particular module. The results of this enrichment analysis are shown in Fig. 5.

	A	B	C	D	G	H	J	L	M	N	P	R	S	V
Approved drugs														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000
Module 2	1.000	1.000	1.000	**0.020**	1.000	0.119	1.000	1.000	1.000	1.000	1.000	**0.005**	**0.024**	1.000
Module 3	1.000	1.000	1.000	1.000	1.000	1.000	**0.006**	0.127	1.000	1.000	1.000	1.000	1.000	0.127
Module 4	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.002**	1.000	1.000	1.000
Module 10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.477	1.000	1.000	1.000	1.000	1.000	1.000
All drugs														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	0.206	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	0.951	1.000	1.000	**0.011**	0.951	**0.015**	1.000	1.000	1.000	1.000	1.000	1.000	0.073	1.000
Module 7	1.000	1.000	1.000	1.000	1.000	1.000	0.178	1.000	0.565	0.950	1.000	0.134	1.000	1.000
Module 10	0.685	0.979	0.547	1.000	1.000	1.000	0.981	1.000	1.000	0.366	0.366	1.000	1.000	1.000
Module 11	0.974	0.743	1.000	0.974	0.743	1.000	1.000	1.000	0.359	0.743	0.974	1.000	0.974	0.743
Module 12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.098	1.000	1.000	1.000	1.000	1.000	1.000
Module 13	1.000	1.000	0.575	1.000	1.000	1.000	1.000	0.492	1.000	1.000	1.000	1.000	1.000	1.000
Module 15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.614	1.000	1.000	1.000	1.000	1.000	1.000
Module 17	1.000	1.000	1.000	0.138	1.000	0.752	1.000	0.614	1.000	1.000	1.000	0.657	0.619	1.000
Module 18	1.000	1.000	1.000	1.000	0.077	1.000	0.513	1.000	1.000	1.000	1.000	0.913	1.000	1.000
MCF7 Cell line														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	0.473	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.049**	1.000	1.000
Module 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.239	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.356	0.091	1.000	1.000	1.000	1.000	1.000
Module 7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.664	1.000	1.000	1.000
Module 8	1.000	1.000	0.207	0.492	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.664	1.000	1.000	1.000
VCAP Cell line														
Module 4	1.000	1.000	1.000	1.000	0.385	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	0.675	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	1.000	1.000	1.000	0.121	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PC3 Cell line														
Module 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.272	1.000	1.000	1.000	1.000	1.000	1.000
Module 2	1.000	1.000	**0.028**	1.000	**0.029**	1.000	1.000	0.816	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	1.000	1.000	0.930	1.000	1.000	1.000	0.087	**0.008**	0.290	1.000	1.000	1.000	1.000	1.000
Module 4	0.150	1.000	1.000	1.000	1.000	1.000	1.000	0.472	1.000	1.000	1.000	1.000	1.000	1.000
Module 5	1.000	1.000	0.499	0.192	0.217	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	0.929	1.000	0.535	1.000	0.695	1.000	1.000	0.173	0.179	0.059	0.521	1.000	1.000	1.000
Module 8	0.122	1.000	0.869	**0.013**	1.000	0.262	0.789	0.999	1.000	0.180	1.000	0.262	**0.004**	1.000
Module 11	0.703	1.000	0.308	0.652	1.000	1.000	0.135	0.991	1.000	1.000	**0.003**	1.000	0.594	1.000
Module 12	0.280	1.000	1.000	1.000	1.000	1.000	1.000	0.060	1.000	1.000	1.000	1.000	1.000	1.000
Module 13	1.000	1.000	1.000	1.000	**0.005**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.272	1.000	1.000	1.000	1.000	1.000	1.000
A549 Cell line														
Module 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.029**	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.062	**0.006**	1.000	1.000	1.000	1.000	1.000
Module 4	1.000	1.000	**0.002**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 8	1.000	1.000	1.000	1.000	0.255	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 9	1.000	1.000	1.000	1.000	0.255	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.757	1.000	1.000	1.000
Module 13	0.251	1.000	1.000	**0.000**	1.000	0.874	1.000	1.000	1.000	1.000	1.000	1.000	0.110	1.000
Module 15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.636	1.000	1.000	1.000	1.000
A375 Cell line														
Module 1	1.000	1.000	1.000	0.600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.029**	1.000	1.000	1.000	1.000	1.000	1.000
Module 6	1.000	1.000	**0.000**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Module 8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.029**	1.000	1.000	1.000	1.000	1.000	1.000
Module 11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.002**	1.000	1.000	1.000	1.000	1.000	1.000

Figure 5. Enrichment of individual modules in the DANs. Shown are the BH corrected q-values of Fisher's exact tests for the enrichment of ATC codes in each of the modules of the DANs. Modules not shown, do not contain any enriched ATC code. The highlighted cells are statistically significant. The horizontal and vertical boxes highlight the multiple occurance of ATC classes in modules and multiple enriched modules for an ATC class respectively.

In N_{approved}, the N (Nervous system) group is overrepresented in first module. The ATC groups R (Respiratory system), S (Sensory organs) and D (Dermatologicals) are enriched to the second module. The ATC group J (Antiinfectives for systemetic use), G (Genito-urinary system and sex hormones) and P (Antiparasitic products, insecticides and repellents) are enriched in 3, 4 and 5 modules. This is interesting to highlight, since the drugs which are overrepresented in the same modules of different classes perturb common genes or a similar subset of genes. This information can be used for further investigation to see if those drugs can perturb common pathways.

In the network (N_{all}), the ATC group L (Antineoplastic and immunomodulating agents) is overrepresented in first module. ATC groups H (Systemic hormonal preparations, excluding sex hormones and insulins) and D

DAN	Used information	Drugs	Edges	Modularity	No. of Modules
N_{approved}	Approved drugs	367	4244	0.318	13
N_{all}	All drugs	2451	22636	0.554	20
gray N_{MCF7}	MCF7 cell line	750	7144	0.623	11
N_{VCAP}	VCAP cell line	520	2727	0.749	25
N_{PC3}	PC3 cell line	612	4314	0.644	17
N_{A549}	A549 cell line	380	2122	0.561	22
N_{A375}	A375 cell line	635	4286	0.636	14

Table 4. Summary of seven DANs constructed from different information. Shown is the information of the giant connected component. Column two describes the used information that characterizes the underlying data for each network.

DAN/ATC code	C	D	G	H	J	L	M	N	P	R	S	SC	SM
Approved drugs		1	1		1			1	1	1	1	7	5
All drugs		1		1		1						3	2
gray MCF7 cell line						1				1		2	2
VCAP cell line												0	0
PC3 cell line	1	1	2			1			1		1	6	5
A549 cell line	1	1				1	1					4	4
A375 cell line	1					3						2	4
SM (all networks)	3	4	3	1	1	7	1	1	2	2	2		

Table 5. Summary of module enrichments shown in Table 5 for all DANs. The columns show ATC classes highlighting if ATC codes are enriched in at least one module in the entire network (see Table 5). SC gives the number of significant ATC classes and SM gives the number of significant modules per network. SM (all networks) gives the number of significant modules in all DANs.

(Dermatologicals) are enriched to the sixth module, however group S (Sensory organs) also show a low q-value (0.073, which is not significant).

For the network N_{MCF7}, it shows the ATC group L (Antineoplastic and immunomodulating agents) and R (Respiratory system) are enriched in the first and third modules. However, the ATC group M show a low q-value (0.090) in module 5.

For the network N_{VCAP}, no ATC group is enriched in any module however, ATC group *D* (Dermatologicals) show a low q-value (0.121) in module 6.

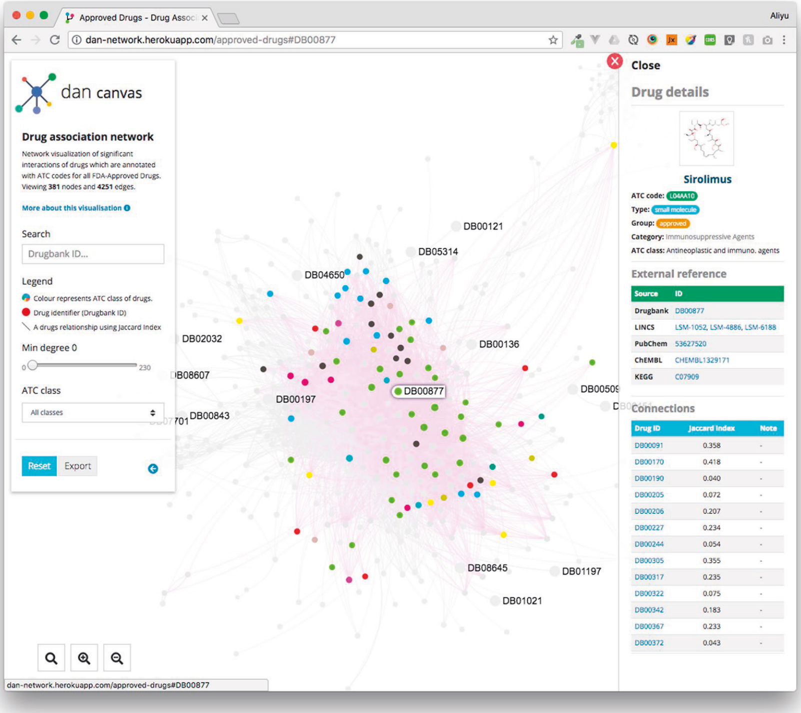
In the network N_{PC3}, the ATC groups G (Genito-urinary system and sex hormones) and C (Cardiovascular system) are enriched in module 2. The ATC group L (Antineoplastic and immunomodulating agents), in module 3, also ATC group J (Antiinfectives for systemic use) has a low q-value (0.087) in module 3. The ATC group N (Nervous system) shows a low q-score (0.059) in module 6. The ATC groups S (Sensory organs) and D (Dermatologicals) are enriched in module 8. The ATC group P (Antiparasitic products, insecticides and repellents) is also enriched in module 11. The ATC group L (Antineoplastic and immunomodulating agents) show a low q-score (0.06) in module 12. The ATC group G (Genito-urinary system and sex hormones) is enriched in module 13.

In the network N_{A549}, the ATC group L (Antineoplastic and immunomodulating agents) is enriched in module 2. The ATC group M is enriched in module 3, ATC group C is enriched in module 4. However, The ATC group L (0.062) and S (0.11) show low q-values in modules 3 and 13 respectively.

In The network N_{A375}, the ATC group L (Antineoplastic and immunomodulating agents) is enriched in modules 3, 8 and 11 respectively. The ATC group C (Cardiovascular system) is enriched in mdoule 6.

The summary of the enrichment analysis of the ATC groups for the modules of the different networks is shown in Table 5. In this table, we highlighted the ATC groups which are enriched in at least one module in different networks. We also include those ATC groups which are not significant but holds low q-values between $0.05 < \alpha < 0.15$.

Web interface for DAN of drugs. Due to complexity of our results making it difficult to communicate all details, we developed an interactive web application. The web application is publicly available at http://dan-network.herokuapp.com/ showing visualizations of all 7 DANs summarized in Table 4. For the technical realization for the visualization of the networks we developed our web interface using the NodeJs[40] and SigmaJS[41] libraries. Each node in the network (drug) has a dedicated pane with a list of the relevant associations and external resources to websites such as: DrugBank, PubChem, LINCS Portal, ChemBL and KEGG Ligand with relevant identifiers. That means, a user can interactively explore the interactions in all 7 DANs obtaining pharmacological information from the linked data resources. A screen shot of our web application is shown in Fig. 6.


Approved Drugs - Drug Associ ×
Aliyu
dan-network.herokuapp.com/approved-drugs#DB00877
DB00121
DB05314
DB04650
DB02032
DB08607
DB00197
DB00877
DB00136
DB00509
DB00843
DB08645
DB01197
DB01021
dan-network.herokuapp.com/approved-drugs#DB00877

dan canvas

Drug association network

Network visualization of significant interactions of drugs which are annotated with ATC codes for all FDA-Approved Drugs. Viewing **381** nodes and **4251** edges.

More about this visualisation

Search: Drugbank ID...

Legend

- Colour represents ATC class of drugs.
- ● Drug identifier (Drugbank ID)
- ╲ A drugs relationship using Jaccard Index

Min degree 0 (0 – 230)

ATC class: All classes

Reset | Export

Close

Drug details

Sirolimus

ATC code: L04AA10

Type: small molecule

Group: approved

Category: Immunosuppressive Agents

ATC class: Antineoplastic and immuno. agents

External reference

Source	ID
Drugbank	DB00877
LINCS	LSM-1052, LSM-4886, LSM-6188
PubChem	53627520
ChEMBL	CHEMBL1329171
KEGG	C07909

Connections

Drug ID	Jaccard Index	Note
DB00091	0.358	-
DB00170	0.418	-
DB00190	0.040	-
DB00205	0.072	-
DB00206	0.207	-
DB00227	0.234	-
DB00244	0.054	-
DB00305	0.355	-
DB00317	0.235	-
DB00322	0.075	-
DB00342	0.183	-
DB00367	0.233	-
DB00372	0.043	-

Figure 6. The website view of the DAN network. This website shows our results of the drug-drug interaction network for all 20,009 drugs and small-molecule compounds profiled in the LINCS L1000 signature gene expression profiles.

Discussion

In our paper, we based our analysis on the LINCS data repository providing compreshensive information about the effect of drugs or compounds on gene expression changes. This means LINCS enables an estimation of the linkage between genotype, phenotype and therapies and to identify key genes which are a significant part of the biological processes related to phenotype differences as approximated by gene expression values.

For our study, we went beyond single genes because we were aiming at a comprehensive overview of the systems relations among all drugs tested in LINCS. In order to accomplish this, we utilized differentially expression profiles to estimate DANs. Specifically, our analysis started by constructing DANs to estimate the similarity between drug pairs using the Jaccard Index, which estimates the proportion of differentially expressed genes that are common in the corresponding expression profiles. If two drugs showed a statistically significant similarity, we connected them by an edge. In this way, we constructed 7 different DANs for 7 different conditions, which we further analyzed. The results of these networks are summarized in Table 4.

We analyzed the DANs on three diffexrnt levels. First we studied the structure of the DANs by identifying network modules. Second, we studied the drugs pairwise by identifying the presence of significant ATC classes in the entire network. Third, we studied the enrichment of the network modules with respect to ATC classes.

The significant pairs in the networks show a variable JI distribution, shown in Fig. 2A,B. In general, the effect of drugs in terms of differentially expressed genes varies, i.e., some drugs show a strong effect, which means a large number of differentially expressed genes, while other drugs have a moderate effect changing the expression of only a small number of genes. If a drug, D_i has a moderate effect, i.e., a small number of differentially expressed genes, but a strong overlap with the drug, D_j, which has a strong effect on the genes, i.e., it causes a larger number of differentially genes, the JI will be significant but not high. In such cases the interaction may not describe the same functionality of both drugs, but it can have a similar effect on some subset of gene targets. On the other hand, if two drugs have a similar proportion of differentially expressed genes and overlap strongly then the corresponding JI is higher.

After the construction of the networks, we identified modules in the networks. For this we employed the multilevel community algorithm[37]. The results of this analysis are summarized in Table 4. In general, the modularity of the networks for the five cell lines is higher than for N_{all} and $N_{approved}$, which has the lowest modularity. For the number of identified modules this distinction is no longer present. It is interesting to note that the number of modules in all networks is of the same order of magnitude as the number of our ATC classes (which is 14).

It is interesting that the modularity of N_{all} and N_{approved} is different to the five cell line DANs because these two network types are indeed quite different from each other due to the different information used for their construction.

These results suggest that the modules in the networks could represent drugs or drug classes effecting similar targets. That means drugs in the same module have a similar effect on some common gene targets, because of their significant overlapping of differentially expressed genes as measured by the JI. This can also be interpreted as follows: The presence of drugs in different modules suggests that each module can identify a different type of target-set, which is independent from other target-sets for different drugs. For instance, for N_{approved}, we identify 13 modules which means that there are 13 distinct effect types of drugs. Interestingly, this number is very close the total number of ATC classes we were using, which is 14 (see Table 2).

In order to test this idea further, we performed an enrichment analysis of the network modules testing for the enrichment of ATC classes. The results are summarized in Fig. 5. Due to the complexity of these results, we discuss them in three steps. First, we discuss results for all networks combined. Second, we discuss network specific characteristics of significant modules and ATC classes. Third, we discuss networks and modules indivdually to identify commonalities.

First, from our results (see Table 5) we see that the total number of significant modules (SM (all networks)) for all networks enriched for the ATC classes is low varying between 7 (for ATC class L) and 0 (for ATC class A, B and V). Most ATC classes are only enriched in 1 or 2 modules in all networks, e.g., ATC class H, J, M, N, P, R and S.

Second, when looking at the networks individually, we found that the total number of enriched modules (SM) per network varies between 5 (for N_{approved}) and 0 (for N_{VCAP}). Similarly, the number of significant ATC classes (SC) per network varies between 7 (for N_{approved}) and 0 (for N_{VCAP}), see Table 5. Taken together, these observations confirm our interpretation of the findings for the number of modules, which did not consider ATC enrichments, underlining the representative character of the modules for ATC classes.

Third, we are looking at networks and modules indivdually. From these we can obtain the following summary for this level. Overall, we can identify four different types of drug-module enrichments discussed in the following.

Single-drug class in individual modules.

For this type of enrichment, we find only one enriched ATC class per module in a DAN. That means there is an unique relation between an ATC class and a module in a network. From our results, we find that the N_{approved} and N_{A549} have four modules which are enriched for a single ATC class, N_{MCF7} and N_{PC3} have two such modules, N_{all} and N_{A375} have one module, and N_{VCAP} has no significant module.

The interpretation for these results is that each module is characteristic for a set of drugs represented by an ATC code and could be used to predict the function of unknown drugs within this module because they are likely to have common targets. This could be used to predict the function of unknown drugs or drug repositining.

Single-drug class in multiple modules.

For this type, an ATC class is enriched in more than one module. For instance, ATC class L is enriched in 3 modules in N_{A375}; see the vertical boxes in Fig. 5. Furthermore, ATC class G is enriched in two modules in N_{PC3}. This suggests that drug class G and L have possibly three, respectively two independent target-sets effected by these drugs. This means ATC classes G and L have multiple target sets which are at least partially independent from each other.

The interpretation is that if in a network a single ATC class is enriched in multiple modules, the drugs from this ATC class are heterogenously separated targeting different subsets of genes.

Multiple-drug classes in a single module.

For this type, we find more than one ATC class enriched in a module. The N_{approved} network has three ATC classes (D, R, and S) enriched in module 2; see the horizontal boxes in Fig. 5. The netwok N_{PC3} has two modules enriched with two drugs. Specifically, module 2 is enriched by ATC class C and G and moduel 8 is enriched by ATC class D and S. Finally, N_{all} has module 6 enriched by ATC class D and H.

Our interpretation for this is if multiple ATC classes are enriched in a single module, this means that, e.g., two drugs from two different ATC classes have at least partially common targets. These targets hight be higher order, i.e., not directly targeted by a drug but further downstream, but enough to change the differential expression of such genes. This could be used to predict a drug repurposing.

Multiple Drug classes in multiple modules.

For this type, we find an ATC class enriched in multiple modules together with further enriched ATC classes; see the intersection of a horizontal and vertical boxe in Fig. 5. For this type, we find merely one network N_{PC3} whereas ATC class G is enriched in module 2 and 13 and the enrichment in module 2 is shared with ATC class C.

This result indicates that a drug class has multiple independent target-sets and could be used for predicting the repurposing of known drugs as well as predicting the function of unknown drugs.

Combining all our findings, our results have a similarity to the conceptual idea of *cancer attractors* introduced by[32,42] and, e.g., studied in[33,34]. The authors analyzed gene regulatory networks and showed that cell types can be seen as attractors in the epigenetic landscape representing the phenotype space of an organism, see Fig. 1A. That means the developmental state of cells giving raise to different cell fates can be seen as dynamical gene networks chaning their structure over time and as a consequence changing their position in the epegenetic landscape. Similar studies have been conducted by[43–45]. In[33] it has been argued that cancer cells are trapped in abnormal attractors allowing in this way the extension of the conceptual idea of *attractors* in gene regulatory networks to general abnormal or tumor cell types in diseaes beyond cancer[46–48].

Our study adds in a non-trivial way to this because we do not study gene regulatory networks but DANs, where the drugs/compounds correspond to the nodes of the network instead of genes. Due to the fact that we determine the similarity between pairs of drugs based on hundreds or even thousands of expression profiles, for certain conditions, a DAN integrates dozens of individual gene regulatory networks, each representing a particular cell state, see Fig. 1A. This includes a temporal integration of the cells due to the perturbation effect to the exposed drugs. This means that despite the fact that the DANs are static they nevertheless represent dynamical states of the underlying cells. Hence, a DAN is capable of representing many different states of cells, corresponding to phenotypes, simultenously and allows the integrated representation of the drug landscape.

It is important to emphasize the difference between the different 'spaces' considered. GRNs are embedded into the genotype space describing the activity of genes, whereas the epigenetic landscape, representing the phenotype space, describes cell states and their transitions. Here a cell state can correspond to a normal cell type or an abnormal tumour or disease cells. These states are the *attractors* of[32,42]. Each cell state has a corresponding GRN and, hence, a projection into genotype space. Our DANs are embedded into the compound space representing therapeutic interventions. Each state in the compound space corresponds to a drug/compound that is connected to the phenotype space to abnormal and normal cell states. The connection between these three spaces is visualized in Fig. 1A.

For our DANs, we found a graph-theoretical correspondence of an 'attractor' state in phenotype space, by the modules in the networks in the compound space. This could be demonstrated by utilizing information about the ATC classification of known drugs. In this way we complemented LINCS with information from DrugBank about known effects of drugs.

For enabling an efficient exploration and reusage of our results, we developed an interactive web interface that can be used to view, explore, and link drug associations for our results. The interface also provides an integration with external resources via added links, curated mappings, and external IDs. Content from other resources such as PubChem has been incorporated into the DAN web interface enabling End users to view information and explore new hypotheses of drug associations. These features could facilitate further research in the field on a large-scale and in addition could provide health care professionals with a valuable systems pharmacogenomics source.

Finally, we would like to note that it appears desirable to integrate different types of genomics data, e.g., transcriptomics, proteomics and metabolomics data, to establish in this way an integrated systems pharmacogenomics landscape of drug similarities. Unfortunately, the LINCS database, on which our analysis is based, nor any other current database, does not provide those different types of data that would allow to realize this approach practically. For this reason, our approach is the most feasible one considering the current practical data constraints and can be as an approximation of thereof. On a more theoretical note, we would like to add that even if one could realize an integrated systems pharmacogenomics landscape it is unclear if all different genomics data types are actually required or if they are, at least partially, redundant. Only future studies can shed light on this conceptual issue.

Conclusion

In this paper, we developed a systems pharmacogenomics approach and applied it to data from the LINCS repository. As a result, we constructed *Drug Association Networks* summarizing hundreds of drugs and thousands of compounds systematically with respect to their therapeutic effects. We showed that the modular structure of the DANs represent enriched ATC classes thus integrating the drug induced changes on the genotype states of the cells.

Materials and Methods

Drug perturbation data from LINCS data.

The LINCS L1000 data comprises of 5806 genetic perturbations (e.g., single gene knockdown and over-expression) and 16,425 perturbations induced by chemical compounds (e.g., drugs)[49]. About 1.3 million gene expression have been profiled and collected for this project using the L1000 technology[50]. The L1000 platform has been developed at the Broad Institute by the connectivity map (CMap) team to facilitate rapid, flexible and high throughput gene expression profiling at a lower cost. However, the L1000 technology only measures expression for 978 *landmark* genes and the expression values for the rest of the transcriptome are estimated using a computational model based on Gene Expression Omnibus (GEO)[51] data. In this paper, we used the level 5 signature data of drug perturbations in various cell lines. Overall, the LINCS data were generated from a multifacturial experimental space, see Fig. 1B.

DrugBank database.

DrugBank is a comprehensive drug data resource that contains records about chemical, pharmacological, and pharmaceutical features of more than 8,000 drugs, including the 2016 FDA-approved drugs[52]. We used version 5.0.11 (released 2017-12-20) of the DrugBank database for our analysis. To make the cross-platform comparisons compatible, we considered the DrugBank ID as the identifier of drugs across the DrugBank and LINCS databases. For our analysis, we used the Anatomical Therapeutic Chemical (ATC) classification codes, controled by the WHO, shown in Table 2. This classification categorizes drugs into different groups/classes according to the organ or system on which they act, their therapeutic effect, and their chemical characteristics. For our analysis we use the first ATC level, which gives 14 main anatomical classes.

Metadata pipeline.

The LINCS data API provides a programmatic pipeline to annotations and perturbational signatures in the L1000 dataset via a collection of HTTP-based RESTful web services. An example of these services includes; Cell Service, which is a service that describes the cell line meta-information. The API services provided by the LINCS API for querying the L1000 metadata support complex queries via simple HTTP GET requests that can be executed in a web browser or most programming languages such as R and Python.

Transcriptional profiles and small molecules diversity. We downloaded the L1000 raw z-score vectors from the GEO repository and pre-processed them using the R L1000 tools[53]. A signature of a small molecule is defined as a vector of z-score values, representing the differential expression between samples treated with small molecules and control samples. That means a z-score signature summarizes the effect of the treatment with a small molecule. This is in depencence on experimental condition, e.g., dosage, time point, cell line etc.

In total, there are 169, 239 z-score signature profiles marked with the highest signature count that satisfied the well- and plate-based quality control. This signature profile subset covers 20, 009 small molecules (out of 49, 400 perturbagens) that were repeatedly measured with 1 to 8 replicates. For our analysis, we select the time points 6, 24 and 48 *h* because they represent by far the majority of conditions. From this we find in total 158, 054 signature profiles (i.e., any combination of the small molecule, time, and cell line) we use for our analysis. In Table 3, we show some summary statistis of this data set.

The z-score signature vectors were used to study the effect of a drug treatment on the differential expression of genes. We used the threshold >2.0 to indicate upregulation and <-2.0 to indicate down-regulation of a gene respectively.

Mapping small molecules to external databases. The L1000 small molecules were assayed across multiple cell lines, experimental replicates, dosages and time points. For this reason, we mapped DrugBank compounds and the directly measured (landmark) genes to calculate a single transcriptional profile across multiple signatures for each L1000 small molecule. We also mapped the L1000 small molecules to external database sources in UniChem database[54]. We achieved this by querying UniChem with the InChIKey of each L1000 compound via UniChem API. This allows us to map the L1000 small molecules not only to DrugBank, but also to PubChem, ChEMBL, and KEGG Ligand covered by UniChem (see Table T1 in Supplementary File 1). The pipeline enables us also to identify FDA-Approved drugs and to map them to the L1000 small molecule identifiers.

After mapping the DrugBank identifiers to small molecules, the identifiers were used to calculate the signature profile consensus for each drug. The purpose for computing consensus is to combine signature profiles for the same perturbation under different conditions (e.g., cell types, different dosages, or time points). The signature profiles consensus were obtained using the following; First, we calculated the Spearman rank correlation of all signatures that belong to a drug identifier in DrugBank. Second, we calculated the weights by taking the mean correlation to normalize the similarities (Total correlation, see Fig. S1 in Supplementary File 1). Third, we multiplied the z-score signatures by their similarity weights. Last, we sum up the weighted z-score vectors to form a single signature consensus.

Drug association network. The basic idea of the drug association network (DAN) is to generate a network where different drugs show a similar effect on gene expressions which means that the number of genes affected by them has the same type of expression profiles compared to the control data. For example, for a particular cell line treated by drug D_i and D_j having observed phenotype changes $\hat{P}_i$ and $\hat{P}_j$, these phenotypes will be similar $\left(\hat{P}_i \sim \hat{P}_j\right)$ if the two drugs influence (overexpression or underexpression compare to a control state) similar genes. In order to estimate the similarity between two drugs we use a Jaccard-like index[55] between two vectors of genes which are characterized as 1 (up), -1 (down) and 0 (no change) by drugs D_i and D_j. In the first step, we obtain a matrix by converting the z-scores of drug-treated expression data to a matrix of categorical data-type whereas rows represent genes and drugs correspond to columns. In this matrix, genes are categorized as differentially expressed and non-differentially expressed genes. The differentially expressed genes are labelled by 1, for up-regulated, and -1 for down-regulated. The non-differentially expressed genes are labelled by 0. In the second step, we measure the overlapping score between pairs of drugs by using a JI as described in Eqn. 1. The JI gives a ratio of differentially expressed genes which are common between a pair of drug-treated data w.r.t. all other genes which are differentially expressed in at least one drug-treated data. In the third step, we test the significance of the Jaccard Index. We perform the significance test with a non-parametric approach by randomizing gene labels of each drug data vector independently. This allows us to estimate the sampling distribution of the null hypothesis. A schematic overview for the construction of a DAN is shown in Fig. 1D.

Jaccard Index. Let D_k and D_l be two drugs with regulation profiles R_i and R_j. R_i and R_j are two vectors of length n, whereas n is the number of genes. Their components correspond to (I) down-regulation (-1), (II) no-change (0) or (III) up-reguation (1). The Jaccard Index (JI) can be estimated from the contingency table (see Table 1) giving the overlap between the two regulation profiles representing the effect of the drugs D_k and D_l:

$$J_{ij} = J(R_i, R_j) = \frac{\left\|G_i \cap G_j\right\|_{/\|0,0\|}}{\left\|G_i \cup G_j\right\|_{/\|0,0\|}} = \frac{n_{11} + n_{33}}{n_t} \tag{1}$$

here $n_t = n_{11} + n_{12} + n_{13} + n_{21} + n_{23} + n_{31} + n_{32} + n_{33}$ is the number of genes showing differential expression.

Construction of the drug association network. The construction procedure for the DAN consists of 11 steps and is based on z-score vectors available in LINCS. Every z-score vector, $Z = \{z_1, z_2 ..., z_n\}$ whereas n is the total number of genes, is a function of experimental conditions, including a drug D_k and a cell line CL_m, which was exposed to drug D_k. For briefity we simply write $Z = Z(D_k,\gamma)$ to indicate that a z-score is a function of drug D_k and further conditions summarized by γ. We call (D_k,γ) a configuration. Due to this dependency, $Z = Z(D_k,\gamma)$ can be seen as a profile for drug D_k.

For reasons of notational simplicity, we index the configurations (D_k,γ) by an integer number. That means we map (D_k,γ) to $c_h \in C = \{c_1, \ldots, c_t\} = \{1, \ldots, t\}$, whereas t is the total number of configurations. This leads to the notation

$$Z = Z(D_k,\gamma) = Z(c_h) \tag{2}$$

we will use in the following.

1. This step is only used for N_{approved}: Summarize the z-scores for all configurations with the same drug, i.e., $DC_k = \{c_i, c, \ldots c_k\}$ whereas every $x \in DC_k$ contains drug D_k. The summarized values are given by

   $$Z' = \frac{1}{n} \sum_{x \in DC_k} Z(x). \tag{3}$$

   In this case the total number of remaining z-scores corresponds to the number of configurations and the number of drugs. Re-indexing of the configurations gives $c_h \in C = \{c_1, \ldots, c_t\}$ whereas t is now the number of different drugs.
2. Convert every z-score vector into a p-value vector, $P = \{p_1, p_2 \ldots, p_n\}$, i.e., $P = P(c_h)$.
3. Convert every p-score vector into a q-value vector (controlling FDR with Benjamini and Hochberg (BH) method[56]), $Q = \{q_1, q_2 \ldots, q_n\}$, i.e., $Q = Q(c_h)$.
4. Construct a matrix R of differentially regulated genes for all configurations c_h, i.e., R is a $(n \times t)$ matrix, whereas the components of this matrix correspond to (I) down-regulation (−1), (II) no-change (0) or (III) up-reguation (1).:
   For each configuration c_h, we have the corresponding z-score vector $Z(c_h)$ and the corresponding q-value vector $Q(c_h)$. The function $f{:}(Z(c_h), Q(c_h))_i \to M$ maps from the q- and z-value of a gene i to its regulation categories, i.e., $M = \{-1, 0, 1\}$. Specifically, the function $f(z_i(c_h), q_i(c_h))$ is defined as follows:

   $$f(z_i(c_h), q_i(c_h)) = \begin{cases} -1 & :q_i(c_h) \leq \alpha \text{ and } z_i(c_h) < 0 \\ 1 & :q_i(c_h) \leq \alpha \text{ and } z_i(c_h) > 0 \\ 0 & :\text{otherwise} \end{cases}$$

   This gives $R_{i,h} = f(z_i(c_h), q_i(c_h))$.
5. Using R to calculate the Jaccard index (J_{ij}) as defined in Eqn. 1 for each pair of configurations c_i and c_j, with $c_i \neq c_j$ and $c_i, c_j \in C$. Specifically, calculate $J_{ij} = J(R_i, R_j)$, whereas the R_i and R_j are the columns of matrix R for the configurations c_i and c_j.
6. Test the significance of a Jaccard Index for each pair of configurations by the following hypothesis.
   H_0: The number of differentially expressed genes overlapping in two dataset treated by drugs D_i and D_j is zero.
   H_1: The number of differentially expressed genes overlapping in two dataset treated by drugs D_i and D_j is not zero.
7. The sampling distribution is obtained from gene-label randomizations for each pair of configuration profiles R_i and R_j from which the corresponding Jaccard index, $J_{ij} = J(R_i, R_j)$, is determined. This results in the permuted Jaccard indices, $J_{perm}(ij) = \{j_{ij}^{perm_1}, j_{ij}^{perm_2} \ldots j_{ij}^{perm_L}\}$ for $L = 2000$.
8. From $J_{perm}(ij)$, we estimate the p-values by:

   $$p_{i,j} = Pr(j_{i,j} > j_{i,j}^{perm}) = \frac{\sum_{k=1}^{L} I(j_{i,j} > j_{i,j}^{perm_k})}{L}$$

   This gives $P^J = \{p_{1,2}, p_{1,3}, \ldots, p_{n,n-1}\}$, containing in total $\frac{t \cdot (t-1)}{2}$ different p-values.
9. Controling the FDR by BH we convert P^J into q-values, $Q^J = \{q_{1,2}, q_{1,3}, \ldots, q_{n,n-1}\}$, consisting in total of $\frac{t \cdot (t-1)}{2}$ different q-values.
10. Construct a matrix B for all configurations C by using the q_{ij} values:

    $$B_{c_i,c_j} = \begin{cases} 1 & :q_{i,j} \leq \alpha \\ 0 & :\text{otherwise} \end{cases} \tag{4}$$

    Here $c_i, c_j \in C$.
11. Construct a DAN by summarizing all configurations with the same drug, i.e., $DC_k = \{c_i, c, \ldots c_k\}$ whereas every $x \in DC_k$ contains dug D_k

    $$A_{D_k,D_l} = \Theta\left(\sum_{x \in DC_k, y \in DC_l} B_{xy}\right) \tag{5}$$

here $\Theta(w)$ is the theta function which gives 1 for $w > 0$ and 0 otherwise.

References

1. Keiser, M. J. *et al*. Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
2. Dunkel, M., Günther, S., Ahmed, J., Wittig, B. & Preissner, R. Superpred: drug classification and target prediction. *Nucleic acids research* **36**, W55–W59 (2008).
3. Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
4. Iorio, F. *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci.* **107**, 14621–14626 (2010).
5. Finley, S. D., Chu, L.-H. & Popel, A. S. Computational systems biology approaches to anti-angiogenic cancer therapeutics. *Drug discovery today* **20**, 187–197 (2015).
6. Lamb, J. *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science* **313**, 1929–1935 (2006).
7. Jiang, W. *et al*. Identification of links between small molecules and mirnas in human cancers based on transcriptional responses. *Sci. reports* **2**, 282 (2012).
8. Subramanian, A. *et al*. A next generation connectivity map: {L1000} platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452. e17, https://doi.org/10.1016/j.cell.2017.10.049 (2017).
9. Wang, Z., Clark, N. R. & Ma'ayan, A. Drug-induced adverse events prediction with the lincs l1000 data. *Bioinformatics* **32**, 2338–2345 (2016).
10. Li, J. *et al*. A survey of current trends in computational drug repositioning. *Briefings bioinformatics* **17**, 2–12 (2015).
11. Musa, A., Tripathi, S., Kandhavelu, M., Dehmer, M. & Emmert-Streib, F. Harnessing the biological complexity of big data from lincs gene expression signatures. *PloS one* **13**, e0201937 (2018).
12. Musa, A. *et al*. A review of connectivity map and computational approaches in pharmacogenomics. *Briefings Bioinforma*. bbw112–bbw112 (2017).
13. Nassiri, I. & McCall, M. N. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic acids research* (2018).
14. Caicedo, J. C., Singh, S. & Carpenter, A. E. Applications in image-based profiling of perturbations. *Curr. opinion biotechnology* **39**, 134–142 (2016).
15. De Wolf, H., De Bondt, A., Turner, H. & Göhlmann, H. W. Transcriptional characterization of compounds: lessons learned from the public lincs data. *Assay drug development technologies* **14**, 252–260 (2016).
16. Aliper, A. *et al*. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. pharmaceutics* **13**, 2524–2530 (2016).
17. Sirci, F. *et al*. Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses. *NPJ systems biology applications* **3**, 23 (2017).
18. Chen, B. *et al*. Relating chemical structure to cellular response: an integrative analysis of gene expression, bioactivity, and structural data across 11,000 compounds. *CPT: pharmacometrics & systems pharmacology* **4**, 576–584 (2015).
19. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
20. Piening, S. *et al*. Impact of safety-related regulatory action on clinical practice. *Drug safety* **35**, 373–385 (2012).
21. Beadle, G. W. & Tatum, E. L. Genetic control of biochemical reactions in neurospora. *Proceedings Natl. Acad. Sci.* **27**, 499–506 (1941).
22. Vidal, M. A unifying view of 21st century systems biology. *FEBS letters* **583**, 3891–3894 (2009).
23. Wang, L. Pharmacogenomics: a systems approach. *Wiley Interdiscip. Rev. Syst. Biol. Medicine* **2**, 3–22 (2010).
24. Yıldırım, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug—target network. *Nat. biotechnology* **25**, 1119 (2007).
25. Hu, G. & Agarwal, P. Human disease-drug network based on genomic expression profiles. *PloS one* **4**, e6536 (2009).
26. Ye, H., Liu, Q. & Wei, J. Construction of drug network based on side effects and its application for drug repositioning. *PloS one* **9**, e87864 (2014).
27. El-Hachem, N. *et al*. Integrative cancer pharmacogenomics to infer large-scale drug taxonomy. *Cancer research* (2017).
28. Sorger, P. K. *et al*. Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. In *An NIH white paper by the QSP workshop group*, vol. 48 (NIH Bethesda, MD, 2011).
29. Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *J. Stat. Mech. Theory Exp.* **2005**, P09008 (2005).
30. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. United States Am.* **99**, 7821–7826 (2002).
31. Tripathi, S., Moutari, S., Dehmer, M. & Emmert-Streib, F. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC bioinformatics* **17**, 129 (2016).
32. Kauffman, S. Differentiation of malignant to benign cells. *J. Theor. Biol.* **31**, 429–451 (1971).
33. Huang, S., Ernberg, I. & Kauffman, S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. In *Seminars in cell & developmental biology*, vol. 20, 869–876 (Elsevier, 2009).
34. Mar, J. C. & Quackenbush, J. Decomposition of gene expression state space trajectories. *PLoS computational biology* **5**, e1000626 (2009).
35. Jiang, W. *et al*. Expression of thyroid hormone receptor alpha in 3t3-l1 adipocytes; triiodothyronine increases the expression of lipogenic enzyme and triglyceride accumulation. *J. endocrinology* **182**, 295–302 (2004).
36. Mai, W. *et al*. Thyroid hormone receptor a is a molecular switch of cardiac function between fetal and postnatal life. *Proc. Natl. Acad. Sci.* **101**, 10332–10337 (2004).
37. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. statistical mechanics: theory experiment* **2008**, P10008 (2008).
38. Chen, L., Zeng, W.-M., Cai, Y.-D., Feng, K.-Y. & Chou, K.-C. Predicting anatomical therapeutic chemical (atc) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS one* **7**, e35254 (2012).
39. Raymond, M. & Rousset, F. An exact test for population differentiation. *Evolution* **49**, 1280–1283 (1995).
40. Tilkov, S. & Vinoski, S. Node. js: Using javascript to build high-performance network programs. *IEEE Internet Comput.* **14**, 80–83 (2010).
41. Wang, R., Perez-Riverol, Y., Hermjakob, H. & Vizcaíno, J. A. Open source libraries and frameworks for biological data visualisation: A guide for developers. *Proteomics* **15**, 1356–1374 (2015).
42. Huang, S. & Kauffman, S. How to escape the cancer attractor: rationale and limitations of multi-target drugs. In *Seminars in cancer biology*, vol. 23, 270–278 (Elsevier, 2013).
43. Cheng, W.-Y., Yang, T.-H. O. & Anastassiou, D. Biomolecular events in cancer revealed by attractor metagenes. *PLoS computational biology* **9**, e1002920 (2013).
44. Li, Q. *et al*. Dynamics inside the cancer cell attractor reveal cell heterogeneity, limits of stability, and escape. *Proc. Natl. Acad. Sci.* **113**, 2672–2677 (2016).
45. Creixell, P., Schoof, E. M., Erler, J. T. & Linding, R. Navigating cancer network attractors for tumor-specific therapy. *Nat. biotechnology* **30**, 842 (2012).

46. Emmert-Streib, F. The chronic fatigue syndrome: a comparative pathway analysis. *J. computational biology* **14**, 961–972 (2007).
47. Del Sol, A., Balling, R., Hood, L. & Galas, D. Diseases as network perturbations. *Curr. opinion biotechnology* **21**, 566–571 (2010).
48. Emmert-Streib, F. & Glazko, G. V. Network biology: a direct approach to study biological function. Wiley Interdiscip. Rev. Syst. *Biol. Medicine* **3**, 379–391 (2011).
49. Duan, Q. *et al*. Lincs canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures. *Nucleic acids research* **42**, W449–W460 (2014).
50. Vidović, D., Koleti, A. & Schürer, S. C. Large-scale integration of small molecule-induced genome-wide transcriptional responses, kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systemslevel drug action. *Front. genetics* **5**, 342 (2014).
51. Barrett, T. *et al*. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991–D995 (2012).
52. Wu, P., Nielsen, T. E. & Clausen, M. H. Small-molecule kinase inhibitors: an analysis of fda-approved drugs. *Drug Discov. Today* **21**, 5–10 (2016).
53. Lincscloud. LINCS L1000 R tools. http://support.lincscloud.org/hc/en-us/articles/202062163-L1000-Code-via-GitHub-(2014). [Online; accessed 19-July-2016].
54. Chambers, J. *et al*. Unichem: extension of inchi-based compound mapping to salt, connectivity and stereochemistry layers. *J. cheminformatics* **6**, 43, https://doi.org/10.1186/s13321-014-0043-5 (2014).
55. Jaccard, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* **37**, 547–579 (1901).
56. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. royal statistical society. Ser. B (Methodological)* 289–300 (1995).

Acknowledgements

A.M. thanks the CIMO foundation Finland for a scholarship. M.D. thanks the Austrian Science Funds for supporting this work (Project P30031).

Author Contributions

A.M., S.T. and F.E.S. conceived the study and conducted the analysis, A.M., M.D. and F.E.S. interpreted the results. All authors wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-44291-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.