Annual Review of Biomedical Data Science

Connectivity Mapping: Methods and Applications

Alexandra B. Keenan, Megan L. Wojciechowicz, Zichen Wang, Kathleen M. Jagodnik, Sherry L. Jenkins, Alexander Lachmann, and Avi Ma'ayan

Department of Pharmacological Sciences and Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; email: avi.maayan@mssm.edu

Annu. Rev. Biomed. Data Sci. 2019. 2:69–92

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-072018-021211

Keywords

systems biology, systems pharmacology, network analysis, drug repositioning, signature commons, responsome

Abstract

Connectivity mapping resources consist of signatures representing changes in cellular state following systematic small-molecule, disease, gene, or other form of perturbations. Such resources enable the characterization of signatures from novel perturbations based on similarity; provide a global view of the space of many themed perturbations; and allow the ability to predict cellular, tissue, and organismal phenotypes for perturbagens. A signature search engine enables hypothesis generation by finding connections between query signatures and the database of signatures. This framework has been used to identify connections between small molecules and their targets, to discover cell-specific responses to perturbations and ways to reverse disease expression states with small molecules, and to predict small-molecule mimickers for existing drugs. This review provides a historical perspective and the current state of connectivity mapping resources with a focus on both methodology and community implementations.

HISTORICAL PERSPECTIVE AND INITIAL SIGNATURE COLLECTIONS

In the year 2000, a comprehensive reference database of gene expression signatures was created from systematic pharmacological and genetic perturbations to enable the characterization of new small molecules and unknown open reading frames (1). This database was constructed from a compendium of *Saccharomyces cerevisiae* gene expression signatures resulting from 300 mutational and chemical perturbations under a single growth condition to enable the direct comparison of all signatures. The authors reasoned that a comprehensive reference database of gene expression profiles following systematic pharmacological, mutational, or disease perturbations would enable unknown drugs and mutations that lead to disease to be characterized based on matching gene expression patterns. In a subsequent perspective article, the authors of this original study conjectured that lower-cost gene expression assays would be needed for generating a larger collection of reference signatures (2). The utility of creating a compendium of gene expression signatures was subsequently demonstrated in additional contexts. For example, a library of signatures was constructed for antidepressants, antipsychotics, and opioid action drugs using expression profiling (3). Around the same time, a collection of gene expression signatures was created for 15 hepatotoxic agents in rats to characterize the dimensionality, or the "space," of various types of phenotypically observed liver toxicities (4). A similar study profiled 26 reference compounds for hepatotoxicity in conjunction with physiological measures such as liver and kidney function in rats (5). Shortly thereafter, such studies were expanded to create the now publicly available resource DrugMatrix (6, 7). To create DrugMatrix, the biotech company Iconix Pharmaceuticals developed a signature compendium for over 600 benchmark drugs and toxicants and tested compounds in rats where gene expression, quantified with microarrays, was used to profile several major tissues. The data set, originally owned by Iconix Pharmaceuticals and Entelos, was acquired by the National Institute of Environmental Health Sciences (NIEHS) in 2010 and was made openly available for researchers. DrugMatrix is still a leading comprehensive publicly available resource for gene expression signatures from an in vivo mammalian system. The main reasons that follow-up studies did not test large collections of compounds in mammals in a similar way were growing concerns about animal welfare and the cost associated with such projects.

In a landmark publication from 2006, the term "connectivity map" or "CMap" was introduced (8). To create the original version of CMap, which we here term CMap 1.0, Affymetrix$^{\text{TM}}$ microarray gene expression profiles were collected for 164 small molecules applied to four human cell lines in different concentrations, and gene expression was measured at two time points for a total of 453 signatures. The CMap 1.0 collection of signatures was later expanded to cover ~1,300 small molecules that include most of the Food and Drug Administration (FDA)-approved drugs, for a total of more than 6,000 signatures. Similarity between signatures was computed using the gene set enrichment analysis method (9), which is an adaptation of the Kolmogorov–Smirnov test for comparing a ranked list to an unordered reference set of matching elements. Importantly, the original CMap 1.0 publication demonstrated how connectivity mapping can be used for generating testable hypotheses regarding uncharacterized small molecules. For example, to better understand how the small molecule gedunin acted to negate androgen receptor activity in prostate cancer cells, researchers queried a gedunin-induced signature against all the signatures in CMap 1.0 to find several known HSP90 inhibitors among the top hits, suggesting that gedunin is also an HSP90 inhibitor.

Following the path set by the CMap 1.0 framework, many drug repositioning studies utilized the resource in a similar way. CMap 1.0 was directly used to reposition drugs for obesity (10), ovarian cancer (11), breast cancer (12), influenza (13), cigarette smoke–induced inflammation (14), gastric cancer (15), inflammatory bowel disease (16), osteogenic differentiation (17), lung adenocarcinoma (18, 19), skeletal muscle atrophy (20), and renal cell carcinoma (21); to find compounds with estrogenic activity (22); to induce bone anabolism (17, 23); to target cancer stemness (24); and to find radiosensitizing agents for treating lung cancer (25). In most of these studies, the identified drugs and small molecules were experimentally validated in vitro, and in some cases in animal models.

AN ILLUSTRATIVE EXAMPLE

Next, we describe an illustrative high-profile example that utilized connectivity mapping to predict that celastrol, a natural compound isolated from the root extracts of *Tripterygium wilfordii*, acts as an antiobesity agent in mice with diet-induced obesity (**Figure 1**) (26). The first step of the study was the construction of a query signature. Given the potential heterogeneity in disease signatures across samples, tissue types, and disease states, the designation of signatures to perform the query is critical. Since it was previously shown that endoplasmic reticulum (ER) stress is linked to obesity, the goal of the signature query was to find small molecules whose signature mimicked signatures that reflected restoration of ER homeostasis. To construct the query signature, the authors utilized microarray data obtained from experiments that treated mice with interventions known to relieve ER stress and increase leptin sensitivity in mouse liver and hypothalamus. Querying the CMap 1.0 connectivity mapping resource with obesity-relevant liver and hypothalamus signatures and then implementing an integration scheme that combines enrichment scores from multiple searches, the authors concluded that celastrol was the top hit. Celastrol was subsequently experimentally

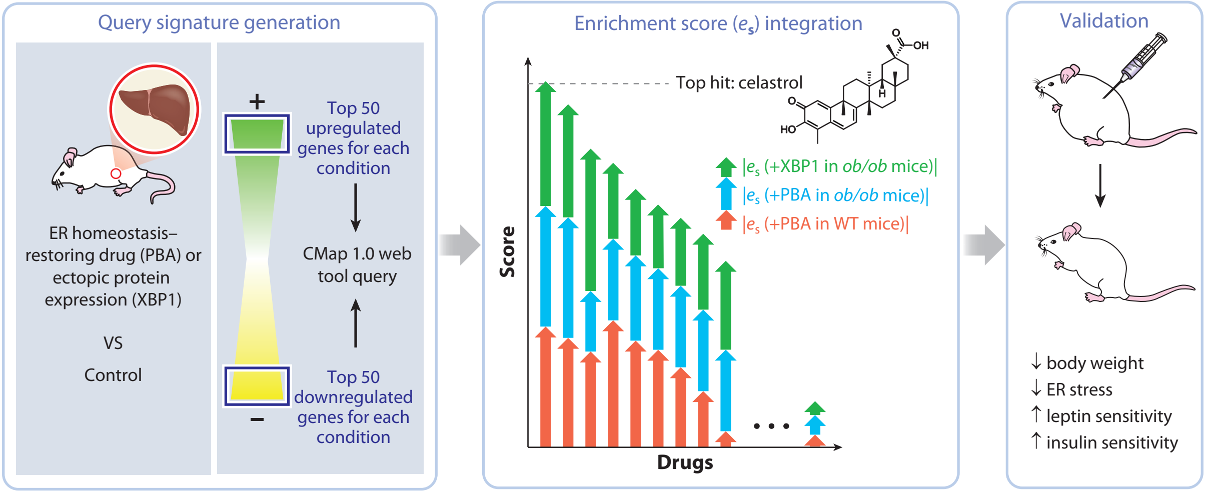

Query signature generation
ER homeostasis–restoring drug (PBA) or ectopic protein expression (XBP1)
VS
Control
+
Top 50 upregulated genes for each condition
CMap 1.0 web tool query
Top 50 downregulated genes for each condition
–
Enrichment score (e_s) integration
Top hit: celastrol
O
OH
H
O
HO
Score
Drugs
• • •
|e_s (+XBP1 in *ob/ob* mice)|
|e_s (+PBA in *ob/ob* mice)|
|e_s (+PBA in WT mice)|
Validation
↓ body weight
↓ ER stress
↑ leptin sensitivity
↑ insulin sensitivity

Figure 1

An illustrative example of the use of connectivity mapping. Gene expression signatures from mouse liver and hypothalamus were obtained by Liu et al. (10) under several experimental conditions designed to reflect pathways that restore ER homeostasis. The top 50 up and down differentially expressed genes from each signature were chosen to query the Connectivity Map (CMap 1.0) online tool. The tool compares the query up/down gene sets against all drug perturbation signatures in the database using a KS (Kolmogorov–Smirnov) statistic. An enrichment score e_s between −1 and 1 is returned for each gene signature, which describes how similar (+1) or dissimilar (−1) the query signature is to the drug-induced signature. The product of the absolute value of the enrichment scores for each signature was used to integrate the search results. The top hit, celastrol, was subsequently shown to reduce body weight, suppress food intake, and increase leptin sensitivity in diet-induced obese mice. Abbreviations: ER, endoplasmic reticulum; *ob/ob*, $Lep^{ob/ob}$ mutant; PBA, phenylbutyric acid; WT, wild-type; XBP1, X-box binding protein 1.

validated to show that it increases leptin sensitivity and glucose homeostasis and reduces body weight and ER stress in obese mice.

While certainly not all computationally repositioned drugs and small molecules identified by the studies that utilized the CMap 1.0 connectivity mapping resource moved to clinical trials, many served as leads for further study and consideration. Drug repositioning through the CMap approach challenges the standard, expensive, risky, and time-consuming approaches associated with bringing a new drug to market (27). Drug repositioning via the CMap approach accelerates the path to drug discovery because existing drugs have pharmacokinetic and clinical data that may include absorption, distribution, metabolism, excretion, and toxicity data; safety data, possibly including postmarket surveillance; accepted formulations; large-scale manufacturing methods; and possibly approval by regulatory bodies for human use. This drastically decreases the cost to bring a drug to market for a new indication because it lowers the risk that the drug will fail the clinical trial process due to undiscovered adverse effects—and it is often one of the only paths for development of therapeutics for rare or neglected diseases. In addition to drug repurposing, CMap 1.0 was used for other applications, for example, confirming dysregulated pathways in Down syndrome (28) and hereditary pulmonary arterial hypertension (29) and suggesting small molecules that can direct human embryonic stem cell differentiation into desired somatic cells (30). Global analysis of CMap 1.0 showed how drugs cluster by their known mechanisms of action (MoAs). Clustering by MoA suggested that novel MoAs can be inferred for new drugs and correct the known MoA for misclassified drugs (31). Using similar approaches, CMap 1.0 was utilized to predict drug targets (32, 33). In another unique application, drug combinations were predicted from CMap 1.0 to maximally reverse the gene expression signatures collected from the kidneys of Tg26 mice, a mouse model for HIV-associated nephropathy (34).

CONNECTIVITY MAPPING PRINCIPLES

Taken together, these initial studies demonstrated the potential of large-scale connectivity mapping. Connectivity mapping emerged as an alternative to the mainstream magic bullet structural and molecular biology approaches for drug discovery. Instead of focusing on targeting a specific protein with a small molecule, with the aim that the small molecule will bind to the target to alter its activity in the expected direction while not binding to other targets unintentionally, connectivity mapping provides a shortcut. With connectivity mapping it is not critical to know the target. The global effects of a compound on the cell, as measured by gene expression or other high-content readout assay, provide a proxy to predict the phenotypic effects of the compound on the cell, the tissue, and the entire organism. Connectivity mapping approaches also offer the bonus of directly providing information about the molecular MoAs of the perturbagen by examining the details of the readout assay, for example, the sets of differentially expressed genes from the signatures induced by the perturbagen.

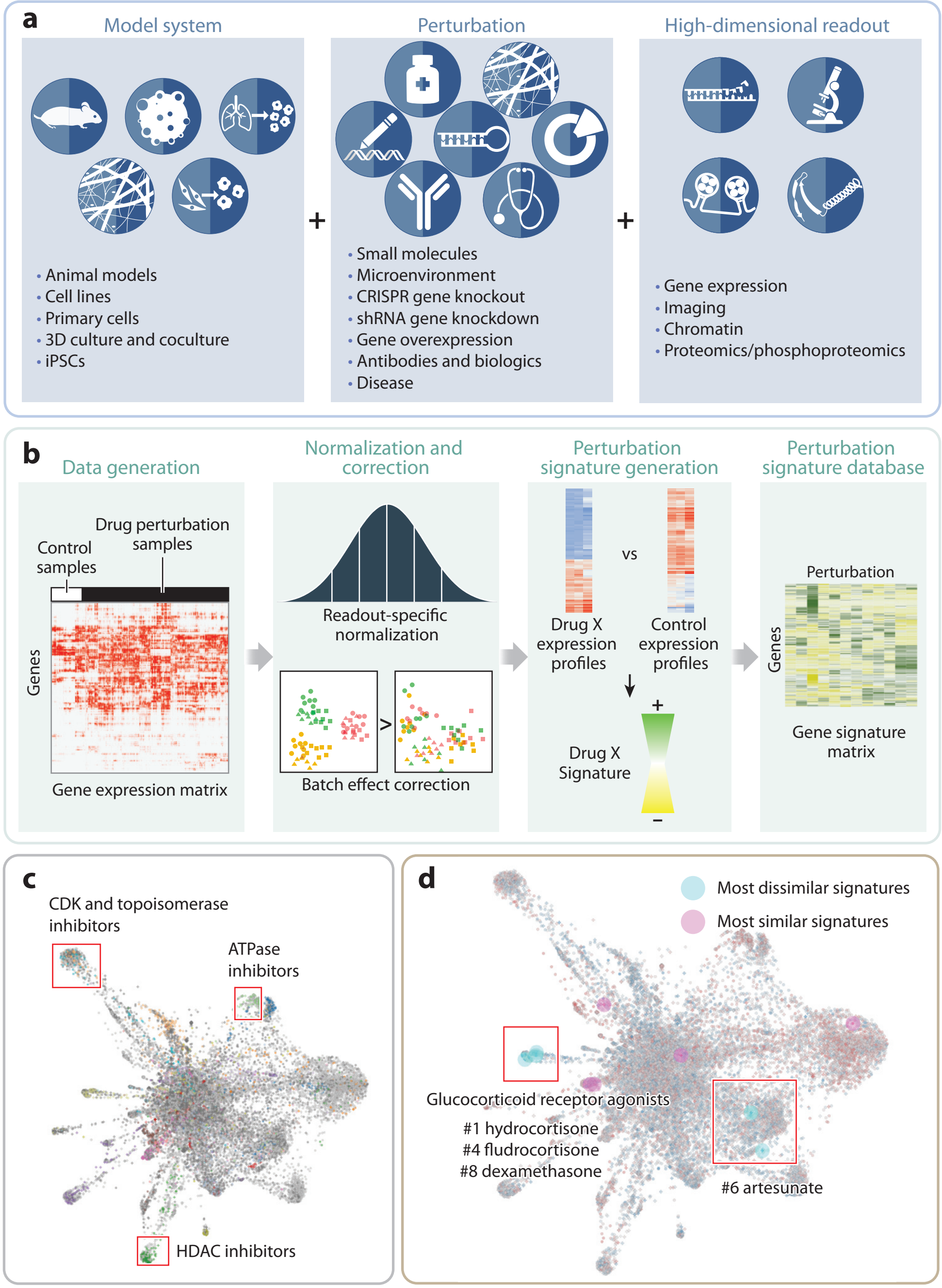
Central to connectivity mapping is the notion that a cellular state compared with some other cellular state, for example, diseased versus healthy cells, can be adequately represented by high-dimensional signatures. These high-dimensional signatures capture a snapshot of the change in cellular state in response to the perturbagen, for example, transcriptional changes induced by some disease, drug, or genetic perturbation. There is a further assumption that signatures from diverse cellular states relate to one another in meaningful ways. Hence, the three central components of a connectivity map resource are (*a*) a large reference database containing signatures representing a change in cellular state in response to a drug, a gene, a disease, or another type of systematic perturbation; (*b*) a query signature; and (*c*) some method to relate the query signature with the entirety of the reference database to discover connections between the query signature and other changes in cellular state. Connections can be defined as opposing/reversing or parallel/mimicking.

CREATING A CONNECTIVITY MAPPING RESOURCE

To create a connectivity mapping resource, researchers conduct systematic perturbations in the model system followed by a high-dimensional readout (**Figure 2*a***). The raw data then needs to be processed and abstracted into signatures (**Figure 2*b***). For example, the first step of analysis of transcriptional readouts typically involves the conversion of the raw data into a matrix format, where the values in the matrix represent levels of the measured transcripts across experimental conditions. Once the data is in a matrix format and appropriately normalized, perturbation signatures are defined and differential expression analysis is conducted. Finally, a signature resemblance measure is used to define similarity between signatures, and then clustering algorithms are applied to the distance matrix to define collections of similar signatures (**Figure 2*c***). This resemblance metric may also be used to return signatures from the compendium most similar or dissimilar to a query signature (**Figure 2*d***).

Batch effects arising from technical variation are some of the complications of experiments utilizing high-throughput technologies (35). Large perturbational compendia are prone to these effects given that they require high-throughput experiments collected by different personnel over long time periods, often across multiple laboratories. When perturbation signatures are compared across batches without corrective measures, spurious biological associations may result while masking true associations. Batch effects can be qualitatively observed by visualizing all gene expression profiles, or computed signatures, using dimensionality reduction techniques such as principal component analysis (PCA) (36) or t-distributed stochastic neighbor embedding (37). Many methods are available to correct for batch effects. For example, Iskar et al. (38) mitigated the batch effect in CMap 1.0 by mean-centering individual perturbation profiles using all of the drug perturbations in the corresponding batch while discarding the vehicle controls. This also enables downweighting common cellular responses to perturbation, such as stress responses. When using a linear model for differential expression analysis (39, 40), it is possible to represent known or suspected batches as covariates (39, 41, 42). These methods correct batch effects as part of the signature generation process. Gene expression values can also be directly corrected during the normalization step, for example, by estimating batch variability using an empirical Bayesian framework (42) or by using a distance-weighted discrimination, which employs support vector machines to find a hyperplane separating the batches (43). Another approach is to remove eigenvectors associated with known batches (44, 45).

Central to connectivity mapping are methods for assessing similarity between signatures. To compare rank-transformed signatures against unordered sets of upregulated and downregulated genes, researchers commonly use the Kolmogorov–Smirnov statistic (8). Similar methods have progressively improved upon the original implementation (46–48). Spearman correlation coefficients (49), the Wilcoxon rank sum test (50), and the cosine distance (51, 52) have been applied to assess similarity between the signatures. Several additional signature similarity methods have been proposed (31, 53, 54). Regardless of the similarity measure, the increasing sizes of signature compendia necessitate fast signature retrieval and comparison. The Blazing Signature Filter transforms compendium data sets into a binary encoding and uses bit operators to compute similarity (55). This method coarsely but quickly searches very large data sets for similar signatures. Advantages and disadvantages of different signature similarity methods have been extensively covered in a previous review article (56).



(Caption appears on following page)

Figure 2 (*Figure appears on preceding page*)

Creating a connectivity map resource. (*a*) Generating a reference compendium of perturbation signatures requires a model system that sufficiently recapitulates the biology relevant to the questions that investigators aim to answer (*left*); some meaningful and useful ways to perturb the cellular state (*center*); and a readout sufficiently granular to represent many different cellular states (*right*). Icons adapted with permission from Pixabay.com, courtesy of Julie McMurry, and from Reference 137. (*b*) Readout data need to be processed and abstracted into signatures after the data are normalized and batch effects are corrected. Once the connectivity map resource is established, query signatures can be projected onto the space of all signatures to find positive and negative connections. (*c*) To illustrate the concept of the space of all signatures, we provide an example from the L1000FWD web-based application (65), which visualizes a subset of the LINCS L1000 drug perturbation signatures. L1000FWD visualizes ~17,000 statistically significant signatures as nodes in a force-directed network, where the edges are not shown. The length of edges represents signature similarity, and colors of nodes/signatures are based on known mechanism of action (MoA) of the drug used to create the signature. Selected clusters of signatures that share the same MoAs are highlighted. (*d*) An example of an unsupervised query of the LINCS L1000 drug perturbation signatures with a chronic lymphocytic leukemia (CLL) disease perturbation signature derived from GEO (Gene Expression Omnibus) (58, 65, 101). L1000 signatures that are most dissimilar (i.e., reverse the disease signature) include corticosteroids, which are used to treat refractory CLL, as well as artesunate, an antimalarial drug that was recently suggested for efficacy in leukemias (138, 139). Abbreviations: CDK, cyclin-dependent kinase; iPSCs, induced pluripotent stem cells; HDAC, histone deacetylase; L1000FWD, L1000 Fireworks Display; LINCS, Library of Integrated Network-Based Cellular Signatures; shRNA, short hairpin RNA.

THE L1000 ASSAY

The initial success of several connectivity mapping efforts in the early to mid-2000s was noticed by pharmaceutical companies, regulatory agencies, and the National Institutes of Health (NIH). Ultimately, an effort to extend CMap 1.0 was established as an NIH Common Fund program called the Library of Integrated Network-Based Cellular Signatures (LINCS) (57). The LINCS program funds research to collect signatures from human cell lines with 21 assays, applied to more than 100 human cell lines and cell types. A central component of the LINCS program is a low-cost gene expression profiling assay called the L1000. The L1000 assay measures the expression level of 978 mRNAs termed the landmark genes via a Luminex bead-based probe hybridization assay (58). These 978 mRNAs are used to infer the expression value of all other human genes using a linear predictive model. These 978 landmark genes were selected based on their orthogonality to capture the variance in the collected gene expression data. The linear inference model was created by analyzing a collection of 12,063 microarray mRNA expression samples, selected from the Gene Expression Omnibus (GEO) (59) and reduced to 384 dimensions with PCA. The initial inference model was later improved by crowdsourcing activities (60) and other community efforts (61). With the L1000 assay, the CMap team at the Broad Institute, which is also one of the NIH-funded LINCS Centers, profiled over 1.3 million samples to generate an initial set of ~400,000 signatures (58). We refer to this resource as CMap 2.0.

Typically, connectivity mapping resources such as CMap only contain transcriptomic signatures. For a LINCS joint project, L1000 signatures were coupled with cell viability measurements under the same conditions. Thousands of signatures were created from applying hundreds of small molecules and drugs to six breast cancer cell lines in different concentrations while gene expression and cell viability were measured at different time points in tandem (62). By comparing growth rate inhibition data with L1000 profiles, it was found that in some cases, breast cancer cell lines rewire their cell signaling pathways to adapt to the insult by the small molecule. Such adaptations can be illuminating to better understand the molecular mechanisms of cancer resistance and recurrence. The authors of this study went a step further to show that targeting independent pathways with combinations of small molecules, determined based on their direction in the expression space of vectors representing the L1000 signatures, can achieve synergistic killing of cells in some contexts. Predicting small-molecule combinations from single-perturbation connectivity mapping data is still underdeveloped. In one study, it was shown that similarity of signatures is predictive of combinations that work (63), while two other studies suggested that complementary signatures work better in some situations (34, 64). All these studies confirmed drug pairs experimentally, one for inducing MCF7 cell death (63), another for killing AML cells (64), and a third for treating kidney disease in a mouse model of HIV-associated nephropathy (34).

Importantly, the compendium of ∼2,000 L1000 and cell viability signatures collected for the LINCS joint project study was visualized as a network where nodes are signatures organized by their similarity (**http://amp.pharm.mssm.edu/LJP**). Coloring the nodes by cell line, cell viability score, drug class, and known drug MoAs results in clusters of signatures with similar effects being highlighted. Such a visualization provides a global view of the dimensionality of the gene expression space for six human cell lines. It shows how many external perturbations converge into few distinct global cellular responses that are likely possible by human cells. Some responses are cell type specific, while others are cell type agnostic. The concept of visualizing a large collection of gene expression signatures as a scattered plot, and then projecting additional prior knowledge on the visualization by changing the color and shape of the points representing signatures, was applied to create the L1000 fireworks display (L1000FWD) software application (65). L1000FWD displays ∼17,000 selected significant L1000 signatures. L1000FWD also serves as a signature search engine where similar, or opposing, signatures are compared with user-submitted signatures. The best matches are highlighted on the map (**Figure 2*d***). Such a visualization provides immediate intuition about the global space of all signatures, and about how a single signature fits within this global space. Following the guide of the L1000FWD visualization, a recent study confirmed the MoAs of two histone deacetylase inhibitors and one topoisomerase inhibitor based on the clustering of these compounds on the L1000FWD map (66).

While the L1000 CMap 2.0 data set is relatively new, several studies have already utilized this massive resource to identify small molecules for drug discovery and repurposing. For example, querying the L1000 data set was used to repurpose the small molecule CGP-60474 as an agent that saves lives of mice infected with a toxic bacterium (67). L1000 CMap 2.0 queries were also utilized to prioritize small molecules for cystic fibrosis (68), melanoma (69), and pancreatic ductal adenocarcinoma (70). In another study, a small-molecule kinase inhibitor was discovered to attenuate the spread of Ebola in human cell lines (52). For that project, the L1000 data were processed with the Characteristic Direction method, a unique method to compute signatures (51). By performing intrinsic and extrinsic benchmarks, the authors showed how the Characteristic Direction method prioritizes more relevant differentially expressed genes for L1000 perturbations compared with the original moderated *Z*-score method from the L1000 data producers. This was determined by examining how signatures cluster by their similarity and are related to known drug targets. Many methods and algorithms have been developed to specifically improve the processing of the L1000 data from CMap 2.0. For example, a two-dimensional spatial bias was found in results from L1000 assays that are carried out in 384-well plates, and a method to correct for such biases was developed and benchmarked (71). Another example is the discovery that there are artifacts in the way expression levels are assigned to genes using the L1000 assay. A method called model-based clustering with data correction was developed to identify and correct these artifacts (72).

The L1000 CMap 2.0 data set can be considered a tensor with cell lines, compounds, compound concentrations, and time points of gene expression measurements as the tensor vectors. This tensor is incomplete because not all possible conditions, or combinations of drugs, cell lines, concentrations, and time points, are covered.

To predict the expression profiles that were not measured, Hodos et al. (73) implemented a tensor completion algorithm to impute expression vectors that filled the missing gaps of experimental conditions. Cross-validation, which involves setting aside some of the collected data to evaluate the predictive model, was applied to calibrate and improve the performance of the tensor completion model (73). The main utility of this approach is that it can obtain expression profiles for specific cell lines that have sparse data. Similarly, the SigMat algorithm (74) attempts to adjust cell-agnostic L1000 search queries according to the background cell type.

The L1000 CMap 2.0 data set was used for other interesting applications; for example, by incorporating L1000 data into a machine learning framework, Wang et al. (75) showed that the L1000 data can improve the prediction of side effects for FDA-approved drugs. Compared with predictions that are based only on the chemical structure of the drugs, or that are based only on image features extracted from cell painting assays before and after compound treatment of a human cell line (76), the incorporation of L1000 data significantly improves side effect predictions. The machine learning model, which was trained on FDA-approved drugs, can be applied to all preclinical compounds profiled by the L1000 assay to predict side effects for these preclinical drugs before they are tested in patients. Hence, connectivity mapping resources can support two types of signature queries, one unsupervised, where the returned results are matched signatures, and another supervised, where the returned results are ranked matching class labels, such as side effects or any other class label determined by prior knowledge about the perturbagen (**Figure 3**).

The L1000 assay has been used by groups outside of the LINCS consortium. For example, a compound screen to induce the expression of the gene that encodes the protein kinase TRIB1 in human HepG2 hepatoma cells used the L1000 to confirm upregulation of TRIB1 and its associated gene neighborhood (77). Additionally, Janssen Pharmaceuticals profiled more than 30,000 compounds in MCF7 cells using the L1000 assay to create an internal library to accelerate their drug discovery platform (78). Importantly, as part of the Tox21 program, the NIEHS, the FDA, the National Center for Advancing Translational Sciences, and the Environmental Protection Agency are currently generating gene expression signatures for 10,000 compounds using a

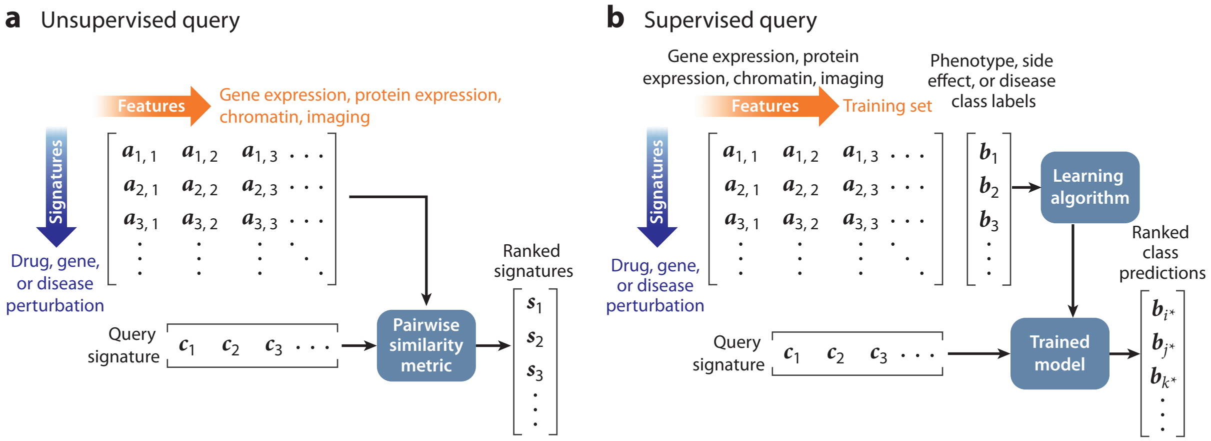

a Unsupervised query
Features
Gene expression, protein expression, chromatin, imaging
Signatures
Drug, gene, or disease perturbation
$a_{1,1}$ $a_{1,2}$ $a_{1,3}$ . . .
$a_{2,1}$ $a_{2,2}$ $a_{2,3}$ . . .
$a_{3,1}$ $a_{3,2}$ $a_{3,3}$ . . .
Query signature
c_1 c_2 c_3 . . .
Pairwise similarity metric
Ranked signatures
s_1 s_2 s_3
b Supervised query
Gene expression, protein expression, chromatin, imaging
Features
Training set
Phenotype, side effect, or disease class labels
Signatures
Drug, gene, or disease perturbation
b_1 b_2 b_3
Learning algorithm
Ranked class predictions
Query signature
Trained model
b_{i^*} b_{j^*} b_{k^*}

Figure 3

Supervised and unsupervised queries with connectivity mapping resources. There are two principal methods for querying a reference database of perturbation signatures with a query signature. (*a*) An unsupervised query involves pairwise comparison of each signature in the resource ($a_{i,j}$) to the query signature (c_j) using a similarity metric such as a Kolmogorov–Smirnov statistic (8) or cosine similarity (52). Query results are displayed as ranked signatures according to their similarity with the query signature. (*b*) A supervised query involves labeling each resource signature ($a_{i,j}$) with a class label (b_i) representing an association of the perturbation with, for example, a phenotype, an adverse side effect, or a disease, as determined by prior knowledge. After training a model on these data with machine learning, a query signature (c_j) can be classified using the trained model, which returns a ranked list of class labels (b_{i^*})—for example, ranked probabilities of a small molecule to induce specific side effects.

modified version of the L1000 assay called S1500+ (79). S1500+ expands the L1000 assay to include an additional ~1,500 genes that represent canonical pathways computationally derived from the MSigDB database (80), with the goal of measuring pathway activity for each compound. While defining the pathways of toxicity is challenging (81), the S1500+ gene selection method promises to cover all the known and most critical molecular toxicology pathways to improve in vitro methods for screening new compounds for potential toxicity.

ALTERNATIVE LOW-COST TRANSCRIPTOMIC TECHNOLOGIES

The L1000 CMap 2.0 approach is currently still the most cost-efficient method to profile gene expression in high throughput for creating a connectivity mapping resource. However, deep sequencing technologies are expected to rapidly become cheaper while providing more accuracy and comprehensiveness. Recently, several methods were described to achieve the goal of lower-cost RNA-seq (RNA sequencing) profiling to create connectivity mapping resources. These methods include RASL-seq (RNA-mediated oligonucleotide annealing, selection, and ligation) (82), TempO-Seq^{TM} (templated oligonucleotide assay with sequencing) (83), PLATE-seq (pooled library amplification for transcriptome expression) (84), and DRUG-seq (digital RNA with perturbation of genes) (85). All four of these new technologies utilize deep sequencing as part of their protocols. The PLATE-seq and DRUG-seq methods multiplex the RNA-seq analysis and produce low-depth reads that can still capture the state of the transcriptome, whereas RASL-seq and TempO-Seq are targeted approaches where the measured genes are predetermined. TempO-Seq covers the entire genome in a similar way to standard cDNA (complementary DNA) microarrays, while RASL-seq only measures a subset of the genome, similar to the way that the L1000 assay can measure only 978 transcripts. Which of these new mRNA expression profiling assays will emerge as the new leader for signature generation to create the next generation of connectivity mapping resources remains to be determined. Recently, members of the NIEHS compared TempO-Seq to S1500+ and concluded that TempO-Seq produces results more consistent with expectations (86).

EXTRACTING SIGNATURES FROM THE GENE EXPRESSION OMNIBUS

Concerted efforts to generate reference signature collections for connectivity mapping have the advantage of measuring perturbations under similar consistent conditions, which yields better reproducibility and comparability. However, there are thousands of publicly available gene expression data sets that can yield valuable connectivity mapping resources. GEO (59) and ArrayExpress (87) are the two leading repositories serving data from published transcriptomics studies. The thousands of data sets in these repositories are organized with accession numbers and metadata. By uniformly reprocessing the data from these repositories, and by identifying the perturbation and control samples, large collections of signatures can be generated.

One of the first efforts toward this end was EXALT (expression signature analysis tool), which automatically extracted thousands of microarray signatures from GEO (88). EXALT relied on the labels assigned by the data submitters to identify groups to be compared. This approach does not define the perturbation or the cell type, nor does it identify the control versus perturbation samples, which is critical for reusability. More toward this end, 790 disease and drug signatures were extracted from microarray experiments from GEO (89). Then, a correlation approach was applied to identify significant disease–disease, drug–drug, and disease–drug relationships. These associations were then visualized as ball-and-stick networks. The authors of this study noted that positive correlations could suggest potential side effects, while negative correlations could suggest novel treatments. In a similar study that appeared a year later, the authors created a disease–drug signature network connecting 99 drugs to 43 diseases via 234 significant associations mined from GEO (90). A more modest study that attempted to improve disease classification through similarity of disease signatures also appeared around the same time (91). Another study combined disease signatures extracted from GEO with CMap 1.0 to suggest repurposing opportunities for approved drugs for 100 diseases (92). These studies were inspired by the arrival of network approaches to biomedical research, for example, those that constructed disease–disease and disease–gene networks based on known disease genes from OMIM (Online Mendelian Inheritance in Man) (93), or drug–drug networks based on shared targets (94, 95).

Another effort to create a connectivity mapping resource from GEO is the Gene Perturbation Atlas (GPA). GPA has a collection of manually curated studies from GEO where gene expression data were collected before and after single-gene, microRNA, and long noncoding RNA perturbations for a total of 3,072 signatures (96). Single-gene perturbations followed by genome-wide expression profiling with RNA-seq is also available from the ENCODE (Encyclopedia of DNA Elements) project (97, 98). In another study, Axelsson et al. (99) compiled a library of 3,852 drug signatures from GEO with the goal of prioritizing drugs to lower hepatic glucose production to improve glucose control in type 2 diabetic patients. Similarly, DrugSig contains ~6,000 signatures for ~1,300 drugs extracted manually from GEO (100).

Since GEO does not require strict metadata standards, extracting gene expression signatures from published studies requires the tedious work of manually visiting the page of each study and identifying the control and perturbation samples. In addition, it is desired to label the perturbagen, the cellular or tissue context, other important experimental parameters such as compound concentration, and gene expression profiling time points, as well as the context of the disease and any other biological process under investigation. These metadata elements then need to be linked to ontologies and controlled vocabularies so that they can be further integrated across studies and with other data sets. A microtask crowdsourcing effort was established to extract gene expression signatures for gene perturbations, diseases, and drugs to create the CREEDS (crowd-extracted expression of differential signatures) connectivity map resource (101). To facilitate the extraction of signatures from GEO, participants used a Chrome browser extension that was developed specifically for the project (102). Together, over 75 volunteers from 25 countries extracted 3,879 signatures. Importantly, these signatures were then used as a gold standard to train a text-mining classifier that processed the entire GEO database to automatically and programmatically extract many more signatures. In a similar study, a crowdsourcing approach was used to tag microarray samples to enable a signature search of over 5,798 studies that include 490,110 samples to create the STARGEO (search tag analyze resource for GEO) resource (103). Together, CREEDS and STARGEO underline the need for improved metadata annotations when new transcriptomics data are submitted to major repositories such as GEO. An alternative approach is to use text mining to enrich and structure the metadata of GEO studies and samples. Natural language processing algorithms were implemented by tools such as GEOracle to identify perturbation and control samples from GEO. GEOracle tags samples programmatically but also enables users to manually adjust the automated results through a web interface (104). Similarly, MetaSRA (metadata for the Sequence Read Archive) (105) utilized manual annotations as a training set to tag all the RNA-seq samples and studies within GEO using a text reasoning graph machine learning algorithm.

Both the CREEDS and STARGEO resources only cover signatures from microarray studies. This is mostly because the RNA-seq data in GEO are provided in a raw, unaligned form. Some RNA-seq data hosted on GEO are available in a more processed form, but the methods, reference genome, and the shape of the data are not standardized. Hence, systematic reprocessing of such data is required before signature extraction can commence. The greatest challenge is performing the sequence alignment step, which was, until recently, computationally demanding and thus expensive. However, several projects were initiated to uniformly reprocess the RNA-seq data in GEO (106–112). ARCHS4 [all RNA-seq and ChIP-seq (chromatin immunoprecipitation and sequencing) sample and signature search] used kallisto (113) and an optimized cloud-based platform to align more than 300,000 GEO samples. Toil Recompute (106) demonstrated that by using a standard workflow language, the sequence alignment task can be distributable, reusable, and reproducible. Expression Atlas (107) is a comprehensive resource for uniformly processed RNA-seq and microarray data for multiple species. Overall, such resources facilitate the next step, which is systematic signature identification and extraction from RNA-seq data. Identification and extraction of signatures from RNA-seq studies can be achieved with BioJupies (114), a platform that automatically generates Jupyter notebook reports for signatures created from raw or processed RNA-seq data.

BENCHMARKING SIGNATURE PROCESSING METHODS

Differences in data processing pipelines can significantly influence the quality of the connectivity mapping resource. Connectivity mapping resources provide the unique opportunity to benchmark algorithms, tools, and pipelines. This is because connectivity mapping resources comprise comprehensive collections of data sets that can be converted into comparable signatures. Hence, different pipelines to process signatures can be applied, and then the resultant collection of signatures can be evaluated by how well the signatures cluster based on what is expected. For comparing data processing algorithms, tools, and pipelines, external background knowledge needs to be incorporated as an independent silver standard. For example, to systematically evaluate the quality of transcriptomics connectivity map resources, researchers can use the external background knowledge data set from each of the following resources to project independent knowledge about expected associations between drugs and small molecules: Anatomical Therapeutic Chemical drug classification (115), structural similarity of compounds (116, 117), or side effects profile similarity (118). These drug similarity aspects were already applied to evaluate the quality of tools, algorithms, and pipelines applied to process the CMap 1.0 and CMap 2.0 resources (31, 52, 53, 119). For evaluating the quality of signatures created from genetic perturbations, gene–gene associations such as those from known protein–protein interactions, disease annotations, or coregulation by transcription factors can be used as an external silver standard.

OTHER TYPES OF QUERIES

Connectivity mapping resources enable other creative applications that do not fall into the previously discussed signature query category, where signatures are assessed in a pairwise fashion for similarity. One example is the Drug Set Enrichment Analysis (DSEA) platform (120). To develop DSEA, researchers first computed consensus signatures for each CMap-profiled drug. Then, enrichment scores for each drug signature were created using gene set libraries from pathway databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) (121) and Reactome (122). Given the enrichment scores for each gene set/drug signature pair, drug set libraries could be generated. Hence, DSEA receives a query of drug sets to return ranked lists of pathways that are putatively modulated by the query set of drugs (120). Another unique example is the web and mobile application Drug Gene Budger (DGB) (123). DGB accepts single-gene queries as input and then ranks small molecules to maximally upregulate or downregulate a specific gene based on three connectivity mapping resources: CMap 1.0, CMap 2.0, and small-molecule signatures extracted from GEO. DGB also provides a measure of specificity for each ranked drug based on how many other genes the drug modulates. This tool serves as a first step in discovering drugs that influence the expression of specific genes in a desired direction.

OTHER TYPES OF SIGNATURES

Connectivity mapping principles are generalizable to other assays and readouts that are sufficiently high-dimensional to meaningfully distinguish between cellular states (76). Above we have covered only transcriptomic-based signature resources (**Table 1**), but other types of assays such as proteomics, phosphoproteomics, cellular imaging, cytokine assays, epigenomics, and metabolomics can be used to create connectivity mapping resources. Similarity between perturbations can be detected even with low-dimensional assays; for example, cell viability or reporter assays can be used to collect profiles of viability across many cell types/lines to create a vector that will distinguish perturbations. It is also not unreasonable to assume that certain assays might highlight some cellular responses while insufficiently representing others; hence, signatures for the same perturbagens can be combined across assays to improve mapping. For example, protein and phosphoprotein levels can be quite distinct from transcriptional responses. This is because mRNA levels and their coding protein levels may be discordant, although the degree of this discordance can vary between studies (124, 125). Furthermore, diverse signaling states, as evidenced by phosphoproteomic readout, intersect with a smaller set of chromatin states (126). As mentioned above, significant transcriptional changes can be observed without seeing a change in cell growth or survival phenotypes (62). The use of orthogonal omics assays can help to alleviate assay-specific biases and provide a more holistic picture. Proteomics assays used by the LINCS consortium include antibody probe–based assays: reverse-phase protein array (127) and microwesterns, as well as targeted mass spectrometry–based phosphoproteomic (P100) (128) and global chromatin profiling (GCP) assays (126, 129). The P100 assay, used by LINCS, measures a reduced representation set of 96 cellular phosphopeptides that act as representatives for clusters of coordinately regulated phosphopeptides (128). Similarly, GCP measures global levels of selected modified histones from bulk chromatin, which provides a unique chromatin signature for a cellular state (129). Together, these two assays were applied to construct a compendium of proteomic signatures for 90 drugs applied to six cell lines. These assays were applied to the same small molecules and cell lines profiled with the L1000 assay. Together with the PRISM (profiling relative inhibition simultaneously in mixtures) method, which profiles cell viability (130), and Cell Painting, which profiles changes in cell morphology (131), the CMap team at the Broad Institute is realizing a more holistic next-generation connectivity mapping resource, covering cell response biology across regulatory layers.

When many cellular properties are simultaneously measured in image-based analysis, the aim is not necessarily to capture well-characterized phenotypes such as cell size or density, but rather to detect features that, taken together, can distinguish a variety of cellular responses and act as a morphological fingerprint in response to varied perturbations. Examples of such imaging assays include the microenvironment microarray assay, where cells are grown with combinations of microenvironment-associated proteins on microwell plates and are subsequently imaged for features of metabolism, cell cycle, nuclear activity, and differentiation status (111). The CycIF (cyclic immunofluorescence) assay enables up to 30× imaging of live cells and up to 60× imaging of paraffin-embedded cells (132, 133), and the Cell Painting assay mentioned above is an image-based analysis of individual cells yielding ~1,500 morphological features including size, shape, texture, and intensity, among others (131). The Cell Painting assay was applied to generate a publicly available compendium of morphological profiles in response to 30,616 compounds (76), as well as 220 exogenously expressed genes (134). One of the caveats with nontranscriptomics connectivity mapping resources is that meaningful queries at the data level are difficult to construct. This is in part because the analytes measured are often assay specific, for example, the examination of a specific subset of phosphosites. Such analytes may not translate well between assays and laboratories, and this limitation prohibits broad community adoption.

Table 1 Connectivity mapping transcriptomics resources

Resource	Perturbation type(s)	Readout	Perturbagens	Cells/ Tissues	Signatures	Species	Access	Reference(s)
CMap 2.0	Small molecules, biologics	L1000	20,125	77 (9 core)	473,647	Human	GSE92742; **https://clue.io**	58
	shRNA, cDNA		18,493 shRNA 3,462 cDNA (5,075 genes)					
	CRISPR		1,331	10	18,619		**https://clue.io**	NA
CMap 1.0	Small molecules	Microarray	1,384	4	3,773	Human	GSE5258; **https://clue.io**	8
Carcinogenome Project (CRCGN)	Small molecules	L1000	500	4	5,996	Human	**https://clue.io**	NA
DrugMatrix	Small molecules	Microarray	657	9	3,938	Rat	GSE59927	6, 7
Fish CMap	Small molecules	Microarray	51	24	55	Zebra fish, fathead minnow	GSE38070, GSE60202, GSE70807, GSE70936	140
Gene Perturbation Atlas (GPA)	Genes	Microarray (mined from GEO)	1,585	1,170	3,072	Human, mouse	**http://biocc.hrbmu.edu.cn/GPA**	96
Axelsson et al.	Small molecules	Microarray (mined from GEO and EBI)	2,426	217	3,852	Human, rat	GEO and EBI; Ref. 99, table S2	99
DrugSig	Small molecules	Microarray (mined from GEO)	1,309	NA	5,997	Human	**http://biotechlab.fudan.edu.cn/database/drugsig**	100
Open TG-GATEs	Small molecules	Microarray	170	2	1,483	Human, rat	**https://dbarchive.biosciencedbc.jp**	141
Senkowski et al.	Small molecules	L1000	22	3 (monolayer and tumor spheroids)	1,065 profiles	Human	**http://data.genometry.com**	142
Reis et al.	Photosensitive HDAC inhibitors	L1000	4	1	24	Human	**http://data.genometry.com**	143
Cusanovich et al.	siRNA	Microarray	59	1	59	Human	GSE50588	144
DRUG-seq	Small molecules	RNA-seq	433	1	3,464	Human	GSE120222	85
ENCODE	shRNA	RNA-seq	421	5	668	Human, fly	GEO	98
Hu & Agarwal	Small molecules	Microarray (mined from GEO and EBI)	127	NA	395	Human	Ref. 89, table S1	89
	Diseases		196	NA	395			
CREEDS	Genes	Microarray (crowdsourced from GEO)	871	1,363	2,176	Human, mouse, rat	**http://amp.pharm.mssm.edu/CREEDS**	101
	Small molecules		271	649	875			
	Diseases		333	501	828			

Abbreviations: cDNA, complementary DNA; CREEDS, crowd-extracted expression of differential signatures; CRISPR, clustered regularly interspaced short palindromic repeats; DRUG-seq, digital RNA with perturbation of genes; EBI, European Bioinformatics Institute; ENCODE, Encyclopedia of DNA Elements; GEO, Gene Expression Omnibus; GSE, GEO series number, HDAC, histone deacetylase; NA, not any/not available; RNA-seq, RNA sequencing; shRNA, short hairpin RNA; siRNA, small interfering RNA; TG-GATEs, Toxicogenomics Project's Genomics-Assisted Toxicity Evaluation System.

ETHICAL CHALLENGES AND EXPERIMENTAL PITFALLS

While the connectivity mapping approach brings hope for accelerating drug and target discovery, much criticism and skepticism still exists. The main concerns are focused around two issues: (*a*) Can the effects of drugs and small molecules on human cell lines translate to the effects of those drugs and small molecules in vivo on the entire organism? (*b*) How reliable are the low-cost transcriptomic profiling methods used so far to create connectivity mapping resources? The fact that the L1000 data were shown to improve the predictions of side effects (75) suggests that there is some information that is captured at the transcriptome signature level that is translatable to the entire organism. Indeed, side effect predictions were possible for only half of the approved and marketed drugs. The distribution of the areas under the curve plot of the model's ability to predict side effects was bimodal. This suggests that there are two groups of drugs, those that can and those that cannot be predictive. Unpredictability might be due to system-level effects of drugs, drugs that target brain circuits, and drugs that target specific receptors that are not present in the cell lines profiled. Reproducibility of signatures is also a serious concern. Cell lines of the same type across labs might be different, and slight differences in experimental conditions can induce very different signatures (135). In addition, many small molecules induce very different signatures in different cell lines or when applied in different concentrations. Hence, querying connectivity mapping resources with signatures collected from different systems, for example, from different cell types or even different organisms, may result in spurious associations. The trust in the quality of the L1000 assay is still a major concern for many biologists and pharmacologists. This is mainly because the idea of gene imputation is not easy to convey to biologists, but is also due to quality issues with some of the data that were openly released. In addition, using a connectivity mapping resource also takes away from figuring out the puzzle on one's own. This is the Betty Crocker principle that adding simple ingredients to a cake mix makes people feel better about their involvement in producing the final product. Finally, investigators may not use connectivity mapping resources appropriately, wasting time and effort in experimentally chasing hypotheses that are based on results they obtained from wrongly utilizing connectivity mapping tools. Hence, adaptation to the connectivity mapping approach has some significant psychological and sociological barriers.

Connectivity mapping has some ethical challenges. One concern is that potential patients will use connectivity mapping resources to self-medicate. Rapid advancements in patient tissue profiling with methods such as RNA-seq and DNA-seq and the increasing ability to query connectivity mapping resources online, together with online access to drugs and small molecules, create a perfect storm, enabling patients to obtain prioritized lists of drugs and small molecules that they may try inappropriately. Predictions for drug repurposing opportunities using connectivity mapping resources can also promote false hope and backlash. Connectivity mapping predictions should be vetted carefully by clinicians before advancing to clinical trials. For example, topiramate, an antiepileptic drug known to cause diarrhea as a major side effect, was proposed for the treatment of irritable bowel syndrome (IBD) by a connectivity mapping publication (16). This prediction raised some concerns by a physician who submitted a blog post comment about the article stating that since the drug is known to cause diarrhea as a side effect, it would be unwise to give it to patients with IBD. It was later demonstrated that topiramate is not efficacious in reducing IBD flares in humans (136).

SUMMARY

A large portion of biomedical research, and in particular bioinformatics, systems biology, and systems pharmacology, involves illuminating the connections between (*a*) genes, proteins, mRNAs, metabolites, and molecular complexes; (*b*) cells, tissues, and organs; (*c*) drugs, small molecules,

Table 2 Connectivity mapping search engines

Name	Purpose	Category	Resource	Link	Reference(s)
CLUE	Query and analyze CMap 2.0 perturbational data sets	Search engine, data portal	L1000, P100, and GCP assays	**https://clue.io**	8, 58
L1000CDS^2	Search for MAL1000 small-molecule perturbations	Search engine	L1000 drug perturbations	**http://amp.pharm.mssm.edu/L1000CDS2**	52
L1000FWD	Visualize similarities between perturbation signatures	Search engine, visualization	L1000 drug perturbations	**http://L1000FWD.net**	65
Drug Gene Budger	Search for drugs to up- or downregulate a gene of interest	Search engine	L1000 drug perturbations	**http://DGB.cloud**	123
LINCS Data Portal	Download LINCS perturbation data sets	Data portal	LINCS data sets	**http://lincsportal.ccs.miami.edu/dcic-portal**	145
SEP-L1000	Visualize and search for predicted adverse drug reactions	Search engine, visualization	L1000 drug perturbations	**http://maayanlab.net/SEP-L1000**	75
Drug Pair Seeker	Search pairs of drug perturbation gene signatures	Search engine	CMap 1.0 microarrays	**http://www.maayanlab.net/DPS**	34
CREEDS	Query perturbation gene signatures mined from GEO by the crowd	Search engine, data portal	GEO drug, disease, and gene perturbations	**http://amp.pharm.mssm.edu/CREEDS**	101, 146
iLINCS	Analyze and query LINCS signatures	Search engine, data portal	LINCS signatures	**http://www.ilincs.org**	NA
QUADrATiC	Search signatures from FDA-approved compounds	Search engine	L1000 drug perturbations	**http://go.qub.ac.uk/quadratic**	147
LINCS Canvas Browser	Query, visualize, and perform enrichment analysis with LINCS L1000 data	Search engine, visualization	LINCS perturbation signatures	**http://www.maayanlab.net/LINCS/LCB**	148
GEN3VA	Aggregate and analyze gene expression signatures crowdsourced from GEO	Search engine, analysis, visualization	Gene, drug, and disease signatures mined from GEO	**http://amp.pharm.mssm.edu/gen3va**	149
DSEA	Enrichment analysis over drug sets derived from drug-induced gene expression signatures	Enrichment analysis tool	CMap 1.0 drug perturbation microarrays	**http://dsea.tigem.it**	120
DrugSig	Drug- and target-based repositioning search tool	Search engine, data portal	Drug perturbations mined from GEO	**http://biotechlab.fudan.edu.cn/database/drugsig**	100
openSESAME	Query thousands of profiles extracted from GEO	Search engine	Gene expression from GEO	**http://opensesame.bu.edu**	50
Transcriptomine	Query drug and gene perturbation gene signatures related to nuclear receptor signaling	Search engine, visualization	Signatures compiled from publicly available data sets	**https://www.nursa.org/nursa/transcriptomine/index.jsf**	150
STARGEO	Crowdsource annotation for GEO samples to enable computation of disease perturbation signatures	Annotation	GEO gene expression data sets	**http://stargeo.org**	103

Abbreviations: CDS, characteristic direction signature; CREEDS, crowd-extracted expression of differential signatures; DSEA, drug set enrichment analysis; FDA, Food and Drug Administration; FWD, fireworks display; GCP, global chromatin profiling; GEN3VA, gene expression and enrichment vector analyzer; GEO, Gene Expression Omnibus; LINCS, Library of Integrated Network-Based Cellular Signatures; NA, not any; QUADrATiC, Queen's University Belfast accelerated drug and transcriptome connectivity; SEP, side effect prediction; STARGEO, search tag analyze resource for GEO.

antibodies, biologics, and endogenous ligands; (*d*) organismal phenotypes, diseases, and side effects; and (*e*) pathways, gene modules, and cellular compartments. Connectivity mapping resources span all possible associations between these five groups of abstract entities. Some connections between these entities have been widely studied, while others remain for future exploration.

This review does not cover all connectivity mapping efforts and resources. Many methodologies, resources, and applications are likely missing due to the sheer volume and scope of the approach. There are fuzzy boundaries to a community-accepted definition of what is a signature. Therefore, it is subjective to determine what constitutes a connectivity mapping resource. An important part of constructing a connectivity mapping resource is the development of intuitive web-based software tools and APIs (application programming interfaces), which provide access and usability of the resource for querying. For this review, we assembled a collection of such software tools and platforms (**Table 2**).

Looking forward, it is expected that in the coming years additional connectivity mapping efforts will emerge. These efforts will augment, or even replace, the more common targeted approaches that have dominated biomedical research and drug discovery in the past four to five decades. While currently most comprehensive gene expression signatures collected uniformly for the creation of a connectivity mapping resource employ the L1000 assay (58), or a variation of it (79), the reduction in cost for RNA-seq and recent efforts to multiplex samples for creating low-cost versions of RNA-seq (82–85) suggest that deep sequencing and imaging technologies will likely dominate the new generation of connectivity mapping efforts in the coming decade.

Another consideration for future connectivity mapping applications is the expected increase in size of signature compendia. It is expected that in the coming years hundreds of millions of signatures will become available. Efficient storage, retrieval, and real-time search over this many signatures will be required.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work is supported by NIH grants U54-HL127624 (LINCS-DCIC), U24-CA224260 (IDG-KMC), and OT3-OD025467 (NIH Data Commons) to A.M. A.B.K. is partially supported by T32-GM062754 (Integrated Pharmacological Sciences Training Program).

LITERATURE CITED

1. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102:109–26
2. Stoughton RB, Friend SH. 2005. How molecular profiling could revolutionize drug discovery. *Nat. Rev. Drug Discov.* 4:345–50
3. Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP. 2003. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *PNAS* 100:9608–13
4. Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, et al. 2001. Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol. Appl. Pharmacol.* 175:28–42
5. Steiner G, Suter L, Boess F, Gasser R, de Vera MC, et al. 2004. Discriminating different classes of toxicants by transcript profiling. *Environ. Health Perspect.* 112:1236–48
6. Engelberg A. 2004. Iconix Pharmaceuticals, Inc.—removing barriers to efficient drug discovery through chemogenomics. *Pharmacogenomics* 5:741–44

7. Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, et al. 2005. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* 119:219–44
8. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313:1929–35
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102:15545–50
10. Liu J, Lee J, Salazar Hernandez MA, Mazitschek R, Ozcan U. 2015. Treatment of obesity with celastrol. *Cell* 161:999–1011
11. Raghavan R, Hyter S, Pathak HB, Godwin AK, Konecny G, et al. 2016. Drug discovery using clinical outcome-based Connectivity Mapping: application to ovarian cancer. *BMC Genom*. 17:811
12. Bhat-Nakshatri P, Goswami CP, Badve S, Sledge GW Jr., Nakshatri H. 2013. Identification of FDA-approved drugs targeting breast cancer stem cells along with biomarkers of sensitivity. *Sci. Rep.* 3:2530
13. Josset L, Textoris J, Loriod B, Ferraris O, Moules V, et al. 2010. Gene expression signature-based screening identifies new broadly effective influenza A antivirals. *PLOS ONE* 5:e13169
14. Vanderstocken G, Dvorkin-Gheva A, Shen P, Brandsma CA, Obeidat M, et al. 2018. Identification of drug candidates to suppress cigarette smoke-induced inflammation via Connectivity Map analyses. *Am. J. Respir. Cell Mol. Biol.* 58:727–35
15. Claerhout S, Lim JY, Choi W, Park YY, Kim K, et al. 2011. Gene expression signature analysis identifies vorinostat as a candidate therapy for gastric cancer. *PLOS ONE* 6:e24662
16. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, et al. 2011. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 3:96ra76
17. Brum AM, van de Peppel J, Nguyen L, Aliev A, Schreuders-Koedam M, et al. 2018. Using the Connectivity Map to discover compounds influencing human osteoblast differentiation. *J. Cell Physiol.* 233:4895–906
18. Wang G, Ye Y, Yang X, Liao H, Zhao C, Liang S. 2011. Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma. *PLOS ONE* 6:e14573
19. Xu S, Liu R, Da Y. 2018. Comparison of tumor related signaling pathways with known compounds to determine potential agents for lung adenocarcinoma. *Thorac. Cancer* 9:974–88
20. Dyle MC, Ebert SM, Cook DP, Kunkel SD, Fox DK, et al. 2014. Systems-based discovery of tomatidine as a natural small molecule inhibitor of skeletal muscle atrophy. *J. Biol. Chem.* 289:14913–24
21. Zerbini LF, Bhasin MK, de Vasconcellos JF, Paccez JD, Gu X, et al. 2014. Computational repositioning and preclinical validation of pentamidine for renal cell cancer. *Mol. Cancer Ther.* 13:1929–41
22. Smalley JL, Gant TW, Zhang SD. 2010. Application of connectivity mapping in predictive toxicology based on gene-expression similarity. *Toxicology* 268:143–46
23. Brum AM, van de Peppel J, van der Leije CS, Schreuders-Koedam M, Eijken M, et al. 2015. Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway. *PNAS* 112:12711–16
24. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, et al. 2018. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 173:338–54.e15
25. Jun HY, Kim TH, Choi JW, Lee YH, Lee KK, Yoon KH. 2017. Evaluation of connectivity map-discovered celastrol as a radiosensitizing agent in a murine lung carcinoma model: feasibility study of diffusion-weighted magnetic resonance imaging. *PLOS ONE* 12:e0178204
26. Greenhill C. 2015. Celastrol identified as a leptin sensitizer and potential novel treatment for obesity. *Nat. Rev. Endocrinol.* 11:444
27. Chong CR, Sullivan DJ Jr. 2007. New uses for old drugs. *Nature* 448:645–46
28. Slonim DK, Koide K, Johnson KL, Tantravahi U, Cowan JM, et al. 2009. Functional genomic analysis of amniotic fluid cell-free mRNA suggests that oxidative stress is significant in Down syndrome fetuses. *PNAS* 106:9425–29
29. Flynn C, Zheng S, Yan L, Hedges L, Womack B, et al. 2012. Connectivity map analysis of nonsense-mediated decay-positive BMPR2-related hereditary pulmonary arterial hypertension provides insights into disease penetrance. *Am. J. Respir. Cell Mol. Biol.* 47:20–27

30. Toscano MG, Navarro-Montero O, Ayllon V, Ramos-Mejia V, Guerrero-Carreno X, et al. 2015. SCL/TAL1-mediated transcriptional network enhances megakaryocytic specification of human embryonic stem cells. *Mol. Ther.* 23:158–70

31. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, et al. 2010. Discovery of drug mode of action and drug repositioning from transcriptional responses. *PNAS* 107:14621–26

32. Wang K, Sun J, Zhou S, Wan C, Qin S, et al. 2013. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. *PLOS Comput. Biol.* 9:e1003315

33. Isik Z, Baldow C, Cannistraci CV, Schroeder M. 2015. Drug target prioritization by perturbed gene expression and network information. *Sci. Rep.* 5:17417

34. Zhong Y, Chen EY, Liu R, Chuang PY, Mallipattu SK, et al. 2013. Renoprotective effect of combined inhibition of angiotensin-converting enzyme and histone deacetylase. *J. Am. Soc. Nephrol.* 24:801–11

35. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11:733–39

36. Clark NR, Ma'ayan A. 2011. Introduction to statistical methods to analyze large data sets: principal components analysis. *Sci. Signal.* 4:tr3

37. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579–605

38. Iskar M, Campillos M, Kuhn M, Jensen LJ, van Noort V, Bork P. 2010. Drug-induced regulation of target expression. *PLOS Comput. Biol.* 6:e1000925

39. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550

40. Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29

41. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet.* 3:1724–35

42. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28:882–83

43. Benito M, Parker J, Du Q, Wu J, Xiang D, et al. 2004. Adjustment of systematic microarray data biases. *Bioinformatics* 20:105–14

44. Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 97:10101–6

45. Nygaard V, Rodland EA, Hovig E. 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17:29–39

46. Zhang SD, Gant TW. 2008. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinform.* 9:258

47. Zhang SD, Gant TW. 2009. sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinform.* 10:236

48. McArt DG, Bankhead P, Dunne PD, Salto-Tellez M, Hamilton P, Zhang SD. 2013. cudaMap: a GPU accelerated program for gene expression connectivity mapping. *BMC Bioinform.* 14:305

49. Tenenbaum JD, Walker MG, Utz PJ, Butte AJ. 2008. Expression-based Pathway Signature Analysis (EPSA): mining publicly available microarray data for insight into human disease. *BMC Med. Genom.* 1:51

50. Gower AC, Spira A, Lenburg ME. 2011. Discovering biological connections between experimental conditions based on common patterns of differential gene expression. *BMC Bioinform.* 12:381

51. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, et al. 2014. The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinform.* 15:79

52. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, et al. 2016. L1000CDS^2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst. Biol. Appl.* 2:16015

53. Cheng J, Yang L, Kumar V, Agarwal P. 2014. Systematic evaluation of connectivity map for disease indications. *Genome Med.* 6:540

54. Iorio F, Tagliaferri R, di Bernardo D. 2009. Identifying network of drug mode of action by gene expression profiling. *J. Comput. Biol.* 16:241–51

55. Lee J-Y, Fujimoto GM, Wilson R, Wiley HS, Payne SH. 2018. Blazing Signature Filter: a library for fast pairwise similarity comparisons. *BMC Bioinform*. 19:221

56. Musa A, Ghoraie LS, Zhang S-D, Glazko G, Yli-Harja O, et al. 2017. A review of connectivity map and computational approaches in pharmacogenomics. *Brief. Bioinform.* 19:506–23

57. Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, et al. 2017. The Library of Integrated Network-Based Cellular Signatures NIH Program: system-level cataloging of human cells response to perturbations. *Cell Syst*. 6:13–24

58. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, et al. 2017. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171:1437–52.e17

59. Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 30:207–10

60. Lakhani K, Garvin D, Lonstein E. 2010. *TopCoder (A): developing software through crowdsourcing*. Harvard Bus. Sch. Case 610-032, Cambridge, MA. **https://www.hbs.edu/faculty/Pages/item.aspx?num=38356**

61. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. 2016. Gene expression inference with deep learning. *Bioinformatics* 32:1832–39

62. Niepel M, Hafner M, Duan Q, Wang Z, Paull EO, et al. 2017. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nat. Commun*. 8:1186

63. Hsu YC, Chiu YC, Chen Y, Hsiao TH, Chuang EY. 2016. A simple gene set-based method accurately predicts the synergy of drug pairs. *BMC Syst. Biol.* 10(Suppl. 3):66

64. Hassane DC, Sen S, Minhajuddin M, Rossi RM, Corbett CA, et al. 2010. Chemical genomic screening reveals synergism between parthenolide and inhibitors of the PI-3 kinase and mTOR pathways. *Blood* 116:5983–90

65. Wang Z, Lachmann A, Keenan AB, Ma'ayan A, Stegle O. 2018. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* 34:2150–52

66. Liu T-P, Hsieh Y-Y, Chou C-J, Yang P-M. 2018. Systematic polypharmacology and drug repurposing via an integrated L1000-based Connectivity Map database mining. *R. Soc. Open Sci.* 5:181321

67. Han H-W, Hahn S, Jeong HY, Jee J-H, Nam M-O, et al. 2018. LINCS L1000 dataset-based repositioning of CGP-60474 as a highly potent anti-endotoxemic agent. *Sci. Rep.* 8:14969

68. Wang Y, Arora K, Yang F, Shin W-H, Chen J, et al. 2018. PP-2, a src-kinase inhibitor, is a potential corrector for F508del-CFTR in cystic fibrosis. bioRxiv 288324. **https://doi.org/10.1101/288324**

69. Fagone P, Caltabiano R, Russo A, Lupo G, Anfuso CD, et al. 2017. Identification of novel chemotherapeutic strategies for metastatic uveal melanoma. *Sci. Rep.* 7:44564

70. Er JL, Goh PN, Lee CY, Tan YJ, Hii L-W, et al. 2018. Identification of inhibitors synergizing gemcitabine sensitivity in the squamous subtype of pancreatic ductal adenocarcinoma (PDAC). *Apoptosis* 23:343–55

71. Lachmann A, Giorgi FM, Alvarez MJ, Califano A. 2016. Detection and removal of spatial bias in multiwell assays. *Bioinformatics* 32:1959–65

72. Young WC, Raftery AE, Yeung KY. 2017. Model-based clustering with data correction for removing artifacts in gene expression data. *Ann. Appl. Stat.* 11:1998–2026

73. Hodos R, Zhang P, Lee H-C, Duan Q, Wang Z, et al. 2017. Cell-specific prediction and application of drug-induced gene expression profiles. *Pac. Symp. Biocomput.* 23:32–43

74. Xiao J, Blatti C, Sinha S. 2018. SigMat: a classification scheme for gene signature matching. *Bioinformatics* 34:i547–54

75. Wang Z, Clark NR, Ma'ayan A. 2016. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 32:2338–45

76. Bray M-A, Gustafsdottir SM, Rohban MH, Singh S, Ljosa V, et al. 2017. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *GigaScience* 6:giw014

77. Nagiec MM, Skepner AP, Negri J, Eichhorn M, Kuperwasser N, et al. 2015. Modulators of hepatic lipoprotein metabolism identified in a search for small-molecule inducers of tribbles pseudokinase 1 expression. *PLOS ONE* 10:e0120295

78. De Wolf H, Cougnaud L, Van Hoorde K, De Bondt A, Wegner JK, et al. 2018. High-throughput gene expression profiles to define drug similarity and predict compound activity. *Assay Drug Dev. Technol.* 16:162–76

79. Mav D, Shah RR, Howard BE, Auerbach SS, Bushel PR, et al. 2018. A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. *PLOS ONE* 13:e0191105

80. Liberzon A. 2014. A description of the Molecular Signatures Database (MSigDB) website. In *Stem Cell Transcriptional Networks*, ed. BL Kidder, pp. 153–60. New York: Springer

81. Kleensang A, Maertens A, Rosenberg M, Fitzpatrick S, Lamb J, et al. 2014. t^4 workshop report: pathways of toxicity. *Altex* 31:53

82. Li H, Qiu J, Fu XD. 2012. RASL-seq for massively parallel and quantitative analysis of gene expression. *Curr. Protoc. Mol. Biol.* 98:4.13.1–4.13.9

83. Yeakley JM, Shepard PJ, Goyena DE, VanSteenhouse HC, McComb JD, Seligmann BE. 2017. A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling. *PLOS ONE* 12:e0178302

84. Bush EC, Ray F, Alvarez MJ, Realubit R, Li H, et al. 2017. PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nat. Commun.* 8:105

85. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, et al. 2018. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun.* 9:4307

86. Bushel PR, Paules RS, Auerbach SS. 2018. A comparison of the TempO-Seq S1500+ platform to RNA-Seq and microarray using rat liver mode of action samples. *Front. Genet.* 9:485

87. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*. 31:68–71

88. Yi Y, Li C, Miller C, George AL Jr. 2007. Strategy for encoding and comparison of gene expression signatures. *Genome Biol*. 8:R133

89. Hu G, Agarwal P. 2009. Human disease-drug network based on genomic expression profiles. *PLOS ONE* 4:e6536

90. Huang H, Liu CC, Zhou XJ. 2010. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *PNAS* 107:6823–28

91. Liu CC, Hu J, Kalakrishnan M, Huang H, Zhou XJ. 2009. Integrative disease classification based on cross-platform microarray data. *BMC Bioinform*. 10(Suppl. 1):S25

92. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, et al. 2011. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77

93. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. 2007. The human disease network. *PNAS* 104:8685–90

94. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. 2007. Drug−target network. *Nat. Biotechnol.* 25:1119–26

95. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R. 2007. Network analysis of FDA approved drugs and their targets. *Mt. Sinai J. Med.* 74:27–32

96. Xiao Y, Gong Y, Lv Y, Lan Y, Hu J, et al. 2015. Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Sci. Rep.* 5:10889

97. ENCODE Proj. Consort. 2004. The ENCODE (encyclopedia of DNA elements) project. *Science* 306:636–40

98. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, et al. 2017. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 46:D794–801

99. Axelsson AS, Tubbs E, Mecham B, Chacko S, Nenonen HA, et al. 2017. Sulforaphane reduces hepatic glucose production and improves glucose control in patients with type 2 diabetes. *Sci. Transl. Med.* 9:eaah4477

100. Wu H, Huang J, Zhong Y, Huang Q. 2017. DrugSig: a resource for computational drug repositioning utilizing gene expression signatures. *PLOS ONE* 12:e0177743

101. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, et al. 2016. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* 7:12846

102. Gundersen GW, Jones MR, Rouillard AD, Kou Y, Monteiro CD, et al. 2015. GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions. *Bioinformatics* 31:3060–62

103. Hadley D, Pan J, El-Sayed O, Aljabban J, Aljabban I, et al. 2017. Precision annotation of digital samples in NCBI's gene expression omnibus. *Sci. Data* 4:170125

104. Djordjevic D, Tang JYS, Chen YX, Shannon SL, Ling RWK, et al. 2019. Discovery of pertubation gene targets via free text metadata mining in Gene Expression Omnibus. *Comput. Biol. Chem.* 80:152–58

105. Bernstein MN, Doan A, Dewey CN. 2017. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics* 33:2914–23

106. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, et al. 2017. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* 35:314–16

107. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, et al. 2015. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* 44:D746–52

108. Fonseca NA, Petryszak R, Marioni J, Brazma A. 2014. iRAP-an integrated RNA-seq analysis pipeline. bioRxiv 005991. **https://doi.org/10.1101/005991**

109. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, et al. 2017. Reproducible RNA-seq analysis using *recount2*. *Nat. Biotechnol.* 35:319–21

110. Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, et al. 2017. Enabling cross-study analysis of RNA-sequencing data. bioRxiv 110734. **https://doi.org/10.1101/110734**

111. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, et al. 2018. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* 9:1366

112. Al Mahi N, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. 2018. GREIN: an interactive web platform for re-analyzing GEO RNA-seq data. bioRxiv 326223. **https://doi.org/10.1101/326223**

113. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34:525–27

114. Torre D, Lachmann A, Ma'ayan A. 2018. BioJupies: automated generation of interactive notebooks for RNA-seq data analysis in the cloud. *Cell Syst.* 7:556–61

115. World Health Organ. (WHO). 2019. *Anatomical Therapeutic Chemical (ATC) Classification Index with Defined Daily Doses (DDDs)*. Oslo, Nor.: WHO Collab. Centre Drug Stat. Method.

116. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, et al. 2013. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42:D1091–97

117. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, et al. 2007. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36:D351–59

118. Food Drug Admin. 2015. *FDA adverse event reporting system (FAERS)*. Public Database, U.S. Food Drug Admin., Silver Spring, MD

119. Iorio F, Shrestha RL, Levin N, Boilot V, Garnett MJ, et al. 2015. A semi-supervised approach for refining transcriptional signatures of drug response and repositioning predictions. *PLOS ONE* 10:e0139446

120. Napolitano F, Sirci F, Carrella D, di Bernardo D. 2016. Drug-set enrichment analysis: a novel tool to investigate drug mode of action. *Bioinformatics* 32:235–41

121. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28:27–30

122. Joshi-Tope G. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33:D428–32

123. Wang Z, He E, Sani K, Jagodnik KM, Silverstein M, Ma'ayan A. 2018. Drug Gene Budger (DGB): an application for ranking drugs to modulate a specific gene based on transcriptomic signatures. *Bioinformatics.* In press

124. Li J, Zhao W, Akbani R, Liu W, Ju Z, et al. 2017. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell* 31:225–39

125. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, et al. 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534:55–62

126. Litichevskiy L, Peckner R, Abelin JG, Asiedu JK, Creech AL, et al. 2018. A library of phosphoproteomic and chromatin signatures for characterizing cellular responses to drug perturbations. *Cell Syst.* 6:424–43.e7

127. Koch RJ, Barrette AM, Stern AD, Hu B, Bouhaddou M, et al. 2018. Validating antibodies for quantitative western blot measurements with microwestern array. *Sci. Rep.* 8:11329

128. Abelin JG, Patel J, Lu X, Feeney CM, Fagbami L, et al. 2016. Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. *Mol. Cell Proteom.* 15:1622–41

129. Creech AL, Taylor JE, Maier VK, Wu X, Feeney CM, et al. 2015. Building the Connectivity Map of epigenetics: chromatin profiling by quantitative targeted mass spectrometry. *Methods* 72:57–64

130. Yu C, Mannan AM, Yvone GM, Ross KN, Zhang Y-L, et al. 2016. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat. Biotechnol.* 34:419–23

131. Bray MA, Singh S, Han H, Davis CT, Borgeson B, et al. 2016. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* 11:1757–74

132. Lin JR, Fallahi-Sichani M, Chen JY, Sorger PK. 2016. Cyclic Immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. *Curr. Protoc. Chem. Biol.* 8:251–64

133. Lin JR, Izar B, Wang S, Yapp C, Mei S, et al. 2018. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife* 7:e31657

134. Rohban MH, Singh S, Wu X, Berthet JB, Bray MA, et al. 2017. Systematic morphological profiling of human gene and allele function via Cell Painting. *eLife* 6:e24060

135. Niepel M, Hafner M, Mills CE, Subramanian K, Williams EH, et al. 2019. A multi-center study on factors influencing the reproducibility of in vitro drug-response studies. bioRxiv 213553. **https://doi.org/10.1101/213553**

136. Crockett SD, Schectman R, Stürmer T, Kappelman MD. 2014. Topiramate use does not reduce flares of inflammatory bowel disease. *Dig. Dis. Sci.* 59:1535–43

137. OpenStax. 2018. *Biology*. Houston, TX: OpenStax CNX

138. Drenberg CD, Buaboonnam J, Orwick SJ, Hu S, Li L, et al. 2016. Evaluation of artemisinins for the treatment of acute myeloid leukemia. *Cancer Chemother. Pharmacol.* 77:1231–43

139. Gruber L, Abdelfatah S, Frohlich T, Reiter C, Klein V, et al. 2018. Treatment of multidrug-resistant leukemia cells by novel artemisinin-, egonol-, and thymoquinone-derived hybrid compounds. *Molecules* 23:841

140. Wang RL, Biales AD, Garcia-Reyero N, Perkins EJ, Villeneuve DL, et al. 2016. Fish connectivity mapping: linking chemical stressors by their mechanisms of action-driven transcriptomic profiles. *BMC Genom.* 17:84

141. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, et al. 2015. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43:D921–27

142. Senkowski W, Jarvius M, Rubin J, Lengqvist J, Gustafsson MG, et al. 2016. Large-scale gene expression profiling platform for identification of context-dependent drug responses in multicellular tumor spheroids. *Cell Chem. Biol.* 23:1428–38

143. Reis SA, Ghosh B, Hendricks JA, Szantai-Kis DM, Tork L, et al. 2016. Light-controlled modulation of gene expression by chemical optoepigenetic probes. *Nat. Chem. Biol.* 12:317–23

144. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. 2014. The functional consequences of variation in transcription factor binding. *PLOS Genet.* 10:e1004226

145. Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, et al. 2018. 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:D558–66

146. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44:W90–97

147. O'Reilly PG, Wen Q, Bankhead P, Dunne PD, McArt DG, et al. 2016. QUADrATiC: scalable gene expression connectivity mapping for repurposing FDA-approved therapeutics. *BMC Bioinform.* 17:198

148. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, et al. 2014. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res*. 42:W449–60

149. Gundersen GW, Jagodnik KM, Woodland H, Fernandez NF, Sani K, et al. 2016. GEN3VA: aggregation and analysis of gene expression signatures from related studies. *BMC Bioinform*. 17:461

150. Becnel LB, Ochsner SA, Darlington YF, McOwiti A, Kankanamge WH, et al. 2017. Discovering relationships between nuclear receptor signaling pathways, genes, and tissues in Transcriptomine. *Sci. Signal.* 10:eaah6275

Annual Review of Biomedical Data Science

Volume 2, 2019

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci