Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINCS

Marcin Pilarczyk [1,6*], Mehdi Fazel Najafabadi [1,6*], Michal Kouril [2,6*], Juozas Vasiliauskas [1,6], Wen Niu [1,6], Behrouz Shamsaei [1,6], Naim Mahi [1,6], Lixia Zhang [1,6] Nicholas Clark [1,6], Yan Ren[1,6], Shana White [1,6], Rashid Karim [1,5,6], Huan Xu[1,6], Jacek Biesiada[1], Mark F. Bennet[1,6], Sarah Davidson[1], John F Reichard [1,7], Vasileios Stathias[3,6], Amar Koleti[3,6], Dusica Vidovic[3,6], Daniel J.B. Clark[4,6], Stephan Schurer[3,6], Avi Ma'ayan[4,6], Jarek Meller [1,2,6], Mario Medvedovic [1,6,§]

[1] Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, Cincinnati, OH 45220

[2] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

[3] Department of Molecular and Cellular Pharmacology, Miller School of Medicine and Center for Computational Science, University of Miami, Miami, FL, USA

[4] Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[5] Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, OH 45220

[6] LINCS Data Coordination and Integration Center

[7] Division of Industrial Hygiene, Department of Environmental Health, University of Cincinnati, Cincinnati, OH 45220

[*] Contributed equally, ordered arbitrarily

[§] Corresponding: medvedm@ucmail.uc.edu

Abstract

iLINCS (http://ilincs.org) is an integrative web-based platform for analysis of omics data and signatures of cellular perturbations. The portal facilitates analysis of user-submitted omics signatures of diseases and cellular perturbations in the context of a large compendium of pre-computed signatures (>200,000), as well as mining and re-analysis of the large collection of omics datasets (>10,000), pre-computed signatures and their connections. Analytics workflows driven by user-friendly interfaces enable users with only conceptual understanding of the analysis strategy to execute sophisticated analyses of omics signatures, such as systems biology analysis and interpretation of signatures, mechanism of action analysis and signature-driven drug re-positioning. iLINCS workflows integrate a range of analytics and interactive visualization tools into a comprehensive platform for analysis of omics signatures. There are only few platforms that integrate multiple omics data types, bioinformatics tools, and interfaces for integrative analyses and visualization that do not require any computer programming skills. Among them, iLINCS is unique in terms of the scope and versatility of the data it provides and the analytics it facilitates.

Keywords: LINCS program, omics data, data analytics, connectivity map, cellular signatures, transcriptomics, proteomics, RNA-seq, P100, GCP, systems biology

Background

Transcriptomics and proteomics (omics) signatures in response to cellular perturbations consist of changes in gene or protein expression levels after the perturbation. An omics signature is a high-dimensional readout of cellular state change that provides information about the biological processes affected by the perturbation which underlie the post-perturbation phenotype of the cell. The signature in itself also provides information, although not always directly discernable, about the molecular mechanisms by which the perturbation causes observed changes. If we consider a disease to be a perturbation of the homeostatic biological system under normal physiology, then the omics signature of a disease are the differences in gene/protein expression levels between disease and non-diseased tissue samples.

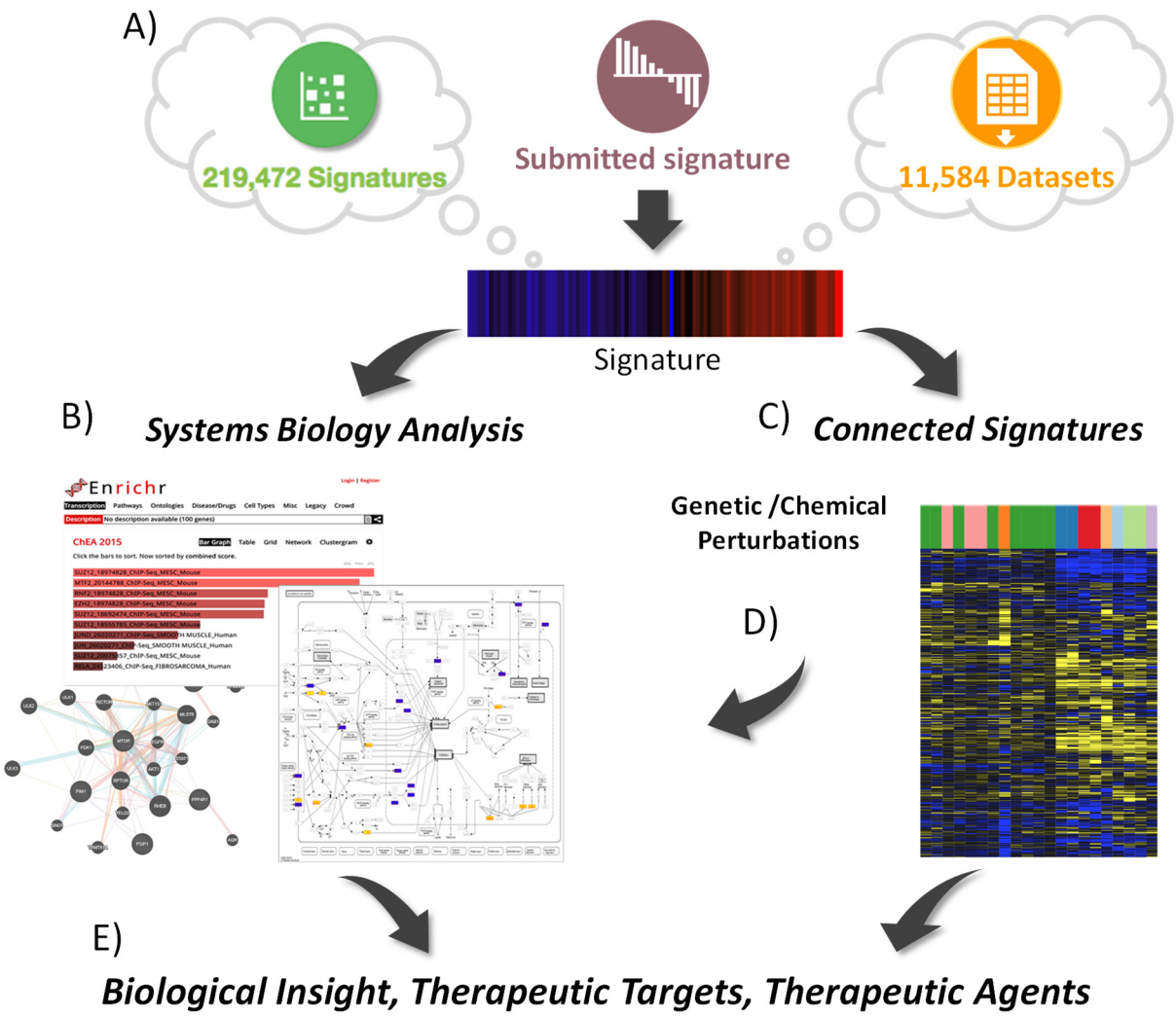
The low cost and effectiveness of transcriptomics assays[1-4] resulted in abundance of transcriptomics datasets and signatures. Beside transcriptomics, recent advances in high throughput proteomics made generation of large numbers of proteomics signatures a reality[5,6]. Several recent efforts were directed at systematic generation of omics signatures of cellular perturbations[7] and generating libraries of signatures by re-analyzing public domain omics datasets[8,9]. The recently released library of integrated network-based cellular signatures (LINCS)[7] L1000 dataset generated transcriptomic signatures at unprecedented scale[2]. The availability of resulting libraries of signatures open exciting new avenues for learning about the mechanisms of diseases and the search for effective therapeutics[10].

The analysis and interpretation of transcriptomic signatures has been intensely researched. Numerous methods and tools have been developed for identifying changes in molecular phenotypes implicated by transcriptional signatures based on gene set enrichment, pathway and network analyses approaches[11-13]. Matching directly transcriptional signatures of a disease with negatively correlated transcriptional signatures of chemical perturbations (CP) underlies the *Connectivity Map* (CMap) approach to identifying potential drug candidates[10,14,15]. Similarly,

correlating signatures of chemical perturbagens with genetic perturbations of specific genes has been used to identify putative targets of drugs and other chemical perturbagens [2].

To fully exploit the information contained within omics signature libraries and within countless omics signatures generated every day by investigators around the world, new user-friendly integrative tools are needed that bring this data together, and are accessible to a large segment

Fig 1. Integrative omics signature analysis in iLINCS. A) A signature can be selected by querying the iILINCS database, submitted by the user, or constructed by analyzing an iLINCS omics dataset. B) The signature can be analyzed using a range of systems biology methods (gene set enrichment, pathway and network analyses). C) Signature "connectivity" analyses can be applied to identify cellular perturbations and biological states with similar (ie connected) signatures. D) The analysis of connected signatures, as well as the identity of the perturbed genes and proteins leading to the connected signatures, can be used to elucidate mechanisms of action. E) Ultimately, the results of the analyses lead to insights about the signature, and can implicate therapeutic targets and putative therapeutic agents.



of biomedical research community. The integrative LINCS (iLINCS) portal brings together libraries of precomputed signatures, formatted datasets, connections between signatures, and integrates them with a bioinformatics analysis engine into a coherent system for omics signature analysis.

Results

iLINCS (available http://ilincs.org) is an integrative user-friendly web platform for the analysis of omics (transcriptomic and proteomic) datasets and signatures of cellular perturbations. The key components of iLINCS are: *Interactive and interconnected analytical workflows for creation and analysis of omics signatures; The large collection of datasets*, pre-computed *signatures* and their *connections; and User-friendly graphical user interfaces for executing analytical tasks and workflows.* The central concept in iLINCS is the omics signature which can be retrieved from the pre-computed signature libraries within the iLINCS database, submitted by the user, or constructed using one of the iLINCS datasets (Fig 1A). The signatures in iLINCS consist of differential gene or protein expression levels and associated p-values between perturbed and baseline samples for all, or any subset of measured genes/proteins. User submitted signatures can also be in the form of a list of genes/proteins, or a list of up- and down-regulated genes/proteins. Analytical workflows facilitate systems biology interpretation of the signature (Fig 1B) and the connectivity analysis of the signature with all iLINCS pre-computed signatures (Fig 1C). Connected signatures can further be analyzed in terms of the patterns of gene/protein expression level changes that underlie the connectivity with the query signature, or through the analysis of gene/protein targets of connected perturbagens (Fig 1D). Ultimately, the multi-layered systems biology analyses, and the connectivity analyses lead to biological insights, and identification of therapeutic targets and putative therapeutic agents (Fig 1E). Below we provide an overview of the key data and analytic components of iLINCS, and then we present three use cases to demonstrate iLINCS' capacity to generate impactful results.

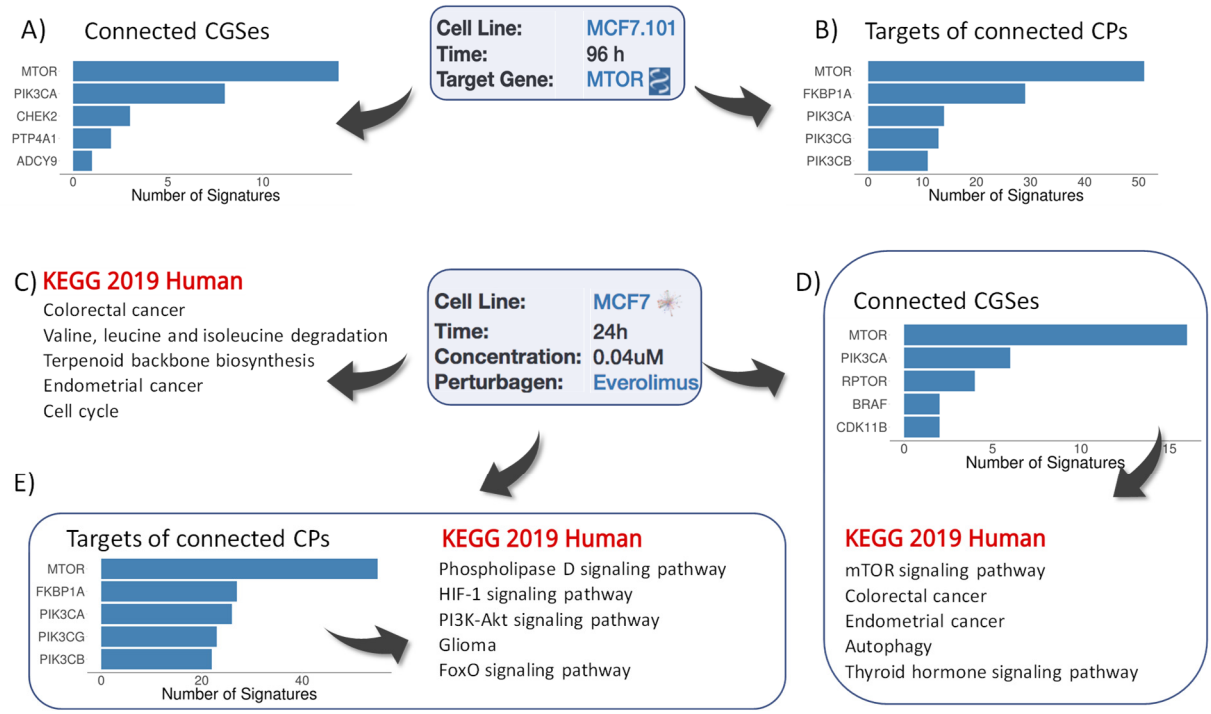
Interconnected workflows for constructing and analyzing omics signatures

Interactive analytical workflows in iLINCS facilitate signature construction through differential expression analysis as well as clustering, dimensionality reduction, functional enrichment, signature connectivity analysis, pathway and network analysis, and integrative interactive visualization. Visualizations include interactive scatter plots, volcano and GSEA plots, heatmaps, and pathway and network node and stick diagram (Supplemental Figure 1). Users can download raw data and signatures, analysis results and publication-ready graphics. iLINCS internal analysis and visualization engine uses R [16], Bioconductor packages[17], the Shiny framework[18], interactive graphics created with *ggplot* [19]and *plotly* [20], and integration of open-source visualization tools such as FTreeView[21] and Morpheous [22]. iLINCS also facilitates seamless integration with a wide range of task-specific online bioinformatics and systems biology tools and resources including Enrichr[23], DAVID[24], ToppGene[25], Reactome[26], KEGG[27], GeneMania[28], X2K Web[29], L1000FWD[30], STITCH[31], Clustergrammer[32], piNET[33], LINCS Data Portal[34], ScrubChem[35], PubChem[36], GEO[37], ArrayExpress[38] and GREIN[39]. Programmatic access to iLINCS data, workflows and visualizations are facilitated by embedding the calls to iLINCS API which are documented with the Swagger community standard. Examples of utilizing the iLINCS API within data analysis scripts are provided on GitHub (https://github.com/uc-bd2k/ilincsAPI). The iLINCS software architecture is described in Supplemental Figure 2.

iLINCS libraries of datasets, signatures and connections

iLINCS backend ***Databases*** contain >10,000 processed omics datasets, >220,000 omics signatures and >10^9 statistically significant "connections" between signatures. ***Omics datasets*** available for analysis and signatures creation include transcriptomic (RNA-seq and microarray) and proteomic (Reverse Phase Protein Arrays [40] and LINCS targeted mass spectrometry proteomics[5]) datasets. Dataset collections include transcriptomic and proteomics data generated by The Cancer Genome Atlas (TCGA) project, GEO GDS datasets, and the complete collection of GEO RNA-seq datasets. ***Omics signatures include***: *LINCS chemical and genetic perturbation signatures* consisting of genome-wide transcriptional response after genetic loss of function perturbation of more than 3,500 genes, or a perturbation by one of more than 4,000 chemical perturbagens based on LINCS L1000 assay data [2], *DrugMatrix* Chemogenomic database of 5,200 transcriptomic profiles of chemical toxicity [41], *Disease Related Signatures* consisting of 9,000 transcriptional signatures constructed by comparing sample groups within the collection of

Fig 2. Analysis of LINCS L1000 signatures of genetic and chemical perturbations. A) Most frequently perturbed genes among the Consensus Genes Signatures (CGS) connected to the MTOR knock-down CGS; B) Most frequent inhibition targets of chemical perturbagens with signatures connected to the MTOR CGS signature; C) Most enriched biological pathways for the everolimus signature; D) Most frequently perturbed genes among CGSes connected with everolimus signature, and pathways most enriched by the perturbed genes; E) Most frequent inhibition targets of chemical perturbagens with signatures connected to the everolimus signature and the pathways most enriched by the genes of the targeted proteins.



curated transcriptomics datasets from GEO[42], EBI Expression Atlas[8] signatures, and 5,000 pharmacogenomics signatures constructed from public domain datasets[4,43].

Use case 1: Identifying chemical perturbagens emulating genetic perturbation of the MTOR gene

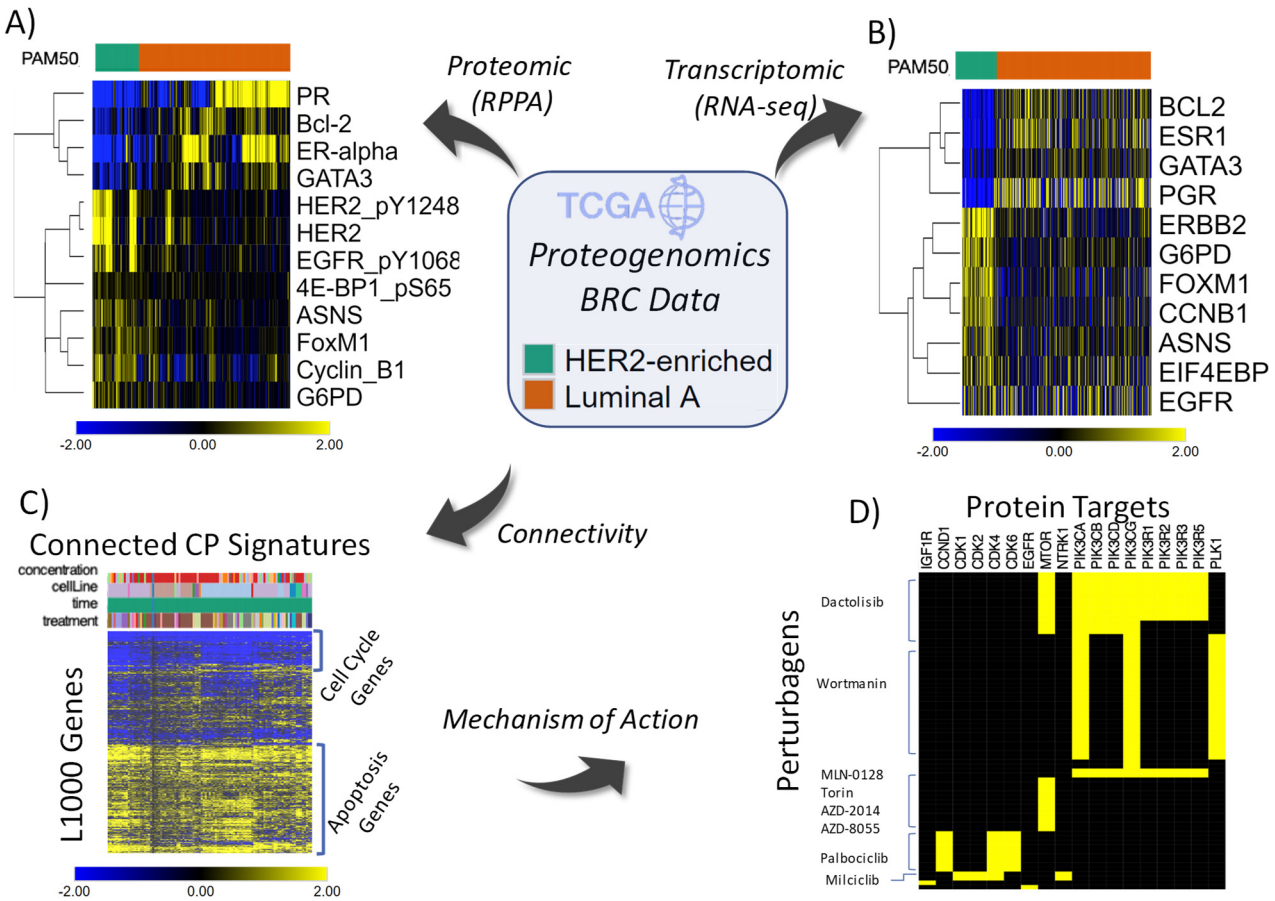
Aberrant activation of mTOR signaling underlies multiple human diseases and numerous efforts in designing drugs that modulate activity of MTOR signaling are under way[44]. Here we use the signature of genetic perturbation (via CRISPR knock-down) of the MTOR genes to identify chemical perturbagens mimicking the perturbation of the MTOR genes. First, we search through iLINCS libraries for Consensus Genes Signatures (CGSes) of MTOR knock-down and use the CRISPR CGS in MCF-7 cell line as the query signature. The connectivity analysis identifies 258 LINCS CGSes and 831 CP Signatures with statistically significant correlation with the query signature. Top 100 most connected CGSes are dominated by the signatures of genetic perturbations of MTOR and PIK3CA genes (Fig 2A), whereas all top 5 most frequent inhibition targets of CPs among top 100 most connected CP signatures are MTOR and PIK3 proteins (Fig 2B). Results clearly indicate that the query MTOR CGS is highly specific and sensitive to perturbation of the mTOR pathway and effectively identifies chemical perturbagens capable of inhibiting mTOR signaling. The full list of connected signatures is shown in Supplemental Table ST1. The connected CP signatures also include several chemical perturbagens with highly connected signatures that have not been known to target mTOR signaling providing additional candidate inhibitors. Step by step instructions for performing this analysis in iLINCS are provided in Supplemental Workflow SW1.

Use case 2: Mechanism of action analysis via connection to genetic perturbation signatures

Identifying small molecules (i.e. chemical perturbagens) that can modulate activity disease-related proteins or pathways is the cornerstone of intelligent drug design. Transcriptional signature of the chemical perturbagens often carry only an echo of such effects since the proteins directly targeted by the chemical and associated signaling proteins are not transcriptionally changed. iLINCS offers the solution for this problem by connecting the CP signatures to LINCS CGSes and follow-up systems biology analysis of genes whose CGSes are highly correlated with the CP signature. This is demonstrated by the analysis of one of the CP signatures of the 24 hour, 0.04µM treatment of the MCF-7 cell line with the mTOR inhibitor everolimus (Fig 2CDE). Traditional pathway enrichment analysis of the transcriptional signature via iLINCS connection to Enrichr (Fig 2C) fails to identify the mTOR pathway as being affected. In the next step, we first connect the CP signature to LINCS CGSes and then perform pathway enrichment analysis of genes with correlated CGSes. This analysis correctly identifies mTOR signaling pathway as the top most affected pathway (Fig 2D). Similarly, connectivity analysis with other CP signatures followed by the enrichment analysis of protein targets of top 100 most connected CPs again identifies the Pi3k-Akt signaling pathway as one of the most enriched (Fig 2E). In conclusion, both pathway analysis of differentially expressed genes in the everolimus signature and pathway analysis of connected genetic and chemical perturbagens provide us with important information about effects of everolimus. However, only the analyses of connected perturbagens correctly pinpoints the direct mechanism of action of the everolimus which is the inhibition of mTOR signaling. Step by step instructions for performing this analysis in iLINCS are provided in Supplemental Workflow SW2.

Use case 3: Proteo-genomics analysis of cancer driver events in breast cancer

Fig 3. Proteo-genomics analysis of cancer driver events in breast cancer. A) Most differentially expressed proteins in the proteomics signatures constructed by comparing RPPA profiles of Her2E and Luminal A BRC samples; B) Gene expression profile of the genes corresponding to proteins in A) based on RNA-seq data; C) Top 100 CP signatures most connected with the transcriptional signature constructed by comparing RNA-seq profiles of Her2E and Luminal A samples; D) Selected chemical perturbagens and their targets for CP signatures in C).



We analyzed TCGA breast cancer RNA-seq and RPPA data using the iLINCS "Datasets" workflow to construct the differential gene and protein expression signatures contrasting Luminal A and Her2 enriched (Her2E) breast tumors[45]. The protein expression signature immediately implicated known driver events distinguishing the two subtypes, the Luminal A cancers being driven by abnormal activity of the estrogen receptor and the Her2E tumors driven by abnormally high activity of the Her2 protein (Fig 3A). In addition to expected proteins, the increased level of phosphorylated EIF4EBP1 protein may indicate increased level of MTOR signaling in Her2E tumors.

The corresponding RNA-seq signature showed similar patterns of expression of key genes (Fig 3B). All genes were differentially expressed (Bonferroni adjusted p-value<0.01) except for EGFR, indicating that the difference in levels of post-translationally activated (phosphorylated) versions of the proteins came from activation of upstream kinases instead of overall increase in gene/protein expression. Analysis of 665 most significantly upregulated genes in Her2E tumors

(p-value<1e-10) identified cell cycle related KEGG pathways (Cell cycle, p-value=1.3e-26; DNA replication, p-value=1.4e-13) to be the most significantly enriched according to the Enrichr combined scores (See Supplemental Table ST2 for all results), implicating known increased proliferation of Her2E tumors in comparison to Luminal A tumors[46]. The connectivity analysis of the RNA-seq signature with LINCS CP signatures shows that treating several different cancer cell lines with inhibitors of PI3K, mTOR, CDK (Fig 3C) and inhibitors of some other more generic proliferation targets (eg. TOP21, AURKA) (see Supplemental Table ST3 for complete results) produces signatures that are positively correlated with RNA-seq Luminal A vs Her2E signature, indicating that such treatments are pushing cancer cell lines toward greater phenotypic similarity with Luminal A tumors.

Group analysis of connected CP signatures (Fig 3C) indicates that results of connectivity analysis may be largely driven by the inhibition of proliferation as evidenced by enrichment of down-regulated genes by cell cycle genes and up-regulated genes by apoptosis related genes in connected signatures. However, the dominance of PI3K and mTOR inhibitor signatures (Fig 3D) indicates that the connections may to some extent also be driven by more specific targeting of PI3K-mTOR signaling which may be more active in Her2E cancers as indicated by increased levels of the phosphorylated EIF4EBP1 protein.

High positive connectivity to Cancer Therapeutic Response Signatures[47] of two PI3K inhibitors in breast cancer cell lines also indicates that Her2E cancers may be more sensitive to mTOR inhibition (Supplemental Table ST3). Connectivity analysis with ENCODE transcription factor targets signatures recapitulated known biology (negative association with E2F4 binding signatures implicating higher proliferation of Her2E tumors and positive associations with ERaplha binding signatures implicating the increase ERalpha activity in Luminal A tumors) (Supplemental Table ST3). Most connected signatures in analysis of Disease related signatures[42] extracted from GEO data and EBI Expression Atlas[8] signatures were all related to comparisons of different breast cancers samples (Supplemental Table ST3). Step by step instructions for performing this analysis in iLINCS are provided in Supplemental Workflow SW3.

Other use cases: The three interconnected iLINCS workflows (Signatures, Datasets, Genes), facilitate a wide range of possible use cases. The three detailed cases above all use either pre-computed iLINCS signatures, or iLINCS omics datasets to construct signatures. Querying iLINCS with user submitted external signatures, genes and gene lists allows identification of connected perturbations and signatures. It also allows users to answer more specific questions about expression patterns of genes or gene lists of interest in specific datasets or across a class of cellular perturbations. For example, a query with a specific gene of interest can identify sets of perturbations that significantly affect the expression of the gene and thus offering a set of chemicals, or genetic perturbations that can be used to modulate the activity of the corresponding protein. A query with a list of genes whose coordinated expression is known to be a hallmark of a specific biological state or process[48] can identify a set of perturbations that can accordingly modify cell phenotype. Additional use cases have also been illustrated in several published scientific studies utilizing iLINCS: identification of putative therapeutic agents for schizophrenia[49,50], developing new strategies for ERα degradation in breast cancers[51], inhibiting the protective effects of stromal cells against chemotherapy in breast cancer[52] and rational drug combination design to inhibit epithelial-mesenchymal transition[53].

Discussion

iLINCS is a unique integrated platform for analysis of omics signatures. The three use cases described here only scratch the surface of the wide range of possible analyses facilitated by the interconnected analytical workflows and the large collections of omics datasets, signatures and their connections. These cases also feature only a small subset of all analytical tools integrated within the iLINCS platform. The user interfaces are streamlined and strive to be self-explanatory to the majority of scientists with conceptual understanding of omics data analysis. All analyses presented here were performed by typing the initial queries and then using the mouse to navigate user interfaces without ever having to copy and/or re-submit any portions of the data and results to a separate analytical tool. iLINCS implements the complete systematic polypharmacology and drug repurposing [54] workflow, and provides new innovative workflows for harnessing the full potential of LINCS omics signatures.

In addition to facilitating standard analyses, iLINCS also implements innovative workflows for biological interpretation of omics signatures via connectivity analysis. For example, in use case 2 we show how connectivity analysis coupled with pathway and gene set enrichment analysis can implicate mechanism of action of a chemical perturbagen when standard enrichment analysis applied to the differentially expressed genes fails to recover targeted signaling pathways. In a similar vein, iLINCS has been successfully used to identify putative therapeutic agents by connecting changes in proteomics profiles in neurons from patients with schizophrenia first with the LINCS CGSes of the corresponding genes, and then with LINCS CP signatures[49,50]. These analyses led to identification of PPAR agonists as promising therapeutic agents capable of reversing bioenergetic signature of schizophrenia, which were subsequently shown to modulate behavioral phenotypes in rat model of schizophrenia[49].

Several online tools have been developed for the analysis and mining LINCS L1000 signature libraries. They facilitate online queries of L1000 signatures[55,56] and construction of scripted pipelines for in-depth analysis of transcriptomics data and signatures[57]. The LINCS Transcriptomic Center at the Broad Institute developed the *clue.io* query tool deployed by the Broad Connectivity Map team which facilitates connectivity analysis of user submitted signatures[2]. iLINCS replicates the connectivity analysis functionality of clue.io, and indeed, the equivalent queries of the two systems may return qualitatively similar results (see Supplemental Results for a use case comparison). However, the scope of iLINCS is much broader. It provides connectivity analysis with signatures beyond Connectivity Map datasets and provides a very large number of primary omics datasets for users to construct their own signatures. Furthermore, analytical workflows in iLINCS facilitate deep systems biology analysis and knowledge discovery of both omics signatures and the genes and protein targets identified through connectivity analysis.

iLINCS removes technical roadblocks for users without programming background to re-use a large fraction of publicly available omics datasets and signatures. Furthermore, all analyses steps behind the iLINCS UI's are driven by API which themselves can and have been already used within computational pipelines based on scripting languages[58], such as R, Python and JavaScript, or to power functionality of other web analysis tools[33,59]. This makes iLINCS a natural tool for analysis and interpretation of omics signatures for scientists preferring point-and-click GUIs as well as data scientists using scripted analytical pipelines.

Methods

Perturbation signatures

All pre-computed perturbation signatures in iLINCS, as well as signatures created using an iLINCS dataset, consist of two vectors: the vector of log-scale differential expressions between the perturbed samples and baseline samples, and the vector of associated p-values. Signatures submitted by the user can also consist of only log-scale differential expressions without p-values, list of up- and down-regulated genes and single list of genes.

Signature connectivity analysis

Depending on the exact type of the query signature, the connectivity analysis with libraries of pre-computed iLINCS signature are computed using different connectivity metric.

If the query signature is selected from iLINCS libraries of pre-computed signatures, the connectivity with all other iLINCS signatures is pre-computed using the extreme Pearson's correlation[60,61] signed significances of all genes, where the signed significance of a gene is equal to -log10(p-value) multiplied by the sign of the log-differential expression. If the number of overlapping genes between significance vectors of two signatures is less than 2,500, 100 overlapping genes with most positive and 100 with most negative significance value are used for the extreme Pearson's correlations.

If the query signature is created from an iLINCS dataset, or directly uploaded by the user, the connectivity with all iLINCS signatures is calculated as the weighted correlation between the two vectors of log-differential expressions and the vector of weights equal to [-log10(p-value of the query) - log10(p-value of the iLINCS signature)][62]. When the user-uploaded signature consists of only log differential expression levels without p-values, the weight for the correlation is based only on the p-values of the iLINCS signatures [-log10(p-values of the iLINCS signatures)].

If the query signature uploaded by the user consists of the lists of up- and down-regulated genes connectivity is calculated by assigning -1 to down-regulated and +1 upregulated genes and calculating Pearson's correlation between such vector and iLINCS signatures. The calculated statistical significance of the correlation in this case is equivalent to the t-test for the difference between differential expression measures of iLINCS signatures between up- and down-regulated genes.

If the query signature is uploaded by the user in a form of a gene list, the connectivity with iLINCS signatures is calculated as the enrichment of highly significant differential expression levels in iLINCS signature within the submitted gene list using the Random Set analysis[63].

Perturbagen connectivity analysis

The connectivity between a query signature and a "perturbagen" is established using the enrichment analysis of individual connectivity scores between the query signature and set of all L1000 signatures of the perturbagen (for all cell lines, time points and concentrations). The analysis establishes whether the connectivity scores as a set are "unusually" high based on the Random Set analysis[63].

iLINCS signature libraries

LINCS L1000 signature libraries (Consensus gene knockdown signatures (CGS), Overexpression gene signatures and Chemical perturbation signatures) : For all LINCS L1000 signature libraries, the signatures are constructed by combining the Level 4, population control signature replicates from two released GEO datasets (GSE92742 and GSE70138) into the Level 5 moderated Z scores (MODZ) by calculating weighted averages as described in the primary publication for the L1000 Connectivity Map dataset[2]. Only signatures showing evidence of being reproducible by having the 75th quantile of pairwise spearman correlations of level 4 replicates (Broad institute distil_cc_q75 quality control metric[2]) greater than 0.2 are included. The corresponding p-values were calculated by comparing MODZ of each gene to zero using the Empirical Bayes weighted t-test with the same weights used for calculating MODZs. The shRNA and CRISPR knock-down signatures targeting the same gene were further aggregated into Consensus gene signatures (CGSes)[2] by the same procedure used to calculate MODZs and associated p-values.

LINCS targeted proteomics signatures: Signatures of chemical perturbations assayed by the quantitative targeted mass spectrometry proteomics P100 assay measuring levels 96 phosphopeptides and GCP assay against ~60 probes that monitor combinations of post-translational modifications on histones[5].

Disease related signatures: Transcriptional signatures constructed by comparing sample groups within the collection of curated public domain transcriptional dataset (GEO DataSets collection)[37]. Each signature consists of differential expressions and associated p-values for all genes calculated using Empirical Bayes linear model implemented in the *limma* package.

ENCODE transcription factor binding signatures: Genome-wide transcription factor (TF) binding signatures constructed by applying the TREG methodology to ENCODE ChiP-seq[64]. Each signature consists of scores and probabilities of regulation by the given TF in the specific context (cell line and treatment) for each gene in the genome.

Connectivity Map Signatures: Transcriptional signatures of perturbagen activity constructed based on the version 2 of the original Connectivity Map dataset using Affymetrix expression arrays[42]. Each signature consists of differential expressions and associated p-values for all genes when comparing perturbagen treated cell lines with appropriate controls.

DrugMatrix signatures: Toxicogenomic signatures of over 600 different compounds[41] maintained by the National Toxicology Program[65] consisting of genome-wide differential gene expression levels and associated p-values.

Transcriptional signatures from EBI Expression Atlas: All mouse, rat and human differential expression signatures and associated p-values from manually curated comparisons in the Expression Atlas[8].

Cancer therapeutics response signatures: These signatures were created by combining transcriptional data with drug sensitivity data from the Cancer Therapeutics Response Portal (CTRP) project[47]. Signatures were created separately for each tissue/cell lineage in the dataset by comparing gene expression between the five cell lines of that lineage that were most and five that were least sensitive to a given drug area as measured by the concentration-response curve (AUC) using two-sample t-test.

Pharmacogenomics transcriptional signatures: These signatures were created by calculating differential gene expression levels and associated p-value between cell-lines treated with anti-

cancer drugs and the corresponding controls in two separate projects: The NCI Transcriptional Pharmacodynamics Workbench (NCI-TPW)[43] and the Plate-seq project dataset[4].

Constructing signatures from iLINCS datasets

The transcriptomics or proteomics signature is constructed by comparing expression levels of two groups of samples (treatment group and baseline group) using Empirical Bayes linear model implemented in the *limma* package[66]. For the *GREIN* collection of GEO RNA-seq datasets[39], the signatures are constructed using the negative-binomial generalized linear model as implemented in the *edgeR* package[67].

Analytical tools, web applications and web resources

Signatures analytics in iLINCS is facilitated via native R, Java, JavaScript and Shiny applications, and via API connections to external web application and services. Brief listing of analysis and visualization tools is provided here. The overall structure of iLINCS is described in the Supplemental Results.

Gene list enrichment analysis is facilitated by directly submitting lists of gene to any of the three prominent enrichment analysis web tools: Enrichr[23], DAVID[24], ToppGene[25]. The manipulation and selection of list of signature genes is facilitated via an interactive volcano plot JavaScript application (shown in Supplemental Workflow 3).

Pathway analysis is facilitated through either general purpose enrichment tool (Enrichr, DAVID, ToppGene), the enrichment analysis of Reactome pathways via Reactome online tool[26], and internal R routines for SPIA analysis[68] of KEGG pathways and general visualization of signatures in the context of KEGG pathways using the KEGG API[27].

Network analysis is facilitated by submitting lists of genes to Genemania[28] and by internal iLINCS Shiny Signature Network Analysis (SigNetA) application.

Heatamap visualizations are facilitated by native iLINCS applications: Java based FTreeView[21], modified version of the JavaScript based Morpheus[22] and a Shiny based HeatMap application and by connection to the web application Clustergrammer[32].

Dimensionality reduction analysis (PCA and t-SNE[69]) and visualization of high-dimensional relationship via interactive 2D and 3D scatter plots is facilitated via internal iLINCS Shiny applications.

Interactive box-plots, scatter plots, GSEA plots, bar charts and pie charts used throughout iLINCS are implemented using R ggplot[19] and plotly[20].

Additional analysis are provided by connection X2K Web[29] (inference of upstream regulatory networks from signature genes), L1000FWD[30] (connectivity with signatures constructed using characteristic dimension methodology), STITCH[31] (visualization of drug-target networks), piNET[33] (visualization of gene-to-pathway relationships for signature genes).

Additional information about drugs, genes and proteins are provided by links to, LINCS Data Portal[34], ScrubChem[35], PubChem[36], Harmonizome[70], GeneCards[71], and several other only databases.

Gene and protein expression dataset collections

iLINCS backend databases provide access to more than 11,000 pre-processed gene and protein expression datasets that can be used to create and analyze gene and expression protein signatures. Datasets are thematically organized into eight collections with some datasets assigned to multiple collections. User can search all datasets, or browse datasets by collection.

LINCS collection: Datasets generated by the LINCS data and signature generation centers[7]

TCGA collection: Gene expression (RNASeqV2), protein expression (RPPA), and copy number variation data generated by TCGA project[45]

GDS collection: A curated collection of GEO Gene Datasets (GDS)[37]

Cancer collection: An ad-hoc collection of cancer related genomics and proteomic datasets

Toxicogenomics collection: An ad-hoc collection of toxicogenomics datasets

RPPA collection: An ad-hoc collection of proteomic datasets generated by Reverse Phase Protein Array assay[72]

GREIN collection: Complete collection of preprocessed human, mouse and rat RNA-seq data in GEO provided by the GEO RNA-seq Experiments Interactive Navigator (GREIN)[73]

Reference collection: An ad-hoc collection of important gene expression datasets.

Reference List

1 Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Computational Biology* **13**, e1005457, doi:10.1371/journal.pcbi.1005457 (2017).

2 Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e1417, doi:10.1016/j.cell.2017.10.049 (2017).

3 Bushel, P. R., Paules, R. S. & Auerbach, S. S. A Comparison of the TempO-Seq S1500+ Platform to RNA-Seq and Microarray Using Rat Liver Mode of Action Samples. *Frontiers in genetics* **9**, 485-485, doi:10.3389/fgene.2018.00485 (2018).

4 Bush, E. C. *et al.* PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nature Communications* **8**, 105, doi:10.1038/s41467-017-00136-z (2017).

5 Abelin, J. G. *et al.* Reduced-representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-scale Comparison of Drug-induced Phenotypes. *Molecular & cellular proteomics : MCP* **15**, 1622-1641, doi:10.1074/mcp.M116.058354 (2016).

6 Zhang, Y. *et al.* A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell* **31**, 820-832.e823, doi:https://doi.org/10.1016/j.ccell.2017.04.013 (2017).

7 Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Systems* **6**, 13-24, doi:https://doi.org/10.1016/j.cels.2017.11.001 (2018).

8 Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic acids research* **46**, D246-D251, doi:10.1093/nar/gkx1158 (2018).
9 Wang, Z. *et al.* Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature communications* **7**, 12846-12846, doi:10.1038/ncomms12846 (2016).
10 Keenan, A. B. *et al.* Connectivity Mapping: Methods and Applications. *Annual Review of Biomedical Data Science* **2**, 69-92, doi:10.1146/annurev-biodatasci-072018-021211 (2019).
11 Tarca, A. L., Bhatti, G. & Romero, R. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLOS ONE* **8**, e79217, doi:10.1371/journal.pone.0079217 (2013).
12 Mitrea, C. *et al.* Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology* **4**, 278 (2013).
13 Ideker, T. & Krogan, N. J. Differential network biology. *Molecular Systems Biology* **8**, 565, doi:10.1038/msb.2011.99 (2012).
14 Strömbäck, L., Jakoniene, V., Tan, H. & Lambrix, P. Representing, storing and accessing molecular interaction data: a review of models and tools. *Briefings in Bioinformatics* **7**, 331-338, doi:10.1093/bib/bbl039 (2006).
15 Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935 (2006).
16 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2016).
17 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
18 shiny: Web Application Framework for R (2016).
19 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).
20 Sievert, C. *et al.* plotly: Create Interactive Web Graphics via 'plotly. js'. *R package version* **4**, 110 (2017).
21 Freudenberg, J. M., Joshi, V. K., Hu, Z. & Medvedovic, M. CLEAN: CLustering Enrichment ANalysis. *BMC Bioinformatics* **10**, 234 (2009).
22 *Morpheus*, <https://software.broadinstitute.org/morpheus> (
23 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90-W97, doi:10.1093/nar/gkw377 (2016).
24 Dennis, G., Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).
25 Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305-W311, doi:10.1093/nar/gkp427 (2009).
26 Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Research* **44**, D481-D487, doi:10.1093/nar/gkv1351 (2016).
27 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353-D361, doi:10.1093/nar/gkw1092 (2017).

28 Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* **38**, W214-W220 (2010).
29 Clarke, D. J. B. *et al.* eXpression2Kinases (X2K) Web: linking expression signatures to upstream cell signaling networks. *Nucleic acids research* **46**, W171-W179, doi:10.1093/nar/gky458 (2018).
30 Wang, Z., Lachmann, A., Keenan, A. B. & Ma'ayan, A. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* **34**, 2150-2152, doi:10.1093/bioinformatics/bty060 (2018).
31 Kuhn, M. *et al.* STITCH 2: an interaction network database for small molecules and proteins. *Nucl.Acids Res.* **38**, D552-D556 (2010).
32 Fernandez, N. F. *et al.* Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. **4**, 170151 (2017).
33 Shamsaei, B. *et al.* piNET: a versatile web platform for downstream analysis and visualization of proteomics data. *bioRxiv*, 607432, doi:10.1101/607432 (2019).
34 Koleti, A. *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**, D558-D566, doi:10.1093/nar/gkx1063 (2018).
35 Harris, J. B. in *Bioinformatics and Drug Discovery* 37-47 (Springer, 2019).
36 Kim, S. *et al.* PubChem substance and compound databases. *Nucleic acids research* **44**, D1202-D1213 (2015).
37 Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.* **37**, D885-D890 (2009).
38 Parkinson, H. *et al.* ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucl.Acids Res.* **37**, D868-D872 (2009).
39 Al Mahi, N., Najafabadi, M. F., Pilarczyk, M., Kouril, M. & Medvedovic, M. GREIN: An interactive web platform for re-analyzing GEO RNA-seq data. *Scientific reports* **9**, 7580 (2019).
40 Li, J. *et al.* TCPA: a resource for cancer functional proteomics data. *Nature methods* **10**, 1046-1047, doi:10.1038/nmeth.2650 (2013).
41 Ganter, B. *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**, 219-244, doi:https://doi.org/10.1016/j.jbiotec.2005.03.022 (2005).
42 Freudenberg, J. M., Sivaganesan, S., Phatak, M., Shinde, K. & Medvedovic, M. Generalized random set framework for functional enrichment analysis using primary genomics datasets. *Bioinformatics* **27**, 70-77, doi:10.1093/bioinformatics/btq593 (2011).
43 Monks, A. *et al.* The NCI Transcriptional Pharmacodynamics Workbench: A Tool to Examine Dynamic Expression Profiling of Therapeutic Response in the NCI-60 Cell Line Panel. *Cancer Research* **78**, 6807-6817, doi:10.1158/0008-5472.Can-18-0989 (2018).
44 Saxton, R. A. & Sabatini, D. M. mTOR Signaling in Growth, Metabolism, and Disease. *Cell* **168**, 960-976, doi:https://doi.org/10.1016/j.cell.2017.02.004 (2017).
45 Consortium, T. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).

46 Bastien, R. R. *et al.* PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. *BMC Medical Genomics* **5**, 44, doi:10.1186/1755-8794-5-44 (2012).
47 Rees, M. G. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology* **12**, 109-116, doi:10.1038/nchembio.1986 (2016).
48 Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).
49 Sullivan, C. R. *et al.* Connectivity Analyses of Bioenergetic Changes in Schizophrenia: Identification of Novel Treatments. *Molecular Neurobiology* **56**, 4492-4517, doi:10.1007/s12035-018-1390-4 (2019).
50 Bentea, E. *et al.* Kinase network dysregulation in a human induced pluripotent stem cell model of DISC1 schizophrenia. *Molecular Omics* **15**, 173-188, doi:10.1039/C8MO00173A (2019).
51 Busonero, C., Leone, S., Bartoloni, S. & Acconcia, F. Strategies to degrade estrogen receptor α in primary and ESR1 mutant-expressing metastatic breast cancer. *Molecular and cellular endocrinology* (2018).
52 Barneh, F. *et al.* Valproic acid inhibits the protective effects of stromal cells against chemotherapy in breast cancer: Insights from proteomics and systems biology. *Journal of cellular biochemistry* **119**, 9270-9283 (2018).
53 Barneh, F. *et al.* Integrated use of bioinformatic resources reveals that co-targeting of histone deacetylases, IKBK and SRC inhibits epithelial-mesenchymal transition in cancer. *Briefings in Bioinformatics* **20**, 717-731, doi:10.1093/bib/bby030 (2018).
54 Liu, T.-P., Hsieh, Y.-Y., Chou, C.-J. & Yang, P.-M. Systematic polypharmacology and drug repurposing via an integrated L1000-based Connectivity Map database mining. **5**, 181321, doi:doi:10.1098/rsos.181321 (2018).
55 Duan, Q. *et al.* L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *Npj Systems Biology And Applications* **2**, 16015, doi:10.1038/npjsba.2016.15 https://www.nature.com/articles/npjsba201615#supplementary-information (2016).
56 Musa, A., Tripathi, S., Dehmer, M. & Emmert-Streib, F. L1000 Viewer: A Search Engine and Web Interface for the LINCS Data Repository. *Frontiers in Genetics* **10**, doi:10.3389/fgene.2019.00557 (2019).
57 Torre, D., Lachmann, A. & Ma'ayan, A. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. *Cell Systems* **7**, 556-561.e553, doi:https://doi.org/10.1016/j.cels.2018.10.007 (2018).
58 Guebila, M. B. & Thiele, I. Predicting gastrointestinal drug effects using contextualized metabolic models. *PLoS computational biology* **15**, e1007100 (2019).
59 Stathias, V. *et al.* LINCS Data Portal 2.0: Next Generation Access Point for Perturbation-Response Signatures. *Nucl.Acids Res.* **In Press** (2019).
60 Iwata, M., Sawada, R., Iwata, H., Kotera, M. & Yamanishi, Y. Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Scientific Reports* **7**, 40164, doi:10.1038/srep40164 https://www.nature.com/articles/srep40164#supplementary-information (2017).
61 Cheng, J. *et al.* in *Pac. Symp. Biocomput.* 5-16 (World Scientific).

62 Engreitz, J. *et al.* Content-based microarray search using differential expression profiles. *BMC Bioinformatics* **11**, 603 (2010).
63 Newton, M. A., Quinatan, F. A., den Boon, J. A., Sengupta, S. & Ahlquist, P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics* **1**, 85-106 (2007).
64 Chen, J. *et al.* Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules. *PLoS Comput Biol* **9**, e1003198, doi:10.1371/journal.pcbi.1003198 (2013).
65 Auerbach, S. DrugMatrix® and ToxFX® Coordinator National Toxicology Program. *National Toxicology Program: Dept of Health and Human Services*.
66 Smyth, G. K. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds R. Gentleman *et al.*) 397-420 (Springer, 2005).
67 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, doi:10.1093/bioinformatics/btp616 (2010).
68 Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75-82, doi:10.1093/bioinformatics/btn577 (2009).
69 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
70 Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, doi:10.1093/database/baw100 (2016).
71 Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database* **2010**, doi:10.1093/database/baq020 (2010).
72 Tibes, R. *et al.* Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular cancer therapeutics* **5**, 2512-2521 (2006).
73 Mahi, N. A., Najafabadi, M. F., Pilarczyk, M., Kouril, M. & Medvedovic, M. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. *Scientific Reports* **9**, 7580, doi:10.1038/s41598-019-43935-8 (2019).