Open access • Posted Content • DOI:10.1101/705228

Predicting mechanism of action of cellular perturbations with pathway activity signatures — Source link

Yan Ren, Siva Sivaganesan, Nicholas A. Clark, Lixia Zhang ...+4 more authors

Institutions: University of Cincinnati

Published on: 16 Jul 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Title: Predicting mechanism of action of cellular perturbations with pathway activity signatures

Author Affiliation:

Yan Ren[1], Siva Sivaganesan[2], Nicholas A. Clark[1], Lixia Zhang[1], Jacek Biesiada[1], Wen Niu[1], David R. Plas[3], Mario Medvedovic[1*]

[1]Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, OH, USA

[2]Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA

[3]Department of Cancer Biology, University of Cincinnati College of Medicine, Cincinnati, OH, USA

[*]Corresponding Author:

Mario Medvedovic: Department of Environmental Health, University of Cincinnati, 160 Panzeca Way, Cincinnati, OH 45267-0056, USA. Tel: +1 513 558 8564. Email: Mario.Medvedovic@uc.edu.

Keywords: signaling pathways, mechanism of action, MOA, transcriptomic signatures, gene expression, LINCS, perturbation signatures, L1000

ABSTRACT

Motivation:

Misregulation of signaling pathway activity is etiologic for many human diseases, and modulating activity of signaling pathways is often the preferred therapeutic strategy. Understanding the mechanism of action (MOA) of bioactive chemicals in terms of targeted signaling pathways is the essential first step in evaluating their therapeutic potential. Changes in signaling pathway activity are often not reflected in changes in expression of pathway genes which makes MOA inferences from transcriptional signatures a difficult problem.

Results:

We developed a new computational method for implicating pathway targets of bioactive chemicals and other cellular perturbations by integrated analysis of pathway network topology, the LINCS transcriptional signatures of genetic perturbations of pathway genes and the transcriptional signature of the perturbation. Our methodology accurately predicts signaling pathways targeted by the perturbation when current pathway analysis approaches utilizing only a transcriptional signature of the perturbation fail.

Availability and Implementation:

Open source R package paslincs is available at https://github.com/uc-bd2k/paslincs.

INTRODUCTION

Misregulation of signaling pathway activity underlies many human diseases (Finkel and Gutkind, 2003; Laplante and Sabatini, 2012; Saxton and Sabatini, 2017). Identifying small molecules (i.e. chemical perturbagens) that can modulate activity of disease-related signaling pathways is the corner stone of intelligent drug design. This concept is exemplified by misregulation of the MTOR signaling pathways in various disorders and the activity of designing drugs to modulate MTOR signaling (Saxton and Sabatini, 2017). In the context of signaling pathways, the mechanism of action (MOA) of a biologically active molecule usually represents the direct effect that the molecule has on the activity of specific proteins in a pathway and therefore on the activity of the downstream elements within the pathway. The pathway MOA of bioactive molecules is important not only in assessing their therapeutic potential, but also their toxicity (Heijne, et al., 2005). In environmental toxicology, the target pathways are the essential component of the adverse outcome pathways framework aiming to predict the adverse health outcomes resulting from exposure to environmental exposures (Ankley, et al., 2010). The recently released dataset of perturbation transcriptional signatures (TS), consisting of genome-wide transcriptional changes after treatment with chemical perturbagens (CP)(Subramanian, et al., 2017), provides an opportunity to define MOAs of a large set of CPs. However, inferring the MOA from a transcriptional signature has been a difficult problem. The transcriptional signature represents a consequence of modulating signaling pathway activity while changes in activity of signaling proteins are often direct consequences of post-translational modifications and are not necessarily reflected in consistent changes in mRNA expression levels of corresponding genes (Dugourd and Saez-Rodriguez, 2019; Geistlinger, et al., 2016).

Nevertheless, there has been intense interest in inferring changes in the biological pathway activities based on the transcriptional signature (Khatri, et al., 2012; Mitrea, et al., 2013; Tarca, et al., 2013). Previous methods have ranged from simple statistical enrichment of differentially expressed genes among genes/proteins in the pathway (Tarca, et al., 2013) to network-based approaches attempting to assess consistency of the gene expression changes with the topology of protein-protein, protein-gene, and gene-gene interactions in the pathway(Mitrea, et al., 2013). Recent benchmarks of these and other methods have shown that the incorporation of pathway topology often yields very limited, if any, positive effect on the performance of different methods (Geistlinger, et al., 2016) which was attributed to the lack of changes in expression of pathway genes. On the other hand, Perturbation-RespOnse GENes (PROGENy) methodology used the transcriptional “footprints” of perturbed pathway genes, but without accounting for the topology of the pathway interaction network (Schubert, et al., 2018).

Gene expression changes after shRNA- or CRISPR- based knockdown of a gene can be used to precisely define a transcriptional signature of protein inactivation (Bild, et al., 2006). The concordance between such a genetic perturbation (GP) transcriptional signature and a transcriptional signature of a chemical perturbagen (CP), indicates the plausibility that the CP is perturbing the activity of the protein (Pilarczyk, et al., 2019; Subramanian, et al., 2017). In addition to CP signatures, recently released L1000 dataset generated by the LINCS project (Library of Integrated Network-based Cellular Signature) provides genetic perturbation signatures consisting of averaged changes in gene expression after knocking down the same gene with multiple shRNA's for more than 3,500 human genes, perturbed in several cancer cell lines (Subramanian, et al., 2017). The new approach presented here leverages LINCS library of protein perturbation signatures to enable identification of the signaling pathways dysregulated by small molecules by integrated analysis of the CP signatures, the pathway network topology and GP signatures of pathway genes.

The key innovation of our methodology is the integration of two distinct strategies for implicating MOA of a chemical perturbagen: the topological pathway analysis (Mitrea, et al., 2013) and use of LINCS GP signatures (Pilarczyk, et al., 2019; Subramanian, et al., 2017). The integration is facilitated by an innovative statistical learning approach that uses the information about the topology of protein-protein interactions within a pathway and the LINCS GP signatures of the genes in the pathway to construct a pathway activity signature (PAS). We show that correlating transcriptional signatures of chemical perturbagens and other cellular perturbations with such PAS can implicate signaling pathways that are affected by the perturbation when standard methods fail to detect a signal. We also show how the new method can be used to refine pathway network models for specific biological contexts.

Methods

Transcriptional pathway activity signature (PAS)

Altered activity of a protein in a signaling pathway responding to chemical or genetic perturbation results in downstream changes in gene expression levels which are captured by the transcriptional signature. Our methods aim to identify the signaling pathways affected by the perturbation by comparing its transcriptional signature to GP signatures in the context of the pathway topology (Fig 1A). The key step in this process is the construction of the pathway activity signature (PAS) by integrating the topology of regulatory relationships within the pathway and the LINCS GP signaturees of genes in the pathway. Fig 1B-E illustrates the construction of the PAS on a small excerpt of the mTOR signaling pathway. The PAS is constructed in two steps: 1) *Signature genes* are selected by quantifying the consistency of changes in expression for each of the 978 measured L1000 landmark genes (Subramanian, et al., 2017) across the LINCS GP signatures of genes in the pathway with the pathway topology via a statistical model (Fig 1B-D); 2) The gene expression profiles of signature genes are summarized into a PAS (Fig 1E).

To assess the consistency of the gene expression profile ($\boldsymbol{y}$) of a single landmark gene across LINCS GP signatures of pathway genes with the pathway topology (Fig 1B), the pathway topology is summarized by the signed adjacency matrix A (Jacob, et al., 2012; Kunegis, et al., 2010). The assignment of positive (1) and negative (-1) weights to the edges in the pathway reflects the underlying assumption that GP signatures of two pathway genes should be positively correlated if the two genes are connected via "activating" interaction and negatively correlated if connected via inhibitory interaction. In the statistical model describing the distribution of the data (Fig 1C), the signed Laplacian (*L*) is then used as the precision matrix of the prior Markov random field for the mean expression changes ($\boldsymbol{\mu}$). The generative model for the data ($\boldsymbol{y}$) is defined as the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and a diagonal variance matrix. Finally, the posterior mean vector ($\widehat{\boldsymbol{\mu}}$) provides gene's expected expression pattern after integrating the observed expression profile ($\boldsymbol{y}$) and the pathway topology. The integration of the pathway topological structure and a gene expression profile by the statistical model is illustrated in Fig 1C for the gene expression profile representing the activation of a single node in the pathway (AKT1). The posterior mean estimate is consistent with the topology in the sense that its direction of the activity is consistent with the assumption that nodes connected by "activation" and "inhibition" relationships have positively and negatively correlated downstream effect on the gene expression changes respectively. Furthermore, the "closer" (in network topological sense) a node is to the initially activated node, the stronger is the activation signal. For each measured gene, we consider the norm of the projection of $\widehat{\boldsymbol{\mu}}$ onto the lower dimensional subspace with the highest information content to represent the measure of consistency (ie *consistency score*) of its expression profile with the pathway topology (Fig 1D). Using the list of the 100 genes with highest consistency scores (*signature genes)*, PAS is constructed as the first principal component of the data matrix consisting of expression changes of signature genes in GP signatures of genes in the pathway (Fig 1D). A mathematical description of the statistical learning model is provided in the Supplementary Materials Section A.

A high-information subspace corresponds to a linear space spanned by the eigenvectors of the signed Laplacian corresponding to small eigenvalues, which have intuitive appeal from graph-theoretic perspective (Jacob, et al., 2012; Kunegis, et al., 2010). This can be shown by analyzing the Bayes factor (Kass and Raftery, 1995) for choosing between two probabilistic models that generated the data ($\boldsymbol{y}$), one being the model in Fig 1C and the other being the model that assumes that topology of the pathway has no effect on the distribution of the data (Supplementary Materials Section B). We show that among the projections onto any one-dimensional subspace, the projection onto the null-subspace, the one corresponding to the eigenvector with the eigenvalue of zero, provides the highest discriminatory power for identifying genes with consistent expression profile (Supplementary Materials Section C). In our testing, going beyond the one-dimensional null-subspace did not improve significantly the discriminative ability of the signature (Supplementary Materials Section D). The eigenvector spanning the null space of the Laplacian ($\lambda = 0$), which is used to derive the gene activity scores, is visualized in Fig 1D. This illustrates the fact that basing the signature only on the projection to the null-space effectively summarizes the topology in terms of direction of the change in activity (increased, or decreased), but omits other aspects, such as distance of nodes in the network and the number of paths between different nodes that are captured by the rest of the eigenvectors.

Node contribution score

For the purpose of assessing the contribution of individual GP signatures of pathway genes to the PAS, we use the *node contribution score.* The *node contribution score* is defined as the decrease in the consistency scores of signature genes after removing the GP signature of the node from the analysis. A positive node contribution score implies that the GP signature of the node improves the consistency of the expression profiles of signature genes with the pathway topology. Wilcoxon signed-rank test is used to test whether the node contribution is statistically significant.

LINCS L1000 Consensus Gene Signatures and CP signatures

To construct GP and CP signatures used in the analyses, Level 4 LINCS L1000 dataset was downloaded from GEO (GSE92742). CP signatures were constructed by averaging Level 4 biological replicates. Level 5 moderated Z (MODZ) signatures of shRNA knock-downs were calculated as a weighted average of Level 4 replicates (Subramanian, et al., 2017). shRNA knock-down signatures were further integrated into GP signatures as weighted (MODZ) averages of individual shRNA signatures targeting the same gene. Only signatures designated as “gold” and generated using the “epsilon” version of L1000 probes were used in the analysis. All LINCS CP and GP signatures used in the analysis can be downloaded via *paslincs* package. The CP information of MOA is obtained from http://clue.io. Signatures of CPs that are activators or agonists of a target proteins were excluded from analyses.

Baseline methods compared with PAS methodology

We considered six baseline methods to compare with our pasLINCS methodology. The baseline methods were designed to establish the benefits of including information from the GP signatures of pathway genes and the pathway network topology in the process of identifying targeted pathways. The first method (KD), defines a pathway signature as the first principal component of all GP signatures of the pathway genes. This method uses the GP signatures, but not utilizes the pathway network topology to identify informative genes. The second method (TP), regards a CP signature of the landmark genes in the pathway proteins as a gene profile, and calculates the consistency score for this profile. Then the consistency score is considered as a measure of the association between a pathway and a CP. This method is meant to represent the class of pathway analysis methodologies that utilize pathway topology to identify affected pathways based on the transcriptional data alone and does not use GP signatures of pathway genes. The third baseline method is the random set (RS) enrichment analysis, a prototypical pathway enrichment analysis method that does not make use of either pathway topology or GP signatures (Newton, et al., 2007). Additionally, the performance of three existing topological pathway analysis methods for analysis of transcriptional signatures was assessed: SPIA (Tarca, et al., 2009), CePa (Gu and Wang, 2013), and PathNet (Dutta, et al., 2012).

ROC curves

For a specific target pathway, we focus on the TSes of CPs that inhibit any gene/protein within the pathway. For each of such TS, we designate all pathways not containing any protein/gene inhibited by the CP as true negatives, and calculate its false positive rate (FPR) as the proportion of correlations between the TS and PASes of true negative pathways that are larger than the correlation between the TS and the target pathway. For each FPR level, the corresponding true positive rate (TPR) is calculated as the proportion of all TSes targeting the pathway with FPR's smaller than the given FPR level. ROC curves are then obtained by plotting FPRs against the corresponding TPRs. For each ROC curve, we calculate the area under the curve (AUC). We also calculated the partial area under the curve (pAUC) corresponding to the FPR<0.05 as this is a better measure of the precision of the methods in the relevant range of the specificity (Cheng, et al., 2014). We report the ratio of the pAUC to the area under the 45-degree line (rpAUC) as the measure of increase in the predictive ability over random predictions.

Analysis of KEGG pathways

We processed *kgml* files corresponding to 328 homo sapiens KEGG pathways(Kanehisa, et al., 2017) using the R package *KEGGRest* to identify 179 "informative" pathways which contain at least two explicit "activation" or "inhibition" interactions and without "conflicting" interactions. For each informative pathway we constructed the signed adjacency matrix by setting the weights for "activation" edges to 1 and the weight for "inhibition" edges to -1, and calculating the signed Laplacian as shown in Fig 1. Supplementary Table S4 provides the summary of topological information contained in informative KEGG pathways. The pathways are grouped based on the secondary level classification in KEGG as: a.) pathways classified with the word "cancer" are grouped as "cancer"; b.) pathways classified related to a disease other than "cancer" are grouped as "disease"; c.) pathways classified with the word "signaling" or "signal" are grouped as "signaling"; d.) all other pathways are grouped as "other"

RESULTS

Transcription signature of mTOR signaling pathway activity

We studied the ability of our pasLINCS methodology to implicate genes whose expression pattern is a telltale sign of changes in pathway activity by constructing the PAS of mTOR pathway and by comparing it with transcriptional signatures of mTOR inhibitors. The protein interaction network representing mTOR signaling pathway was constructed by integrating information from KEGG and two recent papers describing the pathway (Saxton and Sabatini, 2017; Zhang, et al., 2017) (Fig 2A). The corresponding PAS showed a strong correlation with L1000 signatures of mTOR inhibitors in comparison with DMSO signatures (Fig 2B and 2C). To test whether the observed associations are platform independent we correlated PASes constructed from 12 LINCS cell lines with the time-course differential expression signatures of two glioma cell lines after treatment with a dual PI3K and mTOR inhibitor, PI-103 (Guillard, et al., 2009). Differential expression at 24 hours after PI-103 treatment in both glioma cell lines was significantly associated with the mTOR pathway PAS in the majority of the LINCS cell lines. Significant correlations can also be seen at 12 and 6 hours after treatment, but not before (Fig2D). These results are consistent with the expected dynamics of gene expression changes in response to PI-103 treatments (Guillard, et al., 2009). Similar analysis of the dataset studying the response of MCF-7 cell line to amino acid starvation (Tang, et al., 2015) showed consistent results with expected activation of mTOR signaling (Fig 2D). PASes constructed from three cell lines (NPC, SW480 and HCC515) showed lack of correlations in both analyses. SW480 cell line has previously been shown to be resistant to MTOR inhibition (Gulhati, et al., 2009), while the PAS for the NPC cell line was developed from only 6 CGSes. These factors along with relatively weak response (Fig 2E) to MTOR knockdown in HCC515 may explain the poor performance of PASes derived from these three cell lines.

Refining the pathway with node contribution scores

The pasLINCS can also be used to refine the pathway network by examining the changes in the pathway consistency scores for signature genes after removing one specific node. Nodes consistent with their implied role are expected to have a positive, statistically significant *node contribution score*. S6K proteins have been mapped as either upstream negative regulators or downstream positive output of mTORC1 activity in different biological contexts(Magnuson, et al., 2012; Shah and Hunter, 2006). Using node contribution scores, we studied the role played by S6K1 and S6K2 proteins in the upstream negative regulation ("feedback") of mTORC1 and as downstream effectors of mTORC1 signaling. Genetic and biochemical data show that mTORC1 directly phosphorylates and activates S6K1 and on the other hand, S6K1 phosphorylates and destabilizes IRS1, which decouples upstream receptor tyrosine kinases from PI3K-mTORC1 signaling. These two roles result in conflicting positions in the pathway (Fig 2F), and, in any given context, the transcriptional signatures of their activity will be more consistent with only one of these roles. Using the node contribution scores for these two proteins under two topological models, we established that the expression signature of S6K protein knock-downs in L1000 data are consistent with their roles of inhibitors of mTOR signaling in the MCF7 cell line (Fig 2F) and the majority of other 8 cell lines (Supplementary Materials Section F).

Predicting KEGG pathways affected by chemical perturbagens

We studied the ability of our methodology to identify KEGG signaling pathways modulated by a specific chemical perturbagen (CP). The evidence of CP effects on the activity of a pathway was assessed by the correlation between the CP transcriptional signature and the pathway's PAS. For each KEGG pathway and for the our custom mTOR pathway, we constructed ROC curves evaluating the ability of such correlations to implicate pathways targeted by a CP. Fig 3A shows the ROC curve for the new method applied to mTOR signaling pathway (Fig 2A). For comparison, ROC curves are shown for methods that use information from only the LINCS GP signatures (KD), only the pathway topology (TP), and classical gene list enrichment that does not utilize either pathway topology or LINCS GP signatures (RS) (Fig 3A), as well as three topological pathway analysis methods (SPIA, PathNet and CePa). The

ROC curves are summarized by the Area Under the Curve (AUC) and the area under the partial ROC curve (rpROC) for the high specificity (>0.9). Fig 3B shows the comparison of AUCs for all available KEGG pathways and in Fig3C detailed results for only signaling pathways are shown. The detailed information about performance for each pathway is shown in Fig S5 and Table S4, and summary results for rpAUC are shown in Fig S6. ROC analysis results indicate that: 1) pasLINCS methodology outperforms the methods based on simple enrichment analysis and topological pathway analysis methods that do not make use of GP signatures. The exact form of the topological pathway analysis method does not seem to make a big difference, with our simple and fast method (TP) overall performing on par with three state of the art topological pathway analysis methods.; 2) The use of the statistical model to identify signature genes based on the consistency of their expression profiles with the pathway topology improves the performance in comparison to using only GP signatures. 3) The use of GP signatures is particularly important in predicting affected signaling pathways with pasLINCS and KD methods outperforming methods that do not use GP signatures in 87% of signaling pathways

DISCUSSION

pasLINCS methodology integrates two distinct strategies for implicating signaling pathways affected by a CP based on its transcriptional signature: 1) the explicit modeling of shared expression changes implicated by the topology of the protein-protein interactions in the pathway (Mitrea, et al., 2013); and 2) Correlating CP transcriptional signature with the signatures of genetic perturbations of genes in the pathway (Pilarczyk, et al., 2019; Subramanian, et al., 2017). The use of GP signatures provides information about the activity changes in signaling proteins not contained in the CP transcriptional signature alone. Network based modeling integrates the information from different signaling proteins based on the expected interactions encoded by the pathway topology. Our results indicate that the new method is superior to either of the individual strategies in predicting signaling pathways affected by the bioactive chemical.

We compared the performance of our methods to three existing topological pathway methods (SPIA, PathNet and CePa) for analysis of transcriptional signatures that are applicable in analysis of LINCS CP signatures and do not require experimental replicates of treatment and control samples. They performed comparably with our own method (TP) and are likely representative of the results that can be obtained in general by a topological pathway analysis of CP signatures alone. In terms of methods that integrate multiple data types, PARADIGM uses a graphical model to integrate pathway network topology with Copy Number Variants (CNV) and gene expression profiles to identify pathway associated with cancer (Vaske, et al., 2010). Its graphical model, designed to integrate CNV and transcriptomic data, is not directly applicable for integration of GP and CP signatures.

Learning MOA of chemical perturbagens based on their transcriptional signatures opens new avenues for using connectivity map data to search for new therapies. In situations when the disease-related misregulation of signaling pathways is not clearly reflected in any available transcriptional signature, but is learned based on other information (e.g., genetics or proteomics studies), one can “connect” chemicals to disease based on their MOA. In the context of toxicogenomics, use of low-cost, high-throughput transcriptomic technologies (Bush, et al., 2017; Bushel, et al., 2018; Subramanian, et al., 2017), combined with pasLINCS analysis may open alternative avenues for high throughput safety evaluation of commercial chemicals, pesticides, food additives/contaminants, and medical products (Kavlock, et al., 2009; Kleinstreuer, et al., 2014). Previous studies have established the potential of assigning MOA of a chemical perturbagen based on comparison of their transcriptional signatures to the signatures of chemicals with known MOA (Iwata, et al., 2017; Subramanian, et al., 2017; Wang, et al., 2018). For example, the preciously derived PI3K inhibitor signature constructed from TSes of known chemical inhibitors (Zhang, et al., 2017) showed similar level of association with L1000 mTOR pathway inhibitors as we observed with our PAS (Fig 2D). pasLINCS adds another dimension by providing direct mechanistic link between the pathway activity and the effect of the CP without the need for reference signatures of perturbagens with known MOA.

The pasLINCS statistical learning model uses Bayesian inference to integrate the topological information with data on gene expression changes after perturbing nodes in the network. The key step in building the statistical model is the use of the regularized signed Laplacian as the precision matrix of the prior covariance to capture the effects of two basic kinds of protein-protein regulatory interactions in signaling cascades (activation and inhibition) on expression profiles of downstream genes. This simple representation is likely an oversimplification of the complexity of the dynamic biochemical processes taking place in transducing signals. However, our results show that the resulting covariance function captures adequately the static correlation structure of transcriptional signatures of pathway perturbations. The pasLINCS statistical learning model can be re-interpreted in the context of the regularization framework with graph kernels (Smola and Kondor, 2003) where standard Laplacian is replaced with the signed version. Similar strategies using standard graph with only positive edges have been used in the context of non-directed protein-protein interaction network (Cowen, et al., 2017) in general, as well as in predicting drug targets based on transcriptional signatures (Laenen, et al., 2013). The regularization formulation does not depend on the Gaussian distributional assumptions about the data used in our model, indicating that pasLINCS methodology is likely robust with respect to deviations from the distributional assumptions.

Our results demonstrate that pasLINCS methodology can be used to construct different variants of the pathway networks, but also postulate new hypotheses about the role that proteins may play in a signaling pathway. Analysis of the results indicated that the S6Ks GP signatures are more consistent with their role as inhibitors of the PI3K-AKT-mTOR signaling axes, but not as the transducers of mTORC1 activity. S6K1 is well established as a negative feedback regulator of insulin-stimulated AKT-mTORC1 signaling, while studies of S6K2 have revealed context-specific feedback function (Harrington, et al., 2004; Haruta, et al., 2000; Miller, et al., 2017; Pai, et al., 2016; Tremblay, et al., 2007). Optimization of node contribution scores led us to adopt the mTOR pathway network with both S6K1 and S6K2 inhibiting upstream pathway activation. Biologically, the observed results could also be explained by the fact that mTORC1 has multiple downstream transducers that affect transcriptional programs, in addition to S6K. Consequently, the inhibitory role of S6K affects more downstream transcriptional targets than its transducer role, which is then reflected in its GP signatures being dominated by its inhibitory role. Details of the S6K role may not be relevant to the goal of constructing an informative PAS, but it is easy to envision biological contexts in which such predictions would warrant reconfiguring the pathway with follow up experimentation to confirm the predicted role of a specific protein in the given context.

pasLINCS methodology opens a new avenue for functional analysis of transcriptomic data to discover mechanistic underpinnings of observed changes in gene expression levels. Our results indicate that in terms of implicating pathways affected by a CP, results of the pasLINCS analysis are complementary to the established enrichment strategies based on analysis CP transcriptional signatures alone. pasLINCS accurately predicts affected signaling pathways when established enrichment methods fail and should be included within general analytical pipelines for functional assessment of global changes in gene expression patterns.

FUNDING

This work has been supported by the National Institutes of Health [U54HL127624, P30ES006096] .

CONFLICT OF INTEREST

None

REFERENCES

Ankley, G.T., *et al.* Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry* 2010;29(3):730-741.

Bild, A.H., *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;439(7074):353-357.

Bush, E.C., *et al.* PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nature Communications* 2017;8(1):105.

Bushel, P.R., Paules, R.S. and 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* 2018;9:485-485.

Cheng, J., *et al.* Systematic evaluation of connectivity map for disease indications. *Genome medicine* 2014;6(12):95.

Cowen, L., *et al.* Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* 2017;18:551.

Dugourd, A. and Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Current Opinion in Systems Biology* 2019;15:82-90.

Dutta, B., Wallqvist, A. and Reifman, J. PathNet: a tool for pathway analysis using topological information. *Source Code for Biology and Medicine* 2012;7(1):10.

Finkel, T. and Gutkind, J.S. Signal Transduction and Human Disease. Wiley; 2003.

Geistlinger, L., Csaba, G. and Zimmer, R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set-& network-based enrichment analysis. *BMC bioinformatics* 2016;17(1):45.

Gu, Z. and Wang, J. CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics* 2013;29(5):658-660.

Guillard, S.*, et al.* Molecular pharmacology of phosphatidylinositol 3-kinase inhibition in human glioma. *Cell Cycle* 2009;8(3):443-453.

Gulhati, P.*, et al.* TARGETED INHIBITION OF mTOR SIGNALING INHIBITS TUMORIGENESIS OF COLORECTAL CANCER. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2009;15(23):7207-7216.

Harrington, L.S.*, et al.* The TSC1-2 tumor suppressor controls insulin-PI3K signaling via regulation of IRS proteins. *J Cell Biol* 2004;166(2):213-223.

Haruta, T.*, et al.* A rapamycin-sensitive pathway down-regulates insulin signaling via phosphorylation and proteasomal degradation of insulin receptor substrate-1. *Mol Endocrinol* 2000;14(6):783-794.

Heijne, W.H.M.*, et al.* Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. *Expert Rev Proteomics* 2005;2(5):767-780.

Iwata, M.*, et al.* Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Scientific Reports* 2017;7:40164.

Jacob, L., Neuvial, P. and Dudoit, S. More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.* 2012;6(2):561-600.

Kanehisa, M.*, et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 2017;45(D1):D353-D361.

Kass, R.E. and Raftery, A.E. Bayes Factors. *Journal of the American Statistical Association* 1995;90(430):773-795.

Kavlock, R.J., Austin, C.P. and Tice, R.R. Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment. *Risk Analysis* 2009;29(4):485-487.

Khatri, P., Sirota, M. and Butte, A.J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol* 2012;8(2):e1002375.

Kleinstreuer, N.C.*, et al.* Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nature Biotechnology* 2014;32:583.

Kunegis, J.*, et al.* Spectral analysis of signed graphs for clustering, prediction and visualization. In, *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM; 2010. p. 559-570.

Laenen, G.*, et al.* Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Molecular bioSystems* 2013;9(7):1676-1685.

Laplante, M. and Sabatini, David M. mTOR Signaling in Growth Control and Disease. *Cell* 2012;149(2):274-293.

Magnuson, B., Ekim, B. and Fingar, Diane C. Regulation and function of ribosomal protein S6 kinase (S6K) within mTOR signalling networks. *Biochemical Journal* 2012;441(1):1-21.

Miller, W.P.*, et al.* Activation of the Stress Response Kinase JNK (c-Jun N-terminal Kinase) Attenuates Insulin Action in Retina through a p70S6K1-dependent Mechanism. *J Biol Chem* 2017;292(5):1591-1602.

Mitrea, C.*, et al.* Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology* 2013;4:278.

Newton, M.A.*, et al.* Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics* 2007;1(1):85-106.

Pai, C., Walsh, C.M. and Fruman, D.A. Context-Specific Function of S6K2 in Th Cell Differentiation. *J Immunol* 2016;197(8):3049-3058.

Pilarczyk, M.*, et al.* Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINCS. *bioRxiv* 2019:826271.

Saxton, R.A. and Sabatini, D.M. mTOR Signaling in Growth, Metabolism, and Disease. *Cell* 2017;168(6):960-976.

Schubert, M.*, et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications* 2018;9(1):20.

Shah, O.J. and Hunter, T. Turnover of the Active Fraction of IRS1 Involves Raptor-mTOR- and S6K1-Dependent Serine Phosphorylation in Cell Culture Models of Tuberous Sclerosis. *Molecular and Cellular Biology* 2006;26(17):6425-6434.

Smola, A.J. and Kondor, R. Kernels and regularization on graphs. In, *Learning theory and kernel machines*. Springer; 2003. p. 144-158.

Subramanian, A.*, et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 2017;171(6):1437-1452.e1417.

Tang, X.*, et al.* Comprehensive Profiling of Amino Acid Response Uncovers Unique Methionine-Deprived Response Dependent on Intact Creatine Biosynthesis. *PLOS Genetics* 2015;11(4):e1005158.

Tarca, A.L., Bhatti, G. and Romero, R. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLOS ONE* 2013;8(11):e79217.

Tarca, A.L.*, et al.* A novel signaling pathway impact analysis. *Bioinformatics* 2009;25(1):75-82.

Tremblay, F.*, et al.* Identification of IRS-1 Ser-1101 as a target of S6K1 in nutrient- and obesity-induced insulin resistance. *Proc Natl Acad Sci U S A* 2007;104(35):14056-14061.

Vaske, C.J.*, et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010;26(12):i237-i245.

Wang, Z.*, et al.* L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* 2018;34(12):2150-2152.

Zhang, Y.*, et al.* A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell* 2017;31(6):820-832.e823.

TABLE AND FIGURES LEGENDS

Figure 1: Integrating Signaling Pathway Topology and LINCS Consensus Gene Signatures to construct transcriptional Pathway Activity Signatures (PAS). In all panels, shades of red indicate different levels of positive and shades blue different levels of negative numbers. A) Chemical or genetic perturbation affects the activity of a protein in a signaling pathway and dysregulates the activity of the pathway. The pathway disregulation results in downstream changes in gene expression levels (gray arrow) which is captured by the transcriptional signature (TS). pasLINCS aims to identify the changes in pathway activity based on the downstream TS (blue arrow). B) The pathway activity signature (PAS) is constructed by integrating information from the LINCS genetic perturbation (GP) signatures of the pathway genes and the topology of protein-protein interactions in the pathway, summarized by the signed adjacency and Laplacian matrices. Expression profile of gene is consistent with the pathway topology if activation interaction between two nodes results in the expression change in the same direction and the inhibition interaction result in the change in the opposite direction C) Statistical model for integrating pathway topology with the expression profile of a gene; D) The gene-level consistency score between gene expression profile and the pathway topology. Top 100 genes with the highest consistency score are selected as “signature genes”; E) PAS consists of the gene loadings of the first principal component for the data matrix for signature genes across GP signatures of pathway genes.

Figure 2: PAS of the mTOR signaling pathway. A) mTOR signaling pathway constructed from literature consisting of four key modules; B) PAS constructed using the methods in Fig 1 and the LINCS chemical perturbagen (CP) signatures for the pathway inhibitors and vehicle treatment; C) Distribution of differences in average expression levels between positive and negative signature genes for pathway inhibitors and vehicle treatment; D) Statistical significance of differences in expression level of positive and negative mTOR PAS genes after PI3K inhibition and amino acid starvation; E) Using node contribution scores to assess the role of S6K kinase in regulating mTOR pathway activity in the MCF7 PAS.

Figure 3: Predicting pathways perturbed by LINCS chemical perturbagens (CP). A) ROC curves for predicting correctly mTOR signaling pathways for CP’s known to target proteins in the pathway using seven different methods: PAS = pathway activity signatures using our new method; KD = pathway signatures constructed using only CGS data, but not utilizing the pathway topology; TP = using only CP transcriptional signatures and the pathway topology, but not using CGSes; RS = classical enrichment analysis not utilizing GP signatures or the pathway topology; B) Percentage of pathways predicted with the highest AUC for seven different methods across four different types of KEGG pathways; C) Heatmap of AUC’s for predicting affected KEGG signaling pathways for all seven methods.

FIGURES

Figure 1

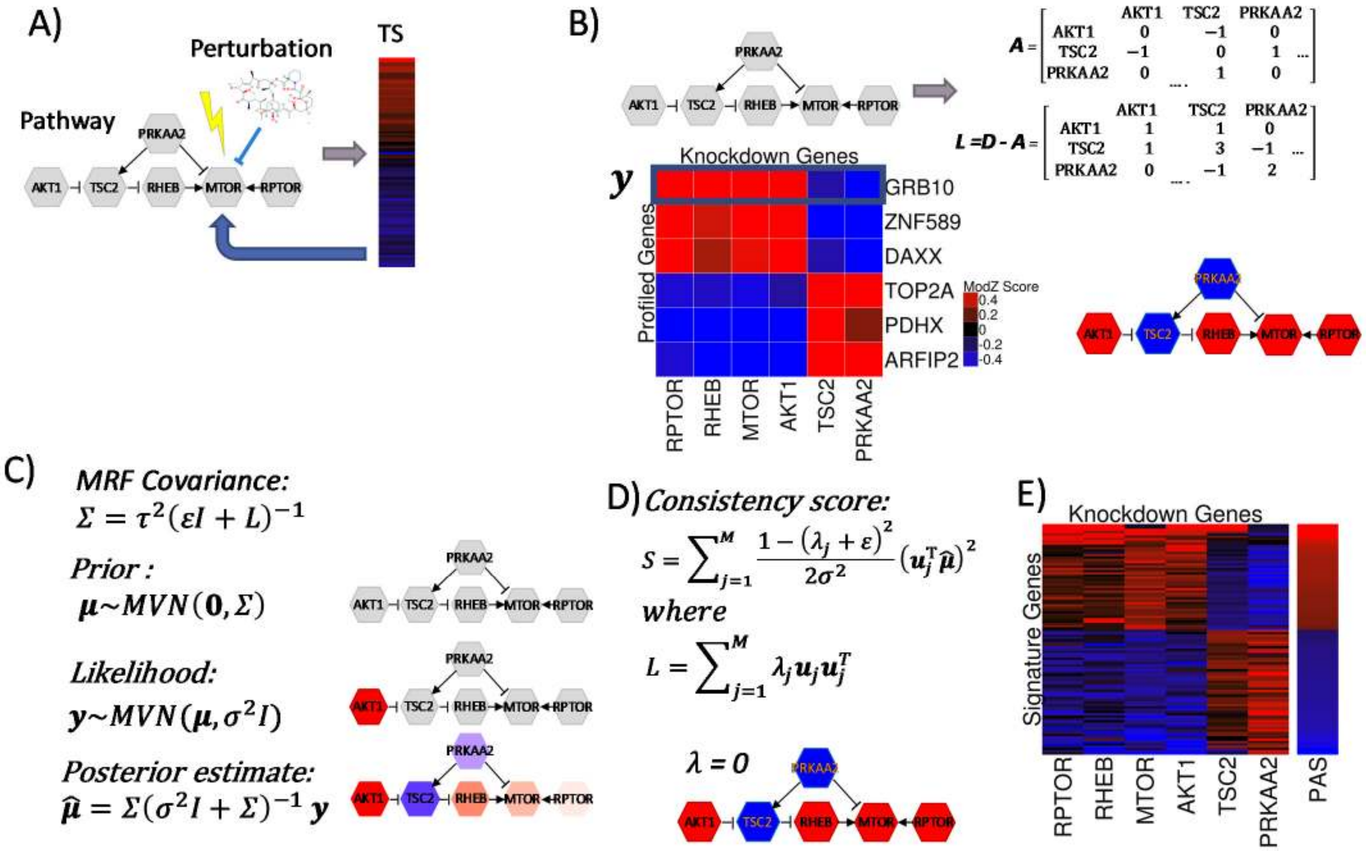
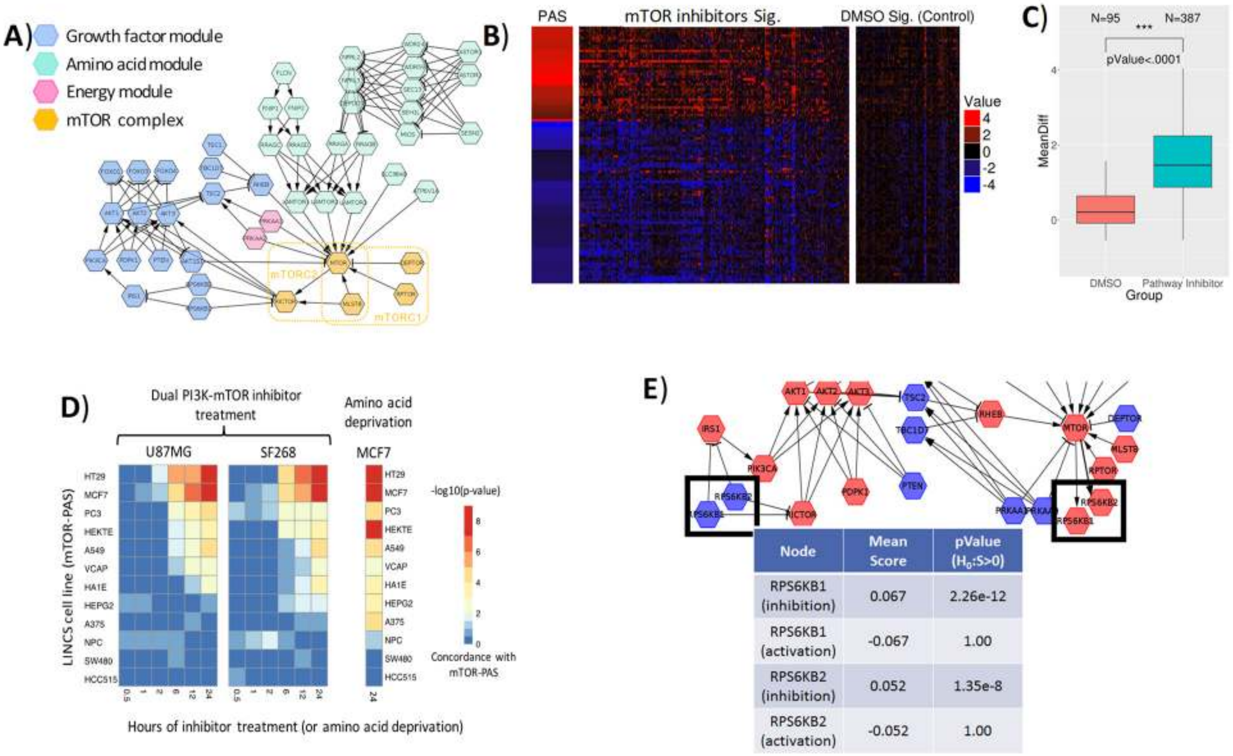
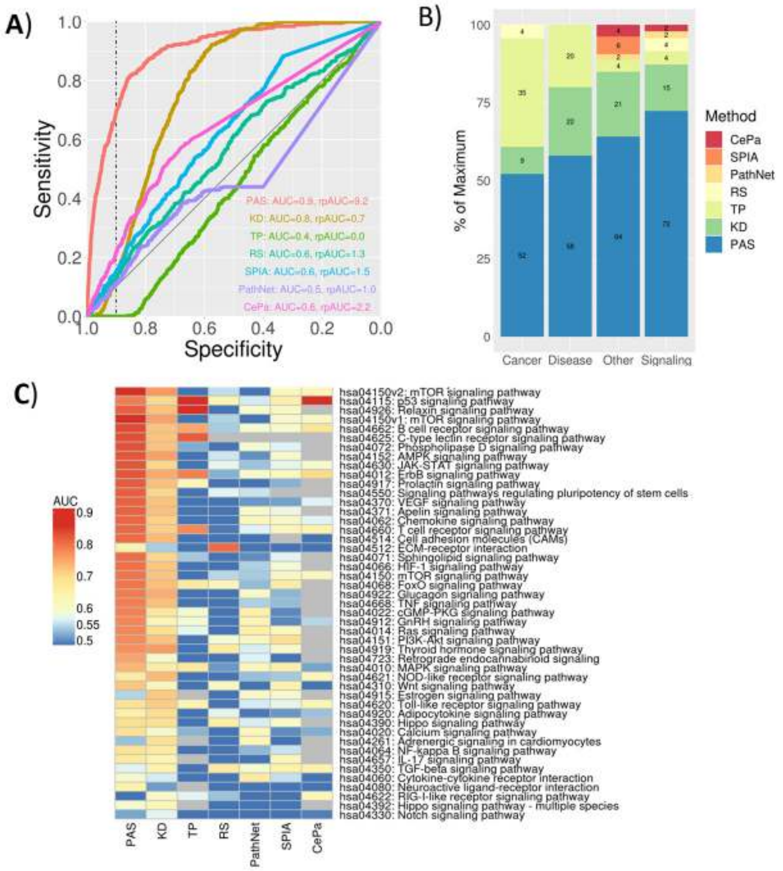

A)
Perturbation
TS
Pathway
PRKAA2
AKT1 TSC2 RHEB MTOR RPTOR
B)
PRKAA2
AKT1 TSC2 RHEB MTOR RPTOR
$A = \begin{array}{c|ccc} & AKT1 & TSC2 & PRKAA2 \\ \hline AKT1 & 0 & -1 & 0 \\ TSC2 & -1 & 0 & 1 \\ PRKAA2 & 0 & 1 & 0 \end{array} \ldots$
$L = D - A = \begin{array}{c|ccc} & AKT1 & TSC2 & PRKAA2 \\ \hline AKT1 & 1 & 1 & 0 \\ TSC2 & 1 & 3 & -1 \\ PRKAA2 & 0 & -1 & 2 \end{array} \ldots$
Knockdown Genes
y
Profiled Genes
GRB10
ZNF589
DAXX
TOP2A
PDHX
ARFIP2
ModZ Score
0.4
0.2
0
-0.2
-0.4
RPTOR RHEB MTOR AKT1 TSC2 PRKAA2
C)
MRF Covariance:
$\Sigma = \tau^2(\varepsilon I + L)^{-1}$
Prior :
$\boldsymbol{\mu} \sim MVN(\mathbf{0}, \Sigma)$
Likelihood:
$\mathbf{y} \sim MVN(\boldsymbol{\mu}, \sigma^2 I)$
Posterior estimate:
$\hat{\boldsymbol{\mu}} = \Sigma(\sigma^2 I + \Sigma)^{-1}\,\mathbf{y}$
D) Consistency score:
$S = \sum_{j=1}^{M} \frac{1-(\lambda_j+\varepsilon)^2}{2\sigma^2}(\mathbf{u}_j^T \hat{\boldsymbol{\mu}})^2$
where
$L = \sum_{j=1}^{M} \lambda_j \mathbf{u}_j \mathbf{u}_j^T$
$\lambda = 0$
E)
Knockdown Genes
Signature Genes
RPTOR RHEB MTOR AKT1 TSC2 PRKAA2
PAS

Figure 2



Node	Mean Score	pValue (H_0:S>0)
RPS6KB1 (inhibition)	0.067	2.26e-12
RPS6KB1 (activation)	-0.067	1.00
RPS6KB2 (inhibition)	0.052	1.35e-8
RPS6KB2 (activation)	-0.052	1.00

Figure 3


A)
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
PAS: AUC=0.9, rpAUC=9.2
KD: AUC=0.8, rpAUC=0.7
TP: AUC=0.4, rpAUC=0.0
RS: AUC=0.6, rpAUC=1.3
SPIA: AUC=0.6, rpAUC=1.5
PathNet: AUC=0.5, rpAUC=1.0
CePa: AUC=0.6, rpAUC=2.2
B)
% of Maximum
100
75
50
25
0
4
35
9
52
20
22
58
4
6
2
4
21
64
2
2
4
4
15
72
Cancer
Disease
Other
Signaling
Method
CePa
SPIA
PathNet
RS
TP
KD
PAS
C)
AUC
0.9
0.8
0.7
0.6
0.55
0.5
hsa04150v2: mTOR signaling pathway
hsa04115: p53 signaling pathway
hsa04926: Relaxin signaling pathway
hsa04150v1: mTOR signaling pathway
hsa04662: B cell receptor signaling pathway
hsa04625: C-type lectin receptor signaling pathway
hsa04072: Phospholipase D signaling pathway
hsa04152: AMPK signaling pathway
hsa04630: JAK-STAT signaling pathway
hsa04012: ErbB signaling pathway
hsa04917: Prolactin signaling pathway
hsa04550: Signaling pathways regulating pluripotency of stem cells
hsa04370: VEGF signaling pathway
hsa04371: Apelin signaling pathway
hsa04062: Chemokine signaling pathway
hsa04660: T cell receptor signaling pathway
hsa04514: Cell adhesion molecules (CAMs)
hsa04512: ECM-receptor interaction
hsa04071: Sphingolipid signaling pathway
hsa04066: HIF-1 signaling pathway
hsa04150: mTOR signaling pathway
hsa04068: FoxO signaling pathway
hsa04922: Glucagon signaling pathway
hsa04668: TNF signaling pathway
hsa04022: cGMP-PKG signaling pathway
hsa04912: GnRH signaling pathway
hsa04014: Ras signaling pathway
hsa04151: PI3K-Akt signaling pathway
hsa04919: Thyroid hormone signaling pathway
hsa04723: Retrograde endocannabinoid signaling
hsa04010: MAPK signaling pathway
hsa04621: NOD-like receptor signaling pathway
hsa04310: Wnt signaling pathway
hsa04915: Estrogen signaling pathway
hsa04620: Toll-like receptor signaling pathway
hsa04920: Adipocytokine signaling pathway
hsa04390: Hippo signaling pathway
hsa04020: Calcium signaling pathway
hsa04261: Adrenergic signaling in cardiomyocytes
hsa04064: NF-kappa B signaling pathway
hsa04657: IL-17 signaling pathway
hsa04350: TGF-beta signaling pathway
hsa04060: Cytokine-cytokine receptor interaction
hsa04080: Neuroactive ligand-receptor interaction
hsa04622: RIG-I-like receptor signaling pathway
hsa04392: Hippo signaling pathway - multiple species
hsa04330: Notch signaling pathway
PAS
KD
TP
RS
PathNet
SPIA
CePa