Classification:

Major: Biological Sciences

Minor: Systems Biology

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], 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

Yan Ren: Department of Environmental Health, University of Cincinnati, 160 Panzeca Way, Cincinnati, OH 45267-0056, USA. (ORCID: 0000-0002-8327-5346)

Siva Sivaganesan: Department of Mathematical Sciences, University of Cincinnati, 5402 French Hall, Cincinnati, OH 45221-0025, USA. (ORCID: 0000-0001-7093-8078)

Nicholas A. Clark: Department of Environmental Health, University of Cincinnati, 160 Panzeca Way, Cincinnati, OH 45267-0056, USA. (ORCID: 0000-0003-0105-9605)

Lixia Zhang: Department of Environmental Health, University of Cincinnati, 160 Panzeca Way, Cincinnati, OH 45267-0056, USA. (ORCID: 0000-0003-4094-0729)

David R. Plas: Department of Cancer Biology, University of Cincinnati, 3125 Eden Avenue, Cincinnati, OH 45267-0521, USA. (ORCID: 0000-0001-7568-1400)

Mario Medvedovic: Department of Environmental Health, University of Cincinnati, 160 Panzeca Way, Cincinnati, OH 45267-0056, USA. (ORCID: 0000-0003-4510-3102)

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

Misregulation of signaling pathway activity is etiologic for many human diseases, and modulating activity of signaling pathways dysregulated by the disease is often the preferred therapeutic strategy. Understanding the mechanism of action of a chemical perturbagen, or any other type of a cellular perturbation on targeted signaling pathways is the essential first step in evaluating its therapeutic potential. The transcriptional signature of a perturbation provides a convenient, high information content readout of changes in the cellular state after perturbation. However, the changes of signaling pathway activity are often not mediated by changes in mRNA levels of pathway constituents, rendering direct enrichment-type analyses ineffective in implicating perturbed signaling pathways. Using a new signaling pathway activity analysis method, we identified changes in signaling activity of targeted pathways with high accuracy. Our method uses LINCS libraries of transcriptional Consensus Gene Signatures (CGS) of global changes in mRNA levels in response to genetic loss-of-function of key nodes in the pathway to construct a novel pathway activity signature (PAS). We show that PASes can be effectively used to evaluate the potential of the perturbagen to modulate pathway activity and to further refine signaling network topology for a specific biological context.

SIGNIFICANCE STATEMENT

Understanding the mechanism of action of cellular perturbation, such as a treatment with a chemical, or deleterious mutation of a gene, has on a biological system is one of the central problems of biomedicine. For example, knowing the disease causing mutation and the chemical that can reverse the effects of such a mutation allows intelligent design of therapeutics. Measuring changes in expression levels of all genes in the genome after a perturbation (ie transcriptional signature) is a convenient, cost effective and highly informative readout of perturbation effects. We describe novel computational methodology capable of predicting mechanism of action of cellular perturbations based on its transcriptional signature and show that it can make crucial and accurate predictions in situations where current approaches fail.

INTRODUCTION

Misregulation of signaling pathway activity underlies many human diseases (1-3). 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 (1). 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 (4). 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 (5). The recently released dataset of perturbation transcriptional signatures (TS), consisting of genome-wide transcriptional changes after treatment with chemical perturbagens (CP)(6), provides an opportunity to define MOAs of a large set of CPs. However, inferring the MOA from a TS has been a difficult problem. The TS 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 (7, 8).

Nevertheless, there has been intense interest in inferring changes in the biological pathway activities based on the TS (9-11). Previous methods have ranged from simple statistical enrichment of differentially expressed genes among genes/proteins in the pathway (10) 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(11). 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(8) which was again attributed to the lack of changes in expression of pathway genes.

Gene expression changes after shRNA- or CRISPR- based knockdown of a gene can be used to precisely define a transcriptional signature of protein inactivation (12). The concordance between such a knockdown (KD) TS and a TS of a CP, indicates the plausibility that the CP is perturbing the activity of the protein (6). In addition to CP signatures, recently released L1000 dataset generated by the LINCS project (Library of Integrated Network-based Cellular Signature) provides a Consensus Gene Signature (CGS) 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 (6). Authors of that study have also demonstrated the utility of LINCS CGSes in implicating the MOA of chemical perturbagens. The new approach presented here leverages this library of protein perturbation signatures to enable identification of the signaling pathways dysregulated by small molecules by integrating CGSes, TSes of CPs, and pathway topology analyses.

Conceptually, our methodology integrates two distinct strategies for implicating CP MOA: the topological pathway analysis (11) and use of LINCS CGSes (6). Key to the integration is the implementation of an innovative statistical learning model to incorporate the information about the topology of protein-protein interactions within a pathway and the LINCS CGSes of the genes in the pathway to construct a new pathway activity signature (PAS). We show that correlating TSes of chemical perturbagens and other cellular perturbations with such PAS is effective in implicating pathways that are affected by the CP, and it can be used to refine pathway models for specific biological contexts.

RESULTS

Transcriptional pathway activity signature overview

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 TS. Our methods aim to identify the signaling pathways affected by the perturbation by comparing its TS to CGSes 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 knock-down CGSes 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 (LM) genes (6) across the LINCS CGSes 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 gene expression profile of a single LM gene across LINCS CGSes of pathway genes ($\boldsymbol{y}$) with the pathway topology (Fig 1B), the pathway topology is summarized by the signed adjacency matrix A (13). The assignment of positive (1) and negative (-1) weights to the edges in the pathway reflects the underlying assumption that CGSes 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 ($\hat{\boldsymbol{\mu}}$) 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 $\hat{\boldsymbol{\mu}}$ 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). 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 (13, 14). This can be

shown by analyzing the Bayes factor (15) for choosing between two probabilistic models that generated the data (***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 A). 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 B). In our testing, going beyond the one-dimensional null-subspace did not improve significantly the discriminative ability of the signature (Supplementary Materials Section C). 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.

Using the list of the 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 CGSes of genes in the pathway (Fig 1D).

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 TSes 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 (1, 16) (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 (17). 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 (17). Similar analysis of the dataset studying the response of MCF-7 cell line to amino acid starvation (18) 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 (19), 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 PAS 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. We call the decrease in the pathway consistency score that is a consequence of removing a node, the *node contribution score*. Nodes consistent with their implied role are expected to have a positive, statistically significant contribution score. S6K proteins have been mapped as either upstream negative regulators or downstream positive output of mTORC1 activity in different biological contexts(20, 21). 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 TS of their function 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 E).

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 TS 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 CGSes (KD), only the pathway topology (TP), and only classical gene list enrichment that does not utilize either pathway topology or LINCS CGSes (RS) (Fig 3A). 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 (>0.95). Fig 3B shows the comparison of AUCs for all available KEGG pathways and in Fig3C these results are further separated by the type of the KEGG pathway. The detailed information about performance for each pathway is shown in Fig S5 and Table S4 and summary results for rpROC are shown in Fig S6. In summary, ROC results indicate that: 1) PAS methodology significantly outperforms the methods based on simple enrichment analysis that does not make use of KD CGSes; 2) The use of our statistical model to identify signature genes improves significantly the performance of the method. Separating pathways by the type further reinforces the premise of this work that effect of a CP on signaling pathways activity would be particularly difficult to detect based only on the transcriptional signatures.

Cell lines as determinants of PAS precision

A comparison across different sets of analyses involving signatures from different cell lines (Fig 2D, 2E, and Fig 3D) indicates a reproducible trend of signatures derived from some cell lines being more informative than from others. The signatures constructed from MCF7, HT29, PC3 and HEKTE cell lines performed very well in all analyses. Signatures derived from NPC, HCC515 and SW480 provided virtually no signal. Some of the results seem to be clearly related to the data characteristics, such as the number of CGSes available. For example, NPC cell line has only 6 signatures (2 in amino-acid activation module) and SW480 is completely missing the amino-acid activation module of the pathway. Additional biological factors may also contribute to reduced performance. For example, HCC515 has a similar number of signatures as MCF7, but performs significantly worse.

DISCUSSION

pasLINCS methodology integrates two distinct strategies for implicating signaling pathways affected by a CP based on its TS: 1) the explicit modeling of shared expression changes implicated by the topology of the protein-protein interactions in the pathway (11); and 2) Correlating CP TS with the signatures of genetic perturbations of genes in the pathway (6). The use of CGSes provides information about the activity changes in signaling proteins not contained in the TS 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 CP.

We showed that the precision in assigning a CP to a pathway is actionable in many cases. For example, for 96% of mTOR pathway inhibitors, mTOR pathway is significantly (fpr<0.05) associated with CP TS, and the association is stronger than with any pathway not containing targeted node(s). This is a striking result when considering that in many cases experimental conditions are not necessarily optimal for detecting the signature of the change in pathway activity. For example, in some cases the dose of the CP may be too low, eliciting no response, or too high, in which case the TS may represent many off-target effects. Consequently, the actual precision of the PAS methodology may be higher than indicated in our results.

Learning MOA of chemical perturbagens based on their transcriptional signatures opens up new avenues for using connectivity map data to search for new therapies. In situations when the disease-related misregulation of signaling pathways are not clearly reflected in any available transcriptional signature, but are 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 technologiesz(6, 22, 23), combined with pasLINCS analysis may open alternative avenues for high throughput safety evaluation of commercial chemicals, pesticides, food additives/contaminants, and medical products (24, 25). Previous studies have established the potential of

assigning MOA of a chemical perturbagen based on comparison of their TS to the TSes of chemicals with known MOA (6, 26, 27). For example, the preciously derived PI3K inhibitor signature constructed from TSes of known chemical inhibitors (16) showed similar level of association with L1000 mTOR pathway inhibitors as we observed with our PAS (Fig 2D). Our methodology adds another dimension by providing direct mechanistic links 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 the static correlation structure of transcriptional signatures of pathway perturbations. The statistical learning model can be re-interpreted in the context of the regularization framework with graph kernels (28) 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 (29) in general, as well as in predicting drug targets based on transcriptional signatures (30). The signed Laplacian was also previously used to capture pathway topology in the context of constraining the covariance matrix for differential gene expression analysis (14).

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 CGSes 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 (31-35). 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 CGS 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 up 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 complemetary to the established enrichment strategies. 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.

MATERIAL AND METHODS

LINCS L1000 Consensus Gene Signatures and CP signatures

To construct CGSes and CP TSes used in the analyses, Level 4 LINCS L1000 dataset was downloaded from GEO (GSE92742). Level 5 moderated Z (MODZ) signatures for chemical perturbagens and individual shRNA knock-downs were calculated as a weighted average of Level 4 replicates (6). Only signatures designated as "gold" and generated using the "epsilon" version of L1000 probes were used in the analysis. shRNA knock-down signatures were further integrated into consensus gene signatures (CGSes) as weighted (MODZ) averages of individual shRNA signatures targeting the same gene. All LINCS CP TSes and CGSes for all cell lines 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 were excluded from analyses.

Statistical learning model for selecting signature genes and constructing PASes

Suppose that for a single pathway with M genes/proteins, $Y = (Y_{ij})$ is the $N \times M$ matrix of gene expression changes in N $(N = 978)$ measured LM genes after genetically perturbing (i.e. knocking down) M pathway genes. That is, the value y_{ij} is the L1000 CGS measuring the differential expression level of the LM gene $i = 1, \ldots, N$ after knocking down pathway gene $j = 1, \ldots, M$. The ith row in the matrix Y, $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iM})^{\mathrm{T}}$ is the expression profile of the changes in expression in LM gene i across CGSes of pathway genes. Our statistical learning model is designed to discriminate between expression profiles of L1000 LM genes that are consistent with the topology of the signaling pathway and those that are not.

Suppose that $A = (A_{ij})$ is the $M \times M$ signed adjacency matrix of the pathway network, that is

$$A_{ij} = \begin{cases} 1, & \text{if gene } i\ (j) \text{ activates gene } j\ (i); \\ -1, & \text{if gene } i\ (j) \text{ inhibits gene } j\ (i); \\ 0, & \text{if genes } i \text{ and } j \text{ are not connected.} \end{cases}$$

Let L and Σ be the signed Laplacian and the variance-covariance matrix of the Markov Random Field as defined in Fig 1. Furthermore, suppose that $Z_i = 1$ if gene i is consistent with the pathway topology and $Z_i = 0$ otherwise. Then, we model the expression profile, $\boldsymbol{Y}_i$, of gene i by $\boldsymbol{Y}_i | Z_i = 1 \sim MVN(\boldsymbol{\mu}_{i1}, \sigma^2 I)$ and $\boldsymbol{Y}_i | Z_i = 0 \sim MVN(\boldsymbol{\mu}_{i2}, \sigma^2 I)$, where $\boldsymbol{\mu}_{i1} \sim MVN(0, \Sigma)$ and $\boldsymbol{\mu}_{i2} \sim MVN(0, \tau^2 I)$. Suppose α is the probability for any one LM gene to be consistent with the pathway, then the probability distribution for the profile of LM gene i is given by the two-component mixture of multivariate Gaussian distribution:

$$\boldsymbol{Y}_i \sim \alpha N(\boldsymbol{\mu}_{i1}, \sigma^2 I) + (1-\alpha) N(\boldsymbol{\mu}_{i2}, \sigma^2 I).$$

The objective of our learning procedure is to evaluate the evidence in data ($\boldsymbol{Y}_i$) of $Z_i = 1$ vs. $Z_i = 0$. We achieve this by following the standard Bayesian learning approach and construct the consistency score based on the Bayes factor for comparing $Z_i = 1$ vs. $Z_i = 0$(15). The consistency score is defined as

$$S_i = \sum_{j=1}^{M} \frac{1}{2\sigma^2} \left[1 - \left(\lambda_j + \varepsilon\right)^2\right] \left(\boldsymbol{u}_j^{\mathrm{T}} \widehat{\boldsymbol{\mu}}\right)^2 \qquad \text{(Eq. 1)}$$

where $\widehat{\boldsymbol{\mu}}$ is the Bayesian estimate of the mean expression profile assuming $Z_i = 1$ (Fig 1b), λ_j and $\boldsymbol{u}_j$ are the jth ($j = 1,2,\cdots,M$) eigenvalue and the corresponding eigenvector of L.

The ratio σ^2/τ^2 in our model defines the relative weight assigned to the pathway topology versus the expression data, and in the derivation of the score we set this ratio to 1. Since σ^2 is a constant multiplier in the consistency score for all genes, the exact value of σ^2 does not affect the comparison among genes (we set it to 1) (Eq. 1). To regularize the Laplacian we set $\varepsilon = 0.1$. However, in our analyses we use only the first eigenvector/eigenvalue ($k = 1$) and the exact value of ε again does not affect the ranking of genes (Eq. 1). It can be shown (Supplementary Materials Section A) that S_i, which is a sum over all Laplacian eigenvectors, is proportional to the logarithm of the Bayes factor for distinguishing $Z_i = 1$ from $Z_i = 0$. It can also be shown that, when ε is small, the components of S_i with the smallest eigenvalues are expected to contribute the most to the Bayes factor (Supplementary Materials Section B). The eigenvectors corresponding to small eigenvalues also capture the desirable properties of the expression profiles in terms of their consistency with the pathway topology (14). Therefore, we use only a subset of k eigenvectors corresponding to the smallest k eigenvalues. In our empirical tests (Supplementary Figure S2), the score based on only the smallest eigenvalue ($\lambda = 0$) performed very close to that with two smallest eigenvalues, as well as that with three smallest eigenvalues, and better than that with all eigenvalues.

For a pathway consisting of multiple connected subgraphs, we calculate the consistency score for each connected subgraph, and take the total of the scores as the final score to assess the consistency of the profiled gene to the pathway topology. Finally, we select G=100 of *signature genes* with the largest consistency scores corresponds to selecting G genes with the highest posterior probabilities that their profiles are consistent with the pathway topology. In our tests (Supplementary Materials Section D) increasing the number of genes did not improve results. Using the knockdown signatures of the selected *signature genes*, we obtain the principal components of their covariance matrix, and use the eigenvector corresponding to the first principal component calculated as the pathway activity signature (PAS) (Fig 1D). The elements of a PAS corresponding to individual genes are referred to as the *activity scores*.

Node contribution score

For the purpose of assessing the contribution of individual CGSes 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 CGS of the node from the analysis. A positive node contribution score implies that the CGS 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.

Baseline methods compared with PAS methodology

We considered three baseline methods to compare with our pasLINCS methodology. The first method (KD), defines a pathway signature as the first principal component of all CGSes of the pathway genes. This method does not consider the pathway topology to identify informative genes. The second method (TP), regards a CP signature of the landmark genes whose corresponding proteins are 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 enriched pathways based on the transcriptional data alone and does not use CGSes of pathway genes. The last 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 CGSes (36).

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 (37, 38). 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.

Associating mTOR PAS with signatures of perturbations of MTOR pathway

To assess the association between a perturbation signature and mTOR PAS, signature genes were divided into two groups based on the direction of their activity scores (positive vs negative). The association between the changes in expression after perturbation for the signature genes and the positive/negative grouping was assessed by the difference in the average TS values of the two groups

(positive/negative) of signature genes. In Fig 2B and 2C, we assess the differences in associations between signatures of chemical perturbagens targeting proteins in the mTOR pathway and the control DMSO signatures. For the 5 CP signatures targeting AMPK, we flipped the signs of the differences in the average TS values to make them comparable with those of the other signatures. In Fig 2D we display the association with signatures of PI-103 treatment and the signature of amino acid starvation as the negative log of the p-value for the t-test comparing average expression changes between positive and negative signature genes.

Creation of external mTOR pathway perturbation signatures

Two microarray datasets profiling transcriptional responses of mTOR pathway perturbations(17, 18) were identified and downloaded from GEO. The first dataset (GSE12175(17)) contained two-channel microarray data profiling the response of two glioma cell lines to treatment by the dual PI3K/mTOR inhibitor, PI-103. Raw GenePix® files (.gpr) from a two-color spotted cDNA microarray were downloaded from GEO. The raw data were corrected for background intensity and quantile-normalized using the *limma* R package(39). PI-103 treated samples were compared to untreated samples using the empirical Bayes linear model as implemented in the *limma* R package to create differential expression profiles. The second dataset (GSE62673) contained Affymetrix® microarray expression data profiling the response of the MCF7 cell line starved of amino acids(18). The raw data were RMA-normalized(40) using the *affy* R package. Differential expression profiles between treated (amino acid starved) and untreated (full amino acid) samples were created using empirical Bayes linear model as implemented in the *limma* R package.

Analysis of KEGG pathways

We processed *kgml* files corresponding to 328 homo sapiens KEGG pathways(41) 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”.

DATA AVAILABILITY

Open source R package paslincs, implementing the pasLINCS methodology, is available at https://github.com/uc-bd2k/paslincs.

FUNDING

This work has been supported by the National Institutes of Health [U54HL127624]. Funding for open access charge: National Institutes of Health.

CONFLICT OF INTEREST

None

REFERENCES

1. R. A. Saxton, D. M. Sabatini, mTOR Signaling in Growth, Metabolism, and Disease. *Cell* **168**, 960-976 (2017).
2. T. Finkel, J. S. Gutkind, *Signal Transduction and Human Disease* (Wiley, 2003).
3. M. Laplante, David M. Sabatini, mTOR Signaling in Growth Control and Disease. *Cell* **149**, 274-293 (2012).
4. W. H. M. Heijne, A. S. Kienhuis, B. van Ommen, R. H. Stierum, J. P. Groten, Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. *Expert Rev Proteomics* **2**, 767-780 (2005).
5. G. T. Ankley *et al.*, Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry* **29**, 730-741 (2010).
6. A. Subramanian *et al.*, A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e1417 (2017).
7. S. Nilsson *et al.*, Mechanisms of Estrogen Action. *Physiological Reviews* **81**, 1535-1565 (2001).
8. L. Geistlinger, G. Csaba, R. Zimmer, Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set-& network-based enrichment analysis. *BMC bioinformatics* **17**, 45 (2016).
9. P. Khatri, M. Sirota, A. J. Butte, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol* **8**, e1002375 (2012).
10. A. L. Tarca, G. Bhatti, R. Romero, A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLOS ONE* **8**, e79217 (2013).
11. C. Mitrea *et al.*, Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology* **4**, 278 (2013).
12. A. H. Bild *et al.*, Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353-357 (2006).
13. J. Kunegis *et al.* (2010) Spectral analysis of signed graphs for clustering, prediction and visualization. in *Proceedings of the 2010 SIAM International Conference on Data Mining* (SIAM), pp 559-570.
14. L. Jacob, P. Neuvial, S. Dudoit, More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.* **6**, 561-600 (2012).

15. R. E. Kass, A. E. Raftery, Bayes Factors. *Journal of the American Statistical Association* **90**, 773-795 (1995).
16. Y. Zhang *et al.*, A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell* **31**, 820-832.e823 (2017).
17. S. Guillard *et al.*, Molecular pharmacology of phosphatidylinositol 3-kinase inhibition in human glioma. *Cell Cycle* **8**, 443-453 (2009).
18. X. Tang *et al.*, Comprehensive Profiling of Amino Acid Response Uncovers Unique Methionine-Deprived Response Dependent on Intact Creatine Biosynthesis. *PLOS Genetics* **11**, e1005158 (2015).
19. P. Gulhati *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* **15**, 7207-7216 (2009).
20. B. Magnuson, B. Ekim, Diane C. Fingar, Regulation and function of ribosomal protein S6 kinase (S6K) within mTOR signalling networks. *Biochemical Journal* **441**, 1-21 (2012).
21. O. J. Shah, T. Hunter, 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* **26**, 6425-6434 (2006).
22. P. R. Bushel, R. S. Paules, S. S. Auerbach, 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 (2018).
23. E. C. Bush *et al.*, PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nature Communications* **8**, 105 (2017).
24. R. J. Kavlock, C. P. Austin, R. R. Tice, Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment. *Risk Analysis* **29**, 485-487 (2009).
25. N. C. Kleinstreuer *et al.*, Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nature Biotechnology* **32**, 583 (2014).
26. M. Iwata, R. Sawada, H. Iwata, M. Kotera, Y. Yamanishi, Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Scientific Reports* **7**, 40164 (2017).
27. Z. Wang, A. Lachmann, A. B. Keenan, A. Ma'ayan, L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* **34**, 2150-2152 (2018).
28. A. J. Smola, R. Kondor, "Kernels and regularization on graphs" in Learning theory and kernel machines. (Springer, 2003), pp. 144-158.
29. L. Cowen, T. Ideker, B. J. Raphael, R. Sharan, Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551 (2017).
30. G. Laenen, L. Thorrez, D. Bornigen, Y. Moreau, Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Molecular bioSystems* **9**, 1676-1685 (2013).
31. T. Haruta *et al.*, A rapamycin-sensitive pathway down-regulates insulin signaling via phosphorylation and proteasomal degradation of insulin receptor substrate-1. *Mol Endocrinol* **14**, 783-794 (2000).
32. L. S. Harrington *et al.*, The TSC1-2 tumor suppressor controls insulin-PI3K signaling via regulation of IRS proteins. *J Cell Biol* **166**, 213-223 (2004).
33. F. Tremblay *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* **104**, 14056-14061 (2007).
34. C. Pai, C. M. Walsh, D. A. Fruman, Context-Specific Function of S6K2 in Th Cell Differentiation. *J Immunol* **197**, 3049-3058 (2016).

35. W. P. Miller, S. Ravi, T. D. Martin, S. R. Kimball, M. D. Dennis, 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* **292**, 1591-1602 (2017).
36. M. A. Newton, F. A. Quinatan, J. A. den Boon, S. Sengupta, P. Ahlquist, Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics* **1**, 85-106 (2007).
37. J. Cheng, L. Yang, V. Kumar, P. Agarwal, Systematic evaluation of connectivity map for disease indications. *Genome medicine* **6**, 540-540 (2014).
38. L. E. Dodd, M. S. Pepe, Partial AUC Estimation and Regression. *Biometrics* **59**, 614-623 (2003).
39. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43** (2015).
40. R. A. Irizarry *et al.*, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).
41. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353-D361 (2017).

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 TS. Our methods aim 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 knockdown Consensus Gene Signatures (CGS) 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) Bayesian 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 CGSes of pathway genes.

Figure 2: PAS of the mTOR signaling pathway constructed de-novo from literature. 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) The significance of individual node contribution scores across PASes derived from 12 cell line-specific CGSes; F) 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 four 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 CGSes or the pathway topology.; B) Box plots of Area Under the ROC Curve (AUC) for different methods across all KEGG pathways; C) AUC for different methods across different types of KEGG pathways; D) AUC for PASes derived from CGSes for different cell lines.

FIGURES

Figure 1

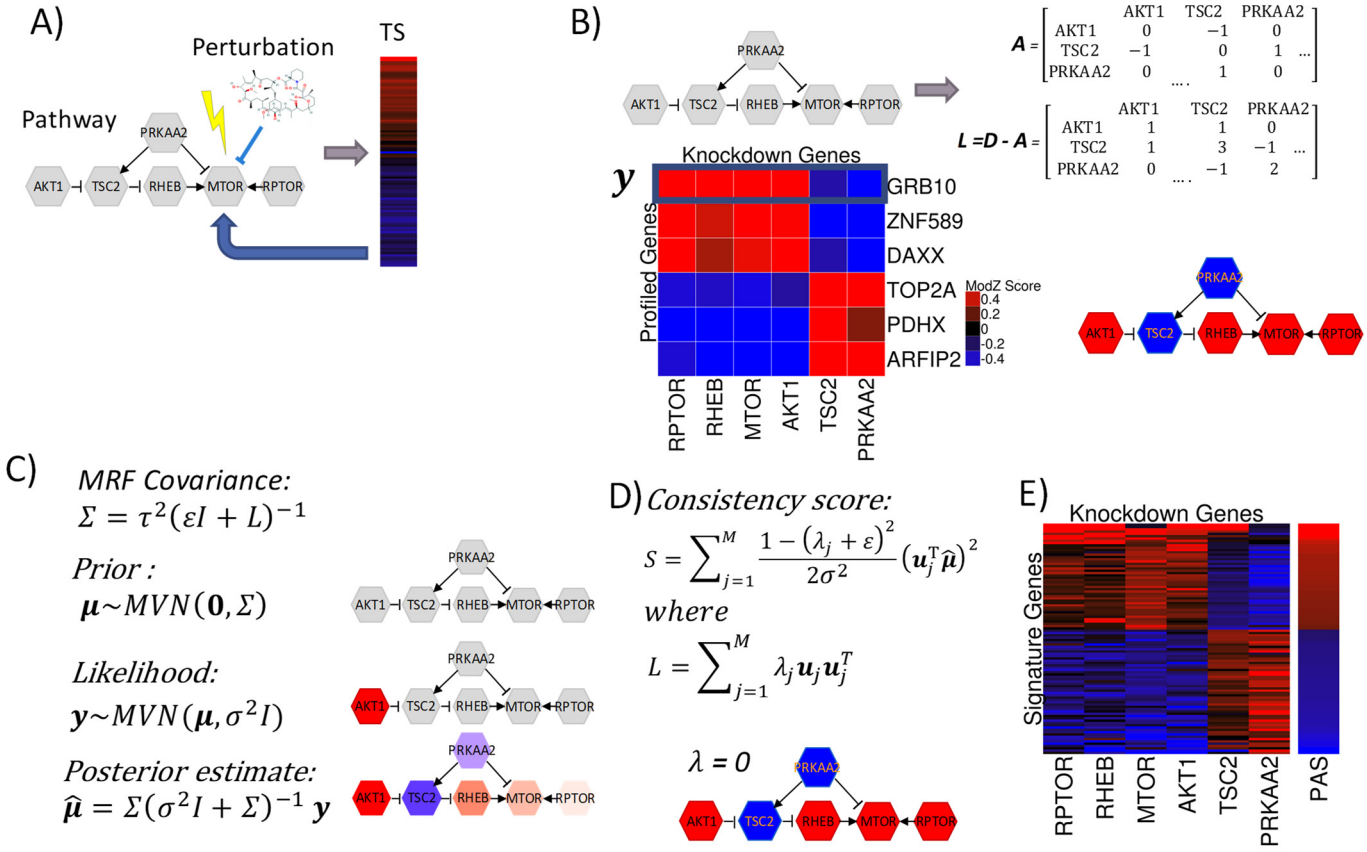
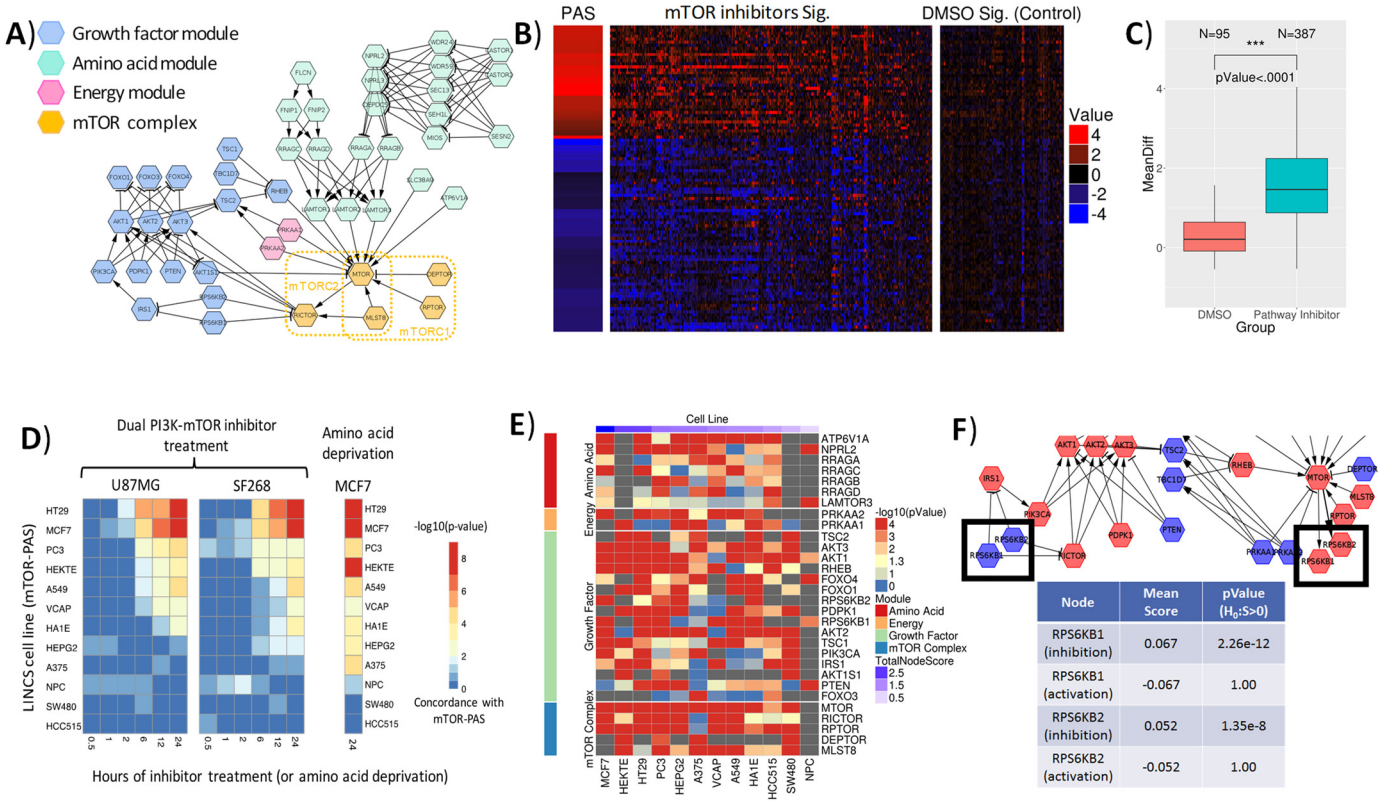
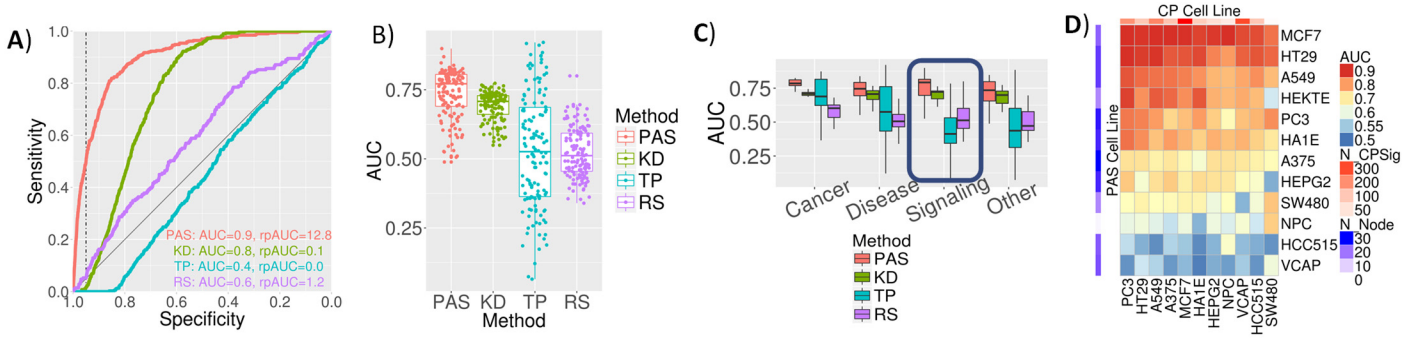


Figure 2



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)
Sensitivity
Specificity
PAS: AUC=0.9, rpAUC=12.8
KD: AUC=0.8, rpAUC=0.1
TP: AUC=0.4, rpAUC=0.0
RS: AUC=0.6, rpAUC=1.2
B)
AUC
Method
PAS KD TP RS
Method
PAS
KD
TP
RS
C)
AUC
Cancer
Disease
Signaling
Other
Method
PAS
KD
TP
RS
D)
CP Cell Line
PAS Cell Line
MCF7
HT29
A549
HEKTE
PC3
HA1E
A375
HEPG2
SW480
NPC
HCC515
VCAP
AUC
N_CPSig
N_Node