Supplementary Material

1. Gene Expression Datasets from Several Experiments

The following datasets and analysis results can be downloaded at: http://software.steenlab.org/fCI/.

1-1: RNA Microarray and TMT LC-MS/MS data of human iPSC differentiation

Directed Differentiation of Suspension ES Cell Culture and RNA Isolation: The hES cell line HUES8 was maintained in mTeSR1 media in 500mL spinner flasks. Cultures were grown in a 37°C incubator with 5% CO_2 and were stirred at a rate of 70 rpm. Cells passaged every three days by dispersing with Accutase and seeding a single cells at 0.5 million cells/mL in mTeSR with 10 μM Y27632. mTeSR media was changed 48 hours after seeding and cells were passaged 72 hours after seeding. For directed differentiation, HUES8 cells were seeded as described above into a 500mL Spinner Flask 48 hours prior to the start of the directed differentiation protocol. Cells were maintained in mTeSR media prior to the start of directed differentiation.

Day 1: S1 + 100ng/ml ActivinA (R&D Systems; 338-AC) + 3μM Chir99021 (Stemgent; 04-0004-10)). Day 2: S1 + 100ng/ml ActivinA. Days 3: S2 + 50ng/ml KGF (Peprotech; AF-100-19)). Days 5, 6: S3 + 50ng/ml KGF + 0.25μM Sant1 (Sigma; S4572) + 2μM Retinoic acid (RA) (Sigma; R2625) + 200nM LDN193189 (only Day 7) (Sigma; SML0559) + 500nM PdBU (EMD Millipore; 524390)). Days 8, 10: S3 + 200nM LDN193189 + 0.25μM Sant1. Days 11, 13: S3 + 200nM LDN193189 + 1μM Alk5i II (Axxora; ALX-270-445). Days 14, 16: S3 (no growth factors).

S1 media: MCDB131 (Cellgro; 15-100-CV) + 8mM D-(+)-Glucose (Sigma; G7528) + 2.46g/L NaHCO3 (Sigma; S3817) + 2% FAF-BSA (Proliant; 68700) + ITS-X (Invitrogen; 51500056) 1:50.000 + 2mM Glutamax (Invitrogen; 35050079) + 1% Pen/Strep (Cellgro; 30-002-CI).

S2 media: MCDB131 + 8mM D-Glucose + 1.23g/L NaHCO3 + 2% FAF-BSA + ITS-X 1:50.000 + 2mMl Glutamax + 1% Pen/Strep.

S3 media: MCDB131 + 8mM D-Glucose + 1.23g/L NaHCO3 + 2% FAF-BSA + ITS-X 1:200 + 2mMl Glutamax + 1% Pen/Strep.

At given time points in the differentiation process, one million differentiating cells were harvested from the differentiation cultures. Cells were washed twice in PBS and RNA was isolated using the RNeasy Mini Kit (Qiagen). Double stranded cDNA was generated by reverse transcription from 100ng of total RNA using the Illumina TotalPrep RNA Amplification Kit (Life Technologies).

Microarray Data Collection: At least 750ng cRNA per sample was hybridized to Human HT-12 Expression BeadChips (Illumina) using the Whole-Genome Expression Direct Hybridization kit (Illumina). Chips were scanned on the Illumina Beadstation 500. Raw data was adjusted by background subtraction and rank-invariant normalization (GenomeStudio software, Illumina).

Sample Preparation for Proteomics: Five different time points during the differentiation of human iPSCs into β-cells and one sample from postmortem crude pancreatic islet tissue were analyzed in the quantitative proteomics analysis using isobaric labeling, namely TMT (Tandem Mass Tags, Thermofisher). Two biological replicates of these experiments were performed. Cells from the different time points were independently resuspended in lysis buffer from the Mammalian protein prep kit (Qiagen, MD) supplemented with protease and phosphatase inhibitors and lysed using sonication. Protein concentration was determined using Pierce BCA Protein Assay kit (Thermo Fisher Scientific, Rockford, IL). Equal amounts of protein (100µg) for each sample was taken and proteins were precipitated using methanol/chloroform. Extracted proteins were reduced with 10mM DTT, alkylated with 1% acrylamide and subjected to overnight trypsin digestion at 37°C using porcine trypsin (Promega, Madison, WI) with a protein:trypsin ratio of 50:1.

TMT Labeling, LC-MS/MS Data Collection and Analysis: Following digestion, samples were labeled separately using isobaric TMT labels (Thermo Fisher Scientific, Rockford, IL) following manufacturer's recommendations. Once the reactions were quenched with 5% hydroxylamine, the samples were combined into one and desalted using Oasis columns (Waters Corp., Milford, MA). Combined sample was further fractionated using off-gel pH 3-10 immobiline dry strips (GE Healthcare, Pittsburg, PA) into 24 fractions. Fractions were desalted using Nestgroup c18 tips (Southborough, MA). Fractions were run on a Q Exactive mass spectrometer (Thermo Fisher Scientific) coupled with Eksigent LC system (AB Sciex, Framingham, MA) over a 60 minute gradient. Collected data files were converted into .mgf files using a modified script with MSConvert (ProteoWizard(1)). Database search was carried out using Mascot(2) (www.matrixscience.com, Matrix Science, Boston, MA). Search results from 24 runs were imported into Scaffold(3) (Proteome Software, Portland, OR). Complete spectrum report resulting in 232,555 PSMs, 51,154 unique peptides and 4,981 proteins at 5% peptide FDR was exported from Scaffold for further analysis. Clustering analysis was performed using GProX software(4).

1-2: RNA-Seq and TMT LC-MS/MS data from Mouse Cortical Neurons

This is an in-house proteogenomic data from mouse cortical neurons for which gene expression levels were quantified by RNA-Seq and TMT LC-MS/MS 6-plex labeling

experiments respectively. Each experiment contained two conditions with two replicates per condition.

Mouse models: PTENf/f mice were crossed with a PTENf/f/synapsin-Cre mice in which Cre is expressed under the neuronal promotor synapsin at E15. In the offspring, PTEN was deleted specifically in neurons thus inducing the activation of the mTOR pathway. To downregulate mTOR activity, wild type C57Bl6 pregnant mice were injected with Rapamycin (LC Laboratories). Rapamycin was first dissolved at 20 mg/ml in ethanol and before each administration, rapamycin was diluted in 5% Tween 80, 5% polyethylene glycol 400 (0.5–1.5 mg/ml)(5). Rapamycin was injected intraperitoneally to pregnant females using a dosage of 6 mg/kg at an of the estimated embryonic stage E15, every day, until E18. Wild type C57Bl6 pregnant mice were used as controls. All experimental procedures were performed in compliance with animal protocols approved by the Institutional Animal Care and Use Committee at Children's Hospital, Boston. AAV preparation was described previously

Preparation of Cortical neurons: For each condition, pregnant mice were sacrificed to extract embryos at an estimated age of E18. The embryos' right and left cortex were dissected out and meninges were removed. Cortices were snap frozen until further processing was initiated. The experiments were repeated twice to generate two biological replicates. Each biological replicate corresponds to a pooled sample of at least 3 different animals. Each biological replicate was separate in two part, one for RNA extraction and one for protein purification.

mRNA extraction: Trizol® was used to extract total RNA from a fraction of each pooled sample. The extracted RNAs were resuspended in RNAse free water and concentration and quality of the RNA was assessed using a Nanodrop and Bioanalyzer respectively.

Protein extraction, TMT labelling and peptide fractionation: Protein extraction was performed on the remaining fraction of the pooled cortical samples. Tissues were dissociated in lysis buffer (250 mM Sucrose, 250 mM KCl,5 mM $MgCl_2$,50 mM Tris-HCl pH7.4, 0.7% NP40) using dounce homogenizer. After complete dissociation, homogenates were subjected to subcellular fractionation according to previous protocols (6). concentrations were estimated using a BCA assay kit (23225, Thermo Scientific). Protein precipitation and digestion was carried out as described by Winter et al., 2011. Briefly, 100 ug of protein was precipitated from each sample using 1ml of ice cold chloroform/methanol. The pellets were re-dissolved in 0.1 % RapiGest (186001861, Waters) in 100 mM TEAB (triethyl ammonium bicarbonate) TEAB, 17902, Sigma), and incubated at 37 °C for 15 min. Trypsin (V5280 Promega) was added to the samples and incubated at 37 °C for 45 min to dissolve the pellet. Samples were reduced with tris (20 carboxyethyl) phosphine (TCEP) and alkylation with a 0.05 M iodoacetamide solution in the dark at room temperature (23 ℃) for 20 minutes. Further

trypsin was added to a final enzyme to protein ratio of 1:100 and the mixture was incubated at 37 °C overnight. Peptide samples from each experimental condition were acidified using 5 ul trifluoroacetic acid (TFA) and incubated for 45 min at 37 °C in order to precipitate the RapiGest followed by centrifugation for 30 min at 20,000 g. Clear supernatants were desalted using Oasis HLB cartridges (186006339 Waters). Briefly, the columns were washed twice with 70 % ACN 0.1 % Formic acid (FA) and twice with 0.1 % FA. The sample pool was passed twice through each individual column, washed with four times with 0.1 % FA and eluted twice with 30 % ACN 0.1 % FA, twice with 50 %ACN 0.1 % FA and twice with 70 % ACN 0.1 % FA. Individual eluated fractions were pooled and samples dried in a vacuum centrifuge. Dried samples were re-suspended in 0.1M TEAB. For each sample, peptides were labeled with one of the 6 TMT labels (PI-90064, Thermo Fischer Scientific) for 3 hours at room temperature (RT) following the manufacturer's protocol. The samples were combined, partially dried using a vacuum centrifuge and desalted using Oasis HLB cartridges as described above. The dried peptides were resuspended in ampholyte solution (pI 3-10) and fractionated overnight into 24 fractions based on their isoelectric point using an OFFGEL fractionator (Agilent) according to the manufacturer's instructions. The fractions were desalted and analyzed using LC-MS/MS.

LC-MS/MS analysis on orbitrap classic**.** 10ug of peptide samples were loaded directly onto in house packed reverse phase columns using 5 um, 200Å particles (magic C18, Michrom) and PicoTip Emmiters (New Objective) with an autosampler / nanoLC setup (2D nanoLC, Eksigent) at a flow rate of 1 µl/min. After loading the column was washed for 5 min at 1 µl/min at 99 %A (water with 0.2 % FA) 1 %B (acetonitrile with 0.2 % FA) followed by elution with a linear gradient from 1 % B to 35 % B at 400 nl/min in 60 min. Peptides eluting from the column were ionized in the positive ion mode and the 6 most abundant ions were fragmented in the PQD-mode to allow for the detection of low mass range reporter ions. Briefly, the LTQ-Orbitrap was run in positive ion mode. Full scans were carried out with a scan range of 395 to 1200 m/z. Normalized collision energy of 35 was used to activate both the reporter ions and parent ions for fragmentation. Scans were carried out with an activation time of 30ms. The isolation window was set to 1.0 m/z.

Database search. The proprietary Thermo Scientific .raw files were converted into .mgf files and MS/MS data was queried against the mouse IPI (version 3.68) protein sequence database, containing common contaminations and concatenated to its decoy version, using MASCOT v2.1 (Matrix Science) with MS peptide tolerance of 10ppm. TMT peptides were searched with enzyme specificity trypsin and TMT-6plex (N-termini and Lys) as variable modifications (oxidation and deamination).

TMT quantification. The TMT reporter ion intensities in MS/MS spectra (m/z 126.12, 127.13, 128.13,129.14,130.14 and 131.14) from raw data were used for quantification. Labeling bias was tested by assessing the $\log_2$ intensities from all the channels. In addition TMT labeling efficiency was evaluated to be close of 99% of all unique and high-confidence peptides (labeled on N-terminal and internal Lysine residues) with TMT reporter ions. All normalization is done on log-transformed TMT intensities among just the active TMT channels. The technical replicates are first separated, then horizontally concatenated, and then all are normalized together. Normalization involves three steps. (1) first the median of all the data is calculated; then (2) the medians of each replicate/channel (RC) is calculated, and (3) all TMT values in each RC are adjusted by the overall median minus each RC's median. This process ensures that each replicate's reporter ion intensities get equal weight, and are normalized to every other replicate's contribution. Following normalization, the RC separation process is reversed, and the replicates are restacked into the original six columns. *All* ratios are calculated at the PSM level, then independently aggregated to the peptide level and the protein level (with outliers optionally removed). To incorpore biological replicate, median PSM-level TMT intensities within each replicate group are first calculated, then separately aggregated into peptides and proteins. The ratios and statistics are then calculated group-to-group. Significance is calculated using the non-parametric, rank-based Wilcoxon test. This is preferred over the student's t-test, which is not appropriate for much of these data, because the t-test assumes a normal distribution for the PSM ratios. This assumption does not hold for most TMT data. In contrast, the Wilcoxon test does not require normality. It yields more conservative but more robust results.

1-3: Rat RNA-Seq Data Collected at Two Developmental Stages

In this dataset, both the control and case samples were generated from rats at the age of two weeks and two months(7) with two replicates per condition. We conducted two different analysis strategies with this dataset. First, we performed fCI analysis on each of the two time points separately as previously described. For both time points, there was one unique empirical null distribution and four case-control distributions indicated by replicate number. Second, we combined the two time points together and produced a bivariate dataset, representing information from the two time points simultaneously (*supplementary pseudo-code)*. In this situation, the bivariate empirical null distribution is a two-dimensional matrix from both time points jointly, and the case-control distribution was formed from the pairwise combinations between treatment and control on the two time points simultaneously.

1-4: Spike-in DNA Microarray Data

This dataset contained 18 Affymetrix microarrays with 9 replicates from both the control and the case conditions(8). There were three pooled samples (biological replicates) each containing three technical replicates in both the control and experimental groups. For each replicate dataset, there were 5,352 cDNAs. The spiked-in DEGs had fold changes ranging from 1.2 to 4 fold. The ratios between the mean of control replicates and the mean of case replicates were computed, showing that 952 cDNAS were up-regulated (ratio greater than 1.2 fold) and 651 cDNAs were down-regulated (ratio less than 0.8333) (see supplementary material 1-4). For each pooled sample, we constructed three empirical null (see Fig 1.c and supplementary pseudo-code) distributions and nine case-control distributions, forming a total of 27 combinations per pool sample. A total of 81 fCI combinations were constructed from all three pool samples. Subsequently, we performed 81 individual fCI analysis and generated the same number of result sets. As each misregulated target could be found from one to all of 81 fCI combinations, we produced 81 lists of counted targets (total number of times a target was found in all tested combinations), and each list contained misregulated targets found at least 1 out of 81 pairwise comparisons.

1-5: Taqman Assays RNA Expression Data

This dataset had 1,043 genes of the original 1,044 genes, with one gene removed because one of its replicates was expressed ~19,000 folds smaller than the other three control replicates. A fold change of 1.4 was used as the cutoff to declare if a gene of interest was significantly changed or not. This RNA dataset contained two conditions, with four replicates per condition. Overall, there were six non-redundant empirical null and 16 case-control ratio distributions (supplementary pseudo-code), forming a total of 96 replicate combinations. Subsequently, we performed fCI analysis individually on all 96 combinations. Similarly, each misregulated target could be found from once to all of fCI analysis. We produced 96 lists of counted target, and each list contained the targets found at least N times(with N ranging from 1 to 96) times out of all pairwise comparisons.

1-6: mRNA and protein levels from bone marrow derived dendritic cells collected at multiple time points.

This dataset contained mRNA and protein expression from DCs which were stimulated for various time points with LPS or Mock (no stimulation) (9). SILAC (stable isotope labeling by amino acids in cell culture) approach was used for obtaining protein levels. In addition, protein levels from SILAC were divided into two measurements, heavy (H/L) and medium (M/L) channels respectively. For SILAC proteomics experiment, protein expression levels were collected from 0h, 0.5h, 1h, 2h, 3h, 4h, 5h, 6h, 9h, 12h and also 24h. For RNA-Seq experiment, mRNA levels were collected from 0h, 1h, 2h, 4h, 6h, 9h, 12h. In this analysis, 0h was used as the reference time points, and only time points 1h, 2h, 4h, 6h, 9h, 12h were used as they appeared in both proteomics and transcriptomics datasets. Because each protein or mRNA level was recorded in two replicates only, it's not feasible to estimate the standard deviation of the records. To remove the impact of

large technical errors which couldn't be estimated in this study, we removed the genes whose technical difference is more than two fold large.

1-7: RNA-Seq Data from Single Cells Cultured in Serum and Two-Inhibitor Medium

This dataset contained RNA-Seq data from 80 mouse embryonic stem cells (mESCs) cultured in serum and in two-inhibitor medium respectively(10). The mESCs cultured in two-inhibitor medium will show less gene expression noises because of their inherent naive pluripotency(10). Only genes with an average of five transcripts or more were considered in the subsequent analysis(10). In the subsequent analysis, fCI performed differential expression analysis on each gene by comparing cells from control (i.e. the two-inhibitor condition) with case (i.e. the serum condition). Different from previous work, gene expression values from all cells of a given condition were directly used to construct the fCI distribution rather than using expression fold changes (ratios) calculated from two samples. This analysis was repeated for all genes in mESCs cultured in serum medium as the control (empirical null) distribution. Similarly, we performed analysis by comparing gene expression values from mESCs cultured in two-inhibitor as the control group and mESCs cultured in serum medium as the case group. If gene expression values are more variable (longer tails are observed in the fCI distribution) i.e. for the mESCs cultured in serum medium, the fCI distribution of expression values will be wider than the two-inhibitor condition for most genes.

Because RNA-Seq data from single cells exhibit large variability as a result of amplification bias and technical noise, the changes in gene expression values for a very small number of cells will result in artifacts due to sampling and technical errors. Several models have been proposed to account for the sampling/technical noise(10). Results showed these genes exhibit an inherent variability of 10-15% due to noise between the models and experimental results. Therefore, only genes with changes greater than the noise were considered to be variable genes(10). In this analysis, we had two conditions where each condition has measurements for 80 cells, only genes that showed differences between conditions and were above the noise were identified as variable genes.

1-8: RNA-Seq data from Fly Embryos

This is a RNA-Seq dataset from fly embryos where two replicates were generated for two conditions. Only a single gene was engineered to be over-expressed. The data were created by Bartek Wilczynski, Ya-Hsin Liu, Nicolas Delhomme and Eileen E. M. Furlong from the Furlong laboratory and was made available from Simon Anders(11).

2. Sample Replicate Normalization

The datasets described in supplementary material session 1-4 (Microarray dataset), section 1-5 (RNA dataset) were directly downloaded from the publications and these data were already properly normalized. Therefore, these datasets were used without any further processing steps.

For the remaining RNA-Seq datasets (see supplementary material session 1-1, 1-2, 1-3, 1-6, 1-7 and 1-8), genes containing zero read counts in any sample were removed for subsequent analysis. We applied the trimmed sum normalization(12) to the time-series data (see supplementary material session 1-3) since the experiment replicates were obtained by the same protocol and an equal library size was expected within each experimental condition. In particular, we normalized each replicate to have the same library size (total read count) after the 5% lowly expressed and the 5% highly expressed genes were removed from each replicate(13).

The normalization approach for the proteogenomic data (see supplementary material session 1-2) was different because the RNA-Seq and TMT labeling LC-MS/MS data had distinct measures of gene expression levels. Therefore, we hypothesized that the genes whose expression was the least affected by the experiment (in the forms of both RNA and protein) should have nearly identical expression levels across different replicates, in both RNA-Seq and proteomic datasets. These unchanged genes will be centered at zero in the logarithm transformed control-control or case-control ratio distributions. Therefore, we normalized proteogenomic dataset's fCI pairwise ratio distribution (Gaussian kernel density approximation) to be centered at zero (14).

3. fCI Model Tuning and Performance Evaluation

fCI can be used to evaluate whether there exists a strong experimental effect or not using the divergence value and spread of control-control replicate distribution. Among the six previous described supplementary datasets, fCI observed that the datasets in Section 1-4, 1-5 and 1-7 were highly reproducible and the divergence value was less than 10^{-2} (per thousand genes). Therefore, fCI used a target coverage (see Fig.1.d-e) of approximately 80% and achieved very stable DEG target lists among pairwise analysis. However, the proteomics datasets 1-2 and 1-6 contained greater variability among replicates and a target coverage of approximately 50% was used instead in compensation for the lost of replicate reproducibility.

1. Chambers,M.C., Maclean,B., Burke,R., Amodei,D., Ruderman,D.L., Neumann,S., Gatto,L., Fischer,B., Pratt,B., Egertson,J., *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

2. Perkins,D.N., Pappin,D.J., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

3. Searle,B.C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, **10**, 1265–1269.

4. Rigbolt,K.T.G., Vanselow,J.T. and Blagoev,B. (2011) GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Mol. Cell. Proteomics MCP*, **10**, O110.007450.

5. Park,K.K., Liu,K., Hu,Y., Smith,P.D., Wang,C., Cai,B., Xu,B., Connolly,L., Kramvis,I., Sahin,M., *et al.* (2008) Promoting axon regeneration in the adult CNS by modulation of the PTEN/mTOR pathway. *Science*, **322**, 963–966.

6. Belin,S., Hacot,S., Daudignon,L., Therizols,G., Pourpe,S., Mertani,H.C., Rosa-Calatrava,M. and Diaz,J.-J. (2010) Purification of ribosomes from human cell lines. *Curr. Protoc. Cell Biol. Editor. Board Juan Bonifacino Al*, **Chapter 3**, Unit 3.40.

7. Hammer,P., Banck,M.S., Amberg,R., Wang,C., Petznick,G., Luo,S., Khrebtukova,I., Schroth,G.P., Beyerlein,P. and Beutler,A.S. (2010) mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res.*, **20**, 847–860.

8. Zhu,Q., Miecznikowski,J.C. and Halfon,M.S. (2010) Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset. *BMC Bioinformatics*, **11**, 285.

9. Jovanovic,M., Rooney,M.S., Mertins,P., Przybylski,D., Chevrier,N., Satija,R., Rodriguez,E.H., Fields,A.P., Schwartz,S., Raychowdhury,R., *et al.* (2015) Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, **347**, 1259038.

10. Grün,D., Kester,L. and van Oudenaarden,A. (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

11. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

12. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

13. Rapaport,F., Khanin,R., Liang,Y., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D. and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.

14. Parzen,E. (1962) On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.*, **33**, 1065–1076.

Supplementary Figure 1. Illustration of control-control empirical null and case-control fCI combinations. The expression levels of a total of n (n=14202) genes were recorded at c (c=2) time points with m (m=2) replicates per gene for the control and case studies. For each individual analysis, fCI will pick up the only empirical null distribution, and compute its divergence with one of the 16 case-control combinations. Therefore, a total of 16 fCI computations will be performed. Each gene will appear from 0 to 16 times as a DEG out of the 16 analysis.

Supplementary Figure 2. Snapshot of the 3D Gaussian kernel density plot of the proteogenomics data (see supplementary material 1-1)**.** The proteogenomics data are composed of ~3000 genes with expression measured in both transcript (microarray) and protein (LC-MS/MS) levels. The left panel showed the kernel distribution of logarithm transformed ratios generated in control-control (empirical null) data using the control replicate 1 and 2 in both transcriptomics and proteomics data, and the right panel showed the distribution from case-control data (using the first replicate of control and experimental sample in both transcriptomics and proteomics data) before differentially expressed genes were removed.

Supplementary Figure 3. Three-way Venn diagram for the fCI differentially expressed genes (DEGs). The DEGs indicated were from Proteomics dataset (orange), RNA-Seq dataset (red) and the joint Proteogenomic dataset (green) respectively, using our in-house dataset (supplementary material 1-2).

Supplementary Figure 4. ROC curve for spike-in microarray data (see supplementary material 1-4). A total of 81 fCI analyses were performed and the resulting DEG detection frequencies were compared with the known DEG labels to obtain the ROC curve.

Supplementary Figure 5. Differential expression analysis using qRT-PCR validated gene set (see supplementary material 1-5). **(a)** ROC analysis was performed using a qRT-PCR $\log_2$expression change threshold of 0.5. The results show a significant advantage for fCI in detection accuracy. **(b)** Histogram of fCI detection frequency based on genes beloing to DEGs and Non-DEGs (non differentially expressed genes). **(c)** At increasing $\log_2$ expression ratios (incremented by 0.1), representing a more stringent cutoff for differential expression, the performances of the fCI methods gradually reduce but it's still has the largest Area under the ROC until it reaches a log fold change of around 2.

Supplementary Figure 6. Analysis of protein and mRNA that changes the most in time-course data using fCI. Protein levels were measured in two treatment conditions (LPS and Mock) and two data recording methods (H/L and M/L) respectively. mRNA levels were were measured in two treatment conditions (LPS and Mock) respectively (see supplementary material 1-6). Both protein and mRNA levels were recorded at 0h,1h, 2h, 4h, 9h and 12h. At each time point, fCI determined whether the given gene is differentially expressed or not compared to reference time point 0h. If no significance was found, a fold change of 0 was assigned. Otherwise, the ratio will be reported for significant changed time points. Effect of gene regulation with respect to the 6 time points were shown on the top-10 genes that have the most fold changes.

Supplementary Figure 7 . **Analysis of Single Cell RNA-Seq dataset** (supplementary material *1-7*) **using fCI**. **(a)** The Control and Case distributions (control and case replicates from gene *Prmt5* was

chosen from 2i and serum treatment conditions respectively) (supplementary material *1-7*). **(b)** The change of fCI divergence values (estimated by cross entropy) between the two distributions according to increasing fold change cutoffs. **(c)** The Control and Case distributions (control and case replicates was chosen from serum and 2i treatment conditions respectively. **(d)** The change of fCI divergence values between the two distributions according to increasing fold change cutoffs.

Supplementary Figure 8 Kernel density plot of control-control (blue) and case-control (red) ratio distributions from the fly RNA-Seq data (supplementary material 1-*8*)**.** The control-control distribution was slightly shifted toward upper y axis to make it separated from the case-control ratio distribution**.**

Supplementary Figure 9. Construction of type-I error rate based on fCI's empirical null distribution. The null hypothesis and alternative hypothesis were constructed from spike-in RNA data (supplementary material 1-4). Idealy, if the cutoff fold change is chosen without error (no false positives), we should not observe any gene in the control-control ratios (empirical null distribution) that has a fold change larger than the chosen cutoff. The proportion of such genes in the empirical null distribution is equivalent to type I error rate (incorrect rejection of a true null hypothesis).

Supplementary Figure 10. False Discovery Rate estimation using fCI's permutation analyses. We generated 100 valid fCI combinations (the target database), and 100 random fCI combinations (the decoy database) by permuting the control and experimental replicates using data from supplementary material 1-5. To achieve this, we randomly permute the replicates between control and experimental samples (i.e. we form a empirical null distribution by computing the ratio between the 2nd control replicate and the 1st experimental replicate), and then we computed the 'DEGs' from this permutated fCI combination. The FDR is computed as the number of total decoy and false hits divided by all fCI predicted DEGs.

Gene					
		Control1	Control2	Case1	Case2
1	Time Point 1	X_{111}	X_{122}	Y_{111}	Y_{122}
…		…	…	…	…
n		X_{11n}	X_{12n}	Y_{11n}	Y_{12n}
1	Time Point 2	X_{211}	X_{222}	Y_{211}	Y_{222}
…		…	..	…	…
n		X_{21n}	X_{22n}	Y_{21n}	Y_{22n}
		$\vec{X}_1$	$\vec{X}_2$	$\vec{Y}_1$	$\vec{Y}_2$

$$\vec{X}_i = \begin{bmatrix} x_{1i1} & x_{1i2} & \cdots & x_{1in} \\ x_{2i1} & x_{2i2} & \cdots & x_{2in} \\ \cdots & \cdots & \cdots & \cdots \\ x_{ci1} & x_{ci2} & \cdots & x_{cin} \end{bmatrix}$$

x_{kij}

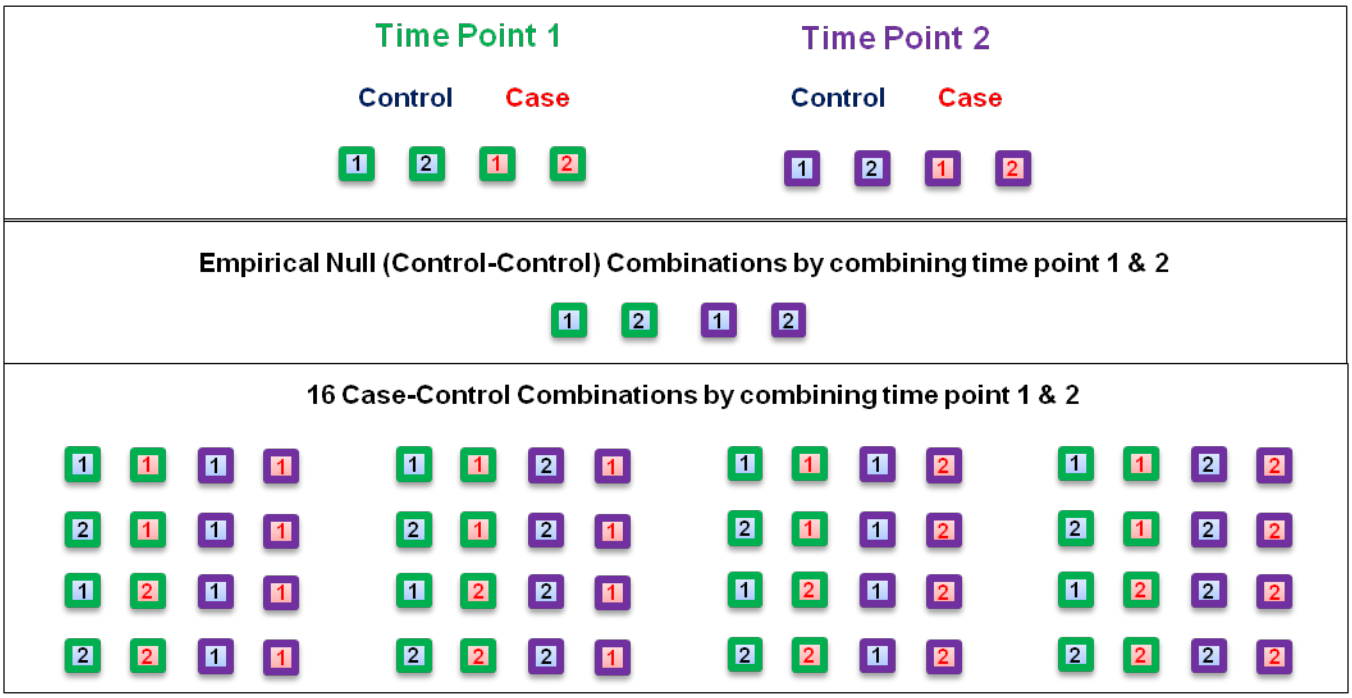
i-th control in the *j*-th gene on the *k*-th dimension

$$\vec{Y}_i = \begin{bmatrix} y_{1i1} & y_{1i2} & \cdots & y_{1in} \\ y_{2i1} & y_{2i2} & \cdots & y_{2in} \\ \cdots & \cdots & \cdots & \cdots \\ y_{ci1} & y_{ci2} & \cdots & y_{cin} \end{bmatrix}$$

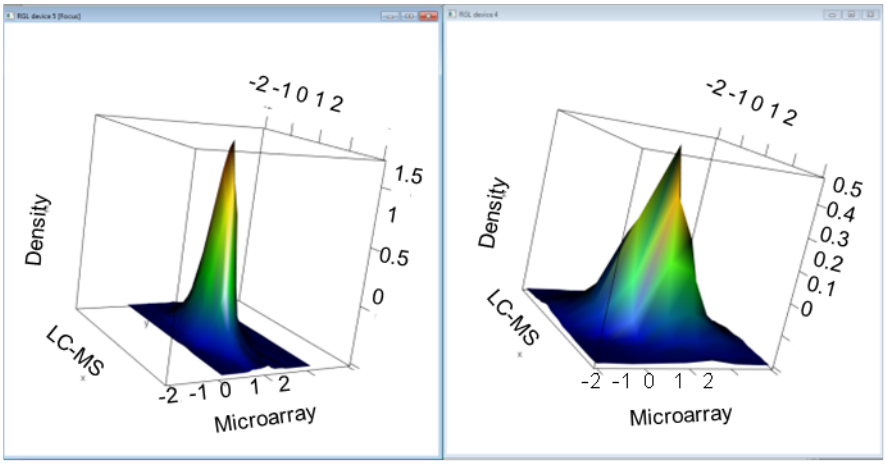
y_{kij}

i-th case in the *j*-th gene on the *k*-th dimension

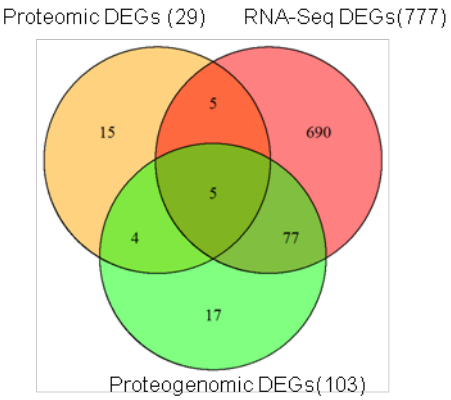
$$k = 1,2,\ldots,c; i = 1,2,\ldots,m; j = 1,2,\ldots,n$$



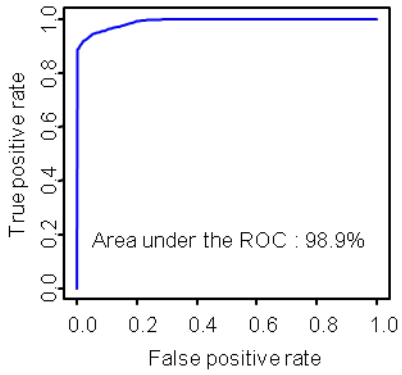
Supplementary Figure 1. Illustration of control-control empirical null and case-control fCI combinations. The expression levels of a total of n (n=14202) genes were recorded at c (c=2) time points with m (m=2) replicates per gene for the control and case studies. For each individual analysis, fCI will pick up the only empirical null distribution, and compute its divergence with one of the 16 case-control combinations. Therefore, a total of 16 fCI computations will be performed. Each gene will appear from 0 to 16 times as a DEG out of the 16 analysis.



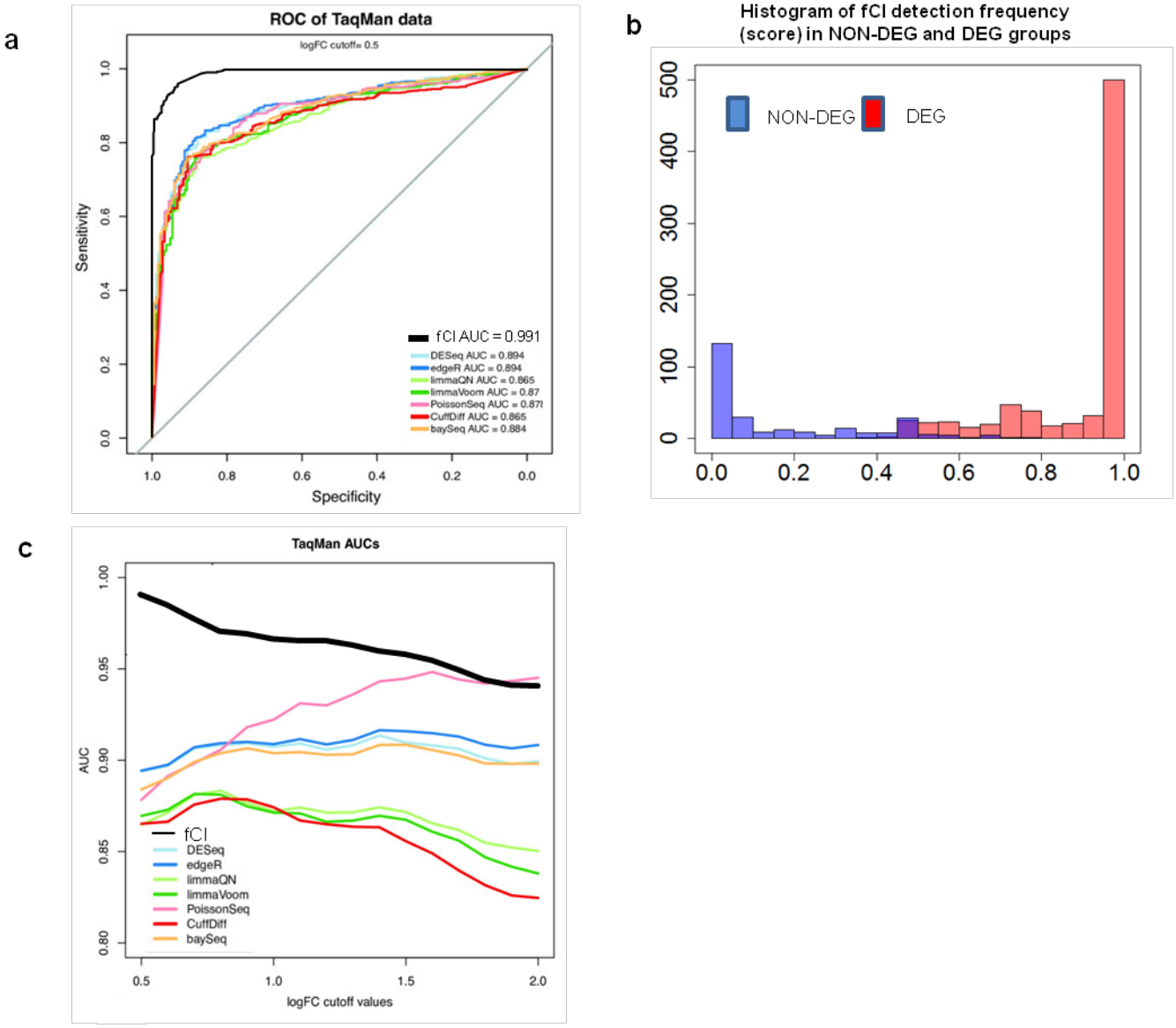
Supplementary Figure 2. Snapshot of the 3D Gaussian kernel density plot of the proteogenomics data. The proteogenomics data are composed of ~3000 genes with expression measured in both transcript (microarray) and protein (LC-MS/MS) levels. The left panel showed the kernel distribution of logarithm transformed ratios generated in control-control (empirical null) data using the control replicate 1 and 2 in both transcriptomics and proteomics data, and the right panel showed the distribution from case-control data (using the first replicate of control and experimental sample in both transcriptomics and proteomics data) before differentially expressed genes were removed.



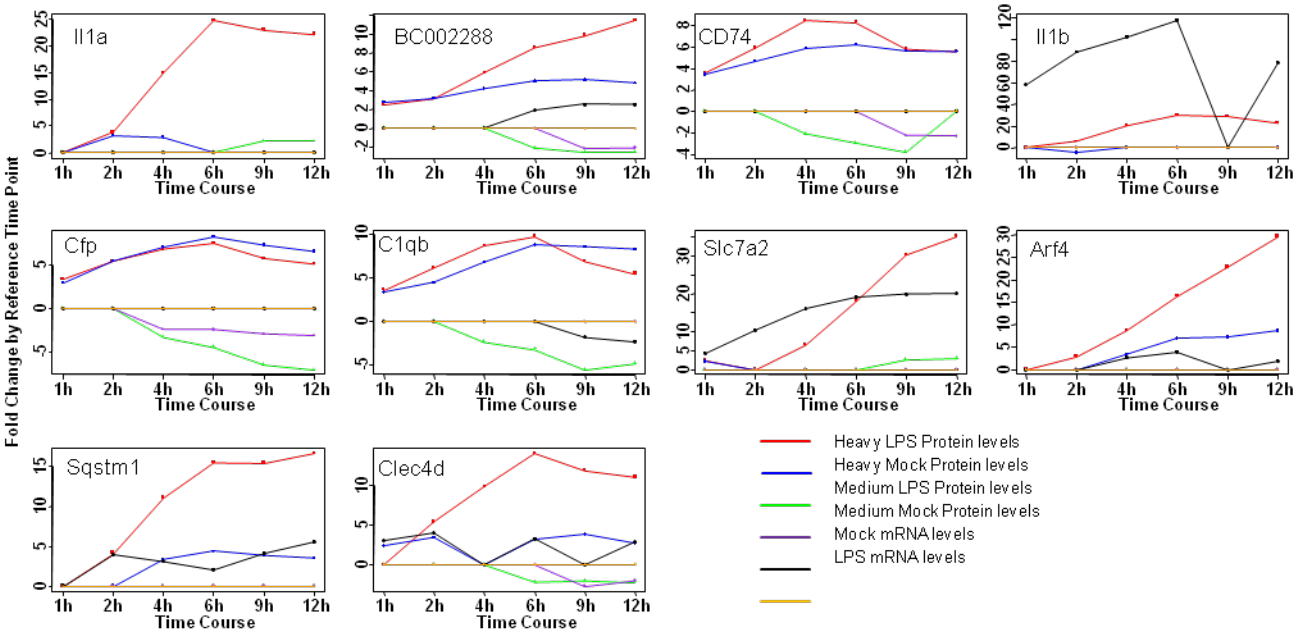
Supplementary Figure 3. 3-way Venn diagram for the fCI differentially expressed genes (DEGs) found on Proteomics dataset (orange), RNA-Seq dataset (red) and the joint Proteogenomic dataset (green) respectively, using our in-house dataset (supplementary material 1-2).



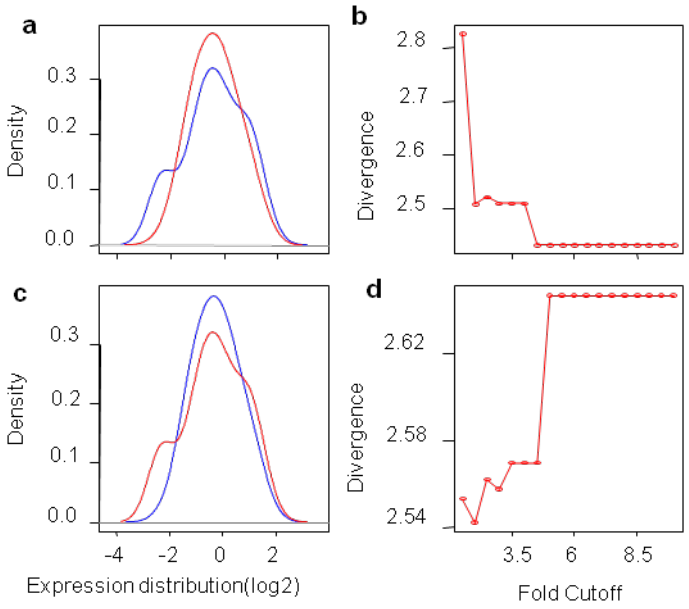
Supplementary Figure 4 ROC curve for spike-in microarray data (see supplementary material 1.3). A total of 81 fCI analyses were performed and the resulting DEG detection frequencies were compared with the known DEG labels to obtain the ROC curve.



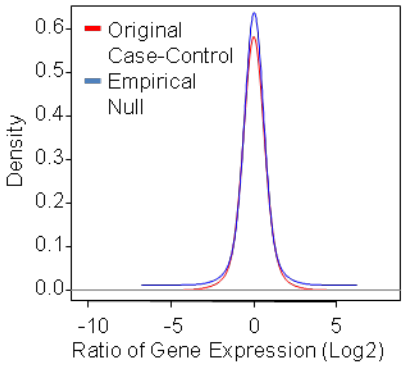
Supplementary Figure 5. Differential expression analysis using qRT-PCR validated gene set. **(a)** ROC analysis was performed using a qRT-PCR $\log_2$expression change threshold of 0.5. The results show a significant advantage for fCI in detection accuracy. **(b)** Histogram of fCI detection frequency based on genes beloing to DEGs and Non-DEGs (non differentially expressed genes). **(c)** At increasing $\log_2$ expression ratios (incremented by 0.1), representing a more stringent cutoff for differential expression, the performances of the fCI methods gradually reduce but it's still has the largest Area under the ROC until it reaches a log fold change of around 2.



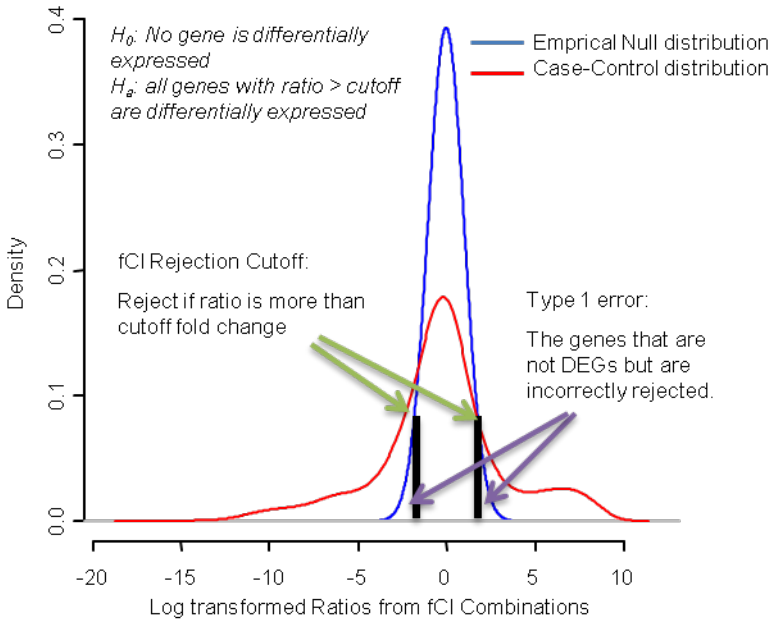
Supplementary Figure 6. Analysis of protein and mRNA that changes the most in time-course data using fCI. Protein levels were measured in two treatment conditions (LPS and Mock) and two data recording methods (H/L and M/L) respectively. mRNA levels were were measured in two treatment conditions (LPS and Mock) respectively. Both protein and mRNA levels were recorded at 0h, 1h, 2h, 4h, 9h and 12h. At each time point, fCI determined whether the given gene is differentially expressed or not compared to reference time point 0h. If no significance was found, a fold change of 0 was assigned. Otherwise, the ratio will be reported for significant changed time points. Effect of gene regulation with respect to the 6 time points were shown on the top-10 genes that have the most fold changes.



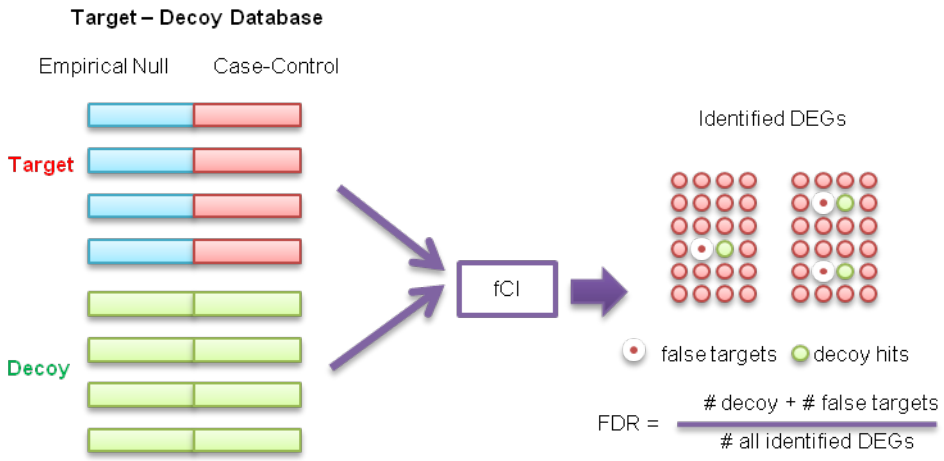
Supplementary Figure 7 . Analysis of Single Cell RNA-Seq dataset (supplementary material *1-7*) using fCI. **(a)** The Control and Case distributions (control and case replicates from gene *Prmt5* was chosen from 2i and serum treatment conditions respectively) (supplementary material *1-7*). **(b)** The change of fCI divergence values (estimated by cross entropy) between the two distributions according to increasing fold change cutoffs. **(c)** The Control and Case distributions (control and case replicates was chosen from serum and 2i treatment conditions respectively. **(d)** The change of fCI divergence values between the two distributions according to increasing fold change cutoffs.



Supplementary Figure 8 Kernel density plot of control-control (blue) and case-control (red) ratio distributions from the fly RNA-Seq data (supplementary material 2.6). The control-control distribution was slightly shifted toward upper y axis to make it separated from the case-control ratio distribution.



Supplementary Figure 9. Construction of type-I error rate based on fCI's empirical null distribution. The null hypothesis and alternative hypothesis were constructed from spike-in RNA data (supplementary material 1.5). Idealy, if the cutoff fold change is chosen without error (no false positives), we should not observe any gene in the control-control ratios (empirical null distribution) that has a fold change larger than the chosen cutoff. The proportion of such genes in the empirical null distribution is equivalent to type I error rate (incorrect rejection of a true null hypothesis).



Supplementary Figure 10. False Discovery Rate estimation using fCI's permtation analyses. We generated 100 valid fCI combinations (the target database) based on Fig 1, and 100 random fCI combinations (the decoy database) by permuting the control and experimental replicates using data from supplementary material1.5. To achieve this, we randomly permute the replicates between control and experimental samples (i.e. we form a empirical null distribution by computing the ratio between the 2nd control replicate and the 1st experimental replicate), and then we computed the 'DEGs' from this permutated fCI combination. The FDR is computed as the number of total decoy and false hits divided by all fCI predicted DEGs.

Supplementary fCI pseudo code

1: Collect a total of k gene expression levels from n control replicates: $\vec{X}_i = [x_{i1}, x_{i2}, x_{i3}, ..., x_{ik}]$ where $i = 1,2,...,n$, and m case replicates $Y_i = [y_{i1}, y_{i2}, y_{i3}, ..., y_{ik}]$ where $i = 1,2,...,m$.

Then Form a total of $\frac{n \times (n-1)}{2}$ control-control replicate combinations $(\vec{X}_i, \vec{X}_j)$ where $i = 1,2,...n-1; j = i+1,...,n$ and $n \times m$ case-control combinations $(\vec{Y}_i, \vec{X}_j)$ where $i = 1,2,...n; j = 1,2,...m$

2: Generate $\frac{n \times (n-1)}{2} \times (n \times m) = \frac{m \times n^2 \times (n-1)}{2} = N$ pairwise fCI control-control and case-control pair : $[(\vec{X}_i, \vec{X}_j) : (\vec{X}_i, \vec{Y}_p)]$ where $i = 1,2,...n-1; j = i+1,...n; p = 1,2,...m$

3: Iterate through the ***d***-th pairwise fCI combination, where $d = 1,2,...,N$

Apply logarithm ratio transformation for the chosen control pair $\vec{P}_d = Log_2(\frac{\vec{X}_i}{\vec{X}_j}) = Log_2([\frac{x_{i1}}{x_{j1}}, ..., \frac{x_{ik}}{x_{jk}}])$ where $j \in 1,2,..., n-1; i \in i+1,..., n$

Apply logarithm ratio transformation for the chosen case pair $\vec{Q}_d = Log_2(\frac{\vec{Y}_i}{\vec{X}_j}) = Log_2([\frac{y_{i1}}{x_{j1}}, ..., \frac{y_{ik}}{x_{jk}}])$ where $j \in 1,2,..., n; i \in 1,2,..., m$

And then apply normalization to $\vec{P}_d$ and $\vec{Q}_d$ (supplementary method 3) if applicable.

Rank the case-control ratio by ascending order so that $\frac{y_{ia}}{x_{ja}} \le ... \le \frac{y_{ib}}{x_{jb}}$ where $i = 1,2,..., m; j = 1,2,..., n; a \in 1,2,..., k; b \in 1,2,..., k; a \ne b$

Set the current minimum divergence $Divergence_{min}$ between $\vec{P}_d$ and $\vec{Q}_d$ to be $+\infty$ and best fold-change cutoff to be $Best\_Cutoff = +\infty$ and the best set of differentially expressed genes to be the empty set $\Delta_d = [.]$

Iterate through the candidate fold-change ratio cutoff ***r*** : $r = 1.1,1.2,1.3,...10$ (or any user-defined range)

Remove gene sets $\Delta_r = (\vec{E}, \vec{F})$ if $\frac{y_{ie}}{x_{je}} \le \frac{1}{r}$ or $\frac{y_{if}}{x_{jf}} \ge r$ where $\vec{E} = e_1,..., e_v; \vec{F} = f_1,..., f_w; e_v \in 1,2,..., k; f_w \in 1,2,..., k$

Obtain the remaining non-differential case-control pair $\vec{Q}^*_d = Log_2(\frac{\vec{Y}^*_i}{\vec{X}^*_j}) = Log_2([\frac{y_{i1}}{x_{j1}}, ..., \frac{y_{ik^*}}{x_{jk^*}}])$ where $\vec{Y}^*_i = \vec{Y}_i - \vec{Y}_{i\Delta_r}; \vec{X}^*_i = \vec{X}_j - \vec{X}_{j\Delta_r}; k^* \notin e_1,..., e_v; k^* \notin f_1,..., f_w$

Compute the divergence between the original control-control empirical null and the case-control distributions $Divergence_r = div(\vec{P}_d, \vec{Q}^*_d)$ by *Hellinger distance*, *Kullback-Leibler* or *Cross Entropy* (supplementary method).

If $Divergence_r < Divergence_{\min}$, set the $Best\_Cutoff = r$ and best differentially expressed sets to be $\Delta_d = \Delta_r$

End inner loop for fold change cutoff selection under the current ***d***-th pairwise combination and report $Best\_Cutoff$ and gene sets Δ_d

End outer loop for all the N pairwise fCI combinations and report all the gene sets $\Delta_{all} = (\Delta_1, ..., \Delta_N)$

4: For the ***t***-th gene of all k genes, count the total number of times C_t and the frequency f_t that appear in the N pairwise fCI results, $C_t = \sum_{i=1}^{N} \Delta_i^t; f_t = \frac{C_t}{N}; t \in 1,2,..., k$, if the ***t***-th gene if shown in the ***i***-th pairwise fCI analysis, $\Delta_i^t = 1$ else $\Delta_i^t = 0$.

Supplementary Table 1. differential gene expression analysis of the proteogenomics data using fCI and *limma* (see supplementary material 1-1). The dataset is analyzed using fCI and *limma* Biocondutor package. Control data from both reference channel (three replicates) and Case data from channel 1 (three replicates) were used. Differentially expressed genes are the targets with an adjusted p-value less or equal than 0.05 in *limma* analysis.

Supplementary Table 2. differential gene expression analysis of the time series RNA-Seq data (see supplementary material 1-3). The dataset is analyzed using DESeq R Biocondutor package. Control data from both time points were combined, which were compared to experimental data from the two treatments that were combined as well. Differentially expressed genes are the targets with an adjusted p-value less or equal than 0.05.