

## Temporal expression divergence of network modules

Yongjin Park<sup>1</sup>, Tae-Hyuk Kang<sup>2</sup>, Theodore Friedmann<sup>3</sup> Joel S. Bader<sup>1,\*</sup>

**1 Department of Biomedical Engineering, High-throughput Biology Center and Institute of Computational Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America**

**2 Department of Neurology, Yale University School of Medicine, New Haven, CT, United States of America**

**3 Department of Pediatrics, Center for Neural Circuit and Behavior and San Diego Rady Children’s Hospital, University of California, San Diego School of Medicine, La Jolla, California, United States of America**

\* E-mail: joel.bader@jhu.edu

### Abstract

Here we propose new module-based approaches to identify differentially regulated network sub-modules combining temporal trajectories of expression profiles with static network skeletons. Starting from modules identified by network clustering of static networks, our analysis refines pre-defined genesets by partitioning them into smaller homogeneous sets by non-parametric Bayesian methods. Especially for case-control time series data we developed multi-time point discriminative models and identified each network module as a mixture or admixture of dynamic discriminative functions. Our results shows that our proposed approach outperformed existing geneset enrichment methods in simulation studies. Moreover we applied the methods to neural stem cell differentiation data, and discovered novel modules differentially perturbed in different developmental stages.

### Introduction

**Dynamics of biological networks** Studies on dynamic networks have heavily focused on node-level analysis. Hub nodes of physical interaction and signaling networks were classified to so called “party” and “date” hubs judged by average correlation with neighboring nodes [1]. Despite criticisms [2–4], dynamic property of network and co-expression modules have provided new biological insights into dynamics and systems (e.g., [1, 5–8]). Nevertheless, there still remains subtlety in the analysis. As pointed out [8], node-based statistics and notion of modular structure can be highly data-dependent, but there is no room for heterogeneity in models.

We propose new module-based approaches to dynamic network analysis. However, our notion of module is based on network topology, not expressions. These structural modules can be inferred from static edges. Previously we estimated a trace of dynamic modules while perturbing edges by expressions [9]. Although modules are usually fixed and static, we strive to relax other aspects. We do not attempt to summarize overall node-node relations by a single metric. Network modules are not fixed, but properties of contained nodes change over time. Moreover, we can circumvent long-standing “ $n \ll p$  problem” by not collecting statistics at node or node-pair level. Instead, we assume genes within well-defined modules are independent identically distributed; then, our  $n$  is not number of samples but number of genes. We turned the problem to opposite  $n \gg p$  regime.

In essence our proposed method resembles geneset analysis methods. We use interaction maps and expression matrix. First we resolve modules from static interactions; then, search for significant dynamic modules by set-based datamining. However, we greatly relaxed assumptions on the sets. Genes are not always identical, i.e., there can be multiple subgroups within genesets. Samples are time-dependent and dynamic. More importantly our goal is to identify significant module at specific timepoints. Modules are not necessarily responsive to environments.

**Time-specific discriminative modules** Regarding gene-wise set-heterogeneity, our solution is quite straightforward. We sought to refine pre-defined genesets by partitioning them into smaller homogeneous sets. We used Bayesian non-parametric method, Dirichlet Process Mixture [10]. However, we need a new paradigm of set-based significance testing to account for sample heterogeneity, especially with time-series samples. Let us put geneset analysis in a classification setting.

Suppose we have a time-series gene expression matrix  $X$ . Each  $x_{it}$  element measures gene  $i$ 's mRNA concentration at time  $t$ . There are  $n$  genes and  $T$  time points. Within this dataset, we may assume genes are independent and identically distributed. We also have each gene  $i$  labeled with  $l_i$  to indicate that was collected in wild-type (WT) cells if  $l_i = 1$ , or abnormal cells if  $l_i = 0$ . Then we can fit a logistic regression model with parameter  $\beta_t$  for each time point to classify the gene labels  $\{l_i\}$ , i.e.,

$$p(l_i|x_{it}) = \sigma(\beta_t x_{it})$$

where the sigmoid function  $\sigma(\xi) = 1/(1 + e^{-\xi})$ . We may estimate  $\beta$  by maximizing likelihood function

$$\mathcal{L}(Z, X, \beta) = \prod_{i,t} \sigma(x_{it}\beta_t)^{l_i} \sigma(-x_{it}\beta_t)^{1-l_i}. \quad (1)$$

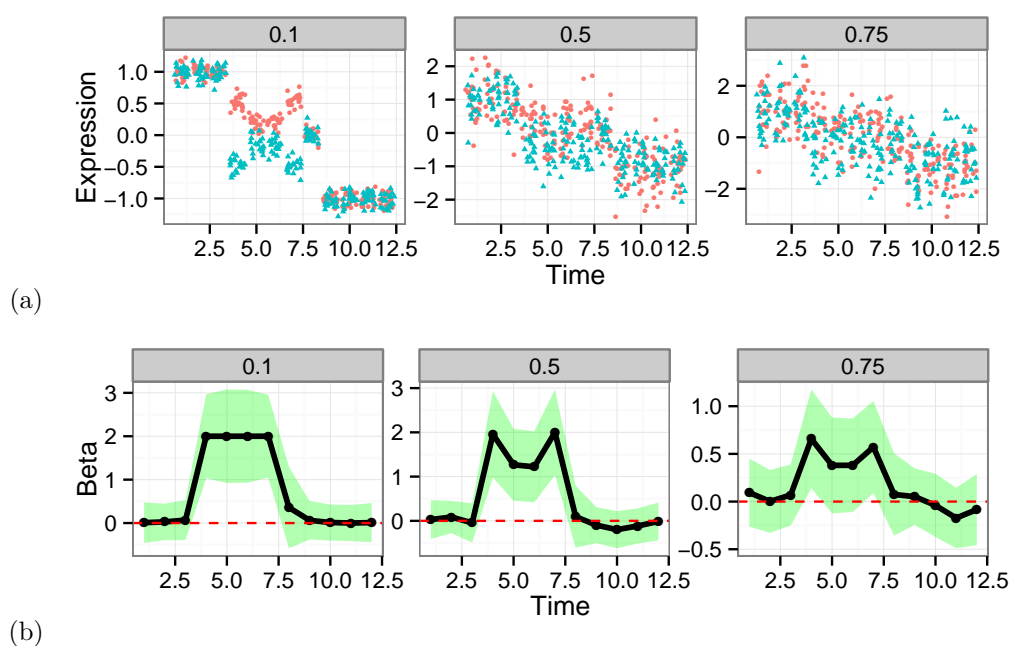
**Temporal Expression Divergence (TED) model** We think of a case-vs-control study in the same classification context. Usually we measure two types of  $n \times T$  gene expression matrices,  $X$  and  $Y$  for case and control, or vice versa. Again, we may assume genes are identical and independent, but later we will relax this assumption using predefined genesets. Although the label variable  $l_i$  not so explicit, we can set all  $x_{it}$  are labeled with  $l_i = 1$  but  $y_{it}$  with  $l_i = 0$ . Then, we want to estimate  $T$  logistic regressors, each parameterized by  $\beta_t$ , maximizing

$$\mathcal{L}(X, Y; \beta) = \prod_{i,t} \sigma(\beta_t x_{it}) \sigma(-\beta_t y_{it}). \quad (2)$$

By this model, we can quantify discriminative power of homogeneous  $X, Y$  time-specifically by magnitude of  $\beta_t$ . Suppose a pair  $(x_{it}, y_{it})$  is discriminative, e.g.,  $x_{it} \gg 0$  while  $y_{it} \ll 0$ . Then we maximize the product of two sigmoids with strongly positive  $\beta_t$ . However, if we have both  $x_{it} \gg 0$  and  $y_{it} \gg 0$ , then we want a neutral  $\beta_t$  that is close to 0 to avoid  $\sigma(-\infty) = 0$ .

Fig.1 demonstrates strength of this framework. We generated control datapoints  $\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i | \mu_{\text{case}}, \sigma^2)$  (colored red) and case  $\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | \mu_{\text{control}}, \sigma^2)$  (colored blue). We tested three toy examples generated under different variance parameters  $\sigma \in \{0.1, 0.5, 0.75\}$  (denoted on the top of subplots). Under small variance  $\sigma = 0.1$ , patterns of diverging time points are visually evident. With increasing variance to  $\sigma = 0.5$ , datasets became fuzzier and harder to separate out, but the beta parameters, corresponding to decision boundary of time-specific logistic regressors, maintain large values deviating from zero. However, with too large variance  $\sigma \geq 0.75$ , two types of datasets became really indistinguishable, then decision boundary was pulled toward zero. This aspect is also desirable. We want to say ‘‘I don’t know’’ for unclear case-versus-control.

**Contributions** Our technical contributions are in two folds: (1) we developed highly sensitive and accurate discriminative model for geneset analysis; (2) we then applied TED models in two well known modeling frameworks, Dirichlet Process Mixture [10] and Latent Dirichlet Allocation [11]. Moreover, from systematic analysis we were able to propose new hypothesis of disease mechanism of neural disorder; and we revealed systematic bias of set- or pathway-based analysis and discussed new approaches.



**Figure 1. Demonstration of Temporal Expression Divergence on datasets with time-dependent batch effects.** (a) We simulated datasets varying standard deviation parameter  $\sigma = 0.1, 0.5, 0.75$  while fixing mean vectors for case (red) and control (blue) samples to  $\mu_{\text{case}} = [1, 1, 1, .5, .2, .2, .5, 0, -1, -1, -1, -1]$  and  $\mu_{\text{control}} = [1, 1, 1, -.5, -.2, -.2, -.5, 0, -1, -1, -1, -1]$ . For each  $\sigma$ , we generated 20 of  $\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i | \mu_{\text{case}}, \sigma^2 I)$  and  $\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | \mu_{\text{control}}, \sigma^2 I)$ . (b) TED model parameters  $\beta$  fitted by our Bayesian inference algorithm (see Methods). Think solid black lines indicate  $\mathbb{E}[\beta]$  and shaded areas show confidence bands within 2 standard deviation.

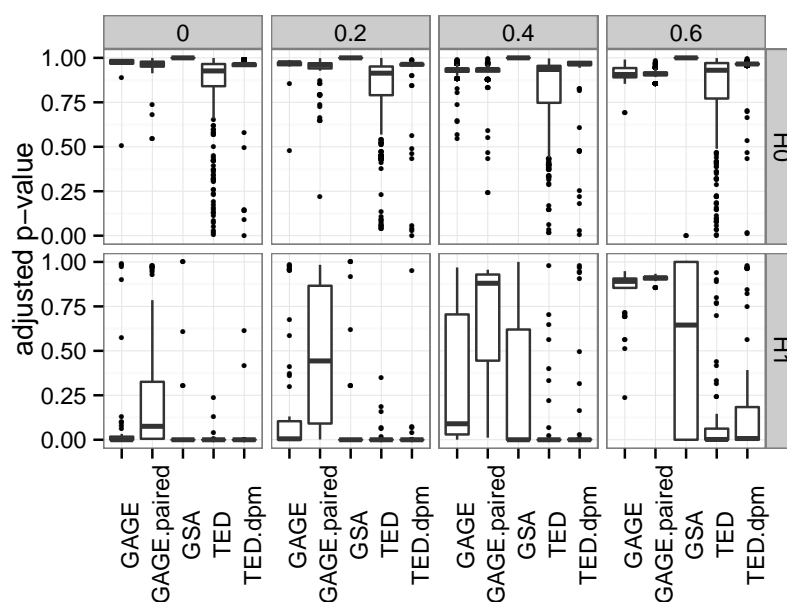
## Results and Discussion

### Comparison with other geneset analysis methods

We compared performance of TED and TED-dpm with existing methods. Although exhaustive comparison with all existing methods could be beneficial, we chose: generally applicable geneset enrichment with full pairwise sample comparison (GAGE), GAGE with paired sample comparison (GAGE-paired) [13], and geneset analysis (GSA) [14]. We chose GAGE over PAGE [15] because of better performance [13]; GSA over others because of generally good performance in empirical comparison [16]. We did not include the work of Goeman [17] based upon generalized linear models (GLMs) [18] fitting since our TED model is also a special case of generalized linear model with temporal specificity.

Two types of setwise expression matrices of 10 samples were generated; one under the null hypothesis (marked  $H_0$ ) and alternative ( $H_1$ ). Under the null hypothesis, both control  $x_{it}$  and case  $y_{it}$  expressions were sampled from the standard Normal distribution, i.e.,  $x_{it} | H_0 \sim \mathcal{N}(x_{it} | 0, 1)$  for all  $t \in [10]$ . However, whitening genesets under the alternative hypothesis, 60% of genes were differentially expressed;  $x_{it} \sim \mathcal{N}(x_{it} | .5, 1)$  while  $y_{it} \sim \mathcal{N}(y_{it} | -.5, 1)$ . All the method performed similarly well on completely differential expression (100%). In addition, we tested conditions where not all time points were differentially expressed under the alternative. We included a certain fraction of non-informative time points, following the null distribution to the differentiating genesets, from 0% to 60%.

Fig.2 summarizes the results. Each column show dispersion of adjusted  $p$ -values found by different



**Figure 2. Performance of geneset enrichment analysis.** Abbreviation of methods are stated in the text. Each method reported  $p$ -values of genesets; adjusted  $p$ -values were computed using Benjamini-Hochberg method [12], implemented in `p.adjust` of R.

methods under the null (H0) and alternative (H1) hypotheses. We expect an oracle method, which gives perfect separation, would generate  $p$ -values clustered to 1 under the null, but 0 under the alternative. Without noisy samples (the 1st column), all the methods work well as expected. However, as we introduce noisy samples more, we may not pick up significant fraction of true positives using GAGE or GSA. Especially when 60 % noisy samples introduced, GAGE and GSA become significantly underpowered compared to TED. However, TED, not using DPM-based set refinement, might have higher false discovery rate than other methods (outliers of the boxplots); yet, the set refinement method corrected this adequately.

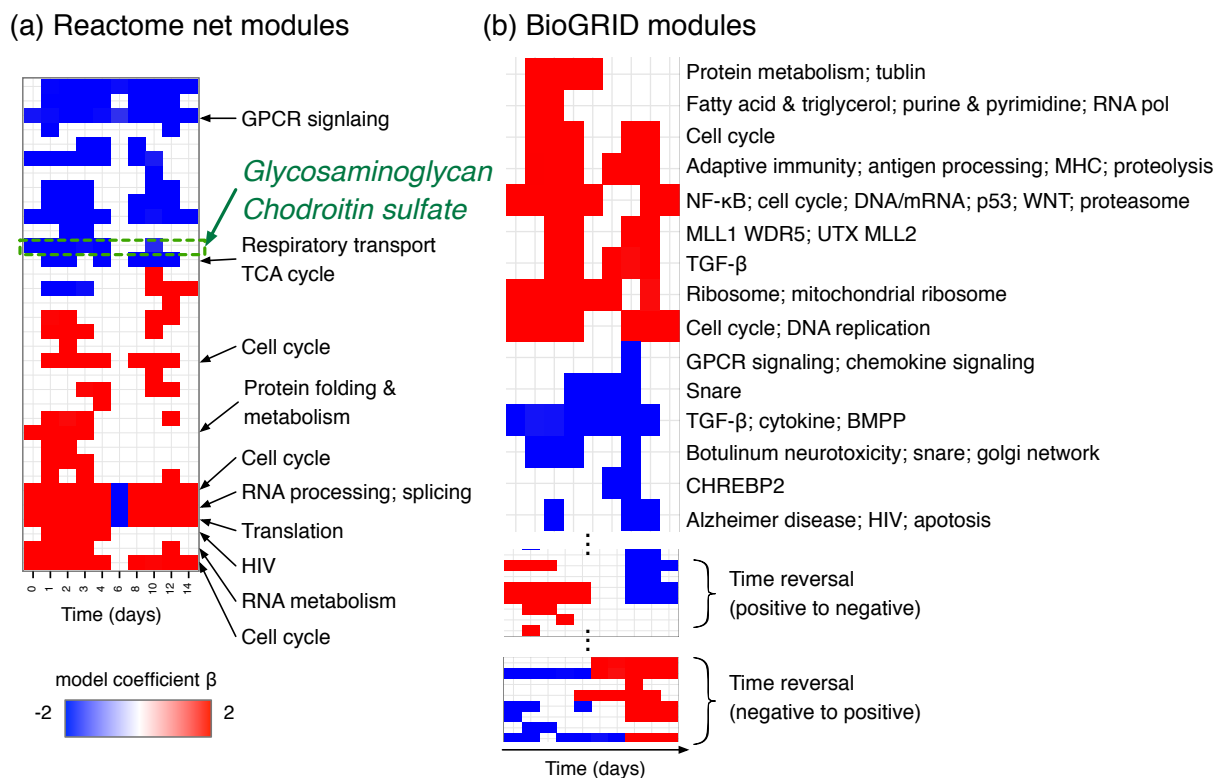
## Neural stem cell differentiation

Lesch-Nyhan Disease hints interesting links between aberrant purine biosynthesis and neurophysiological and neurobehavioral disorder [19]. Although overall mechanisms are yet to be understood, previous studies discovered that mutations of HPRT (hypoxanthine guanine phosphoribosyltransferase) causes defects in neuronal growth and differentiation [20]. Here we focused on changes of network modules/pathways. We used time-series case versus control expression matrices  $X$  and  $Y$  constructed from the RNA-sequencing dataset (GSE42662) of 14 days of Dopaminergic (DA) neuronal differentiation of spherical neural masses (SNMs) [21]. The dataset includes 10 snapshots of days 0, 1, 2, 3, 4, 6, 8, 10, 12 and 14, generated by human DA differentiation protocol [22], where they characterize 14 days as three distinctive phases: (1) neuronal induction from days 0 to 4; (2) DA neuron induction from days 4 to 8; (3) DA neuron maturation from days 8 to 14.

## Significant network modules

**Network clustering** We ran our hierarchical network clustering algorithm, improved upon our previous work [23], on BioGRID physical and Reactome co-reaction networks separately, and resolved 173 of physical and 144 of co-reaction modules (see Methods for details). To ensure high quality of modules, we compared results with other network clustering methods. In extensive benchmark and cross-validation experiments, we found our algorithm substantially outperformed than others (see supplementary results; Fig.7 and 8).

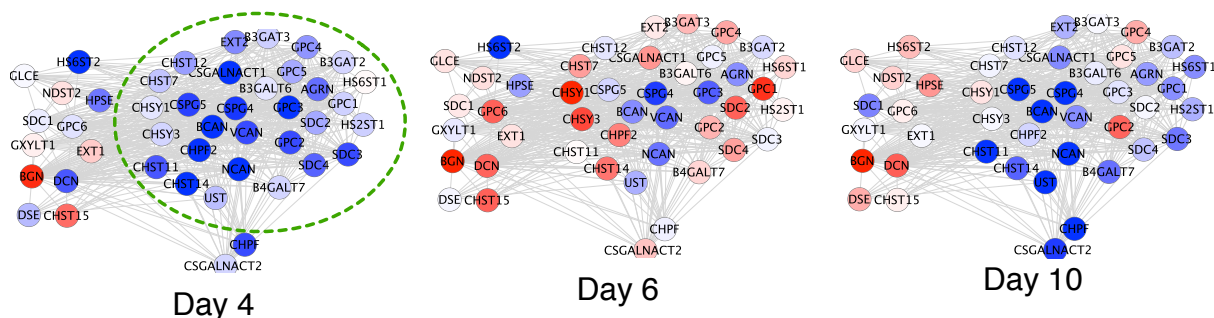
**Time-specific network modules changing the fate of stem cell differentiation** We focus on results of TED-dpm (TED and DPM-based geneset refinement) because of substantial improvement of statistical power and homogeneity of modules. As expected from the result of simulation study (Fig.2), TED-dpm was able to identify more significant modules than other methods. Geneset refinement steps effectively separated out noisy, or weak, gene expressions from genesets, highlighted true signals embedded in the sets. We confirmed that after the refinement other methods such as GSA significantly better than before (see supplementary Table.1, 2, 3 and 4). Fig.3 shows significantly changing modules found by TED-dpm method at familywise error rate 0.01. Due to space limit, we only show those non-trivially overlap with known canonical pathways [24], determined by hypergeometric test (FDR < 0.01 with overlap > 0.25).



**Figure 3. network moduels.** We determined significance of these modules at familywise error rate (FWER) < 0.01;  $p$ -values corrected by Holm procedure (see e.g., [25]).

Throughout all time course, modules related to cell cycle, nucleotide and downstream lipid metabolism

were upregulated in control samples; in other words, activity of these modules was unusually dampened by HPRT gene KD. However, modules of GPCR signaling (Reactome), and a part of TGF- $\beta$  and pathways (BioGRID), were up-regulated in the KD cells. A large fraction of modules did not change direction of regulations either positive (red) or negative (blue). However, there are some modules undergo transient swapping (e.g., RNA processing in Reactome network, Fig.3a), or complete phase transition (e.g., anti-phasing modules in BioGRID network, Fig.3c). Unfortunately anti-phasing modules are poorly characterized in current knowledge; and they may have not been recognized as functional modules. On modules with almost equal amount of up- and down-regulated samples sample average of log<sub>2</sub> ratio would only remain close to zero effect (row-wise summation).



**Figure 4. Glycosaminoglycan complex.** Six sub-networks show transcriptomic dynamics of “Glycosaminoglycan” module in Reactome network. We show all parts of module inferred by network clustering; green dashed circle encloses parts included by TED-DPM. Nodes and edges represent genes and interactions of Reactome network (co-reaction). Genes were colored by expression level. Lowest log<sub>2</sub>(control/KD) ratio was colored blue (RGB 0,0,255); equal amount of control and KD colored white (RGB 255,255,255); highest colored red (RGB 255,0,0).

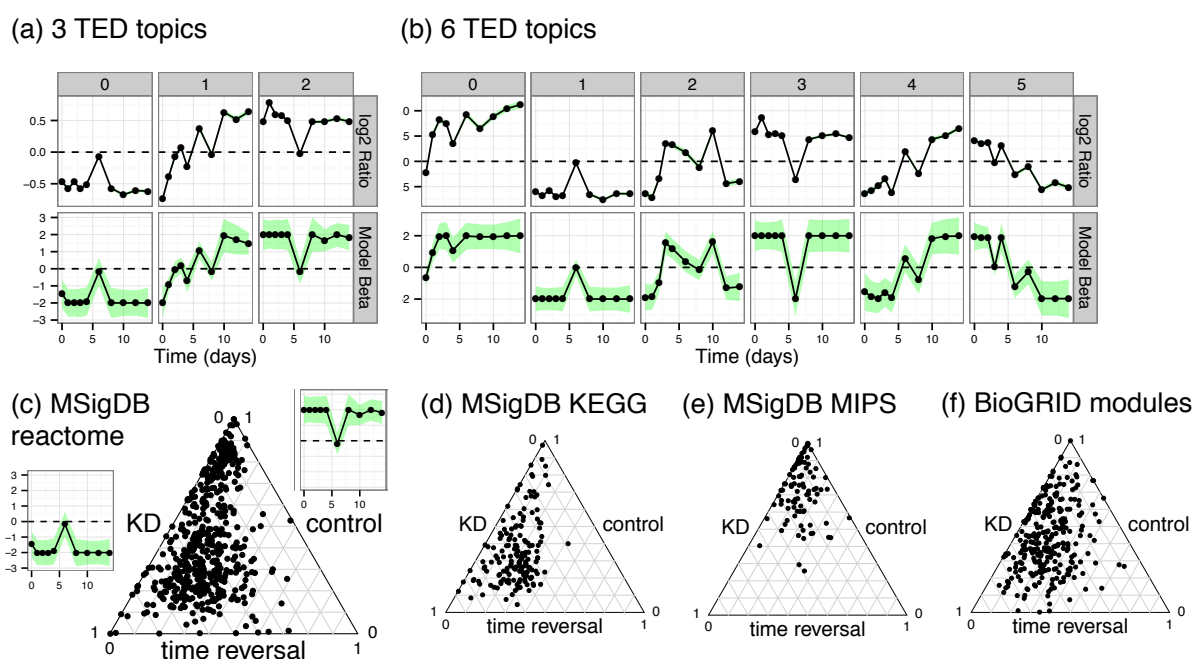
**Glycosaminoglycan complex** In Fig.4, we zoomed-in the glycosaminoglycan module of Reactome network. It is evident that tightly co-expressed core sub-complex (green dashed circle) plays a major role in the complex. The core includes glycan and chondroitin enzymes: B4GALT ( $\beta$ -1,4-galactosyltransferase), B3GAT ( $\beta$ -1,3-glucuronyltransferase), CHST (carbohydrate sulfotransferase), CHSY (chondroitin synthase), and CHPF (chondroitin polymerizing factor). In the knock-down HPRT cells, this core complex was initially up-regulated during neuronal induction phase until day 4, but down-regulated during DA neuron induction, then lost significant synchronous pattern during DA maturation (Fig.3a; Fig.4).

Along the same direction, recent studies show that neural stem cells can be identified specific glycan markers, and many stem cell signaling pathways regulated and marked by post-translational modifications of glycans attached to membrane [26]. Moreover, there was a systematic assay that mRNA level of many glycan enzymes strongly correlates with cell fate change of mouse embryonic stem cells [27]. In some cases, abnormal expression B4GALT family genes has been shown to promote multi-drug resistance in leukemia cells [28]. In sum, we may suggest an interesting hypothesis of LND: upon HPRT knock-down, cells experience dysfunctional purine metabolism, which directly leads to inappropriate exposition of glycans to cell surface; interplay of other signaling pathways (e.g., Wnt and GPCR) completely change cell fates departing from DA neurons. However, we may find better picture by follow-up experimental studies.

## Common signature of regulation

It was our surprise that a substantial fraction of network modules were largely heterogeneous mixtures of TED models. In fact, we modified TED to address this nature. Although our network clustering algorithm recapitulated high-probability modules, we further tested to show heterogeneity of modules is indeed algorithm-independent.

**Pathways are mixed** To answer that question, we incorporated TED models in Latent Dirichlet Allocation (LDA) framework [11]. One may think of LDA as probabilistic principal components. We looked for commonly observed expression patterns across genesets, and estimated mixing proportion of genesets. Common patterns are termed topics, which are TED topics, characterized by  $\beta$  parameters. We collected manually curated  $\approx 1,400$  genesets (c2.cp) from MSigDB [24], which provides extensive coverage of known and manually curated pathways collected from various databases. On top of that, we also added Reactome and Biogrid network modules found in our network clustering.



**Figure 5. Genesets are mixed.** (a) 3 distinctive TED topics. Each column shows each topic. Top row enumerates topic-specific average of  $\log_2(\text{control}/\text{KD})$ . Bottom row shows topic-specific model parameters  $\beta$  with 95% confidence bands (shaded area). (b) 6 TED topics. Top and bottom rows show 6 topic-specific data averages and model parameters. (c) Triplot of topic proportion. Each dot corresponds to MSigDB’s “reactome” genesets. Three vertices of simplex correspond to topics 0 to 2 of 3 TED topics (counterclockwise from the left corner of triangle). (d) Topic proportion of MIPS complexes (e) Topic proportion of BioGRID network modules.

We were able to confirm that trained TED topics (Fig.5a,b) summarized what we repeatedly observed pattern of gene expressions in many genesets and network modules. Models either with 3 or 6 topics showed highest predictive performance in 10-fold cross validation experiments (see Supplementary result; Fig.6). But we mapped genesets and network modules to 3-TED topic space for simplified visualization

(Fig.5c-e). Genesets are apparently mixed. Although many of genesets locate nearby the top corners of triangles (topic 2 in Fig.5a), a remaining portion of genesets are mixture of topics 0 and 1. Notably MIPS complexes are biased toward topic 2, which means that these complexes were overexpressed in control cells. It appears that MIPS complexes were established on normal cells. Therefore, future analysis based solely on MIPSs would have to be limited and may not discover new disease-specific pathways and complexes.

## Conclusion

We developed a novel set-based method to annotate pre-defined genesets or modules inferred from networks. The methods were not only statistically more powerful, but also able to capture transient signals reliably. We designed TED and variants in two philosophy: classification is easier than density estimation (e.g., [29]). However, previous works approach the problem to sift out significant sets by unusual density of observations. We also wanted to do discriminative learning, which is usually easier than generative model fitting (e.g., [30]). TED and TED-variants outperformed in both synthetic and real-world than other geneset analysis methods. Moreover, we empirically demonstrated prevalent heterogeneity during the neural differentiation process. We also showed anti-phasing modules yet to be characterized by subsequent experiments and analysis.

## Methods

### Data preparation

**Interactome** We constructed a physical interaction network dataset of 14,995 proteins and 140,006 interactions from BioGRID 3.1.94 [31], which are edges labeled `physical`. We also constructed a co-reaction network dataset of 4,527 genes and 87,947 interactions. We downloaded Reactome network database [32] from [http://www.reactome.org/download/current/homo\\_sapiens.interactions.txt.gz](http://www.reactome.org/download/current/homo_sapiens.interactions.txt.gz). We used “current version” as of Aug 4 2013. We only included edges labeled `reaction` or `neighbouring_reaction`, so as to identify modules, which could not be found in physical interaction networks.

**Transcriptome** Over 98% of short reads were mapped to NCBI mm9 transcripts using `tophat` [33], counted by `htseq-count` (<http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>) and normalized by `DESeq` taking into accounts of effective size factors [34]. The data was collected on mouse embryonic stem cells but mouse interactomes were rather poorly characterized compared to human. We mapped mouse genes to human genes using `biomaRt` [35]. Resulting  $X$  and  $Y$  matrices contained 10,170 genes of 10 snapshots (days 0, 1, 2, 3, 4, 6, 8, 10, 12 and 14).

### Variational inference of TED model

**Bayesian inference and hypothesis testing** We performed Bayesian inference on unknown parameters, especially the  $\beta$  parameters. We defined likelihood of observed  $X, Y$  under the parameter  $\beta$  of TED model (Eq.2), and the  $\beta$  parameters directly translate to significance of observed  $X, Y$ , quantifying steepness of decision boundary. Since we have  $(\beta_t - \mathbb{E}[\beta_t]) / \sqrt{\mathbb{V}[\beta_t]} \approx \mathcal{N}(0, 1)$ , both approximately and asymptotically, we may construct level  $\alpha$  Wald test of testing  $H_0 : \beta_t = 0$  versus  $H_0 : \beta_t \neq 0$  as:

$$\text{reject } H_0 \quad \text{if} \quad \left| \frac{\mathbb{E}[\beta_t]}{\sqrt{\mathbb{V}[\beta_t]}} \right| > z_{\alpha/2} \quad (3)$$



where  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ . Then,

$$p\text{-value} = \Pr \left( |Z| > \left| \mathbb{E}[\beta_t] / \sqrt{\mathbb{V}[\beta_t]} \right| \right), \quad (4)$$

where  $Z \sim \mathcal{N}(0, 1)$ . In a strict sense, our hypothesis testing is more akin to controlling false discovery rate rather than Type I error. We may consider TED as a “discriminative” version of Sun and Cai [36,37].

**Bayesian sparse prior** Now, let us discuss how to estimate  $\mathbb{E}[\beta_t]$  and  $\mathbb{V}[\beta_t]$  taking into account of temporal smoothness. We introduced the Fused Lasso prior [38] on to  $\beta$  parameters.

$$p(\beta) \propto \exp \left( -\frac{\lambda}{2} \sum_{t=1}^T |\beta_t| - \frac{\gamma}{2} \sum_{t=1}^{T-1} |\beta_{t+1} - \beta_t| \right) \quad (5)$$

where  $\lambda$  and  $\gamma$  control static and kinetic sparsity of parameters. Under appropriate  $\lambda$  and  $\gamma$  we capture sparse and temporally smooth  $\beta$ . However, estimation of  $\beta$  under the Fused Lasso prior may not be technically difficult unlike regular Lasso prior [39]. Instead of direct usage, we used equivalent Bayesian formulation to handle sparsity more gently (see [40] and [41] for details and proofs).

$$\begin{aligned} \beta_j | \tau &\sim \mathcal{N}(\beta_j | 0, \tau_j^2) \\ \beta_j - \beta_{j+1} | \kappa &\sim \mathcal{N}(\beta_j - \beta_{j+1} | 0, \kappa_j^2) \\ \tau_j^2 &\sim \text{Exp}(\tau_j^2 | \lambda/2) \\ \kappa_j^2 &\sim \text{Exp}(\kappa_j^2 | \gamma/2). \end{aligned}$$

**Non-conjugate variational inference** To work around non-conjugate relationship between the likelihood function (Eq.2) and prior (Eq.5), we estimated posterior probability by the non-conjugate variational inference [42]. First we find optimal  $\beta$  by coordinate descent algorithm [43], then construct approximate distribution  $q(\beta)$  by the Laplace method [42].

**Distribution of  $\beta$**  First of all, it is not hard to verify convexity of negative log-likelihood function:

$$f(\beta) = \sum_{t=1}^T \sum_{i=1}^n [\log(1 + e^{-\beta_t x_{it}}) + \log(1 + e^{\beta_t y_{it}})]. \quad (6)$$

Like other GLMs, we can construct local quadratic approximation [43,44]. Using these results, we can construct the following local quadratic form  $g(\beta) \approx f(\beta)$  at current estimate  $\hat{\beta}$ , defined by

$$g(\beta) = \frac{1}{2} \sum_{t=1}^T \sum_{i=1}^n w_{it} (\beta_t - v_{it})^2 \quad (7)$$

where

$$w_{it} = x_{it}^2 \sigma(x_{it} \hat{\beta}_t) \sigma(-x_{it} \hat{\beta}_t) + y_{it}^2 \sigma(y_{it} \hat{\beta}_t) \sigma(-y_{it} \hat{\beta}_t) \quad (8)$$

and

$$v_{it} = \hat{\beta}_t + \frac{x_{it} \sigma(-x_{it} \hat{\beta}_t) - y_{it} \sigma(y_{it} \hat{\beta}_t)}{w_{it}}. \quad (9)$$

With the prior on  $\beta$ , we find the solution by solving

$$g(\beta) + \frac{1}{2} \sum_{t=1}^T \mathbb{E} \left[ \frac{1}{\tau_t^2} \right] \beta_t^2 + \frac{1}{2} \sum_{t=1}^{T-1} \mathbb{E} \left[ \frac{1}{\kappa_t^2} \right] (\beta_t - \beta_{t+1})^2. \quad (10)$$

For each  $t \in \{2, T-1\}$ , we can iteratively update  $\beta_t$  until convergence as follows.

$$\beta_t = \frac{\sum_{i=1}^n w_{it} v_{it} + \hat{\beta}_{t-1} \mathbb{E}[1/\kappa_{t-1}^2] + \hat{\beta}_{t+1} \mathbb{E}[1/\kappa_t^2]}{\sum_{i=1}^n w_{it} + \mathbb{E}[1/\tau_t^2] + \mathbb{E}[1/\kappa_{t-1}^2] + \mathbb{E}[1/\kappa_t^2]}, \quad (11)$$

we drop the  $t-1$  term for  $t=1$  and  $t+1$  for  $t=T$ .

Following Wang and Blei [42], we have

$$q(\boldsymbol{\beta}) \approx \mathcal{N}(\boldsymbol{\beta} | \mathbb{E}[\boldsymbol{\beta}], \Lambda^{-1}) \quad (12)$$

where  $\mathbb{E}[\boldsymbol{\beta}] = \hat{\boldsymbol{\beta}}$  and tridiagonal precision matrix

$$\Lambda_{tt} = \sum_{i=1}^n w_{it} + 1/\tau_t^2 + 1/\kappa_{t-1}^2 + 1/\kappa_t^2, \quad \Lambda_{t,t+1} = \Lambda_{t+1,t} = -1/\kappa_t^2. \quad (13)$$

For subsequent updates, we need covariance matrix  $\Sigma \equiv \Lambda^{-1}$ . Exploiting special structure of  $\Lambda$ , we can easily resolve tri-diagonal elements of  $\Sigma$ , i.e.,  $\Sigma_{tt}$  and  $\Sigma_{t,t+1}$  by linear time algorithms (e.g., [45] or Appendix B of [46]).

$$\mathbb{E}[\beta_t^2] = \hat{\beta}_t^2 + \Sigma_{tt}, \quad \mathbb{E}[(\beta_t - \beta_{t+1})^2] = \mathbb{E}[\beta_t^2] + \mathbb{E}[\beta_{t+1}^2] - 2(\hat{\beta}_t \hat{\beta}_{t+1} + \Sigma_{t,t+1}). \quad (14)$$

**Distribution of  $\tau, \kappa$**  From the general result of Kyung *et al.* [41], although tedious derivations yield the same results, we have  $1/\tau_t^2 \sim \text{inv}\mathcal{N}(1/\tau_t^2 | \sqrt{\frac{\lambda^2}{\hat{\beta}_t^2}}, \lambda^2)$  and  $1/\kappa_t^2 \sim \text{inv}\mathcal{N}(1/\kappa_t^2 | \sqrt{\frac{\gamma^2}{(\hat{\beta}_t - \hat{\beta}_{t-1})^2}}, \gamma^2)$ . Therefore, we can find mean-field solution required for the update of  $q(\boldsymbol{\beta})$  and empirical Bayes estimation penalty terms as:

$$\mathbb{E}[1/\tau_t^2] = \sqrt{\frac{\lambda^2}{\mathbb{E}[\beta_t^2]}}, \quad \mathbb{E}[1/\kappa_t^2] = \sqrt{\frac{\gamma^2}{\mathbb{E}[(\beta_t - \beta_{t-1})^2]}} \quad (15)$$

and

$$\mathbb{E}[\tau_t^2] = \sqrt{\frac{\mathbb{E}[\beta_t^2]}{\lambda^2}} + \frac{1}{\lambda^2}, \quad \mathbb{E}[\kappa_t^2] = \sqrt{\frac{\mathbb{E}[(\beta_t - \beta_{t-1})^2]}{\gamma^2}} + \frac{1}{\gamma^2}. \quad (16)$$

**Empirical Bayes** We adjust a proper degree of penalties,  $\lambda$  and  $\gamma$ , by optimizing the marginal likelihood weighted by the variational distributions over  $\boldsymbol{\beta}$  and  $\tau, \kappa$  (see [40, 41] for details). We update  $\lambda$  and  $\gamma$  as:

$$\frac{1}{\lambda^2} \leftarrow \frac{\sum_{t=1}^T \mathbb{E}[\tau_t^2]}{2T}, \quad \frac{1}{\gamma^2} \leftarrow \frac{\sum_{t=1}^{T-1} \mathbb{E}[\kappa_t^2]}{2(T-1)}. \quad (17)$$

until convergence.

**Overall inference algorithm** Initially we set  $\mathbb{E}[1/\tau_t^2] = 0$  for  $t \in [T]$ , and  $\mathbb{E}[1/\kappa_t^2] = 0$  for  $t \in [T-1]$ . Then, we repeat the following steps until convergence:

1. Optimize  $\hat{\boldsymbol{\beta}}$  (Eq.11) and estimate  $q(\boldsymbol{\beta}|\cdot) \approx \mathcal{N}(\boldsymbol{\beta} | \hat{\boldsymbol{\beta}}, \Lambda^{-1})$  (Eq.12).
2. Calculate covariance matrix  $\Sigma \equiv \Lambda^{-1}$  and  $\mathbb{E}[\beta^2]$  and  $\mathbb{E}[(\beta_t - \beta_{t+1})^2]$  (Eq.14).
3. Update  $\mathbb{E}[1/\kappa^2], \mathbb{E}[1/\tau^2]$  (Eq.15).
4. Alternate update of  $\mathbb{E}[\tau^2], \mathbb{E}[\kappa^2]$  (Eq.16) and empirical Bayes estimation of  $\lambda, \gamma$  until convergence (Eq.17).

C++ implementation of TED and related methods are available in public repository (<https://code.google.com/p/g1mblock>).

## Mixture and admixture of TED models

**Locally collapsed latent update** So far variational inference algorithm assumed a single TED model. For both mixture or admixture models, we need a scoring function to evaluate probabilistic assignment of each object, i.e., a pair of expression vectors  $\mathbf{x}_i$  and  $\mathbf{y}_i$ , to TED models.

We found “locally collapsed” variational inference (LCVI) [47] works better than regular mean-field approximation methods. Suppose for each TED model  $k$  we have estimated variational  $q_k(\boldsymbol{\beta})$ . Essentially LCVI helps avoid bad local optima infusing adequate stochasticity. Given them, latent assignment of  $\mathbf{x}$  and  $\mathbf{y}$  is conditionally independent. Let  $c_i \in [K]$  be a random variable of  $i$ 's model assignment, although sometimes  $K \rightarrow \infty$ . We integrate out uncertainty of  $\boldsymbol{\beta}$  with respect  $q_k(\boldsymbol{\beta}) \approx \mathcal{N}(\boldsymbol{\beta} | \hat{\boldsymbol{\beta}}_k, \Lambda_k^{-1})$  previously determined by each  $k$ -th model.

$$q(c_i = k | \mathbf{x}, \mathbf{y}) \approx \Pr(c_i = k) \prod_{t=1}^T \exp\left(f(\hat{\beta}_{kt}) + \frac{1}{2} \frac{f'(\hat{\beta}_{kt})^2}{\Lambda_{ktt} - f''(\hat{\beta}_{kt})}\right) \left(\frac{\Lambda_{ktt}}{\Lambda_{ktt} - f''(\hat{\beta}_{kt})}\right)^{1/2}$$

where  $\beta_{kt}$  denotes  $t$ -th element of  $\boldsymbol{\beta}_k$  vector;  $\Lambda_{ktt}$  denotes  $(t, t)$  element of  $\Lambda_k$  matrix. We calculate appropriate prior factor  $\Pr(c_i = k)$  depending on DPM or LDA.

But we may not have previous observation at all. For instance, in DPM inference, we can discover a new model which was never explored. We simply treat  $\beta_t \sim \mathcal{N}(\beta_t | 0, \tau^2)$  with  $1/\tau^2 = \infty$ ; we use the fact that  $1/\tau^2 \approx \lambda/|\beta_t| \rightarrow \infty$  as  $\beta \rightarrow 0$ . We elaborated detailed derivations and computationally efficient update strategies in the appendix.

**Dirichlet Process Mixture** Here we briefly mention sketch of overall algorithm. First we randomly assign expression pairs  $(\mathbf{x}_i, \mathbf{y}_i)$  to models and update models given data. Iteratively we sample assignments of  $(\mathbf{x}_i, \mathbf{y}_i)$ , which works like E-step of Expectation Maximization [48]; then, we perform variational update of each TED model including empirical Bayes routines, and this works like M-step. Interested readers may refer to variational inference methods of general DPMs (e.g., [47, 49]). Our algorithm was essentially equivalent to Wang and Blei [47].

**Latent Dirichlet Allocation (admixture model)** LDA [11] models corpus of documents, while assigning topics to words and topic proportions to documents. Each topic is defined by its own topic-specific word frequency vector. Here, we treat genesets as documents and genes as words. Topics are defined by TED models. We sampled each gene's topic assignment according to LCVI score (Eq.18). Each TED model was trained by genes assigned in online stochastic learning [50]. While gradually observing genesets, we updated TED topics. We tested both batch and online learning, but online learning worked in a more scalable way and resolved equally likely models. Again, details can be found in the original paper [11] and algorithm paper [50].

## Network clustering

We extended our previous method, called dynamic hierarchical model (DyHM) [23] in two aspects: (1) we updated base-model to take into accounts of empirical degree distribution [51]; (2) we sped up the overall algorithm using dynamic programming technique. Source codes are available in public repository (<https://code.google.com/p/hsblock>).

## References

1. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430: 88–93.

2. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS biology* 4: e317.
3. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS biology* 5: e154.
4. Agarwal S, Deane CM, Porter MA, Jones NS (2010) Revisiting Date and Party Hubs: Novel Approaches to Role Assignment in Protein Interaction Networks. *PLoS computational biology* 6: e1000817.
5. Jansen R, Greenbaum D, Gerstein M (2002) Relating Whole-Genome Expression Data with Protein-Protein Interactions. *Genome research* 12: 37–46.
6. Komurov K, White M (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Molecular systems biology* 3: 110.
7. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 27: 199–204.
8. Chang X, Xu T, Li Y, Wang K (2013) Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Scientific reports* 3: 1691.
9. Park Y, Bader JS (2012) How networks change with time. *Bioinformatics* 28: i40–i48.
10. Antoniak CE (1974) Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics* 2: 1152–1174.
11. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)* 3: 993–1022.
12. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological* 57: 289–300.
13. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics* 10: 161.
14. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. *The Annals of Applied Statistics* 1: 107–129.
15. Kim SY, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* 6: 144.
16. Maciejewski H (2013) Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics* : bbt002–.
17. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2003) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99.
18. Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. *Journal of the Royal Statistical Society Series A (General)* 135: 370–384.
19. Mastrangelo L, Kim JE, Miyanohara A, Kang TH, Friedmann T (2012) Purinergic signaling in human pluripotent stem cells is regulated by the housekeeping gene encoding hypoxanthine guanine phosphoribosyltransferase. *Proc Natl Acad Sci USA* 109: 3377–3382.

20. Kang TH, Guibinga GH, Jinnah HA, Friedmann T (2011) HPRT deficiency coordinately dysregulates canonical Wnt and presenilin-1 signaling: a neuro-developmental regulatory role for a housekeeping gene? *PloS one* 6: e16572.
21. Kang TH, Park Y, Bader JS, Friedmann T (2013) The Housekeeping Gene Hypoxanthine Guanine Phosphoribosyltransferase (HPRT) Regulates Multiple Developmental and Metabolic Pathways of Murine Embryonic Stem Cell Neuronal Differentiation. *PloS one* 8: e74967.
22. Cho MS, Lee YE, Kim JY, Chung S, Cho YH, et al. (2008) Highly efficient and large-scale generation of functional dopamine neurons from human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* 105: 3392–7.
23. Park Y, Moore C, Bader JS (2010) Dynamic networks from hierarchical bayesian graph clustering. *PloS one* 5: e8118.
24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, et al. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27: 1739–1740.
25. Lehmann EL, Romano JP (2005) *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer.
26. Lanctot PM, Gage FH, Varki AP (2007) The glycans of stem cells. *Current Opinion in Chemical Biology* 11: 373–380.
27. Nairn AV, Aoki K, dela Rosa M, Porterfield M, Lim JM, et al. (2012) Regulation of Glycan Structures in Murine Embryonic Stem Cells: COMBINED TRANSCRIPT PROFILING OF GLYCAN-RELATED GENES AND GLYCAN STRUCTURAL ANALYSIS. *The Journal of Biological Chemistry* 287: 37835–37856.
28. Zhou H, Ma H, Wei W, Ji D, Song X, et al. (2013) B4GALT family mediates the multidrug resistance of human leukemia cells by regulating the hedgehog pathway and the expression of p-glycoprotein and multidrug resistance-associated protein 1. *Cell death & disease* 4: e654.
29. Friedman J (1997) On Bias, Variance,  $0/1$ Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery* 1: 55–77.
30. Ng AY, Jordan MI (2001) On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In: *Advances in Neural Information Processing Systems (NIPS)*.
31. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34: D535.
32. Croft D, O’Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 39: D691–7.
33. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.
34. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11: R106.
35. Zhang J, Haider S, Baran J, Cros A, Guberman JM, et al. (2011) BioMart: a data federation framework for large collaborative projects. *Database : the journal of biological databases and curation* 2011: bar038.

36. Sun W, Cai TT (2007) Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control. *Journal of the American Statistical Association* 102: 901–912.
37. Sun W, Tony Cai T (2009) Largescale multiple testing under dependence. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 71: 393–424.
38. Tibshirani R, Wang P (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics (Oxford, England)* 9: 18–29.
39. Friedman J, Hastie T, Höfling H, Tibshirani R (2007) Pathwise coordinate optimization. *The Annals of Applied Statistics* 1: 302–332.
40. Park T, Casella G (2008) The Bayesian Lasso. *Journal of the American Statistical Association* 103: 681–686.
41. Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* 5: 369–411.
42. Wang C, Blei DM (2013) Variational Inference in Nonconjugate Models. *Journal of Machine Learning Research* 14: 1005–1031.
43. Friedman JH, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33: 1–22.
44. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*. Chapman & Hall / CRC. 1001.4656v2.
45. Yamamoto T, Ikebe Y (1979) Inversion of band matrices. *Linear Algebra and its Applications* 24: 105–111.
46. Rybicki G, Hummer D (1991) An accelerated lambda iteration method for multilevel radiative transfer. I - Non-overlapping lines with background continuum. *Astronomy and Astrophysics* 245: 171–181.
47. Wang C, Blei DM (2012) Truncation-free Online Variational Inference for Bayesian Nonparametric Models. In: P Bartlett, FCN Pereira, CJC Burges, L Bottou, KQ Weinberger, editors, *Advances in Neural Information Processing Systems* 25. pp. 422–430.
48. Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 39: 1–38.
49. Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. *Bayesian Analysis* 1: 121–143.
50. Hoffman M, Blei DM, Bach F (2010) Online Learning for Latent Dirichlet Allocation. In: Lafferty J, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, editors, *Advances in Neural Information Processing Systems* 23. pp. 856–864.
51. Karrer B, Newman MEJ (2010) Stochastic blockmodels and community structure in networks. *Physical Review E* : 11.

## Figure Legends

## Tables

## Appendix: Derivation of update equations

### TED

#### Locally Collapsed Variational Inference

Again, to circumvent non-conjugate relations, we approximate the likelihood by second order Taylor expansion,

$$\log \mathcal{L}(\mathbf{x}, \mathbf{y}; \boldsymbol{\beta}) \approx \sum_{t=1}^T \left[ f(\hat{\beta}_t) + f'(\hat{\beta}_t)(\beta - \hat{\beta}_t) + \frac{1}{2} f''(\hat{\beta}_t)(\beta - \hat{\beta}_t)^2 \right], \quad (18)$$

where  $f(\beta_t) = -\log(1 + e^{-\beta_t x_t}) - \log(1 + e^{\beta_t y_t})$ . Then, posterior probability of  $(\mathbf{x}_i, \mathbf{y}_i)$  assignment to model  $k$  is straightforward.

For simplicity, let  $\tilde{\lambda} \equiv (\lambda - f''(\mu))$  and  $\tilde{\mu} \equiv (\mu + f'(\mu)/\tilde{\lambda})$ .

$$S(x) = \exp\left(f'(\mu)x + \frac{1}{2}f''(\mu)(x - \mu)^2 - \frac{1}{2}\lambda(x - \mu)^2\right) \quad (19)$$

$$= \exp\left(-\frac{\tilde{\lambda}}{2}(x - \mu)^2 + f'(\mu)x\right) \quad (20)$$

$$= \exp\left(-\frac{\tilde{\lambda}}{2}[x - 2\tilde{\mu}x + \tilde{\mu}^2] - \frac{\tilde{\lambda}}{2}\mu^2 + \frac{\tilde{\lambda}}{2}\tilde{\mu}^2\right) \quad (21)$$

$$= \exp\left(-\frac{\tilde{\lambda}}{2}(x - \tilde{\mu})^2 - \frac{\tilde{\lambda}}{2}\mu^2 + \frac{\tilde{\lambda}}{2}\tilde{\mu}^2\right) \quad (22)$$

Then,

$$\int S(x) dx = \exp\left(-\frac{\tilde{\lambda}}{2}\mu^2 + \frac{\tilde{\lambda}}{2}\tilde{\mu}^2\right) \left(\frac{\tilde{\lambda}}{2\pi}\right)^{-1/2} \quad (23)$$

$$\mathbb{E}\left[e^{f(\beta_t)}\right] = \int_{-\infty}^{\infty} \exp(f(\beta)) \mathcal{N}(\beta | \hat{\beta}_t, \lambda_t^{-1}) d\beta \quad (24)$$

$$\approx \int_{-\infty}^{\infty} \exp\left(f(\hat{\beta}_t) + f'(\hat{\beta}_t)(\beta - \hat{\beta}_t) + \frac{1}{2}f''(\hat{\beta}_t)(\beta - \hat{\beta}_t)^2\right) \mathcal{N}(\beta | \hat{\beta}_t, \lambda_t^{-1}) d\beta \quad (25)$$

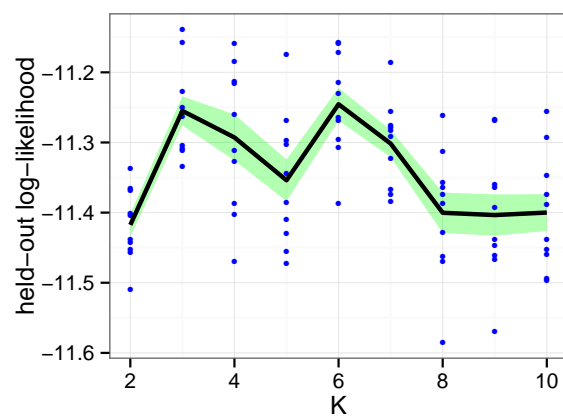
$$= \exp\left(f(\hat{\beta}_t) - f'(\hat{\beta}_t)\hat{\beta}_t\right) \exp\left(-\frac{\tilde{\lambda}_t}{2}\mu^2 + \frac{\tilde{\lambda}_t}{2}\tilde{\mu}^2\right) \left(\frac{\lambda_t}{\tilde{\lambda}_t}\right)^{1/2} \quad (26)$$

$$= \exp\left(f(\hat{\beta}_t) + \frac{1}{2}\frac{f'(\hat{\beta}_t)^2}{\lambda_t - f''(\hat{\beta}_t)}\right) \left(\frac{\lambda_t}{\lambda_t - f''(\hat{\beta}_t)}\right)^{1/2} \quad (27)$$

## Appendix: supplementary results

### Latent Dirichlet Allocation

10-fold cross validation



**Figure 6. Bold the first sentence.** Rest of figure 2 caption. Caption should be left justified, as specified by the options to the caption package.



## Geneset analysis (GSA)

We ran GSA [14]

**Table 1. GSA results testing up-regulation (hi) of Reactome network modules**

$p$ -val cutoff	$N_{\text{obs}}$	$\mathbb{E}[N]$	FDR	$N_{\text{obs}}^*$	$\mathbb{E}[N^*]$	FDR*
0.001	0	0.088	0.733	2	0.075	0.038
0.005	0	0.44	0.733	2	0.375	0.188
0.01	1	0.88	0.733	4	0.75	0.188
0.02	2	1.76	0.733	4	1.5	0.375
0.025	3	2.2	0.733	5	1.875	0.375
0.05	4	4.4	0.859	8	3.75	0.469
0.1	8	8.8	0.859	16	7.5	0.469
0.25	24	22	0.859	28	18.75	0.67
0.4	41	35.2	0.859	31	30	0.968
0.5	50	44	0.88	34	37.5	1

We ran **GSA** routine on Reactome network modules of size  $\geq 10$  with the following parameters set: total 1000 permutations, **Maxmean** statistics, **Two class unpaired**. We had less powerful results with **Two class paired**.

**Table 2. GSA results testing down-regulation (lo) of Reactome network modules**

$p$ -val cutoff	$N_{\text{obs}}$	$\mathbb{E}[N]$	FDR	$N_{\text{obs}}^*$	$\mathbb{E}[N^*]$	FDR*
0.001	0	0.088	0.819	1	0.075	0.075
0.005	0	0.44	0.819	2	0.375	0.188
0.01	0	0.88	0.819	4	0.75	0.188
0.02	2	1.76	0.819	4	1.5	0.375
0.025	2	2.2	0.819	5	1.875	0.375
0.05	4	4.4	0.819	9	3.75	0.417
0.1	9	8.8	0.819	17	7.5	0.441
0.25	26	22	0.819	28	18.75	0.67
0.4	43	35.2	0.819	32	30	0.938
0.5	49	44	0.898	34	37.5	1

We ran **GSA** routine on Reactome network modules of size  $\geq 10$  with the following parameters set: total 1000 permutations, **Maxmean** statistics, **Two class paired**. We had less powerful results with **Two class unpaired**.

**Table 3. GSA results testing up-regulation (hi) of BioGRID network modules**

$p$ -val cutoff	$N_{\text{obs}}$	$\mathbb{E}[N]$	FDR	$N_{\text{obs}}^*$	$\mathbb{E}[N^*]$	FDR*
0.001	0	0.085	0.85	5	0.204	0.041
0.005	0	0.425	0.85	5	1.02	0.204
0.01	0	0.85	0.85	5	2.04	0.34
0.02	2	1.7	0.85	12	4.08	0.34
0.025	2	2.125	0.944	13	5.1	0.392
0.05	4	4.25	0.944	22	10.2	0.434
0.1	9	8.5	0.944	47	20.4	0.434
0.25	19	21.25	1	73	51	0.699
0.4	29	34	1	84	81.6	0.971
0.5	35	42.5	1	92	102	1

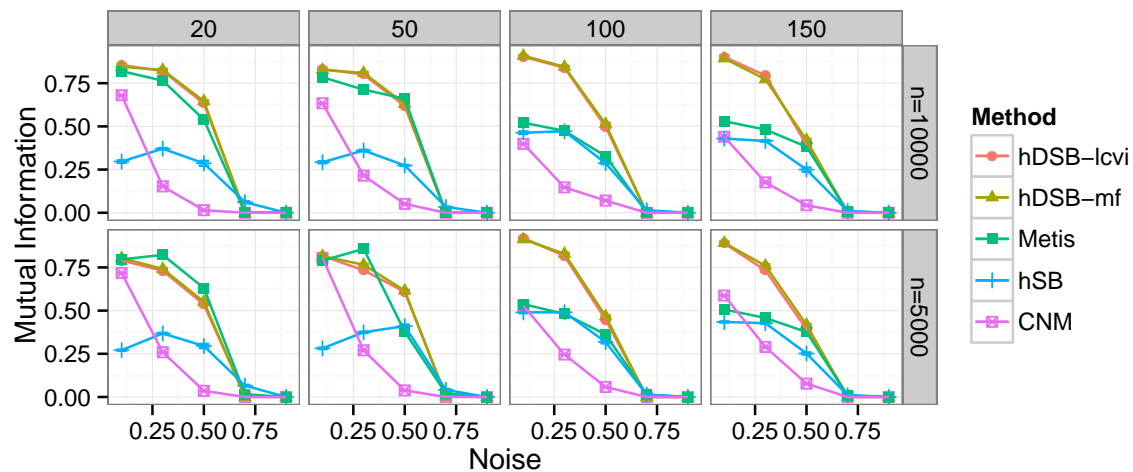
We ran **GSA** routine on BioGRID network modules of size  $\geq 10$  with the following parameters set: total 1000 permutations, **Maxmean** statistics, **Two class unpaired**. We had less powerful results with **Two class paired**.

**Table 4. GSA results testing down-regulation (lo) of BioGRID network modules**

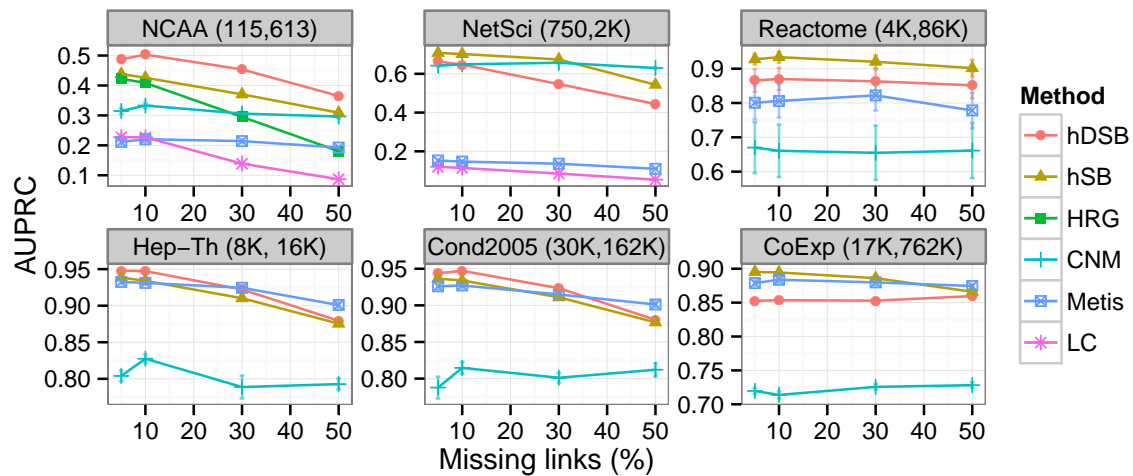
$p$ -val cutoff	$N_{\text{obs}}$	$\mathbb{E}[N]$	FDR	$N_{\text{obs}}^*$	$\mathbb{E}[N^*]$	FDR*
0.001	0	0.085	0.85	3	0.204	0.068
0.005	0	0.425	0.85	5	1.02	0.204
0.01	0	0.85	0.85	7	2.04	0.291
0.02	2	1.7	0.85	12	4.08	0.34
0.025	2	2.125	0.85	12	5.1	0.408
0.05	5	4.25	0.85	25	10.2	0.408
0.1	10	8.5	0.85	44	20.4	0.464
0.25	19	21.25	1	71	51	0.718
0.4	28	34	1	85	81.6	0.96
0.5	34	42.5	1	93	102	1

We ran **GSA** routine on BioGRID network modules of size  $\geq 10$  with the following parameters set: total 1000 permutations, **Maxmean** statistics, **Two class paired**. We had less powerful results with **Two class unpaired**.

## Network clustering methods



**Figure 7. Bold the first sentence.** Rest of figure 2 caption. Caption should be left justified, as specified by the options to the caption package.



**Figure 8. Bold the first sentence.** Rest of figure 2 caption. Caption should be left justified, as specified by the options to the caption package.