Bioinformatics, 36, 2020, i474–i481
doi: 10.1093/bioinformatics/btaa403
ISMB 2020

TopicNet: a framework for measuring transcriptional regulatory network change

Shaoke Lou[1,†], Tianxiao Li[2,†], Xiangmeng Kong[1,†], Jing Zhang[1], Jason Liu[1], Donghoon Lee[1] and Mark Gerstein[1,*]

[1]Department of Molecular Biophysics and Biochemistry and [2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint Authors.

Abstract

Motivation: Recently, many chromatin immunoprecipitation sequencing experiments have been carried out for a diverse group of transcription factors (TFs) in many different types of human cells. These experiments manifest large-scale and dynamic changes in regulatory network connectivity (i.e. network 'rewiring'), highlighting the different regulatory programs operating in disparate cellular states. However, due to the dense and noisy nature of current regulatory networks, directly comparing the gains and losses of targets of key TFs across cell states is often not informative. Thus, here, we seek an abstracted, low-dimensional representation to understand the main features of network change.

Results: We propose a method called TopicNet that applies latent Dirichlet allocation to extract functional topics for a collection of genes regulated by a given TF. We then define a rewiring score to quantify regulatory-network changes in terms of the topic changes for this TF. Using this framework, we can pinpoint particular TFs that change greatly in network connectivity between different cellular states (such as observed in oncogenesis). Also, incorporating gene expression data, we define a topic activity score that measures the degree to which a given topic is active in a particular cellular state. And we show how activity differences can indicate differential survival in various cancers.

Availability and Implementation: The TopicNet framework and related analysis were implemented using R and all codes are available at https://github.com/gersteinlab/topicnet.

Contact: mark@gersteinlab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, large-scale data on the interaction between proteins and DNA has enabled the construction of complex transcriptional regulatory networks (Liu *et al.*, 2015; Zhang *et al.*, 2014). These networks model the molecular program for gene transcription by representing genes and regulatory elements as nodes, and regulatory relationships as edges. Transcription factors (TFs), a major class of protein regulators in gene expression (Thompson *et al.*, 2015), are pivotal regulatory factors in these networks. Under different cellular conditions, TFs may undergo dramatic functional changes (or 'network rewiring') according to the gains and losses of their regulatory target genes. These rewiring events provide insight into differential cellular responses across conditions in the form of altered regulatory programs. Studies have revealed that network rewiring events and the altered regulatory programs they generate have strong phenotypic impacts (Assi *et al.*, 2019; Bhardwaj *et al.*, 2010).

However, quantification of network rewiring is challenging due to the condensed and complex nature of regulatory networks (Gerstein *et al.*, 2012). Genes from various functional modules, pathways and molecular complexes can play varying roles depending on their local associations with other genes. As a result, gains or losses of some gene connections may impact network alterations in a functionally significant way, while others may not. This indicates that identifying the gene functional subgroups and estimating the network rewiring at the subgroup level should be more robust and informative as compared to investigating changes of each individual gene.

These low-dimensional representations of the functional subgroups that underlie network data resemble semantic topics in documents. Based on this consideration, the low-dimensional representation can be constructed using topic modeling techniques, including latent Dirichlet allocation (LDA). LDA was proposed by Pritchard *et al.* (2000) for population genotype inference, and was 'rediscovered' by Blei *et al.* (2003) with applications in natural language processing as a simple and efficient means to extract latent topics from high-dimensional data. This approach has been successfully implemented in several biological scenarios that require decomposition and dimensionality reduction of data (Pinoli *et al.*, 2014; Wang *et al.*, 2011).

Here, we propose a method called TopicNet that makes use of various features of the LDA model to measure the regulatory potential, perturbation tolerance and intranetwork dynamics of TFs in terms of their target gene ‘topics’. To apply LDA, we represent the targets of a TF under a specific condition (cell line or tissue) as a ‘document’, with the TFs’ target genes as ‘words’ and latent functional subgroups as gene ‘topics’ comprised these words. We applied the procedure to a corpus including all of the regulatory networks inferred from 863 chromatin immunoprecipitation-sequencing (ChIP-seq) assays of the ENCODE dataset.

We first applied the LDA model to the regulatory network to characterize gene topics in an unsupervised fashion. From the trained model, the topic composition represents the distribution of words for each topic, and the topic weights, in turn, are the distribution over topics in a given document. The topic compositions can be further annotated for biological significance in terms of their relevance to various biological pathways and processes. The change in topic weights between documents can be used to quantify the network rewiring between two cellular conditions (i.e. between different cell types or different time points). Lastly, we defined a topic activity score by combining the cell-specific expression of target genes with the various topic compositions to characterize the overall activity of each topic, and demonstrated that this quantity is useful for predicting cancer survival.

In summary, our framework provides a straightforward quantitative representation of a TF regulatory network with biological significance, which could be further applied to many downstream analyses.

2 Materials and methods

2.1 Data preprocessing and construction of the regulatory network

We used 863 ChIP-seq experimental results for 387 TFs from the ENCODE portal for model training due to their high-quality control and consensus peak calling. In addition, we included ChIP-Atlas data collections with more than 6000 ChIP-seq experimental results to test the model. The number of target genes included in this dataset ranges from hundreds to thousands (Supplementary Fig. S11), and the TFs with the greatest availability among different cell lines include CTCF, EP300, MYC and REST (Supplementary Fig. S12).

From each ChIP-seq experiment, the regulatory target genes of specific TFs are defined as those with ChIP-seq peaks in proximal regions (±2500 bp) of their transcription start site. The cell type-specific TRN is then defined based on these results.

2.2 TopicNet—topic modeling

Each regulatory network for a TF in a specific cell line is regarded as an independent input document. We treat target genes that exist in these documents as ‘words’, which collectively constitutes the ‘vocabulary’ of the model. Based on existence of all genes as a regulatory target of the TF in the given condition, a document–gene matrix is then constructed. This matrix is used as the input for the LDA model. We use R package topicmodels for LDA learning and inference.

Let $M, \ K, \ V$ be the number of documents, the number of topics and the vocabulary size, respectively. In this scenario, each document is modeled as a mixture of topics, and each topic is a probabilistic distribution over genes. Each document i is represented as a N_i-dimensional vector W_i, where N_i is the number of genes in the document, and each element takes the value $1 \dots V$. The probability of observing a gene w_{ij} in a document W_i is determined by the mixture of topic compositions within the document and the probabilistic distribution of those topics. The existence of a word in a document is modeled as follows (Fig. 1c):

Given two priors (α as the prior for document-topic distribution, and β as the prior topic-gene distribution) we can sample

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

as the probability of all topics appearing in a document i, which constitutes the $M \times K$ matrix for document-topic distribution; and

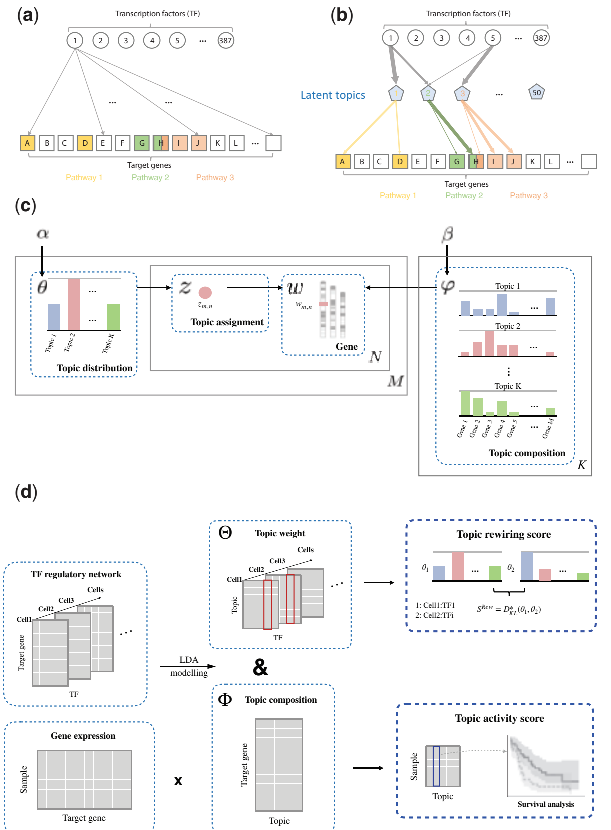

Fig. 1. Overview of method and data. (**a**) An example of a TF–gene regulatory network. (**b**) The high-dimensional regulatory network is decomposed into latent topics related to certain biological pathways. (**c**) Diagram explaining the meanings and biological relevance of the parameters in the LDA model. (**d**) General workflow of our analytical framework

$$\varphi_k \sim \text{Dirichlet}(\beta)$$

as the probability of all genes appearing in a topic k, which constitutes the $K \times V$ matrix for topic-gene distribution.

We can then sample latent topic assignment of each word j in document i as

$$z_{i,j} \sim \text{Multinomial}(\theta_i)$$

which is the topic that generates this gene. Each $z_{i,j}$ can take the value $1 \dots K$.

Given the membership of latent topics, the existence of genes in a document can be drawn as

$$w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$$

which constitutes the observed document-word matrix, where $\varphi_{z_{i,j}}$ is the topic-word distribution for the sampled topic $z_{i,j}$.

2.3 Model inference

Let W and Z be the collection of all aforementioned $w_{i,j}$’s and $z_{i,j}$’s indexed by document and gene position pair. Gibbs sampling can be performed on the Markov chain $\{W, Z\}$ to obtain the estimation of φ and θ. See Supplementary Methods for details.

For a stable and robust topic-gene composition matrix, we averaged the results of 100 runs. As the topics learned by the LDA model for each run were represented in randomized order, topics across different samplings were first mapped against each other based on the correlations of their composition. For any pair of outputs, each topic from the first run was assigned with the same ID as the topic it most strongly correlated with in the other run. We then produced the ensemble model by taking the median of the probability distribution over the composition for all topics that were mapped to the same ID.

Once we obtained the model, we could apply it to unseen documents with the same vocabulary and determine their posterior distribution over topics given the generative processes above.

2.4 Selection of topic numbers

The number of topics K used was selected using three criteria implemented by the R package ldatuning's FindTopicsNumber method:

1. The posterior likelihood of the data given the LDA model of different number choices (Griffiths and Steyvers, 2004) (blue). A higher value was preferred;
2. The Kullback–Leibler (KL) divergence of the document–gene matrix (Arun *et al.*, 2010b) (red). A lower value was preferred;
3. The average cosine distance r within topics (Cao *et al.*, 2009) (green). A lower value was preferred.

All three metrics reached optimal performance at around 50 topics (Supplementary Fig. S1). Based on these results, we used 50 as the number of topics for downstream analysis.

2.5 Reconstructed correlations

We performed LDA, non-negative matrix factorization (NMF) and K-means with 50 topics on each sample to obtain one 50-dimensional embedding vector of each sample for all three models. We represent the raw data for document i as a vector $v_i = [v_{i,1},\ v_{i,2}, \ldots, v_{i,N}]$ with binary values, where N is the number of all genes in the vocabulary. The embedding procedure for each method is as follows:

For LDA, we obtained the document-topic weight matrix Θ as described above. For each document i, the embedding vector is the weight of the 50 topics, which is the ith column of matrix Θ:

$$v_i^{\mathrm{LDA}} = \theta_i.$$

NMF decomposes the input matrix into two non-negative matrices: the feature matrix W, and the coefficient matrix H. For a document i, we use its weights as the embedding vector, which is the ith column of matrix W:

$$v_i^{\mathrm{NMF}} = w_i.$$

K-means identifies $k = 50$ clusters from the dataset, and each cluster is represented as its cluster centroid $\bar{v}_1,\ \bar{v}_2,\ \ldots,\ \bar{v}_{50}$, which is the average of all samples assigned to the respective cluster. Each document i is represented as the Euclidean distances between the raw vector and the 50 cluster centroids:

$$v_i^{\mathrm{KM}} = [d(v_i, \bar{v}_1),\ d(v_i, \bar{v}_2), \ldots, d(v_i, \bar{v}_{50})].$$

For each document pair in the dataset, we could calculate the Pearson correlation using the raw vectors and the 50-dimensional embedded vectors of the three methods. To evaluate whether the embedding retains correlations in the original data, the 'reconstructed' correlation calculated from the embedding vectors was then plotted against the 'original' correlation using the raw vectors. Linear regression was also performed between the reconstructed and original correlation for the three methods.

2.6 Evaluation of topic importance

The importance of the ensemble topics was evaluated by calculating the KL divergence between the topic weights of all of the documents and a background uniform distribution. Given the document-topic weight vector for all documents $\theta = [\theta_1,\ \theta_2,\ \ldots,\ \theta_M]$ (where M is the number of documents), the noise distribution is defined as:

$$\theta_{\mathrm{background}} = [1/M, 1/M, \ldots, 1/M].$$

The distance between the ensemble distribution and the null distribution is defined as the KL divergence:

$$D_{\mathrm{KL}}(\theta \| \theta_{\mathrm{background}}) = \sum\nolimits_i (\theta_i) \log\left(\frac{\theta_i}{\theta_{\mathrm{background}_i}}\right).$$

2.7 Gene set enrichment analysis

We used topic composition as the statistic for gene set enrichment analysis (GSEA). Gene sets in the C2, C5 and C6 categories from Molecular Signatures Database (MSigDB) (Subramanian *et al.*, 2005) were used in the analysis. GSEA was performed with the R package fgsea. The affinity between gene sets is defined by their overlapping of gene sets.

2.8 Quantification of network rewiring: topic rewiring score

For a pair of documents in a rewiring event of a given TF $(X_{\{\mathrm{TF,cell_1}\}},\ X_{\{\mathrm{TF,cell_2}\}})$, we can calculate the KL divergence between their topic weight vectors $\theta_{\{\mathrm{TF,cell_1}\}}$ and $\theta_{\{\mathrm{TF,cell_2}\}}$, here using the latter as reference:

$$D_{\mathrm{KL}}\left(\theta_{\{\mathrm{TF,cell_1}\}}\ \|\ \theta_{\{\mathrm{TF,cell_2}\}}\right) = \sum\nolimits_i \left(\theta_{\{\mathrm{TF,cell_1}\},i}\right) \log\left(\frac{\theta_{\{\mathrm{TF,cell_1}\},i}}{\theta_{\{\mathrm{TF,cell_2}\},i}}\right).$$

Since KL divergence is asymmetric depending on which distribution is used as reference, we consider the symmetrized KL divergence of the two directions to be a better metric for rewiring, which is:

$$D^{*}_{\mathrm{KL}}\ \left(\theta_{\{\mathrm{TF,cell_1}\}},\ \theta_{\{\mathrm{TF,cell_2}\}}\right) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(\theta_{\{\mathrm{TF,cell_1}\}}\ \|\theta_{\{\mathrm{TF,cell_2}\}}\right)\right.$$
$$\left. +\ D_{\mathrm{KL}}\left(\theta_{\{\mathrm{TF,cell_2}\}}\ \|\theta_{\{\mathrm{TF,cell_1}\}}\right)\right).$$

The rewiring score based on KL divergence identifies distinct clusters for closely related cell types like GM cell lines and fibroblasts in hierarchical clustering among CTCF-related documents as shown in Supplementary Fig. S13. This indicates that KL divergence is more robust and interpretable than Jaccard distance on raw vectors.

2.9 Topic activity score

Gene expression data for BRCA, LAML, LIHC and GBM patient samples from TCGA were obtained from GDC data portal (https://portal.gdc.cancer.gov/). The gene expression levels of every sample for each cancer type were formulated into an expression matrix E, where rows represent genes and columns represent samples. For each cancer type, the expression data was first quantile normalized. Only genes that appeared in the topic-gene matrix were retained. We then obtained the expression matrix E by first ranking the expression of the genes for each sample, and then transformed the value for gene i in the matrix of column j (corresponding to patient sample j) to $E_{i,j} = 1/\mathrm{rank}_{i,j}$, where $\mathrm{rank}_{i,j}$ is the rank of gene i in sample j.

For a sample t with expression vector E_t, the topic activity score is calculated as $S_t^{\mathrm{Act}} = \Phi^{\mathrm{T}} E_t$, where each element in the vector is the activity score of the corresponding target.

2.10 Definition of gain or loss genes

Given two TF regulatory networks under two conditions, we arbitrarily assigned one as an 'altered condition' and the other as the reference condition. We define 'gain' genes as those that are only regulated by the TF in the altered condition but not the reference condition (formally called 'gain genes in the altered condition') and the loss genes *vice versa* (called 'loss genes in the altered condition'). For annotation of selected topics, we were particularly interested in the gain or loss genes that are among the top-ranked genes of the topic (i.e. those that have high values in the topic composition). We took the intersection between these two sets and named the resulting set of genes as 'top-ranked gain/loss genes in the altered condition for topic k'.

2.11 PPI network analysis of topic-related target genes

The genes in the corpus were first sorted according to their contributions to the topic. Among the top 500 genes, those that were directly regulated by the given TF (i.e. bound by the TF in the corresponding ChIP-seq experiment) were selected and provided to STRING (Franceschini *et al.*, 2012). The resulting interaction graph contained the selected genes along with their first-layer neighbors.

2.12 Survival analysis

Clinical information for each patient regarding vital status, days to last follow-up and days to death were downloaded from the GDC data portal (https://portal.gdc.cancer.gov/). Records with missing information were discarded. Patients that were still alive in the record were right censored. The values of all 50 topics were used as variables to perform Cox proportional hazards (coxph) regression and implemented with the coxph function from R package survival, with days to death (or days to last follow-up for censored living patients) as the response.

We then selected the topics whose activity score achieved *P*-value <0.05 in the coxph analysis, for further analysis. For each of these candidate topics, the patients were then separated into two groups by the median activity score. The survival curve was then estimated using the Kaplan–Meier (KM) estimator. The topics that achieved the lowest *P*-value were selected and shown.

3 Results

3.1 TopicNet framework

We used an LDA model to decompose the high-dimensional regulatory network into a selected number of latent topics related to certain biological pathways (Fig. 1a and b). Based on the results, we constructed the TopicNet framework to include two parts: topic rewiring score and topic activity score (Fig. 1c and d).

3.2 LDA model

In our study, we treated the regulatory targets of a TF under a specific cellular condition as a ‘document’, denoted as $W_{\{\mathrm{TF,cell}\}}$. The target genes act as ‘words’ and constitute the general ‘vocabulary’ of the corpus. The LDA then identified functional topics from the genes in these documents (see Fig. 1c). The analogy between topic modeling terms and biological terms is elaborated in Supplementary Table S1.

We used published metrics to choose the number of topics *K*; in particular, we used $K=50$ as the optimal number for our model (Supplementary Fig. S1; see Section 2). Two important matrices can be inferred from the trained model:

1. The document-topic weight matrix Θ, which is cellular condition dependent, represents the weight of topics for all documents. Each column $\theta_{\{\mathrm{TF,cell}\}}$ is a vector of the distribution over topics for the corresponding document (note here {TF, cell} represents a single index), and the element $\theta_{\{\mathrm{TF,cell}\},k}$ represents the weight of topic *k* in document $W_{\{\mathrm{TF,cell}\}}$.
2. The topic-gene composition matrix Φ, which is cell independent, indicates the distribution over the target genes within the topics. Each column φ_k represents the composition of a topic (i.e. the distribution of target genes or the contribution of genes to the topic). $\varphi_{k,j}$ represents the contribution of gene *j* to topic *k*.

We further developed the topic rewiring score and topic activity score based on these two matrices.

3.3 Topic rewiring score

Raw network rewiring can be described as the differences between documents of the same TF in two cell types, $W_{\{\mathrm{TF,cell_1}\}}$ and $W_{\{\mathrm{TF,cell_2}\}}$. For comparisons between two documents in terms of topics, we defined the network rewiring score as $S^{\mathrm{Rew}}\left(\theta_{\{\mathrm{TF_1,cell_1}\}},\ \theta_{\{\mathrm{TF_2,cell_2}\}}\right)$ as a symmetrized KL divergence between the topic weights

$$\theta_{\{\mathrm{TF_1,cell_1}\}} \text{ and } \theta_{\{\mathrm{TF_2,cell_2}\}}.\ S^{\mathrm{Rew}}\left(\theta_{\{\mathrm{TF_1,cell_1}\}},\ \theta_{\{\mathrm{TF_2,cell_2}\}}\right) = D^{*}_{\mathrm{KL}}\ \left(\theta_{\{\mathrm{TF,cell_1}\}},\ \theta_{\{\mathrm{TF,cell_2}\}}\right).$$

3.4 Topic activity score

Note that Θ gives an effective weighting to topics in a given condition. However, often the cell type-specific regulatory network is lacking but gene expression is available. In these cases, we can define an effective topic activity score. Given a sample *t* with gene expression vector E_t, we compute the vector of activity score for all topics by multiplying the gene expression vector by the composition matrix Φ (i.e. $S^{\mathrm{Act}}_t = \Phi^{\mathrm{T}} E_t$).

3.5 Validation of the LDA model

We determined $K=50$ as an optimal number of topics using several metrics (Arun *et al.*, 2010a; Cao *et al.*, 2009; Griffiths and Steyvers, 2004). We then tested how similarities in the original data can be preserved compared with two other algorithms, NMF and *K*-means, using methods presented by Guo and Gifford (2017). Each method gives a 50-dimensional representation of the samples. For every pair of samples, we computed their correlation in terms of both the raw data (‘raw’ correlation) and the 50-dimensional representation (‘reconstructed’ correlation) from each method. Among the three algorithms, LDA could reconstruct the raw correlations better than the other two (Fig. 2a and Table 1). T-distributed stochastic neighbor embedding (T-SNE) of the 50-dimensional representation also demonstrates LDA’s ability to preserve similarities because samples about the same TF tend to form distinct clusters in the embedding space (Fig. 2b).

We also investigated the connection of models that were inferred by the varying topic number *K*. We associated topics from models with topic number $K=5,\ 10,\ 20,\ 50$ based on the correlation of their topic-gene composition matrix, and observed a topic–subtopic hierarchical structure among these models (Fig. 2c). This demonstrates that topics may be further split into multiple subtopics when increasing the total number of topics in the model.

The topics can be associated with protein complexes and functional modules. We performed hierarchical clustering on all TF documents in HeLa cells using their topic weights. We observed distinct clusters, indicating coactivation and collaborative binding (Supplementary Fig. S2). Among them, we observed clustering of nuclear transcription factor (NFY) subunits (NFYA and NFYB) and FOS, which have been previously shown to colocalize extensively (Fleming *et al.*, 2013). We observed similar groupings for CTCF and cohesin subunits SMC3 and RAD21, the latter of which frequently cobinds with CTCF (Parelho *et al.*, 2008; Rubio *et al.*, 2008).

3.6 Functional annotation of identified gene topics

We calculated the importance of each topic by measuring the KL divergence between the topic weights across all documents against a background uniform distribution. Important topics should be more specific and, therefore, only highly represented in some documents. In contrast, the topics that show close to a uniform distribution would have almost equal weights in most documents and would be less interesting. The rank of all 50 topics’ importance is shown in Figure 3a. Of particular interest are Topics 3 and 14: from the top genes of these two topics, we identified several functional groups such as transcription regulation, cell proliferation, metabolism and mitosis (Fig. 3b).

To investigate the biological significance of each topic, we annotated their functions using GSEA. For each topic, the probability distribution over target genes can be used directly as the statistics for GSEA. Using C2 and C5 gene sets from the MsigDB (Subramanian *et al.*, 2005), Topic 3, which showed the highest importance, was enriched with gene sets related to breast cancer and glioblastoma tumors (Supplementary Fig. S3).

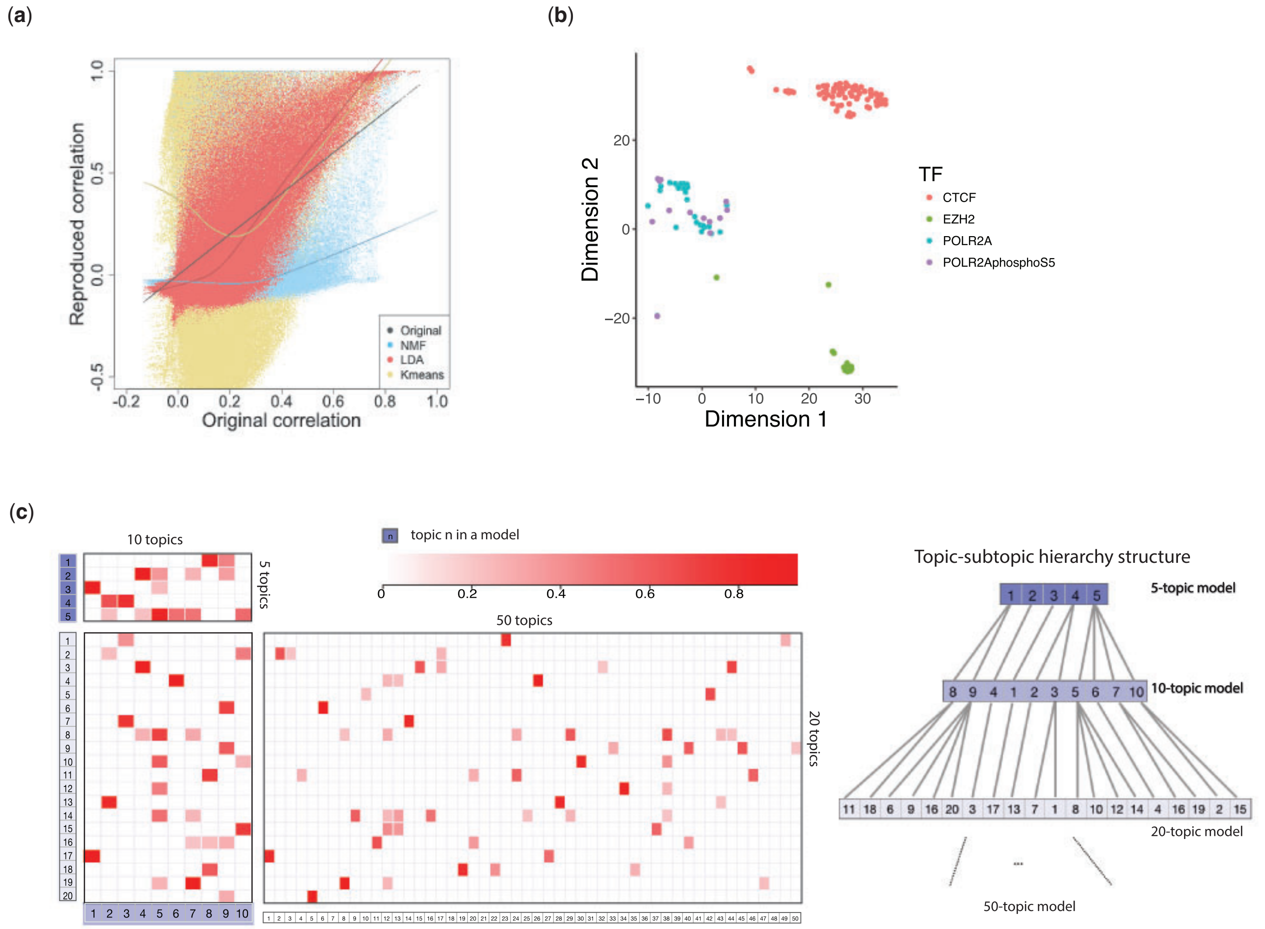

Fig. 2. Tuning and performance evaluation of the LDA model. (**a**) Reproduced pairwise correlations after applying three-dimensionality reduction methods plotted against original correlations. (**b**) T-SNE embedding of topic weights (50 topic model) for data samples on CTCF, EZH2, POL2A and POL2AphosphoS5. (**c**) Left: Correlation between topic compositions identified by LDA models with different topic numbers (5, 10, 20, 50). Right: Hierarchical structure of topics inferred from topic correlations (topic correlation is ignored if it has value <0.4 and is not the maximum in its column)

Table 1. Linear regression of the reconstructed correlation against the raw correlation

Method	Correlation	Linear regression slope	Linear regression R^2
K means	0.1047	0.3828	0.0110
NMF	0.4047	0.4365	0.1638
LDA	0.7886	1.3519	0.6219

3.7 Quantification of TF regulation rewiring using topic weights

For each TF, we calculated the pairwise topic rewiring score for all available cell types. The average rewiring score for a TF reflects its cell type specificity, as higher values correspond to greater difference between cell lines (i.e. higher specificity) (Fig. 4a and Supplementary Fig. S4). We observed that many TFs with higher cell specificity were related to biological processes displaying highly variable regulatory activity across conditions, such as pluripotency, cell cycle regulation, tumor suppression or tumorigenesis, including EP300 (Kim *et al.*, 2013), BCL11A (Dong *et al.*, 2017; Khaled *et al.*, 2015), ZBTB33 (Pozner *et al.*, 2016) and JUND (Caffarel *et al.*, 2008; Millena *et al.*, 2016). On the contrary, TFs with more constant roles such as NR2C2 (O'Geen *et al.*, 2010) showed very little difference between cell types. Interestingly, ZNF274 and SIX5, which have been shown to relate to CTCF binding sites (Hong and Kim, 2017), also showed low specificity, similar to CTCF. Supplementary Figure S5 lists the individual rewiring events with top values. Many of these events involve TFs with high cell-type specificity, such as EP300, SUZ12, ZBTB33 and FOS.

We pinpointed two cell lines, GM12878 and K562, and studied specific rewiring events for several TFs. Among the 69 TFs shared in both cell lines, ZBTB33 and EP300 showed the highest rewiring values. Specifically, Topic 49 and 16 showed the greatest difference in the rewiring of ZBTB33 (Fig. 4b), and Topic 34 and 10 in that of EP300 (Fig. 4f). For these topics, the majority of the top-ranked genes are true targets in the respective cell line (Fig. 4c and g, Supplementary Fig. S6).

Given a specific TF, we defined 'gain' target genes in K562 compared to GM12878 as those that are exclusively present in the former, and 'loss' target genes as those exclusively present in the latter. In this scenario, we were particularly interested in the gain or loss genes among those with high contribution to the topic (i.e. the top-ranked genes) that showed a major difference.

For ZBTB33, Topic 49 showed a very high weight in GM12878. We found that the top-ranked loss genes in K562 for Topic 49 were enriched in the gene set related to cell cycle and cell division function (Fig. 4d). The loss of ZBTB33 regulation resulted in higher expression of its target genes in K562, corresponding to known deacylation and transcriptional suppressive roles of ZBTB33 (Pozner *et al.*, 2016; Fig. 4e). Another highly rewired TF, EP300 (a known

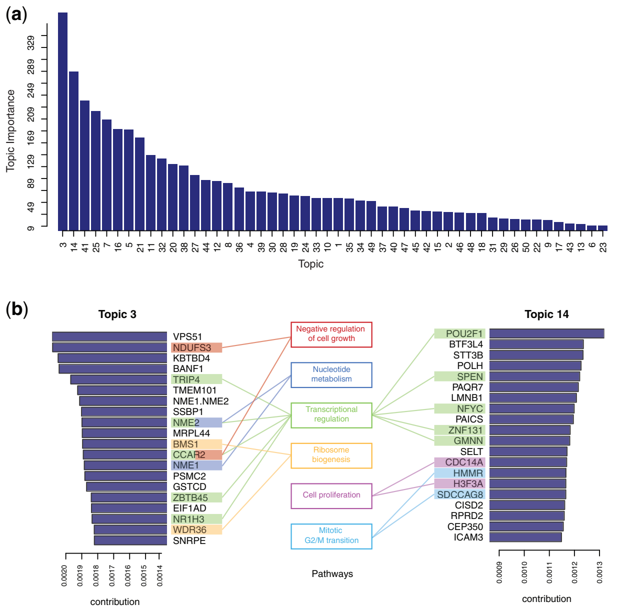

Fig. 3. Annotation of the identified topics. (**a**) All 50 topics ranked by their importance measure (average KL divergence of topic weights against uniform distribution across all samples). (**b**) Genes with the highest contributions to the top-ranked topics (Topics 3 and 14) along with their related functional roles

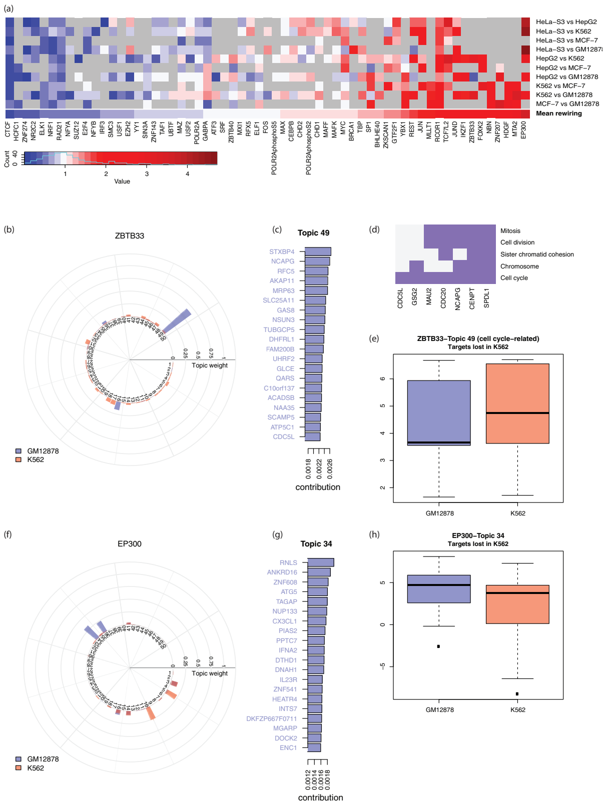

Fig. 4. Quantified rewiring analysis using the identified topics. (**a**) Heatmap of rewiring of TFs in selected cell lines against GM12878 (gray grids correspond to unavailable data). (**b**) Topic weight for the rewiring event of ZBTB33 in GM12878 and K562. (**c**) Top-weighted genes of the two topics with greatest difference in the rewiring event of ZBTB33 in GM12878 and K562. Colored gene names indicate true regulatory targets in the ChIP-seq experiment. (**d**) Functional clustering of top-ranked genes of Topic 49 that are related to cell cycle and cell division. (**e**) Expression of the top-ranked genes of Topic 49 that are lost in K562. (**f**) Topic weight for the rewiring event of EP300 in GM12878 and K562. (**g**) Top-ranked genes of the two topics with the greatest difference in the rewiring event of EP300 in GM12878 and K562. Colored gene names indicate true targets in the ChIP-seq experiment. (**h**) Expression of the top-ranked genes of Topic 34 that are lost in K562

transcriptional activator), regulates a wide range of genes from different functional groups. In concordance with EP300's function, Topics 5 and 10 were deficient in K562 and the top-ranked loss genes from these topics were significantly downregulated in K562. For topics of EP300 that were highly represented in K562 (34, 36), we observed an adverse trend (Fig. 4h and Supplementary Fig. S6). To summarize, topics showing major differences in the rewiring of these TFs were related to the TFs' molecular functions. Comparatively, TFs with low rewiring scores, like CTCF and ZNF274, had almost identical topic distributions (Supplementary Fig. S7).

These results further demonstrate the potential of using a rewiring score derived from LDA as a quantitative measure of change. The rewiring events with high scores could be associated with previously reported biological significance of corresponding TFs.

3.8 Network rewiring shows dynamic topic changes in a time-course study

Temporal changes of topic weights could be used to represent dynamic responses in the cellular regulatory system. To demonstrate this, we further applied our methods to the time series regulatory networks for estrogen receptor (ESR1) in MCF-7 cell lines at 2, 5, 10, 40 and 160 min after estradiol treatment (Guertin *et al.*, 2014). The rewiring score between these time points showed a transition of the topic weights across time points. The first few minutes after estradiol treatment showed dramatic topic changes, followed by a gradual trend to stability (Fig. 5a).

The time-course pattern of the topic membership demonstrated that Topic 3 had the highest weight prior to treatment. Later, Topic 4 became the most prominent topic after a short fluctuation, which experienced a sharp increase at the 10 min time point and then underwent a gradual decrease while remaining dominant until completion (160 min) (Fig. 5b). We then studied the roles of ESR1 target genes that are related to these highly represented topics. At 0 min, true ESR1 target genes that were top-ranked in Topic 3 included genes that are most related to cell proliferation functions: ribosomal functions, protein folding and mRNA splicing (Supplementary Fig. S8). At the 10 min time point for Topic 4, top-ranked target genes included genes related to signal transduction and apoptosis, with some interacting directly with EP300. This result is consistent with a study demonstrating the redistribution of EP300 target genes after the treatment of estradiol (Guertin *et al.*, 2014; Supplementary Fig. S8).

The treatment of estradiol turned the MCF-7 cell line's topic from Topics 3 to 4, which indicates the top-weighted genes in Topic 4, especially for the gained genes, may play a crucial role in the treatment. We compared the nascent gene expression of 10 and 40 min with 0 min for the gained top-ranked genes in Topic 4. These gained top-ranked genes showed significant (P-value $<10^{-5}$) upregulation in 10 and 40 min (Fig. 5c and d).

3.9 Topic activity score and its relationship to tumor survival

The topic activity score incorporates cell type-independent topic composition with cell type-specific gene expression and can be associated with clinical significance. We used patient samples of three cancer types with clinical information from The Cancer Genome Atlas (TCGA) data portal for breast cancer (BRCA), acute myeloid leukemia (LAML) and liver hepatocellular carcinoma (LIHC). We evaluated the topic activity scores for each cancer type and used them for survival analysis. We found that the activity scores of several topics were associated with patient survival (Fig. 6). For each

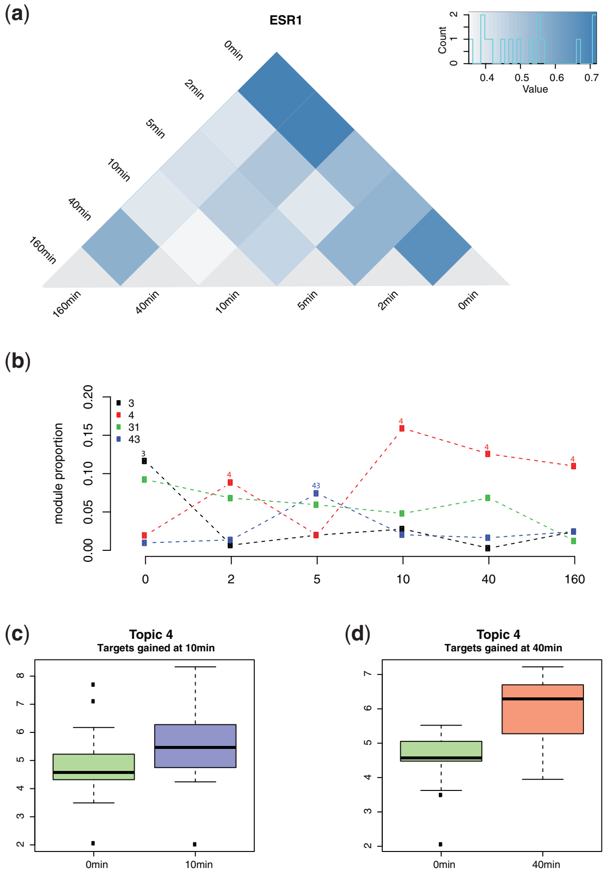

Fig. 5. ESR1 regulation dynamics represented by topic weights. (**a**) Pairwise KL divergence between the topic weights of ESR1 in all time points. (**b**) Time-course change of the topics with the highest weights across the time points. (**c**) and (**d**) Expression of the genes gained in 10 (c) and 40 min (d), as compared to 0 min, that are among the top-ranked genes of Topic 4

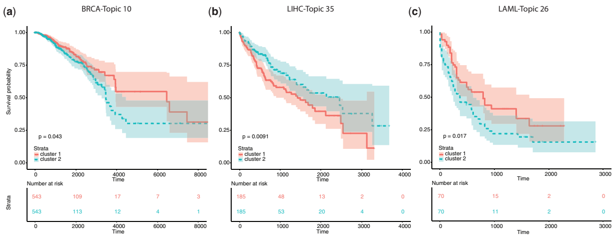

Fig. 6. Topic activity level related to cancer survival. (**a**)–(**c**) KM survival curve of thee cancer types using their related topic activity levels: BRCA with Topic 10 (a), LIHC with Topic 35 (b) and LAML with Topic 25 (c)

cancer type, we characterized the biological relevance of the most predictive topic:

1. The activity score of Topic 10 was predictive for the survival of BRCA patients (Fig. 6a). Correspondingly, Topic 10 was highly represented in the document of {GATA3, MCF-7} (Zhang *et al.*, 2017), and its composition was enriched with the CtIP-associated gene set (Supplementary Fig. S9a). GATA3 and CtIP are known to interact with each other and functionally correlate with breast cancer: GATA3 can regulate BRCA1 (Zhang *et al.*, 2017) and CtIP forms a repressor complex with BRCA1 whose removal accelerates tumor growth (Furuta *et al.*, 2006) (Supplementary Fig. S10a).
2. The activity score of Topic 26 was predictive of LAML patient survival. For Topic 26, which was highly represented in the document {NR2C2, K562} (Fig. 6b and Supplementary Fig. S9b), its compositions were also enriched with genes upregulated in response to activation of the cAMP signaling pathway (van Staveren *et al.*, 2006) (Supplementary Fig. S10b). NR2C2 can be induced by cAMP (Liu *et al.*, 2009) and has been found to be significantly activated in almost all the cancer types (Falco *et al.*, 2016).
3. The activity score of Topic 35 predicted survival outcome of LIHC patients with high accuracy. Topic 35 was highly represented in documents {ATF3, HepG2} and {JUN, HepG2} (Fig. 6c and Supplementary Fig. S9c), and its composition was enriched with gene sets that were upregulated in response to overexpression of the proto-oncogene MYC (Bild *et al.*, 2006) (Supplementary Fig. S10c). Among these factors, ATF3 is a cAMP-responsive element and acts as a tumor suppressor in LIHC (Chen *et al.*, 2018). JUN is a known oncogene and promotes liver cancer (Maeda and Karin, 2003). MYC is also a highly expressed oncogene and correlates with high proliferative activity (Zheng *et al.*, 2017).

In summary, we found associations between the survival-related topics and their biological significance via the activity and function of the TFs that regulate these topics. These results further validate the biological relevance of the identified topics, indicating their potential as prognostic markers and sources for biomarker discovery.

4 Discussion

Rewiring analysis of the regulatory network could provide critical information about the alteration of molecular programs across conditions. Several attempts have been made to derive an effective procedure for identifying network rewiring (Assi *et al.*, 2019; Han and Goetz, 2019; Shou *et al.*, 2011; Xu *et al.*, 2018). In this study, we successfully developed the TopicNet framework. Our framework extracts a low-dimensional representation of a network in the form of functional topics and defines a network rewiring score and topic-weighted activity score. We then demonstrated the application of our framework and showed that the network rewiring score can aid in the identification of the functional rewiring of TFs between cellular conditions. Moreover, the topic-weighted activity score can be applied to sample-specific cohort data for the prediction of patient survival.

To evaluate our framework, we also investigated the biological meaning of the identified topics. We interpreted the learned topics by utilizing two important matrices inferred from the model: document-topic weight and topic-gene composition matrices. The former demonstrates the activity of the topic as a distinctive functional module, and the latter indicates possible biological functions or pathways that the topic represents. Rewiring analysis using topic weights is both efficient and highly interpretable with gene topics serving as bridges between TFs and genes.

Our framework facilitates comparison between regulatory networks under different conditions from multiple sources. Thus, the analysis can be extended to various studies where network changes are of major interest. For example, time-course network changes, such as those after treatment or during the cell cycle, could help pinpoint TFs, genes and pathways that play critical roles in these processes. Having demonstrated the potential application of our method on time- course data, we expect our method to offer valuable insights into network dynamics studies in the future.

Similar to the conclusion from our comparison, LDA has been shown to be advantageous over some other common dimensional reduction techniques (e.g. NMF, SVD and pLSI) in terms of performance and interpretability (Liu *et al.*, 2011; Stevens *et al.*, 2012). The most advantageous feature of LDA is the control of sparsity of low-dimensional representations, which gives a robust representation for the noise reduction.

Furthermore, several extensions could be introduced for future studies. In our framework, we treated all TF–cell line pairs as independent regardless of possible relationships between cell lines and TFs. In addition, the low-dimensional gene topic defined by TopicNet is a simplified representation, which does not take into account the complex and hierarchical gene–gene interactions. These relations can be modeled by reorganizing the data into a cell–TF-target 3D tensor and training on all three dimensions simultaneously. For example, a recent study integrated the incomplete epigenome data as a cell-assay-position 3D tensor and used an artificial neural network to impute the missing data and find latent representations of the epigenome (Schreiber *et al.*, 2019). We are aware that recent advances in topic modeling and other machine learning methods have enabled modeling of more complex dependencies and structures (Blei and Lafferty, 2006; Momeni *et al.*, 2018; Zhou *et al.*, 2017). Though LDA could capture some of these dependencies in an unsupervised fashion, we expect incorporation of such information would help identify even more meaningful patterns from the regulatory network.

Acknowledgements

The authors thank all the comments from the reviewers and members of the Gerstein Lab.

Funding

This work was supported by the National Human Genome Research Institute (NHGRI) [U24 HG 009446] and and Williams professorship funds.

Conflict of Interest: none declared.

References

Arun,R. *et al.* (2010a). On finding the natural number of topics with latent Dirichlet allocation: some observations. In Paper presented at the *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Vol. Part I. Springer-Verlag, Hyderabad, India.

Arun,R. *et al.* (2010b). *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations*, Berlin, Heidelberg.

Assi,S.A. *et al.* (2019) Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat. Genet.*, **51**, 151–162.

Bhardwaj,N. *et al.* (2010) Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci. Signal.*, **3**, ra79.

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

Blei,D.M. and Lafferty,J.D. (2006). Dynamic topic models. In Paper presented at the *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA.

Blei,D.M. *et al.* (2003) Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Caffarel,M.M. *et al.* (2008) JUND is involved in the antiproliferative effect of Delta9-tetrahydrocannabinol on human breast cancer cells. *Oncogene*, **27**, 5033–5044.

Cao,J. *et al.* (2009) A density-based method for adaptive LDA model selection. *Neurocomputing*, **72**, 1775–1781.

Chen,C. *et al.* (2018) ATF3 inhibits the tumorigenesis and progression of hepatocellular carcinoma cells via upregulation of CYR61 expression. *J. Exp. Clin. Cancer Res.*, **37**, 263.

Dong,H. *et al.* (2017) High BCL11A expression in adult acute myeloid leukemia patients predicts a worse clinical outcome. *Clin. Lab.*, **63**, 85–90.

Falco,M.M. *et al.* (2016) The pan-cancer pathological regulatory landscape. *Sci. Rep.*, **6**, 39709.

Fleming,J.D. *et al.* (2013) NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Res.*, **23**, 1195–1209.

Franceschini,A. *et al.* (2012) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Furuta,S. *et al.* (2006) Removal of BRCA1/CtIP/ZBRK1 repressor complex on ANG1 promoter leads to accelerated mammary tumor growth contributed by prominent vasculature. *Cancer Cell*, **10**, 13–24.

Gerstein,M.B. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.

Griffiths,T.L. and Steyvers,M. (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA*, **101**, 5228–5235.

Guertin,M.J. *et al.* (2014) Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Mol. Endocrinol.*, **28**, 1522–1533.

Guo,Y. and Gifford,D.K. (2017) Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics*, **18**, 45.

Han,Y. and Goetz,S.J. (2019) Measuring network rewiring over time. *PLoS One*, **14**, e0220295.

Hong,S. and Kim,D. (2017) Computational characterization of chromatin domain boundary-associated genomic elements. *Nucleic Acids Res.*, **45**, 10403–10414.

Khaled,W.T. *et al.* (2015) BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat. Commun.*, **6**, 5987.

Kim,M.S. *et al.* (2013) Frameshift mutations of tumor suppressor gene EP300 in gastric and colorectal cancers with high microsatellite instability. *Hum. Pathol.*, **44**, 2064–2070.

Liu,N.C. *et al.* (2009) Activation of TR4 orphan nuclear receptor gene promoter by cAMP/PKA and C/EBP signaling. *Endocrine*, **36**, 211–217.

Liu,Z. *et al.* (2011) Performance evaluation of latent Dirichlet allocation in text mining. In Paper presented at the 2011 *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, July 26–28. Shanghai, IEEE.

Liu,Z.P. *et al.* (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database (Oxford)*, **2015**, bav095.

Maeda,S. and Karin,M. (2003) Oncogene at last—c-Jun promotes liver cancer in mice. *Cancer Cell*, **3**, 102–104.

Millena,A.C. *et al.* (2016) JUND is required for proliferation of prostate cancer cells and plays a role in transforming growth factor-beta (TGF-beta)-induced inhibition of cell proliferation. *J. Biol. Chem.*, **291**, 17964–17976.

Momeni,E. *et al.* (2018) *Modeling Evolution of Topics in Large-scale Temporal Text Corpora*. Twelfth International AAAI Conference on Web and Social Media. Palo Alto, California.

O'Geen,H. *et al.* (2010) Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics*, **11**, 689.

Parelho,V. *et al.* (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**, 422–433.

Pinoli,P. *et al.* (2014) Latent Dirichlet allocation based on Gibbs sampling for gene function prediction. In *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. Honolulu, HI. IEEE. <Go to ISI>://WOS:000345738200017.

Pozner,A. *et al.* (2016) Cell-specific Kaiso (ZBTB33) regulation of cell cycle through cyclin D1 and cyclin E1. *J. Biol. Chem.*, **291**, 24538–24550.

Pritchard,J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rubio,E.D. *et al.* (2008) CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. USA*, **105**, 8309–8314.

Schreiber,J. *et al.* (2019) Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *Genome Biol.*, **21**, 81.

Shou,C. *et al.* (2011) Measuring the evolutionary rewiring of biological networks. *PLoS Comput. Biol.*, **7**, e1001050.

Stevens,K. *et al.* (2012). Exploring topic coherence over many models and many topics. In Paper presented at the *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952-961. Association for Computational Linguistics, Jeju Island, Korea.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Thompson,D. *et al.* (2015) Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu. Rev. Cell Dev. Biol.*, **31**, 399–428.

van Staveren,W.C. *et al.* (2006) Gene expression in human thyrocytes and autonomous adenomas reveals suppression of negative feedbacks in tumorigenesis. *Proc. Natl. Acad. Sci. USA*, **103**, 413–418.

Wang,H.J., *et al.* (2011) Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One*, **6**, e17243.

Xu,T. *et al.* (2018) Identifying gene network rewiring by integrating gene expression and gene network data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **15**, 2079–2085.

Zhang,F. *et al.* (2017) The transcription factor GATA3 is required for homologous recombination repair by regulating CtIP expression. *Oncogene*, **36**, 5168–5176.

Zhang,S. *et al.* (2014) Profiling the transcription factor regulatory networks of human cell types. *Nucleic Acids Res.*, **42**, 12380–12387.

Zheng,K. *et al.* (2017) c-MYC-making liver sick: role of c-MYC in hepatic cell function, homeostasis and disease. *Genes (Basel)*, **8**, 123.

Zhou,H. *et al.* (2017) Topic evolution based on the probabilistic topic model: a review. *Front. Comput. Sci.*, **11**, 786–802.