TopicNet: a framework for measuring transcriptional regulatory network change

Shaoke Lou[1,3], Tianxiao Li[2,3], Xiangmeng Kong[1,3], Jing Zhang[1], Jason Liu[1], Donghoon Li[1], Mark Gerstein[1*]

[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[2] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
[3] These authors contribute equally

*Correspondence: pi@gersteinlab.org

Summary

Next generation sequencing data highlights comprehensive and dynamic changes in the human gene regulatory network. Moreover, changes in regulatory network connectivity (network "rewiring") manifest different regulatory programs in multiple cellular states. However, due to the dense and noisy nature of the connectivity in regulatory networks, directly comparing the gains and losses of targets of key TFs is not that informative. Thus, here, we seek a abstracted lower-dimensional representation to understand the main features of network change. In particular, we propose a method called TopicNet that applies latent Dirichlet allocation (LDA) to extract meaningful functional topics for a collection of genes regulated by a TF. We then define a rewiring score to quantify the large-scale changes in the regulatory network in terms of topic change for a TF. Using this framework, we can pinpoint particular TFs that change greatly in network connectivity between different cellular states. This is particularly relevant in oncogenesis. Also, incorporating gene-expression data, we define a topic activity score that gives the degree that a topic is active in a particular cellular state. Furthermore, we show how activity differences can highlight differential survival in certain cancers.

Introduction

In recent years, large-scale data on the interaction between proteins and genes has enabled the construction of complex transcriptional regulatory networks (Zhang et al., 2014, Liu et al., 2015). 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 regulator 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 (Bhardwaj et al., 2010, Assi et al., 2019).

However, quantification of network rewiring is challenging due to regulatory network's condensed and complex nature (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 from the subgroup level should be more robust and informative, as compared investigating change of every individual gene. Thus, a low-dimensional representation of the regulatory network is required.

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

In this study, we propose a method called TopicNet that makes use of various features in the LDA model to measure the regulatory potential, perturbation tolerance, and intra-network dynamics of TFs in terms of their target gene topics. The training corpus includes all the regulatory networks inferred from 863 chromatin immunoprecipitation-sequencing (ChIP-seq) assays of the ENCODE dataset.

We first apply an LDA model to the data to characterize the gene “topics” in an unsupervised fashion. From the trained model we could infer the topic component as the distribution over words for each topic, and the topic weight as the distribution over the topics for a given document. The topic component can be further annotated for biological significance including their relevance to certain biological pathways and processes. The topic weights of each document can be used to quantify the rewiring between two cellular conditions, such as different cell types or time points after treatment. Lastly, we define a topic activity score by incorporating the cell-specific expression of target genes and topic composition and characterize specific topics with expression scores associated with patient survival of several cancer types.

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

Results

TopicNet framework

Based on LDA model, we constructed the TopicNet framework including two parts: topic rewiring score and topic activity score (Fig1).

LDA model

In our study, we treated the regulatory targets of a TF under a specific cellular condition as a document, denoted as $W_{\{TF,cell\}}$. The target genes act as "words" and constitute the general "vocabulary" of the corpus. The LDA then identified the functional topics from the genes in these documents (as described in Fig 1b). We used published metrics to choose the number of topics K; in particular, we use K=50 as the optimal number for our model (Fig s1; see "Methods" for more details).

Two important matrices can be inferred from the trained model, and their meanings are as follows:

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

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

Topic Rewiring score

Raw network rewiring can be described as the differences between documents of the same TF in two cell types, $W_{\{TF,cell1\}}$ and $W_{\{TF,cell2\}}$. For comparison between two documents in terms of topics, we define network rewiring score as $S^{Rew}\left(\theta_{\{TF_1,cell_1\}}, \theta_{\{TF_2,cell_2\}}\right)$ as a symmetrized Kullback-Leibler (KL) divergence between the topic distributions $\theta_{\{TF_1,cell_1\}}$ and $\theta_{\{TF_2,cell_2\}}$.

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

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 more 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 expression score for all topics by multiplying It with the component matrix Φ, i.e. $S_t^{Act} = \Phi E_t$.

Validation of LDA model

We determined T=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, non-negative matrix factorization (NMF) and K-means, using methods presented by Guo et al., 2017 (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).

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

We also investigated the connection of model that inferred by the different topic number T. We associated topics from models with the topic number $T = 5, 10, 20, 50$ based on the correlation of their topic-gene component matrix and observed a natural hierarchical structure among these models (Fig 2c). This demonstrates that changing the topic number adjusts the "resolution" of the pattern recognition process as a model with more topics tends to detect more detailed patterns, while one with fewer topics can detect higher-level ones.

The topics can be associated with protein complex and functional modules. We performed hierarchical clustering on all TF documents in HeLa using their topic weights. Distinct clusters can be observed indicating co-activation and collaborative binding (Fig s2). Among them, nuclear transcription factor (NFY) subunits have been shown to co-localize with Fructooligosaccharides

(FOS) extensively (Fleming et al., 2013), and FOS, NFYA and NFYB are clustered together. Similar grouping is also observed for CTCF and cohesin subunits SMC3 and RAD21, as the latter frequently co-bind with CTCF (Rubio et al., 2008, Parelho et al., 2008).

Functional annotation of identified gene topics

We calculated the importance of each topic by measuring the Kullback-Leibler (KL) divergence between the topic weight of all documents against a background uniform distribution. We expect important topics should be more specific and only highly represented in some documents. The topics that are close to a uniform distribution have almost equal weights in most documents and will be less interesting. The rank of all 50 topics' importance is shown in Fig 3a. Of particular interest are Topic 3 and 14: from the top genes of these two, 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 gene set enrichment analysis (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 Molecular Signatures Database (MsigDB) (Subramanian et al., 2005), Topic 3, the one with the highest importance, is enriched with gene sets related to breast cancer and glioblastoma tumors (Fig s3).

Quantification of TF regulation rewiring using topic weights

For each TF, we calculated the topic rewiring score for every pair of cell type. 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, Fig s4). It can be observed that many TFs with higher cell specificity relate 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 (Khaled et al., 2015, Dong et al., 2017), 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) show very little difference between cell types. Interestingly, ZNF274 and SIX5, which were shown to relate to CTCF binding sites (Hong and Kim, 2017), also have low specificity similar to CTCF. Fig 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 show the highest rewiring values. Specifically, Topic 49 and 16 show 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 their top-rank genes are true targets in the respective cell line (Fig 4c, g, 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 vice versa. 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-rank genes, that show major difference.

For ZBTB33, Topic 49 showed a very high weight in GM12878. We found the top-rank loss genes in K562 for Topic 49 is enriched in the gene set related to cell cycle and cell division function (Fig 4d). The loss of ZBTB33 regulation results in higher expression of its target genes in K562 corresponding to known de-acylation and transcriptional suppressive roles of ZBTB33 (Pozner et al., 2016). (Fig 4e). Another highly rewired TF, EP300 (a known transcriptional activator) regulates a wide range of genes from different functional groups. In concordance with EP300’s function, Topics 5 and 10 are deficient in K562 and the top-rank loss genes from these topics are significantly down-regulated in K562. For topics of EP300 that are highly represented in K562 (34, 36), an adverse trend is observed (Fig 4h, Fig s6). To summarize, topics showing major difference in the rewiring of these TFs are related to the TFs’ molecular functions. Comparatively, transcription factors with low rewiring score like CTCF and ZNF274 have almost identical topic distributions (Fig s7).

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

Network rewiring shows dynamics topic changes of 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 TRNs for estrogen receptor (ESR1) in MCF-7 cell lines at 2min, 5min, 5min, 10min, 40min and 160min after estradiol treatment (Guertin et al., 2014). Rewiring score between these time points shows transition of the topic distributions across time points. The first few minutes right after estradiol treatment have much dramatic topic changes and then it gradually go stable (Fig 5a).

The time course pattern of the topic membership demonstrates that Topic 3 has the highest weight prior to the treatment. Later, Topic 4 becomes the most prominent topic after short fluctuation, which experiences a sharp increase at 10 min time point and then undergoes a gradual decrease but still keep the dominant position until the end (160m) (Fig 5b). We then studied the roles of ESR1 target genes that are related to these highly represented topics. At 0min, true ESR1 target genes that are top-rank in Topic 3 include genes are most related to cell proliferation functions: ribosomal functions, protein folding, and mRNA splicing (Fig s8). At the 10min time point for Topic 4, top-rank target genes include genes related to signal transduction and apoptosis, with some of them interacting directly with EP300, which is consistent with the findings that the redistribution of EP300 target after the treatment of estradiol (Guertin et al., 2014) (Fig s8).

The treatment of estradiol turned the MCF-7 cell line's topic from Topic 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-min and 40-min with 0-min for gained top-rank gene in topic 4. These gained top-rank genes show very significant (p-value < 10^{-5}) up-regulation in 10-min and 40min (Fig 5c, d).

Topic activity score and its relationship to tumor survival

Topic activity score incorporates cell type-independent topic component with cell type-specific gene expression and can be associated with clinical significance. We used patient samples of 3 cancer types with clinical information from the TCGA data portal breast cancer (BRCA), acute myeloid leukemia (LAML) and liver hepatocellular carcinoma (LIHC). Topic activity scores were evaluated for each cancer type and used for survival analysis. We found several topics whose expression score is associated with patient survival (Fig 6). For each cancer type, we characterized the biological relevance of the most predictive topic:

1. Expression score of Topic 10 is predictive for the survival of BRCA patients (Fig 6a). Correspondingly, Topic 10 is highly represented in the document of {GATA3, MCF-7} ,and its component is enriched with CtIP associated gene set (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) (Fig s10a).

2. Expression score of Topic 26 is predictive of LAML patient's survival. Topic 26, which is highly represented in the document {NR2C2, K562} (Fig 6b, Fig s9b), its components is also enriched with genes up-regulated in response to activation of the cAMP signaling pathway (van Staveren et al., 2006) (Fig s10b). NR2C2 can be induced by cAMP (Liu et al., 2009) and are found to be significantly active expressed in almost all the cancers (Falco et al., 2016).
3. Expression score of Topic 35 predict survival outcome of LIHC patients with high accuracy. Topic 35 is highly represented in the documents {ATF3, HepG2} and {JUN, HepG2} (Fig 6c, Fig s9c), and its component is enriched with gene set that are up-regulated in response to over-expression of proto-oncogene MYC (Bild et al., 2006) (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 have 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.

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 identification of network rewiring. In this study, we successfully developed a TopicNet framework. Our framework extracted a low-dimensional representation of network in the form of functional topics, defined network rewiring score and topic-weighted expression score, and then we demonstrated the application of the framework. The network rewiring score can aid in the identification of the functional rewiring of TFs between different cellular conditions. The topic-weighted expression score can be applied to sample-specific cohort data for the prediction of patient survival.

As a verification of our framework, we also investigated the biological meaning of the identified topics. We interpreted the learned topics by utilizing the two important matrices inferred from the model: document-topic weight and topic-gene component 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 pinpointing TFs, genes and pathways that play critical roles in these processes. Having demonstrated the potential application of our method in time course data, we expect our method to offer valuable insights into network dynamics studies in the future.

LDA has been shown to be advantageous over some other common dimensionality reduction techniques in terms of performance and interpretability (Stevens et al., 2012, Liu et al., 2011). Furthermore, several extensions on the LDA model could be introduced for future studies. In our framework, we treated all TF-cell line pairs as independent regardless of possible relationships between documents. Additionally, 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. We are aware that recent advances in topic modelling methods has enabled modelling 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.

Author contributions:

SL: conceptualization, methodology, formal analysis, visualization, writing-original draft; TL: methodology, formal analysis, visualization, writing-original draft; XK: methodology, formal analysis, visualization, JZ, JL and DL: resources, validation;MG: supervision

Declaration of Interests

The authors declare no competing interests.

Methods

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 higher quality control and consensus peak calling. In addition, we included ChIP-Atlas data collections with more than 6,000 ChIP-Seq experimental results to test the model. The number of target genes included in this dataset ranges from hundreds to thousands (Fig s11), and the TFs with the greatest availability among different cell lines include CTCF, EP300, MYC and REST (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 (+/- 2500bp) of their transcription start site. The cell type-specific TRN is then defined based on these results.

TopicNet - Topic modelling

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.

Let M, K, V be the number of documents, the number of topics and the vocabulary size, respectively. In this scenario, each document is modelled 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 components within the document and the probabilistic distribution of those topics. The existence of a word in a document is modelled as follows (Fig 1a):

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

$$\theta_i \sim Dirichlet(\alpha)$$

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

$$\phi_k \sim Dirichlet(\beta)$$

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

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

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

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

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

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 (i, j). The model parameters can be estimated using a collapsed Gibbs sampler (Heinrich, 2005) on a Markov chain of $\{W, Z\}$ where W is the observation and Z is the hidden variable. The joint distribution of the LDA model $\mathcal{P}(Z, W | \alpha, \beta)$ can be obtained by integrating out ϕ and θ:

$$\mathcal{P}(W|Z,\beta) = \int_{\varphi} \prod_{k=1}^{K} \mathcal{P}(\phi_k|\beta) \prod_{i=1}^{M}\prod_{j=1}^{N} \mathcal{P}\left(w_{i,j}\middle|\phi_{z_{i,j}}\right) d\phi = \prod_{k=1}^{K} \frac{\Delta(n_{\cdot,v}^{k}+\beta)}{\Delta(\beta)}$$

$$\mathcal{P}(Z|\alpha) = \int_{\theta} \prod_{i=1}^{M} \mathcal{P}(\theta_i|\alpha) \prod_{j=1}^{N} \mathcal{P}(z_{i,j}|\theta_i) d\theta = \prod_{i=1}^{M} \frac{\Delta(n_{i,\cdot}^{k}+\alpha)}{\Delta(\alpha)}$$

$$\mathcal{P}(Z,W|\,\alpha,\beta) = \mathcal{P}(W|Z;\,\alpha,\beta)\mathcal{P}(Z|\,\alpha,\beta) = \prod_{k=1}^{K} \frac{\Delta(n_{\cdot,v}^{k}+\beta)}{\Delta(\beta)} \prod_{i=1}^{M} \frac{\Delta(n_{i,\cdot}^{k}+\alpha)}{\Delta(\alpha)}$$

where $\Delta(x) = \frac{\prod_t \Gamma(x_t)}{\Gamma(\sum_t x_t)}$ for a t-dimensional vector x. $n_{\cdot,v}^{k}$ is the number of times gene v is assigned to topic k and $n_{i,\cdot}^{k}$ is the number of times genes in document i are assigned to topic k. These sums can all be calculated from the value of $\{W, Z\}$.

The conditional probability of $z_{i,j}$, the topic assignment of the j-th gene in the i-th document, can be inferred as:

$$\begin{aligned}
&\mathcal{P}\left(z_{i,j} = k|z_{\neg(i,j)}, W, \alpha, \beta\right) \\
&\quad = \frac{\mathcal{P}\left(z_{i,j},\ z_{\neg(i,j)},\ W|\,\alpha,\beta\right)}{\mathcal{P}\left(z_{\neg(i,j)}, W|\,\alpha,\beta\right)} = \frac{\mathcal{P}(Z,\ W|\,\alpha,\beta)}{\mathcal{P}\left(z_{\neg(i,j)},\ w_{\neg(i,j)}|\,\alpha,\beta\right)\mathcal{P}\left(w_{(i,j)}|\,\alpha,\beta\right)} \\
&\quad \propto \frac{\mathcal{P}(Z,W;\alpha,\beta)}{\mathcal{P}\left(z_{\neg(i,j)}, w_{\neg(i,j)};\alpha,\beta\right)} = \frac{\Delta\left(n_{\cdot,v}^{k}+\beta\right)}{\Delta\left(n_{\neg(i,j),v}^{k}+\beta\right)} \cdot \frac{\Delta\left(n_{i,\cdot}^{k}+\alpha\right)}{\Delta\left(n_{i,\neg(i,j)}^{k}+\alpha\right)}
\end{aligned}$$

where $n_{\neg(i,j),v}^{k}$ and $n_{i,\neg(i,j)}^{k}$ indicate the same type of count as mentioned above excluding (i,j). This gives the sampling distribution for Z given W.

Using the above formulae, Gibbs sampling can then be performed on the Markov chain $\{W, Z\}$ to obtain the estimation of φ and θ. By definition, we have the probability distribution of $\{W, Z\}$ conditioned on φ and θ. Using Bayes rule, the expectation of φ and θ can be inferred from $\{W, Z\}$:

$$\hat{\phi}_{k,v} = \frac{n_{\cdot,v}^{k}+\beta}{\sum_{v'=1}^{V} n_{\cdot,v'}^{k} + V\beta}$$

$$\hat{\theta}_{i,k} = \frac{n_{i,\cdot}^{k}+\alpha}{\sum_{k'=1}^{K} n_{i,\cdot}^{k'} + K\alpha}$$

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

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

Selection of topic numbers

The number of topics T used was selected using multiple criteria: the posterior likelihood of the data given the LDA model of different choice of number (Griffiths and Steyvers, 2004) (blue); KL divergence of the document-gene matrix (Arun et al., 2010b) (red); and the average cosine distance r within topics (Cao et al., 2009) (green). An optimal T should result in higher likelihood and lower distances. All three metrics reached optimal performance at around 50 topics (Fig 2a). Based on these results, we used 50 as the number of topics for downstream analysis.

Reconstructed correlations

We performed LDA, 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 $v_i = [v_{i,1}, v_{i,2}, \dots, v_{i,N}]$ 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 obtain 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 i-th column of matrix Θ:

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

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 i-th column of matrix W:

$$v_i^{NMF} = w_i$$

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:

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

For document pair in the dataset, we could calculate their 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 is also performed between the reconstructed and original correlation for the three methods respectively.

Evaluation of topic importance

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

$$\theta_{background} = [1/N, 1/N, \ldots, 1/N]$$

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

$$D_{KL}(\theta || \theta_{background}) = \sum_i (\theta_i) \, log(\frac{\theta_i}{\theta_{background_i}})$$

Gene set enrichment analysis

We use topic component as the statistic for gene set enrichment analysis. Gene sets in the C2, C5 and C6 categories from MSigDB (Subramanian et al., 2005) were used in the analysis. Gene set enrichment analysis is performed with R package *fgsea*. Affinity between gene sets is defined by their overlapping of gene sets.

Topic rewiring score - Quantification of network rewiring

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

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

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:

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

Topic activity score

Gene expression data for BRCA, LAML, LIHC and GBM patient samples from TCGA are obtained from GDC data portal (https://portal.gdc.cancer.gov/). The gene expression levels of every sample in each cancer type is formulated into an expression matrix E where rows represent genes and columns represent samples. For each cancer type, the expression data is first quantile normalized. Only genes that appear in the topic-gene matrix are retained. We then obtain the expression matrix E by first ranking the expression of the genes for each sample and then transform the value for gene i in the matrix of column j (corresponding to patient sample j) to $E_{i,j} = 1/rank_{i,j}$ where $rank_{i,j}$ 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 i.e. $S_t^{Act} = \Phi E_t$, where each element in the vector is the expression score of the corresponding target.

Definition of gain or loss genes

Given two TF regulatory networks under two conditions, we arbitrarily assign 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 are particularly interested in the gain or loss genes that are among the top-rank genes of the topic, i.e. those have high values in the topic component. We take the intersection between these two sets and name the resulting set of genes as “top-rank gain/loss genes in the altered condition for topic k”.

PPI network analysis of topic-related target genes

The genes in the corpus are first sorted according to their contributions to the topic. Among the top 500 genes, those that are directly regulated by the given TF (i.e. bound by the TF in the corresponding ChIP-Seq experiment) are selected and provided to STRING (Franceschini et al., 2013). The resulting interaction graph contains the selected genes along with their first-layer neighbours.

Survival analysis

Clinical information for each patient regarding vital status, days to last follow-up, and days to death are downloaded from GDC data portal (https://portal.gdc.cancer.gov/). Records with missing information are discarded. Patients that are still alive in the record are right censored. The values of all 50 topics are 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 select the topics, S^{Exp} score of those have p-value < 0.05 in the coxph analysis, for further analysis. For each of these candidate topics, the patients are then separated into two groups by the median value of S^{Exp}. The survival curve is then estimated using the Kaplan-Meier estimator. The topics that achieves lowest p-value is selected and shown.

References:

ARUN, R., SURESH, V., MADHAVAN, C. E. V. & MURTHY, M. N. N. 2010a. On finding the natural number of topics with latent dirichlet allocation: some observations. *Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part I.* Hyderabad, India: Springer-Verlag.

ARUN, R., SURESH, V., VENI MADHAVAN, C. E. & NARASIMHA MURTHY, M. N. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. 2010b Berlin, Heidelberg. Springer Berlin Heidelberg, 391-402.

ASSI, S. A., IMPERATO, M. R., COLEMAN, D. J. L., PICKIN, A., POTLURI, S., PTASINSKA, A., CHIN, P. S., BLAIR, H., CAUCHY, P., JAMES, S. R., ZACARIAS-CABEZA, J., GILDING, L. N., BEGGS, A., CLOKIE, S., LOKE, J. C., JENKIN, P., UDDIN, A., DELWEL, R., RICHARDS, S. J., RAGHAVAN, M., GRIFFITHS, M. J., HEIDENREICH, O., COCKERILL, P. N. & BONIFER, C. 2019. Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat Genet,* 51**,** 151-162.

BHARDWAJ, N., KIM, P. M. & GERSTEIN, M. B. 2010. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci Signal,* 3**,** ra79.

BILD, A. H., YAO, G., CHANG, J. T., WANG, Q., POTTI, A., CHASSE, D., JOSHI, M. B., HARPOLE, D., LANCASTER, J. M., BERCHUCK, A., OLSON, J. A., JR., MARKS, J. R., DRESSMAN, H. K., WEST, M. & NEVINS, J. R. 2006. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature,* 439**,** 353-7.

BLEI, D. M. & LAFFERTY, J. D. 2006. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning.* Pittsburgh, Pennsylvania, USA: ACM.

BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3**,** 993-1022.

CAFFAREL, M. M., MORENO-BUENO, G., CERUTTI, C., PALACIOS, J., GUZMAN, M., MECHTA-GRIGORIOU, F. & SANCHEZ, C. 2008. JunD is involved in the antiproliferative effect of Delta9-tetrahydrocannabinol on human breast cancer cells. *Oncogene,* 27**,** 5033-44.

CAO, J., XIA, T., LI, J. T., ZHANG, Y. D. & TANG, S. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing,* 72**,** 1775-1781.

CHEN, C., GE, C., LIU, Z., LI, L., ZHAO, F., TIAN, H., CHEN, T., LI, H., YAO, M. & LI, J. 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., SHI, P., ZHOU, Y., YU, Y., GUO, X., YAO, Y., LIU, P. & XU, B. 2017. High BCL11A Expression in Adult Acute Myeloid Leukemia Patients Predicts a Worse Clinical Outcome. *Clin Lab,* 63**,** 85-90.

FALCO, M. M., BLEDA, M., CARBONELL-CABALLERO, J. & DOPAZO, J. 2016. The pan-cancer pathological regulatory landscape. *Sci Rep,* 6**,** 39709.

FLEMING, J. D., PAVESI, G., BENATTI, P., IMBRIANO, C., MANTOVANI, R. & STRUHL, K. 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**,** 1195-209.

FRANCESCHINI, A., SZKLARCZYK, D., FRANKILD, S., KUHN, M., SIMONOVIC, M., ROTH, A., LIN, J., MINGUEZ, P., BORK, P., VON MERING, C. & JENSEN, L. J. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res,* 41**,** D808-15.

FURUTA, S., WANG, J. M., WEI, S., JENG, Y. M., JIANG, X., GU, B., CHEN, P. L., LEE, E. Y. & LEE, W. H. 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., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K. K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R., MIN, R., ALVES, P., ABYZOV, A., ADDLEMAN, N., BHARDWAJ, N., BOYLE, A. P., CAYTING, P., CHAROS, A., CHEN, D. Z., CHENG, Y., CLARKE, D., EASTMAN, C., EUSKIRCHEN, G., FRIETZE, S., FU, Y., GERTZ, J., GRUBERT, F., HARMANCI, A., JAIN, P., KASOWSKI, M., LACROUTE, P., LENG, J. J., LIAN, J., MONAHAN, H., O'GEEN, H., OUYANG, Z., PARTRIDGE, E. C., PATACSIL, D., PAULI, F., RAHA, D., RAMIREZ, L., REDDY, T. E., REED, B., SHI, M., SLIFER, T., WANG, J., WU, L., YANG, X., YIP, K. Y., ZILBERMAN-SCHAPIRA, G., BATZOGLOU, S., SIDOW, A., FARNHAM, P. J., MYERS, R. M., WEISSMAN, S. M. & SNYDER, M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature,* 489**,** 91-100.

GRIFFITHS, T. L. & STEYVERS, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 5228-5235.

GUERTIN, M. J., ZHANG, X., COONROD, S. A. & HAGER, G. L. 2014. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Mol Endocrinol,* 28**,** 1522-33.

GUO, Y. & GIFFORD, D. K. 2017. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics,* 18**,** 45.

HEINRICH, G. 2005. Parameter estimation for text analysis. Technical report.

HONG, S. & KIM, D. 2017. Computational characterization of chromatin domain boundary-associated genomic elements. *Nucleic Acids Res,* 45**,** 10403-10414.

KHALED, W. T., CHOON LEE, S., STINGL, J., CHEN, X., RAZA ALI, H., RUEDA, O. M., HADI, F., WANG, J., YU, Y., CHIN, S. F., STRATTON, M., FUTREAL, A., JENKINS, N. A., APARICIO, S., COPELAND, N. G., WATSON, C. J., CALDAS, C. & LIU, P. 2015. BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat Commun,* 6**,** 5987.

KIM, M. S., LEE, S. H., YOO, N. J. & LEE, S. H. 2013. Frameshift mutations of tumor suppressor gene EP300 in gastric and colorectal cancers with high microsatellite instability. *Hum Pathol,* 44**,** 2064-70.

LIU, N. C., LIN, W. J., YU, I. C., LIN, H. Y., LIU, S., LEE, Y. F. & CHANG, C. 2009. Activation of TR4 orphan nuclear receptor gene promoter by cAMP/PKA and C/EBP signaling. *Endocrine,* 36**,** 211-7.

LIU, Z., LI, M., LIU, Y. & PONRAJ, M. Performance evaluation of Latent Dirichlet Allocation in text mining. 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 26-28 July 2011 2011. 2695-2698.

LIU, Z. P., WU, C., MIAO, H. & WU, H. 2015. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database (Oxford),* 2015.

MAEDA, S. & KARIN, M. 2003. Oncogene at last--c-Jun promotes liver cancer in mice. *Cancer Cell,* 3**,** 102-4.

MILLENA, A. C., VO, B. T. & KHAN, S. A. 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**,** 17964-76.

MOMENI, E., KARUNASEKERA, S., GOYAL, P. & LERMAN, K. 2018. *Modeling Evolution of Topics in Large-Scale Temporal Text Corpora*.

O'GEEN, H., LIN, Y. H., XU, X., ECHIPARE, L., KOMASHKO, V. M., HE, D., FRIETZE, S., TANABE, O., SHI, L., SARTOR, M. A., ENGEL, J. D. & FARNHAM, P. J. 2010. Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics,* 11**,** 689.

PARELHO, V., HADJUR, S., SPIVAKOV, M., LELEU, M., SAUER, S., GREGSON, H. C., JARMUZ, A., CANZONETTA, C., WEBSTER, Z., NESTEROVA, T., COBB, B. S., YOKOMORI, K., DILLON, N., ARAGON, L., FISHER, A. G. & MERKENSCHLAGER, M. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell,* 132**,** 422-33.

PINOLI, P., CHICCO, D. & MASSEROLI, M. 2014. Latent Dirichlet Allocation based on Gibbs Sampling for Gene Function Prediction. *2014 Ieee Conference on Computational Intelligence in Bioinformatics and Computational Biology*.

POZNER, A., TEROOATEA, T. W. & BUCK-KOEHNTOP, B. A. 2016. Cell-specific Kaiso (ZBTB33) Regulation of Cell Cycle through Cyclin D1 and Cyclin E1. *J Biol Chem,* 291**,** 24538-24550.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

RUBIO, E. D., REISS, D. J., WELCSH, P. L., DISTECHE, C. M., FILIPPOVA, G. N., BALIGA, N. S., AEBERSOLD, R., RANISH, J. A. & KRUMM, A. 2008. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A,* 105**,** 8309-14.

STEVENS, K., KEGELMEYER, P., ANDRZEJEWSKI, D. & BUTTLER, D. 2012. Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Jeju Island, Korea: Association for Computational Linguistics.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

THOMPSON, D., REGEV, A. & ROY, S. 2015. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. *Annual Review of Cell and Developmental Biology, Vol 31,* 31**,** 399-428.

VAN STAVEREN, W. C., SOLIS, D. W., DELYS, L., VENET, D., CAPPELLO, M., ANDRY, G., DUMONT, J. E., LIBERT, F., DETOURS, V. & MAENHAUT, C. 2006. Gene expression in human thyrocytes and autonomous adenomas reveals suppression of negative feedbacks in tumorigenesis. *Proc Natl Acad Sci U S A,* 103**,** 413-8.

WANG, H. J., DING, Y., TANG, J., DONG, X. A., HE, B., QIU, J. & WILD, D. J. 2011. Finding Complex Biological Relationships in Recent PubMed Articles Using Bio-LDA. *Plos One,* 6.

ZHANG, F., TANG, H., JIANG, Y. & MAO, Z. 2017. The transcription factor GATA3 is required for homologous recombination repair by regulating CtIP expression. *Oncogene,* 36**,** 5168-5176.

ZHANG, S., TIAN, D., TRAN, N. H., CHOI, K. P. & ZHANG, L. 2014. Profiling the transcription factor regulatory networks of human cell types. *Nucleic Acids Res,* 42**,** 12380-7.

ZHENG, K., CUBERO, F. J. & NEVZOROVA, Y. A. 2017. c-MYC-Making Liver Sick: Role of c-MYC in Hepatic Cell Function, Homeostasis and Disease. *Genes (Basel),* 8.

ZHOU, H., YU, H. & HU, R. 2017. Topic evolution based on the probabilistic topic model: a review. *Front. Comput. Sci.,* 11**,** 786-802.

Figures:

Fig 1 Overview of method and data

(a) Diagram explaining the meanings and biological relevance of the parameters in the LDA model. (b) General workflow of our analytical framework.

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 distributions (50 topic model) for data samples on CTCF, EZH2, POL2A and POL2AphosphoS5. (c) Correlation between topic components identified by LDA models with different topic numbers (5, 10, 20, 50).

Fig 3 Annotation of the identified topics

(a) All 50 topics ranked by their importance measure (average KL divergence of topic distribution against uniform distribution across all samples). (b) Genes with highest contributions to the top ranked topics (Topic 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 (grey 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-rank genes of Topic 49 that are related to cell cycle and cell division. (e) Expression of the top-rank genes of Topic 49 which are lost in K562. (f) Topic weight for the rewiring event of EP300 in GM12878 and K562. (g) Top-rank genes of the two topics with 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-rank genes of Topic 34 which are lost in K562.

Fig 5 ESR1 regulation dynamics represented by topic distributions.

(a) Pairwise KL divergence between the topic distributions of ESR1 in all time points. (b) Time course change of the topics with highest weights across the time points. (c)~(d) Expression of the genes gained in 10min (c) and 40min (d), as compared to 0min, which are among the top-rank genes of Topic 4.

Fig 6 Topic activity level related to cancer survival

(a)~(c) Kaplan-Meier 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).

Supplementary Figure

Fig s1 Selection of topic number using three different metrices.

Fig s2 Hierarchical clustering of TF regulatory roles in HeLa-S3 using pairwise KL divergence as distance measure.

Fig s3 Correlation of the highest represented gene sets in the component of Topic 3. Numbers indicate shared genes among the gene sets.

Fig s4 All TFs ranked by their average pairwise rewiring values across cell types.

Fig s5 Table of rewiring events with highest rewiring values.

Fig s6 Topic component and expression of the top weighted genes that show highest rewiring in EP300 regulation between GM12878 and K562 after Topic 34: Topic 10 (a), 36 (b) and 5 (c).

Fig s7 Topic weight of regulatory targets for the two lowest rewiring events: CTCF (left) and NR2C2 (right) in GM12878 and K562.

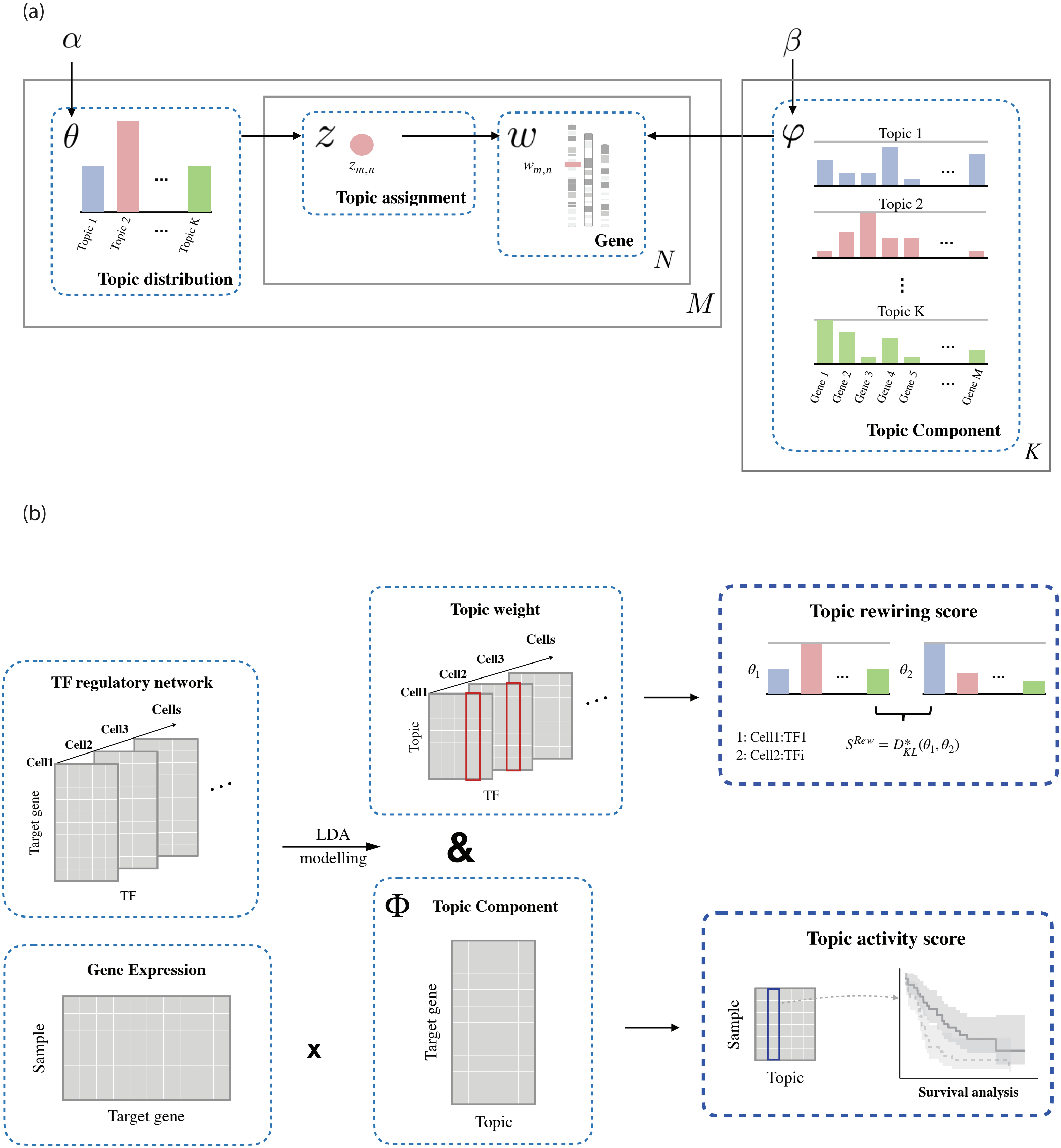
Fig s8 Roles in protein-protein interaction networks of the top-rank target genes of Topic 3 at 0min (a) and Topic 4 at 10min (b). Deep blue nodes are the selected target genes and light blue nodes are their first-layer neighbors in the network.

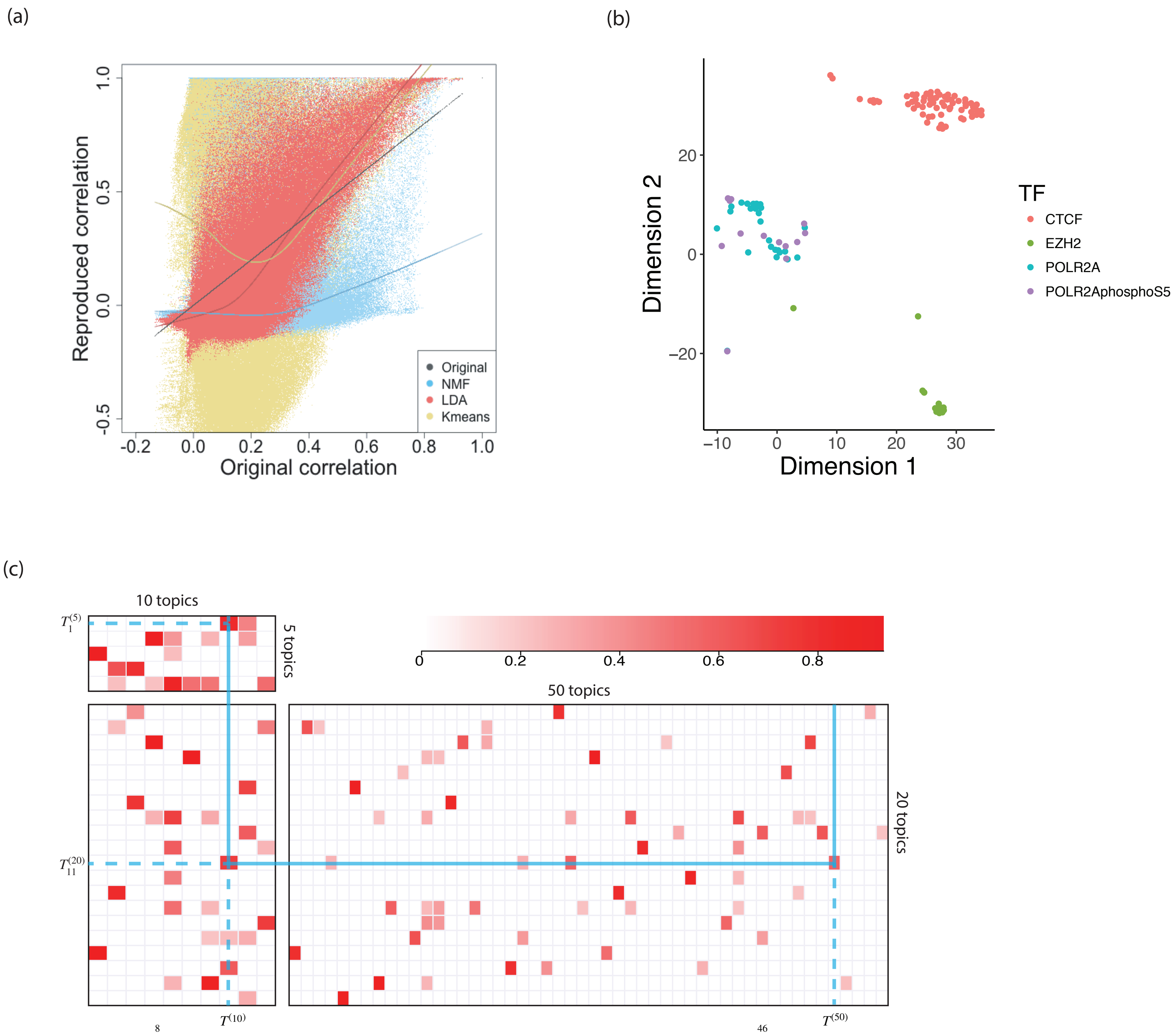
Fig s9 Gene set enrichment of oncogenic signatures (C6) gene sets in topics related to cancer patient survival (topic 10, 26 and 35).

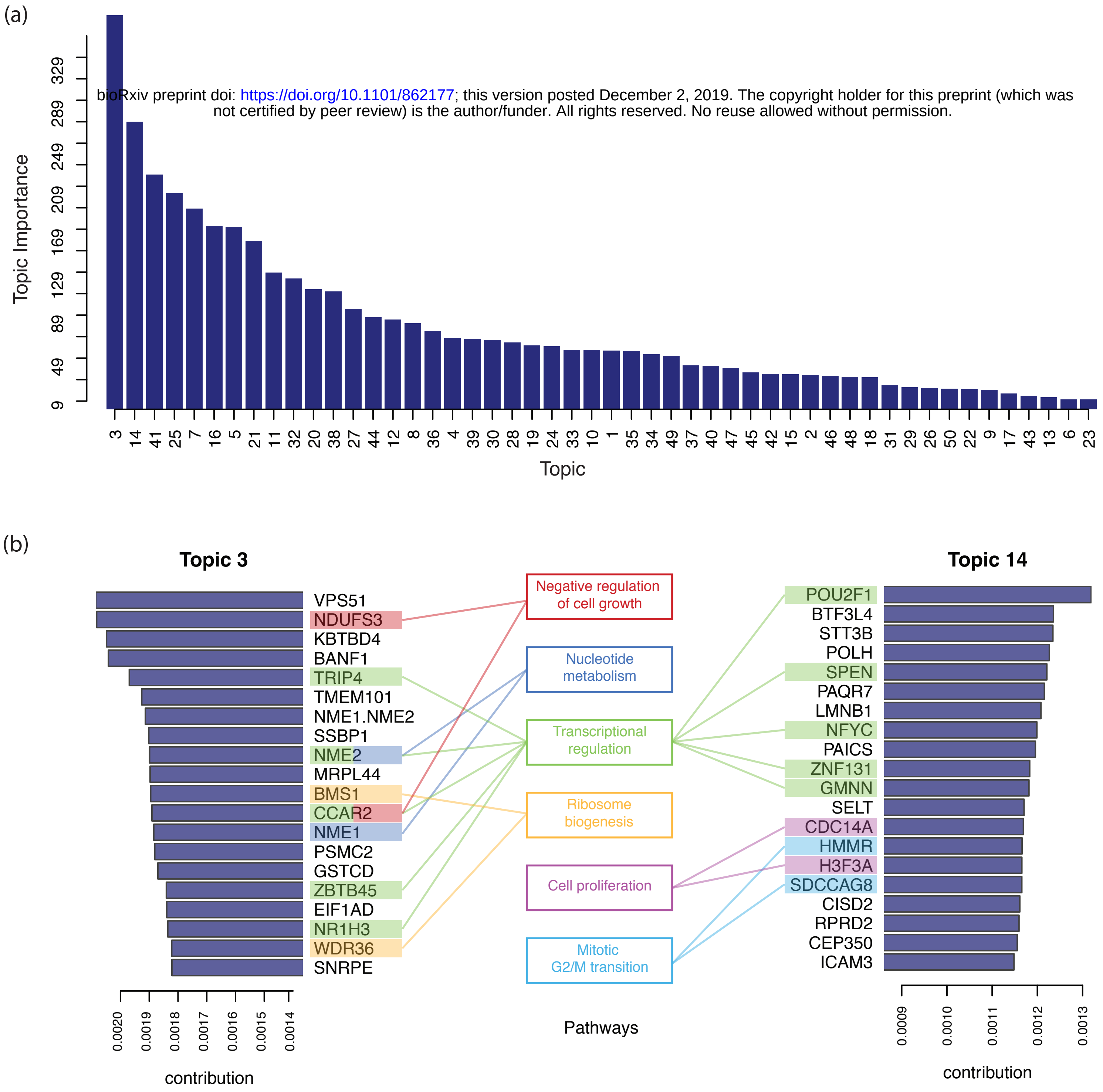
Fig s10 TFs in tumor cell lines that are highly enriched with survival-related topics: Topic 10 in MCF-7 (a), Topic 26 in K562 (b) and Topic 35 in HepG2 (c).

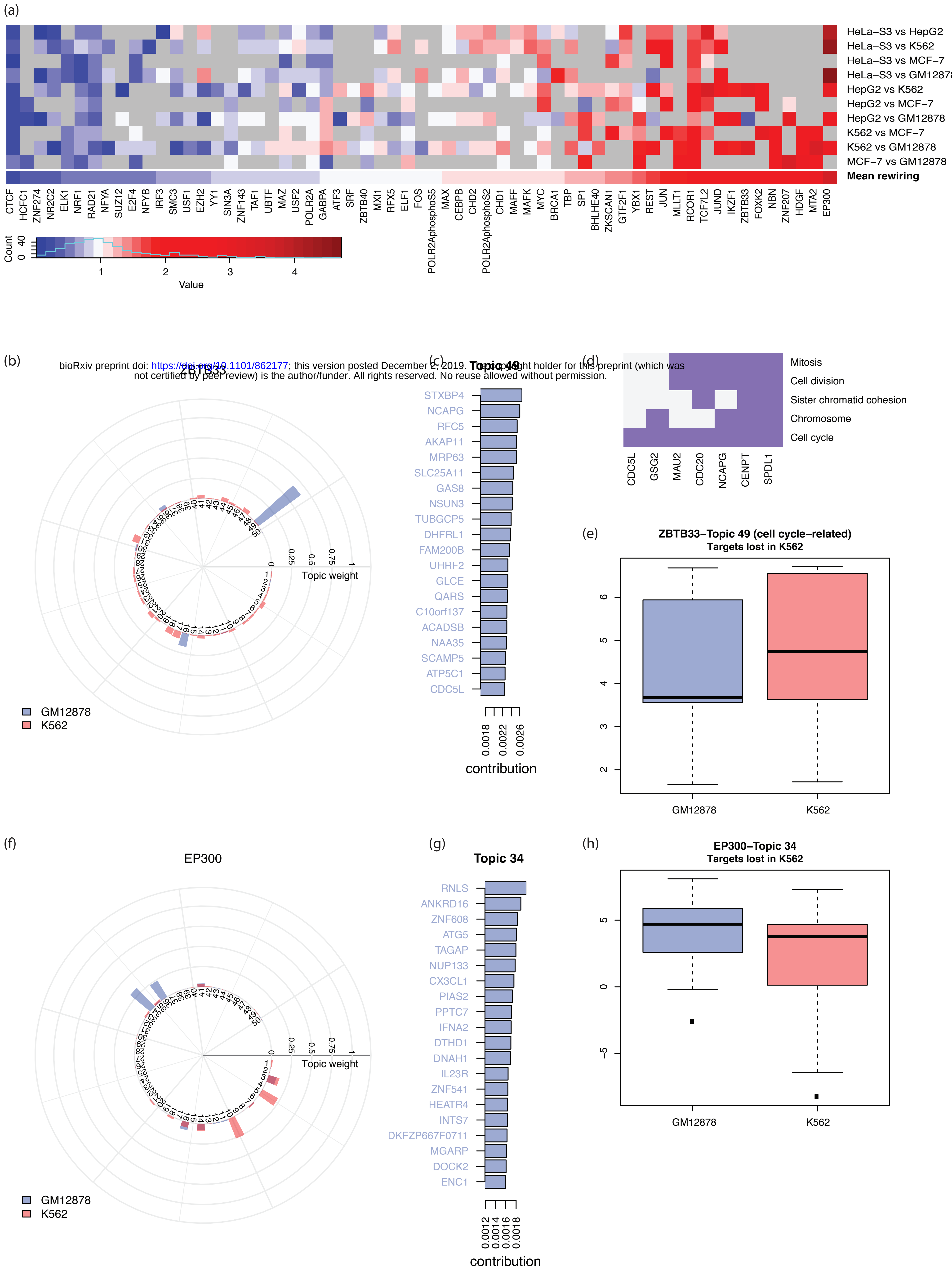
Fig s11 Statistics of identified target genes from ChIP-Seq datasets used in the study.

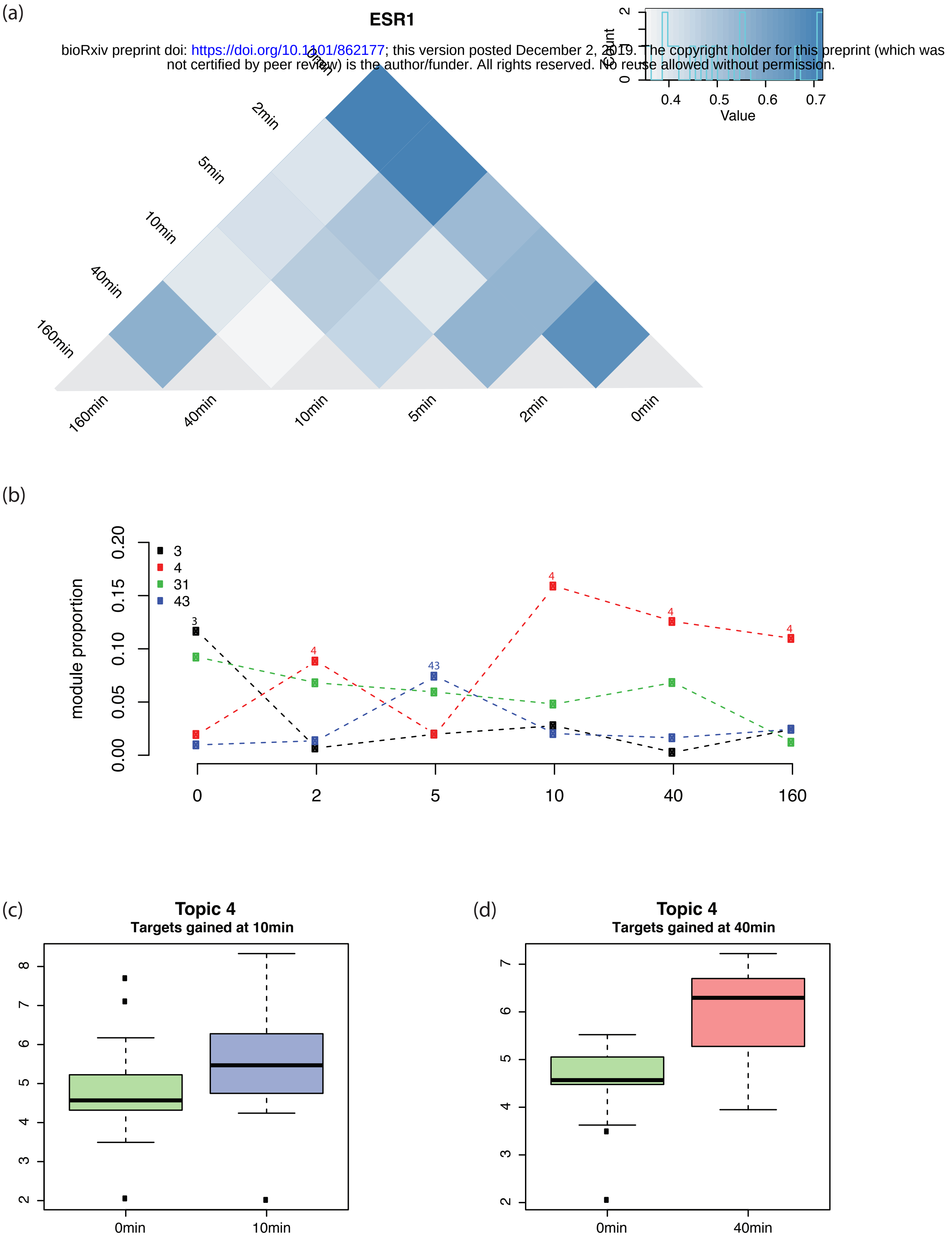
Fig s12 Log frequency of transcription factors with highest occurrence in ChIP-ATLAS and ENCODE datasets.

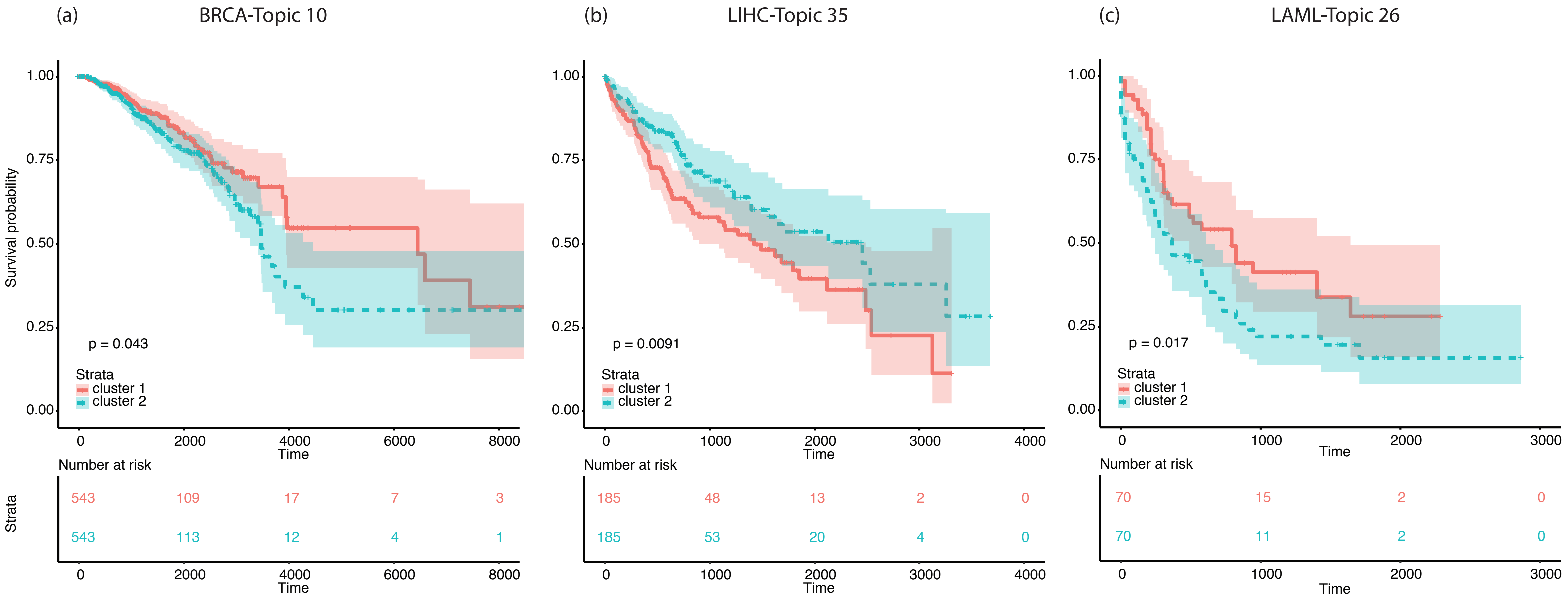

(a)
α
β
θ
Topic 1
Topic 2
...
Topic K
Topic distribution
z
z_{m,n}
Topic assignment
w
w_{m,n}
Gene
N
M
φ
Topic 1
Topic 2
Topic K
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene M
Topic Component
K
(b)
TF regulatory network
Cells
Cell3
Cell2
Cell1
Target gene
TF
LDA
modelling
Topic weight
Cells
Cell3
Cell2
Cell1
Topic
TF
&
Topic rewiring score
θ1
θ2
1: Cell1:TF1
2: Cell2:TFi
S^{Rew} = D*_{KL}(θ1, θ2)
Gene Expression
Sample
Target gene
x
Φ
Topic Component
Target gene
Topic
Topic activity score
Sample
Topic
Survival analysis


(a)
Reproduced correlation
Original correlation
-0.5
0.0
0.5
1.0
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
● Original
● NMF
● LDA
● Kmeans
(b)
Dimension 2
Dimension 1
20
0
−20
−10
0
10
20
30
TF
● CTCF
● EZH2
● POLR2A
● POLR2AphosphoS5
(c)
10 topics
5 topics
50 topics
20 topics
0
0.2
0.4
0.6
0.8
$T_1^{(5)}$
$T_{11}^{(20)}$
8
$T^{(10)}$
46
$T^{(50)}$


(a)
Topic Importance
9
49
89
129
169
209
249
289
329
3
14
41
25
7
16
5
21
11
32
20
38
27
44
12
8
36
4
39
30
28
19
24
33
10
1
35
34
49
37
40
47
45
42
15
2
46
48
18
31
29
26
50
22
9
17
43
13
6
23
Topic
(b)
Topic 3
VPS51
NDUFS3
KBTBD4
BANF1
TRIP4
TMEM101
NME1.NME2
SSBP1
NME2
MRPL44
BMS1
CCAR2
NME1
PSMC2
GSTCD
ZBTB45
EIF1AD
NR1H3
WDR36
SNRPE
0.0020
0.0019
0.0018
0.0017
0.0016
0.0015
0.0014
contribution
Negative regulation of cell growth
Nucleotide metabolism
Transcriptional regulation
Ribosome biogenesis
Cell proliferation
Mitotic G2/M transition
Pathways
Topic 14
POU2F1
BTF3L4
STT3B
POLH
SPEN
PAQR7
LMNB1
NFYC
PAICS
ZNF131
GMNN
SELT
CDC14A
HMMR
H3F3A
SDCCAG8
CISD2
RPRD2
CEP350
ICAM3
0.0009
0.0010
0.0011
0.0012
0.0013
contribution


(a)
HeLa−S3 vs HepG2
HeLa−S3 vs K562
HeLa−S3 vs MCF−7
HeLa−S3 vs GM12878
HepG2 vs K562
HepG2 vs MCF−7
HepG2 vs GM12878
K562 vs MCF−7
K562 vs GM12878
MCF−7 vs GM12878
Mean rewiring
CTCF
HCFC1
ZNF274
NR2C2
ELK1
NRF1
RAD21
NFYA
SUZ12
E2F4
NFYB
IRF3
SMC3
USF1
EZH2
YY1
SIN3A
ZNF143
TAF1
UBTF
MAZ
USF2
POLR2A
GABPA
ATF3
SRF
ZBTB40
MXI1
RFX5
ELF1
FOS
POLR2AphosphoS5
MAX
CEBPB
CHD2
POLR2AphosphoS2
CHD1
MAFF
MAFK
MYC
BRCA1
TBP
SP1
BHLHE40
ZKSCAN1
GTF2F1
YBX1
REST
JUN
MLLT1
RCOR1
TCF7L2
JUND
IKZF1
ZBTB33
FOXK2
NBN
ZNF207
HDGF
MTA2
EP300
Count
0
40
1
2
3
4
Value
(b)
ZBTB33
Topic weight
0
0.25
0.5
0.75
1
GM12878
K562
(c)
Topic 49
STXBP4
NCAPG
RFC5
AKAP11
MRP63
SLC25A11
GAS8
NSUN3
TUBGCP5
DHFRL1
FAM200B
UHRF2
GLCE
QARS
C10orf137
ACADSB
NAA35
SCAMP5
ATP5C1
CDC5L
0.0018
0.0022
0.0026
contribution
(d)
Mitosis
Cell division
Sister chromatid cohesion
Chromosome
Cell cycle
CDC5L
GSG2
MAU2
CDC20
NCAPG
CENPT
SPDL1
(e)
ZBTB33−Topic 49 (cell cycle−related)
Targets lost in K562
2
3
4
5
6
GM12878
K562
(f)
EP300
Topic weight
0
0.25
0.5
0.75
1
GM12878
K562
(g)
Topic 34
RNLS
ANKRD16
ZNF608
ATG5
TAGAP
NUP133
CX3CL1
PIAS2
PPTC7
IFNA2
DTHD1
DNAH1
IL23R
ZNF541
HEATR4
INTS7
DKFZP667F0711
MGARP
DOCK2
ENC1
0.0012
0.0014
0.0016
0.0018
contribution
(h)
EP300−Topic 34
Targets lost in K562
−5
0
5
GM12878
K562


(a)
ESR1
Count
0
1
2
0.4
0.5
0.6
0.7
Value
0min
2min
5min
10min
40min
160min
160min
40min
10min
5min
2min
0min
(b)
module proportion
0.00
0.05
0.10
0.15
0.20
3
4
31
43
0
2
5
10
40
160
(c)
Topic 4
Targets gained at 10min
2
3
4
5
6
7
8
0min
10min
(d)
Topic 4
Targets gained at 40min
2
3
4
5
6
7
0min
40min


(a)
BRCA-Topic 10
(b)
LIHC-Topic 35
(c)
LAML-Topic 26
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.043
p = 0.0091
p = 0.017
Strata
cluster 1
cluster 2
Time
0
2000
4000
6000
8000
1000
3000
Number at risk
Strata
543 109 17 7 3
543 113 12 4 1
185 48 13 2 0
185 53 20 4 0
70 15 2 0
70 11 2 0