BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

DiNeR: a *Di*fferential graphical model for analysis of co-regulation *Ne*twork *R*ewiring

Jing Zhang[1†], Jason Liu[2,3†], Donghoon Lee[2,3], Shaoke Lou[2,3], Zhanlin Chen[4,5], Gamze Gürsoy[2,3] and Mark Gerstein[2,3,5*]

* Correspondence: pi@gersteinlab.org
†Jing Zhang and Jason Liu contributed equally to this work.
[2]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT 06520, USA
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

Abstract

Background: During transcription, numerous transcription factors (TFs) bind to targets in a highly coordinated manner to control the gene expression. Alterations in groups of TF-binding profiles (i.e. "co-binding changes") can affect the co-regulating associations between TFs (i.e. "rewiring the co-regulator network"). This, in turn, can potentially drive downstream expression changes, phenotypic variation, and even disease. However, quantification of co-regulatory network rewiring has not been comprehensively studied.

Results: To address this, we propose DiNeR, a computational method to directly construct a differential TF co-regulation network from paired disease-to-normal ChIP-seq data. Specifically, DiNeR uses a graphical model to capture the gained and lost edges in the co-regulation network. Then, it adopts a stability-based, sparsity-tuning criterion -- by sub-sampling the complete binding profiles to remove spurious edges -- to report only significant co-regulation alterations. Finally, DiNeR highlights hubs in the resultant differential network as key TFs associated with disease. We assembled genome-wide binding profiles of 104 TFs in the K562 and GM12878 cell lines, which loosely model the transition between normal and cancerous states in chronic myeloid leukemia (CML). In total, we identified 351 significantly altered TF co-regulation pairs. In particular, we found that the co-binding of the tumor suppressor BRCA1 and RNA polymerase II, a well-known transcriptional pair in healthy cells, was disrupted in tumors. Thus, DiNeR successfully extracted hub regulators and discovered well-known risk genes.

Conclusions: Our method DiNeR makes it possible to quantify changes in co-regulatory networks and identify alterations to TF co-binding patterns, highlighting key disease regulators. Our method DiNeR makes it possible to quantify changes in co-regulatory networks and identify alterations to TF co-binding patterns, highlighting key disease regulators.

Keywords: Transcription factor co-regulation network, ENCODE, TF dysregulation, Network changes

Background

Thousands of transcription factors (TFs), their cofactors, and chromatin remodelers are employed in a highly coordinated manner to accurately initiate and control the transcriptional process on DNA sequences [1–4]. Precise temporal and spatial coordination among these factors is important for determining cell phenotype and maintaining biological function. Studies have reported that disruption of the co-regulation relationships of TFs can result in gene expression alterations, which can consequently introduce phenotypical variations and even lead to disease [5–9]. However, various computational methods have been proposed to infer dysregulations of an individual TF by exploring differential expressions of the TF itself and its gene targets [10], or investigating the direct TF-gene gain and loss events [11], ignoring the higher-order combinatory binding patterns among TFs. Therefore, large-scale mining of TF co-binding changes in disease and normal states could provide new insights into gene dysregulation and opportunities for targeted therapies.

In this study, we aimed to quantify such alterations to the TF co-regulatory relationships and prioritize regulators associated with pathogenesis. This is a challenging task for many reasons. For instance, many of the ~ 1400 known human TFs have highly dynamic binding profiles depending on the cell state and conditions [12]. Therefore, it is essential to carefully curate a dataset of appropriate cell types to precisely capture disease-specific TF co-regulatory disruptions. Furthermore, joint analysis of binding profiles of numerous factors would be beneficial by maximizing our knowledge of TF co-binding events. Researchers have proposed various models to systematically impute all tissue-specific TF regulomes using features such as DNase accessibility and sequence context, but the accuracy of these methods across TFs is still unknown [13, 14].

In light of this, we propose a multi-step computational framework that we call *Di*fferential graphical model of *Ne*twork *R*ewiring (DiNeR) to infer TF co-binding alterations and pinpoint disease-causing TFs. First, we modeled the cooperative regulation patterns among TFs using a co-regulation network, where the nodes represent TFs and the weighted edges measure the genome-wide co-occurrence between pairwise TFs, all of which are derived from chromatin immunoprecipitation followed by sequencing (ChIP-seq) data. This is distinct from traditional regulatory networks, where edges usually imply the relationship between a TF and the physical interaction it shares with an enhancer or promoter region in order to initiate the transcription process of its target genes. Second, we directly measured the gain or loss of TF co-regulatory events across the genome during the transition from one cellular state to another using a differential graphical model. Intuitively, a higher weight in our differential network indicates a larger alteration in the co-regulatory pattern between two states. As a third step, we adopted the graphical LASSO model and used a stability-based method for penalizing parameter selection to control for the sparsity of the differential network with the goal of removing spurious edges [15–18]. Lastly, we prioritized the TFs according to their degree in the differential network based on the assumption that disease-driving TFs can demonstrate massive binding profile changes and consequently disrupt their cooperative regulation with many other regulators.

To test the effectiveness of our DiNeR framework, we applied this model to the ENCODE Tier 1 cell lines, K562 and GM12878, roughly representing paired disease-to-normal cells for chronic myelogenous leukemia (CML). We used our DiNeR framework to directly estimate the changes in the TF combinatory regulation relationships between disease and normal states. As a result, we identified several TFs that exhibit a high level of disruption to their co-binding partners. Among this list were several well-known risk factors for leukemia, such as BRCA1, RAD51, BMI1, and H3k27me3 [19–22], demonstrating the effectiveness of our method.

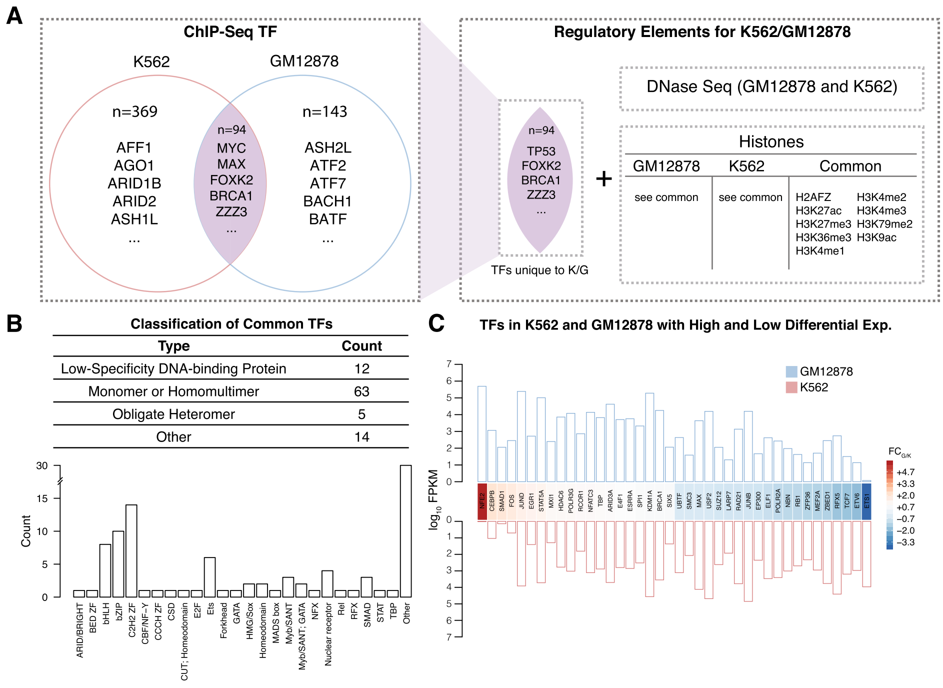
Results

Collecting appropriate large-scale ChIP-seq data to investigate transcriptional regulation

TF binding sites are highly tissue specific and usually change dramatically across cell types. Therefore, it is important to select appropriately matched and representative cell types to investigate a specific disease. Here, we used K562 and GM12878 to represent a rough disease-to-normal pair in order to investigate the transcriptional regulation dynamics during oncogenesis for CML (see details in "ChIP-seq data collection and pre-processing" section). Specifically, we extracted 94 common TFs among these two cell lines (Fig. 1a, Table S1). In order to investigate alterations in the joint activity between TFs and specific chromatin marks in disease, we also extracted nine histone modification marks and chromatin accessibility data sets from these cell lines. Among these TFs, 31 showed significant expression changes between disease and normal states (Fig. 1c).

Building a genome-wide TF co-regulatory network

Aberrant transcriptional regulation is associated with various diseases [5–9]. Network analysis has been proven to be a powerful tool for identifying and prioritizing genes or



Classification of Common TFs

Type	Count
Low-Specificity DNA-binding Protein	12
Monomer or Homomultimer	63
Obligate Heteromer	5
Other	14

Fig. 1 ChIP-seq data collection and pre-processing. **a** TFs shared by both the K562 (red) and GM12878 (blue) cell lines from ENCODE, in addition to DNase sequencing and histone modification data, are pooled together to form the set of regulatory elements used in the analysis. **b** Classification of the common TFs between K562 and GM12878. Specific DBD family classifications are also given. **c** The heatmap shows the top and bottom 20 TFs with high and low log fold change of expression between GM12878 and K562. The bar plot shows the $\log_{10}$ of the FPKM expression values for each of these TFs

regulators associated with pathogenesis. For example, scientists have constructed TF regulatory networks in order to mimic the physical binding of TFs to either enhancer or promoter regions during initiation of the transcriptional process of its target gene (Fig. 2a, b) [11]. Edges in this type of regulatory network only focus on the local interaction between the TF and target gene pair, and do not consider the effect of the rest of the genome. Another approach is to use gene co-expression networks, where a shared edge in the network represents consistent expression patterns between a pair of genes across many samples, which was usually inferred from RNA sequencing or microarray data [23, 24]. However, this output only reflects co-expression patterns, rather than regulatory relationships.

Here, we propose a TF co-regulatory network based on large-scale ChIP-seq data to model a related but different aspect of transcriptional regulation – the cooperative behavior among TFs. Specifically, in our network, nodes represent TFs and the weighted edge between them measures the level of non-random co-binding activity across the genome. It is important to note that this network is distinct from a traditional TF-TF network, which focuses on the mechanism of how one TF gene is regulated by another TF protein. In particular, this type of TF-TF regulatory network describes the binding event of one TF at the non-coding element (e.g., enhancer or promoter) of another TF, but does not indicate whether these two TFs work in a coordinated way to regulate other downstream genes.

We used the Gaussian graphical model (GGM) to construct this TF co-regulatory network [25]. The schematic in Fig. 2c illustrates one example. Two pairs TF_1 and TF_3, TF_2 and TF_3 co-bind over many places in the genome, possibly resulting from

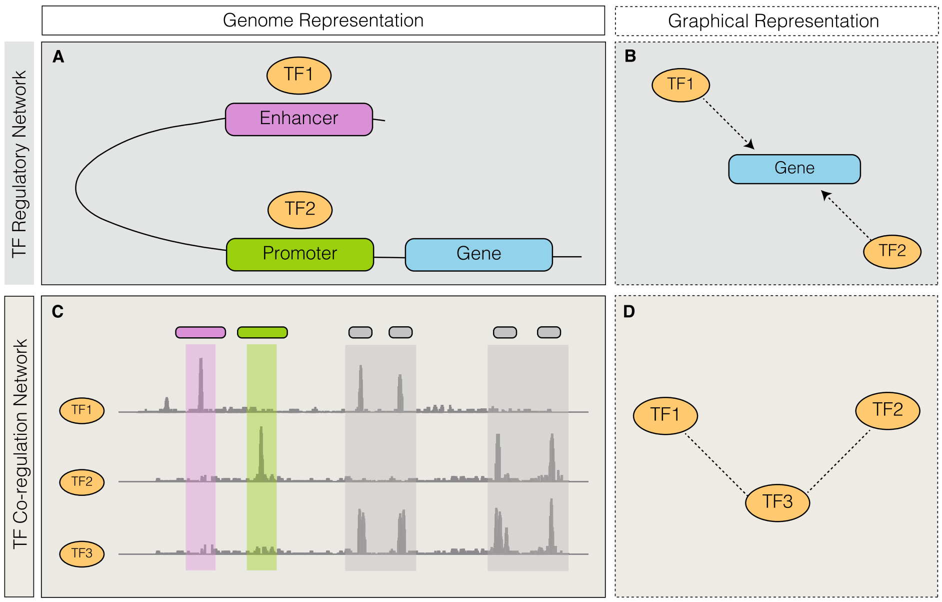


Fig. 2 A schematic of two types of networks are given, in both genomic and graphical form. **a** A regulatory network is given in genomic form where TF1 and TF2 both regulate the downstream gene, binding to an interacting enhancer and promoter, respectively. **b** A directed graph is shown where TF1 and TF2 both regulate the gene. **c** A genomic representation of a co-regulatory network is shown, based on the signal tracks of three TFs. **d** A graphical form of the co-regulatory network is shown, where TF3 co-binds with TF1 and TF2 separately

interacting domains that are necessary to initiate transcription in many genes. Then we draw two edges in our co-regulation network between TF$_1$ and TF$_3$, TF$_2$ and TF$_3$ (Fig. 2d).

Using differential networks to measure TF co-regulation alterations between states

After building the co-regulatory networks under one condition, we aimed to provide a quantitative measure of co-regulation alterations between two conditions (e.g., disease and normal states). Intuitively, our goal is to try and infer a differential co-regulatory network. In this differential network, while nodes still represent TFs, each edge now describes the level of change that each TF pair experiences during disease progression (Fig. 3). Therefore, we extended the GGM for one condition to a differential graphical model for two conditions by estimating the differences between two precision matrices (details in "Differential graphical model for co-regulation alteration in two states" section).

During the estimation process of these two precision matrices, small non-zero values are introduced and may lead to many spurious edges, resulting in a dense differential network with many false positives. Therefore, we introduced a regulation parameter, λ, to penalize the retained edge count in order to remove potential spurious edges. As shown in Fig. 4, higher λ values indicate a larger penalty in the network edge, resulting in a sparser estimation of the differential network.

Selecting an appropriate λ parameter is key to reliably inferring the TF co-regulation gain and loss events while simultaneously removing false positives. We further used a stability-based model selection method to choose the optimal λ based on sub-sampling of the genome [15]. The intuition of our model is that we encourage the network to be more inclusive of edges for the benefit of allowing us to scrutinize many possible

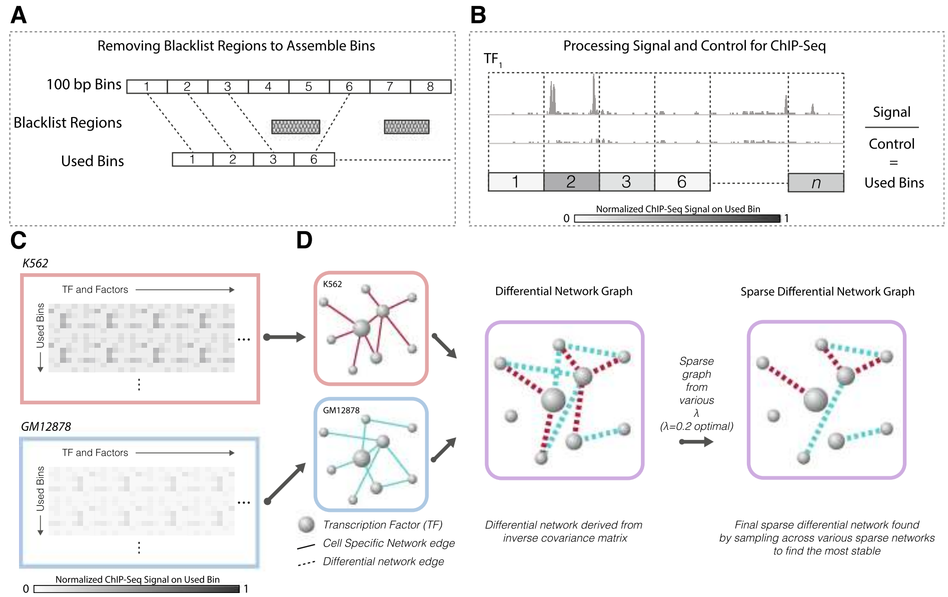

A
Removing Blacklist Regions to Assemble Bins
100 bp Bins
1
2
3
4
5
6
7
8
Blacklist Regions
Used Bins
1
2
3
6
B
Processing Signal and Control for ChIP-Seq
TF1
Signal
Control
=
Used Bins
1
2
3
6
n
Normalized ChIP-Seq Signal on Used Bin
0
1
C
K562
TF and Factors
Used Bins
GM12878
TF and Factors
Used Bins
Normalized ChIP-Seq Signal on Used Bin
0
1
D
K562
GM12878
Differential Network Graph
Sparse Differential Network Graph
Sparse graph from various λ (λ=0.2 optimal)
Transcription Factor (TF)
Cell Specific Network edge
Differential network edge
Differential network derived from inverse covariance matrix
Final sparse differential network found by sampling across various sparse networks to find the most stable

Fig. 3 Workflow of DiNeR method to create sparse differential regulatory network. **a** Bins used for the analysis are created by taking 100 bp bins on the genome and removing any blacklist regions to form the used bins; **b** For each TF and factor, the ChIP-Seq signal fold change is averaged over the used bins; **c** For disease and control cell types K562 and GM12878, respectively, a fold change signal matrix is created with rows or bins and columns of TFs and factors; 4) By using difference of the precision matrix, a differential network is created, which is then made sparse through sampling-based penalization

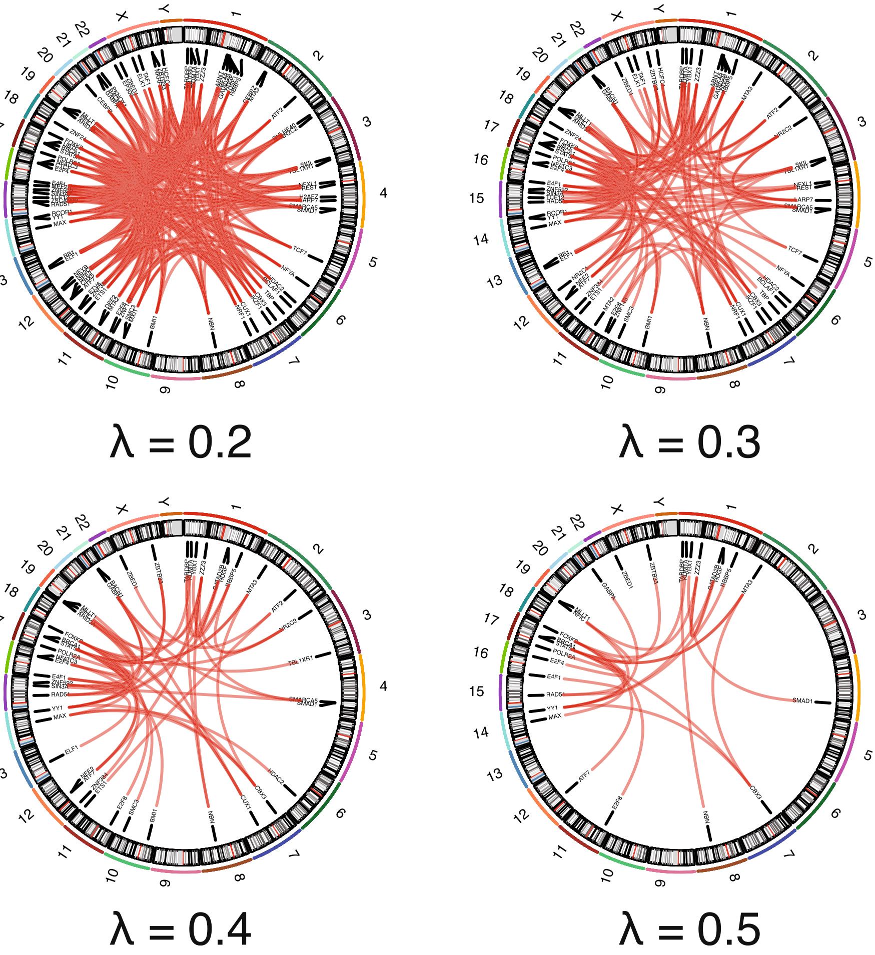


Fig. 4 Circos plots are used to show the sparsity of the network created using various lambda values. λ values of 0.2, 0.3, 0.4, and 0.5 are used and their networks are shown

changes associated with disease, while simultaneously ensuring that the results are highly repeatable across many regions in the genome (Fig. 5, details in "Model selection" section).

Applying DiNeR to prioritize key TFs associated with CML

We applied our DiNeR framework to 104 paired factors in K562 and GM12878 cell lines. After model selection, we set $\lambda_{opt} = 0.2$ (details in "Model selection" section). In total, we included 6.49% of all possible edges in the final network (351 out of 5408). We identified eight out of the 104 factors as consistent network hubs (Table 1), and reliably captured all of them by sub-sampling half of the genome (see Supplement S4). These eight factors include many well-known genes that have been previously associated with leukemia (Table 1), indicating that our method can reliably detect key regulators of disease.

One of the identified factors was the proto-oncogene BMI1, which is a major component of polycomb group complex 1 and plays a central role in DNA damage repair. Although BMI1 showed only moderate expression changes in tumor as compared to normal cells (approximately 20% higher; from 42.90 to 51.19), its co-binding

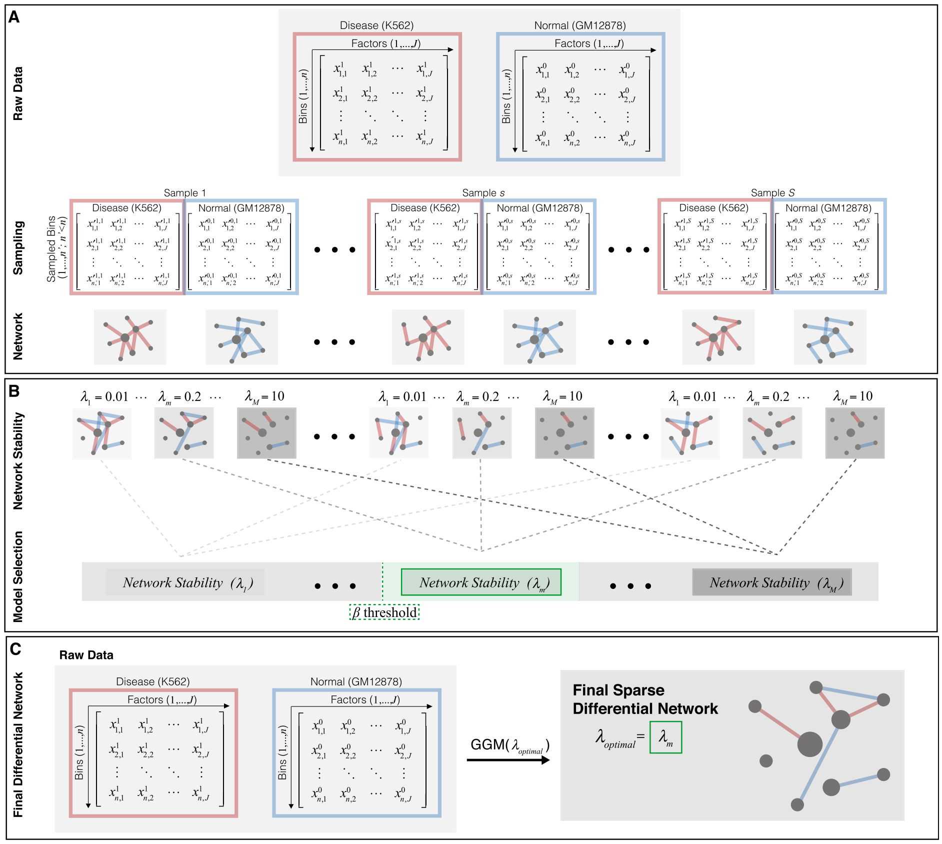


Fig. 5 A flowchart of the model selection process to create a sparse differential network is shown. **a** Raw data is sampled and graphical representations of the network are given for disease and control. The samples are different and therefore have varied network representations. **b** Differential networks are generated using GGM using various λ parameters, $\lambda : \{1, \ldots, \lambda_M\}$, and a model selection is done through stability analysis to see which λ gives the most stable network variance. A threshold for stability, β, allows us to choose the optimal λ, shown in green. **c** The optimal λ, in this case λ_m is used on the original raw data, to generate the final sparse differential network

Table 1 A list of DiNeR prioritized TFs using the network hubs of the K562 vs. GM12878 differential co-regulation networks upon different subsampling of the entire genome. Perc. Incl.: how many rounds of simulations out of 100 this TF has been claimed as a network hub. Lit. Supp: whether there is literature support to link this TF with cancer.

Gene	Perc. Incl.	Lit. Supp.
BRCA1	100	Y
FOXK2	100	Y
RAD51	100	Y
ZZZ3	100	N
RBBP5	100	Y
BMI1	100	Y
H3K27me3	100	Y
BACH1	98	Y

relationship with other factors changed dramatically. BMI1 was a significantly rewired TF and maintained edges with 14 other factors, including a well-known cancer-related histone modification mark H3K27me3. Interestingly, BMI1 is a known biomarker of hematologic malignancies and has been shown to be essential for faithful reprogramming of myeloid progenitors [26, 27]. Studies have reported that levels of BMI1 correlate with prognosis of patients with myelodysplastic syndrome and chronic and acute myelogenous leukemia [21]. Our analysis provides another possible explanation to this phenomenon, suggesting that in addition to aberrant expression patterns, the disruption of the coordination between key regulators during transcription can also contribute to disease progression.

Another interesting factor identified by our method was the tumor suppressor gene BRCA1 (Fig. 6a). BRCA1 has been shown to play a central role in maintaining genomic stability. Moreover, germline defects in this gene have been associated with multiple cancer types, including breast and ovarian cancers [28]. In normal cells, BRCA1 is typically associated with RNA polymerase II across multiple species to faithfully activate transcription of key genes [29]. However, we found that in K562 cells, the key interaction between BRCA1 and RNA polymerase II was significantly disrupted. Specifically, the Jaccard distance between the BRCA1 and POLR2A ChIP-seq peaks decreased by 98% in K562 as compared to GM12878 cells, indicating severe co-localization disruption. Consistent with this finding, we observed a remarkable proximal-to-distal shift in the BRCA1 binding locations in K562 cells (Fig. 6a and b). In particular, approximately 93.6% of the BRCA1 ChIP-seq peaks in GM12878 were located within a 5 kb region of annotated transcription start sites (proximal), and this number was drastically reduced to 16.4% for K562 peaks. Such a remarkable shift indicates that in cancer cells, BRCA1 fails to cooperate with other conserved collaborators to control the transcriptional process. BRCA1 has also been widely reported to affect chromatin structure and introduce chromatin remodeling. Interestingly, we also found that H3K4me1 and open chromatin regions were among the rewired partners of BRCA1. The overlapping peaks between BRCA1 and these factors were significantly reduced during the normal-to-tumor transition (p value $< 2 \times 10^{-16}$ for binomial tests in both cases). Hence, we hypothesize that BRCA1 severely alters not only its transcriptional regulation through promoter region interactions, but also its chromatin remodeling activity.

Further investigating TFs prioritized by DiNeR

To further explore the prognostic value of BRCA1 in leukemia and validate our model, we downloaded 151 patient expression and clinical profiles from The Cancer Genome Atlas (TCGA) for acute myeloid leukemia [30]. If we only consider the expression levels of BRCA1 in K562 and GM12878, we are unable to fully detect the difference in behavior of this gene between the two cell types. However, when considering the regulatory activity, we found that BRCA1 regulatory activity was increased by 50% in K562 as compared to GM12878 cells, suggesting that the regulatory role of BRCA1 could be of interest as compared to just the expression of this gene. In line with large-scale network rewiring events, we combined the BRCA1 regulatory network with patient tumor-to-normal differential expression profiles to measure the regulatory potential of BRCA1 in each patient. In contrast to the survival analysis using expression profiles alone, we

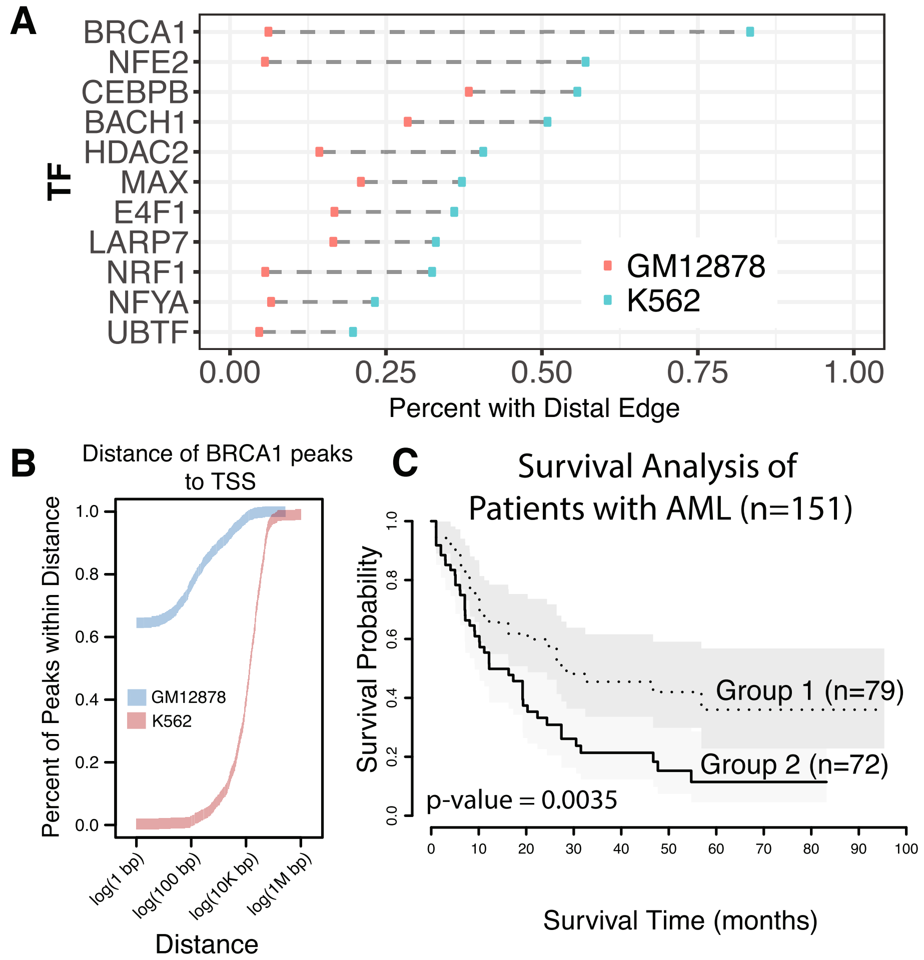


Fig. 6 **a** Summary of the proximal to distal change of binding peaks from GM12878 to K562. **b** Cumulative density of distance of BRCA1 peaks to the nearest TSS in both GM12878 and K562 showcases a dramatic shift in binding profiles between disease and control. **c** Survival analysis in AML using BRCA1 regulatory activity as a prognosis marker, showcasing disrupted BRCA1 regulatory activity results in a worse prognosis. Group 1 includes patients with above average regulatory activity of BRCA1 while group 2 contains those with below average regulatory activity

found that a severe disruption of BRCA1 regulatory activities usually indicated worse patient survival rates as compared to those patients with strong regulatory (p-value = 0.0035) (Fig. 6c). This analysis demonstrates that the differential graphical model can effectively identify key factors affecting patient survival beyond gene expression.

Discussion

Simultaneous binding of many TFs to proximal and distal regulatory regions of the genome is imperative for precise control of spatiotemporal gene expression patterns. Therefore, it is essential to investigate transcriptional regulation alterations in order to prioritize risk factors associated with disease. Many methods have been developed to address this goal. For instance, researchers investigated TF target gene gain and loss events and found distinct patterns between oncogenes and tumor suppressors [11]. Others combined expression changes with TF regulatory networks and identified TFs responsible for aberrant gene expression patterns, and then validated their discoveries through patient survival analysis [10]. Here, we focused on a unique aspect of transcriptional regulation – coordination among TFs, which goes beyond the binding or expression changes of TFs. For instance, genome-wide chromatin remodeling and histone modifications changes may introduce dramatic binding profile alterations for different TF, resulting in alterations of the combinatory coordination among TFs. For instance, some TFs, such as BRCA1 in our analysis, demonstrate massive binding profile changes. As a result, these TFs disrupt the co-binding relationship with many other TFs and chromatin features and are prioritized with the highest importance by our DiNeR framework. We also found that the target genes of TFs showed distinct expression patterns in tumors compared to normal cell lines, most likely due to extensive binding profiles changes. As a result, our analyses offer related but complementary views of transcriptional dynamics compared to previous methods and provide new insights into disease-related transcriptional regulation.

We also emphasize that model selection is a key step of our DiNeR framework. A larger penalty to the edge numbers captures the more confident co-regulatory network alterations at the cost of excluding weaker, but not necessarily insignificant, changes. Researchers have proposed various model selection methods, such as Akaike and Bayesian information criteria, to solve this problem [25]. In our framework, we used the stability-based model selection method, which provides a more interpretable explanation [15]. Specifically, we include an edge in the network as long as we can reliably discover it from numerous random subsamples of different genomic regions.

In addition, we ranked TFs according to the number of gained and lost partner TFs in the differential network. The intuition behind this scheme is that TFs that show a larger degree of genome-wide binding profile changes represent network hubs and are more likely to have disrupted coordination with many other partner TFs, resulting in a higher impact on transcriptional regulation. This is a reasonable assumption without any prior information. However, it is possible that even the disruption of one canonical TF pair may change the expression of key genes, such as in cases of oncogenes and tumor suppressor genes. Hence, it is also valuable to scrutinize the non-hubs of the network provided by DiNeR.

Finally, we showed that using matched ChIP-seq data from disease and normal cells is beneficial to directly capturing the co-localization events of TFs, as compared to other profiles such as expression. However, this approach requires the availability of hundreds of functional characterization datasets. We believe that as high-throughput sequencing technologies continue to develop, especially single-cell sequencing methods, our proposed differential graphical models could be applied to new opportunities to highlight regulators in disease.

Conclusion

We developed a TF-TF network rewiring and regulator prioritization method by applying non-parametric graphical models on large-scale functional genomics data. This approach allows us to identify DNA binding factors demonstrating large co-localization disruptions. Given the number of genome-wide binding profiles from ChIP-seq data in matched disease and control cells, our DiNeR method can reliably and efficiently highlight the significant changes of pairwise coordinated regulations between different factors. We applied our model to 104 common TF, histone modification, and chromatin accessibly data from a loosely paired tumor and normal cell line in CML. We discovered disruptions between well-known partners of transcriptional regulation, such as BRCA1 and RNA polymerase II, signifying the effectiveness of our method.

Methods

Here, we adopted a differential graphical model to investigate the differences in co-binding patterns of TFs between normal and disease conditions. We hypothesized that factors that dramatically change their partners during the transcription process would be altered to a larger degree in disease samples, and hence would have larger effects than TFs showing little difference in co-binding patterns in driving disease progression.

ChIP-seq data collection and pre-processing

We aim to infer the differential TF co-regulation alterations among normal and disease conditions. In order to assess these alterations, we selected disease-to-normal cell types that shared at least 50 common TFs (ChIP-Seq) as well as had a strong expression correlation between the two cell types (Figure S1). Therefore, we decided to use the 369 and 143 ChIP-seq experiments from K562 and GM12878 cell lines, respectively. After de-duplicating and extracting common ChIP-Seq targets, we identified 94 common TFs among these two cell lines (Table S1). In order to investigate alterations in the joint activity between TFs and specific chromatin marks in disease, we also extracted nine histone modifications and chromatin accessibility datasets from these cell lines (Fig. 1a, b). The majority of the 94 common TFs were sequence-specific binding factors (TFSS in Fig. 1c), 31 of which showed significant expression changes between disease and control.

To uniformly process the data, we first divided the autosomal chromosome (hg38 version) into 100 base pair (bp) bins and removed bins that overlapped with genomic regions that have gaps or low mappability using BEDTools (version 2.27.1-foss-2018b) [31]. To remove any artifacts from non-peak regions, we further removed all 100 bp bins that did not overlap with any peaks from these 104 factors (Fig. 3a). In total, we kept 1,351,140 bins in our analysis.

For each factor, we calculated the fold change of read count between the TF ChIP-seq experiment and its matched ChIP-seq control experiment from ENCODE for all bins in order to normalize for read depth (Fig. 3b). We then calculated the average signal from each of the replicates or experiments from different labs when multiple datasets were present for a single factor. Details of the ChIP-seq signal files used has been listed in the supplementary data. We organized the resultant signal data into a matrix for K562 and GM12878 separately, with columns indicating factors and rows indicating bins in the genome (Fig. 3c). We used these two matrices as the inputs for the following differential graphical model (Fig. 3d).

Infer differential graphical model to TF-TF network rewiring

Gaussian graphical model for co-regulation network in one state

In this section, we describe the details of the differential graphical model. Let $G^{(D)} = (V^{(D)}, E^{(D)})$ represent the network with nodes $V^{(D)}$ and edges $E^{(D)}$ for disease status D. Let $X_j^{(D)}$ denote the vector of the average normalized ChIP-seq signal of factor j in status D, where $j = 1\cdots J$. $D = 1$ indicates disease and 0 indicates normal. Next, for each D, we calculated $X_1^{(D)}, X_2^{(D)}, \cdots, X_J^{(D)}$ over 100 bp bins for all the 104 factors. In our analysis, we included 94 TFs, nine histone modification marks, and one chromatin accessibility ($J = 104$).

Under one condition D, we assume that $X_1^{(D)}, X_2^{(D)}, \cdots, X_J^{(D)}$ follows a multivariate Gaussian distribution such that $\boldsymbol{X}^{(D)} = \left(X_1^{(D)}, \cdots, X_J^{(D)}\right)^T \sim N_J\left(\boldsymbol{\mu}^{(D)}, \boldsymbol{\Sigma}^{(D)}\right)$. We can construct the TF co-regulatory network using a traditional GGM. Here, we are aiming to identify TF true physical interactions by highlighting conditional dependent binding profiles among TFs. Therefore, we used the precision matrix represented as $\boldsymbol{\Theta}^{(D)} \coloneqq (\boldsymbol{\Sigma}^{(D)})^{-1}$ to infer whether any TF pair has a non-random co-binding interaction. In other words, If $\boldsymbol{\Theta}_{ij}^{(D)} = 0$, then $X_i^{(D)}$ and $X_j^{(D)}$ are independent of each other, conditioned on all the other TFs. As a result, $(i, j) \notin E^{(D)}$ *if* $\boldsymbol{\Theta}_{ij}^{(D)} = 0$.

Differential graphical model for co-regulation alteration in two states

Next, we used the difference between two networks $G^{(1)}$ and $G^{(0)}$, called the differential network, to represent the degree of TF co-regulation alteration under two conditions ($D = 1$ and $D = 0$). Given the observed data $X_1^{(1)}, X_2^{(1)}, \cdots, X_J^{(1)}$ in the disease cell and $X_1^{(0)}, X_2^{(0)}, \cdots, X_J^{(0)}$ for the normal cell, edges in the differential co-regulatory network can be inferred from the difference between the two precision matrices $\boldsymbol{\Delta} = \boldsymbol{\Theta}^{(1)} - \boldsymbol{\Theta}^{(0)} = (\boldsymbol{\Sigma}^{(1)})^{-1} - (\boldsymbol{\Sigma}^{(0)})^{-1}$, where the co-regulation relationship between TF_i and TF_j changes if $|\Delta_{i,\ j}| \neq 0$. Note that $\boldsymbol{\Sigma}^{(1)}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{(0)} - (\boldsymbol{\Sigma}^{(1)} - \boldsymbol{\Sigma}^{(0)}) = 0$. Hence, we can solve the following equation to estimate a reasonable $\boldsymbol{\Delta}$.

$$\hat{\boldsymbol{\Sigma}}^{(1)}\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}^{(0)} - \left(\hat{\boldsymbol{\Sigma}}^{(1)} - \hat{\boldsymbol{\Sigma}}^{(0)}\right) = 0$$

Here $\hat{\boldsymbol{\Sigma}}^{(D)}$ is the sample covariance matrix. Specifically, we used the penalized D-Trace loss model estimate $\boldsymbol{\Delta}$ [16–18].

$$l(\boldsymbol{\Delta}) = \frac{1}{2}\, tr\left(\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}^{(1)}\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}^{(0)}\right) - tr\left(\boldsymbol{\Delta}\left(\hat{\boldsymbol{\Sigma}}^{(1)} - \hat{\boldsymbol{\Sigma}}^{(0)}\right)\right)$$

tr represents the trace of a matrix. To remove spurious differential edges, we introduced a non-negative regularization parameter λ to penalize the number of edges in the network.

$$l(\boldsymbol{\Delta}) = \frac{1}{2}\, tr\left(\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}^{(1)}\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}^{(0)}\right) - tr\left(\boldsymbol{\Delta}\left(\hat{\boldsymbol{\Sigma}}^{(1)} - \hat{\boldsymbol{\Sigma}}^{(0)}\right)\right) + \lambda \sum\nolimits_{i,j} \left|\Delta_{i,j}\right|_1$$

Here, λ controls the sparsity of the rewired network. For example, $\lambda = 0$ indicates no penalty and usually will result in a very dense network. In contrast, a large λ value will result in a sparse network.

Co-variance matrix inference

Under the Gaussian assumption, $\hat{\boldsymbol{\Sigma}}^{(D)}$ can be directly obtained from the sample covariance matrix. However, One statistical concern of using a differential graphical model is the distribution of $\boldsymbol{X}^{(D)}$. Here, we found that even after log transformation, almost all

TFs severely contradicted the Gaussian assumption (for details see suppl. sect. S2). Therefore, going forward we used a non-parametric model instead of a GGM. Our assumption is that a set of monotonically increasing functions $\{f_j^{(D)}\}_{j=1}^{J}$ exists such that, after transformation, $f_1^{(D)}(X_1^{(D)}), f_2^{(D)}(X_2^{(D)}), \cdots, f_J^{(D)}(X_J^{(D)})$ follow a multivariate normal distribution $N_J(\mathbf{0}, \mathbf{\Sigma}^{(D)})$. Similar to the GGM, we can use the precision matrix $\mathbf{\Theta}^{(D)} \coloneqq (\mathbf{\Sigma}^{(D)})^{-1}$ to infer the conditional dependence between any pair of factors in the network. As described in [17], we adopted the rank-based scheme to estimate the sample covariance matrix without directly estimating $\{f_j^{(D)}\}_{j=1}^{J}$. Specifically, let $r_{li}^{(D)}$ represent the rank of bin l for TF i in status D among all the bins, and n is the total number of bins in the genome. The Spearman correlation of TFs i and j are represented as below.

$$\rho_{ij}^{(D)} = \frac{\sum_{l=0}^{n}\left(r_{li}^{(D)} - \frac{n+1}{2}\right)\left(r_{lj}^{(D)} - \frac{n+1}{2}\right)}{\sqrt{\sum_{l=1}^{n}\left(r_{li}^{(D)} - \frac{n+1}{2}\right)^2 \sum_{l=1}^{n}\left(r_{li}^{(D)} - \frac{n+1}{2}\right)^2}}$$

Then we replace the sample covariance matrix $\hat{\mathbf{\Sigma}}^{(D)}$ by, $\hat{\mathbf{S}}^{(D)}$ with elements

$$\hat{\mathbf{S}}^{(D)} = \begin{cases} 2\,\sin\left(\frac{\pi}{6}\hat{\rho}_{ij}^{(D)}\right) & i \neq j \\ 1 & i = j \end{cases}$$

In cases where $\hat{\mathbf{S}}^{(D)}$ was not positively semi-definite, we used a projection method as described in [17, 32].

Model selection

It is critical to select an appropriate λ to reliably infer network changes in disease samples because different λ values can lead to different conclusions in downstream analyses. Researchers have proposed many methods, cross validation, Akaike information criterion (AIC), and Bayesian information criterion (BIC), to automatically select λ [33–35]. We chose a more interpretable approach in our ChIP-seq-based co-regulation network analysis than AIC and BIC, by using the Stability Approach to Regularization Selection (StARS) approach (Fig. 5) [15]. The key characteristic of this method is that it encourages the differential network to be inclusive to account for the true dynamics between disease and control networks, while guaranteeing an acceptable stability in the resultant differential network.

Specifically, we defined $\Lambda = {}^{1}\!/_{\lambda}$ as an alternative parameter to control network density so that a larger Λ indicates a denser network. During the model section process, we start from subsampling part of the genome for S times. Specifically, during the s^{th} sampling, $\boldsymbol{X}^{(1),\,s}$ and $\boldsymbol{X}^{(0),\,s}$ represent the binding profile matrices. $\psi_{i,j}^{s}(\Lambda) = 1$ if there is an edge between TF i and TF j in the rewired network under Λ, otherwise $\psi_{i,j}^{s}(\Lambda) = 0$. In the $s = 1, 2, \cdots, S$ randomly sampled datasets, we defined $\theta_{i,j}(\Lambda) = \frac{1}{S}\sum_{s=1}^{S}\psi_{i,j}^{s}(\Lambda)$, and $\xi_{i,\,j}(\Lambda) = 2\theta_{i,\,j}(\Lambda)\{1 - \theta_{i,\,j}(\Lambda)\}$ to be the fraction of times the networks disagree with the existence of the edge (i, j). Then, the overall instability of the networks over the sampling sets is

$$\hat{D}(\Lambda) = \frac{\sum_{i<j}\xi_{i,j}(\Lambda)}{\binom{J}{2}}$$

It is clear that $\hat{D}(0) = 0$ in an empty network because there is no instability when there are no edges. In general, the network becomes denser and more instable as Λ goes larger. However, when the network becomes very dense and even fully connected, $\hat{D}(\Lambda)$ goes smaller again and eventually reduces to zero. As suggested in [15], we used the monotone function $\overline{D}(\Lambda) = \ sup_{0\ll t\ll\lambda}\hat{D}(\Lambda)$ to remove such artifact effect in an extremely dense network. As a result, the optimal Λ should be $\hat{\Lambda}_{opt} = \ sup\{\Lambda : \overline{D}(\Lambda) \leq \beta\}$ for a predefine network instability measure β.

In our analysis, we started from a broad spectrum of parameter values from $\lambda_1 = 0.01$, $\lambda_2 = 0.05$, $\cdots\lambda_m$, $\cdots$, $\lambda_M = 10$, representing a wide range of sparse networks, from almost fully connected to empty (Fig. 4). For each λ_m, we randomly selected half of the bins we used in "Infer differential graphical model to TF-TF network rewiring" section to run the LASSO penalized D-Trace loss model. We repeated this process $S = 100$ times for each λ_m and calculated the average network variance. We used $\beta = 0.5\%$ as our stability threshold and selected the optional λ is 0.2.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03605-3.

Additional file 1.

Abbreviations

DiNeR: *Di*fferential Graphical Model of co-regulation *N*etwork *R*ewiring to Infer Transcription Factor Co-binding Alterations; TF: Transcription factor; CML: Chronic myeloid leukemia; GGM: Gaussian graphical model; ChIP-seq: Chromatin immunoprecipitation followed by sequencing; StARS: Stability Approach to Regularization Selection; TCGA: The Cancer Genome Atlas

Acknowledgements

We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure at Yale university.

Authors' contributions

JZ and JL designed the methods; JZ, JL, DL, and ZC performed all the analyses; SL and GG helped with expression data processing; all authors were involved in writing the paper. The author(s) read and approved the final manuscript.

Funding

We acknowledge support from the NIH (U24 HG 009446–02) and from AL Williams Professorship funds.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the ENCODE portal, https://www.encodeproject.org/ [21]. Accession codes can be found in the supplement. Users can directly download the raw data used in this paper from the ENCODE website with the accession code.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

We would like to note that Mark Gerstein is a member of the editorial board of this journal.

Author details

[1]Department of Computer Science, University of California, Irvine, CA 92617, USA. [2]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT 06520, USA. [3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [4]Department of Molecular Cellular and Developmental

Biology, Yale University, New Haven, CT 06520, USA. [5]Department of Computer Science, Yale University, New Haven, CT 06520, USA.

Received: 6 August 2019 Accepted: 15 June 2020
Published online: 02 July 2020

References

1. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–26.
2. Djebali S, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101–8.
3. Gerstein MB, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91–100.
4. Lee TI, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science. 2002;298(5594):799–804.
5. Golson ML, Kaestner KH. Fox transcription factors: from development to disease. Development. 2016;143(24):4558–70.
6. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152(6):1237–51.
7. Nebert DW. Transcription factors and cancer: an overview. Toxicology. 2002;181-182:131–41.
8. Santiago C, Bashaw GJ. Transcription factors and effectors that regulate neuronal morphology. Development. 2014;141(24):4667–80.
9. Bhagwat AS, Vakoc CR. Targeting transcription factors in Cancer. Trends Cancer. 2015;1(1):53–65.
10. Jiang P, et al. Inference of transcriptional regulation in cancers. Proc Natl Acad Sci U S A. 2015;112(25):7731–6.
11. Zhang J, et al. An integrative ENCODE resource for cancer genomics. bioRxiv. 706424. https://doi.org/10.1101/706424.
12. Tsankov AM, et al. Transcription factor binding dynamics during human ES cell differentiation. Nature. 2015;518(7539):344–9.
13. Qian J, et al. Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. Nucleic Acids Res. 2005;33(11):3479–91.
14. Zhong S, He X, Bar-Joseph Z. Predicting tissue specific transcription factor binding sites. BMC Genomics. 2013;14:796.
15. Liu H, Roeder K, Wasserman L. Stability approach to regularization selection (StARS) for high dimensional graphical models. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2. Vancouver: Curran Associates Inc.; 2010. p. 1432–40.
16. Yuan HL, et al. Differential network analysis via lasso penalized D-trace loss. Biometrika. 2017;104(4):755–70.
17. Zhang XF, et al. DiffGraph: an R package for identifying gene network rewiring using differential graphical models. Bioinformatics. 2018;34(9):1571–3.
18. Tian DC, Gu QQ, Ma J. Identifying gene regulatory network rewiring using latent differential graphical models. Nucleic Acids Res. 2016;44(17):e140.
19. Friedenson B. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. BMC Cancer. 2007;7:152.
20. Hamdy MS, et al. RAD51 and XRCC3 gene polymorphisms and the risk of developing acute myeloid leukemia. J Investig Med. 2011;59(7):1124–30.
21. Saudy NS, et al. BMI1 gene expression in myeloid leukemias and its impact on prognosis. Blood Cells Mol Dis. 2014;53(4):194–8.
22. Buchi F, et al. Redistribution of H3K27me3 and acetylated histone H4 upon exposure to azacitidine and decitabine results in de-repression of the AML1/ETO target gene IL3. Epigenetics. 2014;9(3):387–95.
23. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17.
24. Carter SL, et al. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 2004;20(14):2242–50.
25. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007;94(1):19–35.
26. Sahasrabuddhe AA. BMI1: a biomarker of hematologic malignancies. Biomark Cancer. 2016;8:65–75.
27. Yuan J, et al. Bmi1 is essential for leukemic reprogramming of myeloid progenitor cells. Leukemia. 2011;25(8):1335–43.
28. Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nat Rev Cancer. 2011;12(1):68–78.
29. Krum SA, et al. BRCA1 associates with processive RNA polymerase II. J Biol Chem. 2003;278(52):52012–20.
30. Cancer Genome Atlas Research, N, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20.
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
32. Liu H, et al. High-dimensional Semiparametric Gaussian copula graphical models. Ann Stat. 2012;40(4):2293–326.
33. Akaike H. In: Parzen E, Tanabe K, Kitagawa G, editors. Information Theory and an Extension of the Maximum Likelihood Principle, in Selected Papers of Hirotugu Akaike. New York: Springer New York; 1998. p. 199–213.
34. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
35. Efron B. The jackknife, the bootstrap and other resampling plans. https://doi.org/10.1137/1.9781611970319.

Publisher’s Note