SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• RESEARCH PAPER •

November 2014 Vol.57 No.11: 1090–1102
doi: 10.1007/s11427-014-4762-7

Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation

YU Hui[1], MITRA Ramkrishna[1], YANG Jing[2,3], LI YuanYuan[3] & ZHAO ZhongMing[1,4,5,6*]

[1]*Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, 37232, USA;*
[2]*School of Biotechnology, East China University of Science and Technology, Shanghai 200237, China;*
[3]*Shanghai Center for Bioinformation Technology, Shanghai 201203, China;*
[4]*Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee, 37232, USA;*
[5]*Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, Tennessee, 37212, USA;*
[6]*Center for Quantitative Sciences, Vanderbilt University, Nashville, Tennessee, 37232, USA*

Received July 3, 2014; accepted September 22, 2014; published online October 17, 2014

Identification of differential regulators is critical to understand the dynamics of cellular systems and molecular mechanisms of diseases. Several computational algorithms have recently been developed for this purpose by using transcriptome and network data. However, it remains largely unclear which algorithm performs better under a specific condition. Such knowledge is important for both appropriate application and future enhancement of these algorithms. Here, we systematically evaluated seven main algorithms (TED, TDD, TFactS, RIF1, RIF2, dCSA_t2t, and dCSA_r2t), using both simulated and real datasets. In our simulation evaluation, we artificially inactivated either a single regulator or multiple regulators and examined how well each algorithm detected known gold standard regulators. We found that all these algorithms could effectively discern signals arising from regulatory network differences, indicating the validity of our simulation schema. Among the seven tested algorithms, TED and TFactS were placed first and second when both discrimination accuracy and robustness against data variation were considered. When applied to two independent lung cancer datasets, both TED and TFactS replicated a substantial fraction of their respective differential regulators. Since TED and TFactS rely on two distinct features of transcriptome data, namely differential co-expression and differential expression, both may be applied as mutual references during practical application.

differential regulation, differential co-expression, differential expression, simulation, algorithm evaluation, transcription factor

Citation: Yu H, Mitra R, Yang J, Li YY, Zhao ZM. Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation. Sci China Life Sci, 2014, 57: 1090–1102, doi: 10.1007/s11427-014-4762-7

A gene regulatory network is a representation of the relationships between molecular regulators (usually transcription factors, abbreviated as TFs) and their regulating targets, which work collectively to establish simultaneous gene expression profiles. With recent, rapid improvements from both technical and computational aspects, genome-scaled regulatory networks have become accessible as scaffolds in studies of pathophysiological gene regulatory mechanisms [1,2]. Recently, increased awareness that regulatory network scaffolds can differ between conditions has given rise to a new research theme called "differential networking" [3,4]. Underlying the differential networking approach is a "differential regulation" issue, where the loss, gain, or rewiring of regulatory links occur at the localized topology of a baseline gene regulatory network [5,6]. In normal physiological conditions, differential regulation is occasionally observed at switches between different cellular consequences, such as TP53's alternative, contrary regulation

*Corresponding author (email: zhongming.zhao@vanderbilt.edu) effects on KLF4, leading to either cell cycle arrest or cell death [7]. In pathological conditions, sequence mutation or chromosome aberrance might disrupt the normal regulatory links between a regulator and its potential targets, causing the regulator in question to become a “differential regulator.” For instance, TP53 is a major differential TF in many cancer cases, when missense mutations cause it to fail to recognize wild-type binding sites [8]. An illustration of an instance of differential regulation involving single regulator inactivation is shown in Figure 1A.

Precise identification of underlying differential regulators is essential to the elucidation and possibly calibration of many pathophysiological processes. In the last few years, many computational algorithms have been devised to leverage regulatory relationship information towards identifying differential regulators from transcriptome responses. In 2010, Essaghir and colleagues [9] proposed an algorithm, TFactS, and demonstrated that TF regulation can be accurately predicted from the presence of differential expression gene (DEG) signatures in a TF’s target. Almost the same time, two other algorithms, RIF1 and RIF2, were introduced to integrate differential expression (DE) and differential co-expression (DCE) features [10,11]; they successfully identified myostatin as the main cause of muscle growth divergence between two cattle breeds [11]. Last year, we developed two algorithms, TED and TDD, for identifying differential regulators [12]. TED is engineered towards the enrichment of differential co-expression genes (DCGs) within targets, while TDD is directed towards the density of targets’ differential co-expression links (DCLs). We noted that another approach, the “correlation set analysis” [13], calculates the correlation levels between regulatees as an indication of a regulator’s activity, and we employed a simple adaption of its formula to further expand it into two variant forms “dCSA_t2t” and “dCSA_r2t” (see more details in Materials and methods). The two variant algorithms could, in principle, be used to identify differential regulators. These seven algorithms rely more or less on the regulator-target relationships of a baseline regulatory network (Table 1), so we generally refer them as “network-based” algorithms. Each of these algorithms has its own features and was assessed through specific evaluation strategies. It is necessary to align these alternative algorithms and subject them to uniform, systematic evaluation. The evaluation results will help users make the most appropriate methodological selection for their practical application needs.

To perform a systematic evaluation of the seven above-mentioned algorithms in this work, we established a simulation-based schema. Specifically, we used pre-defined, inactivated single or multiple regulators to benchmark the prioritization accuracy of these seven algorithms. The algorithm(s) that performed best using the simulation data were then further evaluated in two independent lung cancer expression profiling datasets, so as to examine the reproducibility of the identified differential regulators. Our simulation-based evaluation tests revealed the complexity of the issues surrounding differential regulator identification and disclosed some factors influencing discrimination accuracy in general. Based on the evaluation of the seven algorithms, two (TED and TFactS) had performance that is relatively better than the others. We found these two algorithms could replicate their findings well in two independent lung cancer datasets. Overall, results from this evaluation work may benefit applications of and future improvements on the related algorithms.

1 Methods

1.1 Seven differential regulator identification algorithms

Each algorithm works on a pair of expression matrices cor-

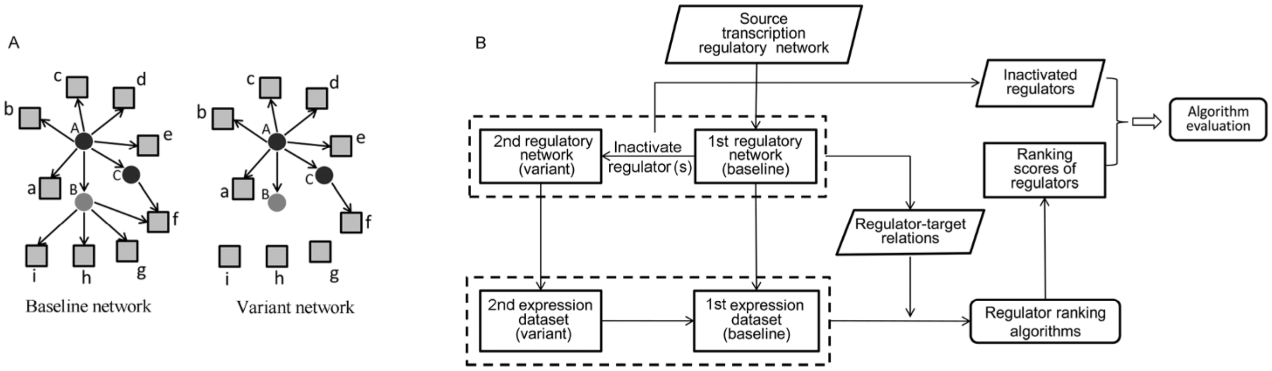
A
c
d
b
A
e
C
a
B
f
i
h
g
Baseline network
c
d
b
A
e
C
a
B
f
i
h
g
Variant network
B
Source transcription regulatory network
Inactivated regulators
Algorithm evaluation
2nd regulatory network (variant)
Inactivate regulator(s)
1st regulatory network (baseline)
Ranking scores of regulators
Regulator-target relations
2nd expression dataset (variant)
1st expression dataset (baseline)
Regulator ranking algorithms

Figure 1 Differential regulation problem and the framework of simulation-based evaluation. A, Illustration of two contrasting regulatory networks involving a single inactivated regulator. The baseline network includes three regulators (genes A, B, and C) and nine targets. In the variant network, all out-going edges of gene B disappear. Therefore, a differential regulation occurs and we call gene B a differential regulator. B, The simulation-based evaluation framework. A medium-scale baseline network was sub-selected from an actual transcriptional regulatory network of *Escherichia coli* (or yeast), and this baseline network was perturbed to characterize single or multiple inactivated regulators in the corresponding variant network. With the two expression datasets simulated based on baseline and variant networks, respectively, regulator ranking algorithms were called on to prioritize all candidate regulators. The prioritization list was examined against the known inactivated regulator(s) and the discrimination accuracy of each algorithm was assessed.

Table 1 Overview of seven algorithms on identifying differential regulators from transcriptome data

Algorithm	Key measurement	Feature[a)]	Network	Target dichotomy	Reference
TED	Enrichment of DCGs in targets	DCE	Required	Yes	[12]
TDD	Density of DCLs among targets	DCE	Required	Yes	[12]
TFactS	Enrichment of DEGs in targets	DE	Required	Yes	[9]
RIF1	Differential co-expression between regulator and DEGs, variant one	DCE & DE	Not required[b)]	Yes	[10]
RIF2	Differential co-expression between regulator and DEGs, variant two	DCE & DE	Not required[b)]	Yes	[10]
dCSA_t2t	Differential co-expression among regulatees	DCE	Required	No	[13][c)]
dCSA_r2t	Differential co-expression between regulator and regulatees	DCE	Required	No	[13][c)]

a) DCE, differential co-expression; DE, differential expression. b) RIF1 and RIF2 do not need a network of regulator-target relationships, but they expect a set of regulators as part of their input. c) dCSA_t2t and dCSA_r2t are two variant forms deriving from the CSA algorithm by [13].

responding to two contrasting experimental conditions. For five algorithms, a network of regulator-target relationships is required, while the other two (RIF1 and RIF2) do not need such a network but still expect a set of regulators as part of their input. Each of these seven algorithms provides output results in a similar format; that is, a scoring list of all candidate regulators. We recapitulate below the key formulae of each algorithm for calculating the prioritization score for a regulator TF_i (eqs. (1)–(7)), where i refers to a specific regulator. Table 1 summarizes the key features of these algorithms.

Five algorithms are characterized with a dichotomy of regulatory target genes (Table 1). This means that, in effect, target genes must be identified as either interesting or non-interesting, in regards to the desired expression feature. The interesting targets of the desired expression feature are either DEGs (in algorithms TFactS, RIF1, and RIF2) or DCGs (in TED and TDD). In this work, the fraction of interesting target genes (DEGs or DCGs) was designated as "key parameter", and we tested a range of key parameter values in the simulation experiments.

1.1.1 TED: Targets' enrichment of differential co-expression genes

$$TED(TF_i) = -\log 10\left(\sum\nolimits_{x=Ti'}^{T_i}\binom{T_i}{x}\left(\frac{K}{N}\right)^x\left(1-\frac{K}{N}\right)^{T_i-x}\right). \quad (1)$$

Here, N denotes the number of targets (also referred to as the "out-degree" in the text below) for all concerned regulators, K denotes the number of all DCG targets, T_i indicates the number of targets for a particular regulator ($\mathrm{TF_i}$ here), and T_i' indicates the number of targets for $\mathrm{TF_i}$ that are DCGs. Of note, here we changed the original base-2 logarithm [12] to the more intuitive base-10 logarithm. It should be noted that all targets are restricted to those contained in the expression data.

In this work, we used the algorithm DCe [14] to determine DCGs, where we adopted the swiftest link filtration method "percent" with a cutoff of 0.1. DCGs were selected based on the DCe's P-value ranking.

1.1.2 TDD: Targets' density of differential co-expression links

$$TDD(TF_i) = \frac{k_i}{N_i(N_i-1)/2}. \quad (2)$$

Here, N_i is the number of targets for TF_i and k_i is the number of DCLs formed within N_i targets. We used the algorithm DCe [14] to determine DCGs and DCLs with the same parameter setting as in TED. DCLs were limited to a default (coarse) fraction of 0.1, but were then further narrowed down to those connected with DCGs. So the TDD result was also dependent on the fraction of DCGs, i.e., key parameter value.

1.1.3 TFactS: Targets' enrichment of differential expression genes

$$TFactS(TF_i) = -\log 10\left(\sum\nolimits_{j=m_i}^{j=n_i}\frac{\binom{m}{j}\binom{N-m}{n_i-j}}{\binom{N}{n_i}}\right). \quad (3)$$

Here, N is the number of total target genes, n_i is the number of TF_i targets, m is the number of DEG targets, and m_i is the number of DEG targets of TF_i. In this work, genes were sorted by their absolute, between-condition mean expression differences, and a number of top-ranking DEGs were selected, depending on their key parameter value.

1.1.4 RIF1: Regulatory impact factor I

$$RIF1(TF_i) = \left|\frac{1}{n_{de}}\sum_{j=1}^{j=n_{de}} e_j d_j (r1_{ij} - r2_{ij})^2\right|. \quad (4)$$

Here, e_j is the mean log expression value of the jth DEG across all samples of the two conditions, and d_j is the difference of the same gene between the two mean log expression values from the two conditions. n_{de} refers to the total number of DEGs. $r1_{ij}$ indicates the Pearson correlation coefficient between $\mathrm{TF_i}$ and the jth DEG in the baseline (or, 1st) condition, and $r2_{ij}$ indicates the counterpart value in the variant (or, 2nd) condition. In this evaluation work, an outmost absolute conversion is added to the original formula [10]. In our application of RIF1, DEGs were determined in the same way as in TFactS.

1.1.5 RIF2: Regulatory Impact Factor II

$$RIF2\left(TF_i\right)=\left|\frac{1}{n_{de}}\sum_{j=1}^{j=n_{de}}\left[(e1_j\times r1_{ij})^2-(e2_j\times r2_{ij})^2\right]\right|. \quad (5)$$

Here, $e1_j$ and $e2_j$ indicate the mean log expression values for the two conditions, respectively; $r1_{ij}$ and $r2_{ij}$ indicate the Pearson correlation coefficient between TF_i and the jth DEG in the 1st condition and the 2nd condition, respectively. n_{de} refers to the total number of DEGs. As in RIF1, here we utilize an outmost absolute conversion and add it to the original formula [10]. In our application of RIF2, DEGs were determined in the same way as in TFactS and RIF1.

1.1.6 dCSA_t2t: Differential correlation set analysis between regulatees

$$\text{dCSA\_t2t}(\text{TF}_i)=\frac{2}{n_i(n_i-1)}\sum_{k=1}^{n_i-1}\sum_{j=k+1}^{n_i}|r1_{jk}-r2_{jk}|. \quad (6)$$

Here, $r1_{jk}$ and $r2_{jk}$ indicate the Pearson correlation coefficients between the jth and the kth targets of TF_i in the 1st condition and 2nd condition, respectively. n_i is the number of targets of TF_i. This index derives from the mean scoring function from the correlation set analysis study [13].

1.1.7 dCSA_r2t: Differential correlation set analysis between regulator and regulatees

$$dCSA_r2t(TF_i)=\frac{1}{n_i}\sum_{j=1}^{n_i}|r1_{ij}-r2_{ij}|. \quad (7)$$

Here, $r1_{ij}$ and $r2_{ij}$ indicate the Pearson correlation coefficients between the TF_i and the jth targets of TF_i in the 1st condition and 2nd condition, respectively. n_i is the number of targets of TF_i. We devised this index as analogous to the dCSA_t2t index.

1.2 Simulation-based evaluation framework

Two tools, SynTReN [15] and GeneNetWeaver [16], were adopted for artificial regulator inactivation and expression dataset simulation. With each tool, we first randomly selected a medium-scale regulatory network, which became the baseline regulatory network. Then, we imposed either single-regulator or multi-regulator inactivation, and as a result, obtained a variant regulatory network. Two expression data matrices corresponding to the two contrasting networks were simulated. The pair of expression data matrices, together with the regulatory relationships extracted from the baseline network, was supplied to each of the seven algorithms for regulator prioritization. Given the ranked lists of the candidate regulators outputted by each algorithm, we judged the accuracy of regulator prioritization by comparing the top-ranking regulators against the artificially inactivated regulator(s). A diagram of the evaluation framework is shown in Figure 1B.

1.2.1 Single-regulator inactivation with GeneNetWeaver

We used GeneNetWeaver [16] to select a 500-gene yeast transcription regulatory network and a corresponding expression dataset ("multifactorial experiment" of GeneNetWeaver's data output), as if it were profiled for multiple biological samples. This baseline network consisted of 1092 directed edges and 500 nodes (genes), of which 47 were candidate regulators (with two or more targets). Since GeneNetWeaver provided expression profiles for a total of 500 individuals, we randomly sampled 20 individuals each time and repeated the subsampling 40 times. In this way we obtained 40 datasets, each consisting of 20 individuals, for the baseline condition.

To simulate a regulator inactivation, we removed all outgoing edges from a particular regulator in the transcription regulatory network. In principle, the perturbed regulator and some of its targets can become orphan nodes and, thus, be absent from the synthetic expression dataset. When such a circumstance happened, we randomly assigned a regulator (not the inactivated regulator) to the would-be orphan nodes. In this way we obtained a variant network in which one particular regulator (the gold standard answer) was inactivated by the suspension of its regulations. As we did for the baseline condition, we also obtained 40 datasets, each consisting of 20 individuals, for this variant condition.

Collating one dataset from the baseline condition and one from the variant condition, we obtained a dataset pair, or two 500×20 data matrices. With this dataset pair and the regulatory relationships extracted from the baseline condition as input, each regulator-ranking algorithm worked to output a ranked list of the 47 candidate regulators. The ranks were converted to decimals so that the regulator at the top was assigned a value 1, while the regulator at the bottom was assigned a value 0. The decimal rank of the inactivated regulator was termed "priority of true answer" (PTA), which was used to compare the performance of the seven algorithms.

Outside the inner module, there were three layers of iterations. First, we tested the algorithms on a range of various key parameter values, specifically, from 0.01 to 0.20, with an increment of 0.01. Second, we applied the algorithms to 40 redundant dataset pairs (mimicking biological variation or technical variation) and evaluated their performances as an average of the 40 tests. Lastly, since 47 total regulators were involved, we inactivated a different regulator each time, that is, repeated the whole procedure 47 times.

1.2.2 Multi-regulator inactivation with SynTReN

We used SynTReN [15] to generate 11 pairs of simulated gene expression datasets based on sub-selected *E. coli* gene regulatory networks. We achieved multi-regulator inactivation in the corresponding variant networks by reducing the number of "external regulators." As explained in the SynTReN work [15], only the explicit external regulators trigger active, condition-specific transcription responses; those "turned-off" external regulators and their downstream cascades were excluded from the major transcription regulation program and they formed a constitutive background (refer to original publication [15] for technical details). The difference set of external regulators between the baseline and the variant networks, as well as their exclusive downstream regulators, were thus regarded as the gold standard differential regulators.

The baseline networks and variant networks were used by SynTReN to simulate data matrices, each involving 1000 genes and 10 samples. Two 1000×10 data matrices from the baseline and variant conditions, respectively, as well as those regulatory relationships extracted from the baseline network, were fed into each algorithm for regulator prioritization. As we did for single-regulator inactivation, here, we also investigated a range of key parameter values. Specifically, we examined these values from 0.05 to 0.5 with an increment of 0.05. We evaluated the prioritization of differential regulators in the output ranking lists with receiver-operating-characteristic (ROC) curves, and the area under the ROC curve (AUC) was used as a quantitative assessment index. Each algorithm was evaluated at its maximum AUC obtained at the optimal key parameter value.

1.3 Lung cancer gene expression datasets and preliminary analyses

We used two publically available gene expression datasets featuring non-small cell lung cancer to test the real data performance of two algorithms, TED and TFactS.

For the first dataset (denoted "Lung-I"), we obtained the transcript per million (TPM) values for 20502 genes from 16 tumor samples and 16 matched normal samples [17]. In a differential expression analysis described elsewhere [18], we identified 1504 DEGs and 3627 non-DEGs, which meant DEGs accounted for 29.3% of the combined set. In another aspect, the expression data containing the total 5131 genes (1504+3627) was analyzed by the differential co-expression analysis method DCe for discriminating DCGs and non-DCGs. All parameters were set as in the simulation experiments, except that the fraction of DCGs was fixed at 29.3% for comparability with the differential expression analysis.

For the second dataset (denoted "Lung-II"), we obtained the microarray log intensity values for 12756 genes from seven tumor samples and seven matched normal samples [19]. The concerned genes were reduced to 3910 after intersecting the genes with dataset Lung-I. A two-group differential expression analysis was conducted on this 3910× 14 data matrix using the Limma tool [20]. Based on the Limma results, we designated the top 1145 genes (29.3%) with an estimated false discovery rate of 0.147 as DEGs and designated the other 2765 genes as non-DEGs. Similar to the procedure used for dataset Lung-I, we adopted DCe to identify DCGs as 29.3% and designated the other genes as non-DCGs.

1.4 Human gene regulatory networks compiled from TRANSFAC

A larger-scale, prediction-based network of human TF-target relationships concerning the 5131 genes from dataset Lung-I was obtained through a search of TFBSs in the TRANSFAC data (release 2011.4) [21], using Match™ software [22]. We basically followed the technical pipeline proposed earlier [23], changing only the matrix score from 0.95 to 1.0. The resultant TRANSFAC-A network comprised 211417 relationships involving 344 TFs. The median number of targets per TF was 124.

A smaller-scale, experimental validation-based network of human TF-target regulatory relationships was derived from the gene annotation file ("gene.dat") in TRANSFAC (release 2013.2). After mapping to the same 5131 genes, this TRANSFAC-B network comprised 3046 relationships involving 572 regulators (mainly TFs, occasionally microRNAs). The median number of targets per regulator was four.

The two networks were reduced further by excluding TFs that were associated with no or only one expression-measured target gene. This constraint reduced the numbers of concerned TFs to 331 (TRANSFAC-A) or 359 (TRANSFAC-B) for the Lung-I dataset, and 330 (TRANSFAC-A) or 349 (TRANSFAC-B) for the Lung-II dataset.

The two networks represented two extremes, one with denser targets and the other with sparser targets. They did not have much overlap, as the TFs shared between the two networks accounted for no more than 45% of either set. These two regulatory networks were used as alternative baseline networks in testing the performance of TED and TFactS on two human lung cancer expression datasets.

2 Results

2.1 Validity of the simulation-based evaluation framework

As shown in the framework diagram (Figure 1B), the initial biological difference between the two conditions lay in the network difference, and then this difference flowed downward through simulated expression data to the resultant ranking scores of candidate regulators. We first attempted to ensure that it was the initial network difference that governed the resultant ranking scores, ruling out the possibility that the observed regulator ranking was attributed to technical biases. For this purpose, we performed a "contrasting" simulation experiment, which involved dataset pairs originating from two contrasting regulatory networks, and a "homogeneous" experiment, which involved dataset pairs originating from two identical regulatory networks. We compiled 40 dataset pairs for each scenario, and hence obtained 40 ranked lists for the homogeneous case and another 40 for the contrasting case. In the homogeneous network comparisons, the resultant regulator ranking was random, and the alternative rankings should not have substantial mutual consistency. In the contrasting network comparisons, however, if the regulator ranking algorithm had sufficient discrimination power, then the persistent structural difference should drive the 40 trial results towards a consensus reflective of the genuine structural divergence.

For each algorithm's multiple results out of the 40 repetitive runs, we calculated the Spearman correlation values for each of the 780 combinations formed from the 40 score lists for the contrasting experiment and the homogeneous experiment, respectively, and we comparatively showed the mean Spearman correlation values as bar plots (Figure 2). Remarkably, each algorithm demonstrated more consistent results in the contrasting case than in the homogeneous case for all or most of the surveyed key parameter values (Mann-Whitney test P-value<0.01; Figure 2). While Figure 2 involved perturbation of one specific regulator, the same conclusion held for the perturbation of every regulator (data not shown). Such an increase in result consistency in the presence of biological difference implied that all existing algorithms were able to reflect regulatory difference signals introduced through regulator inactivation in their ranking results. Therefore, we judged that our simulation framework (Figure 1B) was valid for our purposes of evaluating the discrimination ability of algorithms in regards to the (inac-

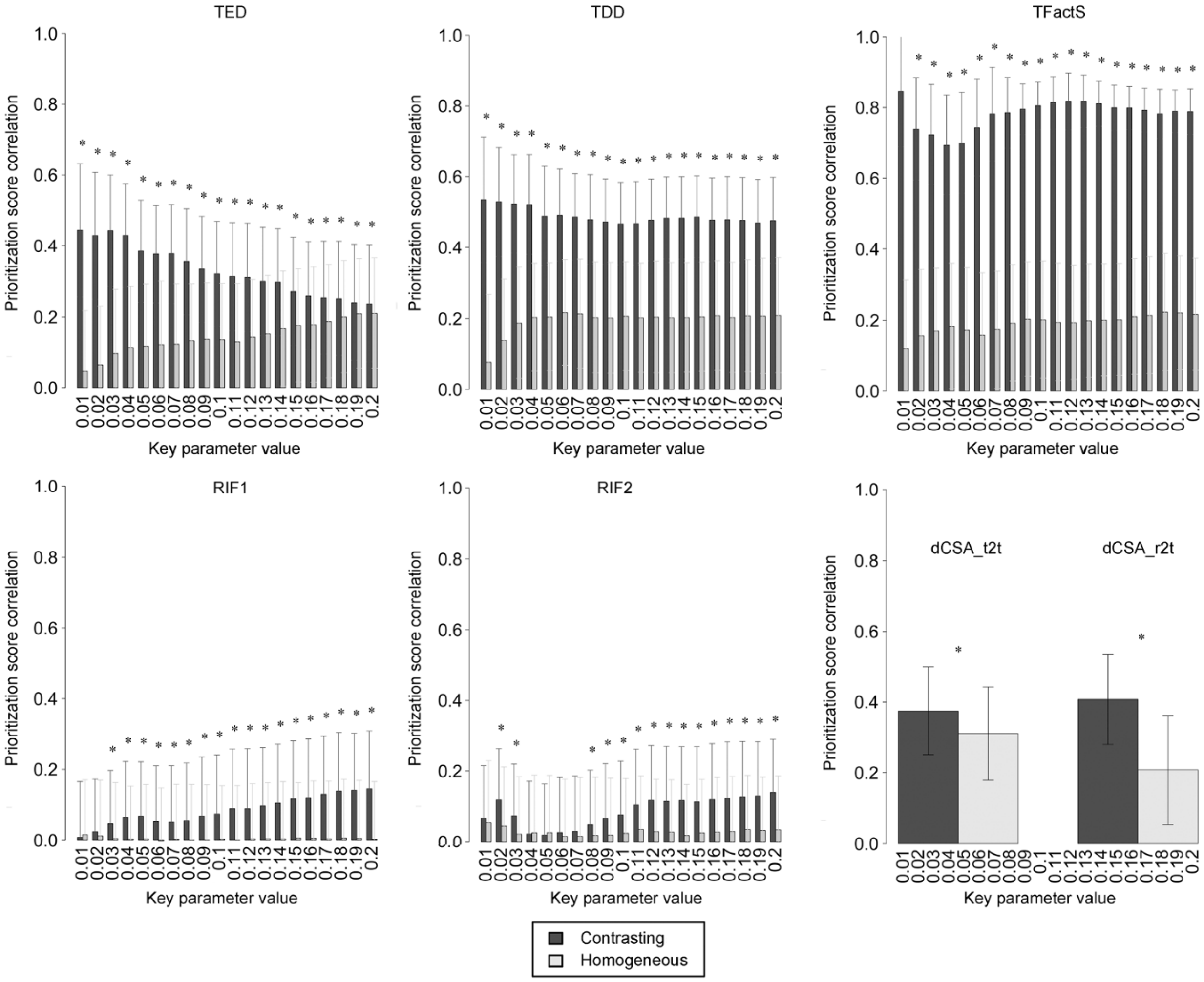

Figure 2 Regulator rankings are more consistent in the presence of biological difference than in the absence of biological difference. Seven regulator-ranking algorithms (subplot titles) were implemented to rank 47 candidate regulators based on a pair of simulated expression datasets, which were derived either from two differential regulatory networks ("contrasting") or from two identical networks ("homogeneous"). Wherever applicable, a series of the algorithms' key parameter values were investigated. With key parameter value being fixed, 40 redundant dataset pairs were simulated for repeated testing. Each subplot shows the mean and standard deviation of Spearman correlations of 780 pairs formed from 40 redundant ranking lists. An asterisk (*) indicates a significant difference between contrasting result consistency and homogeneous result consistency (Mann-Whitney test P-value<0.01).

tivated) differential regulator(s).

While only mutual result consistency is shown in Figure 2, we could have a glimpse of some technical characteristics of the surveyed algorithms therein. The contrasting result consistency actually indicated the algorithms' robustness against variation arising from sample choice or technical noise. As a result, TED, TDD, and TFactS appeared more stable in this regard than other algorithms (Figure 2). The homogeneous result consistency indicated how much an algorithm was biased towards a certain default regulator ranking. And indeed we observed certain result consistency in the absence of differential regulation for TED, TDD, TFactS, dCSA_t2t, and dCSA_r2t (Figure 2). Coincidentally, these five algorithms all require a defined regulator-target network (Table 1), and in particular, TED and TFatS rely on statistical tests in which a regulator's out-degree plays a decisive role (eqs. (1) and (3)). We presume that many of these algorithms may be biased towards regulators with larger out-degrees. This notion was further supported, as shown in the next section. Nevertheless, we noticed that, in general, smaller key parameter values were associated with lower homogeneous result consistency but higher contrasting result consistency, a pattern most evident in TED yet still discernable in TDD and TFactS (Figure 2). This observation may suggest technical biases were less severe with reasonably small key parameter values.

2.2 TED and TFactS outperformed other algorithms in simulation evaluation

In single-regulator inactivation tests, we had 47 sets of results, each for a particular inactivated regulator. In each set of results, PTA scores were retrieved for seven algorithms across a range of tested key parameter values. The two sets of results for the regulators with maximum and minimum out-degrees, respectively, are shown in Figure 3A and B, while all 47 sets of results were averaged and shown in Table 2. In the results for the most extensively regulating regulator (Figure 3A), we observed that TED and TFactS overall outperformed the other algorithms with the highest PTA scores throughout a majority of the key parameter value range. Actually, when all results for the total 47 regulators were summarized, TED and TFactS indeed turned out to be the best and second-best algorithms, respectively, as measured by accuracy (Table 2, column "PTA"). TED and TFactS also had better robustness against data variations than most other algorithms (Table 2, column "RAV"), suggesting that their results might be more stable in varied sample recruitment in real practical usage. However, TED and TFactS were sensitive to key parameter value (Table 2, column "RAP"), which means that special attention must be paid to the fraction of interesting genes (DEGs or DCGs) in real practical application.

As shown in the two subplots of Figure 3, there was a major difference in accuracy between the regulator with the maximum out-degree and the regulator with the minimum out-degree. As we reasoned earlier, differential regulator prioritization accuracy may be correlated with the regulator's out-degree. When we systematically analyzed this relationship for the 47 separately inactivated regulators, we did find a significant positive correlation between PTA scores and TF out-degrees for TED, TDD, TFactS, and RIF1 (Figure 4A), implying that an extensively regulating regulator was likely more discoverable. Another algorithm, dCSA_r2t, demonstrated a significant negative correlation between the PTA scores and mean in-degrees of the targets of perturbed regulators (Figure 4B). As small in-degrees correspond to dominant influences of the single regulator on its targets, it is indicated that an exclusively regulating regulator may also likely be discovered.

Once the PTA scores for four algorithms over each single regulator were plotted on Figure 4A, we could compare the accuracies of the involved algorithms over a broad view. Still, it was evident that the upper portions, characterized by higher PTA scores, were dominated by TED and TFactS. Examining the points on each vertical line led to a degree-

Table 2 Evaluation results based on 47 single-regulator inactivation experiments

	PTA[a)]		RAP[b)]		RAV[c)]	
	PTA value	PTA rank	RAP value	RAP rank	RAV value	RAV rank
TED	0.621±0.171	1	0.038±0.017	5	0.332±0.091	3
TDD	0.571±0.188	5	0.036±0.021	4	0.440±0.055	2
TFactS	0.616±0.225	2	0.110±0.048	7	0.768±0.032	1
RIF1	0.509±0.067	7	0.036±0.014	3	0.035±0.033	7
RIF2	0.550±0.128	4	0.054±0.035	6	0.106±0.054	6
dCSA_t2t	0.515±0.159	6	0[d)]	1	0.317±0.041	4
dCSA_r2t	0.597±0.240	3	0[d)]	1	0.296±0.060	5

a) PTA, a measure of discrimination accuracy. Here, the value is the "priority of true answer" (see detailed explanation in text), summarized across multiple dataset pairs, multiple parameter settings, and multiple regulator manipulations. b) RAP, robustness against parameter variation. Here, the value is the standard deviation of PTA scores over parameter settings, summarized across dataset pairs and varied regulators. c) RAV, robustness against data variation. Here, the value is PTA scores' Spearman correlation values among repetitive dataset pairs, summarized across parameter settings and varied regulators. d) dCSA_t2t and dCSA_r2t did not rely on a key parameter value, so their results were static against any parameter value variation.

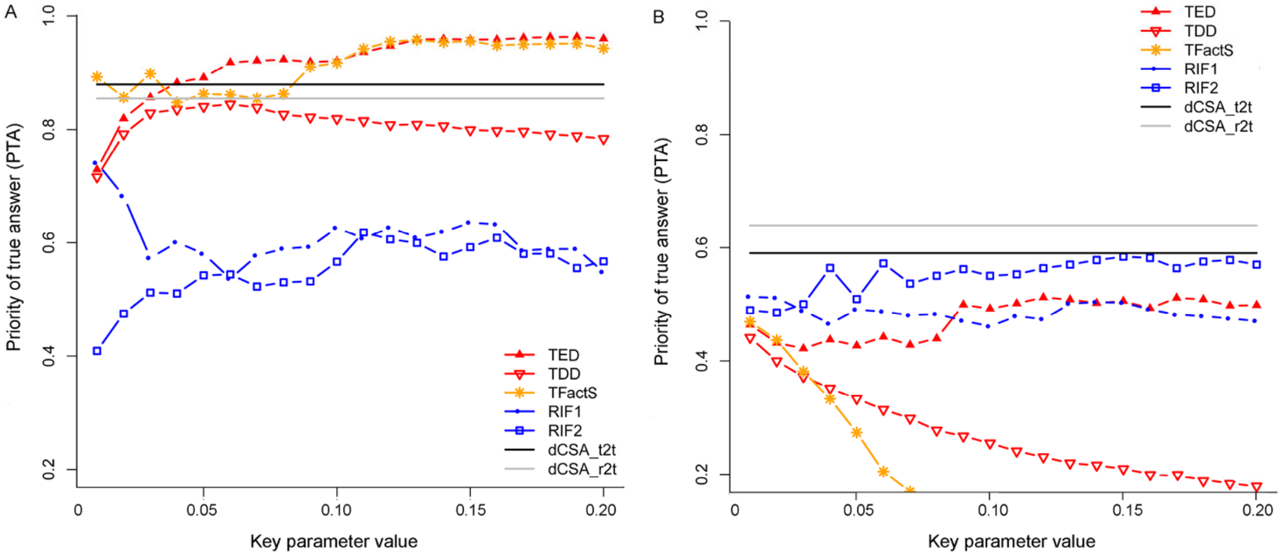

Figure 3 (color online) Performance comparison of seven methods in the scenario of one single regulator being inactivated. A, Results for the regulator with maximum out-degree (72). B, Results for the regulator with minimum out-degree (3).

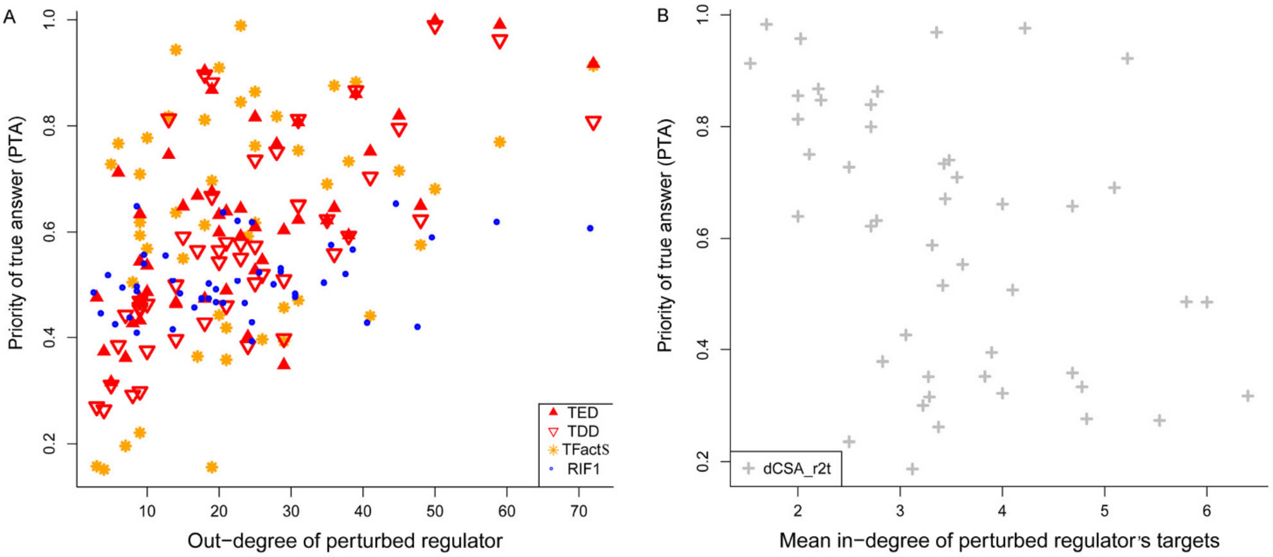

Figure 4 (color online) Correlation between discrimination accuracy of 47 separately inactivated regulators and their regulatory characteristics. A, A significant, positive Pearson correlation (P-value<0.01) was observed between PTA (priority of true answer) scores and out-degrees of the regulators in four algorithms (TED, TDD, TFactS, and RIF1). B, A significant negative Pearson correlation (P-value<0.01) was observed between PTA scores and the mean in-degree of each regulator's targets in algorithm dCSA_r2t.

specific comparison of algorithm accuracies, and the results may suggest an advantage of TED over TFactS for differential regulators with higher-degrees (Figure 4A).

Lastly, we showed the discrimination accuracies when multiple regulators were inactivated (Table 3). In total, we designed 11 simulation cases, where the first eight (Table 3, cases A1–A8) shared a common baseline regulatory network, and the next three had separately selected baseline networks (Table 3, cases B, C, and D). On average, TED and TFactS were ranked the best and the second-best of the seven algorithms, respectively. However, there was considerable variation in the AUC values between cases, in particular, between different baseline networks (Table 3). The AUC value did not appear to correlate with the fraction of differential regulators, as the toughest case (case B, Table 3) for most algorithms happened to have the smallest fraction of differential regulators. We earlier found that the out-degrees of individual regulators, and at times, the in-degrees of the targets, were important factors affecting the discrimination accuracy of the algorithms. Thus, when multiple regulators were simultaneously inactivated, the scenario became much more complex. More rigorous tests are

Table 3 Prioritization accuracies in multi-regulator inactivation experiments

Case	# nodes	# edges	# regulators	DR fraction[a)]	AUC[b)]						
					TED	TDD	TFactS	RIF1	RIF2	dCSA (t2t)	dCSA (r2t)
A1	1000	2309	103	30%	0.69	0.61	0.60	0.62	0.48	0.53	0.57
A2				29%	0.70	0.67	0.62	0.51	0.67	0.51	0.61
A3				28%	0.69	0.58	0.64	0.62	0.43	0.48	0.61
A4				27%	0.77	0.60	0.60	0.48	0.68	0.51	0.54
A5				26%	0.76	0.62	0.63	0.42	0.51	0.58	0.49
A6				24%	0.66	0.64	0.58	0.69	0.77	0.59	0.53
A7				23%	0.69	0.58	0.55	0.63	0.68	0.61	0.52
A8				18%	0.63	0.56	0.55	0.68	0.65	0.54	0.62
B	1000	2293	95	16%	0.50	0.37	0.47	0.44	0.38	0.42	0.78
C	1000	2322	105	49%	0.65	0.71	0.54	0.33	0.63	0.48	0.50
D	1000	2301	98	20%	0.63	0.55	0.69	0.54	0.62	0.49	0.47

a) Fraction of differential regulators of the total regulators. b) Area under the curve (AUC) of receiver-operating-characteristic (ROC).

Table 4 Summary of differential regulators identified from two lung cancer datasets

Network	Dataset	TED	TFactS	TED+TFactS
TRANSFAC-A	Lung-I	21	63	3
	Lung-II	9	30	3
	Replicated	0	16	0
TRANSFAC-B	Lung-I	7	7	0
	Lung-II	10	2	0
	Replicated	2	1	0

warranted to elucidate the mechanisms underlying multi-regulator inactivation scenarios.

2.3 TED and TFactS replicated multiple differential regulators in two lung cancer datasets

Since TED and TFactS were found the most accurate algorithms in the above simulation evaluations, we extended the evaluation of these two algorithms by using two real lung cancer expression datasets. The ranked TF lists outputted by TED and TFactS, respectively, were limited to a threshold of 1.3, corresponding to a nominal *P*-value of 0.05. Depending on the network and the algorithm choice, anywhere from a couple to tens of differential regulators were retrieved (Table 4). In general, more differential regulators were associated with dataset Lung-I (with more samples) than dataset Lung-II, and more differential regulators were associated with the TRANSFAC-A (with more regulatory relationships) than with the TRANSFAC-B network. A minor violation to this general pattern was found when TED run on dataset Lung-II with the TRANSFAC-B network; this combination led to 10 differential regulators, which was slightly greater than that (9) out of the larger network or that (8) out of the larger dataset.

We first compared TED and TFactS in terms of the number of prioritized regulators. As shown in Table 4, TFactS identified more differential regulators than TED with the larger network TRANSFAC-A (63 vs. 21, or 30 vs. 9), but equal or fewer differential regulators with the smaller network TRANSFAC-B (7 vs. 7, or 2 vs. 10). Then, we checked the replication scenario of each algorithm from dataset Lung-I to dataset Lung-II. Using the larger network TRANSFAC-A, those reproduced numbered 0 of TED's initial 21 regulators, and 16 (25.4%) of TFactS's initial 63 regulators. Using the smaller network TRANSFAC-B, those reproduced were two (28.6%) of TED's initial seven regulators, and one (14.3%) of TFactS's initial seven regulators (Table 4). Given these two layers of comparative results, we might speculate that TFactS worked better in a larger-scale, denser regulatory network, while TED is comparable to TFactS in a smaller-scale, sparser regulatory network. However, due to the limited number of datasets, the comparative conclusion may not be generalizable to future cases. Of note, genes contained in dataset Lung-I were more discriminable from the DE perspective than from the DCE perspective, as genes with borderline DE features were not included (see more details in [18]); accordingly, dataset Lung-II was also biased towards the DE feature. Indeed, from Lung-I to Lung-II, we observed significant consistency in the DEG/non-DEG classification (Fisher's exact test, *P*-value<2.2×10^{-6}), but no significant consistency in the DCG/non-DCG classification. Though these two datasets were apparently favorable to TFactS, TED still showed comparable performance under the TRANSFAC-B network. It is expected that TED may show an even better performance in real applications involving unbiased sets of genes.

Regardless of algorithm choice, quite a few differential TFs reproduced from Lung-I to Lung-II. A total of 19 repetitively identified regulators are listed in Table 5 as a reference for other researchers. Of these 19 TFs, five (ARID5B, IRF1, MAX, SPI1, and TCF3) were covered in our two expression data matrices. These five TFs generally had medium to high expression levels in Lung-I dataset as compared to the total genes, but some showed a dramatic decrease of expression level in the other dataset Lung-II. Two TFs were considered as DCGs in dataset Lung-I but not in Lung-II; three TFs were considered as DEGs in dataset Lung-I and two of them (TCF3 and SPI1) were repetitively differentially expressed in dataset Lung-II. According to these observations of specific cases, we might infer that differential regulators might not demonstrate remarkable and stable expression features on their own. The algorithms could discern their importance through analyzing the systematic expression changes among their target genes.

We found that SPI1 was detected as a reproduced differential regulator by TFactS (Table 5). The oncogenic TF SPI1 reportedly accelerates DNA replication and promotes genetic instability in the absence of DNA breakage in leukemia [24]. However, reports on the role of SPI1 in lung cancer development are rare. TED identified two TFs (MAX and E2F1) as reproducible in two independent lung cancer cohorts (Table 5). Intriguingly, MAX inactivation in lung cancer disrupts the MYC-SWI/SNF program, and an aberrant MYC-SWI/SNF network is essential for lung cancer development [25]. Another TF, E2F1, is required for GCN5 (a lysine acetyltransferase that generally regulates gene expression) to mediate lung cancer cell growth and promote the proliferation of a lung cancer cell line [26]. These additional evidences from literature indicate that the repetitively identified differential regulators are highly likely causal to lung cancer development.

Other than the three TFs (SPI1, MAX, and E2F1) discussed above, some other TFs in Table 5 may also be worth noting for follow-up investigation. According to a cancer gene compendium NCG v4.0 [27], *GLI1*, *ZIC3*, *TCF3*, and *HNF1B* are either known or candidate cancer genes, but existing studies have not linked them to lung cancer yet. Four TFs repeatedly identified by TFactS, GTF2I, GLI1, ZIC1, and ZIC3, were also accredited by TED in either the Lung-I or Lung-II dataset. These highlighted TFs likely have more pathogenic potential in the development of lung cancer.

3 Discussion

In a simulation evaluation framework, we defined a specific type of differential regulation, i.e., regulator inactivation, as the loss of all regulations from a single regulator or multiple regulators. With this problem-definition and the pre-defined gold standard differential regulators, we evaluated the accuracy and robustness of seven network and transcriptome-based differential regulator identification algorithms. We found that all algorithms could discern signals arising from genuine differential regulation, indicating the validity of our simulation-based evaluation framework. We inferred that extensively regulating or sometimes exclusively regulating regulators are easier to identify if they are individual-

Table 5 Differential TFs identified from both lung cancer datasets by TFactS or TED

Algorithm/network	TF	Dataset Lung-I		Dataset Lung-II	
		Score	Rank	Score	Rank
TFactS/ TRANSFAC-A	GTF2I	3.8	8	2.2	2
	IRF1	3.4	15	1.7	16
	RBPJ	2.8	19	1.4	26
	GLI1	2.5	23	2.1	3
	NKX2-2	2.2	25	1.3	30
	ZIC1	1.9	35	1.5	24
	ZIC3	1.9	35	1.5	24
	MYOD1	1.8	39	1.8	13
	NR4A2	1.7	40	2.1	1
	ASCL1	1.7	44	1.8	7
	MYF5	1.7	44	1.8	7
	MYF6	1.7	44	1.8	7
	TCF4	1.7	44	1.8	7
	ARID5B	1.6	47	1.9	5
	TCF3	1.6	51	1.8	14
	HNF1B	1.6	53	1.8	15
TFactS/ TRANSFAC-B	SPI1	2.1	3	2.0	1
TED/ TRANSFAC-B	MAX	1.8	1.5	1.3	9.5
	E2F1	1.4	7	1.4	8

ly inactivated. Based on our evaluation results, two algorithms, TED and TFactS, have shown to be more robust, as they excelled in single and multi-regulator inactivation tests. These two algorithms were further found to replicate a substantial fraction of their identified regulators across two independent lung cancer expression datasets.

TFactS and TED rely on two distinct features of gene expression data—differential expression and differential co-expression, respectively. Differential expression is the most intuitive feature of gene expression and may be validated in a straightforward manner. When differential regulation is most characteristic of target genes' differential expression, it corresponds to changes to node properties (i.e., gene expression levels) within the regulatory network, instead of changes to edge wiring (i.e., regulatory links). In many cases, differential edge wiring, or edge loss in particular, occurs in association with node property change, such as the decreased expression of targets resulting from an activating TF's inability to bind targeted DNA segments. This supposition may explain why TED and TFactS, designed from quite different rationales, both stood out from the several other investigated algorithms. However, more essential to general differential wiring is the change in correlation between the expression levels of the TF and its targets [28], and consequently (and probably more legibly), the change of correlation among the TF's target genes. Examining all the simulation-based evaluations together, we noticed that TED showed a slight advantage over TFactS (Tables 2 and 3). Another differential co-expression-based algorithm, TDD, was occasionally comparable or even superior to TFactS (Table 3). Due to their comparable performances demonstrated in the current simulation study, we recommend that TED and TFactS be used as mutual references to each other during practical application.

Our evaluation results are useful for future improvement of such differential regulator identification algorithms. For instance, while it appeared a wise strategy to integrate DE and DCE together, RIF1 and RIF2 surprisingly returned with unsatisfactory discrimination accuracy (Tables 2 and 3). As they are the only algorithms to neglect the accessible regulator-target relationships, we speculated that the results of RIF1 and RIF2 would have improvement if they were modified to accommodate this available information. Also, dCSA_t2t and dCSA_r2t are two algorithms that do not require target dichotomy; they directly deal with the co-expression difference measured between gene pairs involving regulatees and at times, the regulator. Their current suboptimal results remind us that perhaps a soft-thresholding or a half-thresholding of pairwise expression correlation values may enhance their performance [14]. Finally, we observed many zero TDD scores in most simulation experiments (data not shown) and realized that the design of TDD should be adjusted, as it lacks the resolution to quantify the density of DCLs among the targets of a common regulator. In fact, this issue and improvement was recently proposed elsewhere [29]. Despite its limited resolution, TDD showed a comparable performance to TFactS in multi-regulator inactivation tests (Table 3). Of note, the key parameter values tested in this work were the fractions of interesting genes rather than the fractions of interesting gene pairs (links), so TDD, a method dependent more on DCL fraction than DEG fraction, was not evaluated closely enough. Given necessary methodological improvement and a more targeted evaluation strategy, TDD may potentially show better results than what was demonstrated in the present work.

It is advocated to target transcription factors for cancer gene therapy [30,31]. Based on the cancer gene compendium NCG 4.0 (including 2000 cancer genes) and a comprehensive set of human TFs [32] (including 1987 TFs), we found 313 TFs whose genes are cancer-annotated, of which 35 were explicitly associated with lung cancer. The multitude of cancer TFs or lung cancer TFs were statistically significant when a total of 22836 human genes (according to Ensembl statistics as of 3/27/2014, http://www.ensembl.org/index.html) were used as the background of hypergeometric tests for enrichment (P-value=7.1×10^{-4} for general cancer, and 0.0019 for lung cancer). Therefore, it would be intriguing to investigate if the cancer TFs were involved in differential regulating mechanisms. As a tentative exploration, we tested TED and TFactS on two independent lung cancer expression datasets. Both algorithms could reproduce a substantial fraction of differential regulators given their respective favorable regulatory networks, and many of them were previously deemed as cancer genes or related to (lung) cancer in wet-lab studies. Therefore, as we demonstrated in the tentative trial with two lung cancer datasets, differential regulator identification as a specific approach to this long-term goal may contribute to the fight against cancer using TFs.

Although we attempted to perform our evaluation study to as many as the available network-based algorithms for differential regulator identification, there are still some algorithms not covered in our survey. For instance, TFrank [33], a method with characteristic design in network features, was not included, because it exists as a web service. We could not easily adapt it to batch-executable codes. Another limitation is that we have not extended the algorithm testing to other types of molecular regulators, such as microRNA. One major obstacle to this attempt is the lack of simulation models based on microRNA regulatory networks. We used the general term of "differential regulators" in this study because the principles of the surveyed algorithms can be applied to microRNA and other types of regulation as well. However, whether these algorithms, which have been demonstrated for TFs, can indeed effectively identify differential microRNAs has not been tested yet. Lastly the simulated differential regulation in this work is likely too simple to reflect the complex deregulation in a real physiological condition. In reality, a regulator may lose part of its initiated regulations but not all. It may even gain new regulations beyond the loss of wild-type regulations. Sometimes a mutant regulator may cause differential regulation of other regulators through competitive binding or protein-protein interaction, as demonstrated in the mechanisms of mutant p53 [8]. Therefore, in real cases, the differential regulation program is far more complex than conceptualized here in our simulation framework. Hence, wherever possible, the identification of differential regulators should not rely solely on transcriptome data, but ideally should extend to evidence from other sources, such as mutations inferred from DNA-Seq and differential binding from ChIP-Seq. Nevertheless, our evaluation in this study represents the first systematical assessment of the main network-based identification algorithms for differential regulators from transcriptome data.

This work was partially supported by US National Institutes of Health (R01LM011177, R03CA167695, P30CA68485, P50CA095103 and P50CA098131), Ingram Professorship Funds (to Zhao ZhongMing), and The Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation (to Zhao ZhongMing). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

1 Zeng L, Yu J, Huang T, Jia H, Dong Q, He F, Yuan W, Qin L, Li Y, Xie L. Differential combinatorial regulatory network analysis related to venous metastasis of hepatocellular carcinoma. BMC Genomics, 2012, 13(Suppl 8): S14

2 Tu K, Yu H, Hua YJ, Li YY, Liu L, Xie L, Li YX. Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms. Nucleic Acids Res, 2009, 37: 5969–5980

3 Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol, 2012, 8: 565

4 de la Fuente A. From ‘differential expression’ to ‘differential networking’—identification of dysfunctional regulatory networks in diseases. Trends Genet, 2010, 26: 326–333

5 Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, Kumar V. Subspace differential coexpression analysis: problem definition and a general approach. Pac Symp Biocomput, 2010, 15: 145–156

6 Zhang B, Li H, Riggins RB, Zhan M, Xuan J, Zhang Z, Hoffman EP, Clarke R, Wang Y. Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics, 2009, 25: 526–532

7 Zhou Q, Hong Y, Zhan Q, Shen Y, Liu Z. Role for Kruppel-like factor 4 in determining the outcome of p53 response to DNA damage. Cancer Res, 2009, 69: 8284–8292

8 Strano S, Dell’Orso S, Di Agostino S, Fontemaggi G, Sacchi A, Blandino G. Mutant p53: an oncogenic transcription factor. Oncogene, 2007, 26: 2212–2219

9 Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, Demoulin JB. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Res, 2010, 38: e120

10 Reverter A, Hudson NJ, Nagaraj SH, Perez-Enciso M, Dalrymple BP. Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. Bioinformatics, 2010, 26: 896–904

11 Hudson NJ, Reverter A, Dalrymple BP. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. PLoS Comput Biol, 2009, 5: e1000382

12 Yang J, Yu H, Liu BH, Zhao Z, Liu L, Ma LX, Li YX, Li YY. DCGL v2.0: An R package for unveiling differential regulation from differential co-expression. PLoS One, 2013, 8: e79729

13 Huang CL, Lamb J, Chindelevitch L, Kostrowicki J, Guinney J, Delisi C, Ziemek D. Correlation set analysis: detecting active regulators in disease populations using prior causal knowledge. BMC Bioinformatics, 2012, 13: 46

14 Yu H, Liu BH, Ye ZQ, Li C, Li YX, Li YY. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinformatics, 2011, 12: 315

15 Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, De Moor B, Marchal K. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics, 2006, 7: 43

16 Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics, 2011, 27: 2263–2270

17 Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, Stojanov P, McKenna A, Lander ES, Gabriel S, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature, 2012, 489: 519–525

18 Mitra R, Edmonds MD, Sun J, Zhao M, Yu H, Eischen CM, Zhao Z. Reproducible combinatorial regulatory networks elucidate novel oncogenic microRNAs in non-small cell lung cancer. RNA, 2014, 20: 1356–1368

19 Nymark P, Guled M, Borze I, Faisal A, Lahti L, Salmenkivi K, Kettunen E, Anttila S, Knuutila S. Integrative analysis of microRNA, mRNA and aCGH data reveals asbestos- and histology-related changes in lung cancer. Genes Chromosomes Cancer, 2011, 50: 585–597

20 Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, eds. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer, 2005

21 Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res, 2006, 34: D108–110

22 Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res, 2003, 31: 3576–3579

23 Sun J, Gong X, Purow B, Zhao Z. Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma. PLoS Comput Biol, 2012, 8: e1002488

24 Rimmele P, Komatsu J, Hupe P, Roulin C, Barillot E, Dutreix M, Conseiller E, Bensimon A, Moreau-Gachelin F, Guillouf C. Spi-1/PU.1 oncogene accelerates DNA replication fork elongation and promotes genetic instability in the absence of DNA breakage. Cancer Res, 2010, 70: 6757–6766

25 Romero OA, Torres-Diz M, Pros E, Savola S, Gomez A, Moran S, Saez C, Iwakawa R, Villanueva A, Montuenga LM, Kohno T, Yokota J, Sanchez-Cespedes M. MAX inactivation in small cell lung cancer disrupts MYC-SWI/SNF programs and is synthetic lethal with BRG1. Cancer Discov, 2014, 4: 292–303

26 Chen L, Wei T, Si X, Wang Q, Li Y, Leng Y, Deng A, Chen J, Wang G, Zhu S, Kang J. Lysine acetyltransferase GCN5 potentiates the growth of non-small cell lung cancer via promotion of E2F1, cyclin D1, and cyclin E1 expression. J Biol Chem, 2013, 288: 14510–14521

27 An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database (Oxford), 2014, 2014: bau015

28 Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, Garriga-Canut M, Serrano L. Evolvability and hierarchy in rewired bacterial gene networks. Nature, 2008, 452: 840–845

29 Bhattacharyya M, Bandyopadhyay S. Studying the differential co-expression of microRNAs reveals significant role of white matter in early Alzheimer's progression. Mol Biosyst, 2013, 9: 457–466

30 Libermann TA, Zerbini LF. Targeting transcription factors for cancer gene therapy. Curr Gene Ther, 2006, 6: 17–33

31 Darnell JE Jr. Transcription factors as targets for cancer therapy. Nat

Rev Cancer, 2002, 2: 740–749

32 Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, Carninci P, Daub CO, Forrest AR, Gough J, Grimmond S, Han JH, Hashimoto T, Hide W, Hofmann O, Kamburov A, Kaur M, Kawaji H, Kubosaki A, Lassmann T, van Nimwegen E, MacPherson CR, Ogawa C, Radovanovic A, Schwartz A, Teasdale RD, Tegnér J, Lenhard B, Teichmann SA, Arakawa T, Ninomiya N, Murakami K, Tagami M, Fukuda S, Imamura K, Kai C, Ishihara R, Kitazume Y, Kawai J, Hume DA, Ideker T, Hayashizaki Y. An atlas of combinatorial transcriptional regulation in mouse and man. Cell, 2010, 140: 744–752

33 Goncalves JP, Francisco AP, Mira NP, Teixeira MC, Sa-Correia I, Oliveira AL, Madeira SC. TFRank: network-based prioritization of regulatory associations underlying transcriptional responses. Bioinformatics, 2011, 27: 3149–3157