Sikdar and Datta *BMC Bioinformatics* (2017) 18:79
DOI 10.1186/s12859-017-1499-x

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A novel statistical approach for identification of the master regulator transcription factor

Sinjini Sikdar and Susmita Datta*

Abstract

Background: Transcription factors are known to play key roles in carcinogenesis and therefore, are gaining popularity as potential therapeutic targets in drug development. A 'master regulator' transcription factor often appears to control most of the regulatory activities of the other transcription factors and the associated genes. This 'master regulator' transcription factor is at the top of the hierarchy of the transcriptomic regulation. Therefore, it is important to identify and target the master regulator transcription factor for proper understanding of the associated disease process and identifying the best therapeutic option.

Methods: We present a novel two-step computational approach for identification of master regulator transcription factor in a genome. At the first step of our method we test whether there exists any master regulator transcription factor in the system. We evaluate the concordance of two ranked lists of transcription factors using a statistical measure. In case the concordance measure is statistically significant, we conclude that there is a master regulator. At the second step, our method identifies the master regulator transcription factor, if there exists one.

Results: In the simulation scenario, our method performs reasonably well in validating the existence of a master regulator when the number of subjects in each treatment group is reasonably large. In application to two real datasets, our method ensures the existence of master regulators and identifies biologically meaningful master regulators. An R code for implementing our method in a sample test data can be found in http://www.somnathdatta.org/software.

Conclusion: We have developed a screening method of identifying the 'master regulator' transcription factor just using only the gene expression data. Understanding the regulatory structure and finding the master regulator help narrowing the search space for identifying biomarkers for complex diseases such as cancer. In addition to identifying the master regulator our method provides an overview of the regulatory structure of the transcription factors which control the global gene expression profiles and consequently the cell functioning.

Keywords: Master regulator, Transcription factor, Differential connectivity, Regulation, Concordance

Background

Through several scientific findings, it has been suggested that cancer is mainly caused by the mutations in certain genes. So, for effective treatment of cancer, identification of these mutated genes (oncogenes) is very essential. Detailed studies of different cancer datasets often lead to identification of several oncogenes which are directly or indirectly responsible for development and progression of cancer. It is a very challenging task to target and individually study all of these oncogenes as they are large in number. One way to overcome this approach is to group the proteins and genes belonging to the same pathway [1]. These genes and their corresponding pathways are known to form networks that control various cellular functions, and there has been sufficient interest in analyzing such pathway based networks. However recent findings suggest that most oncogenes and tumor suppressor genes encode "transcription factors", deregulations of which play key roles in carcinogenesis [2, 3]. Majority of the cancer signaling pathways seem to

* Correspondence: susmita.datta@ufl.edu
Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA

converge to these sets of transcription factors, and these transcription factors lead to tumor development, progression and cancer metastasis through the controlling of the gene expression patterns [2, 3]. As suggested by [2, 4], three main groups of transcription factors, which have been identified for cancer, are the steroid receptors (e.g. estrogen receptors in breast cancer, androgen receptors in prostate cancer), resident nuclear proteins activated by kinase cascades, and the latent cytoplasmic factors (from the STAT protein family members). Apart from these, the ETS protein family members have also been identified as potential cancer transcription factors for their emerging roles in human cancer [5]. It has been shown that direct suppression of these transcription factor expressions can lead to significant antitumor responses with minimal side-effects, and targeting these transcription factors in tumor-related immune cells can help in recovering from tumor immunoresistance [6]. As a result of these features of the transcription factors, in addition to the facts that they are much smaller in number than the oncogenes and have well-regulated expression and activities, transcription factors are gaining popularity as potential therapeutic targets in anti-cancer drug development [3, 4, 7, 8].

In the recent past, many studies have identified a transcription factor or a group of transcription factors as the driving force behind the development of a biological or disease process [9–12]. In order to facilitate such detection there have been attempts to develop statistical methods for accurate identification of transcription factors that regulate large number of genes. To this end most of these methods have been attempted for identification of transcription factors and transcription factor binding sites in cell cycle of yeast and similar organisms using multiple data sources [13–18], while a few of these methods have been applied for human cell as well [19]. In addition, there have been efforts in developing statistical tools for identifying a cluster of transcription factors that cooperatively regulates a large number of genes and the associated disease process [20, 21]. Methods have also been developed for identifying differentially regulated gene sets by integrating regulatory networks of transcription factors and gene expression data [22]. Also, transcription factor activities have been estimated through their effect on target genes [23]. The importance of transcription factor regulation is also evident from the fact that methods have been developed for identifying coordinately activated functional modules from gene expression data. These methods assume that the transcription factor regulated target genes are differentially expressed from non-target genes in the same functional module [24]. In fact there have been several studies for identifying transcription factors under the assumption that co-expression indicates co-regulation [25–27]. The main idea behind such transcription factor regulation is that genes regulated by such transcription factors should have, on an average, significantly different expression levels during one or more cell cycle phases [28]. Besides, there have been studies for identifying groups of important transcription factors through integration of different genomic and epigenomic features [29] and integration of transcriptional and protein interaction networks [30]. Most of these recent methods, including that of [31] and [32], have been directed towards identification of a group of candidate driver transcription factors. Despite the fact that in most cases there are a group of transcription factors that regulate the oncogenes and hence the disease process, it has been seen that there is a hierarchical structure in the regulatory activities of these transcription factors where a 'master regulator' transcription factor often appears to control most of the regulatory activities of the other transcription factors and the associated genes [33–35]. According to the definition provided in [34] the "master regulator" transcription factor is at the top of a regulatory hierarchy and must not be under the regulatory influence of any other gene or transcription factor. We use this definition and attempt to finding the "master regulator" transcription factor. This master regulator transcription factor can be targeted for proper understanding of the associated disease process and can be used as a biomarker.

In current literature there is a lack of appropriate statistical methods which use a single data source for accurate identification of such a master regulator among a set of identified transcription factors. In this article, we develop a novel two-step statistical approach to test for the existence of a master regulator transcription factor and for subsequent identification of the master regulator, if it exists, from gene expression data alone. The rest of the article is organized as follows: In the Methods section, we develop our test statistic and describe its underlying motivations for identifying the master regulator transcription factor. In the Results section, we describe a set of simulation experiments to evaluate the performance of our method and also apply our proposed method to two real datasets. Finally, we conclude with a discussion on the utility of our proposed method.

Methods

We first discuss the biological considerations that motivated the development of our test statistic. We then provide the methodology to formulate the test statistic and use it for the identification of the regulatory circuit of the transcription factors and genes. Finally, we identify the master transcription regulator at the top of the regulation hierarchy.

Biological considerations

Important biological processes can have multiple layers of regulation and control. A transcription factor is known to control not only genes but also other transcription factors. As discussed before in the Background section, usually there is a hierarchical structure in the regulation of the transcription factors so that the master regulator controls most of the regulatory activities of the other transcription factors and the associated genes. In this article, we aim to identify the master regulator transcription factor which is at the top of the hierarchy for better understanding of the associated disease process. A toy example is shown in Fig. 1 which shows the possible regulatory network across a set of genes and transcription factors in a genome.

In Fig. 1, TF_1, TF_2, $TF_3$and TF_4 denote the transcription factors and g_1, g_2, …, g_{12} denote the set of genes. Suppose TF_1 directly regulates five of the genes, which are g_1, g_2, g_3, g_4 and g_5, and also all the other three transcription factors, TF_2, TF_3 and TF_4. The transcription factor TF_2 regulates the genes g_6, g_7 and g_8. Similarly, the transcription factor TF_3 regulates the genes g_9 and g_{10} and finally, the transcription factor TF_4 regulates the genes g_{11} and g_{12}. In this example, there exists a hierarchical structure with three layers. We have TF_1 at the top of the hierarchical structure as it directly or indirectly regulates the other transcription factors and the genes. So, TF_1 is considered to be the first layer of the hierarchy. Now, TF_1 directly regulates the other transcription factors, TF_2, TF_3 and TF_4 and the genes g_1, g_2, …, g_5. So, TF_2, TF_3 and TF_4 and g_1, g_2, …, g_5 are considered to be at the second layer of the hierarchy. TF_1 regulates the genes g_6, g_7, …, g_{12} indirectly through the transcription factors TF_2, TF_3 and TF_4. Thus, the genes g_6, g_7, …, g_{12} form the third layer of the hierarchy. In this example, TF_1 directly or indirectly regulates all the layers of the hierarchy and is not under the regulatory influence of any other gene or transcription factor. Therefore, according to the definition, TF_1 can be considered as the master regulator transcription factor.

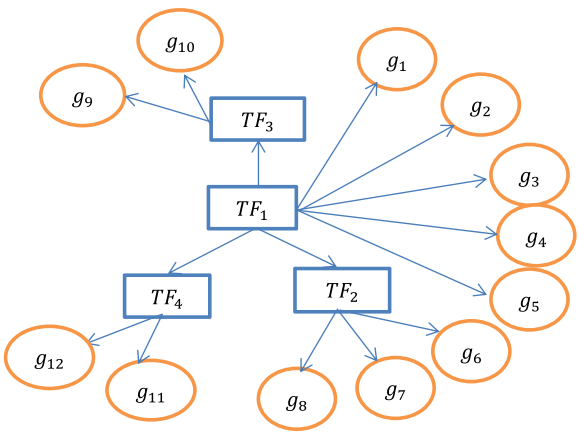

Fig. 1 A toy example showing possible regulatory network across a set of genes and transcription factors

Here, in this article, we attempt to develop a test that can check if there exists any transcription factor that acts as a master regulator in a genome, and identify such a master regulator if present. The details of our proposed method are given in the next section.

Identification of the Master Regulator through a Hypothesis Testing Framework

Let M denote the total number of transcription factors present in a genome. Let the transcription factors be denoted by TF_1, TF_2, …, TF_M. Let the genes, which are not transcription factors, be denoted by g_1, g_2, …, g_N, where N denotes the total number of such genes. So, in total, we have expression data on $M + N$ genes. Let us assume that there are two groups of subjects, for example, the case group (the disease group) and the control group (non-disease group). Let there be r_1 subjects in the case group and r_2 subjects in the control group. So, in other words, we have two groups of subjects with expression levels for $M + N$ features in each group. It is well known that the genes including the transcription factors are expressed differently in the two groups. Additionally, they are connected with one another differently in the networks of the two groups. There are methods to reverse engineer the networks of genes with association measures such as correlations, partial correlations and partial least squares regression scores [36, 37]. These inter genomic connectivity are different in two groups and can be detected by statistical methods such as Differential Network Analysis (DNA) [38].

Since, it is believed that the master regulator maximally controls the other transcription factors as well as the associated genes; it is important to find the regulatory network among the transcription factors and also the degree of regulation of all the transcription factors on the genes. We first measure the degree of regulation of the transcription factors on the genes. The degree of regulation of a transcription factor on the genes is measured by the change in connectivity of the genes it regulates in the two networks. In other words, we find how the connectivity of a transcription factor with the genes differs between the subjects in case and control groups. For this, we estimate the change in connectivity of a transcription factor with the genes in the two groups of samples using connectivity scores of the given transcription factor with all the genes in the case group with that in the control group. The difference in connectivity is measured using the following statistic [38]:

$$x_i = \frac{1}{N}\sum_{g\in\mathcal{G}} \left| s^{case}_{TF_i,g} - s^{ctl}_{TF_i,g} \right| \quad ; \quad i = 1, 2, ..., M , \quad (1)$$

where $\mathcal{G}$ denotes the set of all the genes, i.e., the cardinality of $\mathcal{G}$ is N. Here, $s^{case}_{TF_i,g}$ and $s^{ctl}_{TF_i,g}$ are the connectivity scores between the transcription factor TF_i, $i = 1, 2, ..., M$, and the gene g in the case and control groups, respectively. There are several choices for connectivity scores such as Pearson's correlation scores, partial correlation scores, partial least square based association scores. In this article, we use the Pearson's correlation scores as connectivity scores. So, here, the $x_i's$ give us an idea about the magnitude of the differential regulation of the transcription factors on the genes between the case and control groups.

Next, we find the regulatory structure among the transcription factors. For this, we measure the association between each pair of transcription factors using the Pearson's correlation coefficient scores between them. For each pair (j, k), let y_{jk} denote the absolute value of the Pearson's correlation coefficient score between the transcription factors TF_j and TF_k; $j, k = 1, 2, ..., M$, where $y_{jj} = 1$. Note that this calculation is done by pooling the data from both the groups.

At this stage, for a transcription factor TF_j , $j = 1, 2, ..., M$, we have two measures: a measure of the differential regulation of TF_j on the genes (given by x_j); and a measure of the association of TF_j with all the transcription factors (y_{jk}, $k = 1, 2, ..., M$).

We argue that the degree of change in connectivity of the genes in the two networks is controlled by the transcription factors which are correlated amongst themselves in a hierarchical manner. That is, the hierarchical regulation pattern (as measured by the rank order) among the M transcription factors is the same with the differential connectivity of genes in the two groups that they control. In other words, the rank order of the amount of differential connectivity of a transcription factor with other genes it controls is in line (e.g., *concordant*) with its ordered connectivity with the master regulator. Therefore, we consider two ranked lists. One that ranks the transcription factors by the amount of differential connectivity of the genes it controls and another that puts the master regulator in the first position and ranks the remaining transcription factors by their correlation with the master regulator. We evaluate the concordance of these two sets of ranks using a statistical measure which is described in next paragraph. Since we do not know a priori the identity of the master regulator/s, we maximize this measure of concordance over the set of all transcription factors in candidacy for playing the role of the master regulator. In case the maximal *concordance* is statistically significant, we conclude that there is a master regulator. In addition, we declare the transcription factor for which this concordance measure is maximal amongst all transcription factors to be the master regulator.

We construct a concordance statistic K_j for each transcription factor TF_j that is in candidacy for the master regulator ; $1 \le j \le M$, in the following way:

1) We calculate the Kendall's rank correlation coefficient test statistic given by (2) below based on the pairs of data (x_1, y_{j1}), (x_2, y_{j2}), ..., (x_j, y_{jj}), ..., (x_M, y_{jM}). Note that x_i denotes the average difference in connectivity of transcription factor TF_i between the two groups, and y_{ji} is the absolute correlation of transcription factor TF_i with transcription factor TF_j. This test statistic K_j below conveys whether the differential connectivity of the genes with the transcription factor TF_j in the two experimental groups is concordant with the correlations of the transcription factor TF_j with all other transcription factors. In other words, K_j measures whether the differential connectivity is concordant with the hierarchical regulation of the transcription factors amongst themselves. The Kendall's rank correlation coefficient test statistic for the transcription factor TF_j is given as:

$$K_j = \frac{n_{c,j} - n_{d,j}}{n_0} , \quad (2)$$

where, $n_{c,j}$ = number of concordant pairs in the above paired list,
$n_{d,j}$ = number of discordant pairs in the above paired list,
$n_0 = \frac{M(M-1)}{2}$ = Total number of such paired observations for TF_j.

This statistic can be used to test the null hypothesis that the two sets of ranks produced by differential connectivity x and the correlations with TF_j are non concordant versus the alternative hypothesis that they are concordant.

2) We repeat step 1) for all such transcription factors, so that we have a concordance test statistic for each of the transcription factors which is a potential master regulator.

We believe that the master regulator has the maximum measure of concordance, among all the transcription factors. Since we do not know the identity of the master regulator, we maximize the measure of concordance, given by K_j, over the set of all transcription factors. So, we define K as the maximum of the statistics given

in (2) over all the transcription factors TF_js that are in candidacy for the role of the master regulator, i.e.

$$K = \max_j K_j \tag{3}$$

Thus, statistically significant large values of K would indicate the existence of a master regulator.

Significance of K can be assessed by a bootstrap (re-sampling) based procedure as the sampling distribution of K is not tractable. This will calculate the p-value or the observed level of significance of the value of test statistic K calculated in (3). We draw B bootstrap samples from the original sample each of size $r_1 + r_2$ and consider the first r_1 samples as the case group and the remaining r_2 samples as the control group. We compute the test statistic value for each bootstrap sample. Let K_b denotes the value of our test statistic for the b^{th} bootstrap sample, where $1 \le b \le B$. In order to estimate the p-value, we calculate the proportion of times the test statistic values based on the bootstrap samples exceed the test statistic value obtained from the original sample, i.e.,

$$\text{p-value} = \frac{\sum_{b=1}^{B} I(K_b > K)}{B} \quad , \tag{4}$$

If the p-value obtained from (4) is low then the test is significant and we conclude that there exists a master regulator in the system.

In case we conclude that there exists a master regulator the transcription factor T is claimed to be the master regulator if it has the maximum value of the statistic given in (2), i.e.

$$T = \arg\max_j K_j \tag{5}$$

We evaluate the performance of our master regulator identification procedure using a simulation experiment in the next section. A sample test data where our method can be implemented is available in the Additional file 1, while an associated R code for its implementation can be found in http://www.somnathdatta.org/software.

Results

Simulation

In order to evaluate the performance of our proposed method, we generate synthetic datasets of gene expressions of the case and control groups with the different regulation schemes of the transcription factors. The simulation scheme consists of the following steps:

Data Generation

We consider M transcription factors $TF_1, TF_2, \ldots, TF_M$ and N genes $g_1, g_2, \ldots, g_N$, as described before in the Methods section. Also, let there be r_1 subjects in the case group and r_2 subjects in the control group. The gene expression data for the two groups of subjects are generated as given below. Note that, the choices of all the design parameters considered below are given in later sections depending on whether we are simulating under the null or under the alternative.

1. We assume that (log-transformed) expression values for TF_1 follows a normal distribution with mean μ and variance 1 i.e. $N(\mu, 1)$ in the case group, and $N(\vartheta, 1)$ in the control group.
   We also generate M independent random variables V_i from $N(0, 1)$; $i = 1, 2, \ldots, M$, that are also independent of TF_1.
2. We want to generate all the transcription factors in such a way that there exists a hierarchical regulatory pattern among them. In other words, we want to generate the remaining $M - 1$ transcription factors in such a way that $Corr(TF_j, TF_k) > Corr(TF_j, TF_l)$ $(j = 1, 2, \ldots, M;\ k,\ l = j + 1, \ldots, M;\ k < l)$, where $Corr(TF_j, TF_k)$ denotes the correlation between the transcription factors TF_j and TF_k. One way of achieving this is to simulate the remaining $M - 1$ transcription factors TF_i ; $i \neq 1$ as follows:

$$TF_i = \frac{\rho_i TF_1 + V_i}{\sqrt{1 + \rho_i^2}} \quad ; \quad i \neq 1$$

where, $\rho_i{}'$ s are decreasing in i, $i \neq 1$.

In this case, the correlation structures among all the transcription factors are given by:

$$Corr(TF_1, TF_i) = \frac{\rho_i}{\sqrt{1 + \rho_i^2}} \ ; \ i \neq 1$$

and $Corr(TF_j, TF_k) = \frac{\rho_j \rho_k}{\sqrt{1+\rho_j^2}\sqrt{1+\rho_k^2}} \quad ; \quad j, k \neq 1 \ ; j \neq k.$

3. The next step is to generate the genes. We assume that each of the transcription factors TF_i ; $i = 1, 2, \ldots, M$, regulates m_i genes. Here, $N = m_1 + m_2 + \ldots + m_M$. The genes, $g_1, g_2, \ldots, g_{m_1}$, which are directly regulated by TF_1 alone, are generated as given below:

$$g_j = \begin{cases} TF_1\gamma_1 + \epsilon_j, & \text{for case group} \\ TF_1\gamma_2 + \epsilon_j & \text{for control group} \end{cases} \qquad j = 1, 2, \ldots, m_1$$

where, ϵ_j and $\epsilon_j{}'$ are independent and identically distributed (i.i.d) as $N(0, 1)$, and γ_1 and γ_2 are real numbers.

Here, the correlation between the transcription factor TF_1 and the genes g_k , $k = 1, 2, \ldots, m_1$ is given by

$$Corr(TF_1, g_k) = \begin{cases} \frac{\gamma_1}{\sqrt{1+\gamma_1^2}} & \text{for case group} \\ \frac{\gamma_2}{\sqrt{1+\gamma_2^2}} & \text{for control group} \end{cases} \quad k = 1, 2, \ldots, m_1$$

The genes, regulated by the remaining $M - 1$ transcription factors TF_i; $i \neq 1$, are generated as follows:

$$g_j = \begin{cases} V_i r_{1i} + \epsilon_j & \text{for case group} \\ V_i r_{2i} + \epsilon_j' & \text{for control group} \end{cases} \quad j = m_{i-1} + 1, \ldots, m_i$$

where, ϵ_j and ϵ_j' are i.i.d $N(0, 1)$ and r_{1i} and r_{2i} are real numbers, $i \neq 1$.
In this case, the correlation between a transcription factor TF_i; $i \neq 1$ and the genes regulated by that transcription factor is given by

$$Corr(TF_i, g_k) = \begin{cases} \frac{r_{1i}}{\sqrt{1+\rho_i^2}\sqrt{1+r_{1i}^2}} & \text{for case group} \\ \frac{r_{2i}}{\sqrt{1+\rho_i^2}\sqrt{1+r_{2i}^2}} & \text{for control group} \end{cases} \quad k = m_{i-1} + 1, \ldots, m_i$$

Also, the correlations between a transcription factor TF_i; $i \neq 1$ and the genes which are not regulated by that transcription factors are zero i.e. $Corr(TF_i, g_k) = 0$ for $k \neq m_{i-1} + 1, \ldots, m_i$. Furthermore, $Corr(TF_1, g_k) = 0$; $k \neq 1, 2, \ldots, m_1$. We calculate the size and power of our test in the following sections.

Size of the Test

Recall that, the null hypothesis of interest is that the rank order of the transcription factors based on their differential connectivity with the genes is not statistically concordant with their rank order based on their correlations with the master regulator. So, the null situation can be created by assuming that there exists a hierarchical regulatory pattern among the transcription factors but there is no differential regulation of the genes in the two experimental groups due to the transcription factors. Hence, there is no such master regulator.

In order to follow the null hypothesis in the simulation setup, we assume ρ_is to be decreasing in i , $i = 2, 3, \ldots, M$ and choose $\gamma_1 = \gamma_2$ and $r_{1i} = r_{2i}$, $i = 2, 3, \ldots, M$. The decreasing nature of ρ_is ensures that there exists a hierarchical regulatory pattern among the transcription factors. $\gamma_1 = \gamma_2$ and $r_{1i} = r_{2i}$, $i = 2, 3, \ldots, M$ ensure that the associations of the transcription factors with the genes remain the same in the two groups i.e. there is no differential connectivities of the transcription factors with the genes between the two groups. We generate r_1 samples for the case group and r_2 samples for the control group using the above described scheme. We calculate the value of our test statistic, denoted by K, using Eq. (3) and find its p-value as described in the Methods section.

In order to find the size of the test, we use Monte-Carlo method. We repeat the whole process 1000 times and therefore, get 1000 p-values using Eq. (4). Let the p-value for the i^{th} Monte-Carlo iteration be denoted as p_i , $i = 1, 2, \ldots, 1000$. The size for the test is given by:

$$\text{Size} = \frac{\sum_{i=1}^{1000} I(p_i < 0.05)}{1000} \tag{6}$$

In particular, we consider the following choices of the parameters for calculating the size of the test:

- $M = 10$, $N = 105$, $r_1 = r_2 = 500$, $B = 500$
- $\mu = 50$, $\vartheta = 5$
- $m_1 = 30$, $m_2 = m_3 = \ldots = m_7 = 10$, $m_8 = m_9 = m_{10} = 5$
- $\rho = (\rho_2, \ldots, \rho_{10}) = (0.95, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$
- $\gamma_1 = \gamma_2 = 0.5$
- $r_{12} = r_{22} = 0.45$ and $r_{1i} = r_{2i} = r_{1(i-1)} - 0.05$ for $i = 3, \ldots, 10$.

For the above choices of the parameters, the empirical size of the test came out to be 0.032 which is close to the nominal size of 0.05.

Power of the Test

To calculate the power of our test, we generate a data under the alternative hypothesis H_1. Here the alternative hypothesis is that the rank order of the transcription factors based on their differential connectivity with the genes is concordant with their rank order based on their correlations with the master regulator. So, we generate the data in such a way that TF_1 acts as the master regulator, that is, the connectivity of TF_1 with other transcription factors are most concordant with the differential connectivity of the genes with the transcription factors. We set $\gamma_1 > \gamma_2$ and $r_{1i} > r_{2i}$; $i = 2, 3, \ldots, 10$. Here $\gamma_1 > \gamma_2$ ensures that the connectivity (associations) of TF_1 with the genes, regulated by it, are greater in case group than that in the control group. Similarly, $r_{1i} > r_{2i}$ ensures that the connectivity of TF_i with the genes, regulated by it, are greater in case group than that in the control group, $i = 2, 3, \ldots, 10$. Also, we assume ρ_is to be decreasing in i , $i = 2, 3, \ldots, M$, so that there is a hierarchical regulatory structure among the transcription factors, TF_1 being at the top of the hierarchy. We follow the same steps in calculating the p-value as we did for size calculation in the previous section. We consider the same choices for M, N, r_1, r_2, B, μ, ϑ, ρ and m_i ; $i = 1, 2, \ldots, 10$ as we consider for size calculation. In particular, we choose $\gamma_2 = 0.5$; $r_{12} = 0.45$ and $r_{1i} = r_{1(i-1)}$

– 0.05 for $i = 3, \ldots, 10$. We choose $r_{2i} = (1 - \delta)r_{1i}$, $i = 2, 3, \ldots, 10$ where $0 \le \delta \le 1$. These choices of r_{2i} , $i = 2, 3, \ldots, 10$; ensure that increase in the value of δ also increase the difference between r_{1i} and r_{2i} , $i = 2, 3, \ldots, 10$. In other words, the differential regulations of the transcription factors on the genes between the two groups increase as δ increases.

For the choice of γ_1, we consider the following relation: $\gamma_1 = \gamma_2 + \delta(r_{12} - r_{22})$, which implies $\gamma_1 = \gamma_2 + \delta^2 r_{12}$, $0 \le \delta \le 1$. This choice of γ_1 ensures that increase in the value of δ also increase the difference between γ_1 and γ_2. We draw the power curve for different choices of δ, as shown in Fig. 2.

From Fig. 2, we see that the power steadily increases as the differential connectivity (regulated by δ) of the genes with the transcription factors between the two groups increase. The power curve starts from 3.2% at $\delta = 0$ (no difference in the connectivity of the genes with the transcription factors in the two groups) and reaches its maximum of 100% at $\delta = 1$ (maximum difference in the connectivity of the genes with the transcription factors in the two groups). The power reaches over 80% with a moderate choice of $\delta = 0.6$. Therefore, we can say that our proposed method is a valid test (e.g., size ≤ 0.05) that performs reasonably well (power reaching 100%) in identifying a significant concordance in the differential connectivity of the genes with the transcription factors and the connectivity of a transcription factor with master regulator, if one exists.

We also consider several other choices of the sample sizes in each of the two groups (case and control), and calculate the size and draw the power curves for each of the following choices of the sample sizes: $r_1 = 100$, $r_2 = 70$; $r_1 = 50$, $r_2 = 40$; and $r_1 = r_2 = 50$, representing reduced sample sizes and unequal sample sizes in each treatment group. Overall, from our analyses with different choices of sample size, we find that the power of our test is increasing with increase in the sample size as well as an increase in the differential connectivity of the genes with the transcription factors in the two groups. Details of the variation of the power with sample size can be found in Additional file 2 which shows the power curves for each of the above choices of the sample sizes with different choices of δ, $0 \le \delta \le 1$.

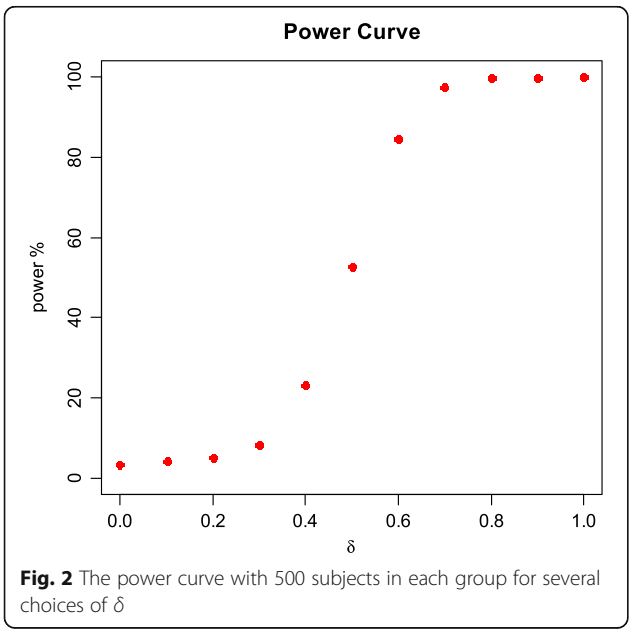

Fig. 2 The power curve with 500 subjects in each group for several choices of δ

In order to check the performance of our test in case there are more than one master regulator transcription factors, we have also studied a simulated scenario where there are two independent master regulator transcription factors regulating two independent sets of genes through transcriptional regulatory networks. Additional file 3 shows the power performance of our test in the presence of two independent master regulators in the system. In this case, too, our test has substantial power performance, similar to the simulated settings of a single master regulator transcription factor. Note that, here we have considered one of the many possible simulation settings. However, our method can also be generalized for several other complicated scenarios.

Application to Real Datasets

Prostate Cancer Data

We apply our test statistic, proposed in Methods section, to a human Illumina expression array dataset GSE18684 of androgen regulated gene expression in the LNCaP prostate cancer cell line [39]. It is believed that androgens and the androgen receptor (AR) play significant roles in prostate cancer cell proliferation and invasion. So, this study was conducted by [39] with an aim to identify the androgen receptor (AR) regulated genes. The LNCaP cells were treated with androgen (R1881) or with vehicle (ethanol) control. There are 10 control and 35 androgen treated samples with expression levels for 17182 probes in the dataset. We identify the set of probes which are differentially expressed in the two groups (androgen treated and vehicle control) using the "limma" package in Bioconductor [40]. After adjusting for false discovery rate (FDR) at 5% significance level, 6054 probes are differentially expressed in the two groups, out of which 542 are transcription factors.

Now, we test whether there exists any master regulator in the above mentioned dataset. For this, we compute the value of our test statistic for this dataset using Eq. (3), which turns out to be 0.49 with a bootstrap based p-value of 0.006. Since the p-value is highly significant we conclude that there exists a master regulator transcription factor in the system which is controlling all the

other transcription factors and the genes. In order to find the master regulator, we use Eq. (5) as given in the Methods section. For this study, the two transcription factors "PEG3" and "ARNT2" have the same value of the test statistic given in Eq. (3). So, we conclude that these two transcription factors maximally control all the other transcription factors and consequently control the connectivity of the genes differently in the two groups. Additionally, the Pearson's correlation coefficient value between the two transcription factors "PEG3" and "ARNT2" is 0.8. This high value of the correlation suggests that these two transcription factors are approximately at the same level of transcriptional regulatory hierarchy. Therefore, it can be concluded that both of them are the master regulators. Among these two master regulators, "PEG3" has often been linked to the development of prostate cancer. It is believed that deregulation of WNT/ β catenin pathway contributes to prostate cancer progression [41–46], and according to [47], inhibition of the transcription factor "PEG3" can lead to enhanced β catenin expression and proliferation in human glioma stem cells. This function of the transcription factor "PEG3" is relevant to prostate cancer [48]. Further, the expression of the transcription factor "PEG3" is known to be associated with the processes of cancer aggressiveness and angiogenesis [49]. The results from our analysis show consistency with these known roles of "PEG3" in prostate cancer and demonstrate the utility of our proposed method to identify the master regulator transcription factor. Besides, the transcription factor "ARNT2" is known to have a critical role in human renal tract development, thereby showing congenital abnormalities of the kidneys and urinary tract [50]. "ARNT2" is also known to have significant roles in many cancers like NSCLC [51], breast cancer [52], etc.

Table 1 shows the list of top 10 transcription factors which are highly correlated with the two potential master regulators "PEG3" and "ARNT2".

From Table 1, it can be seen that the transcription factors "WWC1", "NCOA7", "TSHZ3" and "TCFL5" are highly correlated with both the master regulators. Among these, "WWC1" is known to be associated with prostate cancer. The expression of "WWC1" is influenced by AR signaling and is increased in prostate cancer [53]. The transcription factor "NCOA7" is known to affect AR-mediated transcription [54]. The expression of "TSHZ3" is known to be downregulated in prostate cancer [55]. FOXD4L1 is also implicated in many cancers [56].

Table 1 Top 10 transcription factors having high correlations with the master-regulators in the prostate cancer data

Master Regulators	Top 10 transcription factors correlated with the master regulator
PEG3	WWC1, FOXD4L1, NCOA7, TSHZ3, CTBP1, TCFL5, LHX2, ARID5B, CDCA7L, MAK,
ARNT2	MSRB2, TULP4, TSHZ3, TCFL5, SNAPC5, TFDP1, WWC1, CITED4, NCOA7, GRAMD4

Colorectal Cancer Data

We apply our method to another human microarray dataset GSE4107. This study was conducted by [57] with an aim to identify differentially expressed genes in early onset colorectal cancer (CRC). RNA samples are extracted from colonic mucosa of patients as well as healthy controls and analyzed using GeneChip U133-Plus 2.0 Array. There are 22 subjects involved in the study which included 12 patients and 10 controls. All the patients and the controls in the data are young Chinese who are aged 50 years or less. There are expression levels for 54,675 genes for all the patients in the dataset. We first filter the data in order to find the set of differentially expressed genes between the case and the control groups. For this purpose we use the "limma" package in Bioconductor [40]. After adjusting for FDR at 10% significance level, the number of differentially expressed genes turns out to be 5192, among which 266 are transcription factors.

Next, we apply our method to the filtered dataset. We first test whether there exists a master regulator in the data. The value of our test statistic, given in Eq. (3), is 0.38 for this dataset with a p-value of 0.04 for the bootstrap based test. Since, the p-value is small enough to make the test significant, we conclude that there exists a master regulator in the data. We identify the master regulator using Eq. (5), given in the Methods section. The master regulator in this data is the transcription factor "NFKB2". Hence, we conclude that the transcription factor "NFKB2" maximally controls all the transcription factors and the genes in the data.

The transcription factor "NFKB2" is a subunit of the transcription factor nuclear factor-kappa-B (NFKB). "NFKB" transcription factors are known to be the key regulators of innate immune responses, inflammation, and cell survival [58, 59]. Also, "NFKB" activation has been frequently associated with tumor growth in leukemias and lymphomas, as well as prostate, pancreatic and colorectal cancers [60–62]. It has been widely suggested that "NFKB" activation plays a leading role in regulation of target genes that promote cell proliferation, anti-apoptosis, regulate immune and inflammatory response, and results in pathogenesis of various cancers [59, 63–67]. Further, it has been shown that constitutive activation of "NFKB" instigates strong resistance to chemotherapy and radiotherapy [67], while molecular targeted therapy against "NFKB" activation is believed to be effective in colorectal carcinomas with constitutive "NFKB" activation [59]. According to [66], "NFKB" may contribute to the

promotion of the ongoing inflammatory process in the gut mucosa resulting in the progression of colitis associated colorectal cancer. Besides, it is believed that "NFKB" activation is involved in development of not only colitis-associated cancer, but also sporadic colorectal cancer [68].

From our data, we find that the master regulator "NFKB2" is maximally (negatively) correlated with the transcription factor "PPARGC1A (PGC-1alpha)" with an overall correlation value of -0.76. The correlation of "NFKB2" and "PPARGC1A" is -0.72 in the patients group whereas it is -0.39 in the control group. It is known that "NFKB" directly repress the activity of "PPARGC1A" in cardiac cells. This leads to the increase in glucose oxidation which is observed during pro-inflammatory state [69].

Discussion

In this article, we present a novel approach to identify a master regulator transcription factor in a system using only the gene expression profiles of the patients. We consider a simulation setting which validates our approach with a reasonable power in detecting the existence of a master regulator. We have also checked the power of our test in the presence of two independent master regulator transcription factors in the simulation setup. We apply our approach to two human microarray datasets and detect the existence of master regulators in those. In order to check the robustness of our method in experiments not typically falling under the 'case-control' category, we have applied our method to an additional dataset, namely, Glioblastoma (GBM) TCGA RNA-seq data [70]. Here we compare the two types of GBM tumors: Mesenchymal and Classical. Our method concludes the existence of a master regulator transcription factor (PPRC1) between the two types of GBM tumors (Mesenchymal and Classical) with a p-value of 0.08 (marginally significant).

Our method is aimed to identify a single master regulator, as opposed to identifying a group of transcription factors associated with the disease process as in the case of other existing methods. The method can identify multiple master regulator transcription factors if they are individually at the top of hierarchy of the transcription regulation. This is advantageous in anti-cancer drug development processes which initially target the most potential transcription factor associated with the disease and can be used as a potential biomarker. However, there is a scope of further improvement of our proposed method by incorporating important platforms like ChIP-Seq data. From simulation settings, we see that the performance of our method gets better with the increase in the number of patients in each group. So, our method is expected to be more efficient when there is sufficiently large number (around 100) of patients in each group while it may not be very efficient in case the sample size is very small. Although both the data analyzed in this article have much lower number of subjects in each group, our test was still successful in identifying master-regulator transcription factors from the data. One important assumption of our method is that the ranking of the transcription factors on the basis of their differential connectivity of the genes between two experimental conditions is concordant with the hierarchical order of their own regulation. The fulfilment of the above mentioned condition is a key indicator to the existence of a master regulator transcription factor and its subsequent detection through our method. However, it may be possible that in certain situations, although there exists a master regulator transcription factor, there is no such clear cut concordance between it's regulation on other transcription factors and differential connectivity with the other genes. In such a case, our method may not perform well.

Conclusion

We have developed a method of identifying the 'master regulator' transcription factor using only the gene expression data. This is advantageous in terms of narrowing down the search space for potential candidate transcription factor biomarkers that can be targeted for drug development of complex diseases. Also, the fact that our method uses only a single data source, e.g. gene expression data, for accurately identifying the master regulator transcription factor makes it very useful in case there is limitation in data sources and data from multiple platforms are not available. In addition to identifying the master regulator our method provides an overview of how the transcription factors regulate the global gene expression profiles and consequently the cell functioning. Additionally, with our method, one can identify many other transcription factors involved in the regulatory roles by reporting the hierarchy amongst them using the rankings of the test statistics values. Overall, we believe that our method will give new insight for efficient identification of potential disease biomarker and therapeutic target in drug development processes.

Additional files

Additional file 1: Test Dataset. This file contains an example test dataset where our method can be implemented. This simulated data contains 10 transcription factors, namely TF_1, TF_2, …, TF_{10} along with 105 genes that were regulated by these transcription factors. Among the transcription factors, TF_1 was generated to play the role of the master regulator. (CSV 1382 kb)

Additional file 2: Figure. Plot of the power curves for different choices of the sample sizes with several choices of δ, using simulated datasets. (DOCX 19 kb)

Additional file 3: Figure. Plot showing the power performance of our test in presence of two independent master regulators with varying δ, using simulated datasets. (DOCX 16 kb)

Abbreviations
ChIP-Seq: ChIP Sequencing; FDR: False discovery rate; GBM: Glioblastoma; NSCLC: Non-small cell lung cancer; RNA-Seq: RNA Sequencing; TCGA: The Cancer Genome Atlas

Acknowledgements
We acknowledge Prof. Somnath Datta for insightful comments on the methodology.

Funding
This work was supported by the research funding from the NIH grant CA 170091-01A1 to
Susmita Datta. The motivation for this analysis of high-dimensional data and the protected time for the PI were generated from this grant.

Availability of data and materials
The androgen regulated Illumina expression array dataset (GSE18684) is available from the NCBI website (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18684).
The colorectal cancer microarray dataset (GSE4107) is available from the NCBI website (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4107).

Authors' contributions
SD designed the study and developed the method. SS simulated the datasets, wrote the codes and analyzed the datasets. SD and SS wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Received: 31 August 2016 Accepted: 27 January 2017
Published online: 02 February 2017

References
1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
2. Libermann TA, Zerbini LF. Targeting transcription factors for cancer gene therapy. Curr Gene Ther. 2006;6(1):17–33.
3. Zerbini LF. Oncogenic Transcription Factors: Target Genes. In: eLS. Wiley; 2007. http://www.els.net/WileyCDA/ElsArticle/refId-a0006049.html. doi:10.1002/9780470015902.a0006049.
4. Darnell Jr JE. Transcription factors as targets for cancer therapy. Nat Rev Cancer. 2002;2(10):740–9.
5. Seth RB, Sun L, Ea CK, Chen ZJ. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. Cell. 2005;122(5):669–82.
6. Yeh JE, Toniolo PA, Frank DA. Targeting transcription factors: promising new strategies for cancer therapy. Curr Opin Oncol. 2013;25(6):652–8.
7. Redmond AM, Carroll JS. Defining and targeting transcription factors in cancer. Genome Biol. 2009;10(7):311.
8. Bhagwat AS, Vakoc CR. Targeting transcription factors in cancer. Trends Cancer. 2015;1(1):53–65.
9. Tovar H, García-Herrera R, Espinal-Enríquez J, Hernández-Lemus E. Transcriptional master regulator analysis in breast cancer genetic networks. Comput Biol Chem. 2015;59:67–77.
10. Bae T, Rho K, Choi JW, Horimoto K, Kim W, Kim S. Identification of upstream regulators for prognostic expression signature genes in colorectal cancer. BMC Syst Biol. 2013;7:86.
11. Sawle AD, Kebschull M, Demmer RT, Papapanou PN. Identification of master regulator genes in human periodontitis. J Dent Res. 2016;95(9):1010–7.
12. Gubelmann C, Schwalie PC, Raghav SK, Röder E, Delessa T, Kiehlmann E, Waszak SM, Corsinotti A, Udin G, Holcombe W, et al. Identification of the transcription factor ZEB1 as a central component of the adipogenic gene regulatory network. Elife. 2014;3:e03346.
13. Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. Proc Int Conf Intell Syst Mol Biol. 2000;8:344–54.
14. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003;301(5629):102–5.
15. Tsai HK, Lu HH, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. Proc Natl Acad Sci U S A. 2005;102(38):13532–7.
16. Tsai HK, Huang GT, Chou MY, Lu HH, Li WH. Method for identifying transcription factor binding sites in yeast. Bioinformatics. 2006;22(14):1675–81.
17. Cheng C, Li LM. Systematic identification of cell cycle regulated transcription factors from microarray time series data. BMC Genomics. 2008;9:116.
18. Wu WS, Li WH. Systematic identification of yeast cell cycle transcription factors using multiple data sources. BMC Bioinformatics. 2008;9:522.
19. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res. 2005;33(10):3154–64.
20. Banerjee N, Zhang MQ. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. Nucleic Acids Res. 2003;31(23):7024–31.
21. Hu H. An efficient algorithm to identify coordinately activated transcription factors. Genomics. 2010;95(3):143–50.
22. Ma S, Jiang T, Jiang R. Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data. Bioinformatics. 2015;31(4):563–71.
23. Schacht T, Oswald M, Eils R, Eichmüller SB, König R. Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics. 2014;30(17):i401–7.
24. Petti AA, Church GM. A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. Genome Res. 2005;15(9):1298–306.
25. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol. 1998;16(10):939–45.
26. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics. 2004;5(1):18.
27. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22(3):281–5.
28. Das D, Banerjee N, Zhang MQ. Interacting models of cooperative gene regulation. Proc Natl Acad Sci U S A. 2004;101(46):16234–9.
29. Gevaert O, Plevritis S. Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features. Pac Symp Biocomput. 2013; 123-34.
30. Padi M, Quackenbush J. Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators. BMC Syst Biol. 2015;9:80.
31. Piao G, Saito S, Sun Y, Liu ZP, Wang Y, Han X, Wu J, Zhou H, Chen L, Horimoto K. A computational procedure for identifying master regulator candidates: a case study on diabetes progression in Goto-Kakizaki rats. BMC Syst Biol. 2012;6(1):S2.
32. Saito S, Zhou X, Bae T, Kim S, Horimoto K. Identification of master regulator candidates in conjunction with network screening and inference. Int J Data Min Bioinform. 2013;8(3):366–80.
33. Yang J, Mani SA, Donaher JL, Ramaswamy S, Itzykson RA, Come C, Savagner P, Gitelman I, Richardson A, Weinberg RA. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. Cell. 2004;117(7):927–39.
34. Chan SS, Kyba M. What is a master regulator? J Stem Cell Res Ther. 2013;3:114.
35. De D, Jeong MH, Leem YE, Svergun DI, Wemmer DE, Kang JS, Kim KK, Kim SH. Inhibition of master transcription factors in pluripotent cells induces early stage differentiation. Proc Natl Acad Sci U S A. 2014;111(5):1778-–83.

36. Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiaah PR, editor. Multivariate Analysis. New York: Academic; 1966. p. 391–420.
37. Datta S. Exploring relationships in gene expressions: a partial least squares approach. Gene Expr. 2001;9(6):249–55.
38. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics. 2010;11:95.
39. Massie CE, Lynch A, Ramos-Montoya A, Boren J, Stark R, Fazli L, Warren A, Scott H, Madhu B, Sharma N, et al. The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. EMBO J. 2011;30(13):2719–33.
40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
41. Cronauer MV, Schulz WA, Ackermann R, Burchardt M. Effects of WNT/beta-catenin pathway activation on signaling through T-cell factor and androgen receptor in prostate cancer cell lines. Int J Oncol. 2005;26(4):1033–40.
42. Li H, Kim JH, Koh SS, Stallcup MR. Synergistic effects of coactivators GRIP1 and beta-catenin on gene activation: cross-talk between androgen receptor and Wnt signaling pathways. J Biol Chem. 2004;279(6):4212–20.
43. Song LN, Gelmann EP. Interaction of beta-catenin and TIF2/GRIP1 in transcriptional activation by the androgen receptor. J Biol Chem. 2005;280(45):37853–67.
44. Song LN, Herrell R, Byers S, Shah S, Wilson EM, Gelmann EP. Beta-catenin binds to the activation function 2 region of the androgen receptor and modulates the effects of the N-terminal domain and TIF2 on ligand-dependent transcription. Mol Cell Biol. 2003;23(5):1674–87.
45. Terry S, Yang X, Chen MW, Vacherot F, Buttyan R. Multifaceted interaction between the androgen and Wnt signaling pathways and the implication for prostate cancer. J Cell Biochem. 2006;99(2):402–10.
46. Yang X, Chen MW, Terry S, Vacherot F, Bemis DL, Capodice J, Kitajewski J, de la Taille A, Benson MC, Guo Y, et al. Complex regulation of human androgen receptor expression by Wnt signaling in prostate cancer cells. Oncogene. 2006;25(24):3436–44.
47. Jiang X, Yu Y, Yang HW, Agar NY, Frado L, Johnson MD. The imprinted gene PEG3 inhibits Wnt signaling and regulates glioma growth. J Biol Chem. 2010;285(11):8472–80.
48. Ribarska T, Bastian KM, Koch A, Schulz WA. Specific changes in the expression of imprinted genes in prostate cancer–implications for cancer progression and epigenetic regulation. Asian J Androl. 2012;14(3):436–50.
49. Su ZZ, Goldstein NI, Jiang H, Wang MN, Duigou GJ, Young CS, Fisher PB. PEG-3, a nontransforming cancer progression gene, is a positive regulator of cancer aggressiveness and angiogenesis. Proc Natl Acad Sci U S A. 1999;96(26):15115–20.
50. Webb EA, AlMutair A, Kelberman D, Bacchelli C, Chanudet E, Lescai F, Andoniadou CL, Banyan A, Alsawaid A, Alrifai MT, et al. ARNT2 mutation causes hypopituitarism, post-natal microcephaly, visual and renal anomalies. Brain. 2013;136(10):3096–105.
51. Yang B, Yang E, Liao H, Wang Z, Den Z, Ren H. ARNT2 is downregulated and serves as a potential tumor suppressor gene in non-small cell lung cancer. Tumour Biol. 2015;36(3):2111–9.
52. Martinez V, Kennedy S, Doolan P, Gammell P, Joyce H, Kenny E, Prakash Mehta J, Ryan E, O'Connor R, Crown J, et al. Drug metabolism-related genes as potential biomarkers: analysis of expression in normal and tumour breast tissue. Breast Cancer Res Treat. 2008;110(3):521–30.
53. Stauffer S, Chen X, Zhang L, Chen Y, Dong J. KIBRA promotes prostate cancer cell proliferation and motility. FEBS J. 2016;283(10):1800–11.
54. Heemers HV, Regan KM, Schmidt LJ, Anderson SK, Ballman KV, Tindall DJ. Androgen modulation of coregulator expression in prostate cancer cells. Mol Endocrinol. 2009;23(4):572–83.
55. Yamamoto M, Cid E, Bru S, Yamamoto F. Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. PLoS One. 2011;6(3):e17149.
56. Jackson BC, Carpenter C, Nebert DW, Vasiliou V. Update of human and mouse forkhead box (FOX) gene families. Hum Genomics. 2010;4(5):345–52.
57. Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin Cancer Res. 2007;13(4):1107–14.
58. Karin M, Lin A. NF-kappaB at the crossroads of life and death. Nat Immunol. 2002;3(3):221–7.
59. Sakamoto K, Maeda S, Hikiba Y, Nakagawa H, Hayakawa Y, Shibata W, Yanai A, Ogura K, Omata M. Constitutive NF-kappaB activation in colorectal carcinoma plays a key role in angiogenesis, promoting tumor growth. Clin Cancer Res. 2009;15(7):2248–58.
60. Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet Jr RJ, Sledge Jr GW. Constitutive activation of NF-kB during progression of breast cancer to hormone-independent growth. Mol Cell Biol. 1997;17(7):3629–39.
61. Rayet B, Gelinas C. Aberrant rel/nfkb genes and activity in human cancer. Oncogene. 1999;18(49):6938–47.
62. Tai DI, Tsai SL, Chang YH, Huang SN, Chen TC, Chang KS, Liaw YF. Constitutive activation of nuclear factor kB in hepatocellular carcinoma. Cancer. 2000;89(11):2274–81.
63. Luo JL, Kamata H, Karin M. IKK/NF-kappaB signaling: balancing life and death–a new approach to cancer therapy. J Clin Invest. 2005;115(10):2625–32.
64. Karin M, Ben-Neriah Y. Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity. Annu Rev Immunol. 2000;18:621–63.
65. Hayden MS, Ghosh S. Signaling to NF-kappaB. Genes Dev. 2004;18(18):2195–224.
66. Wang S, Liu Z, Wang L, Zhang X. NF-kappaB signaling pathway, inflammation and colorectal cancer. Cell Mol Immunol. 2009;6(5):327–34.
67. Wang CY, Cusack Jr JC, Liu R, Baldwin Jr AS. Control of inducible chemoresistance: enhanced antitumor therapy through increased apoptosis by inhibition of NF-kappaB. Nat Med. 1999;5(4):412–7.
68. Sakamoto K, Maeda S. Targeting NF-kappaB for colorectal cancer. Expert Opin Ther Targets. 2010;14(6):593–601.
69. Alvarez-Guardia D, Palomer X, Coll T, Davidson MM, Chan TO, Feldman AM, Laguna JC, Vázquez-Carrera M. The p65 subunit of NF-kappaB binds to PGC-1alpha, linking inflammation and metabolic disturbances in cardiac cells. Cardiovasc Res. 2010;87(3):449–58.
70. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155(2):462–77.