BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 4 2015, pages 563–571
doi:10.1093/bioinformatics/btu672

Systems biology

Advance Access publication October 15, 2014

Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data

Shining Ma[1], Tao Jiang[1,2,*] and Rui Jiang[1,3,*]

[1]MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China, [2]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA, [3]Department of Statistics, Stanford University, Stanford, CA 94305, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Although many gene set analysis methods have been proposed to explore associations between a phenotype and a group of genes sharing common biological functions or involved in the same biological process, the underlying biological mechanisms of identified gene sets are typically unexplained.

Results: We propose a method called *Differential Regulation-based enrichment Analysis for GENe sets* (DRAGEN) to identify gene sets in which a significant proportion of genes have their transcriptional regulatory patterns changed in a perturbed phenotype. We conduct comprehensive simulation studies to demonstrate the capability of our method in identifying differentially regulated gene sets. We further apply our method to three human microarray expression datasets, two with hormone treated and control samples and one concerning different cell cycle phases. Results indicate that the capability of DRAGEN in identifying phenotype-associated gene sets is significantly superior to those of four existing methods for analyzing differentially expressed gene sets. We conclude that the proposed differential regulation enrichment analysis method, though exploratory in nature, complements the existing gene set analysis methods and provides a promising new direction for the interpretation of gene expression data.

Availability and implementation: The program of DRAGEN is freely available at http://bioinfo.au.tsinghua.edu.cn/dragen/.

Contact: ruijiang@tsinghua.edu.cn or jiang@cs.ucr.edu

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

Received on April 14, 2014; revised on September 28, 2014; accepted on October 9, 2014

1 INTRODUCTION

With the maturation of high-throughput techniques such as DNA microarrays (Duggan *et al.*, 1999) and RNA sequencing (RNA-seq) (Wang *et al.*, 2009) in gene expression profiling on a genome-wide scale, interpretations of the vast amount of gene expression data have now become a standard task in biological studies (Cookson *et al.*, 2009; Dixon *et al.*, 2007). As a typical example, genes differentially expressed in normal and cancer samples are often identified as associated with the cancer of interest, thereby providing clues for finding biomarkers and drug targets for the diagnosis and treatment of the cancer (Nevins and Potti, 2007).

Nevertheless, a biological function is usually raised by the collaborative effects of multiple gene products instead of individual proteins. With this understanding, methods have been proposed to find or rank sets of differentially expressed genes, forming the research direction of gene set analysis (Goeman and Buhlmann, 2007). To mention a few, Subramanian *et al.* (2005) proposed a method called GSEA (*Gene Set Enrichment Analysis*) to detect pre-defined gene sets that are enriched with differentially expressed genes based on a weighted Kolmogorov–Smirnov statistic. Breitling *et al.* (2004) put forward a method called *Over-Representation Analysis* (ORA) to determine which part of a gene set is differentially expressed by making use of the hypergeometrical distribution. Barry *et al.* (2005) relied on a permutation approach to prioritize gene sets that are composed of a large proportion of marginally differential expressed genes and named their method *Significance Analysis of Function and Expression* (SAFE).

Although the earlier methods are capable of providing a list of differentially expressed gene sets from a collection of pre-defined candidates, two important questions remain largely untouched in the gene set analysis literature. First, how to take regulatory relationships between genes in a candidate gene set into consideration? Second, what is the reason behind the observation that a gene set is differentially expressed? To answer the first question, several network-based methods have been proposed in the literature. For example, netGSA adopted a mixed linear model to test the significance of gene sets (Shojaie and Michailidis, 2009). *Gene Graph Enrichment Analysis* (GGEA) detected consistency between transcriptional regulatory relationships and gene expression levels by using a Petri net with fuzzy logic (Geistlinger *et al.*, 2011). *PAthway Recognition Algorithm using Data Integration on Genomic Models* (PARADIGM) incorporated pathway interactions as well as many types of omics data to infer activities of pathways in patients (Vaske *et al.*, 2010). *Differential Expression Analysis for Pathways* (DEAP) was capable of detecting paths in input pathways with the most differential expression (Haynes *et al.*, 2013). These methods addressed the first question by taking regulatory relationships (or interactions) between genes into consideration, but left the second question unanswered.

The expression of a gene is a complicated process regulated by several factors, among which transcription factors (TFs) play a

*To whom correspondence should be addressed.

crucial role. Therefore, the change of the expression level of a gene in an abnormal phenotype may mainly be attributed to the alteration of the gene's transcriptional regulatory pattern (Cheng *et al.*, 2012). It has been reported that such alterations in transcriptional regulation, usually initiated by mechanisms such as genetic and epigenetic modifications, are often observed among abnormal phenotypes including the vast categories of cancers. For example, mutations occurring in TFs can lead to retargeting on the promoters of specific oncogenes in breast cancer (Zuo *et al.*, 2007). Mutations occurring in the promoter region of the telomerase enzyme-coding gene can drive the development of melanoma (Patton and Harrington, 2013). More generally, TF-encoding genes with chromosomal translocations have been distinguished, resulting in mistargeting and transcriptional dysregulation among cancers (Patel *et al.*, 2012). Furthermore, by investigating transcriptional regulatory networks constructed for different cell lines with ChIP-seq experiments, it has been reported that a large proportion of regulatory interactions vary across different phenotypes (Neph *et al.*, 2012). However, such significant alterations in transcriptional regulation, which typically occur in specific cellular environments in response to stimuli accompanying abnormal phenotypes, can hardly be identified and explained by the traditional analysis of differential expression patterns of the genes involved. Indeed, in order to capture subtle changes of transcriptional regulatory patterns in abnormal phenotypes, a novel computational approach beyond the traditional identification of differentially expressed genes or gene sets is required.

With the above motivation, we propose to connect the observed expression levels of a set of genes with their underlying transcriptional regulatory patterns and further detect alterations in the regulatory relationships from gene expression data. Specifically, we achieve this goal by putting forward an approach called *Differential Regulation based enrichment Analysis for GENe sets* (DRAGEN) that integrates gene expression data and a transcriptional regulatory network to identify differentially regulated gene sets. DRAGEN uses a linear regression model to explain the relationship between the expression level of a target gene and that of a TF, thereby connecting regulatory patterns between TFs and target genes to their expression profiles. With this rigorous statistical model, DRAGEN converts the problem of detecting differentially regulated gene sets to a series of hypothesis testing problems followed by the fusion of multiple *P*-values to obtain the statistical significance of each candidate gene set, thereby enabling the ranking of the candidates. We conduct comprehensive simulation experiments to demonstrate the superior performance of DRAGEN, apply it to an estradiol-treated MCF-7 cell line, an androgen-treated LNCaP cell line and a HeLa cell line at different cell cycle phases, and show the capability of this method in finding gene sets associated with relevant phenotypes.

2 METHODS

2.1 Principles of DRAGEN

The basic premise of this study is that the transcriptional regulatory relationship between a TF and a target gene regulated by the TF may exhibit different patterns in different phenotypic states, and such an alteration in the regulatory relationship can be detected using gene expression profiles of the TF and the target gene. Moreover, for a set of genes that are involved in the biological process underlying a phenotype,

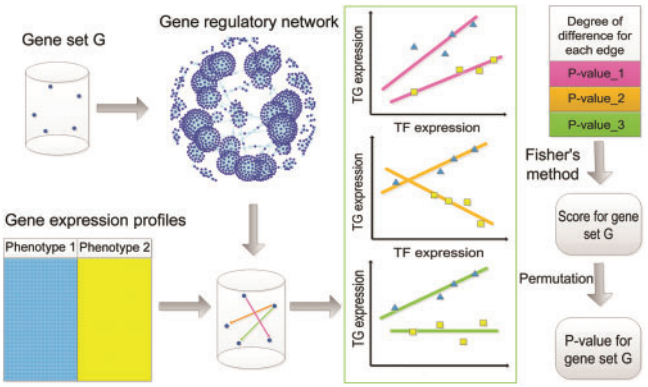

Fig. 1. The workflow of DRAGEN

regulatory patterns of a significant proportion of the genes would be altered when a perturbation is applied to the biological process. Therefore, given a pre-defined gene set, one can quantify the strength of association between the gene set and a phenotype by introducing a perturbation to the phenotype (e.g. treat samples with drugs) and then measuring whether the regulatory patterns in the set of genes change significantly. Besides, given a phenotype and a collection of candidate gene sets (e.g. pathways), one can identify gene sets that are most likely related to the phenotype by ranking the candidate gene sets according to their strength of association with the phenotype. We refer to such studies based on the detection of alterations in regulatory relationships as differential regulation enrichment analysis.

On the basis of the above rationale, we propose an approach called DRAGEN to detect alterations in regulatory patterns in a perturbed phenotype. As illustrated in Figure 1, DRAGEN takes as input three data sources: a transcriptional regulatory network where each edge indicates the regulatory relationship between a regulatory gene and its target, a collection of pre-defined candidate gene sets and gene expression profiles for the normal and perturbed phenotypes. With these input, DRAGEN performs differential regulation enrichment analysis and produces a ranked list of the candidate gene sets as the output. More specifically, for each gene set, DRAGEN carries out the following four steps to calculate a score that indicates the statistical significance of differential regulation in the gene set. First, the gene set is mapped onto the transcriptional regulatory network, resulting in a sub-network corresponding to the gene set. In this procedure, we by default require the inclusion of both the TFs and targets in a gene set. Alternatively, one can relax this constraint and require a gene set to include only target genes. Second, each edge in the sub-network is fitted into two linear regression models with the use of the gene expression data for the normal and perturbed phenotypes, respectively, and a *P*-value indicating the degree of difference between the two models is calculated. Third, *P*-values for all edges in the gene set are combined by using the Fisher's method, and a statistic for the gene set is obtained. Fourth, a permutation procedure is applied to calculate a *P*-value indicating the degree of differential regulation of the gene set. With *P*-values for candidate gene sets calculated by the above four steps, a multiple testing correction procedure is further adopted, and the final *P*-values for gene sets are obtained. The gene sets are then ranked in non-decreasing order according to their final *P*-values.

2.2 Regression model for detecting differential regulation

We represent a transcriptional regulatory network as a directed graph, in which a node denotes either a TF or a target gene, and a directed edge, pointing from a TF to a target gene, indicating the regulatory relationship between the TF and the target gene. Note that although detailed regulatory mechanisms such as activation or inhibition are desired, we found that annotations of such regulatory mechanisms are largely absent in

most databases [e.g. TRANSFAC (Matys *et al.*, 2003) and HTRIdb (Bovolenta *et al.*, 2012)]. Therefore, we do not use any annotations of regulatory mechanisms in our study. We define a candidate gene set typically as a collection of genes that are involved in the same functional pathway of some biological process. A gene expression profile consists of the expression levels of the genes in a sample. Particularly, we are interested in comparing the gene expression profiles of a normal phenotype and those of the counterpart (i.e. a perturbed phenotype).

Given a candidate gene set, we map genes in the set onto the transcriptional regulatory network to obtain a sub-network $G = \{V, E\}$, where E is a set of directed edges connecting genes and V the set of nodes (genes) connected by the edges in E. For each edge in the sub-network, we identify the TF and the target gene, extract their expression profiles for the normal and perturbed phenotypes and build a linear regression model for each phenotype to explain the expression levels of the target gene by those of the TF as

$$Y_j^{(i)} = \alpha^{(i)} + \beta^{(i)} \times X_j^{(i)} + \varepsilon_j^{(i)}, j = 1, \ldots, n_i, \quad (1)$$

where the superscript i indexes the normal ($i = 0$) or the perturbed ($i = 1$) phenotype. For phenotype i, $X_j^{(i)}$ and $Y_j^{(i)}$ denote the expression levels of the TF and the target gene for observation j, respectively, $\alpha^{(i)}$ and $\beta^{(i)}$ the regression intercept and slope, respectively, $\varepsilon_j^{(i)}$ a zero mean Gaussian noise with standard deviation $\sigma_j^{(i)}$ and n_i the sample size.

Clearly, in the above regression model, the slope reflects the regulatory capacity of the TF on the target gene. Therefore, we propose to detect the alteration of the regulatory relationship between the normal and perturbed phenotypes by testing the hypothesis

$$H_0 : \beta^{(0)} = \beta^{(1)} \text{ versus } H_1 : \beta^{(0)} \neq \beta^{(1)}. \quad (2)$$

It is evident that the maximum likelihood estimators of the parameters are $\hat{\beta}^{(i)} = S_{XY}^{(i)}/S_{XX}^{(i)}$ and $\hat{\alpha}^{(i)} = \overline{Y}^{(i)} - \hat{\beta}^{(i)}\overline{X}^{(i)}$, where $S_{XX}^{(i)} = \sum_{j=1}^{n_i}(X_j^{(i)} - \overline{X}^{(i)})^2$ and $S_{XY}^{(i)} = \sum_{j=1}^{n_i}(X_j^{(i)} - \overline{X}^{(i)})(Y_j^{(i)} - \overline{Y}^{(i)})$. Hence, the test statistic

$$T = \frac{\hat{\beta}^{(0)} - \hat{\beta}^{(1)}}{SE_{\hat{\beta}^{(0)} - \hat{\beta}^{(1)}}} \quad (3)$$

has a Student's t distribution with $n_0 + n_1 - 4$ degrees of freedom under the null hypothesis (Armitage *et al.*, 2002). Here, the denominator is the standard error of the difference between the two slopes, calculated as

$$SE_{\hat{\beta}^{(0)} - \hat{\beta}^{(1)}} = \sqrt{S_p^2\left(\frac{1}{S_{XX}^{(0)}} + \frac{1}{S_{XX}^{(1)}}\right)}, \quad (4)$$

with the assumption of equal variance ($\sigma^{(0)} = \sigma^{(1)}$) and the pooled estimate of the residual variance S_p^2 calculated by

$$S_p^2 = \frac{1}{n_0 + n_1 - 4}\sum_{i=0}^{1}\sum_{j=1}^{n_i}(Y_j^{(i)} - \hat{\alpha}^{(i)} - \hat{\beta}^{(i)}X_j^{(i)})^2. \quad (5)$$

The P-value of the proposed two-sided test can then be calculated as

$$p = 2P(T_{n_0+n_1-4} > |t|), \quad (6)$$

where $|t|$ is the absolute value of the realized test statistic.

The above single-TF model only considers the situation that a target gene is regulated by a single TF. To further formulate the regulation of multiple TFs on a target gene, we propose the following multiple-TF model that resorts to principal component analysis to decompose the combined effects of multiple TFs into independent factors (i.e. principal components).

Let $\mathbf{X} = \{x_{ij}\}_{n\times m}$ be expression levels of the m TFs across n conditions, where $n = n_1 + n_2$. We calculate the principal component decomposition of $\mathbf{X}$ as $\mathbf{c} = \mathbf{XV}$, where $\mathbf{V}$ is an m-by-m matrix whose columns are eigenvectors of the covariance matrix of $\mathbf{X}$. The kth column of $\mathbf{C}$ then includes an independent factor $\mathbf{C}_k = \{c_{jk}\}_{n\times 1}$, $j = 1, \ldots, n$, $k = 1, \ldots, m$. Then, we explain the expression level of the target gene by each of the independent factors separately using a regression model, as

$$Y_j^{(i)} = \alpha^{(i)} + \rho_k^{(i)} \times c_{jk}^{(i)} + \varepsilon_j^{(i)}, \quad j = 1, \ldots, n_i, \quad (7)$$

where $Y_j^{(i)}$ and $c_{jk}^{(i)}$ are the expression level of the target gene and the value of the kth independent factor, respectively, both for observation j and phenotype i. Finally, the hypothesis testing problem

$$H_0 : \rho^{(0)} = \rho^{(1)} \text{ versus } H_1 : \rho^{(0)} \neq \rho^{(1)} \quad (8)$$

can be conducted in a similar way to the single-TF model.

2.3 Calibration of statistical significance

With P-values for edges obtained, we calculate the statistical significance of a gene set using Fisher's combined probability test (Fisher, 1925). For the single-TF model, we calculate a statistic as

$$U = -2\sum_{e\in\mathbf{E}} \log p_e, \quad (9)$$

where $\mathbf{E}$ is the set of edges corresponding to a gene set, and p_e the P-value for edge $e \in \mathbf{E}$. It is evident that U has a χ^2 distribution with degrees of freedom being twice the number of edges in the gene set when all null hypotheses are true and all the P-values are independent. For the multiple-TF model, we calculate a statistic as

$$U = -2\sum_{g\in\mathbf{G}}\sum_{k\in\mathbf{C}(g)} w_k \log p_k, \quad (10)$$

where $\mathbf{G}$ is the set of target genes in a gene set, $\mathbf{C}(g)$ independent factors for gene g, p_k and w_k the P-value and weight of the kth factor, respectively. We set $w_k = \lambda_k/\Sigma_{i\in\mathbf{C}(g)}\lambda_i$ by default, with λ_k being the eigenvalue corresponding to the kth principle component.

For each gene set, we further adopt the following permutation procedure to obtain a P-value from the statistic U, based on the recommendation in literature that sampling individuals is superior to sampling genes (Goeman and Buhlmann, 2007).

Step 1: Calculate the realized value u_0 of the test statistic for the given gene set using the original gene expression profiles.

Step 2: Shuffle the phenotype labels in the expression profiles. Calculate a realized value of the test statistic using the permuted expression profiles.

Step 3: Repeat Step 2 a number of N times to obtain a series of realized statistic values $u_{1,\ldots,}u_N$. Count the number of times such values are greater than u_0. Divide this number by N to obtain a P-value for the gene set.

Note that such a permutation procedure also eliminates the influence of the size of the gene set. With P-values for candidate gene sets calculated, we further perform multiple testing corrections by controlling the false discovery rate (Benjamini and Hochberg, 1995).

3 RESULTS

3.1 Data sources

We used three datasets of *Escherichia coli* for simulation studies, including a regulatory network extracted from the RegulonDB database (Gama-Castro *et al.*, 2008), expression profiles of 10 antibiotic-treated samples and 10 untreated samples from the Many Microbe Microarrays Database (M3D) (Faith *et al.*, 2008), and a collection of gene sets extracted from the gene ontology (GO) (Ashburner *et al.*, 2000). We used eight datasets of human for real data analysis, including two regulatory networks

extracted from the HTRIdb database (Bovolenta *et al.*, 2012) and the ENCODE project (Gerstein *et al.*, 2012), three groups of expression profiles and two ChIP-seq datasets extracted from the Gene Expression Omnibus (GEO) database (Edgar *et al.*, 2002) and a collection of manually curated gene sets obtained from the Molecular Signatures Database (MSigDB) (Liberzon *et al.*, 2011). Detailed descriptions of these datasets are given in Supplementary Material (Section S1).

3.2 Simulation experiments

We first performed a simulation study to explore the capability of DRAGEN in detecting the alteration of the transcriptional regulatory relationship between a single TF and a single target gene, based on the *E. coli* data. Results, as shown in Supplementary Material (Section S2), clearly demonstrate the effectiveness of our method in detecting differential regulatory patterns in this situation.

We then conducted a series of simulation experiments to evaluate the performance of our method in finding differentially regulated gene sets. To generate gene sets, we mapped *E. coli* genes onto the biological process domain of GO, obtaining 3334 genes annotated with 1295 GO terms. In this procedure, we regarded a gene as annotated with a term if the gene is annotated with a child of the term in the directed acyclic graph structure of GO. Excluding terms appearing in the top two layers of the GO structure and focusing on those annotated with 5–500 genes, we extracted 318 terms. Further focusing on genes included in RegulonDB, we obtained a total of 59 gene sets, one corresponding to a GO term (Supplementary Table S1 and Fig. S3).

In an experiment, we selected at random a test case from the 59 gene sets and used the remaining 58 sets as templates to generate controls. For the test case, we embedded a differential regulation pattern (power law or logistic) in a proportion of the edges, where the proportion (ρ) and the strength of differential regulation (λ) were treated as parameters. To generate a control case, we replaced the genes in a template with those selected at random from the transcriptional regulatory network. We then applied DRAGEN to calculate *P*-values for simulated gene sets and further ranked the case against the controls.

Repeating this experiment 1000 times, we obtain the same number of ranking lists. We then derive a criterion to quantify the performance of our method. Using the *P*-value as a cut-off, we calculated the sensitivity as the fraction of test cases whose *P*-values were less than or equal to the cut-off value and the specificity as the fraction of negatives controls whose *p*-values were greater than the threshold. Varying the cut-off value, we obtained a series of sensitivities and specificities, and we were able to plot a receiver operating characteristic (ROC) curve and calculate an AUC score as the area under this curve.

The results of this simulation study are shown in Figure 2, from which we clearly see the effectiveness of our method in detecting differentially regulated gene sets. The values of λ indicate the degree of differential regulation for each embedded edge (Supplementary Material, Section S2.1). The larger the absolute value of $(\lambda - 1)$ is, the higher degree of differential regulation is. Taking the regulation loss case $(\lambda - 0)$ in the logistic model (Fig. 2e) as an example, when the proportion of embedded differentially regulated edges is 10%, the AUC score is 0.568. When

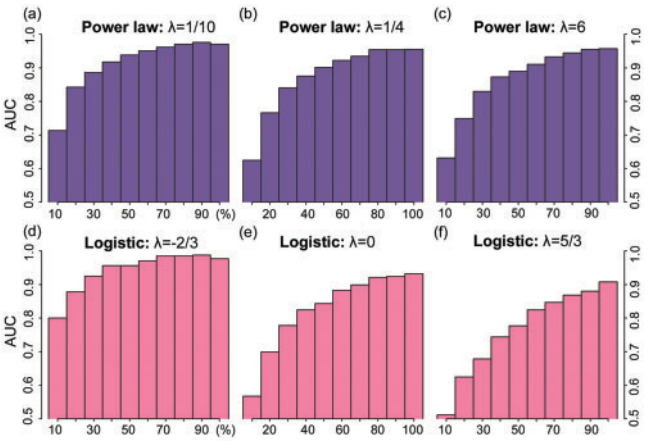

Fig. 2. Results of gene set simulation studies. Each row indicates a model (power law or logistic), and each column indicates a parameter value (λ). *X*-axes indicate proportions of embedded differentially regulated edges

the proportion of differentially regulated edges increases to 50%, the AUC score increases to 0.843. When the proportion of differentially regulated edges keeps increasing toward 100%, the AUC score tends to be stable at around 0.926. Similar phenomena are observed for the other values of $_{\lambda}$. These results suggest that even in the case that regulatory relationships are significantly nonlinear, our methods are still capable of identifying some differentially regulated gene sets that include a proportion of differentially regulated interactions.

3.3 Application to an estradiol treated MCF-7 cell line

We assessed the capability of DRAGEN in detecting differential regulatory events concerning estrogen receptor α (ESR1). From ChIP-seq data (GSE23701) of MCF-7 cells, we identified 1584 genes with different ESR1 binding events before and after the treatment of estradiol-17-β (E2). From expression data (GSE11352), we collected nine E2-treated and nine untreated samples. We then applied DRAGEN to quantify the strength of differential regulation for these genes and compared their *P*-values with those of interactions selected at random from HTRI. Results, as detailed in Supplementary Material (Section S3.1), show that *P*-values of the differentially regulated genes are significantly smaller than random cases (one-sided Wilcoxon test $P\text{-value} < 2.2 \times 10^{-16}$), indicating the effectiveness in detecting differential regulation.

We then evaluated the performance of our method in detecting differentially regulated gene sets that included ESR1. Mapping each of those gene sets from MsigDB category c2 onto the regulatory network and focusing on gene sets containing at least three edges, we obtained a total of 711 gene sets, among which 71 contained the ESR1 gene and its target genes. Treating these 71 gene sets as positive cases and the rest 640 as negative controls, we applied our method for these gene sets and further generated a ranking list. And then we plotted an ROC in Figure 3a.

We first compared the performance of DRAGEN with a simple correlation coefficient approach, and observed the better performance of our method (Supplementary Material,

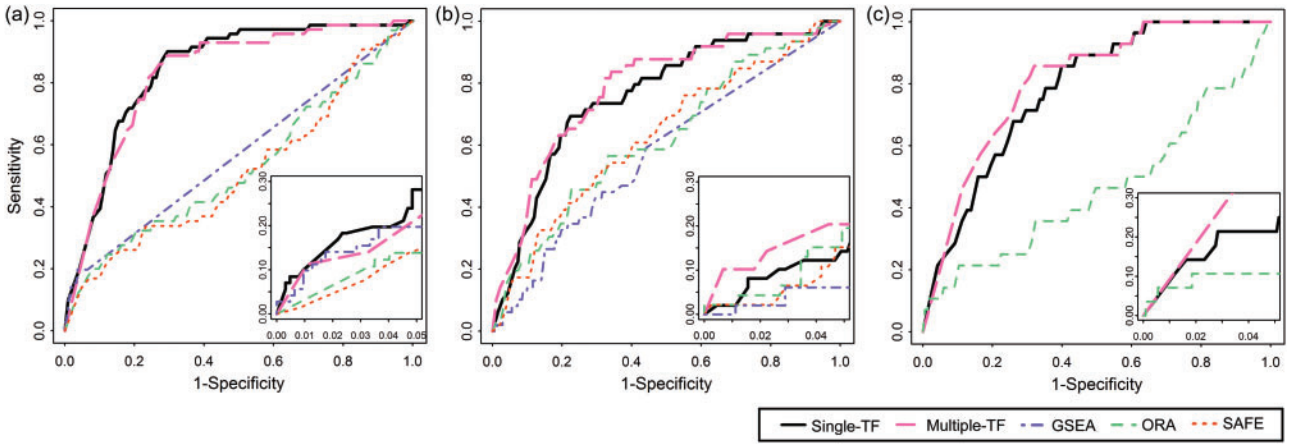

Fig. 3. ROC curves of different methods on (a) MCF-7, (b) LNCaP and (c) HeLa cells. In the zoomed-in plots, the ROCs of DRAGEN climb faster than those of the other methods toward the top-left corner, indicting its better performance

Section S3.7). We then compared our method with three existing methods (GSEA, ORA and SAFE) and plotted their ROCs in Figure 3a. Results show that the curves of DRAGEN climb toward the top-left corner of the zoomed-in plot rapidly, suggesting the high performance of our method. AUC scores (0.843 for the single-TF model of DRAGEN, 0.569, 0.527 and 0.512 for GSEA, ORA and SAFE, respectively) further support the effectiveness of our method.

We next compared our method with GGEA that also utilized network information. Since the annotations of the activation or inhibition mechanism of regulation required by this method were largely absent in databases (Bovolenta *et al.*, 2012; Matys *et al.*, 2003), we first used the untreated samples to infer the activation or inhibition mechanism for each edge in the regulatory network by fitting a linear regression model and hypothesis testing whether the regression slope was positive (for activation) or negative (for inhibition), respectively. We then applied GGEA to prioritize the 711 gene sets. Nevertheless, our results showed that all adjusted *P*-values of the gene sets turned out to be 1, making further analysis infeasible. We conjectured that this poor performance of GGEA was partly caused by the quality of annotations of regulatory mechanisms and hence left GGEA out of the plots.

We also compared our method with DEAP that integrated pathway and expression data. As this method needed pathway structures as input, the comparison was performed on a subset of the gene sets that represent pathways in KEGG (Ogata *et al.*, 1999) and Reactome (Croft *et al.*, 2011), and discussed separately in Supplementary Material (Section S3.5).

We further analyzed the ranking list of the 711 gene sets generated by DRAGEN and found that 20 out of the 71 positive cases were ranked among the top 50 (Table 1 and Supplementary Table S3). By comparison, GSEA, ORA and SAFE found 14, 11 and 11 positive cases among their top 50, respectively (Supplementary Tables S4–S6). The comparison among the top 10 (and top 20) gene sets of these methods and the enrichment significance by Fisher's exact test was in Supplementary Material, Supplementary Table S2, which clearly showed the better performance of DRAGEN. In hormone-dependent breast cancer epithelial cells such as MCF-7, ERα-mediated estrogen acts as a dominant stimulus for the proliferation and survival of the cells (Risbridger *et al.*, 2010). Particularly, MCF-7 cells stimulated with estrogen could progress to a more malignant stage (Levenson and Jordan 1997). With DRAGEN, four gene sets with both ESR1 and its target genes were ranked at the top (the first), and six positive gene sets were ranked among the top 10 (Supplementary Fig. S5). The connections between several top ranked gene sets and estrogen treatment on MCF-7 cells reported in the literature were discussed in Supplementary Material (Section S3.2).

3.4 Application to an androgen-treated LNCaP cell line

From ChIP-seq data (GSE28126) of an LNCaP cell line, we identified 237 genes with different androgen receptor (AR) binding events before and after the treatment of R1881 (Metribolone, a potent androgen). From expression data (GSE18684), we collected 20 R1881-treated and 20 untreated samples. We then applied DRAGEN to quantify the strength of differential regulation. Results show that *P*-values of the differentially regulated genes are significantly smaller than random cases (one-sided Wilcoxon test $P\text{-value} < 9 \times 10^{-6}$), further supporting the effectiveness of our method (Supplementary Material, Section S3.3).

We then evaluated the capability of DRAGEN on distinguishing differentially regulated gene sets that included AR. Again we mapped the gene sets from MsigDB category c2 onto the HTRI network. We obtained a total of 499 gene sets containing at least three edges, among which 50 contained the AR gene and its target genes. We checked the functional annotations of the 50 gene sets and removed one that was meaningful only for females. Finally, we obtained 49 gene sets as positive cases and the rest 450 as negative controls to perform the following experiment.

We applied our method to these gene sets and eventually generated a ranking list. Similar to the breast cancer case, we plotted the ROC in Figure 3b and obtained the AUC score as 0.761. By comparing with the other methods (the AUC is 0.582 for GSEA, 0.625 for ORA and 0.631 for SAFE), we clearly saw the high performance of our method. With DRAGEN, 14 out of the 49 gene sets containing AR were ranked among the top 50 in a list

Table 1. List of 20 gene sets that contain ESR1 and at least one of its target genes

Gene set name	*P*-value	FDR	Rank
GINESTIER_BREAST_CANCER_20Q13_AMPLIFICATION_UP	0.0000	0.0000	1
WESTON_VEGFA_TARGETS_6HR	0.0000	0.0000	1
WESTON_VEGFA_TARGETS	0.0000	0.0000	1
WESTON_VEGFA_TARGETS_3HR	0.0000	0.0000	1
POOLA_INVASIVE_BREAST_CANCER_DN	0.0001	0.0102	7
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	0.0017	0.0195	9
SMID_BREAST_CANCER_LUMINAL_B_UP	0.0015	0.0198	12
FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	0.0014	0.0203	14
HATADA_METHYLATED_IN_LUNG_CANCER_UP	0.0014	0.0203	14
GOZGIT_ESR1_TARGETS_DN	0.0014	0.0203	14
SMID_BREAST_CANCER_BASAL_DN	0.0016	0.0203	14
TOYOTA_TARGETS_OF_MIR34B_AND_MIR34C	0.0016	0.0203	14
REACTOME_SIGNALING_BY_ERBB4	0.0019	0.0208	27
DOANE_BREAST_CANCER_ESR1_UP	0.0012	0.0213	30
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	0.0021	0.0220	41
BENPORATH_ES_WITH_H3K27ME3	0.0030	0.0220	45
JISON_SICKLE_CELL_DISEASE_DN	0.0030	0.0220	45
VANTVEER_BREAST_CANCER_ESR1_UP	0.0026	0.0223	49
SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A4	0.0026	0.0223	49
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	0.0026	0.0223	49

These gene sets were ranked among the top 50 by DRAGEN out of 711 gene sets in the analysis of MCF-7 cells

composed of a total of 499 gene sets (Table 2 and Supplementary Table S8). In comparison, GSEA, ORA and SAFE found 6, 10 and 8 positive gene sets among their top 50 gene sets, respectively (Supplementary Tables S9–S11). From Table 2, we observed that the androgen signaling pathway was correctly ranked as the top 1, which was not detected by the other methods (i.e. found to be among the top 50). All three positive gene sets ranked among the top 10 by DRAGEN are presented in Supplementary Figure S7. The connections between several top ranked gene sets and androgen treatment on LNCaP cells reported in the literature were discussed in Supplementary Material (Section S3.4).

3.5 Application to a HeLa cell line

This experiment intended to examine gene regulation patterns altered during the cell cycle progression on a HeLa cell line. Differential regulations by many TFs were observed across the cell cycle progression in the literature (Burkhart *et al.,* 2010; Takahashi *et al*., 2000). For example, the E2F and pRB families were known to play pivotal roles in timely regulating gene expression during the cell cycle progression and their promoter binding affinities change during the process (Takahashi *et al*., 2000). Hence, we used the expression data at multiple time points in the cell cycle of HeLa cells to test the performance of DRAGEN. The expression data were obtained at six time points, so we manually separated the data into two classes. The first class covered the first three time points, all contained in interphase. The second class consisted of the other three time points, spanning mitosis. Since some of the members of the E2F family (e.g. E2F1) were found to be enriched in the ENCODE regulatory network, we combined the ENCODE network with the HTRI network as our regulatory network in the experiment. As before, we considered gene sets from MSigDB category c2 and focused on gene sets that contain at least three edges. We used the functional annotation in MSigDB as an evidence of association to the cell cycle progression, and treated gene sets whose functional annotations explicitly mention the involvement in the cell cycle progression as positive cases. This resulted in a total of 1193 gene sets, of which 28 gene sets were directly associated with cell cycle and defined as positives.

We applied GSEA, SAFE, ORA and DRAGEN to the HeLa dataset. The ROCs of DRAGEN and ORA were plotted in Figure 3c. Their AUC scores were 0.783 and 0.464, respectively. The ROCs of GSEA and SAFE were omitted here but presented in Supplementary Figure S8 because they were unable to find any significant gene sets (FDR < 0.25). Their detailed findings were given in Supplementary Tables S14 and S16. To understand the poor performance of these two methods, we performed two-sided *t*-test on all expressed genes to estimate their degree of differential expression, and we found that the proportion of significantly differentially expressed genes was less than 1% (with Bonferroni corrected *P*-value < 0.05). This might partly explain the poor performance of SAFE and GSEA, since both were based on differential expression. DRAGEN ranked 6 of the 28 positive gene sets among the top 50 of the 1193 gene sets (Table 3 and and Supplementary Table S13) while ORA ranked three positive gene sets among the top 50 (Supplementary Table S15). The results in Figure 3c suggested that DRAGEN was sensitive and effective in detecting differential regulation in the situation of low differential expression.

It is interesting to observe that two of the top ranked positive gene sets from DRAGEN (BIOCARTA_P27_PATHWAY and BIOCARTA_RACCYCD_PATHWAY) contain both genes RB1 and E2F1, and the regulation of RB1 by E2F1 changes significantly during the cell cycle progression (*P*-value = 0.0003).

Table 2. List of 14 gene sets that contain AR and at least one of its target genes

Gene set name	*P*-value	FDR	Rank
PID_AR_TF_PATHWAY	0.0000	0.0000	1
PID_HES_HEYPATHWAY	0.0003	0.0374	7
PID_AR_PATHWAY	0.0021	0.0374	8
KEGG_PATHWAYS_IN_CANCER	0.0041	0.0418	11
ZWANG_EGF_INTERVAL_DN	0.0009	0.0449	15
KEGG_PROSTATE_CANCER	0.0014	0.0466	19
CHANG_IMMORTALIZED_BY_HPV31_DN	0.0037	0.0543	28
HOSHIDA_LIVER_CANCER_SURVIVAL_DN	0.0028	0.0582	31
MASSARWEH_TAMOXIFEN_RESISTANCE_UP	0.0049	0.0582	32
FARMER_BREAST_CANCER_APOCRINE_VS_LUMINAL	0.0024	0.0599	35
PID_HNF3APATHWAY	0.0059	0.0613	43
MOHANKUMAR_TLX1_TARGETS_UP	0.0031	0.0619	45
LEE_LIVER_CANCER_SURVIVAL_UP	0.0067	0.0619	46
REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	0.0056	0.0635	48

These gene sets were ranked among the top 50 by DRAGEN out of 499 gene sets in the analysis of LNCaP cells

Table 3. List of six gene sets that are directly associated with cell cycle

Gene set name	*P*-value	FDR	Rank
BIOCARTA_P27_PATHWAY	0.0000	0.0000	1
BIOCARTA_RACCYCD_PATHWAY	0.0000	0.0000	1
WHITFIELD_CELL_CYCLE_S	0.0000	0.0000	1
BIOCARTA_G1_PATHWAY	0.0001	0.0066	18
BIOCARTA_CELLCYCLE_PATHWAY	0.0003	0.0112	32
SA_G1_AND_S_PHASES	0.0004	0.0126	38

These gene sets were ranked among the top 50 by DRAGEN out of 1193 gene sets in the analysis of HeLa cells

This differential regulation has been confirmed by ChIP experiments at different cell cycle phases in the literature (Burkhart *et al*., 2010).

3.6 Performance of the multiple-TF model

We repeated experiments for the MCF-7, LNCaP and HeLa datasets using the multiple-TF model and showed the results in Figure 3. For MCF-7, the AUC scores of the single-TF and multiple-TF models were 0.843 and 0.832, respectively. For LNCaP, the AUC scores were 0.761 and 0.783. For HeLa, the AUC scores were 0.783 and 0.818. These results suggested that the multiple-TF model outperformed the single-TF one on the LNCaP and HeLa datasets. The corresponding positive gene sets among their top 50 gene sets were shown in Supplementary Tables S7, S12 and S17, respectively.

3.7 Robustness of DRAGEN

Because of the dynamics of transcriptional regulatory relationships and limitation of existing experimental techniques (Bovolenta *et al*., 2012), the HTRI network, though built from numerous other databases that contain experimentally verified human regulatory interactions, might still be far from being complete and fully correct. With this consideration, we tested the robustness of our method with respect to the incompleteness and possible errors in the transcriptional regulatory network by entirely replacing the HTRI network with a network constructed according to a series of ChIP-seq experiments in the ENCODE project (Gerstein *et al*., 2012). See Supplementary Material (Section S3.6) for a detailed description of the HTRI and ENCODE networks.

Focusing on gene sets containing at least three edges in the ENCODE network, we obtained a total of 1000 gene sets, among which 51 sets contained at least one gene regulated by ESR1. Considering that the network was sparse and possibly incomplete, and being aware of the fact that a small number of differentially regulated edges inside a gene set might not be enough to represent the overall differential regulation mechanism of the entire gene set, we only selected gene sets with at least k ESR1 regulatory edges as the phenotype-associated gene sets, instead of all the 51 gene sets. For relatively small values of k (3 and 5), we collected 33 and 26 gene sets as the test gene sets, respectively. We then applied DRAGEN to rank these gene sets against the rest of the 1000 gene sets and presented the resulting ROCs in Supplementary Figure S9.

The AUC scores are 0.718 and 0.794 for $k = 3$ and 5, respectively. Though not as high as the one in the previous study using the HTRI network (AUC = 0.843), these AUC scores still strongly suggest the effectiveness of DRAGEN when used with the ENCODE network. Moreover, we notice that a larger k produces a better ROC with a higher AUC score (Supplementary Fig. S9). These observations are consistent with the expectation that gene sets containing more differentially regulated edges tend to rank toward the top. We further collect a total of 559 gene sets shared by the HTRI and ENCODE data, and find the Spearman's correlation coefficient of the relative ranks of these gene sets is only 0.42, suggesting that the relative ranks of these gene sets are actually very different.

3.8 Performance on random regulatory networks and gene sets

In order to assess the actual contributions of the regulatory network and gene sets to the performance of DRAGEN, we tested DRAGEN on regulatory networks and gene sets with randomly introduced noise. First, we perturbed the sub-network of each gene set by randomly shuffling N% of its nodes while it kept the same topology. For each parameter N ($=10, 20, \ldots, 80$), we perturbed the sub-network of every gene set 1000 times to generate 1000 test datasets. The original 71, 49 and 28 positive gene sets were still used as the positive cases. The performance of DRAGEN on these datasets is shown in Supplementary Figure S10. The figure clearly illustrates that the performance of DRAGEN declines as N increases. The median AUC scores are around 0.814, 0.750 and 0.763 for the MCF-7, LNCaP and HeLa data when N is equal to 10, but fall to 0.724, 0.691 and 0.722 when a half of the nodes in each sub-network are shuffled ($N = 50$). This demonstrates that the network information plays a key role in DRAGEN.

Then we perturbed the gene sets by randomly choosing N% genes of each gene set and replacing them with an equal number of genes selected randomly from all expressed genes. The regulatory network from the HTRI database was fixed as the input regulatory network. As before, for each parameter N ($= 10, 20, \ldots, 80$), we perturbed every gene set 1000 times to generate 1000 test datasets and used the original 71 and 49 positive gene sets as the positive cases. The performance of DRAGEN on these datasets is shown in Supplementary Figure S11. Again, the performance of DRAGEN clearly declines as N increases. The median AUC scores are around 0.828, 0.749 and 0.768 for the MCF-7, LNCaP and HeLa data when N is 10, but fall to 0.754, 0.696 and 0.674 when a half of the genes in every gene sets are shuffled ($N = 50$). Therefore, high-quality gene sets are also critical to the performance of DRAGEN.

4 CONCLUSION AND DISCUSSION

In this article, we have proposed a method, named DRAGEN, to integrate gene expression data and a regulatory network for differential regulation enrichment analysis. Through a series of comprehensive simulation and real data experiments, we demonstrate its power in detecting alterations of regulation not only between individual TFs and their targets but also among a set of genes.

The effectiveness of our method is due to a combination of two aspects. First, we systematically use gene expression data and transcriptional regulation information in an integrative way. Existing methods largely depend on gene expression data alone to detect differentially regulated gene sets that may be associated with a phenotype of interest, overlooking the fact that alterations of gene expression levels may actually result from the changes of regulatory patterns. Therefore, our method, which directly detects changes in regulatory relationships, is in principle more powerful in discovering gene sets in response to a perturbed phenotype. Second, we ground our method on a rigorous statistical model, which successfully connects the regulatory relationship between a TF and its target gene to their expression profiles. With this model, the detection of differentially regulated gene sets is converted to a series of hypothesis testing problems followed by the fusion of individual P-values, thereby enabling our method to quantify subtle changes in regulatory relationships for achieving a strong performance in differential regulation enrichment analysis.

On the other hand, several aspects of DRAGEN can perhaps be further improved. First, it is known that the expression of a gene is regulated by not only multiple TFs, but also post-transcriptional factors such as microRNAs. Besides, the expression level of a TF may not truly reflect the activity of the TF. Therefore, our method, though having taken the effects of multiple TFs into consideration, may fail to incorporate the contributions of other factors and thus produce additional false positives in a real application. In this sense, our method is more suitable to be used as an exploratory tool. Second, although the linear regression model demonstrates reasonably good power in detecting differentially regulated patterns, the true relationships between expression levels of TFs and target genes might not be ideally linear. Hence, a more sophisticated model that takes non-linearity into consideration might be desirable for capturing more general regulatory relationships. How to achieve a reasonable control over the model complexity to avoid overfitting when the sample size is limited is a question that needs to be carefully addressed. Third, sub-networks obtained by mapping gene sets to an underlying network may be incomplete and thus may miss important interactions. A plausible solution is trying to generate a more complete regulatory network. For example, the methods from DREAM Challenge Five (Marbach *et al*., 2010) can be used to construct a regulatory network from gene expression profiles, which could be further combined with the one derived from the biological experiments. How to combine experimental and computational resources to construct a high-quality comprehensive regulatory network and extend our current model taking into account all above considerations is a direction that we plan to pursue in the immediate future.

ACKNOWLEDGEMENTS

The authors thank Prof. Xuegong Zhang, Prof. Shirley Liu and Prof. Zhiping Wen for several helpful suggestions. We are also grateful to Winston Haynes for sharing with us programs and the anonymous reviewers for their constructive comments.

Funding: This research was partially supported by the National Basic Research Program of China [2012CB316504], the National High Technology Research and Development Program of China [2012AA020401], the National Science Foundation grant [DBI-1262107], the National Natural Science Foundation of China [61175002].

Conflict of interest: none declared.

REFERENCES

Armitage,P. *et al.* (2002) *Statistical methods in medical research*. 4th edn. Blackwell Scientific Publications, Oxford.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barry,W.T. *et al.* (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.
Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
Bovolenta,L.A. *et al.* (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, **13**, 405.
Breitling,R. *et al.* (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.
Burkhart,D.L. *et al.* (2010) Regulation of RB transcription in vivo by RB family members. *Mol. Cell Biol.*, **30**, 1729–1745.
Cheng,C. *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658–1667.
Cookson,W. *et al.* (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
Croft,D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
Dixon,A.L. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.
Duggan,D.J. *et al.* (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, **21**, 10–14.
Edgar,R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
Faith,J.J. *et al.* (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.
Fisher,R.A. (1925) *Statistical Methods for Research Workers*. Oliver & Boyd, Edinberg.
Gama-Castro,S. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.
Geistlinger,L. *et al.* (2011) From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, **27**, i366–i373.
Gerstein,M.B. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
Haynes,W.A. *et al.* (2013) Differential expression analysis for pathways. *PLoS Comput. Biol.*, **9**, e1002967.
Levenson,A.S. and Jordan,V.C. (1997) MCF-7: the first hormone-responsive breast cancer cell line. *Cancer Res.*, **57**, 3071–3078.
Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, **107**, 6286–6291.
Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
Neph,S. *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.
Nevins,J.R. and Potti,A. (2007) Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat. Rev. Genet.*, **8**, 601–609.
Ogata,H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
Patel,M. *et al.* (2012) Tumor-specific retargeting of an oncogenic transcription factor chimera results in dysregulation of chromatin and transcription. *Genome Res.*, **22**, 259–270.
Patton,E.E. and Harrington,L. (2013) Cancer: trouble upstream. *Nature*, **495**, 320–321.
Risbridger,G.P. *et al.* (2010) Breast and prostate cancer: more similar than different. *Nat. Rev. Cancer*, **10**, 205–212.
Shojaie,A. and Michailidis,G. (2009) Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol.*, **16**, 407–426.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545–15550.
Takahashi,Y. *et al.* (2000) Analysis of promoter binding by the E2F and pRB families in vivo: distinct E2F proteins mediate activation and repression. *Genes Dev.*, **14**, 804–816.
Vaske,C.J. *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.
Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
Zuo,T. *et al.* (2007) FOXP3 is an X-linked breast cancer suppressor gene and an important repressor of the HER-2/ErbB2 oncogene. *Cell*, **129**, 1275–1286.