Published in final edited form as:

J Mol Biol. 2017 February 03; 429(3): 446–453. doi:10.1016/j.jmb.2016.10.030.

BFDCA: A Comprehensive Tool of Using Bayes Factor for Differential Co-expression Analysis

Duolin Wang^{1,2}, Juexin Wang², Yuexu Jiang^{1,2}, Yanchun Liang^{1,2}, and Dong Xu^{2,1,*}

¹College of Computer Science and Technology, Jilin University, China

²Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, USA

Abstract

Comparing the gene-expression profiles between biological conditions is useful for understanding gene regulation underlying complex phenotypes. Along this line, analysis of differential co-expression (DC) has gained attention in recent years, where genes under one condition have different co-expression patterns compared with another. We developed an R package BFDCA for DC analysis. BFDCA is unique in integrating various aspects of DC patterns (including Shift, Cross, and Re-wiring) into one uniform Bayes factor. We tested BFDCA using simulation data and experimental data. Simulation results indicate that BFDCA outperforms existing methods in accuracy and robustness of detecting DC pairs and DC modules. Results of using experimental data suggest that BFDCA can cluster disease-related genes into functional DC subunits and estimate the regulatory impact of disease-related genes well. BFDCA also achieves high accuracy in predicting case/control phenotypes by using significant DC gene pairs as markers. BFDCA is publicly available at <http://dx.doi.org/10.17632/jdz4vtvnm3.1>.

Graphical Abstract

*Correspondence: xudong@missouri.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Gene expression; gene regulation; Bayes factor; R package; multivariate normal distribution

Introduction

To infer a set of genes associated with a particular condition or phenotype, comparing gene-expression profiles between different biological conditions or phenotypes is a common approach [1, 2]. The genes that are found with significant differential expression (DE) are often useful for understanding the gene regulation underlying complex phenotypes and may serve as biomarker candidates. For the last two decades, the analyses of gene-expression data have been dominated by univariate-based methods, which analyze expression data under the assumption that genes are independent. However, such methods may overlook the interactions and coordination among genes in performing their functions [3]. In particular, the small changes in multiple differentially regulated genes may not be captured by single gene test (e.g. t-test or ANOVA). This is especially true for transcription factors, which are often stably expressed at baseline levels [4] but have major causal regulatory effects in biological processes. For example, myostatin (GDF8) was not a DE gene between Piedmontese and Wagyu cattle, but it is a differential hub gene in a DE network [5]. Another example is the “switching mechanism”, where a gene can be an activator or a suppressor depending on another gene’s activities under different conditions. A well-known case is the Max gene, a transcription factor, which activates transcription when it heterodimerizes with Myc and represses transcription when it heterodimerizes with Mad [6].

Due to the above facts, a promising approach of differential co-expression (DC) analysis emerged in recent years. It seeks gene modules that are co-expressed under one condition but have different co-expression patterns in another condition. In addition to investigating gene expression changes from single genes independently, DC takes gene-gene interactions into consideration and examines the changes in gene correlations between different conditions. Thus, DC can reveal the regulatory impact of a gene on other genes via identifying condition-specific patterns in the context of regulatory relationships. There are various scenarios of differential co-expression presented between two genes, and three extreme cases are summarized [3, 7] as follows: (1) **shift**, the correlation of genes does not change between two conditions, but expression values under one condition are clearly different from those under the other (Fig. 1a); (2) **cross**, genes are positively correlated under one condition but negatively correlated under the other (Fig. 1b); (3) **re-wiring**, genes are positively (or negatively) correlated under one condition, while not correlated (or less correlated) under the other (Fig. 1c). More complex changes in co-expression patterns often occur, which from a mathematical perspective, can be represented as different distributions of gene expression profiles between two potentially co-expressed genes under different conditions.

Several computational approaches have been developed for DC analysis. They can be divided into two categories: targeted and untargeted approaches. In targeted approaches, e.g., Gene Set Co-expression Analysis (GSCA) [8] and Gene Sets Net Correlations Analysis

(GSNCA) [9], pre-defined gene sets are required to test correlation changes between conditions. In untargeted approaches, no pre-defined gene sets are required and gene modules are detected on the basis of their differential co-expression profiles. Unlike targeted approaches, untargeted approaches are capable of detecting novel modules. A group of untargeted approaches aimed at detecting DC pairs, such as EBcoexpress, an empirical Bayesian approach for identifying DC gene pairs [10] and ROS-DET, a robust detector of switching mechanisms [11]. Another group of untargeted approaches aimed at detecting DC modules, such as DiffCoEx [12], which uses the Weighted Gene Coexpression Network Analysis (WGCNA) framework [13] for DC modules detection, and DICER [14], which extracts differential gene clusters and meta-modules by analyzing a labeled gene-expression matrix. Many of these existing methods [11, 12, 14] only aimed at detecting significant changes in the correlation of gene pairs, but few of them considered the “shift” scenario. However, evidence from previous studies shows that considering shifts in mean and changes in correlations could produce complementary information for DC modules detection and improve identification of functionally important genes [5, 15]. Beside this, another shortcoming of these existing methods is that they always produce gene sets including tens of thousands of genes. Though they can always obtain significant biological meaning, the huge number of the resulting genes is too many for practical applications.

Here we present BFDCA (Bayes Factor approach for Differential Co-expression Analysis), a novel untargeted approach for DC analysis. Besides detecting DC modules, BFDCA quantitatively characterizes the regulatory impact of genes and provides a relatively small number of gene pairs, which can lead to a high accuracy classifier. In BFDCA, the pairwise DC is estimated by integrating different aspects of DC into one uniform Bayes factor, a ratio of marginal likelihood of the data between the alternative hypothesis and the null hypothesis. As opposed to using *p-values* [11, 14, 15], making inference with Bayes factor has many advantages [16]. Particularly, *p-value* simply indicates the chance of rejecting the null hypothesis, but ignores the acceptance of the alternative hypothesis. In contrast, Bayes factor considers the likelihood of both the null and the alternative hypothesis. Instead of focusing on the significant changes in the correlations of gene pairs [11, 12, 14], BFDCA detects gene sets that possess different distributions of gene expression profiles among different conditions, so that it can cover many scenarios of DC and has the capacity to combine different DC scenarios comprehensively. It can also make use of the integrated information of shifts in mean and changes in correlations for identifying functionally important genes. We demonstrate the performance of BFDCA by comparing with several existing methods and applying it to two real case studies.

Results

Overview of BFDCA

BFDCA is implemented in a general R package freely available for download from <http://dx.doi.org/10.17632/jdz4vtvnm3.1>. It takes gene expression data under different conditions as input, without any other information needed. BFDCA has three main functions: (1) clustering condition-specific genes into functional DC subunits; (2) quantitatively characterizing the regulatory impact of genes based on their differential connectivity within

DC structures; and (3) providing a DC-based model to predict case/control phenotypes by taking DC significant gene pairs as markers.

Figure 1 summarizes the algorithm used by BFDCA, which consists of five phases (details can be found in Text S1):

1. BFDCA calculates a Bayes factor (a ratio of marginal likelihood of the data between the alternative hypothesis and the null hypothesis) to estimate the DC of gene pairs. For all gene pairs to be tested, the alternative hypothesis is that the under test gene pair g manifests DC associated with the condition, wherein g (1) has different distributions of gene expression profiles between two conditions and (2) correlates at least in one condition. The null hypothesis, as complementary to the alternative hypothesis, is that g manifests no DC associated with the condition. Our proposed Bayes factor integrates different scenarios of DC into one single term and represents the significance level of evidence in favor of the alternative hypothesis, where the larger the Bayes factor, the more DC a gene pair manifests.
2. DC edges with low values of Bayes factors will be discarded. A generally accepted interpretation to the scale of Bayes factor is provided in [17]. Specifically, edges with Bayes factors less than 6 are filtered out.
3. WGCNA [13] is applied to the remaining DC pairs to infer DC modules using the edge weights based on the Bayes factors calculated by BFDCA.
4. Genes are assigned with weights based on their connectivity to other genes within DC modules. Intra-modular hub genes often obtain high weights, which indicates that they may play important roles in the regulatory network compared with other genes.
5. Significant DC gene pairs are selected according to their end-node importance and edge values. These gene pairs can be used as edge markers and we will demonstrate that they can achieve high accuracy in predicting case/control phenotypes.

Simulation study

To compare the performance of BFDCA in detecting DC pairs and DC modules with other methods in a fully controlled setting, we designed a simulation study to mimic real DC as close as possible. The details of simulation process are described in Text S2. In brief, the simulation data set contains 7 co-expression modules, each with 50 genes. Modules 1 and 2 represent the stable co-expression modules, in which genes have the same co-expression patterns under two conditions; modules 3 and 4 represent the intra-module DC, where genes in module 3 are only correlated under one condition and genes in module 4 are highly correlated under one condition, but less correlated under the other condition; module 5 represents the “shift” scenario of DC, in which the correlations of genes under two conditions do not change while their expression levels have an overall increase or decrease; modules 6 and 7 represent the inter-module DC, where genes within each module have

conserved co-expression patterns under each condition but have a “cross” correlation between the two modules.

We compared the performance in detecting DC pairs on simulation data with existing methods DiffCoEx, ROS-DET and other methods described in [18], including difference of correlation coefficients (DCC-P and DCC-S), entropy-based approaches (ENT-P and ENT-S), considering both Pearson and Spearman correlation coefficients, difference of mutual information (DMI), and basic method for “shift” pattern (Sshift) [7, 18]. Here module 5 is excluded since the “shift” scenario cannot be detected by most of other approaches. The comparisons were carried out by different sample sizes ($N=30, 50, 100$). The precision-recall curve is used to show the comparisons. For each sample size, 100 independent simulations were generated and the average result is shown. The comparisons were conducted under two scenarios: when all the genes between conditions have no change in mean (Fig. 2a, scenario 1); when the genes within modules 3 and 4 have small changes in mean besides the changes in correlations under the two conditions while other modules are unchanged (Fig. 2b, scenario 2). The comparison results for different sizes of gene modules (6 and 20 in each module) are as shown in Figs. S1 and S2. We also compared BFDCA with these methods on an independent simulation study for “shift” and “cross” patterns following exactly the same procedure as in [18]. The results are shown in Figs. S3 and S4. Overall, BFDCA shows consistently reliable results under different scenarios, and it is the only method that can detect both the shift in mean and changes in correlations, which is especially useful in identifying functionally important genes.

The performance of detecting DC modules was compared with DiffCoEx, which used a similar strategy to generate DC modules. Both methods apply the WGCNA [13] framework to extract DC modules from DC pairs. The framework contains two main steps: (1) a dissimilarity matrix is constructed by calculating the topological overlap between each gene pair from an adjacency matrix; (2) a hierarchical clustering followed by tree cutting using dynamicTreeCut [19] is applied on the dissimilarity matrix to cluster genes into DC modules. The major difference between BFDCA and DiffCoEx lies in how the adjacency matrix is calculated. The comparison was conducted on the scenario 1, with module 5 included. The dendrograms for each method under different softPower values in the range from 1 to 7 are shown in Fig. S5. From the dendrograms, we can infer that BFDCA can detect higher compact gene clusters since the heights of the detected modules are smaller, so that they are not sensitive to the cutHeight parameter of the dynamicTreeCut [19] method. The results also confirm that DiffCoEx cannot find module 5 since the correlation of genes in module 5 does not change between two conditions and it is hard to detect module 3 without increasing false positive rate. On the contrary, BFDCA can detect modules 3 and 5 correctly in a wider parameter range. The comparison with DiffCoEx using the same simulation procedure as described in DiffCoEx paper [12] was shown in Fig. S6. The overall comparison results show that BFDCA can detect DC modules with high sensitivity and is robust in a wider parameter range compared with DiffCoEx.

Case study

Experimental expression data of Acute Lymphoblastic Leukemia (ALL) and the breast cancer are used as two cases to test the performance of BFDCA. Data collection and preprocessing are described in Text S3.

ALL data

For the ALL dataset, we identified 61 DC modules consisting of 3682 genes in total. Module assignment and weight assignment are in Table S1. A full network built based on the 61 DC modules is shown in Table S2. The essential links among genes, which were from a union of the first and second minimum spanning trees (MST2) representing the skeleton of the intra-module DC are described in Table S3. Figure 3a shows examples of detected DC gene pairs from “shift”, “cross” and “re-wiring” scenarios, where we can clearly see the different patterns of gene pairs under two conditions. Figures 3b and 3c shows a DC module contains gene PDGFRB, a known hub gene also found in the breast cancer [20]. This is an example showing that BFDCA can detect a DC hub gene that cannot be detected by single gene test (t-test, p -value=0.09). The genes in this module are enriched (p -value 0.05) with cancer-related mutations, whereas 7 out of 61 genes are annotated as cancer-related mutation genes [21]. In Figs. 3b and c, known interactions from GENEMANIA [22] and changes of correlations between case and control samples are represented. The GENEMANIA analysis reveals that there are many known interactions within the DC module. The different patterns of correlations shown in Figs. 3b and 3c illustrate that genes within this DC module are generally more correlated in case samples than in control samples.

To characterize DC modules at the biological functional level, we carried out pathway enrichment analysis for gene sets within each DC module, respectively. The significantly enriched pathways of each DC module are listed in Table S4, which shows that all the detected DC modules are significantly enriched to relevant KEGG pathways.

We extracted 211 significant DC gene pairs from the DC modules and applied these gene pairs as edge markers to a DC-based model to validate their power in predicting BCR/ABL mutation (details are in Text S1). We used an effective cross-validation method, leave-one-out cross-validation (LOOCV) [23] to validate the results. After 79 times of LOOCV, the accuracy rate reached 97%. It means that for 79 samples only 2 samples were misclassified. All the selected significant DC edges are provided in Table S5.

The significantly enriched KEGG pathways for genes in the edge markers are listed in Table S6. Among these pathways, some are cancer-related or even targets for cancer therapies. The most significant one is the Endocytosis pathway. Several recent studies [24–27] have reported that the abnormalities of the Endocytosis pathway might play a role in cancer and it may constitute a novel target for cancer therapy [26]. This is also supported by the presence in leukemia of genetic alterations involving endocytic proteins [27]. Other examples are Glycosaminoglycan biosynthesis pathway [28, 29] and Renin pathway [30], both of which have been reported to play roles in cancer progression.

Since the selected DC gene pairs are network markers, the network ontology analysis (NOA) [31] was used to conduct enrichment analysis, as shown in Table S7. Interestingly, the association between gene CAV1 and KCND2 was significantly annotated with the terms “regulation of membrane potential” and “ion transport”; meanwhile, many other DC edges are significantly enriched in membrane-related cellular component GO terms. In fact, membrane potential, controlled by the regulation mechanism of ion channels/transporters, is a key bioelectrical signal that modulates important cellular activities, such as proliferation and differentiation [32, 33]. Emerging data provides consistent results and shows that cancer cells possess distinct properties of membrane potential, suggesting that the level of membrane potential has functional roles in cancer cell [33], and might be a valuable clinical marker for tumor detection and treatment.

Breast Cancer

From the ALL dataset, we show that by using significant DC gene pairs as edge markers, BFDCA can achieve very high LOOCV classification accuracy rate. To further validate the generalization ability and the reproducibility of BFDCA, we carried out deeper analyses on three independent breast cancer datasets.

To validate the generalization ability, we trained a model on one data set, and performed prediction on the other two independent datasets. We extracted 248 significant DC gene pairs (Table S8) from **GSE7390** and trained a DC-based model on this dataset. The LOOCV prediction accuracy for the training dataset is 91%. When using the trained model to predict the ER status for the other two independent GEO breast cancer datasets (**GSE21653** and **GSE2034**), the prediction accuracies for **GSE21653** and **GSE2034** were 89% and 85%, respectively. The significantly enriched KEGG pathways for the genes included in the edge markers are listed in Table S9. Significantly enriched annotated GO terms for edge markers are listed in Table S10. Among these enriched pathways and GO terms, many are related to cancer and important biological processes.

We also validated the reproducibility of BFDCA. In one aspect, we estimated the significance of the overlapping edges between DC modules generated by different datasets. The test DC modules were generated from **GSE7390**, and then the other two DC module sets were generated from **GSE21653** and **GSE2034**, respectively. The representation factor for the overlapping edges between **GSE7390** and **GSE21653** was 97, and for **GSE7390** and **GSE2034** was 117. The probability of finding an overlap of DC gene pairs was calculated by hypergeometric probability. The corresponding *p-values* are both nearly 0. On the other hand, we estimated the significance of the overlapping top weighted genes between datasets. The comparisons of top-100, top-500, top-1000, and top-1500 between datasets were performed and the probability of finding an overlap of top weighted genes was calculated by hypergeometric probability, where corresponding *p-values* for GSE7390~GSE21653 is 1.73e-15 (top-100), 1.77e-27 (top-500), 4.80e-23 (top-1000) and 1.82e-45 (top-1500) and for GSE7390~GSE2034 is 9.75e-08 (top-100), 2.84e-23 (top-500), 7.36e-53 (top-1000) and 5.05e-75 (top-1500).

All these analyses above reveal the generalization ability and the reproducibility of BFDCA on independent datasets.

Conclusion and Discussion

In this study, we proposed a novel approach for differential co-expression analysis. This approach is implemented in a general R package, which can be easily used by researchers using only gene expression data. BFDCA is different from other existing methods, since it integrates different aspects of DC patterns into a uniform Bayes factor and aims to detect gene sets that possess different distributions of gene expression profiles between two different conditions. Compared with approaches focusing only on different expression correlations, BFDCA has the ability to integrate node information with edge information. The approaches that comprehensively analyze nodes and edges are proven to be valuable in analyzing biological data [5, 15], especially for the regulatory importance of nodes.

For experimental gene expression data analysis, BFDCA provides three main functions. First, it clusters condition-specific genes into functional DC subunits. This has been shown from the ALL dataset, that all the detected DC modules are significantly enriched to relevant KEGG pathways. Second, it quantitatively characterizes the regulatory impact of genes by assigning a weight to each gene. The hub genes are usually assigned with high weights, revealing their important roles in the context of regulatory network. The reproducibility of BFDCA in identifying DC gene pairs and top weighted genes has been demonstrated in the validation of using independent breast cancer datasets. BFDCA also provides a set of significant DC gene pairs, which can be used as edge markers. We applied these edge markers in different datasets to distinguish disease phenotypes. For the ALL dataset, we used the edge markers to distinguish samples in terms of having BCR/ABL mutation or not. The LOOCV accuracy of the prediction is high.

Remarkably, the edge markers extracted for breast cancer dataset **GSE7390** also achieved high prediction accuracies for the two independent-dataset validations (**GSE21653** and **GSE2034**) in ER positive and ER negative classification, with similar classification accuracies as the LOOCV accuracy of the training dataset. This result suggests that the edge markers extracted by BFDCA can be generalized across samples and they might capture biological meaningful features to distinguish ER status for breast cancer. Unlike edge markers alone as in [34], BFDCA extracts both significant DC gene pairs that possess high Bayes factor and important end-nodes estimated in the network context. BFDCA does not remove the redundant information or correlated markers in analyses. The method does not require the independency among markers, which is likely to be the case in real data. This may be the reason why our method is robust with good generalization ability in predicting independent datasets.

BFDCA also has some limitations. Like many other statistical methods, BFDCA may suffer from the “small sample” problem, since our approach needs enough samples to estimate the hyperparameters; however, it can be compensated by the topological overlap calculated by involving the topological relationships with neighbor genes. For typical experimental data, this should not be a problem since N is typically much larger than 30.

In conclusion, BFDCA provides a novel way to conduct differential co-expression analysis, and it can provide a comprehensive analysis of the gene expression data, which may generate biologically meaningful hypothesis for researchers in a novel way.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work has been supported by the US National Institutes of Health (GM100701) and the National Natural Science Foundation of China (Grants 61272207, 61472158, 61572228).

Abbreviations

DC	differential co-expression
DE	differential expression
BFDCA	Bayes Factor approach for Differential Co-expression Analysis

References

1. Asyali MH, Colak D, Demirkaya O, Inan MS. Gene expression profile classification: a review. *Current Bioinformatics*. 2006; 1:55–73.
2. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Reviews Genetics*. 2007; 8:601–9.
3. de la Fuente A. From ‘differential expression’ to ‘differential networking’—identification of dysfunctional regulatory networks in diseases. *Trends in genetics*. 2010; 26:326–33. [PubMed: 20570387]
4. Zampieri M, Soranzo N, Altafini C. Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics*. 2008; 24:1510–5. [PubMed: 18467346]
5. Hudson NJ, Reverter A, Dalrymple BP. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput Biol*. 2009; 5:e1000382. [PubMed: 19412532]
6. Ayer DE, Eisenman RN. A switch from Myc: Max to Mad: Max heterocomplexes accompanies monocyte/macrophage differentiation. *Genes & Development*. 1993; 7:2110–9. [PubMed: 8224841]
7. Kayano M, Shiga M, Mamitsuka H. Detecting differentially coexpressed genes from labeled expression data: a brief review. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*. 2014; 11:154–67.
8. Choi Y, Kendziorski C. Statistical methods for gene set co-expression analysis. *Bioinformatics*. 2009; 25:2780–6. [PubMed: 19689953]
9. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics*. 2014; 30:360–8. [PubMed: 24292935]
10. Dawson JA, Ye S, Kendziorski C. R/EBcoexpress: an empirical Bayesian framework for discovering differential co-expression. *Bioinformatics*. 2012; 28:1939–40. [PubMed: 22595207]
11. Kayano M, Takigawa I, Shiga M, Tsuda K, Mamitsuka H. ROS-DET: robust detector of switching mechanisms in gene expression. *Nucleic Acids Research*. 2011; 39:e74–e. [PubMed: 21459849]
12. Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*. 2010; 11:497. [PubMed: 20925918]
13. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008; 9:1. [PubMed: 18173834]

14. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol*. 2013; 9:e1002955. [PubMed: 23505361]
15. Lui TW, Tsui NB, Chan LW, Wong CS, Siu PM, Yung BY. DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. *BMC Bioinformatics*. 2015; 16:182. [PubMed: 26026612]
16. Jarosz AF, Wiley J. What are the odds? A practical guide to computing and reporting Bayes Factors. *The Journal of Problem Solving*. 2014; 7:2.
17. Kass RE, Raftery AE. Bayes factors. *Journal of the american statistical association*. 1995; 90:773–95.
18. Ho Y-Y, Cope L, Dettling M, Parmigiani G. Statistical methods for identifying differentially expressed gene combinations. *Gene Function Analysis*. 2007:171–91.
19. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*. 2008; 24:719–20. [PubMed: 18024473]
20. Chai F, Liang Y, Zhang F, Wang M, Zhong L, Jiang J. Systematically identify key genes in inflammatory and non-inflammatory breast cancer. *Gene*. 2016; 575:600–14. [PubMed: 26403314]
21. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*. 2011; 39:D945–D50. [PubMed: 20952405]
22. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*. 2010; 38:W214–W20. [PubMed: 20576703]
23. Torgo, L. Data mining with R: learning with case studies. Chapman & Hall/CRC; 2010.
24. Polo S, Pece S, Di Fiore PP. Endocytosis and cancer. *Current Opinion in Cell Biology*. 2004; 16:156–61. [PubMed: 15196558]
25. Rosell R. Cancer and alterations in the endocytic pathway. *Future Medicine*. 2007
26. Mosesson Y, Mills GB, Yarden Y. Derailed endocytosis: an emerging feature of cancer. *Nature Reviews Cancer*. 2008; 8:835–50. [PubMed: 18948996]
27. Floyd S, De Camilli P. Endocytosis proteins and cancer: a potential link? *Trends in cell biology*. 1998; 8:299–301. [PubMed: 9704404]
28. Yip GW, Smollich M, Götte M. Therapeutic value of glycosaminoglycans in cancer. *Molecular Cancer Therapeutics*. 2006; 5:2139–48. [PubMed: 16985046]
29. Afratis N, Gialeli C, Nikitovic D, Tsegenidis T, Karousou E, Theocharis AD, et al. Glycosaminoglycans: key players in cancer cell biology and treatment. *Febs Journal*. 2012; 279:1177–97. [PubMed: 22333131]
30. Wegman-Ostrosky T, Soto-Reyes E, Vidal-Millán S, Sánchez-Corona J. The renin-angiotensin system meets the hallmarks of cancer. *Journal of Renin-Angiotensin-Aldosterone System*. 2015; 16:227–33.
31. Wang J, Huang Q, Liu Z-P, Wang Y, Wu L-Y, Chen L, et al. NOA: a novel Network Ontology Analysis method. *Nucleic Acids Research*. 2011:gkr251.
32. Becchetti A, Munaron L, Arcangeli A. The role of ion channels and transporters in cell proliferation and cancer. *Frontiers in Physiology*. 2013; 4:312. [PubMed: 24194722]
33. Yang M, Brackenbury WJ. Membrane potential and cancer progression. *Front Physiol*. 2013; 4:185. [PubMed: 23882223]
34. Zhang W, Zeng T, Liu X, Chen L. Diagnosing phenotypes of single-sample individuals by edge biomarkers. *Journal of Molecular Cell Biology*. 2015:mjv025.

Highlights

- Differential co-expression (DC) analysis is useful in understanding functional and regulatory relationships among genes, but few tools are available for effective DC analysis.
- By integrating different aspects of DC patterns into a uniform Bayes factor, BFDCA can estimate DC between two conditions with high sensitivity.
- BFDCA clusters condition-specific genes into functional DC subunits and quantitatively characterizes their regulatory impact on genes.
- BFDCA identifies significant DC gene pairs and achieves high accuracy in predicting case/control phenotypes by using these gene pairs as markers.
- BFDCA is implemented in a general R package, which can be easily used by biological researchers and generate biologically meaningful hypotheses.

Fig. 1.

Overview of the algorithm used by BFDCA. The input (left) is a set of gene expression profiles from two classes of samples. The algorithm performs the following: (1) estimating the DC of gene pairs by calculating Bayes factors; (2) filtering DC edges whose Bayes factor values are too small; (3) inferring DC modules using the remaining edges; (4) assigning each DC with a weight to reflect its regulatory importance; (5) selecting significant DC gene pairs according to their end-node importance and edge values. (a), (b) and (c) represent shift, cross and re-wiring scenarios of differentially co-expressed gene pair. X-axis and y-axis represent gene expression levels of gene X and gene Y, respectively. Dots in different colors and shapes represent samples under two different conditions. Taken together, these scenarios can be interpreted as different distributions of gene expression profiles between two potentially co-expressed genes under different conditions.

Fig. 2.

Precision-recall curves from different methods for DC pair detection. The comparison is generated under different sample sizes, $N=30, 50, 100$. The X-axis represents the recall and the Y-axis represents the precision. The results are for the following scenarios: (a) all the genes between conditions have no change in mean; (b) the genes within modules 3 and 4 have small changes in mean besides the changes in correlations.

Fig. 3.

DC patterns in ALL dataset. (a) Examples of different DC scenarios. The X-axis and y-axis represent gene expression levels for two genes, respectively. Dots in blue represent samples with BCR/ABL mutation, and dots in red represent control samples. Known interactions from GENEMANIA for each gene pair are indicated (bottom). The circles represent the 95% contours of the bivariate normal density estimated from case and control samples. This function is provided in BFDCA. (b) and (c) DC patterns for hub gene PDGFRB. Co-expression in BCR/ABL mutation samples (b) and control samples (c) represented by the correlation coefficient. The highlight nodes represent genes which are annotated as cancer-related mutation genes. Known interactions from GENEMANIA within the module are represented in different line styles. The co-expression and predicted interactions were excluded.