BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 20 2010, pages 2637–2638
doi:10.1093/bioinformatics/btq471

Gene expression

Advance Access publication August 26, 2010

DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data

Bao-Hong Liu[1,2,3,†], Hui Yu[2,3,†], Kang Tu[4], Chun Li[5], Yi-Xue Li[1,2,3,*] and Yuan-Yuan Li[2,3,*]

[1]School of Life Science and Technology, Tongji University, Shanghai 200092, [2]Bioinformatics Center, Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, [3]Shanghai Center for Bioinformation Technology, Shanghai 200235, P. R. China, [4]National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD 30105 and [5]Department of Biostatistics, Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: Gene coexpression analysis was developed to explore gene interconnection at the expression level from a systems perspective, and differential coexpression analysis (DCEA), which examines the change in gene expression correlation between two conditions, was accordingly designed as a complementary technique to traditional differential expression analysis (DEA). Since there is a shortage of DCEA tools, we implemented in an R package 'DCGL' five DCEA methods for identification of differentially coexpressed genes and differentially coexpressed links, including three currently popular methods and two novel algorithms described in a companion paper. DCGL can serve as an easy-to-use tool to facilitate differential coexpression analyses.

Contact: yyli@scbit.org and yxli@scbit.org

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

Received on February 25, 2010; revised on August 9, 2010; accepted on August 12, 2010

1 INTRODUCTION

From the perspective of systems biology, gene coexpression analysis is useful for investigating gene interconnection at the expression level. Differential coexpression analysis (DCEA), which examines the change in expression correlation of gene pairs between two conditions, helps to explore the global transcriptional mechanisms underlying phenotypic changes. Compared with traditional differential expression analysis (DEA), the development of DCEA tools is lagged. In this work, we developed an R package, DCGL, implementing three previously proposed DCEA methods and two new algorithms reported in a companion paper (Yu,H. *et al.*, submitted for publication).

Log Ratio of Connections (LRC) calculates the logarithm of the ratio of the connectivities of a gene between two conditions (Reverter *et al.*, 2006). Average Specific Connection (ASC) counts the 'specific connections' that exist in only one coexpression network (Choi *et al.*, 2005). The weighted gene coexpression network analysis (WGCNA) weights links with correlation coefficients and compares the sums of the correlation coefficients of a gene (Mason *et al.*, 2009; van Nas *et al.*, 2009). In contrast, our two methods, differential coexpression profile (DCp) and differential coexpression enrichment (DCe), are designed based on the exact coexpression changes of gene pairs, and thus can differentiate significant coexpression changes from relatively trivial ones, and identify coexpression reversal between positive and negative (Yu,H. *et al.*, submitted for publication). All the five methods are able to identify differentially coexpressed genes (DCGs) from microarray datasets, and DCe is also able to pick out differentially coexpressed gene pairs or links (DCLs).

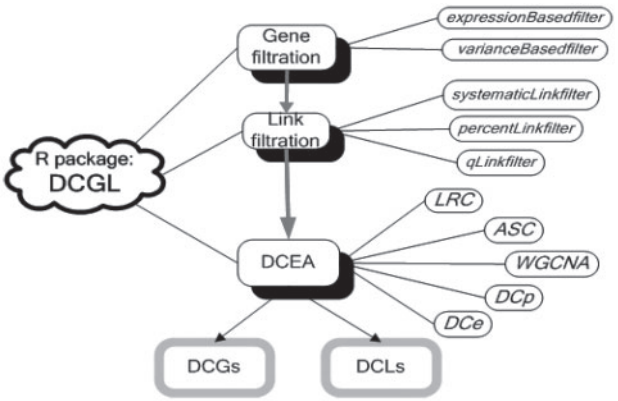

Fig. 1. DCGL design. Function names are shown in italic texts.

2 DESIGN

A typical DCEA workflow involves three successive procedures: gene filtration, link filtration, DCG and DCL identification. Correspondingly, DCGL consists of three parts of functions (Fig. 1). For gene filtration, one choice is based on the expression level (*expressionBasedfilter*) and the other based on its variability (*varianceBasedfilter*). For link filtration, we provide three functions for cutting off coexpression values (*systematicLinkfilter*, *percentLinkfilter* and *qLinkfilter*). A gene pair (link) is filtered out

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

if both of its coexpression values for two conditions are lower than the cutoff.

The third part, also the core of the package, includes five methods for identifying DCGs and DCLs, which mainly differ in the measure of differential coexpression (dC) of a gene. After the steps of gene filtration and link filtration, suppose gene i is associated with n_i links whose coexpression values are projected to $X=\{x_{i1},x_{i2},\ldots,x_{in_i}\}$and $Y=\{y_{i1},y_{i2},\ldots,y_{in_i}\}$ for two conditions. The dC measures of different methods are given in the following equations.

$$\mathrm{dC}_i(\mathrm{DCp})=\sqrt{\frac{(x_{i1}-y_{i1})^2+(x_{i2}-y_{i2})^2+\cdots+(x_{in_i}-y_{in_i})^2}{n_i}} \quad (1)$$

$$\mathrm{dC}_i(\mathrm{DCe})=\sum_{x=k_i}^{n_i} C_{n_i}^{x}\left(\frac{K}{N}\right)^{x}\left(1-\frac{K}{N}\right)^{n_i-x} \quad (2)$$

where N and K indicate the numbers of total links and total DCL links in the coexpression network, respectively, and n_i and k_i indicate the links and DCLs connected to gene i (see Yu,H. *et al.*, submitted for publication).

$$\mathrm{dC}_i(\mathrm{WGCNA})=\frac{s_1(i)}{\max\limits_j(s_1(j))}-\frac{s_2(i)}{\max\limits_j(s_2(j))} \quad \text{where}$$
$$\begin{cases} s_1(i)=x'_{i1}+x'_{i2}+\cdots+x'_{in_i} \\ s_2(i)=y'_{i1}+y'_{i2}+\cdots+y'_{in_i} \end{cases}, \quad (3)$$

with x′ and y′ are transformed from original x and y values with a 'soft-thresholding' strategy (Mason *et al.*, 2009; van Nas *et al.*, 2009).

$$\mathrm{dC}_i(\mathrm{LRC})=\log_{10}(\#(C_1(i))/\#(C_2(i))) \quad (4)$$

Link sets $C_1(i)$ and $C_2(i)$ for two conditions are determined by screening the coexpression values according to a certain threshold.

$$\mathrm{dC}_i(\mathrm{ASC})=(c_1(i)+c_2(i))/2 \quad \text{where}$$
$$\begin{cases} c_1(i)=\#(C_1(i)-C_1(i)\cap C_2(i)) \\ c_2(i)=\#(C_2(i)-C_1(i)\cap C_2(i)) \end{cases} \quad (5)$$

3 IMPLEMENTATION

DCGL is released as an R package including two gene filtering functions, three link filtering functions and five DCEA functions (Fig. 1). These functions generally expect gene expression matrices (with genes in rows and samples in columns) as a major input, and the ultimate output are genes ranked by dC measure or *P*-value, from which one can obtain a DCG list. DCe has an additional output of classified DCLs. DCGL can be obtained from the supplementary data to this manuscript, or at http://cran.r-project.org/web/packages/DCGL/index.html.

We tested the five DCEA methods using dataset GSE3068 obtained from GEO (Table 1). Note that this test was carried out with the most time-efficient option of link filtration (setting thresholds on coexpression value directly). For the memory analysis, we tested DCp and DCe with the most memory-intensive filtration option 'qLinkfilter'. We approached a memory limit of around 5.7 GB at a gene total of 7000. So it is anticipated that, if *qLinkfilter* is evoked, a gene expression matrix generally should undergo a gene filtration step beforehand so that the gene total is cut down to a few thousands or less.

Table 1. Execution time (in seconds) of five DCEA methods in handling different subsets of GSE3068

	1000	3000	5000	7000	8799
DCp	1	10	6	50	82
DCe	6	38	88	161	257
WGCNA	1.2	9.6	26.4	51	82
ASC	1.2	9.6	26.4	53	86.2
LRC	1	8.4	24.6	48.8	78

Different subsets, with a reduced number of rows, were taken from GSE3068 by favoring genes with top-ranked expression variability. The computing platform is a Linux system with five nodes, each having a dual quad-core Intel Xeon 2.33 GHZ CPU and a RAM of 16 GB. Execution time was averaged over five repetitive runs.

4 EXAMPLE

Three simulated datasets are included in the package for exploring the functions. For example, 'dataC' gives expression values of 1000 genes in 20 samples divided equally into two groups corresponding to two conditions. Since this dataset contains a moderate number of genes, the gene filtration step can be skipped. The link filtration procedure is wrapped as a sub-function in the DCEA functions, so one can specify the link filtration choice in the arguments of DCEA functions.

If the DCEA function DCe is called, one can get a resulted variable with four components. The gene names ranked by the dC measure (*P*-value) make up the first '$DCGs' component, while DCLs of different types are given in other three components.

Funding: Shanghai Institutes for Biological Sciences; Chinese Academy of Sciences (2008KIP207); the National '973' Basic Research Program (2006CB0D1203, 2006CB0D1205); the National Natural Science Foundation of China (30770497, 31000380); National Key Technologies R&D Program (2007AA02Z331 and 2009ZX10603).

Conflict of Interest: none declared.

REFERENCES

Choi,J.K. *et al.* (2005) Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, **21**, 4348–4355.

Mason,M.J. *et al.* (2009) Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, **10**, 327.

Reverter,A. *et al.* (2006) Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics*, **22**, 2396–2404.

van Nas,A. *et al.* (2009) Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. *Endocrinology*, **150**, 1235–1249.