CeTF: an R package to Coexpression for Transcription Factors using Regulatory Impact Factors (RIF) and Partial Correlation and Information (PCIT) analysis

Carlos Alberto Oliveira de Biagi Jr[1,2*], Ricardo Perecin Nociti[2,4], Breno Osvaldo Funicheli[2], Patrícia de Cássia Ruy[2,5], João Paulo Bianchi Ximenez[2], Wilson Araújo Silva Jr[1,2,3,*]

[1]Department of Genetics at Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil;

[2]Center for Cell-Based Therapy (CEPID/FAPESP); National Institute of Science and Technology in Stem Cell and Cell Therapy (INCTC/CNPq), Regional Blood Center of Ribeirão Preto, Ribeirão Preto, Brazil;

[3]Center for Integrative Systems Biology (CISBi) – NAP/USP, Ribeirão Preto, Brazil;

[4]Laboratory of Molecular Morphophysiology and Development, Department of Veterinary Medicine, Faculty of Animal Science and Food Engineering, University of São Paulo, Pirassununga, Brazil;

[5]Center for Medical Genomics, HCFMRP/USP, Ribeirão Preto, Brazil

Abstract

Summary: Finding meaningful gene-gene associations and the main Transcription Factors (TFs) in co-expression networks is one of the most important challenges in gene expression data mining. CeTF is an R package that integrates the Partial Correlation with Information Theory (PCIT) and Regulatory Impact Factors (RIF) algorithms applied to gene expression data from microarray, RNA-seq, or single-cell RNA-seq platforms. This approach allows identifying the transcription factors most likely to regulate a given network in different biological systems — for example, regulation of gene pathways in tumor stromal cells and tumor cells of the same tumor. This pipeline can be easily integrated into the high-throughput analysis.

Availability: CeTF is available as R package in Bioconductor (https://bioconductor.org/packages/CeTF), GitHub (https://github.com/cbiagii/CeTF) and as docker image (https://hub.docker.com/r/biagii/cetf). More information on how to use the package can be found in the Supplemental File 1.

Corresponding authors:

Wilson A Silva Jr (wilsonjr@usp.br) and Carlos AO Biagi Jr (biagi@usp.br)

Department of Genetics at the Ribeirão Preto Medical School, University of São Paulo
Av. Bandeirantes 3900
Monte Alegre
CEP: 14049-900, Ribeirão Preto, SP, Brazil.
Phone: +55 16-3315-3293

1. Introduction

Gene expression data analysis has become crucial to biological sciences, one of the most interesting forms of analyzing this type of data is the gene-to-gene network interaction analysis, aiming to highlight which gene interactions are the most relevant to the study. Despite the plethora of tools, new methods are needed to evaluate all possible interactions and their significance (Yu et al., 2013). In addition to identifying pairwise gene-to-gene interaction, finding the transcription factors (TFs) is of great interest, mainly because they can play an essential regulatory role (Farnham, 2009). Furthermore, integrating network generation with the identification of main TFs brings a deciding view of the data. In this article, we provide an R package that enables the performing of the Regulatory Impact Factors (RIF) and Partial Correlation with Information Theory (PCIT) analysis separately, or by applying the full pipeline. This package will be useful for creating a network from gene expression data identifying the most significant pairwise interactions and main TFs.

2. Implementation and main functions

CeTF is an implementation in R for PCIT (Reverter and Chan, 2008) and RIF (Reverter et al., 2010) algorithms, which initially were made in FORTRAN language. From these two algorithms, it was possible to integrate them in order to increase performance and results. Input data may come from microarray, RNA-seq, or single-cell RNA-seq. As seen earlier, the input data can be read counts or expression (TPM, FPKM, normalized values, and others). The main pipeline (Fig. 1) consists of the following steps.

Step 1: Data adjustment

If the input data is a count table, data will be converted to TPM by each column (x) as follows:

$$TPM = \frac{10^6 * x}{sum(x)} \qquad (1)$$

The mean for TPM values different than zero, and the mean values for each gene are used as a threshold to filter the genes. Genes with values above half of the previous averages will be considered for subsequent analyses. Then, the TPM data is normalized using:

$$Norm = \frac{log(x+1)}{log(2)} \qquad (2)$$

If the input already has normalized expression data (TPM, FPKM, etc), the only step will be the same filter for genes that consider half of the means.

Step 2: Differential Expression analysis

For differential analysis of gene expression, there are two options, the *Reverter* method (Reverter *et al.*, 2006) and DESeq2 (Love *et al.*, 2014). In both methods, two conditions are required (i.e., control *vs.* tumor samples). In the *Reverter* method, the mean between samples of each condition for each gene is calculated. Then, subtraction is made between the mean of one condition concerning the other conditions. The variance of the subtraction is performed, then is calculated the difference of expression using the following formula, where *s* is the result of subtraction and *var* is the variance:

$$diff = \frac{s - \frac{sum(s)}{length(s)}}{\sqrt{var}} \qquad (3)$$

The DESeq2 method applies the differential expression analysis based on the negative binomial distribution. Although both methods can be used on count data, it is strongly recommended to use only the *Reverter* method on expression input data.

Step 3: Regulatory Impact Factors (RIF) analysis

The RIF algorithm is well described in the original paper (Reverter *et al.*, 2010). This step aims to identify critical Transcription Factors calculating for each condition the co-expression correlation between the TFs and the Differentially Expressed (DE) genes (from Step 2). The result is RIF1 and RIF2 metrics that allow the identification of critical TFs. The RIF1 metric classifies the TFs as most differentially co-expressed with the highly abundant and highly DE genes, and the RIF2 metric classifies the TF with the most altered ability to act as predictors of the abundance of DE genes. A main TF is defined if:

$$\sqrt{RIF1^2}\ or\ \sqrt{RIF2^2} > 1.96\ (4)$$

Step 4: Partial Correlation and Information Theory (PCIT) analysis

The PCIT algorithm is also well described in the original paper from Reverter and Chan (Reverter and Chan, 2008). Moreover, it has been used for the reconstruction of Gene Co-expression

Networks (GCN). The GCN combines the concept of Partial Correlation coefficient with Information Theory to identify significant gene-to-gene associations defining edges in the reconstruction of the network. At this stage, the paired correlation of three genes is performed at the same time, thus making the inference of co-expressed genes. This approach is more sensitive than other methods and allows the detection of functionally validated gene-gene interactions. First, is calculated for every trio of genes x, y, and z the partial correlation coefficients:

$$r_{xy,z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \qquad (5)$$

And similarly, for $r_{xz,y}$ and $r_{yz,x}$. After that, for each trio of genes is calculated the tolerance level (ε) to be used as a threshold for capturing significant associations. The average ratio of partial to direct correlation is computed as follows:

$$\varepsilon = \frac{1}{3}\left(\frac{r_{xy,z}}{r_{xy}} + \frac{r_{xz,y}}{r_{xz}} + \frac{r_{yz,x}}{r_{yz}}\right) \qquad (6)$$

The association between the genes *x* and *y* is discarded if:

$$|r_{xz}| \leq |\varepsilon r_{xz}| \; and \; |r_{xy}| \leq |\varepsilon r_{yz}| \qquad (7)$$

Otherwise, the association is defined as significant, and the interaction between the genes *x* and *y* is used in the reconstruction of the GCN.

The final output includes the network with gene-gene and gene-TF interactions for both conditions, besides generating the main TFs identified in the network.

3. Additional functionalities

The CeTF package also has some additional features that includes plots to visualize the data distribution, the distribution of differentially expressed genes/TFs that shows the average expression (in log2) by the difference of expression, and the network for both conditions. It is also possible to perform the grouping of ontologies (Carbon *et al.*, 2019) without statistical inference and functional enrichment for several databases with statistical inference of any organism. Finally, the network that integrates the genes, TFs and pathways is generated.

3. Conclusions

CeTF is a tool that assists the identification of meaningful gene-gene associations and the main TFs in co-expression networks. It offers functions for a complete and customizable workflow from count or expression data to networks and visualizations in the form of a freely available R package. We expect that CeTF will be widely used by the genomics and transcriptomics community and scientists that work with high-throughput data that aims to understand how main TFs are working in a co-expression network and what are the pathways involved in this context.

Funding

This work was financed by the Coordination for the Improvement of Higher Education Personnel (CAPES), grant #88882.378695/2019-01; São Paulo Research Foundation (FAPESP), #2013/08135-2, and by Research Support of the University of Sao Paulo, CISBi-NAP/USP Grant #12.1.25441.01.2.

Conflict of Interest: none declared.

References

Carbon,S. *et al.* (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.

Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 1–21.

Paul Shannon,1 *et al.* (1971) Cytoscape: A Software Environment for Integrated Models. *Genome Res.*, **13**, 426.

Reverter,A. *et al.* (2010) Regulatory impact factors: Unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics*, **26**, 896–904.

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

Reverter,A. and Chan,E.K.F. (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, **24**, 2491–2497.

Weinstein,J.N. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Yu,D. *et al.* (2013) Review of Biological Network Data and Its Applications. *Genomics Inform.*, **11**, 200.

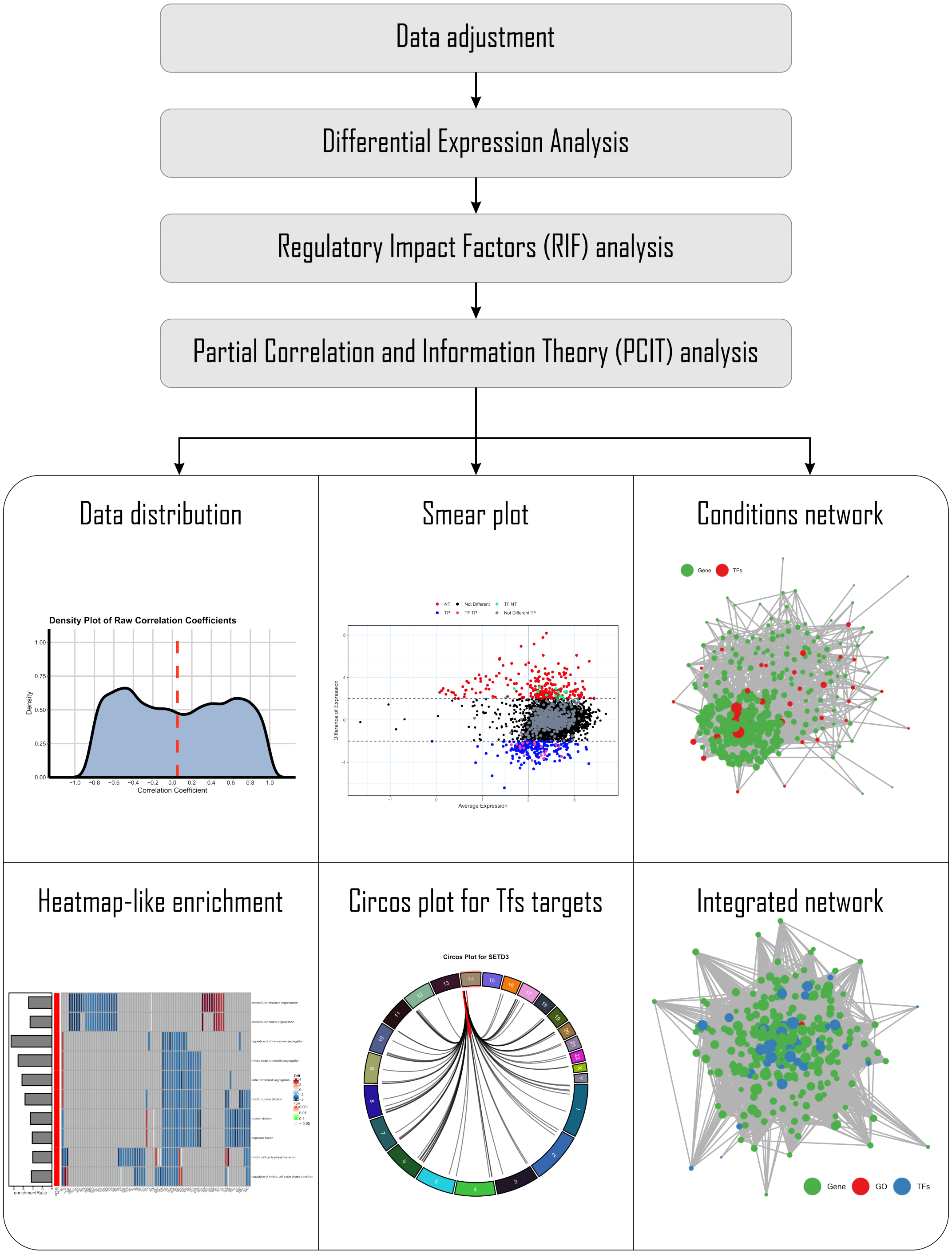

Fig. 1: Schematic of a CeTF workflow. The main functions of CeTF and their logical order are illustrated. Grey boxes signify the mandatory steps for the main analysis. The white box signifies the optional outputs from CeTF package, i.e. data distribution, smear plot, conditions network, heatmap-like enrichment, circus plot for TFs targets, integrated network, etc.