Bioinformatics Advance Access published August 19, 2016

Systems Biology

DEsubs: an R package for flexible identification of differentially expressed subpathways using RNA-seq experiments.

Aristidis G. Vrahatis[1,2], Panos Balomenos[1,2], Athanasios K. Tsakalidis[1] and Anastasios Bezerianos[2,3,*]

[1]Department of Computer Engineering and Informatics, University of Patras, Patras, 26500, GR.

[2]Department of Medicine, University of Patras, Patras, 26500, GR.

[3]SINAPSE Institute, Center of Life Sciences, National University of Singapore, Singapore 117456

*To whom correspondence should be addressed.

Associate Editor: Dr. Jonathan Wren

Abstract

Summary: DEsubs is a network-based systems biology R package that extracts disease-perturbed subpathways within a pathway network as recorded by RNA-seq experiments. It contains an extensive and customized framework with a broad range of operation modes at all stages of the subpathway analysis, enabling so a case-specific approach. The operation modes include pathway network construction and processing, subpathway extraction, visualization and enrichment analysis with regard to various biological and pharmacological features. Its capabilities render it as a tool-guide for both the modeler and experimentalist for the identification of more robust systems-level drug targets and biomarkers for complex diseases.

Availability and implementation: DEsubs is implemented as an R package following Bioconductor guidelines (permanently available in URL: http://biosignal.med.upatras.gr/wordpress/desubs).

Contact: tassos.bezerianos@nus.edu.sg

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

1 Introduction

In the last years, systems-level network-based approaches have gained ground in the research field of systems biology (Leung et al., 2013). Towards this direction the research community has been focusing on the analysis of subpathways, namely local gene sub-regions within a pathway network (Chen *et al.,* 2011). Approaches in this direction are essential for identifying more accurate system-level drug targets and biomarkers for complex diseases by monitoring the local functions of networks perturbed by diseases (Barabási *et al*., 2011). Hence, several subpathway-based tools have been developed recently focusing though predominantly on differentially expressed (DE) genes (Haynes *et al.,* 2013; Martini *et al.,* 2012; Nam *et al*., 2014; Li *et al*., 2009; Li *et al*., 2012; Li *et al*., 2013; Liu et al., 2013; Judeh *et al*., 2013; Zhang *et al*., 2014; Vrahatis *et al*., 2016).

Meanwhile, since the amount of RNA-seq transcriptome studies is increasing rapidly year by year, there is high demand for tailored differential expression (DE) analysis tools. Although most DE tools are crucial components in pathway analysis workflows, there is a gap in addressing this issue through an extensive and customized subpathway-based approach using DE analysis for RNA-seq data.

Towards this orientation, we developed an R package, called DEsubs, which extracts differentially expressed subpathways based on RNA-seq expression profiles enabling so a customized analysis to the problem under investigation through numerous operation modes. DEsubs offers a systematic way to uncover the local perturbations within a net-

work that could lead to human diseases through differentially expressed subpathways.

2 Methods

2.1 Pathway network construction

All available organism-specific KEGG pathway maps are converted to a pathway network, maintaining both their topology and information flow (Vrahatis *et al*., 2016). Organism level networks are constructed by joining all related processed graphs based on the user-input list of genes. DEsubs operates for six well-known organisms three pathway types and several gene label systems (for details see package vignette).

2.2 Pathway network processing

Next, the RNA-seq data are mapped onto the nodes and edges of the pathway network. Two pruning rules are used to isolate robust and validated interactions (edges) among statistically significant differentially expressed genes (DEGs). $NodeRule$ selects the core set of genes with statistically significant differential expression and $EdgeRule$ further prunes the interactions among these genes according to prior biological evidence.

More specifically, DEGs are identified using any of the eight well-established DE tools (edgeR, DESeq, NBPSeq, TSPM, voom, vst, SAM-seq, EBSeq) by considering the FDR-adjusted P-value of each gene ($Qvalue$) based on the work of (Soneson *et al*., 2013). For a graph $G(V, E)$ with V nodes and E edges, the $NodeRule$ is as follows: $NodeRule$: $Qvalue(i) < threshold,\ \forall i \in V(G)$. Among the edges of

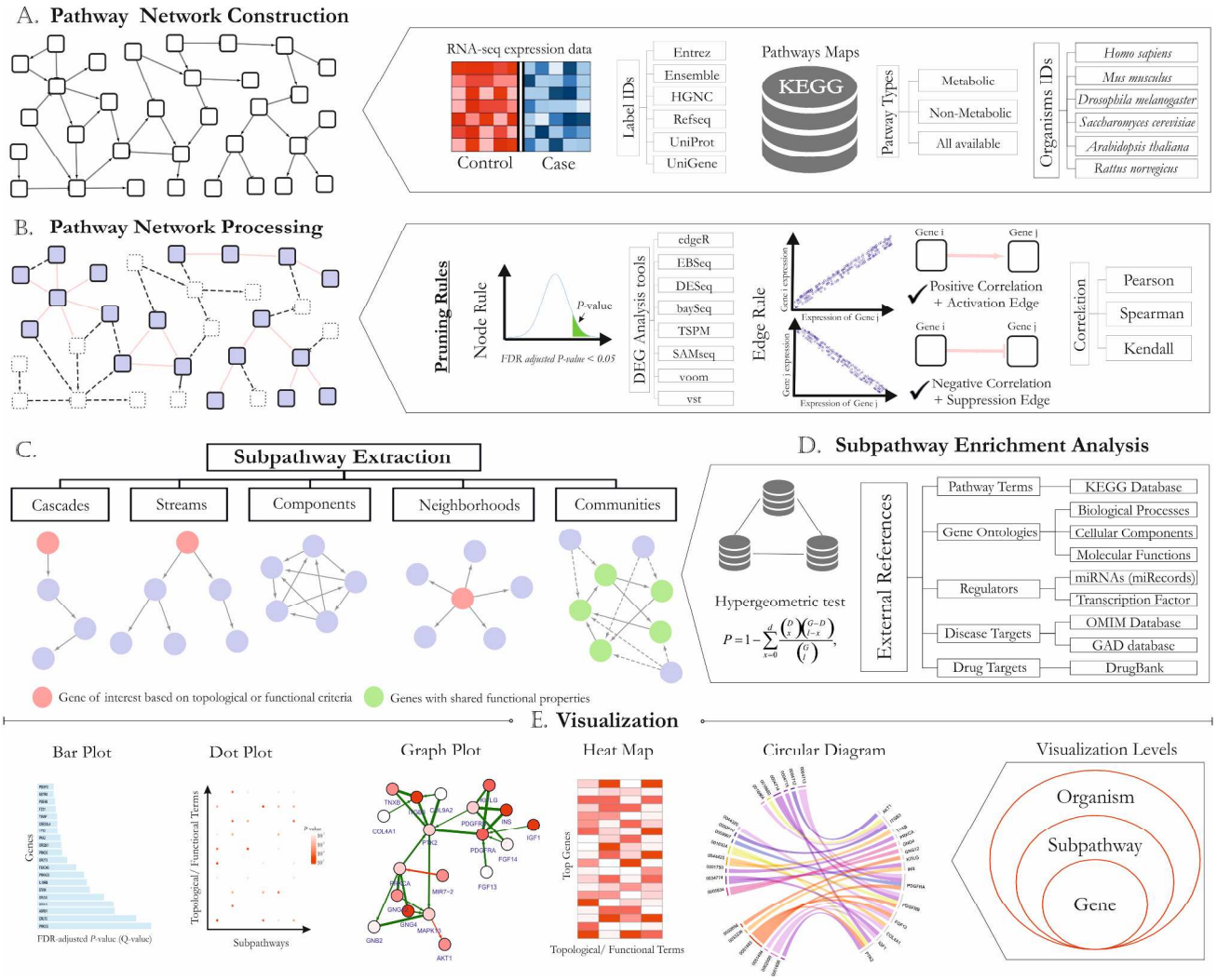

Fig. 1. DEsubs overview. (A) User inputs the RNA-seq control-case expression dataset and selects the organism, the gene label system and the pathway type and subsequently the appropriate pathway network is constructed. (B) The pathway network processing is based on two pruning rules by retaining the statistically significant DEGs (*NodeRule*) and interactions which are in accordance with the flow of information imposed by KEGG pathway maps (*EdgeRule*). Pruning rules are applied through several user-defined DE tools and correlation measures. (C) Subpathways are extracted from the processed pathway network based on a broad range of sub-structures, resulting in 124 extraction types in total. (D) A hypergeometric test is used to estimate the subpathway's associations with various biological and pharmacological features. (E) The enriched associations of a subpathway to each feature are illustrated using several visualization schemes. Furthermore, by zooming in and out of subpathways, DEsubs provides a customized view of the resulting subpathways from the gene up to organism level. Heat maps, circular diagrams, dot plots, bar plots and graph plots are supported.

isolated genes, we highlight those where the expression profiles of the adjacent genes comply with prior biological knowledge. These are edges with highly positive or negative correlated adjacent gene expression profiles and are considered a priori as edges with activating ($reg = 1$) or suppressing ($reg = -1$) regulation role based on KEGG pathway maps (Vrahatis *et al*., 2016). The correlation is estimated based on cor value which uses any of the three well-established correlation measures (*Pearson*, *Spearman*, *Kendall*). For a pair of interacting gene nodes i, j the rule is formulated as follows: $EdgeRule: cor(i,j) * reg(i,j) > threshold,\ \forall\ (i,j) \in E(G)$. The thresholds are user-defined and the default values are 0.05 and 0.7 respectively. For more details, see package vignette.

2.3 Subpathway extraction

The processed pathway network is subsequently frittered based on five main topological structures. We extract (i) group of genes sharing common roles or properties within the graph (communities), (ii) strongly connected genes (components), (iii) gene streams, (iv) cascades and (v) neighborhoods. The three latter types are generated starting from a gene of interest with specific topological or functional properties within the network. From this point on, subpathways are generated following either forward or backward propagation. Thus, the user can observe the local perturbations within the network starting from (or ending to) different points of interest through various topological schemes. As a result, a total of 124 different extraction types are supported thus providing the user with a broad repertoire of the sub-structures existing within the pathways. To assess the performance of DEsubs, we examine all operation modes based on synthetic benchmark data and provide guidelines for the user to select the appropriate parameter setting. For detailed description in the subpathway extraction stage, see package vignette.

2.4 Subpathway enrichment analysis

The extracted subpathways are further examined for enrichment in various biological and pharmacological features. A hypergeometric test is used to estimate the subpathway associations with (i) pathway terms, (ii) gene ontology (GO) terms of molecular function, biological processes and cellular components, (iii) disease terms, (iv) drug substances, (v) microRNA targets and (vi) transcription factors. The aforementioned associations are extracted from seven databases (KEGG, GO, miRecords, Transfac, Jaspar, GAD, OMIM, DrugBank) using the approach of (Barneh *et al*., 2015; Chen *et al*., 2013; Li *et al*., 2011; Vrahatis *et al.,* 2015). For details see package vignette.

2.5 Visualization

The enriched associations of a subpathway to each feature are illustrated through circular diagrams. Furthermore, by zooming in and out of subpathways, DEsubs provides a customizable view of the resulting subpathways from the gene up to organism level. Feature-rich genes are imprinted in bar plots and heat maps. Also, the resulted subpathways along with their associations to biological and pharmacological features are summarized in dot plots providing a comprehensive illustration of each experiment (see package vignette).

3 Conclusion

DEsubs R package is a flexible, computationally fast and efficient tool (see supplementary file) for exploring disease-associated gene perturbations contextualized by subpathways in a graph-theoretic framework. The numerous operation modes provide the user a highly customizable setup, rendering the package as a guide for the systematic exploration of the local subpathways within a pathway network perturbed by disease.

Acknowledgements

The authors thank Dr. Andrei Dragomir and Dr. Konstantina Dimitrakopoulou for comments that greatly improved the manuscript.

Conflict of Interest: none declared.

References

Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, **12(1)**, 56-68.

Barneh,F. *et al*. (2015) Updates on drug–target network; facilitating polypharmacology and data integration by growth of DrugBank database. *Briefings in bioinformatics*, bbv094.

Chen,X. *et al.* (2011) A sub-pathway-based approach for identifying drug response principal network. *Bioinformatics*, **27(5)**, 649-654.

Chen,E.Y. *et al*. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, *BMC bioinformatics*, **14(1)**, 128

Haynes,W.A. *et al*. (2013) Differential expression analysis for pathways. *PLoS Comput. Biol*, **9(3)**, e1002967.

Judeh,T. *et al.* (2013) TEAK: topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic Acids Res.*, **41**, 1425-37.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, **8(2)**, e1002375.

Leung,E.L. *et al*. (2013). Network-based drug discovery by integrating systems biology and computational technologies. *Briefings in bioinformatics*, ***14*(4)**, 491-505

Li,X. *et al.* (2011) The implications of relationships between human diseases and metabolic subpathways. *PloS one*, **6(6)**.

Li,C. *et al.* (2013) Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res*., **41**, e101.

Li,C. *et al.* (2009) SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res*., **37**, e131.

Li,C. *et al.* (2012) Identifying disease related sub-pathways for analysis of genome-wide association studies. *Gene*, **503(1)**, 101-109.

Liu,W, *et al*. (2013) Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* **29(17)**, 2169-2177.

Martini,P. *et al.* (2014) timeClip: pathway analysis for time course data without replicates. *BMC bioinformatics*, **15(5)**, 1-10.

Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, **14(1)**, 1.

Vrahatis,A.G. *et al.* (2016) CHRONOS: A time-varying method for microRNA-mediated sub-pathway enrichment analysis. *Bioinformatics*, **32(6)**.

Zhang,C. *et al.* (2014) Identification of miRNA-mediated core gene module for glioma patient prediction by integrating high-throughput miRNA, mRNA expression and pathway structure. *PloS one*, **9(5)**, e96908.