Bioinformatics, 31(10), 2015, 1683–1685
doi: 10.1093/bioinformatics/btu853
Advance Access Publication Date: 9 January 2015
Applications Note

Systems biology

ENViz: a Cytoscape App for integrated statistical analysis and visualization of sample-matched data with multiple data types

Israel Steinfeld[1,2,*], Roy Navon[1,†], Michael L. Creech[3], Zohar Yakhini[1,2] and Anya Tsalenko[4]

[1]Agilent Laboratories, Tel-Aviv, Israel, [2]Technion – Israel Institute of Technology, Haifa, Israel, [3]Blue Oak Software and [4]Agilent Laboratories, Santa Clara, CA, USA

*To whom correspondence should be addressed.
†Present address: MeMed Diagnostics Ltd., Haifa, Israel

Associate Editor: Alfonso Valencia

Received on August 14, 2014; revised on November 27, 2014; accepted on December 23, 2014

Abstract

Summary: ENViz (Enrichment Analysis and Visualization) is a Cytoscape app that performs joint enrichment analysis of two types of sample matched datasets in the context of systematic annotations. Such datasets may be gene expression or any other high-throughput data collected in the same set of samples. The enrichment analysis is done in the context of pathway information, gene ontology or any custom annotation of the data. The results of the analysis consist of significant associations between profiled elements of one of the datasets to the annotation terms (e.g. miR-19 was associated to the cell-cycle process in breast cancer samples). The results of the enrichment analysis are visualized as an interactive Cytoscape network.

Availability and implementation: ENViz is publically available in the Cytoscape App Store (http://apps.cytoscape.org/apps/enviz). For additional information please visit the tool website: http://www.agilent.com/labs/research/compbio/enviz/

Contact: israel_steinfeld@agilent.com

1 Introduction

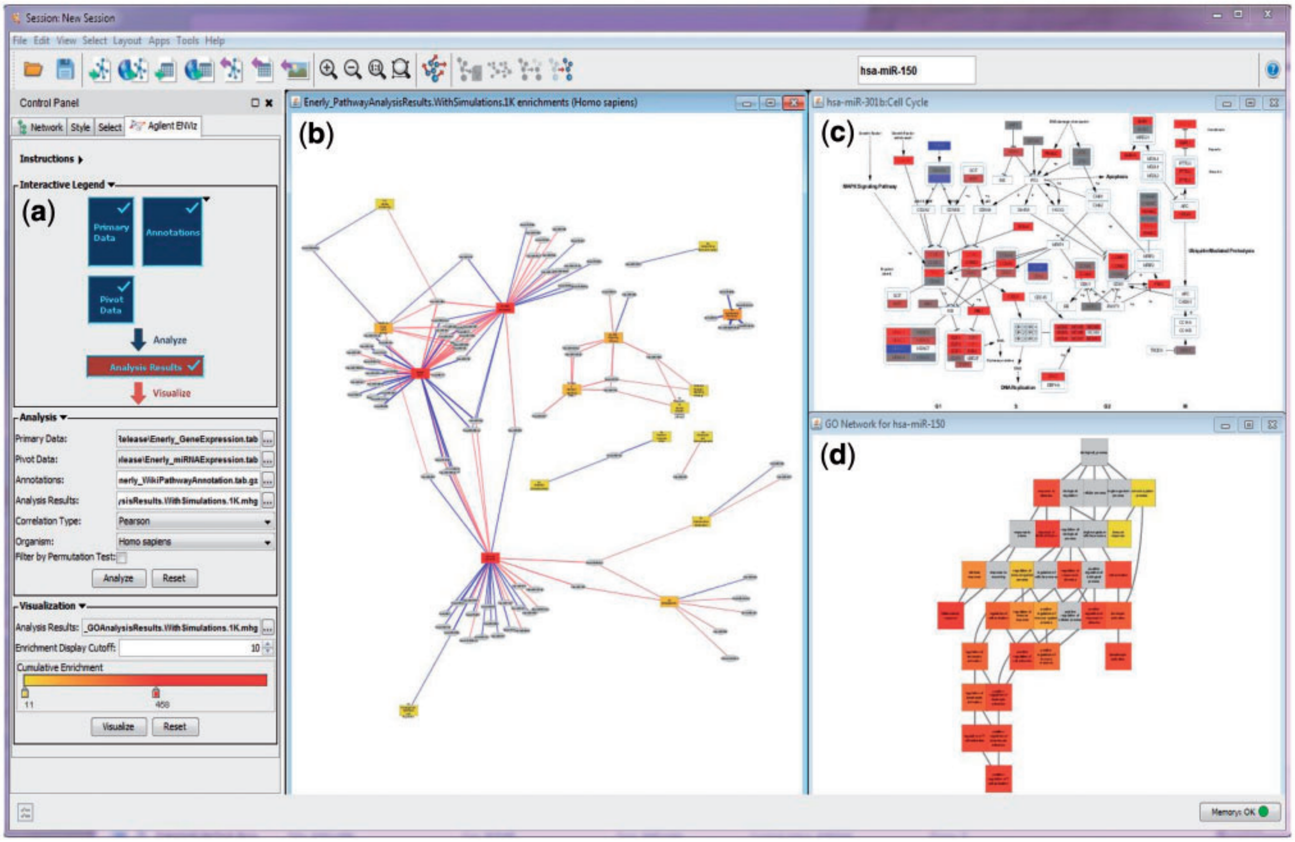
The recent emergence of novel high-throughput technologies enables the quantification of different types of biological features in a genome-wide scale (e.g. mRNA expression levels, miRNA expression levels and DNA copy numbers). In parallel with these technologies, various methodologies have been developed to handle integrated analysis of functional genomics data, mainly by studying the transcriptional programs and global organization of biological processes. Still, only a few tools support routine joint analysis of sample cohorts with multiple genomic measurement results (Gomez-Cabrero *et al.*, 2014). Even fewer tools provide the visualization strength of Cytoscape in this context (Cline *et al.*, 2007; Bindea *et al.*, 2013; Xia *et al.*, 2010).

The ENViz approach to integrated data analysis uses the power of enrichment statistics combined with genomic annotation databases to statistically assign relevant function annotations to explored profiled elements. It thus provides a better understating of the relationship between different molecular entities in cells or organisms. Visualizing ENViz results as Cytoscape networks provide compact structured representation of enrichment results.

Even though the development of ENViz was motivated by available modern biological measurements, joint analysis of two sample matched datasets and systematic annotations may be applied to other similarly structured.

2 Tool description

ENViz follows an enrichment analysis approach, driven by three input matrices: (i) the primary data matrix (e.g. genes expression measurement across a set of samples), (ii) the annotation matrix



Primary Data: Release\Enerly_GeneExpression.tab

Pivot Data: elease\Enerly_miRNAExpression.tab

Annotations: nerly_WikiPathwayAnnotation.tab.gz

Analysis Results: ysisResults.WithSimulations.1K.mhg

Correlation Type: Pearson

Organism: Homo sapiens

Filter by Permutation Test: □

Analysis Results: _GOAnalysisResults.WithSimulations.1K.mhg

Enrichment Display Cutoff: 10

Fig. 1. Cytoscape session with ENViz application running. **(a)** Interactive legend, Analysis, and Visualization control panels. The Interactive legend shows a schematic of the analysis and the overview of the data loaded into Cytoscape for ENViz analysis. The Analysis panel controls data input and analysis parameter settings. The Visualization panel controls the significance threshold for the enrichment network generation and the color coding of the annotation categories based on enrichment scores. **(b)** Bi-partite network for enrichment analysis of breast cancer data. Nodes in this network correspond to WikiPathways (colored yellow->red) and miRNAs (gray), and edges represent significant enrichments of genes in the pathway correlated (red) or anti-correlated (blue) to the miRNA. **(c)** Cell Cycle WikiPathway with genes color coded according to their correlation to mir-301 b. **(d)** GO terms enriched in genes correlated to miR-150

providing binary annotation on each of the primary data matrix elements [e.g. pathway or gene ontology (GO) annotation] and (iii) the pivot data matrix providing information on the same set of samples of the primary data matrix (e.g. miRNAs expression measurement) (Fig. 1a).

For each pivot data element ENViz performs these steps:

- Compute the correlation to each element of primary data.
- Rank primary data elements based on above correlations.
- Compute the statistical enrichment of annotated elements (e.g. pathways) in the top of above ranked list based on a minimum-hypergeometric (mHG) statistics.

Details of mHG statistics are explained in (Eden *et al.*, 2007, 2009). Briefly, given a ranked binary annotation vector we compute enrichment of this annotation in the top k ranked elements based on the hypergeometric statistics, where k is selected to optimize this enrichment. Finally, the mHG score [$-\log$(mHG P-value)] for the pivot-annotation association is reported. The calculated significance level is valid for every individual pivot annotation pair, but is not corrected for the number of pairs tested.

Significant results are represented in Cytoscape as an enrichment network—a bipartite graph with nodes corresponding to pivot and annotation elements, and edges corresponding to significant pivot-annotation associations, where significance threshold is user defined (Fig. 1b). In addition ENViz supports extended visualization for association to:

- The WikiPathways (Kelder *et al.*, 2012) database (Fig. 1c), where top correlated genes (from the primary data) are visually overlaid on the relevant pathway.
- The GO (Ashburner *et al.*, 2000) database (Fig. 1d), where all GO term associated with a particular pivot element can be visually overlaid on the GO graph, which may point to functionally relevant mechanisms.

To address multiple testing issues, as well as some potential dependencies between primary data elements, ENViz implements filtering by permutation correction. For each permutation, samples in the pivot data matrix are randomly shuffled, and an enrichment score S_{rand} is calculated for each pivot with at least one significant association, as defined by the user. If, for a given pivot-annotation pair with enrichment score S, we observe $S_{\mathrm{rand}} \geq S$ more than a user-defined number of times across all permutations, then this pivot-annotation element pair is considered not significant. For pivot-annotation pairs that survive this permutation test filtering, the original mHG score is reported as the enrichment score.

More details can be found in the user tutorial (http://www.agilent.com/labs/research/compbio/enviz/ENVizUserTutorial.pdf).

3 Example

An example dataset, based on data published in (Enerly *et al.*, 2011) and formatted for ENViz, can be downloaded from http://www.agilent.com/labs/research/compbio/enviz/data.html. This dataset consists of 100 breast tumor samples with various characteristics. Primary data is gene expression profiles from Agilent microarray experiments, pivot data is Agilent microarray-based miRNA profiles, and the annotation matrix is taken from WikiPathways and GO database. As shown in Figure 1 using ENViz we identify a significant association between miR-301 b and the cell-cycle pathway.

On a standard laptop (i7 chip), the analysis of the example data with default parameters and the WikiPathways annotations takes ~1 min; analysis with GO annotation takes ~25 min.

Acknowledgements

We thank Allan Kuchinsky who identified the potential for weaving a joint data analysis approach into Cytoscape. Even though Allan was constantly fighting cancer and its complications, he led our team with great enthusiasm to cross countless obstacles and make ENViz a reality. This work is dedicated to the memory of Allan Kuchinsky, a Cytoscape enthusiast and pioneer.

Conflict of Interest: I am an employee and hold stock of Agilent Technologies, the manufacturer of genomic microassays and library preparation assays upstream of next generation sequencing, and am currently conducting research sponsored by the company as part of my employment. Enviz analysis supports all relevant data, independent of the measurement technology provider.

References

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

Bindea,G. *et al.* (2013) CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*, **29**, 661–663.

Cline,M.S. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.

Eden,E. *et al.* (2007) Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput. Biol.*, **3**, e39.

Eden,E. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Enerly,E. *et al.* (2011) miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PLoS One*, **6**, e16915.

Gomez-Cabrero,D. *et al.* (2014) Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, **8**, I1.

Kelder,T. *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.

Xia,T. *et al.* (2010) OmicsAnalyzer: a Cytoscape plug-in suite for modeling omics data. *Bioinformatics*, **26**, 2995–2996.