DOI 10.18699/SBB-2022-049

GESECA: Gene Set Co-regulation Analysis

Sukhov V.[1], Artyomov M.[2], Sergushichev A.[1]*
[1] *ITMO University, St. Petersburg, Russia*
[2] *Washington University in St. Louis, School of Medicine, St. Louis, MO, USA*
* *alserg@itmo.ru*

Key words: pathway enrichment, gene signature, gene expression omnibus

Pathway enrichment is a standard approach to interpret transcriptional data. Commonly, in this approach, two conditions of interest are considered, for which differential gene expression is calculated, followed by one of the existing tools to find pathways that have coordinated positive or negative regulation of gene expression levels. However, currently the experiments contain many conditions and samples, ultimately, single-cell RNA-sequencing datasets contain thousands of individual cells measured without an inherent annotation for which cell corresponds to which cell type. For such datasets, classical pathway enrichment methods cannot be directly applied.

Here we propose GESECA (Gene Set Co-regulation Analysis): a method to find co-regulated gene sets in transcriptomic data without requirement for an explicit sample annotation. The method uses the percentage of variance explained by the queried gene set as an enrichment statistic: for gene sets with high pairwise correlation this value is high, compared to gene sets with uncorrelated genes. Similarly to the FGSEA approach, the statistical significance is assessed with a Multilevel Monte Carlo scheme, which allows empirical estimation of arbitrarily low P-values. Together, this establishes a method that can find co-regulated pathways in the data without explicit knowledge of the dataset annotation, and thus applicable both to bulk RNA-seq datasets with many conditions, as well as to scRNA-seq datasets.

Development of GESECA opens an opportunity to systematically re-analyze/re-utilize *all* of the published transcriptional profiling data. Specifically, we developed a search engine https://ctlab.itmo.ru/geseca/ that allows to find gene expression datasets in which the query gene set is co-regulated. For this purpose we preprocessed a collection of more than 50000 gene expression datasets from Gene Expression Database (both RNA-seq and microarray). We have benchmarked the search engine against manually annotated signatures from the CRowd Extracted Expression of Different Signatures (CREEDS) web portal, which showed high dataset recovery rate. Furthermore, we have developed a keyword extraction algorithm for the search results, that allowed us to validate the search quality on the collection of MSigDB Hallmark gene sets.

To conclude, here we developed GESECA method to find co-regulated gene sets in transcriptomic data that provides (1) a quick and straightforward approach for annotation-free gene set enrichment, paving a path to (2) unbiased search of biologically relevant datasets based on a specific query gene set.

Acknowledgements: This research was supported by the Ministry of Science and Higher Education of the Russian Federation, Priority 2030 Federal Academic Leadership Program.