A NOVEL APPROACH FOR DETERMINING THE OPTIMAL NUMBER OF INDEPENDENT COMPONENTS FOR REPRODUCIBLE CANCER TRANSCRIPTOMES DATA ANALYSIS

U. Kairov [1], L. Cantini [2], A. Greco [2], A. Molkenov [1], U. Czerwinska [2], E. Barillot [2], A.Zinovyev [2]

[1] *Laboratory of bioinformatics and computational systems biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University (Astana, Kazakhstan)*
[2] *Institut Curie, INSERM U900, PSL Research University, Mines ParisTech (Paris, France)*

Keywords: Transcriptome, Independent Component Analysis, reproducibility, cancer

Introduction: Independent Component Analysis (ICA) is a method that models gene expression data as an action of a set of statistically independent hidden factors. The output of ICA depends on a fundamental parameter: the number of components (factors) to compute. The optimal choice of this parameter, related to determining the effective data dimension, remains an open question in the application of blind source separation techniques to transcriptomic data.

Methods: fastICA algorithm accompanied by the icasso package have been used to improve the independent components estimation and to rank the components based on their stability. ICA was applied to each transcriptomic dataset separately. For each analysed transcriptomic dataset, we computed M independent components (ICs), using pow3 nonlinearity and symmetrical approach to the decomposition. In our analysis, we used Docker with packaged compiled MATLAB code for fastICA together with MATLAB Runtime environment, which can be readily used in other applications and does not require MATLAB installed.

Results: Here we address the question of optimizing the number of statistically independent components in the analysis of transcriptomic data for reproducibility of the components in multiple runs of ICA (within the same or within varying effective dimensions) and in multiple independent datasets. To this end, we introduce ranking of independent components based on their stability in multiple ICA computation runs and define a distinguished number of components (Most Stable Transcriptome Dimension, MSTD) corresponding to the point of the qualitative change of the stability profile.

Conclusions: We propose a new approach of ICA application to cancer transcriptomics data with a possibility of prioritizing components with respect to their reproducibility that strengthens the biological interpretation.