INDEPENDENT COMPONENT ANALYSIS OF CANCER TRANSCRIPTOMES: OPTIMIZATION OF PARAMETERS AND IMPROVEMENT OF INTERPRETABILITY

M. Chepeleva [1,2], A. Kakoichankava[3], M. M. Yatskou[1] and P. V. Nazarov[2]

[1]Belarusian State University, Minsk, Belarus
[2]Luxembourg Institute of Health, Strassen, Luxembourg
[3]Vitebsk State Medical University, Vitebsk, Belarus
E-mail: maryna.chepeleva@gmail.com

Here we estimated the optimal parameters of the independent component analysis method regarding the tasks of identification of glioblastoma and pancreatic cancer subtypes, prediction of patient survival and characterization of active biological processes. Analysis of deconvolution results of bulk and single-cell data allows sharing annotation between highly correlated components and improving interpretability of the results.

Key words: transcriptomics; RNA-seq; independent component analysis; single-cell; cancer; survival.

Introduction. Independent component analysis (ICA) allows decomposing heterogeneous transcriptomic data and extracting signals that correspond either to relevant biological processes or to technical biases [1]. Weights of independent components can be used as features for downstream analysis allowing to improve quality compared to gene-based methods and to carry out interpretation of meaningful signals. Here we investigated the effect of the number of independent components on quality of prediction and extracted biological content. Furthermore, by simultaneous analysis of single cell RNA-seq (scRNA-seq) data from tumor and normal tissues, we improved our understanding of the biology of the components linked to survival and patient classes.

Method. The parallel consensus ICA implemented in consICA [2] was applied to RNA-seq gene expression data of glioblastoma (GBM) [3] and pancreatic adenocarcinoma (PAAD) [4] bulk samples, as well as to two scRNA-seq datasets originated from a cancer glioblastoma cell line [5] and a normal pancreas [6] (Fig. 1A).

Weight matrix and the most significant differentially expressed genes were used as input features to random forest classifiers. Final balanced accuracy of multi-class classification was calculated as an average of balanced accuracies for each class. We also carried out survival prediction with Cox regression using the weights of independent components with the following risk score for each *j*-th patient:

$$RS_j = \sum_{i=1}^{k} H_i \, R_i^2 M_{i,j}^*, \qquad (1)$$

where H_i is log-hazard ratios for the components significantly (FDR < 0.05) linked to survival on training data (5-fold cross-validation was used) and 0 for other, R_i^2 is the stability of *i*-th component (mean squared correlation between runs of ICA), $M_{i,}^*$ is a standardized row of the weight matrix *M*, (Fig. 1B), as in [2].

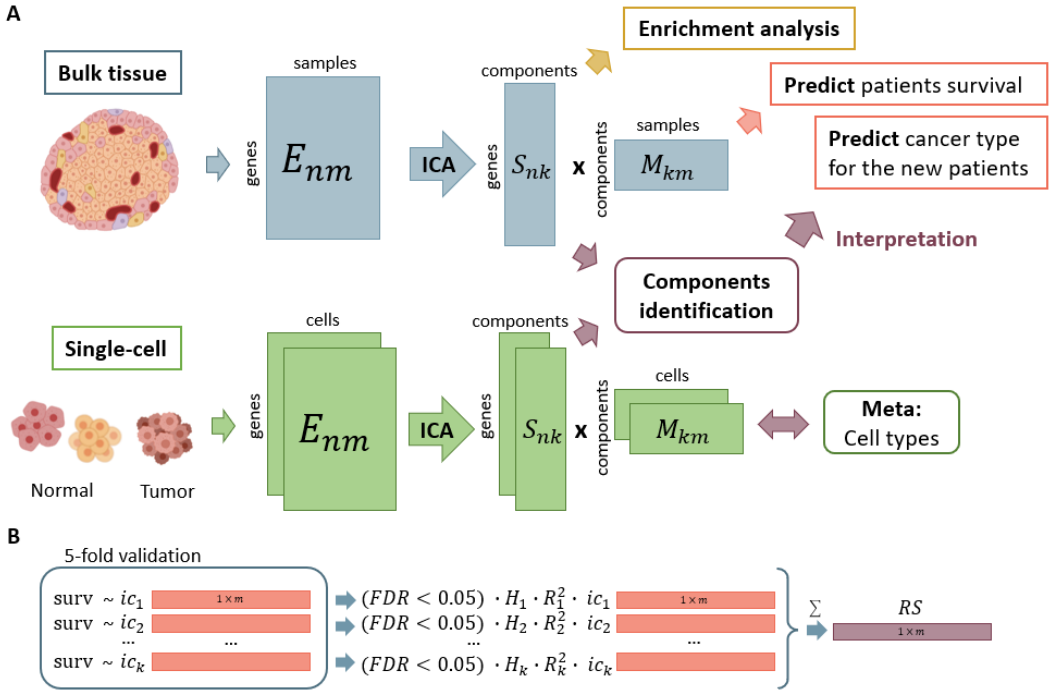

Fig.1. (A) ICA decomposes gene expression matrix into meaningful signals (metagenes, S) and weights (weight matrix, M). Biological processes can be found by analysis of S, while M could be linked to patient cancer groups and patient survival. Matching S-matrices of bulk and single-cell deconvolution allows annotating bulk components with specific cells signals. (B) Cox regression models were built for each independent component. Then components with FDR < 0.05 were used to calculate the risk score (Eq. 1).

Then we calculated correlation between *S* matrices of ICA decompositions for bulk tissues and single-cell data to identify components representing similar signals (Fig.1). Some components were linked to specific cells types according to distribution of the weights of each component in respect to the cell types.

Results. We validated multi-class models to predict cancer subtypes and calculated average multi-class combined balanced accuracy for several experiments (Fig. 2A). Predicted cancer subtypes were classical/mesenchymal/neural/proneural (for GMB), adex/immunogenic/progenitor/squamous (for PAAD). Comparing to the models trained on the 100 most significantly differentially expressed genes (horizontal line), the top ICA-based features improved the results.

To identify the optimal number of the independent components for the survival analysis, we performed 5-fold cross-validations of ICA-based cox-regression models. As it can be seen from Fig. 2B, the optimal number of components depends on the dataset: 80 components were required for GBM, whereas only 20 were enough for PAAD.

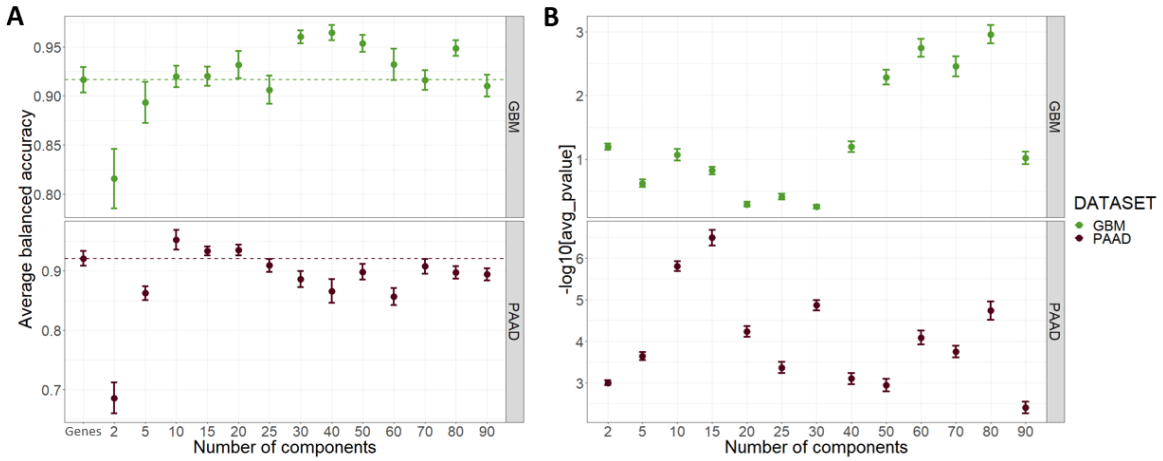

Fig.2. (A) Average balanced accuracy depending on the number of independent components for random forest cancer subtypes prediction. Horizontal line corresponds to the model trained on the most significant differentially expressed genes. (B) Average p-value of ICA-based Cox regression models, trained on the components with FDR < 0.05, depending on the number of independent components.

Matching deconvolution results (matrices *S*) obtained on bulk cancer samples and single-cell data, we found highly correlated components. In the case, when 7 components were extracted from cancer cell lines (single-cell) and 100 from GBM data (bulk), the two most correlated components were linked to the cell cycle and two others represented technical effects (overall gene expression). Single-cell data required 5 or more components to extract the cell cycle. In the more complex bulk sample dataset, the cell cycle was detected with at least 9 components.

Next, the results of ICA deconvolution of scRNA-seq from normal pancreas samples were matched with the ICA results on bulk PAAD samples. Required number of components to extract signals of cell subtypes were: acinar cells and pancreatic ductal cells – 4 components, mesenchymal cells – 9, pancreatic A cells – 14. Technical effects were isolated when 11 or more independent components were used. All these signals were detected in the analysis of bulk PAAD data with 20 components.

Conclusions. Analysis of deconvolution results of bulk and single-cell data allows sharing annotation between highly correlated components of bulk and single-cell experiments. This annotation can be used to improve interpretation of predictive models. Two the most correlated components from tumor single

cell data represented cell cycle. Finally, we estimated the optimal number of components for extracting individual cell types in pancreas scRNA-seq data.

REFERENCES

1. Sompairac N., Nazarov P. V., Czerwinska U. et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets // Int. J. Mol. Sci. 2019. Vol. 20, № 18. P. 4414–4441.
2. Nazarov P. V., Wienecke-Baldacchino A. K., Zinovyev A. et al. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients // BMC Med. Genomics. 2019. Vol. 12, №1, P. 132–149.
3. McLendon R., Friedman A., Bigner D. et al. Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways // Nature. 2008. Vol. 455, № 7216. P. 1061–1068.
4. Bailey P., Chang D. K., Nones K. et al. Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer // Nature. 2016. Vol. 531, № 7592. P. 47–52.
5. Dirkse A., Golebiewska A., Buder T. et al. Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment // Nat Commun. 2019. Vol. 10, № 1787.
6. Enge M., Arda H. E., Mignardi M. et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns // Cell. 2017. Vol. 171, №2. P. 321–330.

СОСТАВ И ВАРИАБЕЛЬНОСТЬ ГЕНОВ GST И UDT В ГЕНОМЕ APHIS CRACCIVORA

Р. С. Шулинский, Н. В. Воронова

Белорусский государственный университет,
биологический факультет, Минск, Беларусь
E-mail: bio.shulinsk@bsu.by, nvoronova@bsu.by

В работе представлены результаты аннотации генов GST и UDT в геноме *Aphis craccivora*, также проведен сравнительный анализ генного состава с другими геномами тлей. Показано, что генный состав семейств UDT и GST в геноме *A. craccivora* является типичным для тлей рода Aphis, однако отличаясь от такового у тлей других родов. Среди особенностей этого рода можно выделить уникальные гены 2A2 и 2B33, а также отсутствие генов 1_1.

Ключевые слова: геномика; тли; системы детоксикации.

Введение. В детоксикации ксенобиотиков участвуют три группы ферментов, которые поочерёдно преобразуют субстрат. При изолированном изучении генов, продукты которых участвуют в детоксикации на ее первой стадии (CYP450 и/или эстеразы и карбоксил-эстеразы) невозможно получить полную картину работы системы детоксикации. В связи