12. Коуэн К. Ф. Н., Грант П.М. Адаптивные фильтры. Пер. с англ. Лихацкой Н.Н., Ряковского С.М. М.: Мир, 1988. 392 с.
13. Уидроу Б., Стирнз С. Адаптивная обработка сигналов. Пер. с англ. под ред. Шахгильдяна В.В. М.: Радио и связь, 1989. 440 с.
14. Пистолькорс А.А., Литвинов О. С. Введение в теорию адаптивных антенн. М.: Наука, 1991. 200 с.
15. Тараканов А. Н., Хрящев В. В., Приоров А. Л. Адаптивная цифровая обработка сигналов: учебное пособие. Ярославль: ЯрГУ, 2001. 134 с.
16. Джиган В.И. Адаптивная обработка сигналов в радиотехнических системах: учебное пособие. М.: МИЭТ, 2012. 148 с.
17. Джиган В.И., Лесников В.А. Адаптивная цифровая фильтрация в радиотехнике и связи: учебное пособие. Киров: Вятский государственный университет, 2014. 100 с.
18. Джиган В.И. Адаптивные алгоритмы и устройства радиотехнических систем: учебн. пособие. М.: МИЭТ, 2016. 104 с.

DECONVOLUTION OF "BIG DATA" IN CANCER GENOMICS: FROM PAN-CANCER LEVEL TO SINGLE CELLS

M. Chepeleva[1,2)], Y. Wang[2)], A. Kakoichankava[3)], A. Muller[2)], T. Kaoma[2)], P. V. Nazarov[2)]

1) Belarusian State University, Minsk, Belarus
2) Luxembourg Institute of Health, Strassen, Luxembourg
3) Vitebsk State Medical University, Vitebsk, Belarus
Correspoding author: P. V. Nazarov (petr.nazarov@lih.lu)

Large genomics pan-cancer datasets that were made publically available in the last decade are now complemented with measurements at single cell level and may include up to a billion data points. Here we show how deconvolution method based on independent component analysis can process transcriptomes measured for bulk samples at pan-cancer level and for single-cell measurements from normal tissues and neoplasia.

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

Introduction. The majority of tumor samples collected from patients and studied by high-throughput transcriptomics are heterogeneous at three levels. First, bulk tissue samples contain a mixture of several cell types. Their proportions vary from one specimen to another and are difficult to control. Second, cancers naturally develop inter and intra-tumor heterogeneity of malignant cells. Third, the evolving technology may introduce technical biases and limit comparison of data originated from new patients to large publicly available datasets, such as The Cancer Genome Atlas (TCGA) [1].

One of experimental methods, developed to disentangle the complexity of bulk biological samples and characterize various cell populations is a single-

cell RNA sequencing (scRNA-seq). It has a strong application potential in oncology. At the same time, single cell data are susceptible to the same technical issues as bulk RNA-seq, including the effects of batches, number of reads per cell (cell "size"), gene length, and patient-to-patient variability. In addition, various biological processes in the cells can be masked by the cell cycle, which is the strongest contributor to transcriptome variability in proliferating tumor cells. Another challenge of single cell data is its dimensionality: one dataset may easily describe ~10^4 features for ~10^4 cells.

In this work, we applied an *in silico* data-driven deconvolution method called consensus independent component analysis (ICA) [2, 3] to resolve the complexity of bulk and single cell data. The methods were able to separate technical biases from signals of biological interest, isolate signals from different biological processes and cell subpopulations in different components integrate several levels omics data with recorded patient clinical data. The developed methods were implemented in R scripts and as an R package *consICA* and were applied individually to several cancers included into TCGA dataset, as well as to entire TCGA and, in addition, to several scRNA-seq experiments.

Methods. The parallel multiplatform consensus ICA was implemented in *consICA* [3] package. The ICA algorithm (R-package *fastICA*) was performed with multiple initial estimations, excluding random samples for each run. At each run the centered and scaled initial data matrix was presented as a matrix product of independent signals (***S***) and their weights (***M***) (Fig.1). The results of individual runs were mapped and consensus matrices of the signals and weights were calculated [3]. Finally, matrix ***S*** can be interpreted based on the most contributing genes, e.g. by gene-set enrichment analysis, and weight matrix ***M*** can be linked to clinical data using ANOVA and Cox-regression. We also used random forest as a classification tool to predict subtypes of the tumor based on component weighs. The method was tested on 10456 TCGA samples and several independent datasets. Additional level of interpretation of independent components was achieved by including scRNA-seq data into consideration. Two scRNA-seq datasets composed of 2544 normal pancreas cells [4] and 3304 cancer cells from two patient-derived glioblastoma cell-lines [5] were considered.

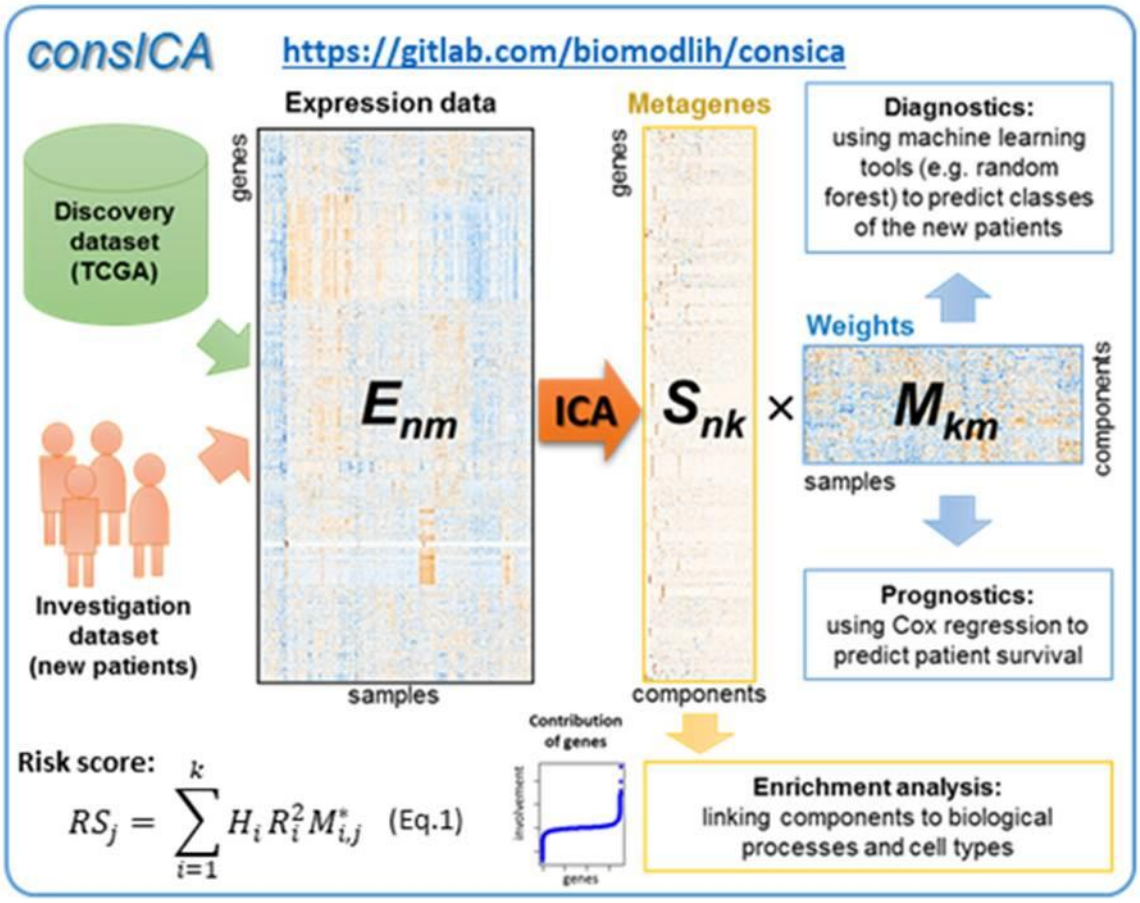

Fig. 1. Consensus ICA implemented in consICA tool decomposes combined gene expression matrix from multiple samples into meaningful statistically independent signals in the space of genes (metagenes, S) and weights in the space of samples (weight matrix, M). Biological processes and signatures of cell subtypes can be found in S, while M could be linked to patient groups and patient survival (using Cox regression).

Results. We first applied the method independently to several cancers: skin cutaneous melanoma (SKCM) [3], lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), low grade gliomas (LGG), glioblastoma multiform (GBM), pancreatic cancer (PAAD) and prostate cancer (PRAD). We showed that ICA-based deconvolution engineers features with predictive power and can be used to classify subtypes of the tumors, predict patient survival and characterize intensity of biological processes within the bulk patient samples. In the majority of datasets a cell cycle component was associated with poor survival. Similar behavior was observed for a component linked to keratinization and epidermis development. Interestingly, the presence of immune-related components showed an antinomic effect: they were associated with better survival for SKCM, LUAD, LUSC patients, showed no effect in PRAD and PAAD, and an increased the risk for brain cancer patients (GBM and LGG).

We investigated predictions of the biological processes on in-house data from GBM patient tumor specimens, patient-derived orthotopic xenograft (PDOX) models *in vivo*, and patient-derived classical and stem-like cell lines cultured *in vitro* and xenografted *in vivo* (collaboration with Dr. A. Golebiewska and Prof. S. Niclou, Luxembourg Institute of Health).

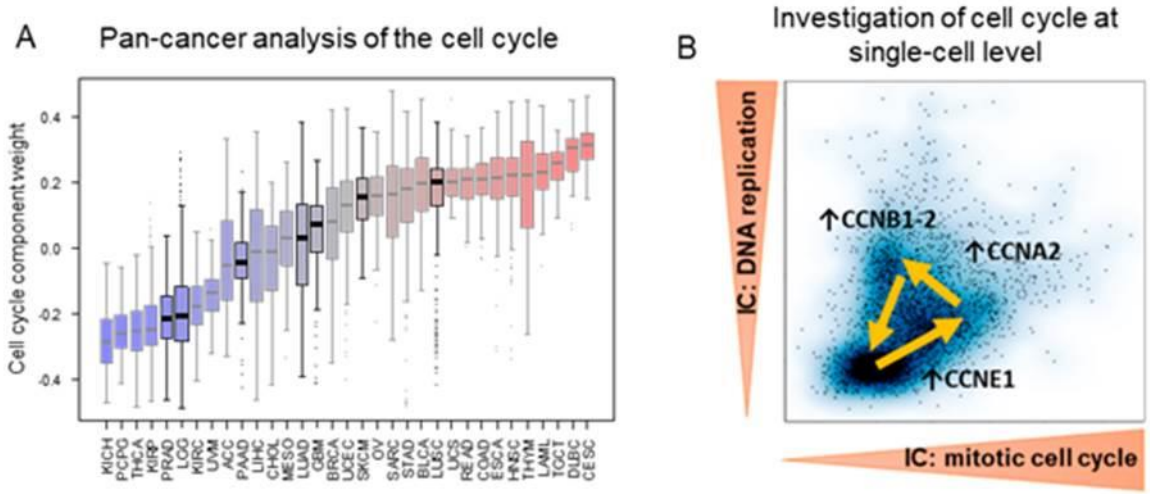

Fig. 2. Consensus ICA detects cell cycle in bulk pan-cancer TCGA dataset composed of 10456 samples from 33 cancers (A) and in a single cell data of 3304 cancer cells (B) (from Sompairac et al., 2019 [2]). Pan-cancer analysis shows the proliferation level of different cancers, averaged over the cells collected in each sample (highlighted cancers were investigated specifically). At the same time, scRNA-seq allows zooming into the cycling process itself and discriminate between subpopulations of dividing and resting cells. Up-regulated gene markers of the cell cycle are highlighted.

The highest cell cycle component was found for the stable and most proliferative cell lines U87, U251 that showed minimal cell migration, while preclinical models showed strong migrating signals. This confirmed the benefits of replacement of human stromal compartment with rodent counterparts. The method was also able to reproduce the phenotype of patient-derived cell lines: angiogenic (forming solid tumors with blood vessels), migrating (spreading in mesenchyme without vascularization) and mixed.

Simultaneous analysis of the entire TCGA dataset (pan-cancer) was performed using 100 independent components. We identified components linked to increased risk in all considered 33 tumors: keratinization/cornification, cell cycle, inflammation, increased glycolysis, cell motility and angiogenesis. Average level of the cell cycle can be a characteristic of tumor aggressiveness (Fig. 2A).

On the single-cell level, our method successfully identified technical factors (batches, number of reads per cell) and biological signals (cell cycle in tumor tissue, mRNA processing, ribosome biogenesis, translation) and isolated their effects in different components. We showed that it is possible to exclude technical components, such as batch effect and “cell size”, and thus

normalize the single cell data by removing correspondent components. Similarly to bulk samples, some components can also be used to discriminate between cell types. Finally, we investigated two components linked to the cell cycle and reconstructed trajectory of the dividing cells using the information about the markers of the cell cycle (Fig. 2B). Interestingly, we were able to correlate the ICA results of scRNA-seq and bulk sample RNA-seq and show that similarity is preserved between estimated components both for cancer and stromal cells. The later fact was used to identify stromal cells in PAAD cancer.

Conclusions. We demonstrated that ICA-based deconvolution can be applied at the level of large datasets from bulk samples as well as at the level of single cells. In both cases, the developed *consICA* was able to separate technical effect from biologically related signals. Biological signals can be related to cellular processes, such as cell cycle, transcription, translation and metabolic processes; or to cell types: various immune, endothelial, muscle cells, as well and highly proliferative cancer cells. By mapping of the components between bulk and single cell datasets, we can improve the annotation of the components and therefore establish a link between observation at cellular level and clinical data.

REFERENCES

1. Weinstein J. N., Collisson E. A., Mills G. B., Shaw K. R. et al. The Cancer Genome Atlas Pan-Cancer analysis project // Nat Genet. 2013. V. 45, № 10. P. 1113–1120.
2. Sompairac N., Nazarov P. V., Czerwinska U., Cantini L. et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets // Int. J. Mol. Sci. 2019. V. 20, №18. P. 4414–4441.
3. Nazarov P. V., Wienecke-Baldacchino A. K, Zinovyev A., Czerwinska U. 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. V. 12. P. 132.
4. Enge M., Arda H. E., Mignardi M., Beausang J. et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns // Cell. 2017. V. 171, №2. P. 321–330.
5. Dirkse A., Golebiewska A., Buder T., Nazarov P. V. et al. Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment // Nat Commun. 2019. V. 10, №1. P. 1787.