THE STATISTICAL STABILITY OF CONSENSUS INDEPENDENT COMPONENT ANALYSIS FOR RNA-SEQ DATA IN CANCER RESEARCH

Chepeleva M., Yatskou M., Nazarov P.
Department of System Analysis and Computer Modeling, Belarusian State University
Multiomics Data Science Research Group, Quantitative Biology Unit, Luxembourg Institute of Health
Minsk, Belarus; Strassen, Luxembourg
E-mail: maryna.chepeleva@gmail.com, yatskou@bsu.by, petr.nazarov@lih.lu

Independent component analysis (ICA) became a part of the standard machine learning pipeline for genomics data analysis. The approach allows to correct technical biases and batch effects in transcriptomics datasets. Separated signals are successfully used to characterize biological functions, their weights might be used for diagnostics (cancer subtypes classification) and prognostics (survival prediction). Using weights of independent components as features for downstream analysis requires high reproducibility of decomposition. Here we investigated the stability of extracted components depending on ICA parameters and validated the optimal number of parallel consensus ICA runs that provided reproducible deconvolution. Also, we estimated the effect of parallel runs on the quality of lung cancer type classification (LUSC/LUAD) and gene enrichment analysis results. Finally, we estimated the boundary values for the number of components that allows detecting biologically relevant signals in smaller patient cohorts.

Introduction

Independent component analysis (ICA) allows decomposing heterogeneous transcriptomics data and extracting relevant transcriptional signals that correspond either to relevant biological processes or to technical biases [1]. Using independent components as features for downstream analysis requires high reproducibility of decomposition. Here we investigated the stability of ICA and tested reproducibility of its results for single and multiple runs, and in the case of reduced number of samples.

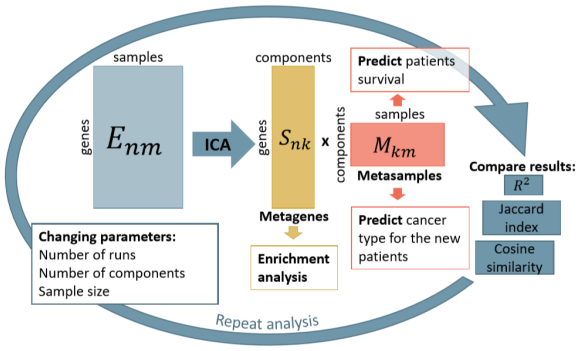

Рис. 1 – ICA decomposes gene expression matrix into meaningful signals S and weights M. Biological processes can be found in S, while M could be linked to patient cancer groups and patient survival. Changing input data and ICA parameters results can be compared using correlation and similarity metrics.

Method

We applied the developed parallel consensus Independent Component Analysis (ICA) algorithm [2] to TCGA [3] RNA-seq gene expression data on patients with skin cutaneous melanoma (SKCM) and non-small cell lung cancers: squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD). ICA finds a robust decomposition of an expression matrix: $E = S \times M$, where S is a matrix of statistically independent and biologically meaningful signals (metagenes) and M – their weights (metasamples). Functional annotation of components was performed by enrichment analysis using R package topGO [4] (biological processes are considered).

Two predicting models were used to classify the patients based on ICA results: random forest [5] and xgboost [6] from corresponding R packages. Metasamples (M-matrix), the most significant differentially expressed genes (by limma package [7]) and all the genes were used as input features to classifiers. To estimate patient survival, Cox-regression model was trained using ICA-based risk score, as in [1]. To explore how the stability of ICA depends on the parameters, we repeated the analysis on subsampled data sets and carried out pairwise comparison of calculated metagenes (S-matrix) or enriched gene ontologies (GOs) corresponding to each component. Independent components from different ICA runs were matched by maximum correlation. GO lists were compared using Jaccard index and cosine similarity between ranks of GO-term significance (Fig. 1).

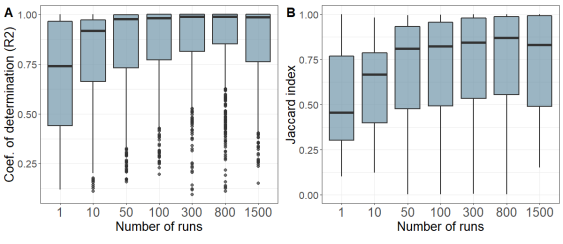

Рис. 2 – Stability of independent components (A) and significant gene signatures (B).

Results

Multiple runs. Exploring dependence of the stability on the number of consensus ICA runs, we observed a strong increase of a squared correlation R^2 between corresponding independent signals (columns of S) and Jaccard indexes between contributing genes signatures (FDR < 0.05) as is shown in Fig. 2 A, B. GO similarity showed growing trend as well (Fig. 3A). For lung cancers 18 of components did not have enriched GOs (FDR < 0.05). These components may be linked to technical artefacts. Based on the presented profiles (Fig 2), the number of consensus parallel runs > 100 was enough to provide high reproducibility of ICA.

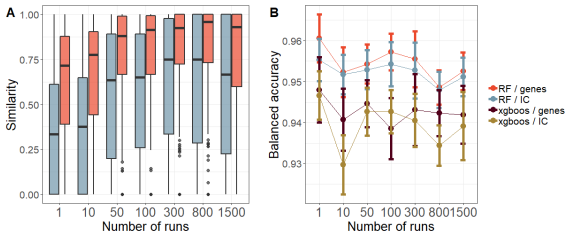

Рис. 3 – Stability of independent components (A) and significant gene signatures (B).

To validate the increase of reproducibility we performed two additional verifications. First, elements of each row in the data matrix were permuted, thus preserving distribution for each gene. Then ICA was performed on the permuted data. Second, we repeated permutations before each ICA. Slight increase was observed in the first case and no significant increase in the second (data not shown).

Average balanced accuracy did not have a considerable variation (Fig 3B). Therefore increasing number of ICA runs raises the stability but have no influence on classification accuracy.

Sample size. To investigate the dependency of ICA stability on the sample size, we fixed the number of components, as it can not be larger than the number of the samples. We selected equal number of LUSC and LUAD patients and a smaller number of normal samples from the dataset. The number of normal patients did not show any effect on the stability and classification accuracy. Fig. 4, 5 present the required sample size in order to achieve the median stability R^2 above 0.5 ICA with a small number of components (≤ 30) required less samples to reach plateau in the stability. However ICA with the low number of components may be not sensitive enough to detect all important biological signals and technical artefacts. Thus it is necessary to keep a balance between the number of components and the sample size. Interestingly, after a certain number of components, there is no improvement in classification accuracy with their increase.

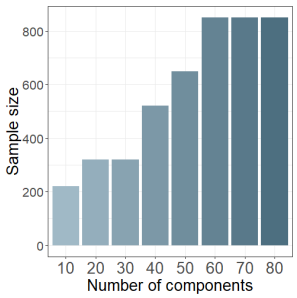

Рис. 4 – Required sample size to get median 0.5 stability (R^2) of metagenes for fixed number of components.

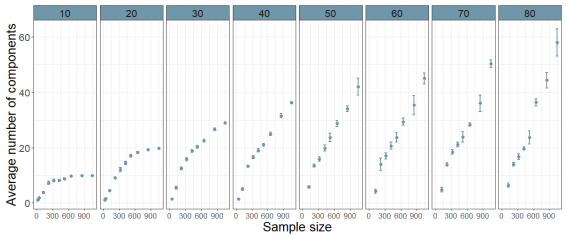

Рис. 5 – An average number of components with $R^2 > 0.5$ depending on the sample sizes.

Conclusion

To guarantee a high reproducibility of ICA 100 runs is sufficient. Random forest provides the highest accuracy on the significant genes but this approach loses in working time and interpretability. In order to detect more independent biological signals, more components should be used in ICA. However it requires a larger dataset.

Acknowledgement

This work was supported by the Luxembourg National Research Fund (C17/BM/11664971/DEMICS).

1. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients / P. V. Nazarov [et al.] // BMC Med Genomics. – 2019. – V. 12, № 1. – P. 132–149.
2. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets / N. Sompairac [et al.] // Int J Mol Sci. – 2019. – V. 20, № 18. – P. 4414–4441.
3. Tomczak, K. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge / K. Tomczak, P. Czerwińska, M. Wiznerowicz // Contemp Oncol (Pozn). – 2015. – V. 19, № A1. – P. 68–77.
4. Alexa, A. Enrichment Analysis for Gene Ontology with topGO [Electronic resource] / A. Alexa, J. Rahnenfuhrer // Bioconductor. – Mode of access: https://bioconductor.org/packages/release/bioc/html/topGO.html – Date of access: 4.10.2019.
5. Liaw, A. Classification and Regression by randomForest / A. Liaw, M. Wiener // R News. – 2002. –V. 2, № 3. P. 18–22.
6. Chen, T. XGBoost: A Scalable Tree Boosting System / T. Chen, C. Guestrin // KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016. – P. 785–794.
7. Ritchie, M. E. Limma powers differential expression analyses for RNA-sequencing and microarray studies / M. E. Ritchie, B. Phipson // Nucleic Acids Research. – 2015. – Vol. 43. – P. e47.