DECONVOLUTION OF CELL AND ENVIRONMENT SPECIFIC SIGNALS AND THEIR INTERACTIONS FROM COMPLEX MIXTURES IN BIOLOGICAL SAMPLES

Czerwinska Urszula[1], Barillot Emmanuel[1], Soumelis Vassili[2], Zinovyev Andrei[1]

[1] Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, Paris

[2]Institut Curie, PSL Research University, Inserm U932, Paris

PROBLEM OF MIXED SIGNALS IN TUMOR MICROENVIRONMENT

- studied system is a complex mixture of signals of various origins
- tumors are engulfed in a tumor microenvironment (TME) that critically impacts progression and response to therapy[1]
- TME includes tumor cells, fibroblasts, and a diversity of immune cells[1]
- it is possible to separate complex signal mixtures, using classical and advanced methods of source separation and dimension reduction[2]

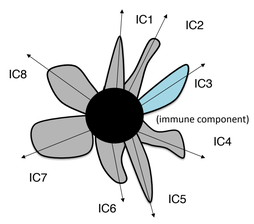
OUR APPROACH

- independent components analysis (ICA)[3] to decipher sources of signals shaping transcriptomes (global quantitative profiling of mRNA molecules) of tumor samples
- particular focus on immune system-related signals
- application of ICA iteratively decomposing components into sub-components that can be interpreted using pre-existing immune signatures through correlation, enrichment[4] or overdispersion[5] analysis

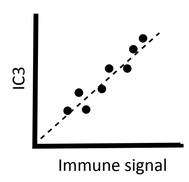
PIPELINE

TRANSCRIPTOMIC
DATA
BREAST CARCINOMA PUBLIC DATASET

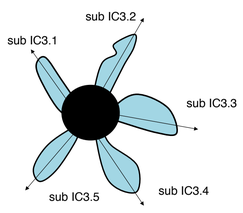
UNSUPERVISED
DECONVOLUTION
INDEPENDENT COMPONENTS ANALYSIS



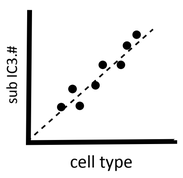
IDENTIFICATION OF
IMMUNE COMPONENT
CORRELATION/ GENE ENRICHMENT/ OVERDISPERSION ANALYSIS

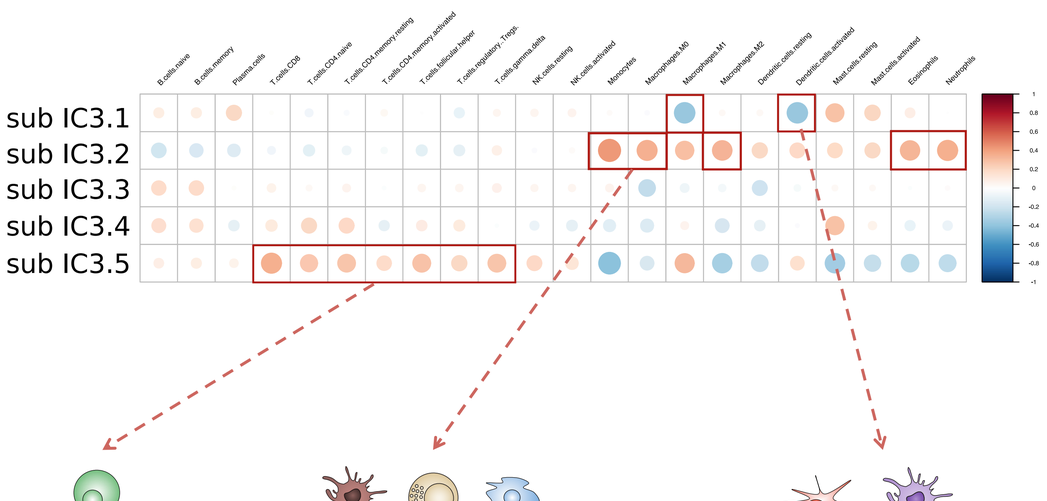


UNSUPERVISED
DECONVOLUTION
INDEPENDENT COMPONENTS ANALYSIS OF **IMMUNE COMPONENT**



ASSOCIATION WITH
IMMUNE CELL TYPES
CORRELATION/ GENE ENRICHMENT/ OVERDISPERSION ANALYSIS





Pipeline of our deconvolution procedure. Started with public transcriptomic data of breast carcinoma[6]. Normalized matrix included measure of 14709 transcripts with corresponding HUGO names for 277 patients treated with tamoxifen. Data were centered and transformed into log10 space. FastICA[3] algorithm was applied to the matrix. Number of components was decided based on stability criterium. Resulting independent components (metagenes) were interpreted through GESEA analysis using MSigDB[7], GO[8], KEGG[9] and Reactome[10], correlation analysis with known metagenes[11] and through overdispersion analysis[5]. Therefore the matrix was reduced to the genes the most contributing to the component characterized as immune-related (IC3) and the FastICA algorithm was applied to centered reduced matrix. Resulting components (metagenes) were again characterized using mentioned tools with database completed with a set of immune signatures (see Toolbox: immune signatures). Three out of 5 components were associated with groups of cell types. Here visualized through spearman correlation with marker genes expression of 22 immune cell (sub)types from LM22 matrix[12,13] where negative colors correspond to negative correlations and the size of the point to the correlation coefficient. Only significative (p-value< 0.05) correlations are represented. The three identified groups were (sub IC3.5) T-cells, (sub IC3.2) Monocytes/ Macrophages M0 and M2/ Eosynophiles/Neutrophiles, (sub IC3.1) DC/Macrophages M0. Remaining sub IC3 (sub IC3.3 and sub IC3.4) components were featureless.

KEY MESSAGES

Deconvolution of mixed signal in TME gives better insight into TME lanscape

Unsupervised technique = robustness & no overfitting

It allows better exploration of most available and accesssible resource : transcroptomic data

In the case of success, the results will be used in the diagnosis and cancer therapy, especially immunotherapies

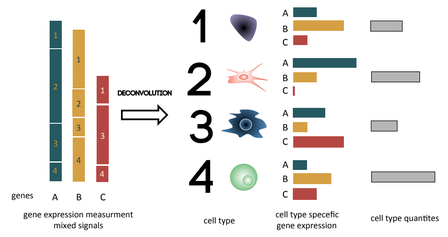
TOOLBOX

deconvolution

objective is to find cell type specific expression and cell type aboundance from complex tissue total gene expression (RNA levels) through a mathematical manipulation of the matrix

unsupervised deconvolution

algorithm does not have any previous information about cell types number, quantities or marker genes



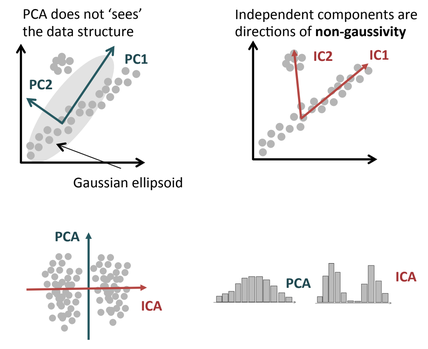
Independent Components Analysis (ICA)

a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals based on the notion of mutual information between components; source of factors must be independent.

Main **ICA theorem**[14] :
Minimize mutual information = maximize non-gaussivity

Comparing to **Principal Component Analysis (PCA)**[15], ICA does not impose othogonality of components.

Comparing to **Negative Matrix Factorisation (NMF)**[16], ICA does not impose any constraints, while NMF impose non-negativity of the weights and data. In our ICA analysis, negative projections are interpreted in terms of absolute values. Tests performed with NMF for immune cell types deconvolution gave results hard to interpret (data not shown).



immune signatures

also called 'marker genes'; set of genes allowing a differentiation of one immune cell type from an other. In our analysis we are using sets of marker genes published in R package 'Cell Mix'[17] as well as more recent published[12,13, 18-24] and in house curated signatures.
We accord higher accuracy to signatures coming from cancer studies[23]. We also value weighted signatures[11,12,19], as they can be explored and interpreted in more quantitative way.
We believe there is a need to establish **gold standards** for 'immune signatures' depending on the signature context and purpose of application.

PERSPECTIVES

- evaluation of robustness of the represented groups
- application in several types of cancer
- characterization the immune infiltration degree in the cancer transcriptome dataset
- further correlate the immune infiltrate with patients' survival and tumor characteristics
- compare to alternative (supervised) methods

LIMITATIONS

- the level of definition of deconvolution (cell groups – cell types – cell subtypes)
- evaluation of cell type proportion
- data quality

CONCLUSIONS

- identification of signals related to groups of immune cell types with unsupervised learning approach in a Breast Carcinoma dataset
- enrichment, correlation and oversipersion analysis identified significative groups corresponding to three out of five sub-signals: (1) T-cells, (2) DC/Macrophages, (3) Monocytes/ Macrophages/ Eosynophiles/Neutrophiles

REFERENCES

1. Fridman et al. (2012)
2. Zinovyev, A.et al. (2013
3. Hyvärinen et al. (2001)
4. Subramanian et al. (2005)
5. Martignetti et al. (2016)
6. Loi et al. (2007)
7. Liberzon et al. (2011)
8. Gene Ontology Consortium (2004)
9. Kanehisa (2002)
10. Joshi-Tope (2005)
11. Biton et al. (2014)
12. Newman et al. (2015)
13. Gentles et al. (2015)
14. Comon and Jutten (2010).
15. Pearson (1901)
16. Lee and Seung (1999)
17. Gaujoux et al. (2013)
18. Abbas et al. (2005)
19. Abbas et al. (2009)
20. Palmer et al. (2006)
21. Watkins et al. (2009)
22. Grigoryev et al. (2010
23. Bindea et al. (2013)
24. Doig et al. (2013)

ABOUT AUTHOR

Urszula Czerwinska
PhD candidate 2015-2018
I am passionate about **Data Science** and **Analytics**.
Hope you enjoyed the poster and do not hesitate to contact me!

Paris, France
+ 33 698210283
urszula.czerwinska@cri-paris.org
@UlaLaParis
/urszulaczerwinska
UrszulaCzerwinska
http://sysbio.curie.fr/

NO TIME TO READ IT NOW? SCAN THE QR CODE AND HAVE A LOOK LATER!!!