COMPUTATIONAL DECONVOLUTION OF MIXED SIGNALS IN TUMOR MICROENVIRONMENT USING INDEPENDENT COMPONENTS ANALYSIS

Urszula Czerwinska[1], Ulykbek Kairov*[2], Laura Cantini*[1], Alessandro Greco[1], Emmanuel Barillot[1], Vassili Soumelis[3], Andrei Zinovyev*[1]

1 Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, Paris
2 Laboratory of bioinformatics and computational systems biology, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan
3 Institut Curie, PSL Research University, Inserm U932, Paris

WHY STUDY TME?

TME critically impacts cancer prognosis and response to treatment[1]

TME is composed of tumor cells, fibroblasts ana a diversity of immune cells[1]

Estimating immune infiltration and its impact remains a challenging task[2]

HOW DO WE STUDY TME?

Apply ICA to reduce tumor transcriptomes into essential factors[3]

Identify the immune-related genes and their importance in each transcriptome

Identify cell-type specific independent components

Develop and validate the method and the pipeline

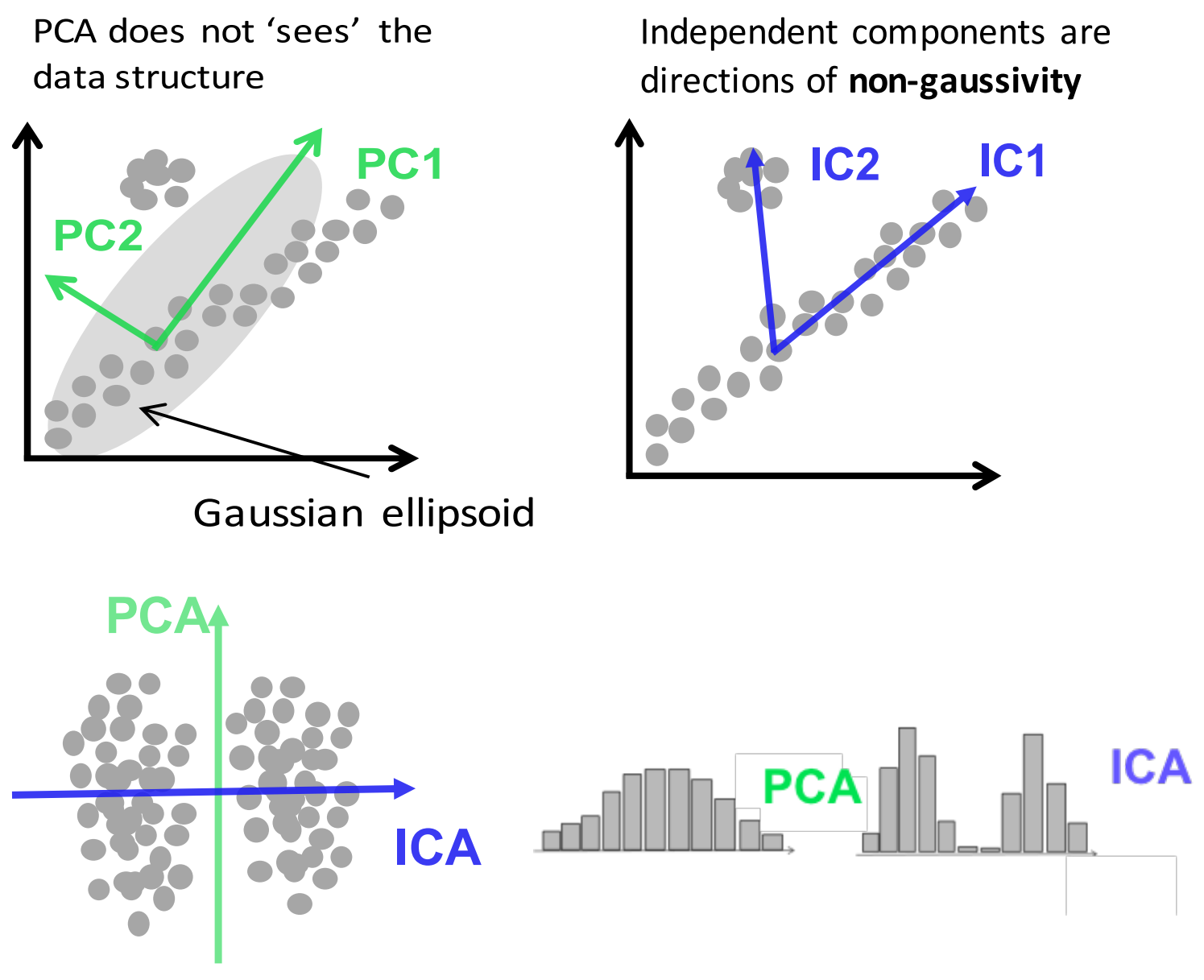
1. WHAT IS ICA[4]?

matrix factorisation

blind source deconvolution

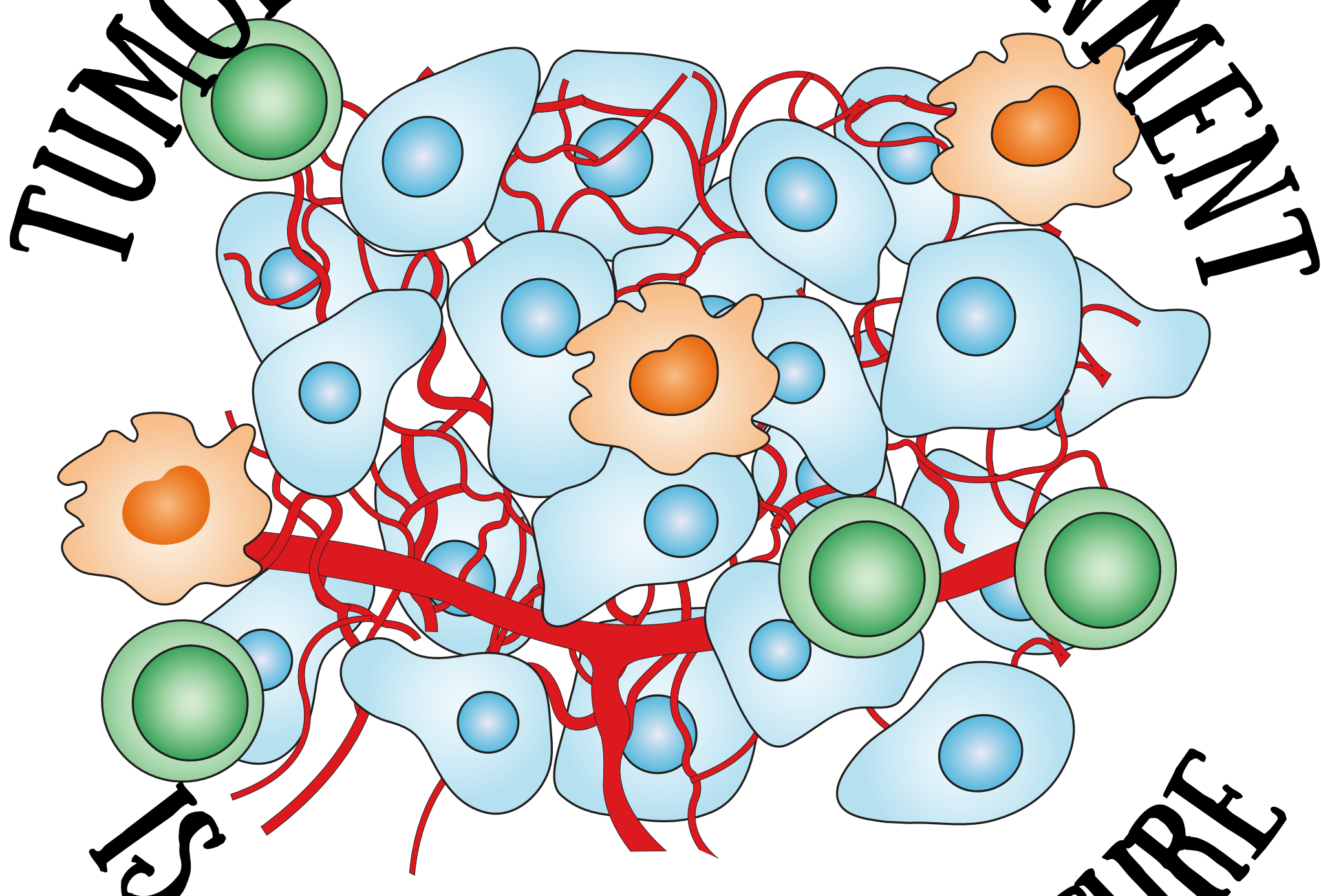
minimize mutual information = maximize non-gaussivity[5]

compared to Principal Component Analysis[6] (PCA), ICA does not impose orthogonality of components


PCA does not 'sees' the data structure
PC1
PC2
Gaussian ellipsoid
Independent components are directions of non-gaussivity
IC2
IC1
PCA
ICA
PCA
ICA

compared to Negative Matrix Factorization (NMF)[7], 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)

TUMOR MICROENVIRONMENT IS A COMPLEX MIXTURE



2. BUT HOW TO DEFINE NUMBER OF COMPONENTS?

Independent components cannot be naturally ordered

the independent components are only defined as local minima of a non-quadratic optimization function = runs can give different

icasso[8] method have been developed to improve the stability of the independent components

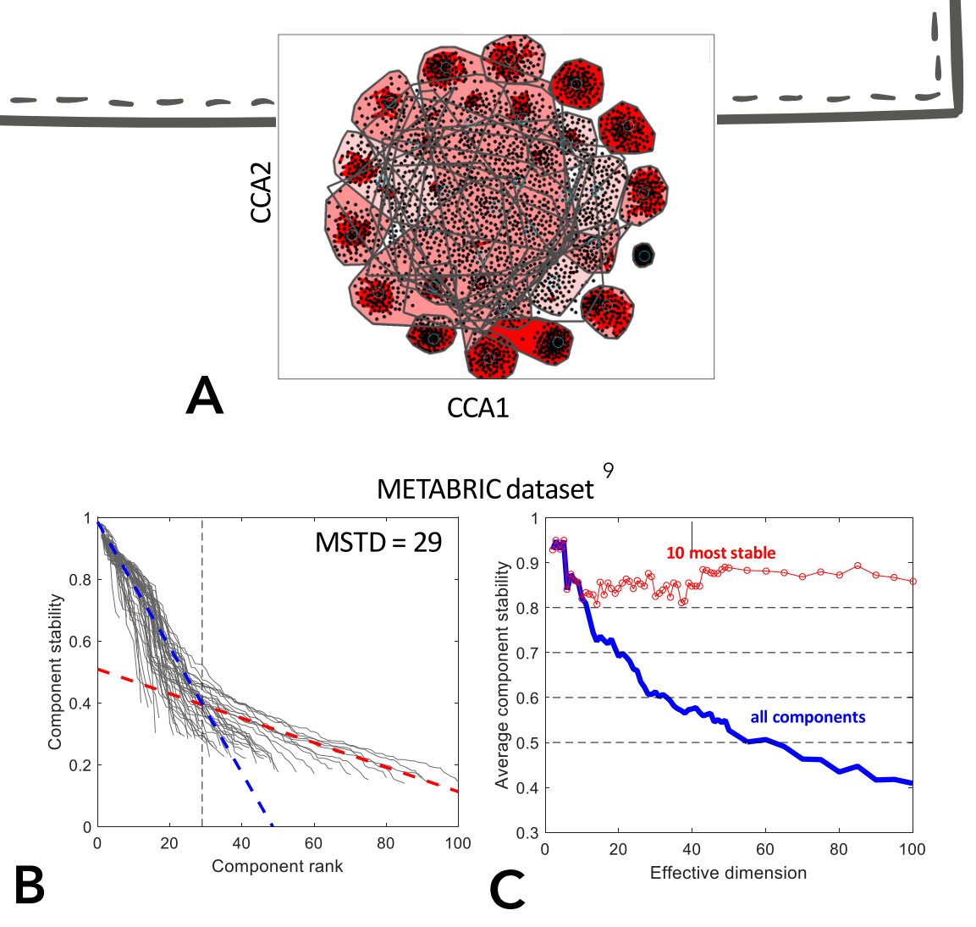


Fig 1. A) visualizing the results of computing ICA 100 times with MSTD=29 components in METABRIC dataset and component clustering (icasso package, Canonical Correlation Analysis (CCA) plot). Each black point represents a component, red lines show significant correlations between them, polygons show the convex hull area of the clusters. B) stability profiles for ICA decompositions in various dimensions (from 2 to 100) shown by grey lines. Two-line clustering result is shown by blue and red dashed lines, with MSTD determined as the point of their intersection (vertical dashed line). C) average stability profile SMtotal (blue line) and the average stability of 10 most stable components SM(10) (red line).

Maximally Stable Transcriptome Dimension (MSTD), a novel criterion for choosing the optimal number of ICs in transcriptomic data analysis

Determining the optimal number of reproducible independent components for transcriptomic data analysis
Ulykbek Kairov, Laura Cantini, Alessandro Greco, Askhat Molkenov, Urszula Czerwinska, Emmanuel Barillot, Andrei Zinovyev
[Under revision]

Compute stability index of each cluster:

$$I_q(C_k) = \frac{1}{|C_k|^2} \sum_{i,j \in C_k} |r_{ij}| - \frac{1}{|C_k| \sum_{l \neq k} |C_l|} \sum_{i \in C_k} \sum_{j \notin C_k} |r_{ij}|$$

C_k kth cluster $|C_k|$ kth cluser size r_{ij} Pearson correlation coeff between components

Compute average stability index:

$$S(M) = \frac{1}{M} \sum_k I_q(C_k) \quad M \text{ number of clusteres}$$

MSTD = the point of intersection of the two lines approximating the distribution of stability profiles

CONCLUSIONS

ICA is a reproductible and unsupervised manner to decompose transcriptomes into biological functions

ICA revealed components realted to immune cell types in tumor transcriptomes

Estimating immune cell types aboundance, better validation framework and user-friendly pipeline to perform our analysis are ongoing progress

4. VALIDATION

Lack of gold standard

Possible partial validation with FACS of blood, IHC, methylome

In most publications for simulated mixtures:

Cell type profiles from blood or cell-lines

Simplistic, do not take into account gene covariance and plausible proportions of cell types

Our simulation ideas

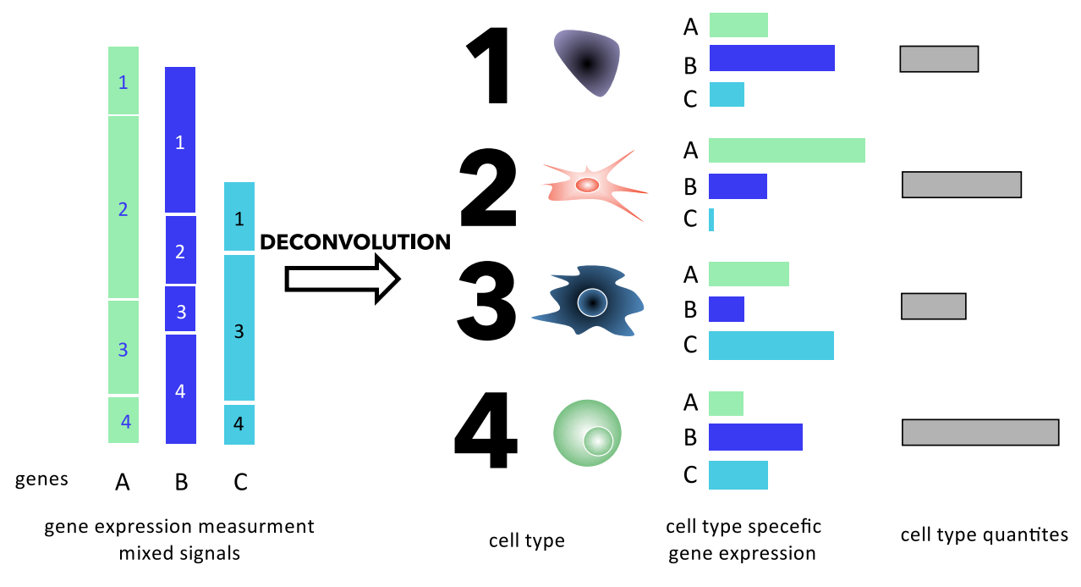
using single cell profiles from Melanoma to simulate bulk

implicit: estimation of distribution parameters copulas?

explicit : mimic existing ditribution Generative Adversarial Networks (GAN) ?

3. ICA TO STUDY IMMUNE INFILTRATION

3.1 Model


1
2
3
4
genes
A
B
C
gene expression measurment mixed signals
DECONVOLUTION
cell type
cell type specefic gene expression
cell type quantites

In the basic hypothesis[10], mixture of signals from TME in transcriptomic samples can be described as a linear mixture.

$$AX=B$$

B microarray data matrix of one biological sample, X are mixing proportions and A is the matrix of expression of genes in each cell type.

Blind source separation[11] separates the set of mixed signals $x(t)$, through determination of an 'unmixing' matrix $B=[b_{ij}] \in R^{n \times m}$, to 'recover' an approximation of original signals, $y(t) = (y_1(t), \dots, y_n(t))^T$

$$y(t) = B \cdot x(t)$$

3.2 Application to Breast carcinoma

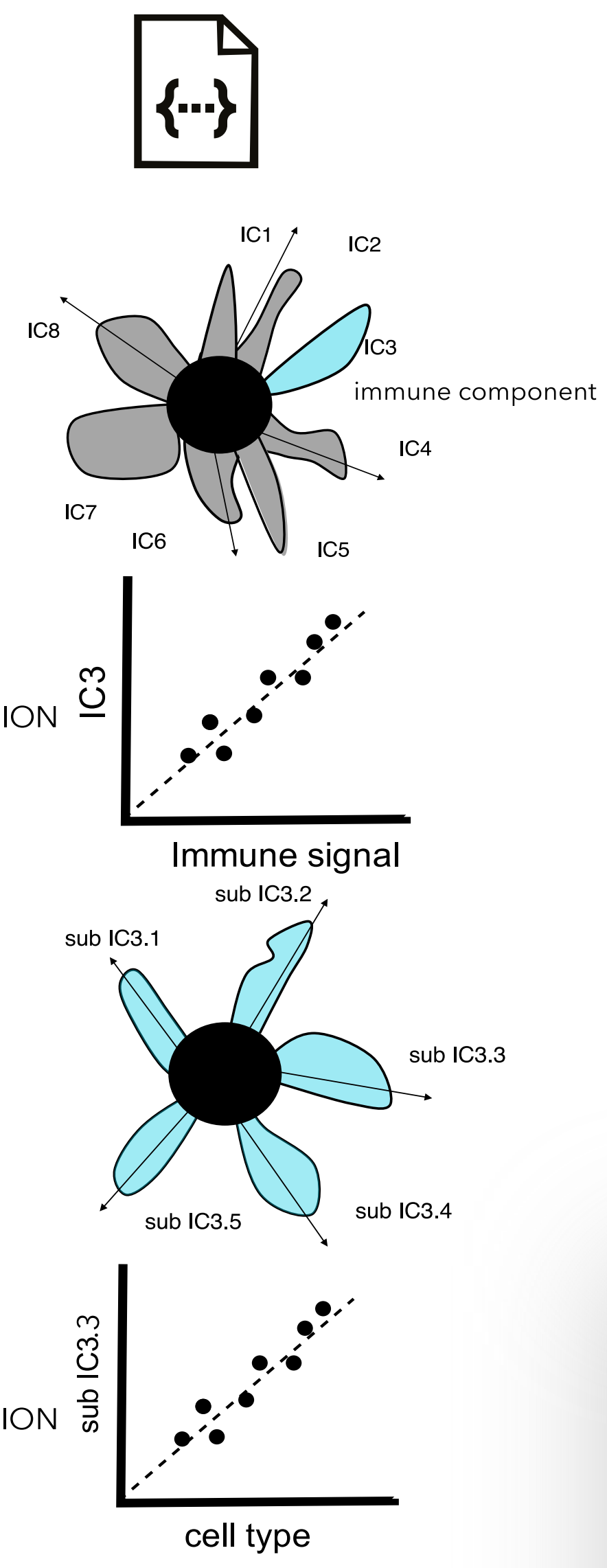
TRANSCRIPTOMIC DATA[12]
BREAST CARCINOMA PUBLIC DATASET

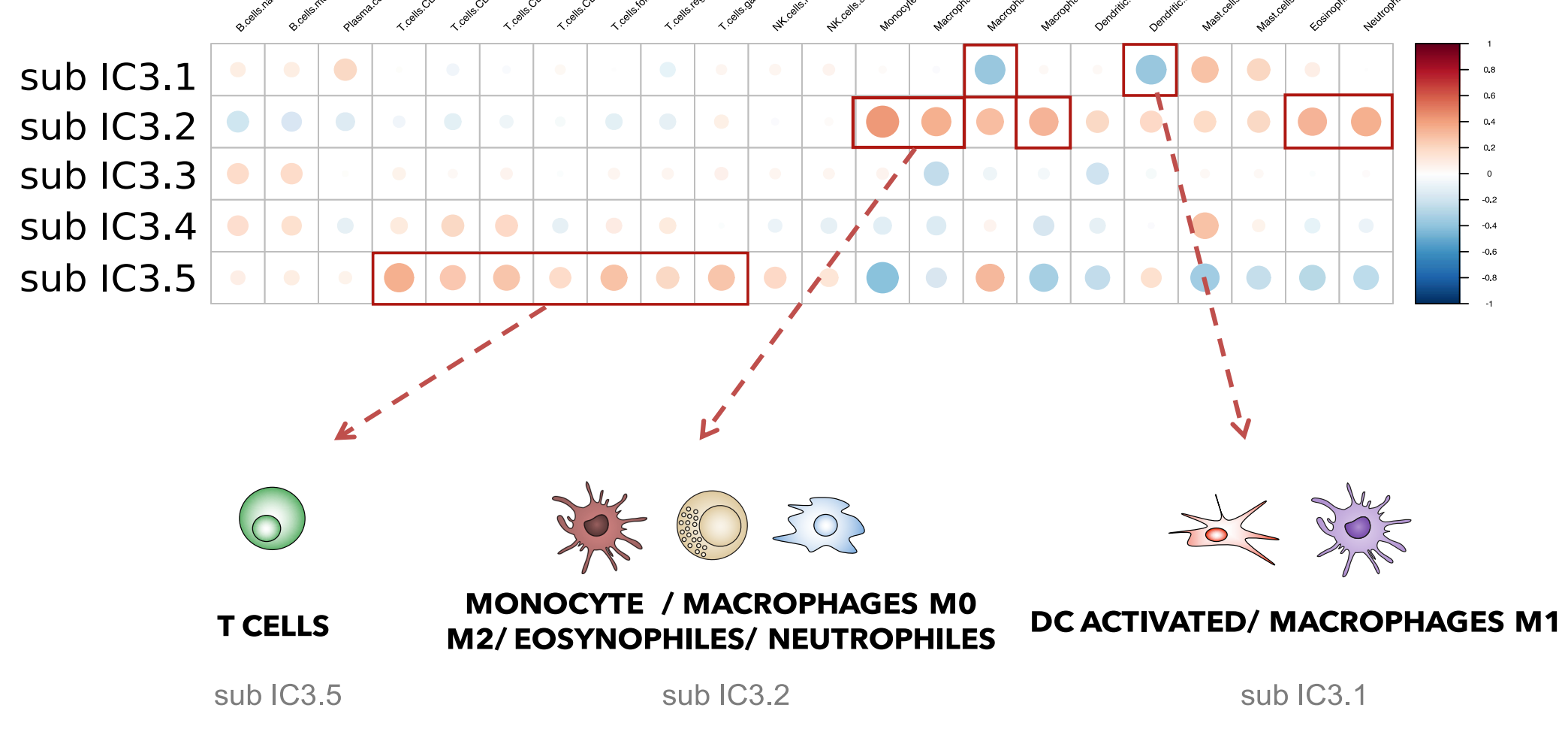
UNSUPERVISED DECONVOLUTION
INDEPENDENT COMPONENTS ANALYSIS

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





T CELLS
sub IC3.5

MONOCYTE / MACROPHAGES M0 M2/ EOSYNOPHILES/ NEUTROPHILES
sub IC3.2

DC ACTIVATED/ MACROPHAGES M1
sub IC3.1

3.3 Decomposition of Metabric

Correlation with immune metagene

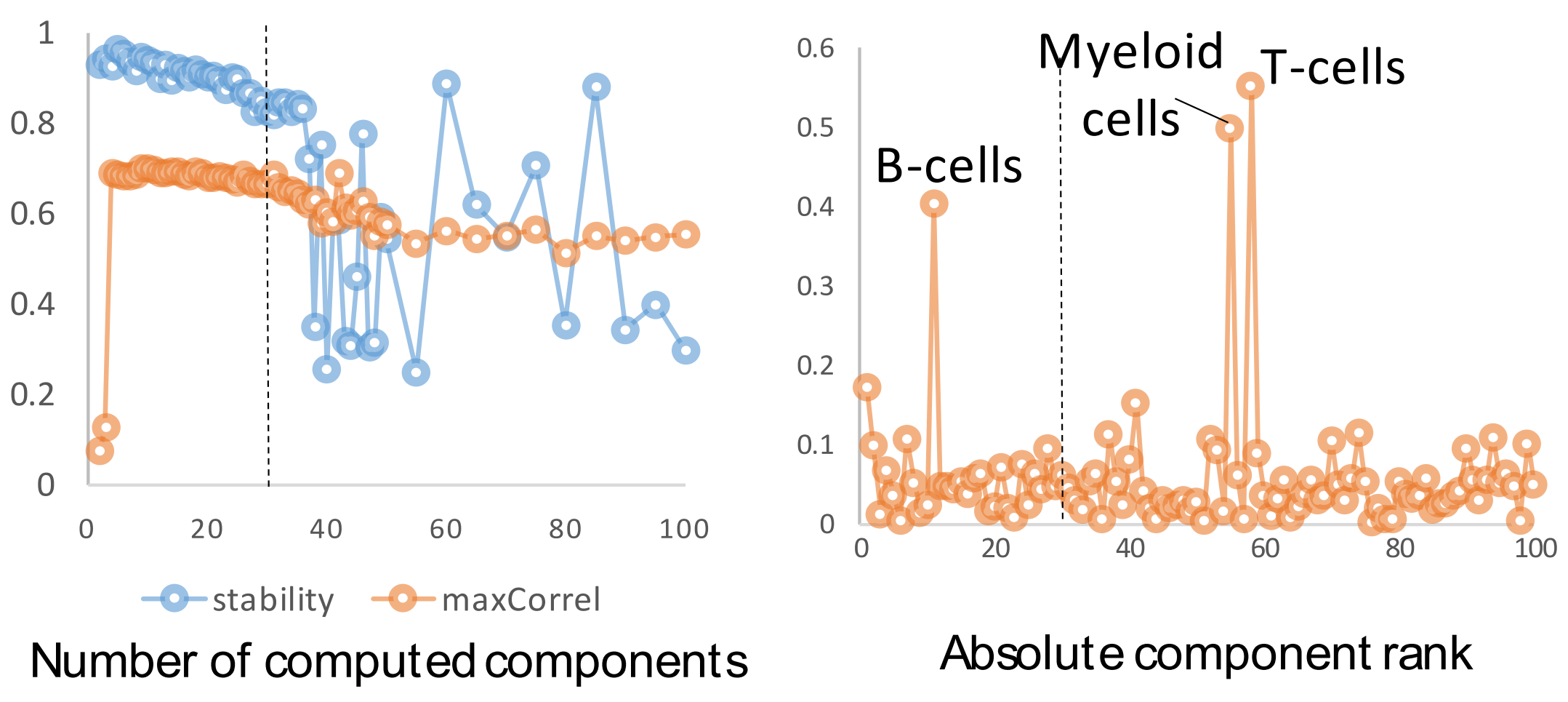


Figure 2. Analysis of reproducibility of previously identified metagenes in independent components of the METABRIC dataset. Enrichment in ImmGene[16] signatures, performed with ToppGene (p-value < 0.001)

ABOUT AUTHOR

Urszula Czerwinska

PhD candidate
graduating in 2018
Passionate about Data Science and Analytics
http://urszulaczerwinska.github.io/about/

SCAN ME

DOWNLOAD POSTER!

BIBLIOGRAPHY

1. Fridman et al. (2012)
2. Gnjatic et al. (2017)
3. Zinovyev et al. (2013)
4. Hyvärinen et al. (2001)
5. Comon and Jutten (2010)
6. Pearson et al. (1901)
7. Lee and Seung (1999)
8. Himberg et al. (2003)
9. http://www.cbioportal.org/
10. Abbas et al. (2009)
11. Cardoso et al. (2009)
12. Loi et al. (2007)
13. Newman et al. (2015)
14. Tamayo, et al. (2005)
15. Martignetti et al. (2016)
16. Heng et al. (2008)
17. Chen et al. (2009)

ACKNOWLEDGEMENTS

PSL RESEARCH UNIVERSITY PARIS
institut Curie
Inserm
MINES ParisTech