Application of Independent Component Analysis to Tumor Transcriptomes Reveals Specific and Reproducible Immune-Related Signals

Urszula Czerwinska[1,3(✉)], Laura Cantini[1], Ulykbek Kairov[2], Emmanuel Barillot[1], and Andrei Zinovyev[1]

[1] Institut Curie, INSERM U900, PSL Research University, Mines ParisTech, 26 rue d'Ulm, Paris, France
urszula.czerwinska@curie.fr
[2] Laboratory of Bioinformatics and Computational Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan
[3] Center for Interdisciplinary Research, Paris Descartes University, Paris, France
https://sysbio.curie.fr/

AQ1

Abstract. Independent Component Analysis (ICA) can be used to model gene expression data as an action of a set of statistically independent hidden factors. The ICA analysis with a downstream component analysis was successfully applied to transcriptomic data previously in order to decompose bulk transcriptomic data into interpretable hidden factors. Some of these factors reflect the presence of an immune infiltrate in the tumor environment. However, no foremost studies focused on reproducibility of the ICA-based immune-related signal in the tumor transcriptome. In this work, we use ICA to detect immune signals in six independent transcriptomic datasets. We observe several strongly reproducible immune-related signals when ICA is applied in sufficiently high-dimensional space (close to one hundred). Interestingly, we can interpret these signals as cell-type specific signals reflecting a presence of T-cells, B-cells and myeloid cells, which are of high interest in the field of oncoimmunology. Further quantification of these signals in tumoral transcriptomes has a therapeutic potential.

Keywords: Blind source separation · Unsupervised learning
Genomic data analysis · Cancer · Immunology

1 Introduction

In many fields of science (biology, technology, sociology) observations on a studied system represent complex mixtures of signals of various origins. It is known that tumors are engulfed in a complex microenvironment (TME) that critically impacts progression and response to therapy. In the light of recent findings [1],

Y. Deville et al. (Eds.): LVA/ICA 2018, LNCS 10891, pp. 1–13, 2018.
https://doi.org/10.1007/978-3-319-93764-9_46

many cancer biologists believe that the state of tumor microenvironment (in particular, the composition of immune system-related cells) defines the long-term effect of the cancer treatment.

In biological systems information is coded in a form of DNA that do not vary a lot between different individuals of the same species. In order to trigger a function in an organism, a part of the DNA is transcribed to RNA, depending on the intrinsic and extrinsic factors, and after additional modification messenger RNA (mRNA) is translated into a protein (i.e. digestive enzyme) that fulfill a role in the organism. The mRNA information (also called transcriptome) can be captured with experimental methods at high throughput (transcriptomics) and provides an approximation of the state of the studied system (i.e. a tissue).

Given the way transcriptomic data is collected, in the resulting dataset, for each observation or sample, the measured transcripts' expression (a putative gene expression that is transcribed to mRNA, and before it is translated to a protein) level is affected by a mixture of signals coming from various sources. Thus, we adopt a hypothesis that a transcriptome is a mixture of different signals (that can be biological or technical), including cell-type specific signals.

Recent works [2–4] showed that expression data from complex tissues (such as tumor microenvironment) can be used to estimate the cell-specific expression profiles of the main cellular components present in a tumor sample. This methodology is based on a linear model of a mixture of signals and their interaction and termed cell-type deconvolution. The mentioned methods take advantage of the prior knowledge (and, at the same time, heavily depend) on the specific transcriptomic signatures (characteristic genes and their weights) of cell types composing TME; therefore, they fall into supervised learning category.

A methodology using an unsupervised data decomposition was applied, so far, in the context of tumor clonality deconvolution by Roman et al. [5]. Some attempts were made to apply Non-negative Matrix factorization to transcriptomic data as well. However, they were either applied in very simplified context of *in vitro* cell mixtures [6] or without a specific focus on the immune signals [7].

In our work, we propose to apply an unsupervised method that will decompose mixture into hidden sources, which will be as independent as possible, based uniquely on data structure and without any prior knowledge. For this purpose, we apply Independent Component Analysis (ICA) [8] that solves blind source separation problem. ICA defines a new coordinate system in the multidimensional space such that the distributions of the data point projections on the new axes become as mutually independent as possible. To achieve this, the standard approach is maximizing the non-gaussianity of the data point projection distributions.

As a result of ICA, conventionally, data matrix X can be approximated: $X \approx AS$, where X is a matrix of data of size $m \times n$, A is a $m \times k$ matrix, $k < m$ and S is $k \times n$ matrix [9]. In our pipeline, input data matrix $n \times m$ (n genes/probes in rows and m samples in columns) is first transposed before applying ICA to $m \times n$. Thus columns of A ($m \times k$) can be named components (m-dimensional vectors) of mixing proportions for each sample m. The S matrix

$(k \times n)$ is transposed to $n \times k$ where rows are projections of data vectors onto the components (a k-dimensional vector for each of n data points).

ICA has been applied for the analysis of transcriptomic data for blind separation of biological, environmental and technical factors affecting gene expression [9–13].

The interpretation of the results of any matrix factorization-based method applied to transcriptomics data is done by the analysis of the resulting pairs of metagenes and metasamples, associated to each component and represented by sets of weights for all genes and all samples, respectively [7,9]. Standard statistical tests applied to these vectors can then relate a component to a reference gene set (e.g., cell cycle genes), or to clinical annotations accompanying the transcriptomic study (e.g., tumor grade). The application of ICA to multiple expression datasets has been shown to uncover insightful knowledge about cancer biology [11,14]. In [11] a large multi-cancer ICA-based metaanalysis of transcriptomic data defined a set of metagenes associated with factors that are universal for many cancer types. Metagenes associated with cell cycle, inflammation, mitochondria function, GC-content, gender, basal-like cancer types reflected the intrinsic cancer cell properties.

In our previous work, we introduced a ranking of independent components based on their stability in multiple independent components computation runs and define a distinguished number of components (Most Stable Transcriptome Dimension, MSTD) corresponding to the point of the qualitative change of the stability profile [15].

However, an interesting observation can be made employing a number of components going far beyond the MSTD ($M \gg$ MSTD), that we call here *overdecomposition*. Applying this approach, one can discover more specific components that remain reproducible between independent datasets. In this work, we present results of overdecomposition with focus on the fine decomposition of the immune signal into cell-type specific signals.

In this analysis, we used a set of six independent breast cancer transcriptomic datasets (BRCATCGA [16], METABRIC [17], BRCACIT [18], BRCABEK [19], BRCAWAN [20] and BRCABCR [21]) to evaluate a detectability and a reproductibility of the immune cell-type related signal. Each dataset contains gene expression measured in breast tumor biopsy for a number of patients. Therefore each measured gene expression here can be a mix of expression from different cells: tumor cells, stroma cells (fibroblasts), immune cells or normal connective tissue.

Throughout this publication we employ terms: *stability*, *conservation* and *reproductibility* that we define as follows. Stability of an independent component, in terms of varying the initial starts of the ICA algorithm, is a measure of internal compactness of a cluster of matched independent components produced in multiple ICA runs for the same dataset and with the same parameter set but with random initialization. Conservation of an independent component in terms of choosing various orders of the ICA decomposition is a correlation between matched components computed in two ICA decompositions of different orders (reduced data dimensions) for the same dataset. Reproducibility of an independent

component is an (average) correlation between the components that can be matched after applying the ICA method using the same parameter set but for different datasets. We claim that if a component is reproduced between the datasets of the same cancer type, then it can be considered a reliable signal less affected by technical dataset peculiarities. If the component is reproduced in datasets from many cancer types, then it can be assumed to represent a universal cancerogenesis mechanism, such as cell cycle or infiltration by immune cells.

2 Methods

2.1 ICA Overdecompostion Procedure

Our pipeline can be described as follows. Started with six public transcriptomic data of breast cancer, we apply the fastICA algorithm [8] accompanied by the icasso package [22] to improve the components estimation and to rank the components based on their stability. In order to run the analysis we used open source BIODICA tool (ICA applied to BIOlogical Data), available from https://github.com/LabBandSB/BIODICA. It provides both a command line and a user-friendly Graphical User Interface (GUI) for high-performance ICA analysis, including bootstraping and further stability analysis. It also allows the computation of MSTD index, introduced in [15]. BIODICA software links to downstream analysis enabling the interpretation of components, such as standard statistical methods, i.e. enrichment test, and non-standard methods, such as using projection on top of molecular maps (InfoSigMap, [23]). The downstream analysis was not exhaustively employed in this publication as we focused on specific immune signals.

ICA was applied to each transcriptomic dataset separately. For each analyzed transcriptomic dataset, we computed M independent components (ICs), using *pow*3 nonlinearity and symmetrical approach to the decomposition. The number of dimensions was set to 100 ($M = 100$) as it is significantly greater than MSTD for these datasets (that is in the order of $M = 30$). Each component of the resulting S matrix was oriented in the direction of its heavy tail, being defined as the tail with the maximum sum of absolute weight values, so that it always has the positive sign.

2.2 Interpretation of Components

In order to confirm that we can recover expected known signals performing the overdecomposition procedure, we correlate reference metagenes with the S matrix. Correlations are performed on common genes for each component and metagene. The result was graphically represented using R package *ggplot*2 [24]. An interpretation is assigned to a component only if its assignment is reciprocal. In our analysis reciprocity is defined as follows. Given correlations between the set of metagenes $M = \{M_1, ..., M_m\}$ and S matrix $S = \{IC_1, ..., IC_N\}$, if $S_i = argmax_k(corr(M_j, S_k))$ and $M_j = argmax_k(corr(S_i, M_k))$, then S_i

and M_j are reciprocal. In this way, the breast cancer metagenes were matched against the following set of previosuly defined metagenes [11] - reference metagenes: MYOFIBROBLASTS, BLCAPATHWAYS, STRESS, GC CONTENT, SMOOTH MUSCLE, MITOCHONDRIAL TRANSLATION, INTERFERON, BASALLIKE, CELLCYCLE, UROTHERIALDIFF. Details about construction of reference metagenes and their interpretation can be found in Biton et al. 2014 [11]. The correlation plot was visualized in Cytoscape 2.8 [25].

2.3 Selecting Immune-Related Components

In order to preselect immune-related signals, we focused on all Independent Components (ICs) with Pearson correlation > 0.1 between IMMUNE metagene and ICs (columns of the S matrix). The interpretation was given using Fisher exact test on 100 top-ranked genes of each of the preselected components and Immgen [26] signatures containing in total 6467 genes of six immune cell types: $\alpha\beta$T-cells, $\gamma\delta$T-cells, B-cells, CD+, Myeloid cells, NK cells and four non-immune cell types: Fetal-Liver, Stem cells, Stromal cells and Pasmocytoid, 241241 signatures in total, each of 480 genes in average.

2.4 Comparing Independent Components from Different Datasets

Following the methodology developed previously in [11], the metagenes computed in two independent datasets were compared by computing a Pearson correlation coefficient between their corresponding gene weights. Since each dataset can contain a different set of genes, the correlation is computed on the genes which are common for a pair of datasets. Note that this common set of genes can be different for different pairs of datasets. The same correlation-based comparison was done with previously defined and annotated metagenes. In all correlation-based comparisons, the absolute value of the correlation coefficient was used.

3 Results

3.1 Most of Known Metagenes Can Be Found in Overdecomposed Datasets

In all six overdecomposed datasets of breast cancer, we could find major reference metagenes [11]. As an example, we present results for METABRIC dataset [17] (Fig. 1) where we can observe correlations between metagenes and all 100 ICs. For some metagenes (MYOFIBROBLASTS, INTERFERON, MITOCHONDRIAL TRANSLATION, CELL CYCLE), there is only one reciprocal and strongly (>0.3) correlated component, which can be understood as a good signal reproducibility. Some other as STRESS, BASALLIKE and SMOOTH MUSCLE can have two similarly correlated components. This is probably due to component split in higher-order decomposition. Importantly, reference metagenes were

defined in significantly lower dimensional space ($M = 25$) and as a result of high-dimensional decomposition, these signals are decomposed to more specific sources that can still be interpreted in biological terms. For few components, no strong correlations with metagenes were found (UROTHELIALDIFFERENTIATION and BLCPATHWAYS). As these metagenes are more specific to Bladder cancer, we can consider them as negative control here. Also, GC Content and IMMUNE metagenes have several corresponding components. The IMMUNE metagene is considered here as a special case as we can find several components correlated to it and, in addition, their interpretation can be interesting for biological applications. We investigate more about the immune-related components in the Subsect. 3.3.

3.2 Reproducibility of the Signals in Breast Cancer Datasets

It would be reasonable to expect that the main biological signals are characteristic for a given cancer type. Thus, they should be the same when one studies molecular profiles of different independent cohorts of patients. For this reason, we expect that for multiple datasets related to the same cancer type, the ICA decompositions should be somewhat similar; hence, reciprocally matching each other.

We correlated the ICA overdecompositions of all six datasets with each other and with the forementioned metagenes [11]. One can notice from the correlation graph (Fig. 2A), that some pseudo-cliques characterized with strong correlation coefficient (thick edges) and reciprocal (green) edges are present in the mass of low correlation coefficients edges. If the edges with correlation coefficient < 0.4 are filtered out, we can better visualize a collection of pseudo-cliques (Fig. 2B). Some of those pseudo-cliques are connected to a metagene and can be given an interpretation directly, some others would need a further investigation of the gene signature in order to attribute a meaning to them. We can see that in some pseudo-cliques not all datasets are represented. It may suggest that some signals, still reproducible, are not representative for all datasets. In order to explain, why a signal is missing, one should first interpret the signal, then try to understand the similarities or differences of samples based on provided metadata. From our previous analysis [11], the components that do not find reciprocity (absent from the pseudo-cliques) are either dataset specific or they correspond to unknown batch effects that cannot be guessed without an additional knowledge. It is remarkable that despite overdecompostion, the metagenes conceived in lower-dimensional space are highly conserved and reproducible, which suggests the overdecompostion does not diminish strong signals conceived in "optimal" dimensional space (i.e. MSTD). Of note, these datasets were produced using various technologies of transcriptomic profiling.

3.3 Three Pseudo-cliques Related to Three Immune Cell Types

To better understand the reproductibility of the immune-related signal, we extracted only components correlated with IMMUNE > 0.1. Hence, we obtain

three strongly connected cliques (Fig. 3) and some disconnected components. We interpreted each of the ICs with an enrichment test. The results of Fisher exact test indicate mainly three cell types T-cell, B-cell and Myeloid cells with a p-value < 0.05 as indicated in the Fig. 3. While T-cell and Myeloid cell are indicated with very high certainty, the B-cell signal seems to be more complex. The results of the enrichment test for the B-cell component are less explicit as among the most enriched pathways, different cell types (T-cells and Natural Killers) are listed together with dominating B-cell signal. However, this can be explained by functional and phenotypic similarities between NK and B cells [27]. Also, T cell and B cell as they are both lymphocytes, they share common features. It is worth highlighting that definition of cell type signature is a part of ongoing debate [28] and here we use them as an indicator of possible signal definitions. Also, some ICs belonging to one pseudo-clique are correlated (with lower coefficients) with ICs from another pseudo-clique (i.e. BRCABCR IC2). It may suggest an inclination of the signal towards the other phenotype. As far as components not included in pseudo-cliques are concerned, through interpretation BRCACIT IC42 can be associated with B cells, METABRIC IC28 with Myeloid cells, BRCAWAN IC68 and BRCABEK IC27 with T-cells. Thus, the correlations of the disconnected components, even though they are low, they are most probably not spurious. Some other components not included in the pseudo-cliques like BRCAWAN IC28 and BRCABCR IC19 seem to contain stroma elements. It would be worth understanding more deeply the nature of each signal and interpret in terms of biological functions or sub-phenotypes.

4 Discussion

The overdecomposition of six breast cancer datasets, where different normalization methods and different transcriptome profiling platforms were used, showed that even in high order blind source separation, the ICA-based analysis can be reproducible between datasets. Moreover, the most stable signals are conserved and not affected by the number of dimensions. Interestingly, for some signals we can observe a split into more specific signals that can still be interpreted in biological terms. In the case of the immune-related signals, it allows robust reproduction of three main signals that form pseudo-cliques on the correlations graph in the Fig. 3. This result let us believe that ICA allows separation of signals in cancer transcriptomes in an unsupervised manner and detect the most represented immune cell-types. We found highly interesting that technically non-stable signal is found reproducible and interpretable in the six breast cancer datasets.

The question about the choice of ICA over other available blind source separation methods can be asked. We address this question more extensively in a publication in preparation comparing NMF, ICA and PCA for transcriptome BSS. From our expertise (unpublished data) NMF applied to transcriptomes can effectively separate sources and their proportions (proven in controlled mixtures of different cell types or tissues). However, when NMF was applied to noisy tumor

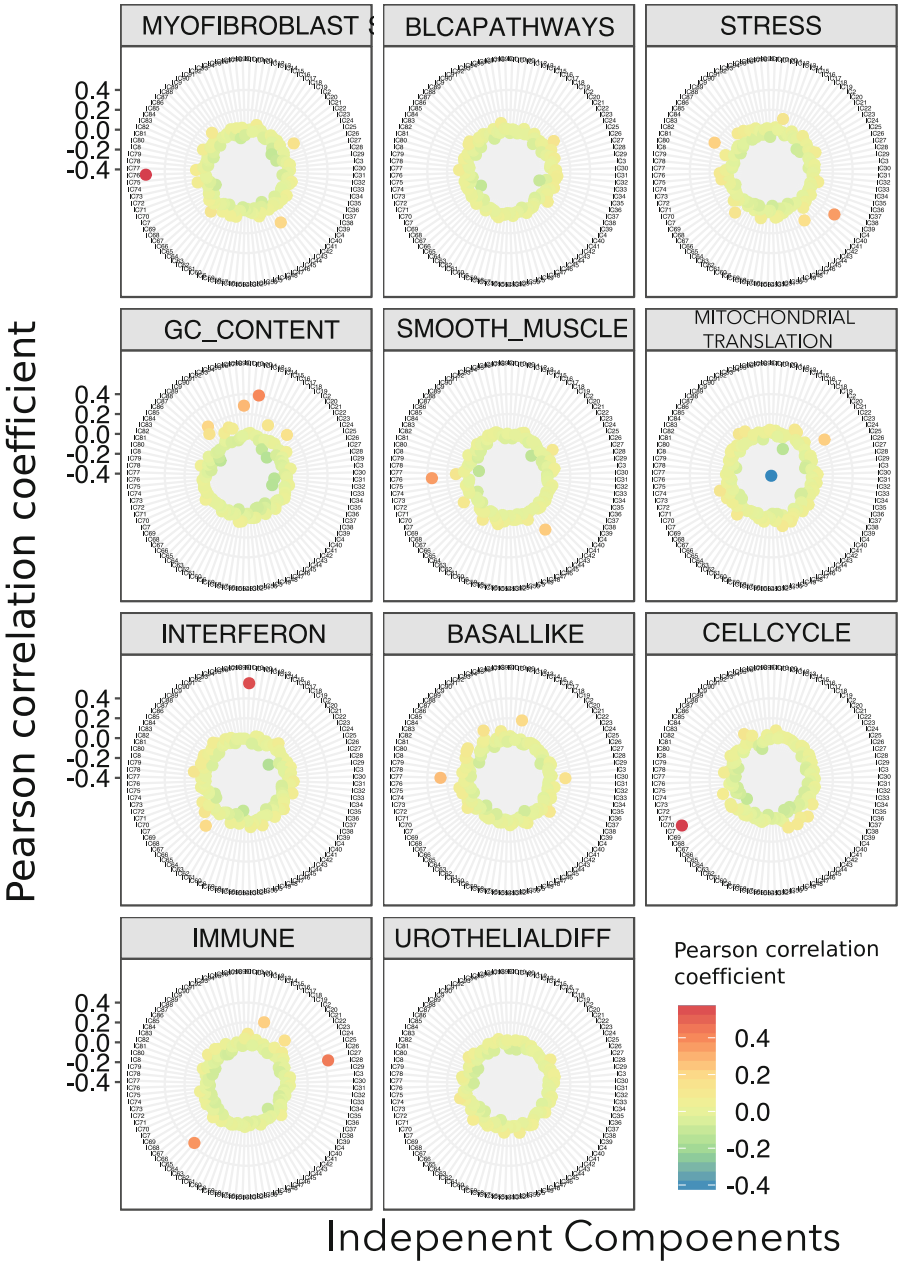

Fig. 1. Correlations between 11 metagenes [11] **and 100 independent components of METABRIC dataset** [17]. Each panel shows correlation coefficients between a given metagene and 100 ICs of METABRIC, the components are ordered in the same manner for all panels from 1 to 100 in a circle. For a high correlation coefficient, the point is red, for low, it is blue (see legend). (Color figure online)

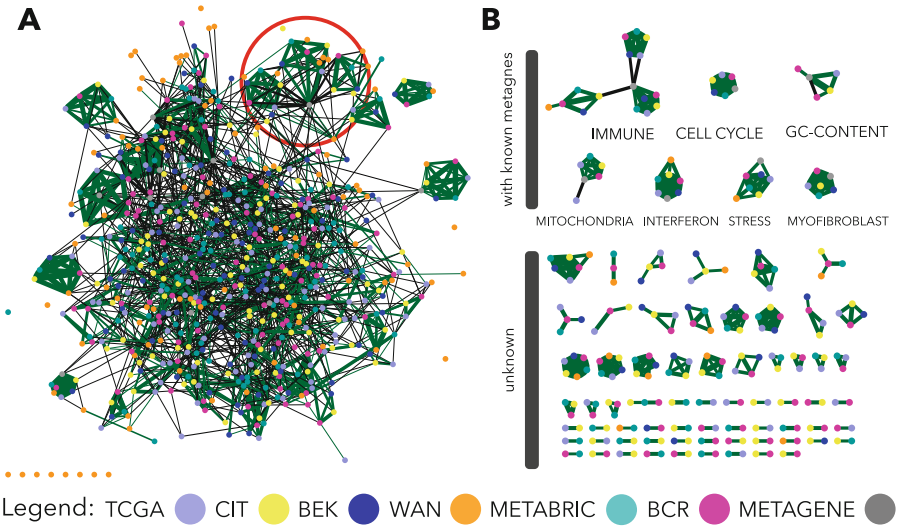

Fig. 2. Correlation plot of six tumor datasets and the reference metagenes [11] A- Correlation graph between decompositions into 100 ICs of the six transcriptomic datasets and the 11 reference metagenes. The IMMUNE metagene and related ICs in encircled; B - collection of pseudo-cliques extracted from the correlation graph A through filtering out edges of the Pearson correlation coefficient < 0.4. They were split in two groups, the ones that are directly interpretable via their correlation with a metagene and cliques that are not related to any known metagene; The thickness of edges is proportional to the Pearson correlation coefficients, green color indicates reciprocity of edges, colors of nodes indicate dataset (see legend). (Color figure online)

transcriptomes, obtained source profiles were not highly reproducible between different datasets. Our unpublished research showed that NMF profiles are highly affected by mean gene expression. Therefore, NMF decomposition applied to breast cancer transcriptomes followed by correlation of obtained profiles did not reveal meaningful pseudo-cliques as the ICA-based analysis discussed in this article.

In order to translate our findings into real biomedical application, more time should be dedicated to analyze ICA signatures in details, to report their similarities and differences. As well as, this analysis could be applied in a pan-cancer manner to observe the reproducibility of the signal among different tumor types. Such an analysis would possibly identify components and/or genes linked with patients' survival or response to treatment and eventually, use them to compose a predictive score for tumor immune therapy outcome.

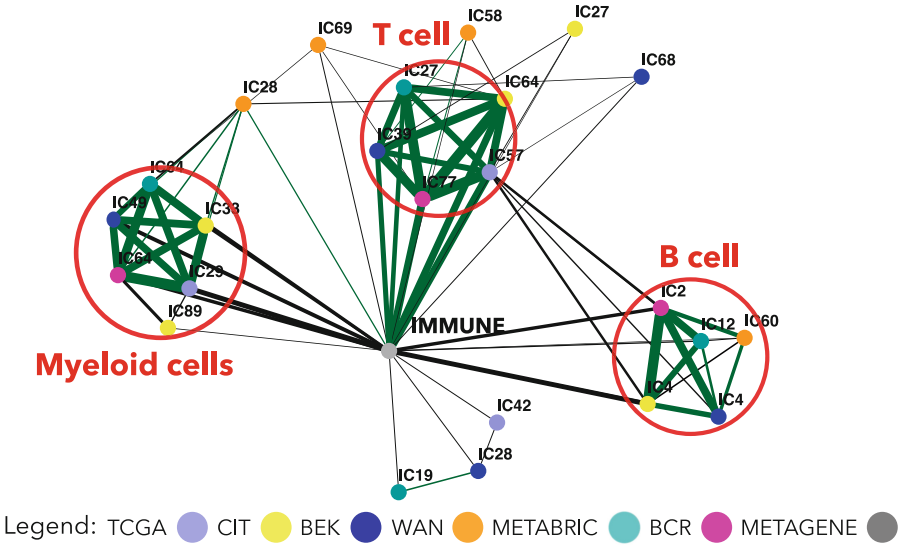

Fig. 3. Correlation graph of ICs correlated with IMMUNE metagene > 0.1. Three pseudo-cliques are encircled and labeled according to the results of Fisher exact test. The thickness of edges is proportional to the Pearson correlation coefficients, green color indicates reciprocal edges, colors of nodes indicate dataset (see legend). (Color figure online)

5 Conclusions

We applied overcomposition into one hundred components of six transcriptomic datasets using Independent Components Analysis, a blind source separation algorithm. We used a known collection of ranked ICA-derived genetic signatures (that we call reference metagenes) to conclude that most of the signals are conserved in the higher dimensions. We noticed that some of the components split into more specific signals. Our correlation analysis of the ICA overdecompositons of the transcriptomes stated that majority of components are reproducible between datasets. Our more focused investigation of immune-related ICs demonstrated that three cell types can be named: T-cell, B-cell and myeloid cells as a reproducible source signal in the breast cancer datasets. Further interpretation of those cell-type related genomic signatures can find application in immuno-oncology therapeutics as predictive biomarkers for immunotherapies.

Acknowledgments. We thank Vassili Soumelis for discussions on multidimensionality of biological systems. This work has been funded by INSERM Plan Cancer N BIO2014-08 COMET grant under ITMO Cancer BioSys program and by ITMO Cancer (AVIESAN) who provided 3-year PhD grant. We would like to acknowledge as well foundation Bettencourt Schueller and Center for Interdisciplinary Research funding for the training of the PhD student.

References

1. Swartz, M.A., Iida, N., Roberts, E.W., Sangaletti, S., Wong, M.H., Yull, F.E., Coussens, L.M., DeClerck, Y.A.: Tumor microenvironment complexity: emerging roles in cancer therapy (2012)
2. Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H., et al.: Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. **17**(1), 218 (2016)
3. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods **12**(5), 453–457 (2015)
4. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D.E., Gfeller, D.: Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife **6**, e26476 (2017)
5. Roman, T., Xie, L., Schwartz, R.: Automated deconvolution of structured mixtures from heterogeneous tumor genomic data. PLoS Comput. Biol. **13**(10), e1005815 (2017)
6. Gaujoux, R., Seoighe, C.: Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. Infect. Genet. Evol. **12**(5), 913–921 (2012)
7. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. **101**(12), 4164–4169 (2004)
8. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**(45), 411–430 (2000)
9. Zinovyev, A., Kairov, U., Karpenyuk, T., Ramanculov, E.: Blind source separation methods for deconvolution of complex signals in cancer biology. Biochem. Biophys. Res. Commun. **430**(3), 1182–1187 (2013)
10. Teschendorff, A.E., Journée, M., Absil, P.A., Sepulchre, R., Caldas, C.: Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput. Biol. **3**(8), 1539–1554 (2007)
11. Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., DeReyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A., Radvanyi, F.: Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. **9**(4), 1235–1245 (2014)
12. Gorban, A., Kegl, B., Wunch, D., Zinovyev, A.: Principal Manifolds for Data Visualisation and Dimension Reduction. Lecture notes in Computational Science and Engineering, vol. 58, p. 340. Springer, Heidelberg (2008)
13. Saidi, S.A., Holland, C.M., Kreil, D.P., MacKay, D.J.C., Charnock-Jones, D.S., Print, C.G., Smith, S.K.: Independent component analysis of microarray data in the study of endometrial cancer. Oncogene **23**(39), 6677–6683 (2004)
14. Bang-Berthelsen, C.H., Pedersen, L., Fløyel, T., Hagedorn, P.H., Gylvin, T., Pociot, F.: Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. BMC Genomics **12**, 97 (2011)
15. Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., Zinovyev, A.: Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics **18**(1), 712 (2017)

16. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., et al.: The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**(10), 1113 (2013)
17. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Aparicio, S., Brenton, J.D., Ellis, I., Huntsman, D., Pinder, S., Murphy, L., Bardwell, H., Ding, Z., Jones, L., Liu, B., Papatheodorou, I., Sammut, S.J., Wishart, G., Chia, S., Gelmon, K., Speers, C., Watson, P., Blamey, R., Green, A., MacMillan, D., Rakha, E., Gillett, C., Grigoriadis, A., De Rinaldis, E., Tutt, A., Parisien, M., Troup, S., Chan, D., Fielding, C., Maia, A.T., McGuire, S., Osborne, M., Sayalero, S.M., Spiteri, I., Hadfield, J., Bell, L., Chow, K., Gale, N., Kovalik, M., Ng, Y., Prentice, L., Tavaré, S., Markowetz, F., Langerød, A., Provenzano, E., Purushotham, A., Børresen-Dale, A.L., Caldas, C.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346–352 (2012)
18. Guedj, M., Marisa, L., De Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A.L., Feugeas, J.P., Bièche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., De Thé, H., Theillet, C.: A refined molecular taxonomy of breast cancer. Oncogene **31**(9), 1196–1206 (2012)
19. Bekhouche, I., Finetti, P., Adelaïde, J., Ferrari, A., Tarpin, C., Charafe-Jauffret, E., Charpin, C., Houvenaeghel, G., Jacquemier, J., Bidaut, G., Birnbaum, D., Viens, P., Chaffanet, M., Bertucci, F.: High-resolution comparative genomic hybridization of Inflammatory breast cancer and identification of candidate genes. PLoS ONE **6**(2), e16950 (2011)
20. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-Van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M., Atkins, D., Foekens, J.A.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet **365**(9460), 671–679 (2005)
21. Reyal, F., Rouzier, R., Depont-Hazelzet, B., Bollet, M.A., Pierga, J.Y., Alran, S., Salmon, R.J., Fourchotte, V., Vincent-Salomon, A., Sastre-Garau, X., Antoine, M., Uzan, S., Sigal-Zafrani, B., de Rycke, Y.: The molecular subtype classification is a determinant of sentinel node positivity in early breast carcinoma. PLoS ONE **6**(5), e20297 (2011)
22. Himberg, J., Hyvärinen, A.: ICASSO: software for investigating the reliability of ICA estimates by clustering and visualization. In: Neural Networks for Signal Processing - Proceedings of the IEEE Workshop, vol. 2003, pp. 259–268, January 2003
23. Cantini, L., Calzone, L., Martignetti, L., Rydenfelt, M., Blüthgen, N., Barillot, E., Zinovyev, A.: Classification of gene signatures for their information value and functional redundancy. npj Syst. Biol. Appl. **4**(1), 2 (2018)
24. Wickham, H.: ggplot2 Elegant Graphics for Data Analysis, vol. 35 (2009)
25. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software Environment for integrated models of biomolecular interaction networks. Genome Res. **13**(11), 2498–2504 (2003)
26. Shay, T., Kang, J.: Immunological Genome Project and systems immunology (2013)

27. Kerdiles, Y.M., Almeida, F.F., Thompson, T., Chopin, M., Vienne, M., Bruhns, P., Huntington, N.D., Raulet, D.H., Nutt, S.L., Belz, G.T., Vivier, E.: Natural-Killer-like B cells display the phenotypic and functional characteristics of conventional B cells. Immunity **47**(2), 199–200 (2017)
28. Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B., Raue, A.: Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nature Commun. **8**(1), 2032 (2017)