# Assessing reproducibility of matrix factorization methods in independent transcriptomes

Laura Cantini, Ulykbek Kairov, Aurélien de Reyniès, Emmanuel Barillot, François Radvanyi, Andrei Zinovyev

*Subject Section*

# Assessing reproducibility of matrix factorization methods in independent transcriptomes

Laura Cantini[1,2,3,4*], Ulykbek Kairov[6], Aurélien de Reyniès[7], Emmanuel Barillot[1,2,3], François Radvanyi[8,9] and Andrei Zinovyev[1,2,3,5*]

[1]Institut Curie, PSL Research University, F-75005 Paris, France,[2]INSERM U900, F-75005 Paris, France,[3]Mines ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France,[4]Computational Systems Biology Team, Institut de Biologie de l'École Normale Supérieure, CNRS UMR8197, INSERM U1024, École Normale Supérieure, PSL Research University, 75005 Paris, France,[5]Lobachevsky University, Nizhni Novgorod, Russia,[6]Laboratory of bioinformatics and systems biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan, [7]Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France, [8]Institut Curie, PSL Research University, CNRS, UMR144, Equipe Labellisée Ligue Contre le Cancer, 75005 Paris, France, [9]Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR144, 75005 Paris,

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** Matrix factorization (MF) methods are widely used in order to reduce dimensionality of transcriptomic datasets to the action of few hidden factors (metagenes). MF algorithms have never been compared based on the between-datasets reproducibility of their outputs in similar independent datasets. Lack of this knowledge might have a crucial impact when generalizing the predictions made in a study to others.

**Results:** We systematically test widely-used MF methods on several transcriptomic datasets collected from the same cancer type (14 colorectal, 8 breast and 4 ovarian cancer transcriptomic datasets). Inspired by concepts of evolutionary bioinformatics, we design a novel framework based on Reciprocally Best Hit (RBH) graphs in order to benchmark the MF methods for their ability to produce generalizable components. We show that a particular protocol of application of Independent Component Analysis (ICA), accompanied by a stabilization procedure, leads to a significant increase in the between-datasets reproducibility. Moreover, we show that the signals detected through this method are systematically more interpretable than those of other standard methods. We developed a user-friendly tool BIODICA for performing the Stabilized ICA-based RBH meta-analysis. We apply this methodology to the study of colorectal cancer (CRC) for which 14 independent transcriptomic datasets can be collected. The resulting RBH graph maps the landscape of interconnected factors associated to biological processes or to technological artifacts. These factors can be used as clinical biomarkers or robust and tumor-type specific transcriptomic signatures of tumoral cells or tumoral microenvironment. Their intensities in different samples shed light on the mechanistic basis of CRC molecular subtyping.

**Availability:** The RBH construction tool is available from http://goo.gl/DzpwYp

**Contact:** laura.cantini@curie.fr andrei.zinovyev@curie.fr

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Large-scale cancer genomics projects, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), are generating an overwhelming amount of transcriptomic data. These data offer us the unprecedented opportunity to understand cancer, its onset, progression and response to treatment. To deal with the high-dimensionality of transcriptomic data, Matrix factorization (MF) approaches, reducing high-dimensional data into low dimensional subspaces, are widely employed (Stein-O'Brien *et al.*, 2017; Kim and Tidor, 2003). Given the natural representation of a transcriptomic dataset as a matrix X (n x m) with n genes in the rows and m samples in the columns, MFs decompose X into the product of an unknown mixing matrix A (n x k) and an unknown matrix of source signals S (k x m). In the following, we denote the columns of A as "metagenes" and the rows of S as "metasamples". The rationale behind MF usage in biology is that the state of a biological sample, such as a tumor sample, is determined by multiple concurrent biological factors, from generic processes such as proliferation and inflammation to cell-type specific ones. Transcriptomic data can be thus interpreted as a complex mixture of various biological signals convoluted with technical noise of various kind (Avila Cobos *et al.*, 2018; Brunet *et al.*, 2004).

The MF methods most widely applied to trascriptomic data are Principal Component Analysis (PCA), Non Negative Matrix Factorization (NMF) and Independent Component Analysis (ICA) (Biton *et al.*, 2014; Ma and Dai, 2011; Devarajan, 2008; Alter *et al.*, 2000). We will here consider the original NMF algorithm by Lee and Sung (Lee and Seung, 1999; Ochs *et al.*, 1999), while for ICA three variants of the same fastICA algorithm (Himberg and Hyvarinen; Hyvarinen, 1999) will be considered: "Stabilized ICA (sICA)" the protocol previously proposed by us that maximizes kurtosis of metagenes and searches for stable components (Biton *et al.*, 2014; Kairov *et al.*, 2017); "ICA" that maximizes kurtosis of metagenes without stabilization and " ICA' " the application of ICA that maximizes kurtosis of metasamples (see Supp Text 1). A component output of any of these MF methods potentially recapitulates a biological signal that can be rediscovered in another independent dataset of the same kind (e.g., in independently profiled cohort of the same cancer type). If this is the case, we call such a component *reproducible*. Here we will evaluate the reproducibility of the above MF methods, i.e. their capability to identify many reproducible components. Note that this definition is different from other metrics of MF reproducibility, such as subsampling and cross-validation (Molinaro *et al.*, 2005). Surprisingly, little is known about the level of between-dataset reproducibility of various MF methods when applied to transcriptomic data. Lack of this knowledge might have a crucial impact when extrapolating predictions made in a particular study to future transcriptomic studies of the same kind.

In this manuscript we developed a framework for assessing the reproducibility of MF methods. The metrics is based on exploiting Reciprocal Best Hit (RBH) relations between MF metagenes and quantifying structural properties of the RBH graph. Given its ultimate aim, our framework evaluates the reproducibility of components independently identified from multiple datasets, differently from multi-level factorizations that co-factorize multiple datasets as a whole (Argelaguet *et al.*, 2018; Tenenhaus *et al.*, 2017).

We applied our framework based on the RBH graph to compare the performances of various MFs (PCA, NMF, sICA, ICA and ICA') in three biological contexts: colorectal, breast and ovarian cancer. We found marked differences in terms of reproducibility among the various MFs. Stabilized ICA (sICA) remarkably outperformed alternative approaches and it valuably reconstructed the landscape of factors shaping cancer transcriptomes.

## 2 Methods

### 2.1 Biological contexts chosen for the comparison

The large number of carefully annotated transcriptomic datasets available in cancer biology and the wide heterogeneity of these data are the reasons that motivated our choice toward using cancer trascriptomes for assessing MF reproducibility. We here use colorectal cancer (CRC), breast cancer (BRCA) and ovarian cancer (OVCA) for our comparison.

CRC and BRCA have been chosen as being among the most studied cancers, especially in the context of transcriptional subtyping (Guinney *et al.*, 2015; Parker *et al.*, 2009). We employed 14 independent CRC datasets and 8 BRCA datasets. In these two test cases both the profiling platform and the cohort of patients are changing across the various datasets. In addition, we chose OVCA to test to which extent the type of profiling platform affects the reproducibility of the different MF methods. Four TCGA ovarian cancer datasets profiled with four different platforms: Affymetrix U133, Agilent and Affymetrix HuEx, plus RNAseq (Bell *et al.*, 2011) have been used. The 418 samples common to all four datasets have been used for our analysis. The samples have been organized into four datasets each of them associated to one of the four platforms and composed of the same samples. See Table S1 for data availability.

### 2.2 Computational framework for metagene comparison

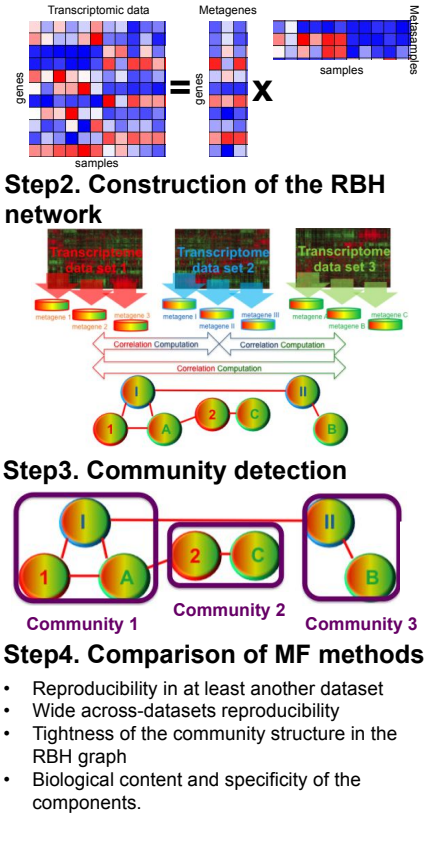



**Fig.1. Schematic representation of MF comparison framework.**

We here introduce a framework to compare four standard MF algorithms: PCA, NMF, ICA, ICA' and sICA (see Figure 1, Supp Text 2). First, the number *k* of components in which the expression matrix is decomposed should be chosen for all the compared MFs. We overdecomposed the matrices and we fixed the same number of components for all the MFs (see Table S1). Overdecomposition here stands for the fact that the selected number of

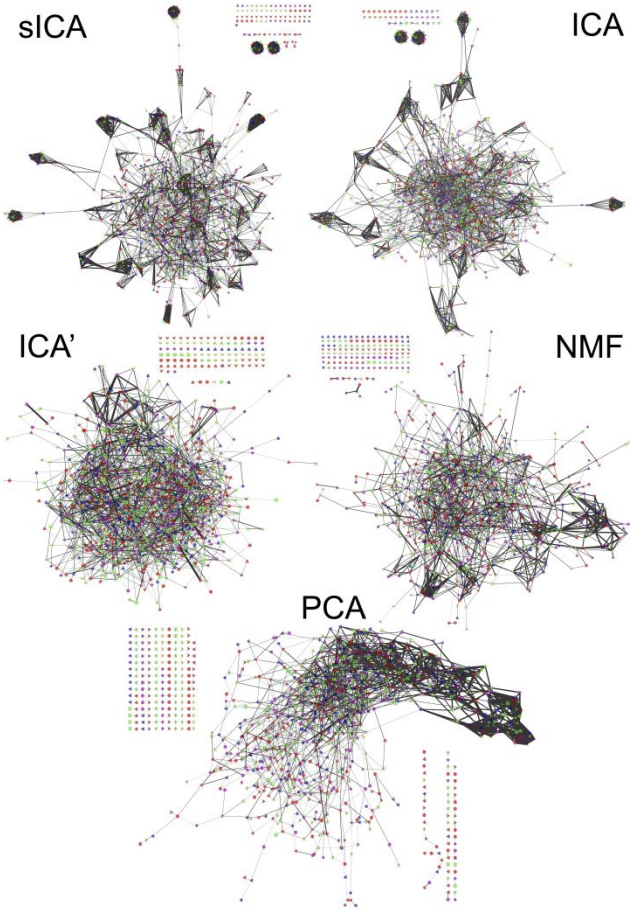
components is taken larger than the estimation of the effective transcriptome dimension.

In our previous work, we have shown that in case of ICA, overdecomposition is not detrimental for the interpretability of the resulting components (Kairov *et al.*, 2017). The same is true for PCA, since the higher-order components do not alter the lower order ones. For NMF the number $k$ of components in which a dataset should be decomposed is frequently decided by looking at the last local maximum of the cophenetic coefficient, summarizing the results of a consensus over different runs of the algorithm (Brunet *et al.*, 2004). We thus chose to also compare our four algorithms against the version of NMF whose number of components is chosen based on the cophenetic coefficient, called in the following "cophNMF". Such comparison, reported in Table S2, did not affect our conclusions.

As shown in Figure 1, our framework is composed of 4 main steps to be separately performed for each MF algorithm. The only inputs required to perform the comparison are as many independent transcriptomic datasets as possible for the same biological context. At step 1, each dataset is decomposed into a set of metagenes and metasamples. At this step, when the variants of ICA and PCA are applied to the input datasets we first perform a centering step, i.e. for each gene expression value we subtract its average expression across all samples. This is a standard procedure aimed at avoiding to capture the signal connected to the genes' average expression, i.e. the vector containing the mean gene expression across all the samples of the dataset, as first component. Of note, the centering could not be applied to NMF due to the non-negativity constraint. In step 2, the graph of reciprocal correspondences between the metagenes obtained from the various independent datasets is reconstructed.

Given the two sets of metagenes $\{M_1 \dots M_k\}$ and $\{N_1 \dots N_k\}$ obtained in step 1 from the trascriptomic datasets $T^m$ and $T^n$, respectively. We here define $M_i$ and $N_j$ as a Reciprocal Best Hit (RBH) iff

$$\max\left(cor\left(M_i, \{N_t\}_{t=1}^{k}\right)\right) = \max\left(cor\left(\{M_t\}_{t=1}^{k}, N_j\right)\right). \quad (1)$$




The procedure (1) is then repeated for all couples of available trascriptomic datasets $T^m$ and $T^n$ and the obtained RBHs are merged into a single graph whose nodes are the metagenes of all transcriptomic datasets and whose links correspond to their RBHs. Here and in the following we will refer to this graph as "RBH graph". This name is chosen in analogy with the namesake common definition of orthology in comparative genomics (Tatusov *et al.*, 1997; Bork *et al.*, 1998). The idea behind our approach is thus to identify orthologous biological factors across different transcriptomic datasets. The RBH approach is free of necessity to define a threshold as opposed to correlation graph construction procedure and it leads to relatively sparse graphs. In Figure S1 we compare the number of RBHs and the dimension of the largest connected component of the correlation graph for various thresholds vs. the RBH network in all the MFs. The RBH construction tool is available from http://goo.gl/DzpwYp as part of "ICA for Big Omics Data (BIODICA)" tool (see Supp Text 3).

Following the reconstruction of the RBH graph, we observed that the components detected by NMF were strongly biased toward the genes' average expression (see Figure S2), i.e. the vector containing in each row the average expression of a gene across all the samples of the dataset. As a further standardization, we thus regressed each metagene over the genes' average expression of the associated dataset and we used the resulting residues in place of the original metagenes to construct the RBH graph. Alternative normalizations of the datasets before the application of NMF have been also considered, but they appeared detrimental for the reproducibility of its metagenes (see Supp Text 4).

At step 3, differently from previous works (Biton *et al.*, 2014; Kairov *et al.*, 2017), communities are detected in the RBH graph using the Markov Clustering algorithm (MCL) (Enright *et al.*, 2002). Such communities reflect the existence of factors strongly reproduced across different transcriptomes. Finally, at step 4 different objective measures are computed to compare the results obtained by the various MFs. The idea in this last step is to evaluate the performances of the different algorithms focusing on measures that are of practical interest to researchers when analyzing high-throughput data. In particular, we evaluated the ability of the different MFs to (i) produce components reproducible in at least one other dataset; (ii) determine widely reproducible components; (iii) derive an RBH graph characterized by a tight community structure and (vi) identify components biologically meaningful and specific, i.e. accurately and univocally predicting known biological signals.

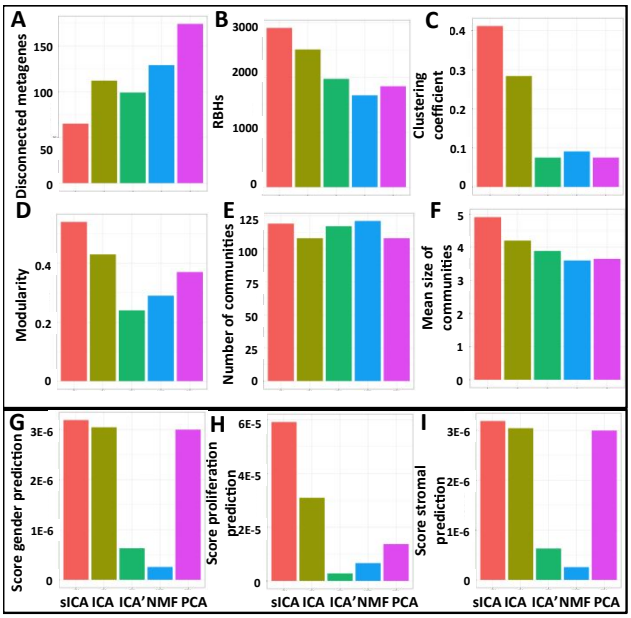
In CRC, we also employed our framework to compare the various MFs to Regularized Generalized Canonical Correlation Analysis (RGCCA), which co-factorizes all the datasets together by explicitly maximizing inter-dataset correlations (Tenenhaus *et al.*, 2017). To this end, we had to restrict the number of genes to 11300 common to all datasets, which is not needed in case of independent MF applications. This evaluation of the performances of RGCCA is aimed at exploring the consistency of our framework, that should in this case achieve the maximal match between components and thus the maximal scores in criteria (i)-(iii). Finally, we characterized the communities obtained in the RBH graph of sICA using the available biological and clinical annotations as described in Supp Text 5.

## 3 Results

**Fig.2. RBH graphs of widely-used MFs built for 14 independent CRC datasets.**

Once steps 1 and 2 have been performed, as discussed in the Methods section, we obtained the RBH graphs visualized in Figure 2. The nodes of these graphs are the metagenes obtained by the different MFs while the links correspond to the presence of an RBH. The topological structure of the obtained graphs is substantially different. The RBH graphs of sICA and ICA are characterized by tight communities and less disconnected nodes in respect to the others. NMF has some areas of densely connected nodes but these are less pronounced in respect to those of sICA. The graph of PCA reflects the hierarchical structure of the principal components (PC). A densely connected area can be indeed identified in the lower part of the graph, where the first, second and third PCs are localized. This topological organization is lost when going toward higher-order components. Finally, the graph of ICA' has a surprisingly divergent structure in respect to the one of sICA, with a much lower number of tight communities. This last result suggests that the protocol used to apply ICA has a strong impact in the obtained RBH graph. Similar conclusions on the RBH graph topology have been made when we tested the effect of subsampling onto MFs applied to the same dataset (Supp Text 6 and Figure S3).

The qualitative characteristics here discussed will be extensively tested in the next sections, devoted to the comparison of the measures defined as step 4 of our framework.




### 3.1 Reproducibility in at least one other dataset

Having multiple independent transcriptomic datasets from the same biological condition (in our case CRC, BRCA or OVCA), we can expect to have similar biological factors captured by the MF in at least few datasets. As a consequence, a metagene should find a RBH in at least one other dataset. This may not happen if the metagene captures a technical dataset-specific bias or a rare subpopulation of tumors uniquely present in one dataset or due to the inability of an MF method to generalize to other cohorts.

To measure this aspect, we evaluated the number of disconnected nodes/metagenes in the results of the various MFs (Supp Text 2). As shown in Figure 3, sICA, with 65, 224 and 36 disconnected metagenes in CRC, BRCA and OVCA, respectively, outperforms other approaches (see Figures 3A, S4A, S5A). For example, NMF and PCA had respectively 129 and 173 disconnected nodes in CRC. Finally, cophNMF obtained 12% of disconnected nodes against the 6.7% of sICA (see Table S2). As expected, RGCCA-based RBH graphs has less disconnected components than any other MF method independently applied to each dataset (Figure S6A).

### 3.2 Wide across-datasets reproducibility

To evaluate the reproducibility of the metagenes output of the different MFs we computed the number of links in their RBH graphs (Supp Text 2). For example, working with 14 CRC datasets, in an optimal scenario a metagene should find 13 RBHs corresponding to the metagenes that reflect the same biological factor in the remaining 13 datasets. In reality, this is not always the case given that a biological factor can be underrepresented in some datasets due to the choice of the samples or to their number. However, higher is the number of RBHs lower is the deviation of the performances of a MF approach from the optimal scenario. As shown in Figures 3B, S4B and S5B sICA, with 2900 RBHs in CRC 1605 in BRCA and 390 in OVCA, strikingly outperforms alternative approaches. In CRC, for example, sICA identified approximately 1000 RBHs more than the other MFs, including also cophNMF (see Table S2). At the same time, RGCCA-based RBH graph for CRC was characterized by 3730 RBH links (Figure S6B). Interestingly, sICA, without forcing the correlation between the components of different datasets, provides only 830 RBHs less (corresponding to 22% less) than RGCCA.

**Fig3. Comparison of MFs in CRC.** Different measures are here plotted for the comparison of the various MFs: Stabilized ICA (red), ICA (dark green) ICA' (green), NMF (blue) and PCA (violet).

### 3.3 Tightness of the community structure in the RBH graph

Concerning the topological structure of the RBH graph, the best MF algorithm should derive a cluster-graph like graph, i.e. a disjoint union of tight communities. Indeed as discussed above an optimal MF algorithm should find a component for each relevant biological factor underlying the transcriptome. Working with various transcriptomic datasets obtained from the same disease (for example CRC), those components associated to the same biological factor should cluster together forming a tight community. The final structure of the optimal RBH graph should be thus composed of various tight communities sparsely connected one to each other.

In order to verify how the RBH graphs resulting from the different MF approaches are close to this optimal topology, we considered four well-established measures (Supp Text 2): (i) clustering coefficient; (ii) modularity; (iii) number of communities and (iv) average size of the communities. The first two are standard measures in network theory for evaluating how evident is the presence of communities in a graph (Fortunato, 2010). The average size and the number of the communities are instead used to evaluate how consistently each MF algorithm merges components obtained from different datasets. From the results reported in Figures 3C-F, S4C-F and S5C-F the superior performances of sICA with respect to alternative approaches can be clearly appreciated. Especially the clustering coefficient and modularity are strikingly higher in sICA in respect to its alternatives. Of note, concerning the number of communities, in CRC NMF performs as sICA and, in OVCA, PCA outperforms sICA. However while PCA detects more communities than sICA in OVCA, these are smaller and in two cases they merge

metagenes coming from the same dataset. As shown in Table S2, also concerning the topology of the RBH graph, the performances of NMF do not improve if considering cophNMF. RGCCA-based RBH graph for CRC was characterized by tighter communities as expected (Figure S6C-F).

### 3.4 Biological content and specificity of the components

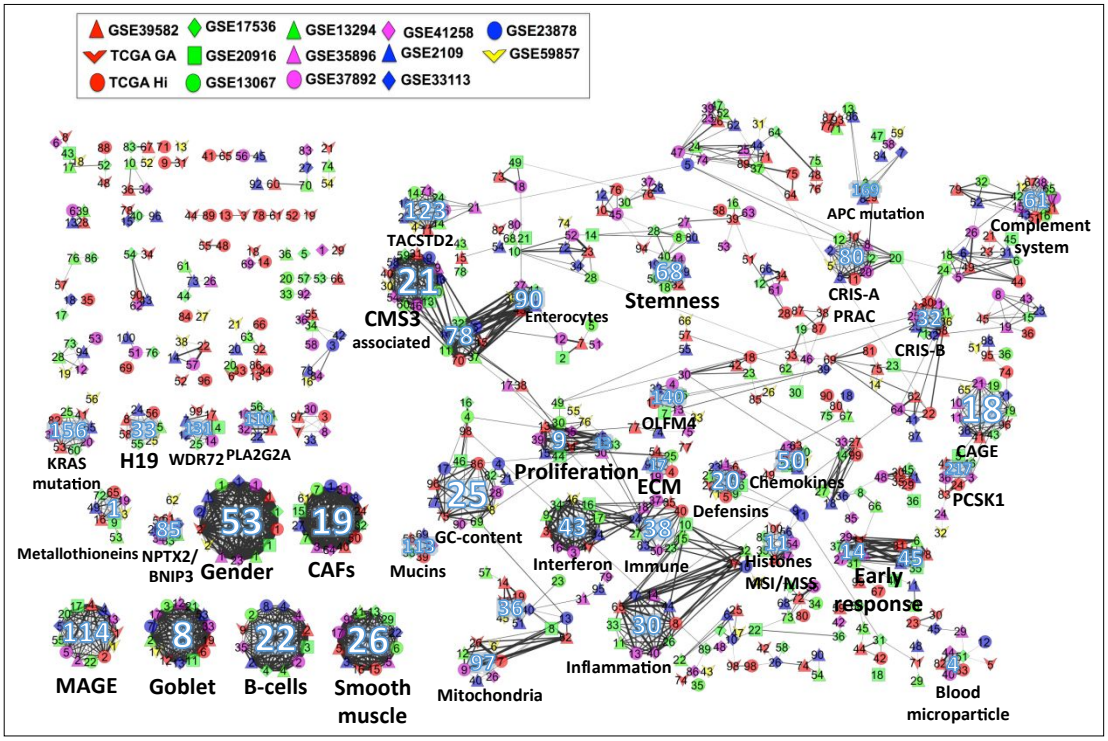
Finally, we checked if the communities identified in the RBH graph were effectively associated to specific biological factors. In particular, we tested the ability of the communities of the different MFs in predicting three biological factors that are expected to influence cancer transcriptomic profiles: patient gender, proliferation status of a tumor and the level of stromal infiltration. For this test we performed a regression analysis of the metasamples obtained from the different MFs.

The gender annotation is composed of discrete values M/F obtained from the available clinical annotations: in this case, we thus performed a logistic regression. Proliferation was evaluated averaging the expression of the genes belonging to a well-known proliferation signature (Giotti *et al.*, 2017) and it is thus a vector of continuous weights. Finally, stromal infiltration was estimated using the average expression of the genes belonging to the stromal signature of ESTIMATE tool (Yoshihara *et al.*, 2013).

The results of this first test are summarized in Figures 3G-I, S4G-H and S5G-H. We focused on the community that predicted the best the specified biological signal. The community was selected as the one with the highest percentage P of metasamples whose regression on the biological signal was significant. We used three parameters commonly used to evaluate the quality of a linear regression: $R^2$, Bayesian information criterion (BIC) and Akaike's information criterion (AIC). We finally define a score to combine them in a single value as (P*R2) / (BIC*AIC). The higher this score the stronger is the association between the community and the biological factor. Indeed a good regression would correspond to $R^2$ value near to 1 and low BIC and AIC values. Such scores are reported in Figures 3G-I, S4G-H and S5G-H. The specific values obtained by the single scores are reported in Table S3. As shown in Figures 3G-I, S4G-H and S5G-H, sICA better approximates all three tested biological factors. In particular, NMF does not identify any component that can significantly predict the gender signal. We then investigated the specificity of such predictions, meaning the ability of the MF approach to define a clear one-to-one association between a biological signal and a component. To test for the specificity of the different MFs we focused on the components obtained on the GSE39582 dataset (see Table S1) and considered the $R^2$ obtained in the previously computed regressions by all the 100 components. As shown in Figure S7, sICA resulted to be far more specific than the alternative MFs. In particular for all the three biological factors (gender, proliferation and stromal infiltration) sICA found only one component strongly associated to them. On the opposite, NMF and ICA' identified multiple components with similar regression performances. Finally PCA resulted to be specific in stromal infiltration and proliferation prediction. However, PC1 was the component predicting simultaneously both signals, confirming the already observed limitation of PCA of conflating multiple biological processes into a single component.

### 3.5 Impact of the technical platform on the MFs

We used OVCA as a case study to evaluate the impact of the profiling platform on the results of the various MF algorithms. Indeed having four OVCA datasets composed of the same samples we are sure that no biological variability is present across them. In the optimal scenario, all the metagenes of an MF algorithm should find a RBH with a metagene of the other three datasets. At step 2 of our framework applied to OVCA we checked the number of RBH links of the different MFs together with their average absolute correlation. sICA resulted to perform better than alternative approaches also in this case, with 390 links and average correlation of 0.396 (see Table S4 and Figure S5B). Finally, we evaluated if a specific agreement could be identified between profiling platforms (see Figure S8 and Supp Text 2). The correlations among the obtained across different platforms are highly variable, depending on the MF method employed. Agilent seems to show the lower correlation with



Affymetrix microarray and RNAseq platforms. From such analysis, together with the results of BRCA and CRC, we can conclude that RNAseq and microarray platforms give similar results in terms of extracted components.

**Fig.4. RBH graph of sICA built in CRC with the main biological annotations.** The node colors indicate the dataset from which the components have been computed. The edge thickness indicates the magnitude of the correlation. Communities with more than six elements are marked with an integer number. For details on the community annotations see Table S5.

### 3.6 sICA identifies biological insights on CRC consistent with previous knowledge

In the previous sections we showed that sICA has more reproducible results than alternative approaches according to multiple measures of practical interest for high-throughput data analysis. We now concentrate more deeply on the biological insights that can be derived from the RBH graph of this MF algorithm in CRC. To this aim we added to the analysis other four datasets: single-cell RNAseq from normal and tumoral CRC tissue (Li *et al.*, 2017), Patient-derived Xenograft (PDX) CRC Models and liver metastasis (LM) (Isella *et al.*, 2017). Combining sICA components from scRNASeq data together with those obtained in bulk RNA-seq transcriptomes through the RBH network allows better characterization of cell-type specific signals in bulk transcriptomes while PDX and liver metastasis data help to better discriminate tumor cell-specific signals from microenvironment signals. Given the different nature of such data in respect to the previous 14 we only employed them for the biological characterization and not in the assessment of MF algorithm performances. We then biologically annotated the communities of the RBH graph by using consensus metagenes and metasamples according to the procedure described in the Supp Text 5. The consensus metagenes obtained for the communities of sICA are reported in Table S5 and represent a useful resource for further analyses. Figure 4 reports the RBH graph of Stabilized ICA and the main biological informations extracted from it. Four main categories of biological factors can be distinguished in the graph: factors intrinsic to the tumor, microenvironment signals, technical signals, effects of small groups of genes and unknown factors. Concerning the tumor-specific factors, some communities were found to be associated to core tumoral functions, such as proliferation, inflammation, stemness, interferon response and mitochondria. Other tumor-specific communities resulted instead to be associated to CRC-specific tumoral signals, such as MSI/MSS (microsatellite instability/microsatellite stable), goblet cells (a differentiated cell of the colon) and KRAS mutation. Finally, one community was found to be related to chromatin silencing and histones. The stromal communities instead include microenvironment signals, such as cancer-associated fibroblasts (CAFs), smooth muscle, immune, complement system and B-cells. Of particular interest is the identification of the communities related to B-cells and CAFs whose association to these cell types was evident not only using MSigDB signatures, but also from single-cell data (see Supplemtnary text 4 and Figure S9). The technical factors included instead GC-content and gender. Finally, 10 communities have been found to be associated with small groups of genes. In this last case, the consensus metagenes associated to these communities contained few genes having a much higher weight than the others.

Concerning the association with the predefined CRC Consensus Molecular Subtypes (CMS) we could clearly match CMS1 with our immune component, concordantly to what previously observed. Communities associated to CMS3 and CMS4 were also identified. Of note, the CMS4 subtype resulted from our analysis to be associated to both smooth muscles and CAFs. A strong CAFs infiltration had been already observed in this CRC subtype (Isella *et al.*, 2015; Guinney *et al.*, 2015).

## 4 Discussion

In this manuscript we compared the three most commonly used matrix factorization methods for their ability to detect reproducible and biologically interpretable signals in independent transcriptomic datasets of the same cancer type (CRC, BRCA, OVCA). For one of the methods, Independent Component Analysis, we also compared three protocols of its application to transcriptomic data, named ICA, ICA' and sICA. We designed a framework based on the concept of Reciprocal Best Hit (RBH), for assessing the reproducibility of any MF method. From our study we can conclude that minimizing mutual information between metagenes (ICA, sICA) rather than metasamples (ICA') results in better metagene reproducibility and interpretability. Moreover, using multiple runs of ICA for stabilisation and prioritizing stable components (as done by sICA) significantly improves reproducibility. By contrast, PCA components appear to systematically mix multiple sources of transcriptome variability, reducing interpretability. Also, the higher-order PCA components are regularly not reproducible which is partly expected given rotational invariance of the linear subspaces spanned by the principal components (Ochs and Fertig, 2012). From previous studies it is known that NMF shows a good performance in the analysis of mutation data (Alexandrov *et al.*, 2013) and cancer subtyping (Isella *et al.*, 2017). However, the NMF components are less frequently selectively associated with biological factors compared to ICA. Moreover, to the best of our knowledge, we lack validated tools for stabilizing NMF components, similarly to sICA, in transcriptomic data analysis.

We demonstrated that the meta-analysis of the results of sICA, based on constructing the RBH graph, provides a biologically rich image of the signals shaping tumoral transcriptomes and their interconnection. Tight communities, existing in the RBH graph, whose meaning can be compared to the Clusters of Orthologous Genes (COGs) in evolutionary bioinformatics, can be matched to previously known and/or expected highly reproducible biological signals (such as proliferation and immune infiltration) but also highlights novel biological mechanisms which require further investigation and interpretation.

The metagenes obtained through application of MF methods can be compared to other methods, sharing similar spirit. In particular, attractor metagenes were suggested in order to serve as surrogates of cancer phenotypes (Cheng *et al.*, 2013). Attractor metagenes were used as variables in the DREAM Challenge winning approach for predicting breast cancer clinical outcome (Margolin *et al.*, 2013). We find ICA-based framework for identifying metagenes more computationally elegant and potentially producing less poorly generalizable signatures; however, further study is required to compare the results of both approaches and their computational performances. INSPIRE method uses the latent variable approach to infer modules of co-expressed genes and the dependencies among the modules from multiple expression datasets that may contain different sets of genes (Celik *et al.*, 2016). Therefore, INSPIRE shares general objectives of MF-based meta-analysis but significantly differs in terms of methodology. For example, INSPIRE is based on the assumption of Gaussianity in the data distributions and uses disjoint module definitions rather than metagenes, where each gene can contribute to several biological functions.

Lastly, here we compared MF methods in application to cancer transcriptomic datasets. However, the suggested approach can be easily extrapolated to other data types (methylomic, proteomic) or other fields of research collecting massive transcriptomic datasets (such as drug screenings).


## Acknowledgements

We thank Urszula Czerwinska, Askhat Molkenov and Alexander Gorban for their advices and help. We thank Justin Guinney and the whole CRCSC for supporting the idea and providing the normalized datasets.

## Funding

This work has been partially supported by the "Pan-cancer deconvolution of omics data using Independent Component Analysis" project (AP05135430) from the Ministry of Education and Science of the Republic of Kazakhstan, ITMO Cancer within the framework of the Plan Cancer2014–2019 and convention Biologie des Systèmes N°BIO2015–01 (M5 and MOSAIC projects), ERACoSysMed ERA-NET programme (COLOSYS project), Ministry of education and science of Russia (project No. 14.Y26.31.0022), the European Union's Horizon 2020 programme (grant No 826121, iPC project).

*Conflict of Interest:* none declared.


*Conflict of Interest:* none declared.


## References


Alexandrov,L.B. *et al.* (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.

Alter,O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 10101–10106.

Argelaguet,R. *et al.* (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.

Avila Cobos,F. *et al.* (2018) Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinforma. Oxf. Engl.*, **34**, 1969–1979.

Bell,D. *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Biton,A. *et al.* (2014) Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.*, **9**, 1235–1245.

Bork,P. *et al.* (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.

Brunet,J.-P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 4164–4169.

Celik,S. *et al.* (2016) Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer. *Genome Med.*, **8**, 66.

Cheng,W.-Y. *et al.* (2013) Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput. Biol.*, **9**, e1002920.

Devarajan,K. (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.*, **4**, e1000029.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Fortunato,S. (2010) Community detection in graphs. *Phys. Rep.*, **486**, 75–174.

Giotti,B. *et al.* (2017) Meta-analysis reveals conserved cell cycle transcriptional network across multiple human cell types. *BMC Genomics*, **18**.

Guinney,J. *et al.* (2015) The consensus molecular subtypes of colorectal cancer. *Nat. Med.*, **21**, 1350–1356.

Himberg,J. and Hyvarinen,A. Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. 259–268.

Hyvarinen,A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, **10**, 626–634.

Isella,C. *et al.* (2017) Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat. Commun.*, **8**, 15107.

Isella,C. *et al.* (2015) Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.*, **47**, 312–319.

Kairov,U. *et al.* (2017) Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics*, **18**.

Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, **13**, 1706–1718.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Li,H. *et al.* (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**, 708–718.

Ma,S. and Dai,Y. (2011) Principal component analysis based methods in bioinformatics studies. *Brief. Bioinform.*, **12**, 714–722.

Margolin,A.A. *et al.* (2013) Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.*, **5**, 181re1.

Molinaro,A.M. *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinforma. Oxf. Engl.*, **21**, 3301–3307.

Ochs,M.F. *et al.* (1999) A New Method for Spectral Decomposition Using a Bilinear Bayesian Approach. *J. Magn. Reson.*, **137**, 161–176.

Ochs,M.F. and Fertig,E.J. (2012) Matrix factorization for transcriptional regulatory network inference. IEEE, pp. 387–396.

Parker,J.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, **27**, 1160–1167.

Stein-O'Brien,G.L. *et al.* (2017) Enter the matrix: Interpreting unsupervised feature learning with matrix decomposition to discover hidden knowledge in high-throughput omics data.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tenenhaus,M. *et al.* (2017) Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. *Psychometrika*.

Yoshihara,K. *et al.* (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**.