UNIVERSITÉ PARIS DESCARTES

ED 474 Frontières du vivant

Institut Curie, PSL Research University, Mines Paris Tech, Inserm U900
26, rue d'Ulm
75005 Paris, France

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

Par Urszula Czerwińska

Thesis Advisory Committee Report 2018

Thèse dirigée par Andrei Zinoviev et Vassili Soumelis

Paris, le 7 février 2018

Devant un jury composé de :

Andrei Zinoviev	directeur de thèse - PSL
Vassili Soumelis	directeur de thèse - PSL
Denis Thieffry	advisor - ENS
Frank Pagès	advisor - Université Paris Descartes

Title: Computational deconvolution of cell and environment specific signals and their interactions from complex mixtures in biological samples

Abstract: In many fields of science (biology, technology, sociology) observations on a studied system represent complex mixtures of signals of various origin. Tumors are engulfed in a complex microenvironment (TME) that critically impacts progression and response to therapy. It includes tumor cells, fibroblasts, and a diversity of immune cells. Most studies have focused on individual cell types in model tumor systems, and/or on individual molecules mediating a crosstalk between two cells. Unraveling the complexity, organization, and mutual interactions of TME cellular components represents a major challenge. Methods for deconvolution of complex mixtures of signals have been developed in signal processing field. It is known that under some assumptions, it is possible to separate complex signal mixtures, using classical and advanced methods of source separation and dimension reduction. Our recent large-scale analysis of more than 6500 tumor transcriptomes, applying classical blind source separation methods showed that we can reliably separate signals coming from tumor microenvironment from the tumor-specific signals and various technical artifacts. However, the precise composition of the immune-related signals in a tumor sample remains to be deciphered.

In this project, we develop and apply the advanced methodology of signal deconvolution to decipher sources of signals shaping transcriptomes of tumor samples, with a particular focus on immune-related signals. So far, we managed to deconvolute successfully immune-related signal into groups related to immune cell-types in six breast cancer datasets. However, the precise composition of the immune-related signals and their interactions in a tumor sample remains to be deciphered and our method needs to be calibrated.

We are going to release our processing pipeline in a form of an R package. This will allow the scientific community profit from our analytical pipeline and easily reproduce our results.

In the case of success of this project, the results will be helpful in the detemining diagnosis and treatment of cancer, especially for immunotherapies.

Keywords: tumor microenvironment, cancer systems biology, transcriptome data analysis, single cell data analysis, bioinformatics, heterogeneity, blind deconvolution, unsupervised learning, cancer, immunology

Dédicace

And now, let's repeat the Non-Conformist Oath!

I promise to be different!
I promise to be unique!

I promise not to repeat things other people say!

— Steve Martin, *A Wild and Crazy Guy (1978)*

Avertissement

Note to TAC committee

This is a draft of PhD thesis realized for a purpose of a Thesis Advisory Committee meeting of 3rd year. Please, forgive possible incoherence in the form and blanks that you will find in this report. The shape of this work will probably change many times before reach its final form. Don't mind the citation and references errors that will be fixed at the very end.

Enjoy the reading!

Contents

List of Tables

List of Figures

Chapter 1

Immuno-biology of cancer

This chapter will introduce a basic topic of cancer and participation of stroma in cancer development, progression and response to treatment. It will also describe most important types of data used to study cancer microenvironment.

1.1 Cancer seen as complex environment

For a long time studying tumor was focused on tumor cells, their reprogramming, mutations. Cancer was seen as disease of uncontrolled cells. Recent discoveries moved research focus from tumor cells to tumor cells in their context: tumor microenvironment. We will describe here what is the composition and role of the TME in tumor progression, diagnosis and response to treatment.

1.1.1 Our understanding of cancer over time

cancer is a disease touching blah blah many ppl over the word. it has been known that blah blah and then types

CANCER STATISTICS

Tumor was seen for decades as a disease with mutations as a main source of heterogeneity.

1.1.2 Tumor micro environment: fiend or foe?

1.1.2.1 What is Tumor Microenvironment (TME)

Tumor Microenvironment is a complex tissue that surrounds tumor cells. It is composed of blood and lymphatics vessels, epithelial cells, mesenchymal stem cells, fibroblast, adipocytes and a wide variety of immune cells. Their proportion and specific roles vary significantly with tumor type and stage. Communication between the environmental cells and the tumor is critical for tumor development and its impact on patient's response to treatment.

1.1.2.2 TME as tumor allay

In 1863 Rudolf Virchow observed a link between chronic inflammation and tumorigenesis. Accoriding to Virchov theory genetic damage would be the "match that lights the fire" of cancer, and the inflammation or cytokines produced by immune cells should be the "fuel that feeds the flames". (**?** and cancer: back to Virchow?). Therefore lymphocyte infiltration was confirmed by subsequent studies as a hallmark o cancer. The question one may ask is why our immune system does not defend the organism from tumor cells as it does in a range of bacterial and viral infections? It is mainly because of the ability of tumor cells to inhibit immune response through activation of negative regulatory pathways (so called immune checkpoints).

Many examples can be cited on how TME facilitates tumor development. For instance, in the early stages of tumorigenesis some macrophage phenotypes support tumor growth. Also, it was shown that myeloid-derived suppressor cells (MDSCs) have an ability to suppress immune defence i.e. immunosurveillance by dendritic cells (DCs), T cell activation and macrophage polarisation and they promote tumor vascularisation as well. (@ Talmadge, J.E. & Gabrilovich, D.I. History of myeloid-derived suppressor cells. Nat. Rev. Cancer 13, 739–752 (2013).44. @ Gabrilovich, D.I., Ostrand-Rosenberg, S. & Bronte, V. Coordinated regulation of myeloid cells by tumours. Nat. Rev. Immunol. 12, 253–268 (2012). Tregs and myeloid-derived suppressor cells can negatively impact natural immune defence and by these means allow growth and invasion of tumor cells. (**?** of the tumor immune microenvironment for staging and therapeutics Janis M Taube1,2,3, Jérôme Galon). Another cell type, a part of ECM, fibroblast, or more precisely Cancer Associated Fibroblasts (CAFs) have proven pro-tumor functions in breast cancer where they enhance metastasis. (@ Dumont, N. et al. Breast fibroblasts modulate early dissemination, tumorigenesis, and metastasis through alteration of extracellular matrix characteristics. Neoplasia 15, 249–262 (2013).)

In addition, it is worth mentioning the role of ECM as an integral part of TME and its

impact on tumorigenesis and metastatis. It is usually anit-tumor in early stages and pro-tumor at the metastatic stages. The blood and lymphatic vessels maintain tumor growth providing necessary nutritive compound to malignant cells.

HALLMARKS?

Fig 2 The micronvironment supports metastatic dissemination and colonistaion at secondary sites ? (**?** regulation of tumor progression and metastasis)

1.1.2.3 Two-faced nature of immune cells

Recent studies unveil ambivalent nature of immune cells of TME. While some as cytotoxic T cells, B cells and macrophages can manage to eliminate tumor cells. Treg cells role is to regulate expansion and activation of T and B cells. Depending on cancer type, they can be either pro- or anit- tumor. For example as it has been shown for Tregs, they can be also associated with improved survival (i.e. in colorectal cancer(@. Frey, D.M. et al. High frequency of tumor-infiltrating FOXP3+ regulatory T cells predicts improved survival in mismatch repair-proficient colorectal cancer patients. Int. J. Cancer 126, 2635–2643 (2010).). For innate immunity, there are widely accepted M1 (ani-tumor) and M2 (pro-tumor) extreme macrophages phenotypes in TME. (**?**, B.Z. & Pollard, J.W. Macrophage diversity enhances tumor progression and metastasis. Cell 141, 39–51 (2010). Most of the statements seem to be context dependent and not valid universally across all cancer types. We already mentioned Macrophages phenotypic plasticity as well as different behaviour of EMC depending on tumor stage.

From more general point of view, it has been observed that immunodeficiency can correlate with high cancer incidence. Results of analysis based on observations of 25,914 female immunosuppressed organ transplant recipients, the tumor incidence was higher than predicted for multiple cancers. However, the number of breast cancer cases decreased which can be really disturbing if we need to decide on the role of immune defence in tumor progression (**?** T., Tsai, S.C., Grayson, H., Henderson, R. & Opelz, G. Incidence of de-novo breast cancer in women chronically immunosuppressed after organ transplantation. Lancet 34, 796–798 (1995).) This trend was confirmed through a study on individuals with AIDS and other studies. This indicates that immune microenvironment can be cancer stimulating or inhibiting depending on the type of cancer.

- review hallmarks of cancer immune

1.1.3 Cancer immune phenotypes

There can be distinguished cancer phenotypes depending on immune infiltration how they are measured, defined, indexes, types of cancer, impact

> In further support of a role for memory T cells in antitumour responses, tumour-infiltrating lymphocytes that express CD4 or CD8 extracted from experimental tumour models typically have the features of memory T cells and can possess an activated or exhausted phenotype, expressing markers such as PD-1, T-cell immunoglobulin and mucin-domain containing protein 3 (TIM-3) and lymphocyte activation gene 3 (LAG-3). (**?** CANCER CIRCLE)
>
> Anticancer immunity in humans can be segregated into three main phenotypes: the immune-desert phenotype (brown), the immune–excluded phenotype (blue) and the inflamed phenotype (red). (**?** CANCER CIRCLE Fig 3)
>
> Inflamed versus non-inflamed tumours
>
> What is the basis for the three immune profiles observed in tumours? To a first approximation, differences between the profiles can be ascribed to whether tumours harbour an inflammatory microen-vironment, which can reflect variations in a number of cellular and other factors (Fig. 4). The degree of inflammation can be gauged by the cellular content of the tumour — for example, the presence of immune cells, either in the parenchyma or at the invasive margin of the tumour78,79. Inflamed tumours also contain proinflammatory cytokines that should provide a more favourable environment for T-cell activa-tion and expansion, including type I and type II IFNs, IL-12, IL-23, IL-1□, tumour-necrosis factor (TNF)-□ and IL-2. However, it is unclear whether the presence of these cytokines is the cause or consequence of the cellular influx. The production of tropic chemokines by lympho-cytes and myeloid cells is therefore likely to be an important feature of inflamed tumours.Non-inflamed tumours generally express cytokines that are associated with immune suppression or tolerance. They can also contain cell types associated with immune suppression or tissue homeostasis. As well as regulatory T cells, these cells include the lesser characterized populations of myeloid-derived suppressor cells (for example, immature granulocytes) and tumour-associated macrophages, which are unacti-vated and often called M2 macrophages. However, regulatory T cells are not associated uniquely with non-inflamed tumours as they typically accompany effector T cells into inflammatory sites and are important for maintaining immune homeostasis, even in the presence of an active antitumour immune response

immunoscore

immunophenoscore

ML based scoring scheme. Random forest

link to *precision medecine*

> Together, these observations support the notion that gene expression-based correlates of immune involve- ment could hold valuable clinical utility for a number of prognostic and therapy-predictive applications. However, to date, mRNA-based diagnostics that quantify immune involvement in tumors do not exist. Multi-gene diagnos- tics that simultaneously measure mRNA transcripts of multiple genes represent a class of In Vitro Diagnostic Multivariate Index Assay (IVDMIA) that has in recent years gained wide clinical acceptance for the diagnosis and stratification of patients into risk groups to guide therapeutic decisions [80, 81]

> These observations raise the question of the underlying molecular mechanisms that explain the differences in immunogenicity of the tumors. The question can be reduced to the notion of sources of immunogenic differences, which can be divided into two categories: tumor-intrinsic factors and tumor-extrinsic factors. Tumor-intrinsic factors include the mutational load, the neoantigen load, the neoantigen frequency, the expression of immunoinhibitors and immunostimulators (e.g., PD-L1), and HLA class I molecule alterations. Tumor-extrinsic factors include chemokines that regulate T cell trafficking, infiltration of effector TILs and immunosuppressive TILs, and soluble immunomodulatory factors (cytokines) (**?** et al., 2006Immune resistance orchestrated by the tumor microenvironment.)

> For each of the studied cancers, the analysis revealed only immune-related factors, which we classified into four categories: (1) infiltration of activated CD8+/CD4+ T cells and Tem CD8+/CD4+ cells; (2) infiltration of immunosuppressive cells (Tregs and MDSCs); (3) expression of MHC class I, class II, and non-classical molecules; and (4) expression of certain co-inhibitory and co-stimulatory molecules (Figure 5A).To visualize the information, we constructed an immunophenogram that includes these four categories (Figure 5B). We then calculated an aggregated score, immunophenoscore, based on the expression of the representative genes or gene sets comprising four categories: MHC molecules, immunomodulators, effector cells (activated CD8+ T cells and CD4+ T cells, Tem CD8+ and Tem CD4+ cells), and suppressor cells (Tregs and MDSCs) Multivariate analysis showed that the immunophenoscore was associated with survival in 12 solid cancers, of which 4 were significant: KIRC, SKCM, breast cancer (BRCA), and bladder cancer (BLCA) (Figure 5C).

> The immunophenoscore we developed was derived in an unbiased manner using the TCGA data and machine learning, but it reflects current understanding of the categories of genes that determine immunogenicity of the tumors: effector cells, immunosuppressive cells, MHC molecules, and immunomodulators. The immunophenoscore is similar to the conceptual immunogram that was recently proposed to represent the status of the immune system (Blank et al., 2016). Another advantage of the immunophenoscore is that it represents a standardized value because Z scores are used, and is therefore more robust compared with the use of expression values. However, because presently only limited data are available, additional studies are required to validate the immunophenoscore. Notably, the method can be further improved by optimizing the immunophenoscore for specific cancers. Finally, for routine applications, other techniques for gene expression profiling like microarrays and qPCR can be used instead of RNA-sequencing.

1.1.4 Immune signatures

definition of signature: marker genes, list of genes, weighted list we can talk about the general immune signature of signature of immune infiltration and stroma or immune signature of a specific cell type of functional subpopulation purpose of signatures

availability of immune signatures

the problem of not consistency of immune signatures origin of signatures

“the gene expression profiles of tumour-associated immune cells differ considerably from those of blood derived immune cells”(**?** et al. Estimation of immune cell content using single cell data)

1.2 Immunotherapies

This section outlines progress in cancer therapies with a focus on immune therapies. It will link the ongoing research on TME with therapeutical potential.

1.2.1 Cancer therapies

1.2.2 Recent progress in immuno-therapies

The immunotherapies, in contrast with other types of cancers therapies discussed in the previous chapter, aim to trigger or restart the immune system to defend the organism and attack the malignant cells. All this, however without provoking persisting inflammation state (**?**, J., (2013). Changes in the local tumor microenvironment in recurrent cancers may explain the failure of vaccines after surgery. Proc. Natl. Acad. Sci. USA 110, E415–E424.)

The idea of stimulating immune system to fight malignant cell was not born recently. Since a long time a possibility of development of an anti-cancer vaccine has been investigated. Unfortunately, this idea faced two important limitations 1) lack of knowledge of antigens that should be used in vaccine to successfully stimulate cytotoxic T cells 2) the ability of cancer to block the immune response also called *immunostat*. Despite those impediments works on anti-tumor vaccines do not cede. (**?**, K., and Banchereau, J. (2013). Dendritic-cell-based therapeutic can- cer vaccines. Immunity 39, this issue, 38–48.)

Another idea involving using immune system as a weapon to fight cancer, would be the use of genetically modified patient's T-cells, carrying *CARs* (chimeric antigen receptors) (**?** Driving CAR T-cells forward nat rev 2016). After a long period of small unsuccessful trials, recently in 2017, two CAR T-cell therapies were accepted, one to "treat adults with certain type of large B-cell lymphoma" (of Health and Services (2017b)), other to treat "children with acute lymphoblastic leukemia (ALL)" (of Health and Services (2017a)) , which are, at the same time, the first two gene therapies accepted by FDA.

However, the two most promising immuno-related strategies are based on blocking so called immune check point inhibitors: cytotoxic T-lymphocyte protein 4 (CTLA4) and programmed cell death protein 1 (PD-1). The anti-CLTA4 antibodies blocks repressive action of CLTA4 on T-cells and they become therefore activated. It was shown efficient in melanoma patients and accepted by FDA in 2015 as adjuvant therapy for stage III metastatic melanoma patients (of Health and Services (2015)). PD-1 is a cell surface receptor of T cells, that binds to PD-L1/PD-L2. After binding, an immunosuppressive pathway is activated and T cells activity is dampened. An action of an anti-PD-L1 antibody is to prevent this immune exhaustion.(**?** CANCER CIRCLE). A stepping stone for anti-PD-L1 therapies was approval of Tecentriq (atezolizumab) for Bladder cancer (of Health and Services (2016a)) and anit-PD1 Keytruda (pembrolizumab) initially accepted for NSCLC and further extended to head and neck cancer, Hodgkin's lymphoma, gastric cancer and microsatellite instability-high cancer (of Health and Services (2016b)). Since other anti-PD-L1 or anti-PD1 antibodies were accepted or entered advanced stages of clinical trials (**?**, 2015,PD-1 blockers). A short history of immunotherapy FDA-accepted treatments can

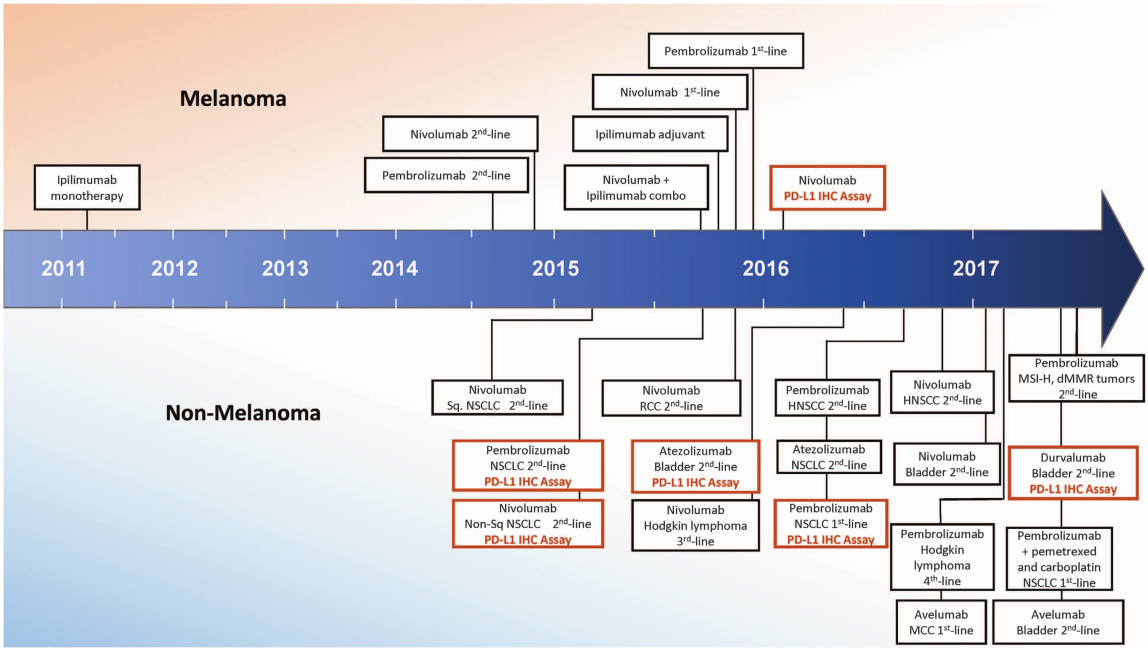


Figure 1.1: This timeline describes short history of FDA approval of checkpoint blocking immunotherapies up to 2017. Reprinted by permission from Springer Nature, (**?** of the tumor immune microenvironment for staging and therapeutics Janis M Taube1,2,3, Jérôme Galon), © 2017 Macmillan Publishers Limited, part of Springer Nature. All Rights Reserved.

be found in Fig. 1.1

The main drawback of immunotherapies is a heterogeneity of response rate, which can vary i.e. from 10–40% in case of PD-L1blocking(@ Zou, W., Wolchok, J. D. & Chen, L. PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: mechanisms, response biomarkers, and combinations Sci. Transl. Med. 8, 328rv4 (2016).) , suggesting that some patient can have more chances than others to respond to an immune therapy. So far, it has been shown that anti PD-L1 therapies works more effectively in T cell infiltrated tumors with exclusion of Tregs because of lack of difference in expression of FOXP3 in responding and non-responding group of patients. (**?** role PD_L1_. Also some light has been shade by Rizvi et al (@. Rizvi, N. A.et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348, 124–128 (2015). who connected mutational rate of cancer cells to the chances of response to an immunotherapy.

Despite those fundings, the precise qualifications of patients that should be sensitive to an immunotherapy are not defined (**?** et al., 2016,Resistance mechanisms to immune-checkpoint blockade in cancer: tumor-intrinsic and -extrinsic factors.). As most patients do not answer to immunotherapies, it stimulates researches to look for better biomark-

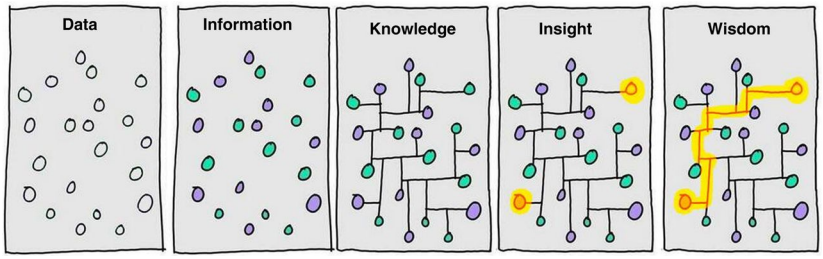


Figure 1.2: **From *Data* to *Wisdom***. Illustration of different steps that it takes to go from *Data* to generating *Wisdom*. It highlights that generating data is not equal to understanding it and additional efforts are needed to generate value. Image authored by Clifford Stoll and Gary Schubert published by Portland Press Limited on behalf of the Biochemical Society and the Royal Society of Biology and distributed under the Creative Commons Attribution License 4.0 (CC-BY) in (**?** DATAOMICShttp://www.emergtoplifesci.org/content/1/3/245.article-info).

ers and patient stratifications, and pharmaceutical industries to discover new immune checkpoints based therapies.

1.2.3 Potential of development of new immunotherapies

?

1.3 Quantifying immune infiltration (data)

Nowadays, more and more biological data is produced. However, this proliferation of accessible resources is not proportional to generated insights and wisdom. In this thesis, we wok mostly generate *Knowledge* and *Insights* and we hope to generate some *Wisdom* (Fig. 1.2). However, in this part, we will introduce the foundation of our analysis: different data types that will be further discussed in chapters that follow.

We will introduce most relevant data types that are used to study immune infiltration of tumors.

1.3.1 Facs

Flow cytometry : Laser-based technology that allows for simultaneous quantification of the abundance of up to 17 cell surface proteins using fluores-

cently labelled antibodies. (**?** RNA sequencing to explore immune cell heterogeneity) cost : 0.05$/ cell

Mass cytometry(commercial name CyTOF). Mass spectrometry technique used as an alternative to flow cytometry that allows for the quantification of cellular protein levels by using isotopes that overcome problems associated with the spectral overlap of fluorophores. 40 prot per cell (**?** RNA sequencing to explore immune cell heterogeneity) 35$/cell

1.3.2 staining (histopathology, immunoscore!!! , multiplex immunofluorescence)

The standardized Immunoscore was based on the quantification (cells/mm2) of two lymphocyte populations (CD3 and CD8) within the central region and the invasive margin of colorectal carcinoma tumors and provides a scoring system ranging from Immunoscore 0 (I0) to Immunoscore 4 (I4) (Figure 4).41 (**?** of the tumor immune microenvironment for staging and therapeutics Janis M Taube1,2,3, Jérôme Galon)

The immune cell content of a tumour sample can also be determined by using more established multiplexed methods like immunohistochemistry (IHC) or immunofluorescence (IF)20 or newer methods like imaging mass cytometry using FFPE tissue samples21. The advantage of these techniques is that a larger number of cells can be analysed and that these techniques also provide information about the spatial distribution of the different cell types. However, these methods are limited to the number of proteins that can be analysed simultaneously currently (ranging from ~10 to 100), advantage of the deconvolution approach is that it is unbiased (i.e., hypothetical response markers do not need to be pre-specified). It allows one to link both the cellular characteristics and the cellular content with treatment response. We anticipate that this approach will aid in the discovery of novel predictive response biomarkers for both conventional and immune-directed therapy by taking cellular composition into account. (**?** et al. Estimation of immune cell content using single cell data).

1.3.3 omics

Some kind of sequencing explanation needed for non-biologists

1.3.3.1 transcriptome

Transcriptomics is the large-scale study of RNA mole- cules by use of high-throughput techniques. It examines the abundance and makeup of a cell's transcriptome1,2. In contrast to DNA, which is largely identical across all cells of an organism, the actively transcribed RNA is highly dynamic, reflecting the diversity of cell types, cellular states and regulatory mechanisms. Because a transcrip- tome profile can be regarded as a signature or snapshot of the underlying cell state, the experimental profiling of samples and specimens can provide insights into their unique biology.

Transcriptome profiling can detect changes in gene activity and regulation by capturing quantitative expres- sion patterns and has the capacity to describe the under- lying phenotypes in great detail. Furthermore, many genetic and epigenetic events can be either directly observed or indirectly inferred from transcriptomic data. The primary readouts of modern-day cancer tran- scriptomics can be broadly categorized as genetic and functional (FIG. 2). Whereas functional measurements benefit mostly from the breadth of genome-wide assays, the detection of genetic events required increased depth and base pair resolution.

Although most transcriptomic platforms are highly reproducible by themselves, reproducibility across platforms is limited. Unfortunately, the biggest challenge is in the measurement of absolute expression levels263, which is the input to many biomarkers and signatures. Therefore, signatures cannot be expected to translate verbatim between platforms. Few studies have explored this topic. Fumagalli et al.264 concluded that single-gene expression biomarkers and established prognostic signatures generalize well between microarrays and RNA sequencing (RNA-seq). Zhang et al.265 reached a similar conclusion in that "technological platforms (RNA-seq versus microarrays) [...] do not significantly affect performances of the [predictive] models." However, these studies were done on high-quality samples and did not explore whether RNA degradation or crosslinking had a detrimental effect.

Types of gene expression analyses. Transcriptome- wide gene expression profiles are now available for the majority of cancer types and their corresponding tissues of origin. In general terms, there are two cancer-centric paths to analyse these data: the differential approach, which interprets tumour expression profiles relative to the patient-matched or unmatched normal tissue sam- ples; and the relative approach, which compares transcript levels across tumours or other samples (FIG. 3). Inherently, these strategies have unique advantages and applications. Differential analyses

are designed to detect cancer-specific changes, but if the normal samples are not comparable164, the results will be difficult to inter- pret, for example, if the cancer cell of origin is rare or unknown. In general terms, differential analyses tend to be underpowered in the clinical setting. Comparisons at the single-patient level are often limited by the dearth of replicates due to cost and sample availability, while at the cohort level, they are often confounded by interpatient heterogeneity. Relative analyses are useful to charac- terize individual samples but typically depend on the availability of external knowledge or reference data sets. The validity of any relative comparison is contingent on how well a query sample is matched to the reference in terms of technical (for example, type of data processing) and biological (for example, molecular subtype) biases. Therefore, relative analyses often necessitate advanced normalization techniques165 and batch correction166. Overall, the differential approach is more common in the research setting to generate hypotheses, whereas the relative approach drives many clinical applications, such as precision medicine. Differential approaches. The simplest type of differen- tial analysis is the identification of genes that are upreg- ulated or downregulated in cancer (that is, differentially expressed genes (DEGs)), and established methods to detect DEGs are available for both microarray167 and RNA-seq data168–170. A typical result is a long list of DEGs that is difficult to interpret without additional functional annotation, as demonstrated by a landmark study in breast cancer171. Differential methods have also been proposed for splicing93 or isoform usage172. Although transcriptomes have very high dimensionality, there is also substantial correlation among the genes, which can be leveraged to simplify or summarize the data173. A common strategy is to break down the transcrip- tome-wide gene expression profile into a set of modules that are less interdependent, more generalizable and simpler to understand. The specifics for each method differ substantially, but in general, it is possible to test for differential gene sets174, pathways, gene regulatory networks or modules in co-expression networks175. Ideally, testing multiple related genes will improve sen- sitivity and yield results that are easier to understand. For example, using a simple gene set method, Majeti et al.176 were able to identify the dysregulation of the WNT pathway in acute myeloid leukaemia. Beyond upregulation or downregulation, methods have been developed to detect less-uniform changes in gene expres- sion177; for example, detecting mechanism of action by network dysregulation (DeMAND) leverages changes in correlation to prioritize dysregulated or ‘rewired’ modules178

Cellular composition and microenvironment. The study of the heterogeneous cellular composition of tumours is one of the most recent applications

of cancer tran- scriptomics. Approaches typically involve either directly isolating and characterizing individual cells (using, for example, single-cell sequencing) or indirectly inferring cell compositions in silico from bulk expression data. From bulk expression data (which are currently more readily available from clinical samples than are sin- gle-cell data), the computational task is often referred to as sorting, or deconvolving, the gene expression pro- file. Deconvolution is a difficult problem that requires methodological constraints in order to converge on plausible solutions. A large number of algorithms have been proposed197 that make different trade-offs on the basis of the available data and the desired output. In general, the methods can be divided into those that use cell-type-specific gene signatures and can be applied to a single tumour sample and those that require multiple tumour and normal samples (matched or unmatched). Currently, the most important applications are to esti- mate tumour clonality and purity, which are affected by intrinsic tumour cell heterogeneity or infiltration by stro- mal or immune cells198. For example, in silico purification of gene expression profiles has been applied to improve their performance in prognosis199 and classification200. Deconvolution also provides a unique opportunity to study the tumour microenvironment, for exam- ple, to unravel tumour–stromal paracrine crosstalk201. In the future, single-cell transcriptomics is bound to revolutionize our understanding of the tumour microenvironment202, heterogeneity203 and evolution204.

The clinical utility of RNA- seq has been demonstrated by a number of sequencing programmes where RNA-seq identified a large num- ber of actionable genetic events221–223. Still, targeted DNA sequencing is currently the method of choice for many clinical applications in precision oncology. DNA is a highly stable analyte and is therefore well suited for molecular diagnostics.

Transcriptomics in immuno-oncology. The needfor RNA-based companion diagnostics is particu-larly acute in immuno-oncology250. Cancer immunephenotypes were shown to broadly reflect the activityof the host immune system and to generalize remark-ably well across cancer types. Numerous studies haveinvestigated the association of immune infiltrationwith survival251–253 and found significant correlationsat the level of immune cell types, inflammation signa-tures and individual genes. Although immune check-point inhibitors are broadly beneficial across cancertypes, the response rates are highly variable. It isbecoming increasingly clear that positive responses toimmunotherapy are associated with tumour immuno-genicity and host immune infiltration253–255. However,given the complexity of adaptive immune responsesand the dynamic nature of tumour–immune evasion,it is unrealistic to expect that a single gene will

besufficient to accurately predict outcomes or guidetreatment.

The clinical utility of transcriptome profiling forimmunotherapy was demonstrated in a landmarklongitudinal study that demonstrated that signaturesof adaptive immunity are predictive of response to immune checkpoint blockade254. As both prognos-tic and predictive approaches require the expressionlevels of hundreds of genes, their clinical translationwill depend on the routine use of whole-transcrip-tome profiling or custom-targeted panels256. We haveshown that comprehensive immunophenotypic datacan be obtained from clinical transcriptomes and thatthey provide unique insights into the immunologicalheterogeneity of metastatic tumours across all majorprimary tissue types223. RNA-seq data are also particu-larly valuable for the development of personalized can-cer vaccines257,258, where they can be used to identifychimeric fusion proteins that contain putative mutantepitopes245 and help in the selection of potentiallyhighly abundant neoantigens.

The complexity of tumour–immune cell interactionsis mirrored by the diversity of bioinformatics approachesto characterize them. Both data-driven198 and know-ledge-driven253 approaches have been proposed to quan-tify the overall level of tumour–immune infiltration. Inaddition, recent methodological advances made it pos-sible to estimate cell-type fractions from bulk tumourexpression profiles in a process referred to as in silicocell sorting259,260, which is similar to the ‘purification’ ofthe tumour cell expression profiles discussed above199,200.Finally, clonal expansion of antitumour T cells can bedetected by the presence of somatically rearranged TCRsequences, that is, clonotypes261. An analogous strategycan be applied to B cells and immunoglobulin loci262.As neoantigen prediction remains a daunting problem,RNA-seq data are useful for both the detection of pro-tein-altering genetic aberrations and their prioritizationbased on expression levels.

(**?** transcriptome profiling at the juncture of clinical translation)

Bulk RNA-seq data can easily be obtained from either flash-frozen or formalin-fixed, paraffin-embedded (FFPE) tissue samples, including both surgically resected material and core needle biopsies. (**?** et al. Estimation of immune cell content using single cell data).

1.3.3.2 methylome

Changes in gene expression in tumours owing to epigenetic modifi-cations and the expression of microRNAs probably contribute directly to determining the immune microenvironment and immunogenicity of a tumour. Cytokine expression during T-cell development is regu-lated by epigenetic alterations to both DNA and chromatin96. Cancer can also be accompanied by epigenetic changes, which makes it prob-able that such changes will influence cytokine profiles that modulate the immune microenvironment. In fact, DNA methylation in lung-cancer cells has been shown to reduce the expression of IL-1□97. And PD-L1 expression can be modulated by microRNAs, with miR-200 (a repressor of epithelial-to-mesenchymal transition) and possibly others decreasing its expression98. Methylation of the promoter for the gene PD-L1 itself also seems to repress PD-L1 expression; demethylation can result in constitutive expression in tumours, especially non-small cell lung cancer99..

Another influence on the immune profile of a tumour that has an epigenetic mechanism involves the tissue of origin of the tumour. Colorectal cancer tumours commonly express elevated levels of transforming growth factor (TGF)-□100. Presumably, this reflects the importance of the TGF-□ pathway in intestinal biology and, especially, its role in maintaining tolerance to the gut microbiota by favour-ing the development of regulatory T cells101. Elevated expression of TGF-□ may also contribute to the development of abundant stromal elements in these tumours that can restrict the access of immune cells to the tumour parenchyma, as has been demonstrated in pancreatic cancer102. Although other factors also contribute, it is interesting to note that pancreatic cancer and most forms of colorectal cancer (except for the mutationally rich microsatellite-instability-high sub-group85) respond poorly to single-agent inhibition of PD-L1/PD-1 (refs 62, 83 and 103–105).

(**?** CANCER CIRCLE)

1.3.3.3 single cell

Described above methods of process DNA from hundreds of thousands of cells simultaneously and report averaged gene expression of all cells. In contrast, scRNA-seq technology allows getting results for each cell individually. This is tremendous step forward enhancement of our understanding of cell heterogeneity and opens new avenues of research questions.

Continuous discovery of new immune subtypes has proven that cell surface markers

that are used for phenotyping by techniques like FACS and immunohistochelistry cannot capture the full complexity. ScRNA-seq methods allow to cluster known cell types in subpopulations based on their genetic features. (**?** RNA sequencing to explore immune cell heterogeneity). ScRNA-seq is also able to capture particularly rare cell types as it requires much less of RNA material (1 ng isolated from 100-1000 cells) compared to 'bulk' RNA-seq (~ 1 µg of total mRNA transcripts)

new cellular states

> In summary, these studies have established that sur-face phenotypes are not sufficient to define cellular states in disease and have proposed new scRNA-seq methods to study innate immunological processes as well as dis-ease pathogenesis and progression at high resolution (**?** RNA sequencing to explore immune cell heterogeneity)

This new data type also brings into the field new challenges related to data processing due to the volume, distribution, noise, and biases. Experts highlight as the most "problematic" "batch effect" and noise and "dropout effect" (**?**). So far, there are no official standards that can be applied which makes data comparison and post-processing even more challenging. Up to date, there are around 70 reported tools and resources for single cell data processing (@ GitHub, called 'Awesome Single Cell' (go.nature.com/2rmb1hp)) .

A limited number of single-cell datasets of tumors are made publicly available (**?**).

One can ask why then developing computational deconvolution of transcriptome if we can learn relevant information from single-cell data. Today's reality is that single cell data does not provide a straightforward answer to the estimation of cell proportions. The coverage is not full and sequenced single cells are not fully representative of the true population. For instance, neutrophiles are not found in scRNA-seq data because of they are "difficult to isolate, highly labile *ex vivo* and therefore difficult to preserve with current single-cell methods" (**?** et al. Estimation of immune cell content using single cell data). In addition, a number of patients included in published studies of range <100 cannot be compared to thousand people cohorts sequenced with bulk transcriptome methods. This is mostly because single cell experiments are challenging to perform, especially in clinicsal setting as fresh samples are needed.(**?** et al. Estimation of immune cell content using single cell data). Today, single cell technology brings very interesting "zoom in" perspective, but it would be incautious to make fundings from a restricted group of individuals universal to the whole population. Major brake to the use of single cell technology more broadly might be as well the price that is neatly 10x higher for single cell sample compared to bulk (**?**—June-2017.pdf).** (A table?)**

Technology	Price per sample
scRNA-seq	3000 $

Technology	Price per sample
RNA-seq	200 $

In this work, we are using single cell data in two ways. Firstly, in Comparative... chapter we compare immune cell profiles defined by scRNA-seq, blood and blind deconvolution (problem introduced in Immune signatures section). Secondly, in Heterogeneity of immune... we use single call data of Metastatic melanoma generated by Tirosh et al. (**?** Melanoma sc) to demonstrate subpopulations of Macrophages and NK cells.

Chapter 2

Mathematical foundation of cell-type deconvolution of biological data

In this chapter, we will discuss how mathematical models can be used to extract information about specific cell-types from 'bulk' data or how to unmix mixed sources. It will introduce you to basic concepts of data analysis as well as most popular advanced solutions adapted for estimating presence and proportion of immune cells within cancer biopsies.

2.1 Introduction to supervised and unsupervised learning

2.2 Blind source sepration

(ICA, NMF etc)

2.3 Finding optimal number of components and over-decomposition of transcriptopmes

(adapted from BMC article)

Kairov et al. BMC Genomics (2017) 18:712
DOI 10.1186/s12864-017-4112-9

BMC Genomics

RESEARCH ARTICLE Open Access

Determining the optimal number of independent components for reproducible transcriptomic data analysis

Ulykbek Kairov[2†], Laura Cantini[1†], Alessandro Greco[1], Askhat Molkenov[2], Urszula Czerwinska[1], Emmanuel Barillot[1] and Andrei Zinovyev[1*†]

Abstract

Background: Independent Component Analysis (ICA) is a method that models gene expression data as an action of a set of statistically independent hidden factors. The output of ICA depends on a fundamental parameter: the number of components (factors) to compute. The optimal choice of this parameter, related to determining the effective data dimension, remains an open question in the application of blind source separation techniques to transcriptomic data.

Results: Here we address the question of optimizing the number of statistically independent components in the analysis of transcriptomic data for reproducibility of the components in multiple runs of ICA (within the same or within varying effective dimensions) and in multiple independent datasets. To this end, we introduce ranking of independent components based on their stability in multiple ICA 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. Based on a large body of data, we demonstrate that a sufficient number of dimensions is required for biological interpretability of the ICA decomposition and that the most stable components with ranks below MSTD have more chances to be reproduced in independent studies compared to the less stable ones. At the same time, we show that a transcriptomics dataset can be reduced to a relatively high number of dimensions without losing the interpretability of ICA, even though higher dimensions give rise to components driven by small gene sets.

Conclusions: We suggest a protocol of ICA application to transcriptomics data with a possibility of prioritizing components with respect to their reproducibility that strengthens the biological interpretation. Computing too few components (much less than MSTD) is not optimal for interpretability of the results. The components ranked within MSTD range have more chances to be reproduced in independent studies.

Keywords: Transcriptome, Independent component analysis, Reproducibility, Cancer

Background

Independent Component Analysis (ICA) is a matrix factorization method for data dimension reduction [1]. ICA defines a new coordinate system in the multi-dimensional 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 [1]. ICA has been widely applied for the analysis of transcriptomic data for blind separation of biological, environmental and technical factors affecting gene expression [2–6].

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 [6, 7]. 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

* Correspondence: Andrei.Zinovyev@curie.fr
†Equal contributors
[1]Institut Curie, PSL Research University, INSERM U900, Mines ParisTech, Paris, France
Full list of author information is available at the end of the article

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 [3, 8]. In [3] 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. ICA was also able to unravel the organization of tumor microenvironment such as the presence of lymphocytes B and T, myofibroblasts, adipose tissue, smooth muscle cells and interferon signaling. This analysis shed light on the principles underlying bladder cancer molecular subtyping [3].

It has been demonstrated that ICA has advantages over the classical Principal Component Analysis (PCA) with respect to interpretability of the resulting components. The ICA components might reflect both biological factors (such as proliferation or presence of different cell types in the tumoral microenvironment) or technical factors (such as batch effects or GC-content) affecting gene expression [3, 5]. However, unlike principal components, the independent components are only defined as local minima of a non-quadratic optimization function. Therefore, computing ICA from different initial approximations can result in different problem solutions. Moreover, in contrast to PCA, the components of ICA cannot be naturally ordered.

To improve these aspects, several ideas have been employed. For example, an *icasso* method has been developed to improve the stability of the independent components by: (1) applying multiple runs of ICA with different initializations; (2) clustering the resulting components; (3) defining the final result as cluster centroids; and (4) estimating the compactness of the clusters [9]. The resulting components can be then naturally ordered from the most stable to the least stable ones. This ranking is usually different from more commonly used independent component rankings based on the value of the used non-gaussianity measure (such as kurtosis) or the variance explained by the components.

The fundamental question is the determination of the number of independent components to produce. This problem can be split into two parts: a) what dimension should be selected for reducing the transcriptomic data before applying ICA (determining the effective data dimension); and b) which is the most informative number of components to use in the downstream analysis?

Determining the optimal effective data dimension for application of signal deconvolution was a subject of research in various fields. For example, ICA appeared to be a powerful method for analyzing the fMRI (functional magnetic resonance) data [9–12]. In this field, it was shown that choosing a too small effective data dimension might generate "fused components," not reflecting the heterogeneity of the data, leading to a loss of interesting sources (under-decomposition). At the same time, choosing the effective dimension too high might lead to signal-to-noise ratio deterioration, overfitting and splitting of the meaningful components (over-decomposition) [10–12]. The influence of the effective dimension choice on the ICA performance has not been well studied in the context of transcriptomic data analysis. For example, in [3] each dataset was decomposed into a number of components in an ad hoc manner (n = 20).

Several theoretical approaches for estimating effective data dimension exist. The simplest ones, developed for PCA analysis, are represented by the Kaiser rule aimed at keeping a certain percentage of explained variance and the broken stick model of resource distribution [13]. More sophisticated approaches employ the information theory (e.g., Akaike's information or Minimal Description Length criteria) [13] or investigate the local-to-global data structure organization [14]. Also, computational approaches based on cross-validation have been suggested in the literature [15]. Specifically for ICA analysis, few methods have been proposed to optimize the effective dimension. For example, the Bayesian Information Criterion (BIC) can be applied to the Bayesian formulation of ICA for selecting the optimal number of components [16].

Although many of the above theoretical methods are "parameter-free," selecting the best method for choosing an effective dimension for transcriptomic data can be challenging in the absence of a clearly defined validation strategy. One possible approach to overcome this limitation is to apply the same computational method to multiple transcriptomic datasets derived from the same tissue and disease. In this situation, it is reasonable to expect that a matrix factorization method should detect similar signals in all datasets. By taking advantage of the rich collection of public data such as The Cancer Genomic Atlas (TCGA) [17] and Gene Expression Omnibus [18], it is possible to compare and contrast the parameters of different gene expression analysis methods such as ICA.

In this study, we used TCGA pan-cancer (32 different cancer types) transcriptomic datasets and a set of six independent breast cancer transcriptomic datasets to evaluate the effect of the number of computed independent components on reproducibility and biological interpretability of the obtained results. We evaluated the reproducibility of ICA on three aspects: First, we analyzed the stability of the computed components with respect to multiple runs of ICA; second, we analyse the conservation of the computed components by varying the choice of the reduced data dimension; and third, we consider the reproducibility of the resulting set of ICA

metagenes across multiple independent datasets. Our reproducibility analysis thus explores 13,027 transcriptomic profiles in 37 transcriptomic datasets, for which more than 100,000 ICA decompositions have been computed.

We finally defined a novel criterion adapted for choosing the effective data dimension for ICA analysis of gene expression, which takes into account the global properties of transcriptomic multivariate data. The Maximally Stable Transcriptome Dimension (MSTD) is defined as the maximal dimension where ICA does not yet produce a large proportion of highly unstable signals. By numerical experiments, we showed that components ranked by stability within the MSTD range tend to be more reproducible and easier to interpret than higher-order components.

Results

Definition of component reproducibility measures used in this study

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*. The exact index used for quantifying the clustering is documented in the Methods section. Conservation of an independent component in terms of choosing various orders of 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*. For example, 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. The details on computing correlations between components from different datasets are described in Methods.

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

We used 37 transcriptomic datasets to analyze the stability and reproducibility of the ICA results conditional on the chosen number of components. ICA has been applied separately to 37 cancer transcriptomic datasets following the ICA application protocols as described in Methods.

The proposed protocol depends on a fundamental parameter M (effective dimension of the data and, at the same time, the number of computed independent components) whose effect on the stability of the ICs is investigated. For each transcriptomic dataset, the range of M values 2–100 has been considered. For each value of M, the data dimension is reduced to M by PCA and then data whitening is applied. Subsequently, the actual signal decomposition is applied in the whitened space by defining M new axes, each maximizing the non-gaussianity of data point projections distribution.

For transcriptomic data, ICA decomposition provides: (a) M metagenes ranked accordingly to their stability in multiple runs (n = 100) of ICA; and (b) a profile of stability of the components (set of M numbers in [0,1] range in descending order). Considering the largest dataset METABRIC as an example, the behavior of the stability profile as a function of M is reported in Fig. 1a. The results for stability analysis for other breast cancer datasets are similar (See Additional file 1: Figure SF2). To recapitulate the behaviour of many stability profiles, the average stability of the first k top-ranked components $S_M(k)$ is used (See Fig. 1b). For k = M, the average stability of all computed components is denoted as S_M^{total}. Three major conclusions can be made from Fig. 1. First, the average stability of the computed components S_M^{total} decreases with the increase of M, while the average stability of the first few top ranked components, e.g., $S_M(10)$, weakly depends on M (Fig. 1b). Moreover, S_M^{total} is characterized by the presence of local maxima, defining certain distinguished values of M that correspond to the (locally) maximally stable set of components (Fig. 1b). Third, the stability profiles for various values of M can be classified into those for which the stability values are distributed approximately uniformly and those (usually, in higher dimensions) forming a large proportion of the components with low stability (I_q between 0.2 and 0.4) (Fig. 1a).

Considering these observations, we hypothesized that the optimal number of independent components – large enough to avoid fusing meaningful components and yet small enough to avoid producing an excessive amount of highly unstable components – should correspond to the inflection point in the distribution of the stability profiles (Fig. 1a). To find this point, the stability measures have been clustered along two lines, which is analogous of 2-means clustering but with lines as centroids. In this clustering, the line with a steeper slope (Fig. 1a, blue line) grouped the stability profiles with uniform distribution, while another line (Fig. 1a, red line) matched the mode of low stability components. The intersection of these lines provided a consistent estimate of the effective

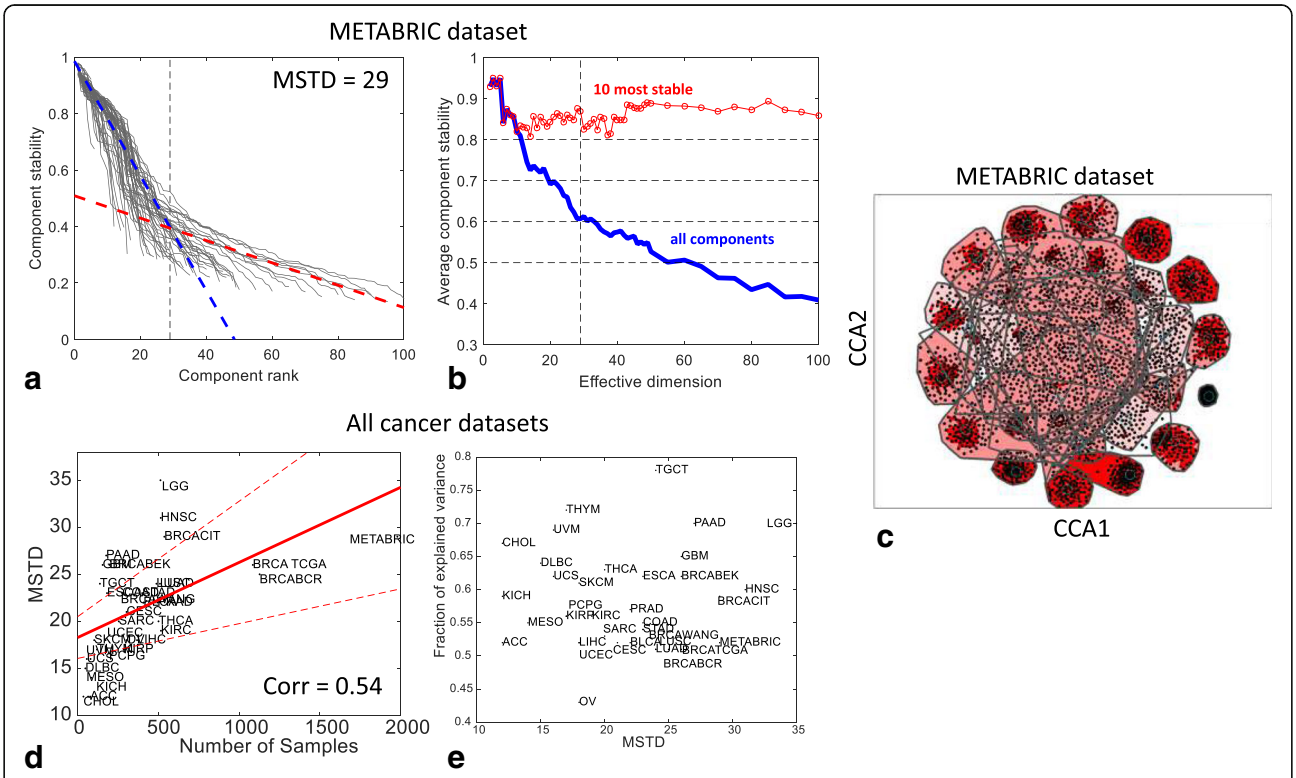


Fig. 1 Defining Maximally Stable Transcriptomic Dimension (MSTD) value in 37 transcriptomics cancer datasets (13,027 samples in total). In **a**-**c**) an example of the analysis is presented for the largest breast cancer dataset METABRIC. **a** stability profiles for ICA decompositions in various dimensions (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*). **b** average stability profile S_M^{total} (*blue line*) and the average stability of 10 most stable components $S_M(10)$ (*red line*). **c** visualizing the results of computing ICA 100 times with MSTD = 29 components in the 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. **d** Dependence of MSTD on the number of samples for all datasets. **e** Dependence of the fraction of explained variance on MSTD for all datasets

number of independent components. We call this estimate Maximally Stable Transcriptome Dimension (MSTD) and in the following we investigated its properties. We note that, as in various information theory-based criteria (BIC, AIC), this estimate is free of parameters (thresholds), and it only exploits the property of the qualitative change in the character of the stability profile in higher data dimensions for transcriptomic data.

In most of the cancer transriptomics datasets used in our analysis, MSTD was found to correspond roughly to the average stability profile $S_M^{total} \approx 0.6$ (Additional file 1: Figure SF2). In Fig. 1d, the dependence of MSTD on the number of samples contained in the transcriptomic dataset is investigated for all the 37 transcriptomic datasets. As shown in Additional file 2: Figure SF1, MSTD increased with the number of samples; however, this trend was weaker than other estimates of an effective dimension such as Kaiser rule and broken stick distribution-based data dimension estimates. Finally, the fraction of variance explained by the linear subspace spanned by MSTD number of components was evaluated (Fig. 1e), and it was observed that the fraction of variance explained varied from 0.45 to 0.75 with a median of 0.56.

Underestimating the effective dimension (M < MSTD) leads to a poor detection of known biological signals

Previous large-scale ICA-based meta-analyses [3] have shown that some of the ICs derived from the decomposition of a cancer transcriptomic data were clearly and uniquely associated with known biological signals. For example, one of these signals was the one connected to proliferative status of tumors. Another example was given by the signals related to the infiltration of immune cells that were also strongly heterogeneous across cancer patients.

We have checked the reproducibility of several metagenes obtained in previous meta-analyses [3] for all ICA decompositions as a function of M. For this analysis, we employed the METABRIC breast cancer dataset, which was not included in the input data of the previous publication [3] and thus it had not been used to derive the metagenes of that work. In addition, we checked how

the significance of intersections between the genes defining the components and several reference gene sets (produced independently of the ICA analyses) behaved as a function of M.

We applied the previously developed correlation-based approach to match previously identified metagenes with the ones computed for a new METABRIC dataset (see Methods section). The components were oriented accordingly to the direction of the heaviest tail of the projection distribution. When matching an oriented component to the previously defined set of metagenes, we verified that the resulting maximal correlation should be positive, i.e. large positive weights in one metagene should correspond to large positive weights in another metagene.

One of the most important case studies is reproducibility of the "proliferative" metagene in different data dimensions. It is investigated in Fig. 2a-c. For this metagene, we computed correlations with M newly identified independent components. As an example, the profile of correlations for M = 100 is shown in Fig. 2b. It can be seen that one of the components (ranked #7 by stability analysis) is much better correlated to the proliferative metagene than any other component. Therefore, component #7 is called "best matched" in this case, for M = 100, and "well separable." Repeating this analysis for all M and reporting the observed maximal correlation coefficient and the corresponding stability value gives a plot shown in Fig. 2a. Separability of the best matched component from the other components is visualized in Fig. 2c.

As it can be seen from Fig. 1a, the biologically expected signals (i.e., cell cycle) can be poorly detected for M < MSTD; however, once the best matching component with significant correlation was found, it remained unique and was detected robustly even for very large values of M> > MSTD. For example, even when 100 components (M) were computed, the correlation between the previously defined proliferative metagene and the best matched independent component did not diminish (Fig. 2a). Moreover, the separability of the best matched component from the rest of the components was not ruined (Fig. 2c). In this example, the identification of cell cycle component remained clear (large and well-separated correlation coefficient) for M> > MSTD. This result was consistent and complementary when compared with the previously observed weak dependence of $S_M(10)$ on M. Indeed, the "proliferative" best matched component had stability rank k in the range [6, 11]. That is, it remained stable in ICA decompositions in all dimensions. Moreover, the intersection of a recently established proliferation gene signature [19] with the set of top contributing genes of the best matched component improved with increasing M and saturated (Fig. 2d). This proves that the detection of the proliferation-associated signal with ICA does not depend on the ICA-based definition of the proliferative metagene.

Together with the proliferative signal, other metagenes from the previously cited ICA-based meta-analysis [3] were robustly identified in our analysis. In Fig. 2e-h, we showed the correlation with the best matching component for the metagenes associated with the presence of myofibroblasts, inflammation, interferon signaling and immune system, as a function of M. These plots illustrated different scenarios that can result from such analysis. The myofibroblast-associated metagene was robustly detected for all values of M > 7 (Fig. 2f). However, the stability of the best matching component was deteriorated in higher-order ICA decompositions (M > 45). For the inflammation-associated metagene, an ICA decomposition with M > 38 was needed to robustly detect a component that correlates with the metagene (Fig. 2e).

Interestingly, the immune-associated metagene was found robustly matched starting from M = 4. However, in higher-order decompositions (starting from M = 30) it could be matched to several components that can be associated with specific immune system-related signals (Fig. 2h-i). Hypergeometric tests applied to the sets of top-contributing genes (weights larger than 5.0) allowed us to reliably interpret these components as being associated with the presence of three types of immune-related cells: T cells (corrected enrichment p-value = 10^{-39} with "alpha beta T cells" signature [20], other immune signatures are much less significant), B cells (p-value = 10^{-7} with "B cells, preB.FrD.BM" signature) and myeloid cells (p-value = 10^{-78} with "Myeloid Cells, DC.11cloSer.Salm3.SI" signature).

Overestimating the number of components (M> > MSTD) produces multiple ICs driven by small gene sets

We observed that the higher-order ICA decompositions (M> > MSTD) produced a larger number of components driven by small gene sets (frequently, one gene), such that the projections of the genes in this "outlier" set is separated by a relatively large gap with the rest of the projections. We thus designed a simple algorithm to distinguish such components driven by a small gene set from all the others. The names of the genes composing these small sets were used for annotating the corresponding components (Fig. 3a, right part).

It was observed that the presence of such "small gene set-driven" components is a characteristic of higher-order ICA decompositions (M> > MSTD), much less present in ICA decompositions with M ≤ MSTD (compare Fig. 3a and Additional file 1: Figure SF2).

To check the biological significance of the outlier genes, we considered as a case study the higher-order (M = 100) ICA decomposition of the METABRIC breast cancer dataset. We collected all those genes found to be

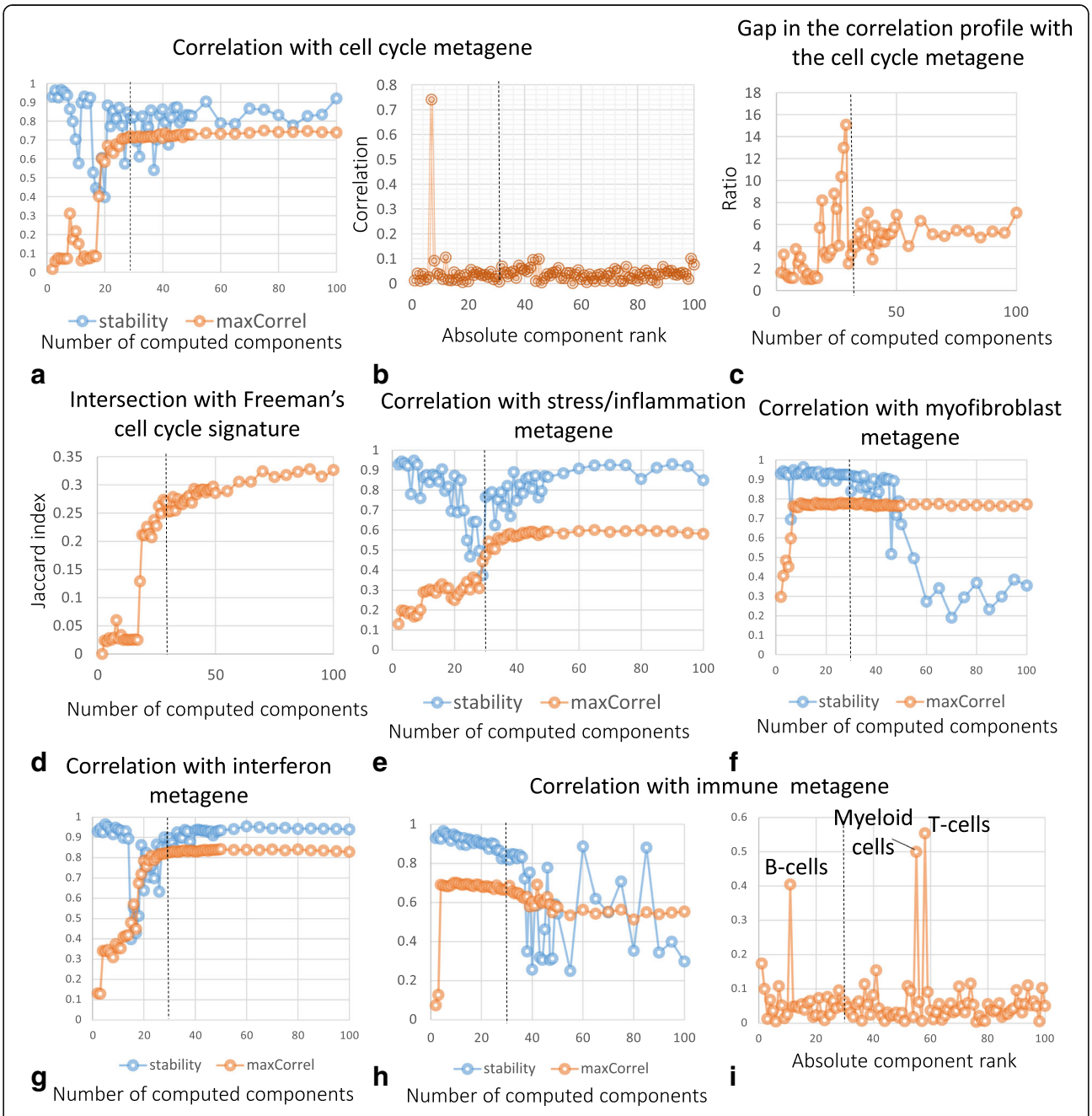


Fig. 2 Analysis of reproducibility of previously identified metagenes in independent components of the METABRIC dataset. **a**,**e**,**f**,**g**,**h** show correlations (see Methods section) with the cell cycle, inflammation, myofibroblast, interferon signaling, immune-related metagenes from [3] as a function of the chosen data dimension M, and the stability of the best matched component. **c** shows the ratio between the correlation value of the proliferation metagene with the best matched component and the second best correlation (gap). **d** shows an intersection (Jaccard index) of the Freeman's cell cycle signature [19] and the set of top-contributing genes (projection > 5.0) from the proliferation-associated independent component. **b**,**i** correlation of the cell cycle and immune-related metagene with the best matched component in the M = 100 ICA decomposition as a function of the stability-based component rank. In all plots, the *vertical dashed line* shows the MSTD value for the METABRIC dataset

drivers of at least one "small gene set-driven" component. We obtained in this way a set of 98 genes listed in Additional file 3: Table ST2. This list appeared to be strongly enriched (p-value = 10^{-12} after correction for multiple testing) in the genes of the signature DOANE_BREAST_CANCER_ESR1_UP "Genes up-regulated in breast cancer samples positive for ESR1 compared to the ESR1 negative tumors" from Molecular Signature Database [21] and several other specific to breast cancer gene signatures. This analysis thus

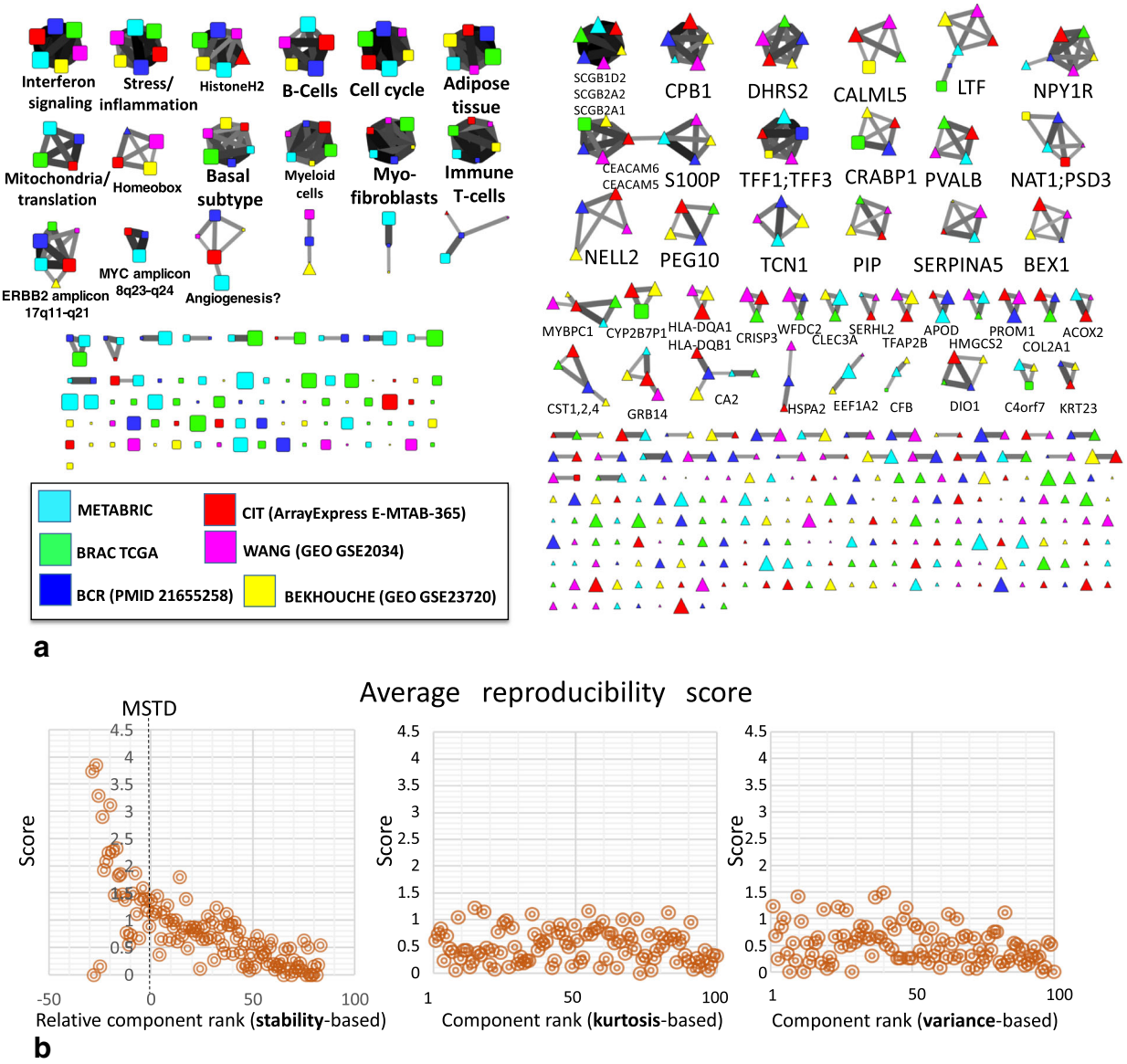


Fig. 3 Analysis of component reproducibility in independent datasets. **a** Graph of reciprocal correlations showing the reproducibility of the metagenes of independent components in 6 independent breast cancer datasets. Each node here is an independent component, represented by a metagene, from an ICA decomposition with M = 100 components. Edges show only reciprocal correlations between metagenes with Pearson correlation >0.3. Triangles (on the right) show the components driven by the expression of a small group of genes (frequently, one gene). Node size reflects the rank of the component based on the stability in multiple runs of fastICA (larger nodes are more stable ones). The edge width and the color reflect the value of the correlation coefficient between two metagenes, with thicker edges showing larger correlation values. Several pseudo-cliques of highly reproducible components are annotated either by the dominating small group of genes (pseudo-cliques of triangle nodes), or by comparing to the results of the previously published large-scale ICA-based analysis of gene expression [3] or by performing the hypergeometric test using the set of top-contributing genes (with projection larger than 5.0 onto the component). The analogous correlation graph computed for MSTD number of components is provided in Additional file 7: Figure SF3. **b** average reproducibility score (sum of reciprocal correlation coefficients for an independent component) for the correlation graph shown in a), as a function of the relative (component rank minus MSTD value for a given dataset, for stability-based ranking) or absolute (for other ranking types) component rank. It is clear that only stability-based ranking matches the reproducibility score

suggested that at least some of the identified "small gene set-driven" components are not the artifacts of the ICA decomposition, but they can be biologically meaningful and reproducible in independent datasets (Fig. 3a, right part).

Most stable components with stability rank ≤ MSTD have more chances to be reproduced across independent datasets for the same cancer type

It would be reasonable to expect that the main biological signals characteristic for a given cancer type should be the same when one studies molecular profiles of different independent cohorts of patients. Therefore, we expect that for multiple datasets related to the same cancer type, ICA decompositions should be somewhat similar; hence, reciprocally matching each other. We called this expected behavior "reproducibility," and here we studied this by applying ICA to six relatively large breast cancer transcriptomic datasets. Of note, these datasets were produced using various technologies of transcriptomic profiling (Additional file 4: Table ST1).

To identify the reproducible components, we applied the same methodology as in the previously published ICA-based gene expression meta-analysis [3]. We decomposed the six datasets separately and then constructed a graph of reciprocal correlations between the obtained metagenes. Correlation between two sets of components is called reciprocal when a component from one set is the best match (maximally correlated) to a component from another set, and vice versa (see Methods for a strict definition).

Pseudo-cliques in this graph, consisting of several nodes, correspond to reproducible signals detected by ICA. As shown in Fig. 3, multiple reproducible signals were identified in the analysis. Some of them correspond to signals already identified in [3] (e.g., cell cycle, interferon signaling, microenvironment-related signals), and some correspond to newly discovered biological signals (e.g., ERBB2 amplicon-associated). Some other pseudo-cliques are associated with "small gene set-driven" components (frequently, one gene-driven), such as TFF1–3-associated or SCGB2A1–2-associated components.

The genes driver of reproducible and "small gene set-driven" components (S100P, TFF1, TFF3, SCGB2A1, SCGB1D2, SCGB2A2, LTF, CEACAM6, CEACAM5 being most remarkable examples) have been investigated in detail, to further check their biological interest. They were found to be the genes known to be associated with breast cancer progression [22]. For example, seven of the nine previously mentioned genes form a part of a gene set known to be up-regulated in the bone relapses of breast cancer (M3238 gene set from MSigDB).

To quantify the reproducibility of the components, we computed a reproducibility score. It is a sum of correlation coefficients between the component and all reciprocally correlated components from other datasets. By construction, the maximum value of the score is 5, which meant that a component with such a score would be perfectly correlated with the reciprocally related components from five other datasets. We studied the dependence of this score as a function of the relative to MSTD component stability-based rank (Fig. 3b). From this study, it follows that even for the high-order ICA decompositions, the components ranked by their stability within MSTD range, have an increased likelihood of being reproduced in independent datasets collected for the same cancer type.

To show that the stability-based ranking of genes is more informative compared with the standard rankings of independent components, we performed a computational analysis in which we compared the stability-based ranking with the rankings based on non-gaussianity (kurtosis) and explained variance. These two measures are frequently used to rank the independent components [6]. From Fig. 3b it is clear that the stability-based ranking of independent components corresponds well to the reproducibility score, while two other simpler measures do not.

It can also be shown that the total number of reciprocal correlations with relatively large correlation coefficients ($|r| > 0.3$) between ICA-based metagenes computed for several independent datasets is significantly bigger when the component stabilization approach is applied (Additional file 5: Figure SF4). This proves the utility of the applied stabilization-based protocol of ICA application to transcriptomic data.

Computing large number of components (M> > MSTD) does not strongly affect the most stable ones

We lastly used ICA decompositions of 37 transcriptomic datasets to compare the ICA decompositions corresponding to M = MSTD with the higher-order decompositions, M = 50 or M = 100.

It was found that the components calculated in lower data dimensions can be relatively well matched to the components from higher-order ICA decompositions (Fig. 4). More precisely, 90% of the components defined for M = MSTD had a reciprocal best matched component in the M = 100 ICA decomposition. Most stable components had a clear tendency to be reproduced with high correlation coefficient ($r > 0.8$). Only 10% of the components had only non-reciprocal or too small correlations between two decompositions (in other words, *not conserved* in higher-order ICA decompositions).

Approximately 15% of the components in M = MSTD ICA decomposition together with reciprocal maximal correlation also had a non-reciprocal correlation to one of the components in M = 100 ICA decomposition (Fig. 4). This case can be described as splitting a component into

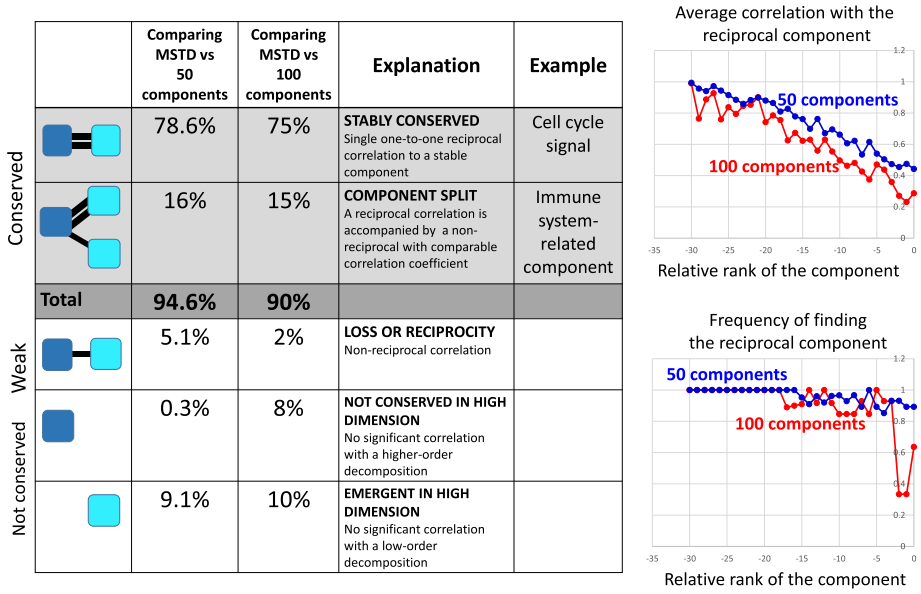


Fig. 4 Conservation or non-conservation of independent components in higher than MSTD dimensions. ICA decomposition in MSTD data dimension is compared with the one of dimension 50 and 100. Table on the left: list of different scenario with the relative frequency of each of them estimated for a pan-cancer TCGA dataset. Plots on the right show the frequency of finding the reciprocally correlated component in the higher dimension and the average correlation coefficient of the reciprocal correlation, as a function of the relative rank of the component (component rank minus MSTD value for a given dataset)

two or more components in the higher-order ICA decompositions. At least one such split had a clear biological meaning, namely the splitting of the component representing the generic "immune infiltrate." The resulting "split" components more specifically represented the role of T cells, B cells and myeloid cells in the tumoral microenvironment (see the "*Underestimating the effective dimension…*" Results section).

Discussion

Our results shed light on the organization of the multivariate distribution of gene expression in the high-dimensional space. It appears that the organization contained two relatively well separated parts: *the dense one* of a relatively small effective dimension and *the sparse one*. The former contained the genes from within co-regulated modules that contained from few tens to few hundreds of genes. The latter was spanned by the genes with unique regulatory programs (perhaps tissue-specific) weakly shared by the other genes. Here the sparsity was understood in the sense of low local multivariate distribution density.

Independent Component Analysis can capture both these parts of the multivariate distribution. However, while the dense part defined independent components with approximately uniformly distributed stabilities, starting from highly stable to less stable, the sparse part was spanned by the components characterized mostly by small stability values.

This organization of the gene expression space is captured in the distribution of ICA stability profiles for varying M, which allowed us to define the Maximally Stable Transcriptome Dimension (MSTD) value, roughly reflecting the dimension of the dense part of the gene expression distribution. In one hand, when underdecomposing (compressing too much by dimension reduction, M < MSTD) a transcriptomic dataset, the resulting independent components are hard to interpret. In the other hand, overdecomposing transcriptomes (choosing the effective dimension much bigger than MSTD) is not dramatically detrimental: one can choose to explore a relatively multi-dimensional subspace of a transcriptomic dataset, taking into account that applying matrix factorization methods in higher dimensions becomes computationally challenging and prone to bad algorithm convergence. Nevertheless, higher-order decompositions might allow capturing the behavior of some tissue-specific or cancer type-specific biomarker genes from the sparse part of the distribution, which can be found reproducible in other independent studies.

In our computational experiments, we selected 100 as the maximum order of ICA decomposition (M) to test. However it is possible to examine even higher orders of ICA decompositions, reducing the data to more than 100 dimensions, but not more than the total number of samples, of course. In practice, computing ICA in such high dimension leads to significant deterioration of the fastICA algorithm convergence, so exploring M > 100

might be too expensive in terms of computational time. Moreover, our study suggests that the most interesting for interpretation components are usually positioned within the first few ten top ranks: therefore, 100 seems to be a reasonable limit for dimension reduction when applying ICA to transcriptomic data.

Our proposed approach can be used for comparing intrinsic reproducibility, at different levels, of various matrix factorization methods. For example, it would be of interest to compare the widely used Non-negative matrix factorization (NMF) method [6, 7] with ICA to assess reproducibility of extracted metagenes in independent datasets of the same nature.

More generally, systematic reproducibility analysis can be a useful approach for establishing the best practices of application of the bioinformatics methods.

Conclusion

By using a large body of data and comparing 0.1 million decompositions of transcriptomic datasets into the sets of independent components, we have checked systematically the resulting metagenes for their reproducibility in several runs of ICA computation (measuring *stability*), for their reproducibility between a lower order and higher-order ICA decompositions (*conservation*), and between metagene sets computed for several independent datasets, profiling tumoral samples of the same cancer type (*reproducibility*).

From the first of such analyses, we formulated a minimally advised number of dimensions to which a transcriptomic dataset should be reduced called Maximally Stable Transcriptome Dimension (MSTD). Reducing a transcriptomic dataset to a dimension below MSTD is not optimal in terms of the interpretability of the resulting ICA components. We showed that for relatively large transcriptomic datasets, MSTD could vary from 15 to 30 and that the number of samples matters relatively weakly.

From the second analysis, we concluded that the suggested protocol of ICA application to transcriptomic data is conservative, i.e., the components identified in a higher dimension (for example, in one hundred dimensional space) can be robustly matched with those components obtained in the dimensions comparable with MSTD. Moreover, we described an effect of interpretable component splitting in higher dimensions, leading to detection of finer-grained signals (e.g., related to the decomposition of the immune infiltrate in the tumor microenvironment). At the same time, the application of ICA in high dimensions resulted in a greater proportion of unstable components, many of them were driven by expression of small (one to three members) gene sets. Yet, some of these small gene set-driven components were highly reproducible and biologically meaningful.

From the third analysis, we established that the used protocol of ICA application, with ranking the independent components based on their stability, prioritized those components having more chances to be reproduced in independent transcriptomic datasets. Moreover, when ICA was applied in higher dimensions, the components within the MSTD range still have more chances to be reproduced.

In sum, our results confirmed advantageous features of ICA applied to gene expression data from different platforms, leading to interpretable and quantifiably reproducible results. Comparing ICA analyses performed in various dimensions and multiple independent datasets for the same cancer types allow prioritizing of the most reliable and reproducible components which can be quantitatively recapitulated in the form of metagenes or the sets of top contributing genes. We expect that ICA will demonstrate similar properties in other large-scale transcriptomic data collections such as scRNA-seq data.

Methods

Transcriptomics cancer data used in the analysis

Expression data derived for 32 solid cancer types (ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM) were downloaded from the TCGA web-site and internally normalized. Normalized breast cancer datasets from CIT, BCR, WANG, BEKHOUCHE were re-used from the previous study [3]. Normalized METABRIC breast cancer expression dataset was downloaded from cBioPortal at this link http://www.cbioportal.org/study?id=brca_metabric. When it was not already the case, the data values were converted into logarithmic scale.

The list of breast cancer transcriptomic datasets used for reproducibility study is available in Additional file 4: Table ST1.

ICA decompositions computation

We applied the same protocol of application of ICA decomposition as in [3]. In the ICA decomposition $X \approx AS$, X is the gene expression (sample vs gene) matrix, A is the (sample vs. component) matrix describing the loadings of the independent components, and S is the (component vs. gene matrix) describing the weights (projections) of the genes in the components. To compute ICA, we used the *fastICA* algorithm [1] accompanied by the *icasso* package [23] to improve the components estimation and to rank the components based on their stability. ICA was applied to each transcriptomic dataset separately.

For each analysed transcriptomic dataset, we computed *M* independent components (ICs), using *pow3* nonlinearity and *symmetrical* approach to the decomposition, where M = [2…50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]. In those

cases, when M exceeded the total number of samples, the maximum M was chosen equal to 0.9 multiplied by the number of samples (moderate dimension reduction improves convergence). We found that the MATLAB implementations of *fastICA* performs superior to other implementations (such as those provided in *R* [24]). The computational time required for performing all the 0.1 million ICA decompositions used in this study is estimated in ~1500 single processor hours using MATLAB while other implementations would not make this analysis feasible at all. In our analysis, we used Docker with packaged compiled MATLAB code for fastICA together with MATLAB Runtime environment, which can be readily used in other applications and does not require MATLAB installed [25]. An example of computational time needed for the analysis of two transcriptomic datasets of typical size (full transcriptome, from 200 to 1000 samples) is provided in Additional file 6: Figure SF5. As a rough estimate, it takes 3 h to analyze a transcriptomic dataset with 200 samples and 7 h to analyze a dataset with 1000 samples, using an ordinary laptop. In each such analysis, more than 2000 ICA decompositions of different orders have been made.

The algorithm for determining the most stable Transcriptome dimension (MSTD)

1) Define two numbers [*Mmin*, *Mmax*] as the minimal and maximal possible numbers of the computed components.
2) Define the number *K* of ICA runs for estimating the components stability. In all our examples, we used K = 100.
3) For each *M* between M_{min} and M_{max} (or, with some step) do

3.1) Compute *K* times the decomposition of the studied dataset into *M* independent components using the *fastICA* algorithm. This results in computation of *M*x*K* components.

3.2) Cluster *M*x*K* components into *M* clusters using agglomerative hierarchical clustering algorithm with the measure of dissimilarity equal to 1-$|r_{ij}|$, where r_{ij} is the Pearson correlation coefficient computed between components.

3.3) For each cluster C_k out of *M* clusters (C_1, C_2,..., C_N) compute the stability index using the following formula

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

where $|C_k|$ denotes the size of the *k*th cluster.

3.4) Compute the average stability index for *M* clusters:

$$S(M) = \frac{1}{M}\sum_k I_q(C_k)$$

4) Select the MSTD as the point of intersection of the two lines approximating the distribution of stability profiles (Fig. 1a). The lines are computed using a simple k-lines clustering algorithm [26] for k = 2, implemented by the authors in MATLAB, with the initial approximations of the lines matching the abscissa and the ordinate axes of the plot.
The index used in 3.3 is a widely used index of clustering quality defined as a difference between the average intra-cluster similarity and the average inter-cluster similarity. In [9] this index was introduced to estimate the quality of clustering of independent components after multiple runs with random initial conditions, and tested in application to fMRI data. In the case of clustering independent components, I_q = 1 corresponds to the case of perfect clustering of components such that all the components in one cluster are correlated with each other with $|r|$ = 1, and that all components in the same cluster are orthogonal to any other component (in the reduced and whitened space).

Comparing metagenes computed for different datasets and in different analyses

Following the methodology developed previously in [3], 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. We computed the correlation only between those genes having projection value more than 3 standard deviations in the identified component.

When comparing two sets of metagenes $\mathbf{A} = \{A_1,\ldots,A_M\}$ and $\mathbf{B} = \{B_1,\ldots,B_N\}$, in order to do component matching, we focused on the maximal correlation of a metagene from one set with all components from another set. If B_i = arg max(corr(A_j,**B**)) then B_i is called *best matched*, for A_j, metagene from the set **B**. If B_i = arg max(corr(A_j,**B**)) and A_j = arg max(corr(B_i,**A**)), then the correlation between B_i and A_j is called *reciprocal*.

In all correlation-based comparisons, the absolute value of the correlation coefficient was used.

The orientation of independent components was chosen such that the longest tail of the data projection

distribution would be on the positive side. Then, for quantifying an intersection between a metagene and a reference set of genes (e.g., cell cycle genes), simple Jaccard index was computed between the reference gene set and the set of top-contributing genes to the component, with positive weights >5.0.

Determining if a small gene set is driving an independent component

To distinguish whether an independent component is driven by a small gene set, the distribution of gene weights W_i from the component was analyzed. For each tail of the distribution (positive and negative), the tail weight was determined as the total absolute sum of weights of the genes exceeding certain threshold W^{top}. The heaviest tail of the distribution was identified as the tail with the maximum weight. For the heaviest tail and for the set of genes P with absolute weights exceeding W^{top}, sorted in descending order by absolute value, we studied the gap distribution of values $G_i = W_i/W_{i+1}$, $i \in P$. If there was a single value of G_i exceeding a threshold G^{max}, then the component was classified as being driven by a small set of genes corresponding to the indices $\{i; i \le \max(k; G_k \le G^{max})\}$. The values $W^{top} = 3.0$, $G^{max} = 1.5$ collected the maximal gene set size = 3 in all ICA decompositions. These are few genes with atypically high weights separated by a significant gap from the rest of the distribution (note that these genes cannot always be considered outliers since they and the resulting independent components can be reproducible in independent datasets).

Additional files

Additional file 1: Figure SF2. Estimating MSTD dimension for six breast cancer datasets. The notations are the same as in Fig. 1. (PDF 479 kb)

Additional file 2: Figure SF1. Standard estimations of intrinsic dimensionality (by Keiser rule or by broken stick distribution) of cancer datasets. (PDF 288 kb)

Additional file 3: Table ST2. Genes associated with ICA components of the METABRIC dataset, in the case when a component is driven by a small group of genes (frequently, one gene). Gene names marked in bold also drive independent components in several other breast cancer datasets and the corresponding components are reciprocally reproducible in terms of the correlation of the whole ICA-based metagenes. (XLSX 10 kb)

Additional file 4: Table ST1. Breast cancer transcriptomic datasets used for the analysis of component reproducibility in independent datasets. (XLSX 13 kb)

Additional file 5: Figure SF4. The histograms of the total number of reciprocal correlations in the correlation graph such as the one shown in Fig. 3, with and without applying the component stabilization approach. (PDF 164 kb)

Additional file 6: Figure SF5. Computational time for ICA decomposition of different orders from 2 to 100 with step 5, using compiled MATLAB fastICA implementation and stability analysis by re-computing fastICA from 100 various initial conditions. The computation is made using an ordinary laptop with Intel Core i7 processor and 16Gb of memory, in a single thread. The BRCA BEK dataset (from [27]) contains 10,000 genes in 197 samples, and the BRCA TCGA dataset (from [28]) contains 20,503 genes in 1095 samples. The overall timing for computing all ICA decomposition with their stability analysis is 3.0 h for BRCA BEK dataset, and 6.5 h for BRCA TCGA dataset. These computations can be repeated using BIODICA software [29] (https://github.com/LabBandSB/BIODICA), by launching ICA computation in scanning mode. (PDF 361 kb)

Additional file 7: Figure SF3. Graph of reciprocal correlations between components computed with MSTD choice for the reduced dimension and the number of components. The size of the points reflects their stability (larger points corresponds to more stable components). The color and the width of the edges reflect the Pearson correlation coefficient. Propositions of annotations of the pseudo-cliques in the graph are made based on the comparison with previously annotated metagenes [3] and the analysis of the top contributing genes using hypergeometric test and the *toppgene* web tool [30]. (PDF 315 kb)

Abbreviations

IC: Independent Component; ICA: Independent Component Analysis

Acknowledgements

We thank Dr. Anne Biton for sharing the normalized public transcriptomics data for four breast cancer datasets. We also thank Prof. Joseph H. Lee (Columbia University) for critical reading and improving the manuscript text.

Funding

This study is supported by "Analysis of cancer transcriptome data using Independent Component Analysis" project from the budget program "Creation and development of genomic medicine in Kazakhstan" (0115RKO1931) from the Ministry of Education and Science of the Republic of Kazakhstan. This work was partly supported by ITMO Cancer within the framework of the Plan Cancer 2014–2019 and convention Biologie des Systèmes N°BIO2015–01 (M5 project) and MOSAIC project.

Availability of data and materials

The results shown in this paper are in part based upon publicly available data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The provenance of the public data used in this study is indicated in the Method section and Additional file 4: Table ST1.

Authors' contribution

UK LC EB AZ designed the study and developed the methodology, UK LC AG AM UC AZ performed the computational experiments, UK LC UC AZ wrote the manuscript, all authors read, approved and edited the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Publisher's Note

Author details

[1]Institut Curie, PSL Research University, INSERM U900, Mines ParisTech, Paris, France. [2]Laboratory of bioinformatics and computational systems biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan.

Received: 16 April 2017 Accepted: 4 September 2017
Published online: 11 September 2017

References

1. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000;13(4-5):411–30.

2. Teschendorff AE, Journée M. Absil P a, Sepulchre R. Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol. 2007;3(8):e161.
3. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, et al. Independent component analysis uncovers the landscape of the bladder tumor Transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 2014;9(4):1235–45.
4. Gorban A, Kegl B, Wunch D, Zinovyev A. Principal Manifolds for Data Visualisation and Dimension Reduction. Lect notes Comput Sci Eng. 2008;58:340p.
5. Saidi SA, Holland CM, Kreil DP, MacKay DJ, Charnock-Jones DS, Print CG, et al. Independent component analysis of microarray data in the study of endometrial cancer. Oncogene. 2004;23(39):6677–83.
6. Zinovyev A, Kairov U, Karpenyuk T, Ramanculov E. Blind source separation methods for deconvolution of complex signals in cancer biology. Biochem Biophys Res Commun 2013;430(3):1182–7.
7. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004; 101(12):4164–9.
8. Bang-Berthelsen CH, Pedersen L, Fløyel T, Hagedorn PH, Gylvin T, Pociot F. Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. BMC Genomics. 2011; 12:97.
9. Himberg J, Hyvärinen A, Esposito F. Validating the independent components of neuroimaging time-series via clustering and visualization. NeuroImage 2004;22(3):1214–22.
10. Li Y-O, Adali T, Calhoun VD. Estimating the number of independent components for functional magnetic resonance imaging data. Hum Brain Mapp. 2007;28(11):1251–66.
11. Hui M, Li R, Chen K, Jin Z, Yao L, Long Z. Improved estimation of the number of independent components for functional magnetic resonance data by a whitening filter. IEEE J Biomed Heal Informatics. 2013;17(3): 629–41.
12. Majeed W, Avison MJ. Robust data driven model order estimation for independent component analysis of fMRI data with low contrast to noise. PLoS One. 2014;9(4):e94943.
13. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. Biol Direct. 2007;2:2.
14. Kégl B. Intrinsic dimension estimation using packing numbers. Symp. A Q. J Mod Foreign Lit. 2003;15:681–8.
15. Bro R, Kjeldahl K, Smilde AK, Kiers HA. Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem. 2008;390(5): 1241–51.
16. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. J Proteome Res. 2012;11:4120–31.
17. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20.
18. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–10.
19. Giotti B, Joshi A, Freeman TC. Meta-analysis reveals conserved cell cycle transcriptional network across multiple human cell types. BMC Genomics. 2017;18(1):30.
20. Heng TSP, Painter MW, Consortium IGP. The immunological genome project: networks of gene expression in immune cells. Nat Immunol. 2008; 9(10):1091–4.
21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
22. Dhivya P, Harris L. Circulating Tumor Markers for Breast Cancer Management. Mol. Pathol. Breast Cancer. Springer International Publishing; 2016. p. 207–18.
23. Himberg J. Hyvärinen A. ICASSO: Software for investigating the reliability of ICA estimates by clustering and visualization. Neural Networks Signal Process. - Proc. IEEE Work; 2003. p. 259–68
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2017. https://www.R-project.org/.
25. BIODICA docker web-page [Internet]. 2017. Available from: https://hub.docker.com/r/auranic/biodica/
26. Agarwal S, Lim J, Zelnik-Manor L, Perona P, Kriegman D, Belongie S. Beyond pairwise clustering. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2005. p. 838–45.
27. Bekhouche I, Finetti P, Adelaide J, Ferrari A, Tarpin C, Charafe-Jauffret E, et al. High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. PLoS One. 2011;6(2): e16950.
28. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
29. Kairov U, Zinovyev A, Kalykhbergenov Y, Molkenov A. BIODICA GitHub page [Internet]. 2017. Available from: https://github.com/LabBandSB/BIODICA/.
30. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37:W305–11.

2.4 Cell-type deconvolution models

(families of approaches)

2.4.1 basis matrix

2.4.2 regression algorithm

2.4.3 others

2.5 Short review of most popular cell-type deconvolution tools

tools will already be mentioned in the section above. However a comment on other than mentionned aspects are needed

Chapter 3

Study of sensitivity of known methods

3.1 Reproducibility of NMF versus ICA

3.2 Impact of modification of signatures list on result for signature-based deconvolution methods

Chapter 4

Deconvolution of transcriptomes and methylomes

We describe our methods in this chapter.

4.1 From blind deconvolution to cell-type quantification: general overview

4.1.1 The ICA-based deconvolution of Transcriptomes

4.1.2 Interpretation of Independent components

4.1.2.1 Correlation based identification of confunding factors

4.1.2.2 Identification of immune cell types with enrichment test

4.1.3 Transforming metagenes into signature matrix

4.1.4 Regression-based estimation of cell-type proportions

4.2 *DeconICA* R package for ICA-based deconvolution

Chapter 5

Comparative analysis of cancer immune infiltration

This chapter will include biological interpretation of Pan-cancer analysis with DeconICA

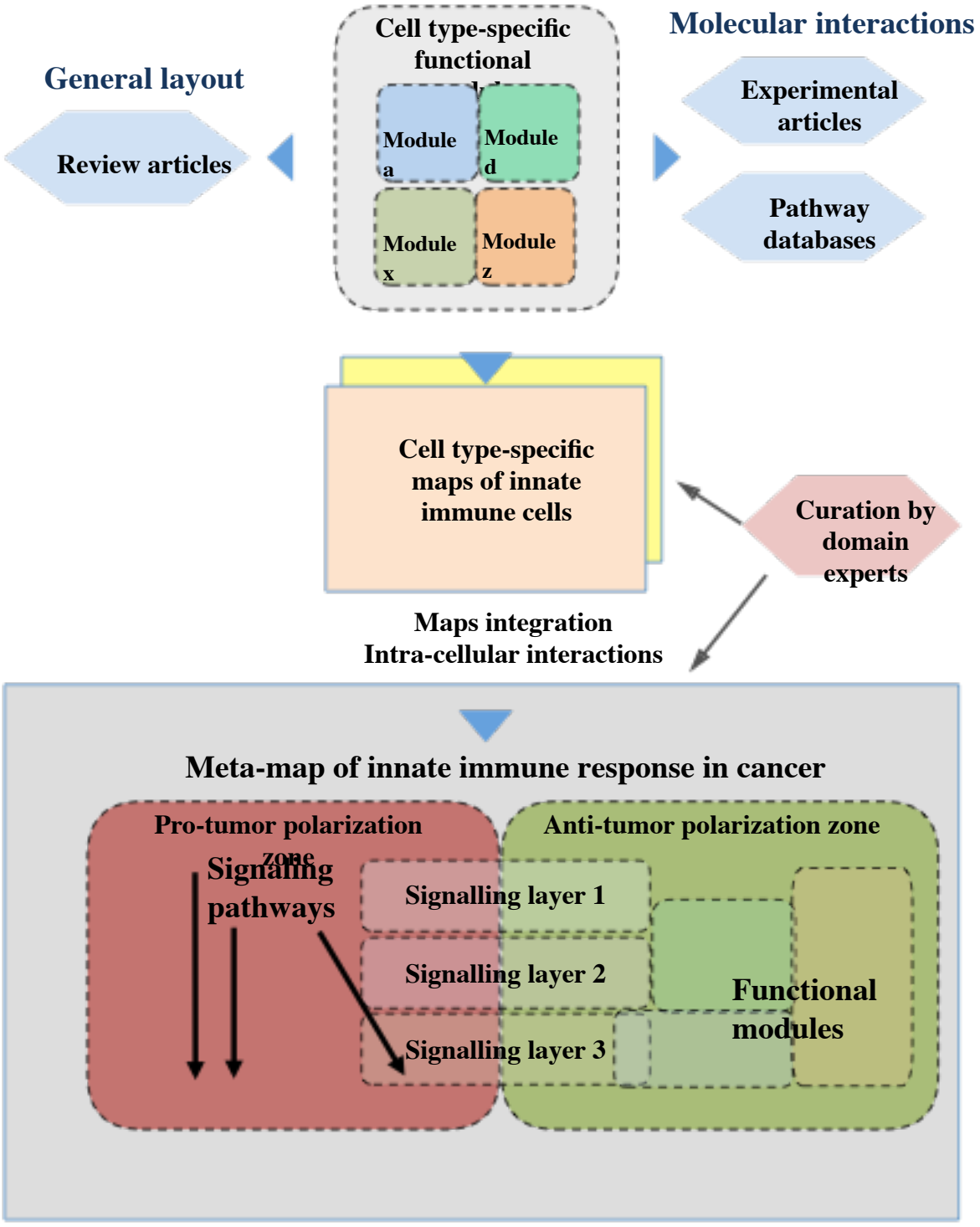
5.1

Chapter 6

Heterogeneity of immune cell types

Adapted from *submitted* article of Kondratova et al.

Figure 1


General layout
Review articles
Cell type-specific functional
Module a
Module d
Module x
Module z
Molecular interactions
Experimental articles
Pathway databases
Cell type-specific maps of innate immune cells
Curation by domain experts
Maps integration
Intra-cellular interactions
Meta-map of innate immune response in cancer
Pro-tumor polarization zone
Anti-tumor polarization zone
Signaling pathways
Signalling layer 1
Signalling layer 2
Signalling layer 3
Functional modules

Annexes

PhD timeline for defence before the end of October 2018 (Fig. 6.1)

In order to defend before 31 October, I need to follow the guidelines of the University.

- ~29 June - officially submlit the jury proposal and a draft of the thesis to the university
- ~end of July - send manuscript to reviewers
- 24 Septemeber - 31 October - defend

Note: This annex will not be a part of final manuscript

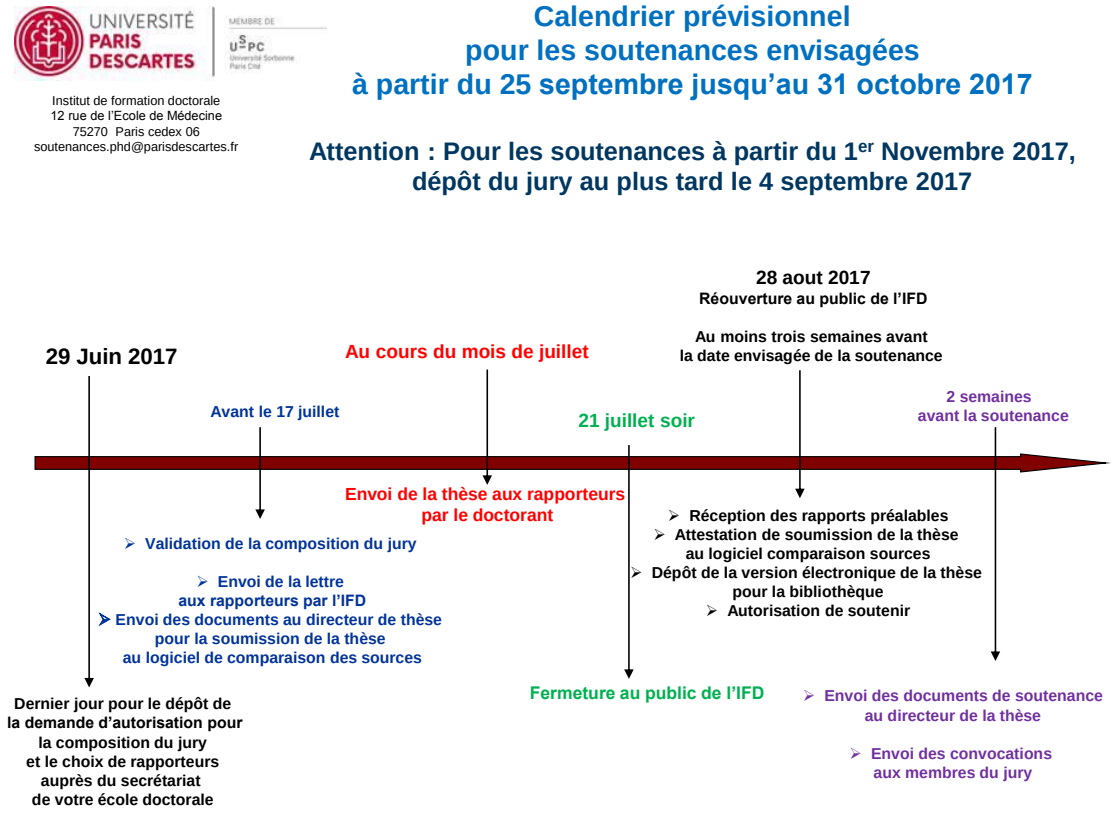


Figure 6.1: Timeline privided by University Paris Descartes for 2017

Bibliography

of Health, U. D. and Services, H. (2015). Fda approves yervoy to reduce the risk of melanoma returning after surgery.

of Health, U. D. and Services, H. (2016a). Fda approves new, targeted treatment for bladder cancer.

of Health, U. D. and Services, H. (2016b). Pembrolizumab (keytruda) checkpoint inhibitor.

of Health, U. D. and Services, H. (2017a). Fda approval brings first gene therapy to the united states.

of Health, U. D. and Services, H. (2017b). Fda approves car-t cell therapy to treat adults with certain types of large b-cell lymphoma.