Unsupervised hierarchical deconvolution of gene expression data to unravel the tumor micro-environment complexity

Nicolas Sompairac

Université de Paris

ED474 Frontières de l'Innovation en Recherche et Education

Unité 900 - Cancer et génome : bioinformatique, biostatistiques et épidémiologie

Unsupervised hierarchical deconvolution of gene expression data to unravel the tumor micro-environment complexity

par Nicolas Sompairac

Thèse de doctorat: Gène, Omiques, Bioinformatique et Biologie des Systèmes

Dirigée par Andrei Zinovyev et Inna Kuperstein

Présentée et soutenue publiquement le 10 Décembre 2021

Devant un jury composé de:

Andrei ZINOVYEV	directeur de thèse - Institut Curie
Inna KUPERSTEIN	co-directrice de thèse - Institut Curie
Elana FERTIG	rapportrice - John Hopkins University
Tatiana POPOVA	rapportrice - Institut Curie
Aurélien DE REYNIÈS	examinateur - Université de Paris
Wolfram LIEBERMEISTER	examinateur - Université Paris Saclay
Anna NIARAKIS	examinatrice - Université d'Evry

Titre (Français): Deconvolution hiérarchique non supervisée appliquée aux données d'expression génique pour élucider la complexité du micro-environment tumoral

Résumé: Les tumeurs solides sont caractérisées par une organisation complexe de l'écosystème dans lequel les cellules tumorales résident et se développent, appelé le Micro Environment Tumoral (TME). Ce TME est la cible privilégiée de l'immunothérapie qui cible à impacter de manière critique la croissance d'une tumeur ou son potentiel invasif et métastatique. De ce fait, caractériser le contenu et l'état du TME d'un patient atteint du cancer et une priorité. Cependant, dû à la large variabilité du TME et de sa complexité cellulaire et moléculaire, il est parfois difficile d'exploiter les connaissances pré-existantes sur les propriétés de ses continuants, souvent obtenues dans des contextes différents. Pour cette raison, il devient intéressant de tirer profit des approches non supervisées ou exploratoires en se basant sur les données de cancer disponibles qui ne requièrent pas de fixer une forte connaissance a priori par avance. Les outils mathématiques de machine learning comme les différentes catégories de méthodes de factorisation matricielle ont démontré leur utilité dans ce but. Dans mon travail, c'est au travers de l'utilisation d'une méthode de factorisation matricielle nommée Analyse par Composantes Indépendantes (ICA) que j'ai développé une méthode computationnelle visant à disséquer l'expression des gènes et d'autres types de données omiques, ainsi que pour extraire les signaux liés à l'infiltration immunitaire dans le TME. L'ICA récupère les sources indépendantes venant de la variation d'expression des gènes sous la forme de poids associés à tous les gènes mesurés. Mais même si cette méthode à prouvé son efficacité pour la tache de déconvolution computationnelle ainsi que d'autres applications sur des données du cancer, dû à sa nature non supervisée, elle comporte certaines complications lorsque vient le besoin de sélectionner le nombre de signaux que nous attendons dans les données ou lorsqu'on veux interpréter ces signaux. Pour soulager ce problème de choisir une dimension spécifique pour la décomposition des données, une nouvelle méthode HACK (Hierarchical Analysis of Component linKs) a été développée pour permettre d'analyser les signaux sur un assortiment de plusieurs dimensions en tant qu'une hiérarchie interconnectée ainsi que de caractériser le transcriptome comme un groupe de métagènes persistants, reproductibles sur plusieurs ordres de décomposition. Cette approche permet non seulement d'avoir une idée sur la qualité et la reproductibilité des signaux récupérés mais aussi d'aider à reconstruire les relations parmi eux. Pour l'interprétation des signaux extraits, je proposes d'exploiter les reconstructions complètes des voies de signalisation pour tirer des conclusions sur le sens biologique des signatures moléculaires dérivées des données. Par conséquences, dans ce projet j'ai participé à la production et l'exploitation de plusieurs cartes moléculaires détaillées reliées à la biologie du cancer comme la carte du rôle du système immunitaire inné dans le cancer ou la carte sur la régulation de la mort cellulaire. En définitive, c'est au travers l'utilisation d'analyses de données non supervisées, couplées à une description détaillées des interactions moléculaires que nous pouvons commencer à démêler la complexité du TME, d'une manière complémentaire aux autres méthodes.

Mots clefs: Micro-environment tumoral, Factorisation matricielle, Données d'expression de cancer, Deconvolution non-supervisée, Analyse par Composantes Indépendantes, Voies de signalisation, Biologie des systèmes

Title (English): Unsupervised hierarchical deconvolution of gene expression data to unravel the tumor micro-environment complexity

Abstract: Solid tumours are characterised by a complex organisation of the cellular ecosystem, in which the tumor cells reside and progress, called tumor microenvironment (TME). This TME is the primary target of immunotherapy that aims to critically impact tumour growth or its invasive and metastatic potential. Thus, characterising a cancer patient's TME content and state becomes a priority. However, due to the large variability of the TME and its cellular and molecular complexity, it is sometimes difficult to exploit pre-existing knowledge about the properties of its constituents, frequently obtained in unrelated contexts. For this reason, it appears interesting to take advantage of unsupervised or exploratory approaches based on available cancer data that don't require strongly fixed a priori knowledge. Mathematical tools from machine learning such as various flavours of matrix factorisation showed to be helpful for this purpose. In our work, through the use of a matrix factorisation method named Independent Component Analysis (ICA) we developed a computational methodology aiming at dissecting the gene expression and other types of omics data and extracting signals related to the immune infiltration into TME. ICA retrieves the independent sources of gene expression variation in the form of weights associated with all measured genes. Although this method proved to be efficient for the computational deconvolution task and in other applications, due to its unsupervised nature, it bears some complications when it comes to the need to select the number of signals we expect to be in the data or when we want to interpret those signals. To alleviate the problem of choosing a specific dimension for the data decomposition, a novel method HACK (Hierarchical Analysis of Component linKs) was developed that allows us to analyse signals over a range of multiple reduced data dimensionalities as an interconnected hierarchy and characterise the transcriptome as a set of persistent metagenes, reproducible across multiple decomposition orders. This approach provides not only an idea of the quality and reproducibility of the retrieved signals but also can help reconstruct relations between them. For the interpretation of the extracted signals, I suggest exploiting comprehensive signalling pathway reconstructions to draw conclusions on the biological meaning of the molecular signatures derived from the data. Therefore, in this project I participated in producing and exploiting several comprehensive molecular maps related to cancer biology such as the map of the role of innate immunity in cancer and the map of the regulated cell death. In definitive, it is through the use of unsupervised data analyses coupled with a detailed description of molecular interactions that we can approach the unraveling of the complexity of the TME, in a way complementary to other methods.

Keywords: Tumor micro-environement, Matrix factorisation, Cancer expression data, Unsupervised deconvolution, Independent Component Analysis, Signalling pathway, Systems biology

Remerciements

Une thèse est peut être quelque chose que l'on défend seul, mais c'est loin d'être un travail solitaire, bien heureusement. C'est avec le soutien de collègues, amis et familles qui, tout au long de ces 3 années, aident à progresser dans le bon sens et garder courage dans les moments de doute. Je voudrais donc adresser mes remerciements aussi bien professionnels que personnels à tous ceux, sans qui, cette thèse n'aurait pas été la même ou n'aurait peut être juste pas été.

Mes remerciements vont tout d'abord aux membres du jury, Wolfram Liebermeister, Aurélien De Reyniès, Anna Niarakis et en particulier aux membres rapporteurs Tatiana Popova et Elana Fertig, d'avoir accepté d'évaluer mon travail, ainsi que pour leur questions et remarques enrichissantes.

Merci également aux membres du comité de suivi, Vassili Soumelis et Tatiana Popova pour avoir suivi ce travail et su m'orienter avec leurs remarques pertinentes.

Je voudrais remercier ensuite Andrei Zinovyev et Inna Kuperstein pour avoir supervisé cette thèse et pour m'avoir accompagné et apporté leur soutien lorsque j'en avais besoin. Merci à Inna pour m'avoir aidé et pour m'avoir permis de découvrir de nombreux sujets de recherche qui m'ont donné une vision plus large de mon travail. Merci en particulier à Andrei pour toujours avoir trouvé le temps de m'aider, me guider, m'encourager et me soutenir tout au long de ces années, notamment pendant ces deux dernières années compliquées.

Mes remerciements suivants vont aux membres de l'équipe de biologie des systèmes, avec qui j'ai collaboré ou juste passé des moments d'échanges chaleureux et intéressants. Merci à Maria Kondratova pour avoir partagé ses connaissances poussées en biologie qui m'ont aidé à mieux comprendre mes analyses. Merci à Jane Merlevede avec qui j'ai collaboré plus étroitement et sur qui je pouvais toujours compter si je me sentais perdu dans mon travail. Merci à Vincent Noel pour ses astuces et son savoir de programmation, ainsi que pour sa bonne humeur constante et ses nombreuses anecdotes. Merci à Urszula Czerwinska pour m'avoir grandement aidé à prendre mes bases sur le sujet de l'ICA de ma thèse et merci à Nicolas Captier pour avoir éclairci de nombreux points de mathématiques.
Merci plus largement à tous les membres de l'équipe qui ont fait de cette thèse une période plus qu'agréable: Arnau, Cristobal, Mihaly, Luca, Loic Chadoutaud et Loic Verlingue, Jonas, Laura, Loredana, Laurence, Marco, Andrea, Alexander, Aziz, Christine, Marianyela, Jonathan.

Merci aussi aux membres de la plateforme, Julien, Choumouss, Laetitia et Henri pour tous les moments passés ensemble en dehors du travail qui m'ont toujours mis de bonne humeur.

Merci à Caroline, Kati, Yasmina, Eugenia, Camille et Elodie pour leur disponibilité et leur dynamisme et qui ont toujours été là pour répondre à mes questions et qui m'ont facilité de nombreux aspects organisationnels.

Merci aux autres membres de l'U900 avec qui j'ai eu des échanges plus brefs mais toujours plaisants.

Même s’il n’a pas été présent lors de ma thèse, je voudrais remercier Joël Pottier, qui a été le meilleur professeur que j’ai eu la chance et le plaisir d’avoir lors de mes années universitaires. C’est grâce à ses enseignements et nos discussions que j’ai acquis un intérêt pour les mathématiques et l’informatique et, dans un sens plus large, la recherche en science. Et c’est grâce à vous que je me suis lancé dans cette voie avec une réelle envie et une grande curiosité.

Je me dois d’exprimer la plus grande partie de ma gratitude à ma famille. Merci tout spécialement à mes parents pour m’avoir soutenu et conforté dans les moments difficiles. Merci de ne jamais avoir douté de mes choix et de m’avoir encouragé tout au long de ma voie.

Merci également à ma seconde famille, Sylvie, Emmanuel, Alice et Flore pour la curiosité qu’ils ont porté à mon travail et leur

Et enfin, merci Elise, merci pour tous ces moments où tu me mettais de bonne humeur lorsque j'étais triste ou fatigué, en étant la personne adorable que tu es. Merci du fond d’un cœur qui t’appartient…

Content

List of Figures

List of Tables

I - Introduction

1. Cancer Systems Biology applied to immunotherapy

Before tackling any problem related to cancer, it is important to first know what is really meant and understood about it. For this purpose, this chapter will briefly introduce the historical understanding of what scientists thought cancer was and what the community of researchers came to understand it now. It will then describe the existence of its cellular environment called the Tumoral Micro-Environment (TME) and its importance and participation in cancer development, progression or regression. It will touch upon the use of our understanding about TME for curing therapies. It will also describe the different types of data used to better understand the TME content and state. Lastly, it will describe different bioinformatics approaches used to extract this needed information from data and interpret it biologically.

1.1 Cancer disease, Tumor microenvironment and Immunotherapy

1.1.1 A brief overview of historical evolution in cancer understanding

Cancer can be seen as a group of cells with uncontrolled growth and division that can spread to other parts of the body via a phenomenon called metastasis. Cancer can be of many types and can appear almost anywhere in the body. Animals have had cancer throughout recorded history and the oldest evidence of cancer in mammals goes as far as the Cretaceous era where tumors were found in vertebrae of Hadrosaurs (Rothschild et al., 2003) which lived approximately 70 million years ago. The oldest human description of cancer was found in ancient Egyptian papyri written around 1.600 B.C. and may probably be based on even older material from 2.500 B.C. The papyrus in question contains a description of a bulging tumor in the breast with a mention of possible spread and classification as a case with no possible treatment (Breasted, 1991).

Different kinds of cancer have also been described by Hippocrates (460-360 B.C) who referred to them for the first time as ***karkinos***, a term that is know now as carcinoma (Hajdu, 2011). This Greek word for crab was used to describe the appearance of solid malignant tumors that looked like the stretched feet of a crab because of the veins spreading on all sides. As a treatment, Hippocrates recommended different diets and if this approach didn't work, the summarised possibilities are, as mentioned in his Aphorism 165-6:

> *"That which medicine does not heal, the knife frequently heals; and what the knife does not heal, cautery often heals; but when all these fail, the disease is incurable."*

Aulus Cornelius Celsus (25 B.C. - 50 A.D.), a Roman physician could be considered as a Hippocrates successor. He described the evolution of tumors as two states: ***cacoethes*** for tumors capable of being removed surgically and ***carcinomas*** for later stages that should be left alone or else they would cause death of the patient. In his writing Celsus mentioned these different stages and pointed out their implications for the patient:

> *"It is only the cacoethes which can be removed; the other stages are irritated by treatment; and the more so the more vigorous it is. Some have used caustic medicaments, some the cautery, some excision with a scalpel; but no medicament has ever given relief; the parts cauterized are excited immediately and increase until they cause death.*
> *No one, however, except by time and experiment, can have the skill to distinguish a cacoethes which admits of being treated from a carcinoma which does not."*

Cancer observations has continued to be expanded by physicians across the centuries with a constant refining of diagnosis techniques and treatments of surgical or chemical nature (Faguet, 2015). But it is really with the advent of molecular biology, helping to bridge the gap between biochemistry and genetics, that the cancer mechanisms could finally be described in details. The discovery of the DNA structure by Wilkins, Franklin, Watson, Crick and Pauling (Klug, 2004) allowed the establishment and the beginning of the molecular biology of cancer.

1.1.2 Cancer as a complex disease coming from a deregulated machinery

In 1958, Crick formulated what is known now as the "central dogma of molecular biology" (Crick, 1970). The central dogma is an explanation of the genetic information flow in a biological system. It states that the information contained in DNA has to sequentially pass through RNA and finally a protein. However, once it has passed this final point, it cannot get out again (Figure 1.1).

A simple way to understand this is to consider how cellular functions are encoded and expressed in a simple cell. A function is encoded as a genetic information (gene) stored in the form of DNA, contained in the nucleus present in all cells. However, functions have to happen outside of the nucleus so the information contained in the DNA has to be transcribed into another form of nucleic acid, the RNA. Once outside the nucleus, with the help of ribosomes that serve as cell factories, the information is extracted from the RNA and converted into a chain of amino-acids, the building blocks of proteins. Once in this protein form, cellular functions can finally be expressed. These processes are constantly regulated in the cell by either gene expression being turned off and on or via more refined events such as modifications of the RNA by splicing or modifications of proteins by some post-translation alterations of chemical nature.

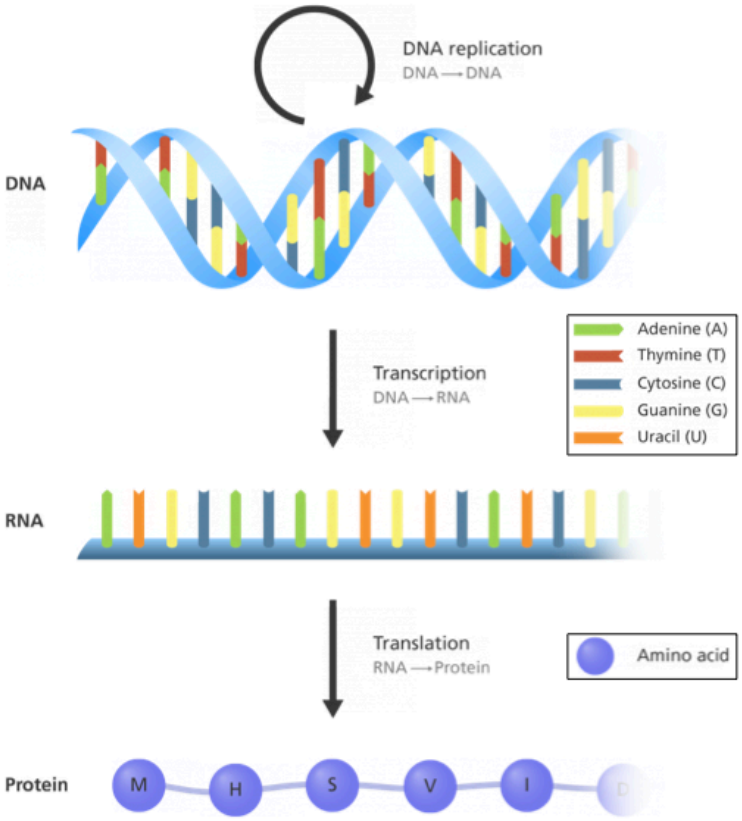


Figure 1.1. Schematic illustration of the central dogma of molecular biology.
The genetic information flow contained in the DNA is transferred sequentially through RNA to Proteins. Adapted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from https://www.yourgenome.org/ (Image credit Genome Research Limited).

It is when this machinery becomes deregulated, that cells can acquire a cancerous state. A gene which when altered can give rise to cancer is called an oncogene. Oncogenes override normal regulatory controls and induce a cancer state of a healthy cell. Many studies have been performed to find the possible actors implicated in the causes of cancer (Bos, 1989). These oncogenes can have many different functional properties and can be caused by different phenomena, as listed in (Croce, 2008).

The possible alterations giving rise to an oncogene are the following:

- **Chromosomal rearrangement** is a changing in the structure of the native chromosome. This type of event is caused by a break of the DNA helix at two locations followed by the rejoining of the broken ends, thus leading to a new chromosomal arrangement.
- **Mutations** are modifications of the nucleic acid content in the DNA strand. A modification of the DNA sequence leads to a different structure of the encoded protein, changing its functional properties.
- **Gene amplifications** is an increase in the number of gene copies. It can occur as a defect prior to a tumor appearance but more often happens during tumor progression (Albertson, 2006). An increase in gene copies leads to an increase in the protein level that can deregulate the balanced functions of a cell which can for instance increase the rate of cell divisions.

Once activated, oncogenes can give birth to a panel of products that can be classified into six groups:

- **Transcription factors** participate in the regulation of DNA transcription by binding to specific sequences in enhancer or promotor regions. When a transcription factor is abnormal, it can increase the expression of several genes. For example, the AP1 transcription factor controls cell division and can become abhorrent in cancer cells (Shaulian and Karin, 2001) leading to an uncontrolled cellular growth.
- **Chromatin remodellers** change the compaction level of chromatin and consequently control gene expression, replication and the repair of chromosome segregations. If such elements become mutated, they can harm the normal processes of DNA repair, leading to more anomalies with time.
- **Growth factors** stimulate cells growth, participate in the control of certain cell cycle phases and sometimes play a role in wound healing. It can be easily seen that an abnormal activity of growth factors can provoke an increased cell proliferation.
- **Growth factor receptors** are proteins located on the cell membrane which when activated control various cellular activities such as cell division, survival, differentiation and migration. Modified receptors can become active even in the absence of the necessary ligand and lead to cells escaping their normal fate.
- **Signal transducers** participate in the exchange of molecular signals between cells and their environment. Mutated transducers can become either fully activated or no longer sensitive to the signal. These changes can induce modifications in the cells themselves by activating downstream pathways but also impact their environment (Juliano, 2020).
- **Apoptosis regulators** help control the “cell suicide mechanisms” and are active in cases where a cell has heavy DNA damage or has an abnormal metabolic behaviour. They can also be activated in cases of cell damage or through the unfavourable state of the environment surrounding the cell. In cases where these regulators malfunction, a cell can escape the regulated death controls and continue to live and multiply while accumulating other oncogenic functions.

Despite all these possible anomalies that a cell can be facing with, protection mechanisms remain to help getting rid of abnormal cells. The cellular environment can force an abnormal cell to pass through apoptosis. Some mutations can lead to the appearance of oncogenes but some mutations can be too deleterious and lead to an incapacity of a cell to grow further. The mechanism of telomeres shortening with each cell division can also help passively control the rate of uncontrolled cell divisions. Therefore, for a cell to become cancerous, it would have to find a way to avoid all these protection mechanisms. It is only after it was able to acquire a defence against them that a cancer cell can truly drift from the normal cell machinery and progress in the organism.

A seminal article listed the properties required for a cell to be considered as cancerous (Hanahan and Weinberg, 2011), called the hallmarks of cancer, that can be seen in Figure 1.2. These hallmarks correspond to the various capabilities a cancer cell has to acquire to be able to grow and provide a solid foundation for understanding the biology of cancers. The hallmarks' essence can be summarised as follows:

- **Sustaining Proliferative Signaling**: In a normal cell, the growth happens in a controlled and balanced manner and from the deregulation of this balance, cancer cell have to deal with the unbalance that follows. The lack of elements such as growth factors can lead to a premature growth stop but to mitigate that, cancer cells can send signals to surrounding cells that may in turn provide them with the necessary elements to continue growing. Or it can also be possible that a cancer cell becomes growth factor independent by a set of mutation of its regulatory elements.
- **Evading Growth Suppressors**: As I mentioned before, natural and powerful mechanisms help to negatively regulate cell growth thanks to tumors suppressive genes that can trigger an apoptosis when deemed necessary. This control can also be performed by cell-to-cell proximity but cancer cells are known to abolish this "contact inhibition". This can be done by modifying the state of membrane proteins that normally normally involved in in the structure and cell integrity.
- **Resisting Cell Death**: The mechanisms of apoptosis is a natural barrier of tumor development. It can be triggered from sensors looking at the DNA damage state or environmental stress. By losing either these sensor genes or tumor suppressors such as TP53, cancer cell are able to circumvent the apoptotic machinery.
- **Enabling Replicative Immortality**: To enter an "immortal" state, cells have to survive the two natural barriers that are the senescence and crisis phases. It is the constant shortening of telomere ends of the DNA after each cell division that often leads to the crisis state. Immortalised cells combat this by having specialised DNA polymerases that can add telomere repeats, thus ignoring the crisis phase entirely.
- **Inducing Angiogenesis**: Because of their increased replication rate, tumor cells require much more nutrients than the surrounding environment can normally offer as well as evacuating metabolic wastes. These needs can be achieved by stimulating angiogenesis which consists in augmenting the vascularisation of nearby blood vessels. In fact, the angiogenic switch is almost always activated during tumor progression, making all surrounding vessels to sprout and sustain said progression.
- **Activating Invasion and Metastasis**: One particularity of cancer cells is their capacity to detach from their environment of origin and invade other parts of the organism. This phenomenon happens because of modified membrane proteins that no longer maintain the cancer cells attached to their surrounding and some migration factors are turned on and help tumors to leave their original territory which is normally impossible for cells of their original type.

- **Genome Instability and Mutation**: To achieve all these different characteristics the genome has to undergo a series of alterations all while evading the reparation mechanisms. The accumulation of these alterations can happen when “caretaker” genes that generally participate in the maintenance of a stable genome acquire defects. These defects can damage functions that help detecting DNA damage, repairing said damage or intercepting and inactivating mutagenic molecules.
- **Tumor-Promoting Inflammation**: Tumors can sometimes be infiltrated by cells of the immune system. However, paradoxically, instead of of destroying cancer cells, the inflammatory response has the effect of enhancing the tumor growth and progression. Such an effect can be explained by the capacity of immune cells to supply tumors with molecules such as growth, survival and angiogenic factors.
- **Reprogramming Energy Metabolism**: Because of the uncontrolled proliferation of cancer cells, tumors can sometimes end up in a normally unfavourable environment such as a very tight space with limited access to oxygen. Still, cancer cells manage to adapt by modifying the energy metabolism. One of the most famous adaptation is known as the “Warburg effect”, which corresponds to the capacity to limit the energy metabolism to glycolysis even in the presence of oxygen, leading to what could be called an “aerobic glycolysis”. Although being a counterintuitive phenomenon, the Warburg effect has been also observed in embryonic cells because of the metabolic benefits this can give, such as biomass production and redox regulation (Krisher and Prather, 2012).
- **Evading Immune Destruction**: The immune defence is supposed to constantly monitor the organism for any threats to deal with them early on. However, some tumors manage to avoid detection or can defend efficiently against immune attacks.

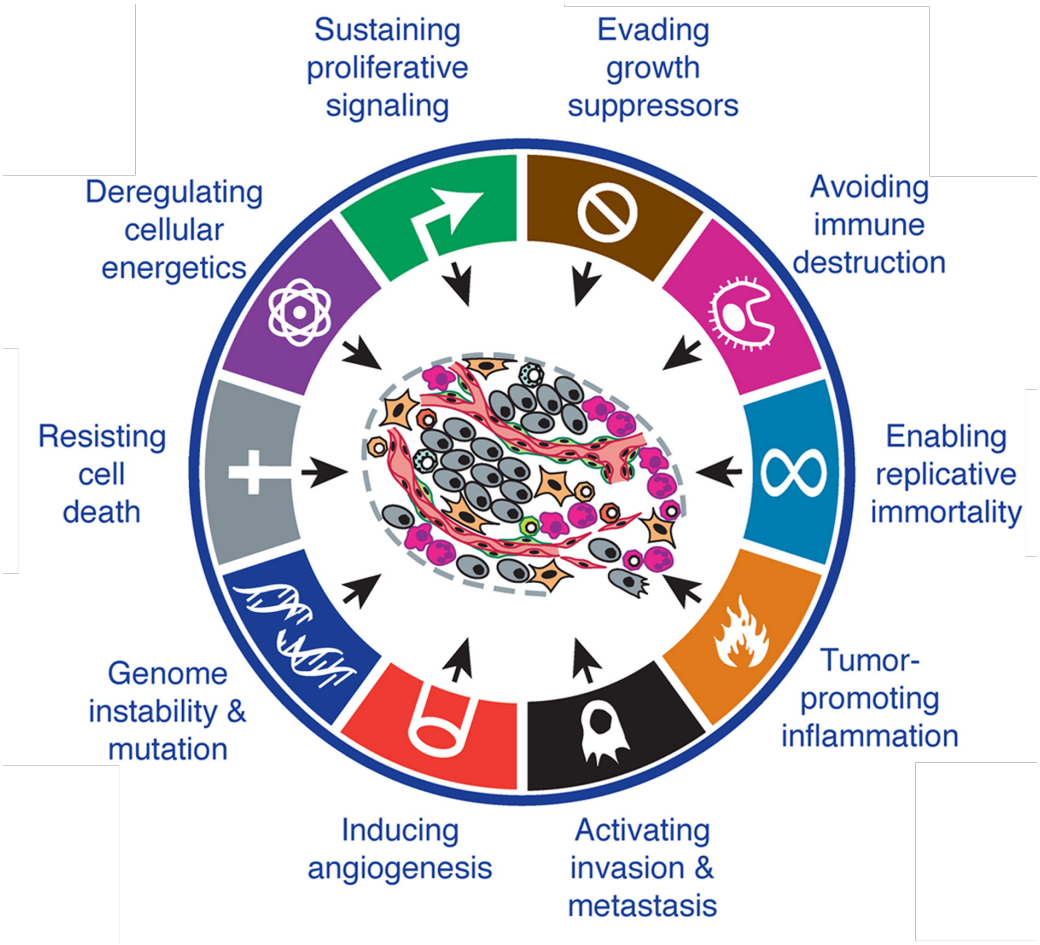


Figure 1.2. The 10 Hallmarks of cancer.
List of the 10 different capabilities acquired by cancer cells necessary for tumor growth and progression. Adapted with permissions from Cell (Hanahan and Weinberg, 2011).

Each of these hallmarks result from a series of of genetic alterations, with some being tissue specific or not (Hanahan and Weinberg, 2011). Most cancer present a combination of these characteristics giving them multiple ways of evading the protective systems put in place by the organism. But one important part always remains and that is the fact that tumors always evolve in an active environment and that this environment can be beneficial or deleterious to its growth. It is the study of this environment that gives us hope to better understand how tumors can form and how we can deal more efficiently against this. I will discuss about the implication of the tumoral environment more in detail in the next section.

1.1.3 Tumor Micro-Environment: a complex ecosystem

Solid tumors are not composed of exclusively malignant cancer cells. In reality, they are characterised by a complex organisation of the cellular ecosystem, in which the tumor cells reside and progress, called the Tumor Micro-Environment (TME). Each tumor has not only a unique combination of mutations and genomic alterations but also a unique composition of infiltrating non-tumoral cells, which depends on the tumor type, stage, host and other factors. Recent progress in cancer biology clearly showed that TME critically impacts tumor growth, invasive and metastatic potential,

as well as the response to treatment including both short-term and long-term disease outcomes (Quail and Joyce, 2013; Whiteside, 2008).

The relations between tumors and their environment are many as there is a great diversity of cells composing the TME (Balkwill et al., 2012) (Figure 1.3). Cells present in the TME can be separated in two categories:

- **The Stroma**, which is composed of the extracellular matrix, fibroblasts, endothelial and epithelial cells, adipocytes, mesenchymal stem cells and lymphatic vessels, etc.
- **The Immune cells** such as T and B lymphocytes, NK and NKT cells, Dendritic cells, Macrophages, etc.

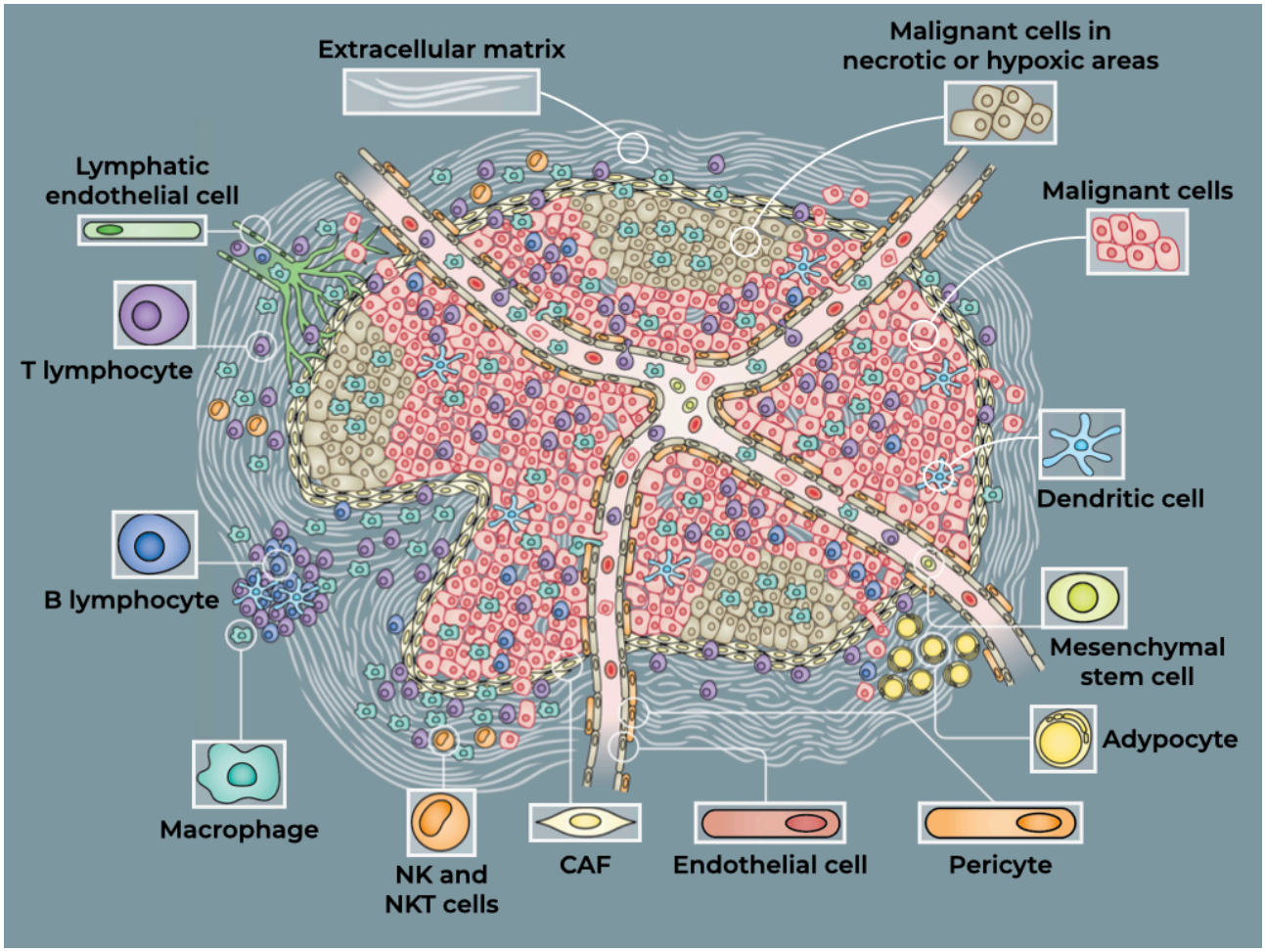


Figure 1.3. Schematic illustration of the composition of the tumor microenvironment. The TME can be seen a complex system of interacting cells in an evolving environment. Adapted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Balkwill et al., 2012).

The TME can have multiple faces in the presence of tumors:

- Positive: It can defend the organism by detecting and suppressing tumors
- Neutral: It can be oblivious to the presence of tumors and not interact
- Negative: The TME can participate in tumor growth and promote metastases.

Firstly, let's start by mentioning the positive sides.

At the first stages of cancer development, the TME plays a barrier role for the organism with the help of immune cell types. The presence of T-cells and NK cells has been proved to be correlated with good prognostic markers (Tachibana et al., 2005) as well as B-cells (Milne et al., 2009). However, since tumors and immune cells are in perpetual interaction and can remodel the functions of each other, Dunn et al. (Dunn et al., 2002) considered the term of immunosurveillance inappropriate and proposed instead the term of immuno-editing.

This immuno-editing being the result of three processes: Elimination, Equilibrium and Escape. These tree processes can be understood as the sequential order of interactions evolution between the TME and tumors.

First, the elimination process takes place where developing tumors are deleted successfully in chronological order. In the first phase, cancer cells grow and release proteins alerting immune cells (Hanahan and Folkman, 1996). Then, in phase 2, tumor deaths start to happen by apoptosis or proliferation arrest due to the release of TNF-gamma in their surrounding (Bromberg et al., 1996). This leads to the secretion of chemokines in the environment which blocks the angiogenic potential of close vessels, leading to more tumor cell deaths (Qin and Blankenstein, 2000). Next, recruited macrophages and NK cells in phase 3 can now target tumor cells (Smyth et al., 2000). Finally, in the last forth phase, DC and T cells destroy the remaining tumor with visible antigenes on their surface (Shankaran et al., 2001).

After that comes the stage of equilibrium where only surviving tumor cells are capable of functioning despite the constant immune defence. New variants continue to appear and be destroyed following the mechanism of the Red Queen hypothesis, until we reach the final stage of escape where tumors acquired a resistance to immune detection and destruction.

At this stage, patients are left with a defence system incapable of functioning correctly, leaving the existence of local tumoral spots containing remaining surviving variants that can now slowly continue to grow. This stage can be considered as the neutral moment where the TME and tumors do not interact as actively as they did anymore. This is probably the stage that explains why the TME effect on tumors have eluded researchers for so long. With such a limited amount of interactions during this period, TME was wrongly deemed as not playing an interesting part for the tumor progression.

It is only after such respite that malignant cells can fully express and cause much damage. Cells that came first to the tumor site are incapable of suppression and in certain cases can even participate in the development of cancer cells (Quail and Joyce, 2013). And example can be seen in Figure 1.4 where Tregs participate in the immune suppression and CAFs enhance angiogenesis.

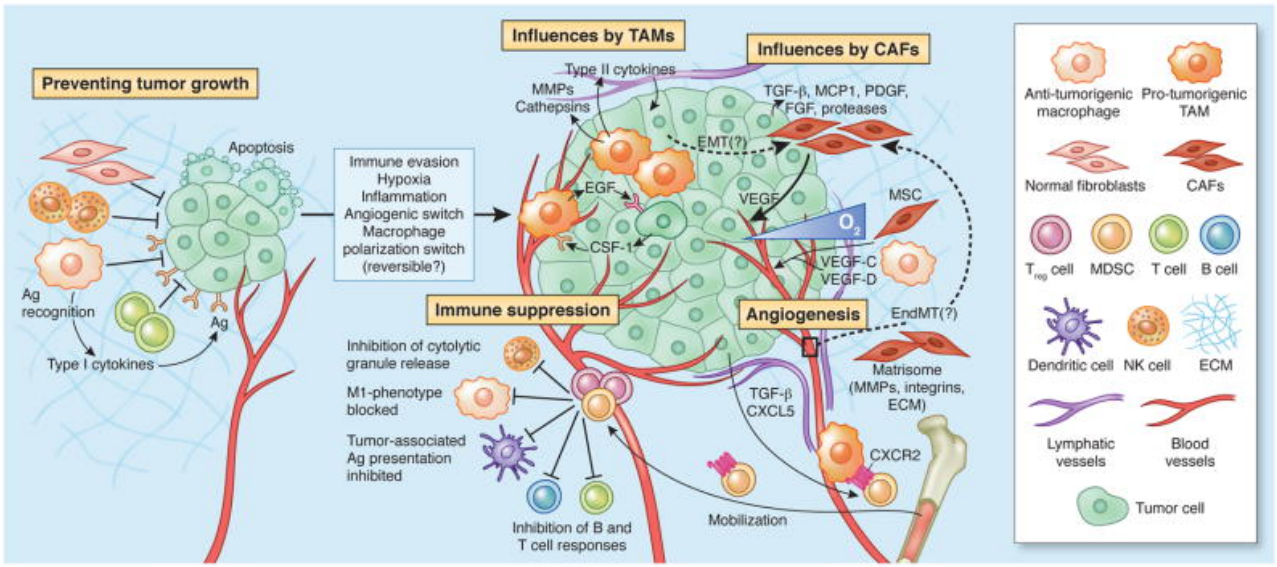


Figure 1.4. From elimination to tumor support.
In the first stages, defence cells come to destroy cancerous cells. But with time, tumors are capable to evade the immune function and can even modify immune cell properties to help angiogenesis and tumor growth. Reprinted with permissions from Springer Nature from (Quail and Joyce, 2013).

The presence and the amount of these TME cells can vastly vary between cancer types making it difficult to correctly estimate the contribution of each cell type. But even if a lot of mechanisms and interactions between cancer cells and their TME still remain unknown, I can still briefly enumerate some known cell types that promote the tumor growth:

- Cancer associated Fibroblasts (CAFs) are fibroblasts present in tumors and can help them grow in various ways. Firstly, it was observed that CAFs can secrete growth factors that can be mitogenic for malignant cells (Spaeth et al., 2009). Different types of CAFs have been identified, with some promoting metastases (Pelon et al., 2020), while others have been implicated into the formation of an immuno-suppresive environment (Erez et al., 2010).
- Dendritic cells have been shown to contribute to an immuno-suppresive environment and even promote tumor progression by becoming incapable to activate CD8 T cells, thus initiating their anti-tumor immunity (Fu and Jiang, 2018).
- Immune cells such as macrophages have been shown to promote tumor angiogenesis (Zumsteg and Christofori, 2009) while some myeloid cells (Murdoch et al., 2008) and Tregs (Campbell and Koch, 2011) have shown an immuno-suppressive activity by producing factors such as TGF-β or CTLA4.
- Lymphatic endothelial cells can help the dissemination of malignant cells (Tamella and Alitalo, 2010) but have also demonstrated an effect of altering the immune response to the tumor (Swartz and Lund, 2012).

1.1.4 Immunotherapy: using the TME to help fighting cancer

I have shown in the previous Section 1.1.3 that the tumoral environment plays a major role in the tumoral development. Interactions between cancer cells and the TME work both ways and cellular participants can shift sides and increase the tumorgenesis instead of fighting it. Many types of therapies were proposed to deal with this complex disease. Surgery was among the first, followed by chemical treatments and then radiotherapy. One way to fight against tumor proliferation without attacking cancer cells directly can be achieved by limiting angiogenesis which after all was the first characteristic observed in cancers. By limiting the growth of blood vessels, we can hamper the access to nutrients and oxygen to the cancer cells. Many anti-angiogenic therapies have been developed and approved such as anti vascular endothelial growth factors (anti-VEGF) (El-Kenawi and El-Remessy, 2013). However, such therapies represent only an indirect method of fighting tumors and can be connected with serious side effects.

Since immune cells in their original state are programmed and recruited to deal with cancer cells, it is by boosting their efficiency or by reactivating them that we can hope to deal with cancer in a more natural way with lesser side-effects. This type of therapy called Immunotherapy is aiming at stimulating the existing immune system to either target cancer cells more efficiently by triggering targeting mechanisms or by helping some parts of the machinery, blocked by cancer cell emissions, to restart. As interactions are many, targets are so as well and I will describe next some of the promising treatments and their targets.

Adoptive T cell (ATC) therapies are therapies that introduce allogenic T cells into a cancer patient. Those T-cells being derived from tumors are meant to recognise tumor associated antigenes more efficiently (Rosenberg et al., 1994). However, such approach remains limited with a high treatment cost since it requires a patient specific tailored design. It can also show adverse reactions such as inflammatory responses or in certain cases even organ damages (Brudno and Kochenderfer, 2016).

Cancer vaccines can be of two categories: prophylactic or therapeutics. Two well-known prophylactic vaccines have already been used with great success to protect against hepatitis B which leads to hepatocellular carcinoma and against papilllomavirus which can lead to the development of cervical cancer (Guo et al., 2013). Therapeutic vaccines however are based on the existence of tumor associated antigens (TAA) found on cancer cells only (Zhang et al., 2009). By sequencing tumors and finding the potential target antigenes and their coded peptides, it is possible to inject these neoantigenes as vaccines to force an increased immune response targeting the cancer cells of interest. But obviously, the most troublesome drawbacks are the time needed to develop such treatment but also the difficulty in detecting the right antigenes. As tumors have a very diverse genetic variations, it is indeed really problematic to find neoantigens that would result in the strongest antitumor response.

Immune checkpoint therapy has been proved to be a most promising approach in immunotherapy. The most potent examples have been targeting two immune checkpoint inhibitor proteins: cytotoxic T-lymphocyte 4 (CTLA4) and programmed cell death 1 (PD-1), which both play important roles for T cell activity (Fife and Bluestone, 2008). CTLA4 acts as a negative regulator of T-cell activation while PD-1 activates an immuno-suppressive pathway, diminishing the activity of T-cells. It is by releasing specific anti-bodies of these two proteins that their activity can be inhibited, allowing

T-cells to be fully active to fight against cancer cells. But it is of course expected that removing a naturally occurring immune blockade can unchain powerful immune responses going beyond the normal boundaries of immune tolerance. Because of that, auto-immune responses can occur and target different organs of the patient (Michot et al., 2016).

1.2 Omics data: existing types and information they contain

Biological systems are complex and it is difficult to capture this complexity in a straightforward way. Cancer especially is accepted now as a supreme complex disease. Its phenotypes are diverse and can only be accurately described by integrating a multitude of interconnected elements as well as understanding the environment in which cells proliferate and exchange with (Knox, 2010). It is a complex disease that can have many precursors leading a multitude of genotypes (Forbes et al., 2015). It still remains unclear which anomalies in the process of genetic information flow causes a cell to enter a cancerous state and the task is made even more difficult because of the genetic state in total disarray in cancer cells (Wishart, 2015). But one thing is sure, cancer is more than a simple genetic disease a requires us to look at different levels to fully understand its way of functioning.

The central dogma of molecular biology can be seen as a progressive transfer of a coded biological information that is transferred from DNA to a protein and passing through RNA. However, this vision can be expanded to include a step of availability of such information prior to its initiation and a step of the first manifestation of a function coming from the molecular processes of a cell (Figure 1.5).

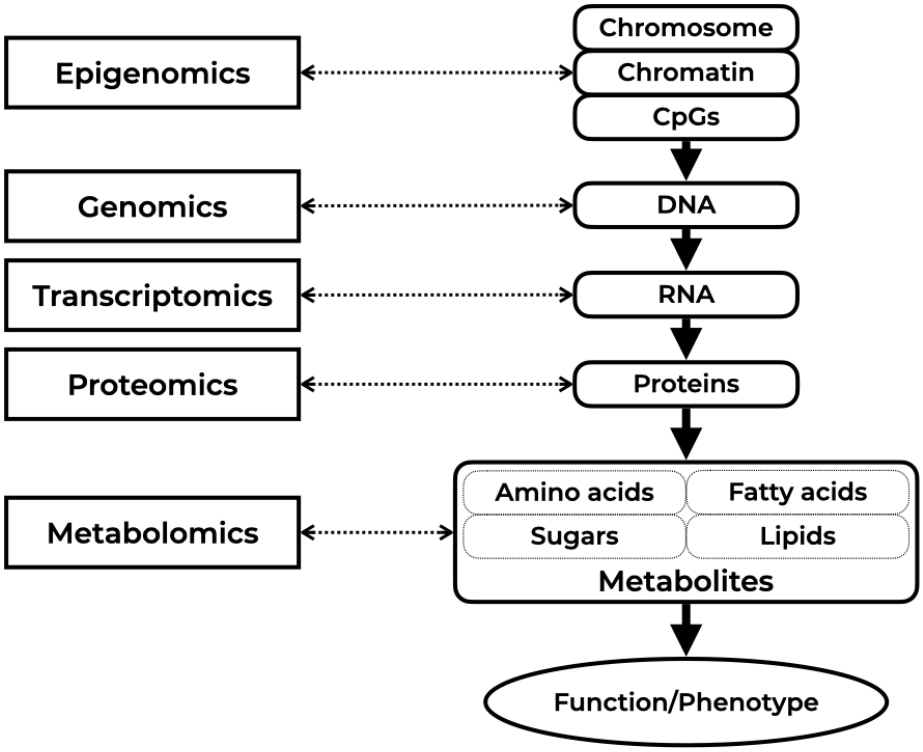


Figure 1.5. Simplified schema of genetic information flow from the chromosome to the activation of a function or observation of a phenotype.
Each step on the right is accompanied by the type of omics data that can be used to identify the corresponding molecules.

This whole system can thus be described at different levels interacting with each other. But to have a glance at these levels, this requires the use of unique technologies and a corresponding strategy to analyse and measure it. Because of this, various technologies have been developed and all gained popularity (Figure 1.6) across the years. And thanks to the emergence of systems biology, scientists realised the need to look at a problem at different angles since the whole system is often more complex that the simple sum of its parts studied independently (Barillot et al., 2020).

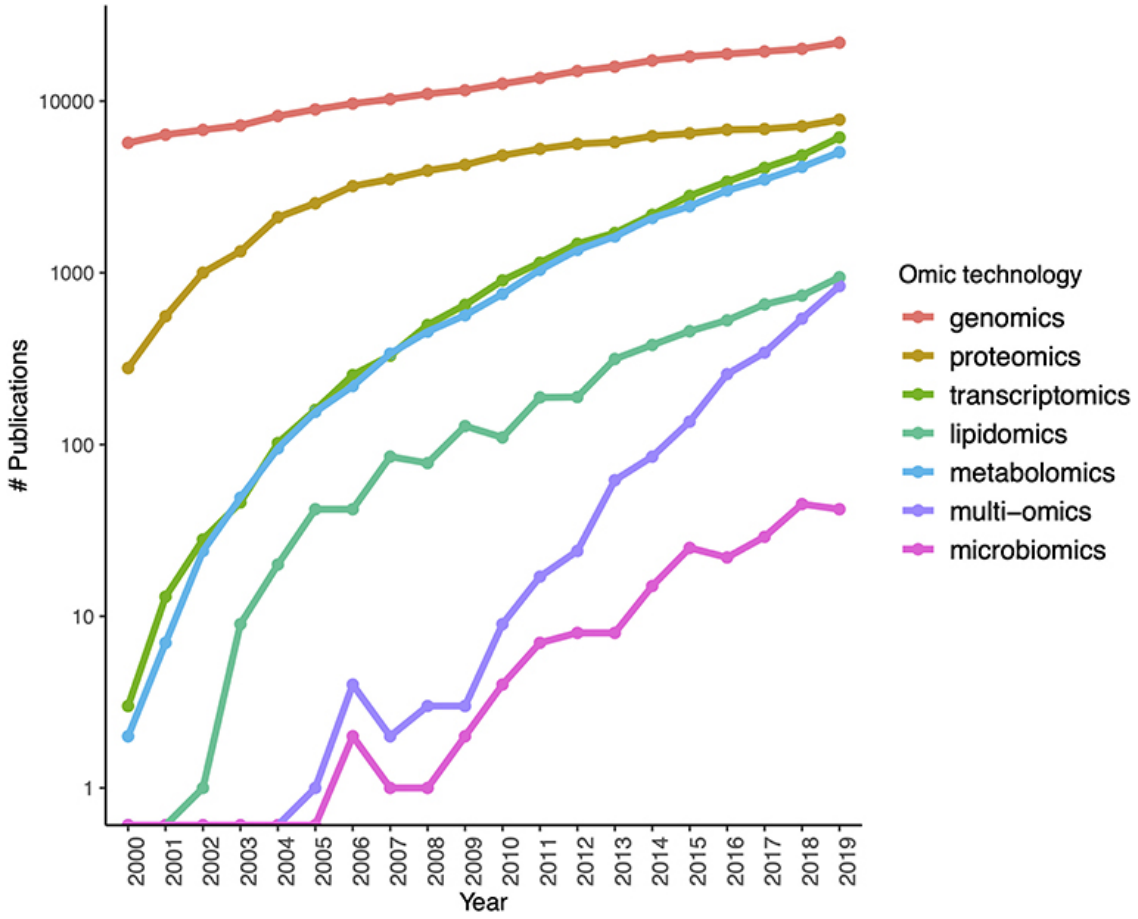


Figure 1.6. Evolution of interest in omics technologies in the 21st century.
The number of publications related to a certain type of omics was counted for publications mentioning the corresponding technology in its title or abstract. Reprinted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (de Anda-Jáuregui and Hernández-Lemus, 2020).

Most current omics are generated using technologies that are mainly sequence-based, mass-spectrometry-based or array-based. In this section, I will introduce the different types of data at our disposition and briefly discuss about the level and type of understanding they can bring us to elucidate disease mechanisms and how to detect and treat them.

1.2.1 Genomics

Genomics is one of the most mature of omics technologies and part of the most interdisciplinary field of biology. Contrary to genetics, which concern the study of a unique gene, genomics try to characterise all the genes of an organism. Genomics analyses focus on the structure, function, evolution mapping, modification and states of genomic sequences. A genomic sequence is a part of DNA, which when transcribed into RNA if modified into a messenger RNA (mRNA) and then translated into a protein that triggers a certain function in the organism. It is then possible to analyse a function by looking at the different states of the genomic sequence of origin by focusing on its variations such as insertions and deletions (INDELs) (Fan et al., 2007), Copy Number Variations (CNVs) (McCarroll and Altshuler, 2007), Single Nucleotide Polymorphism (SNPs) (Lander et al., 2001), etc. The associated technologies to retrieve the state of these sequence variations are genotype arrays (Ragoussis, 2009), exomes sequencing (Ng et al., 2009) and Next Generation Sequencing (NGS) for whole genome sequencing analyses (Koboldt et al., 2013) which became commonplace in genomic studies.

Cancer genomics helped discover major subtypes of diseases, is now an essential part of clinical analyses and treatments (Lehmann-Che et al., 2017) and the analysis of these genomic alterations gave us a great insight at how cancers functioned and developed (Stratton et al., 2009). Some studies showed that even small alterations such as INDELs have a big impact on cancer state (Ye et al., 2016) and some where even found to contribute to the immunogenic phenotype (Turajlic et al., 2017).

1.2.2 Epigenomics

Epigenomics studies focus on the characterisation of reversible chemical changes of DNA and DNA associated proteins such as histones. These changes often come in the form of CpG island methylations and histone acetylations and are responsible for structure modifications of the chromatin as well as function changes of the genome (Bernstein et al., 2007).

DNA methylation of a certain region can alter its transcriptional activity. Mutations in genes involved in epigenetic regulations have been found in multiple tumor types and some immune responses were found associated with the state of DNA-methylation (Jeschke et al., 2017). The methylation state of the DNA can be measured using the technique of Whole-Genome Bisulfite Sequencing (WGBS) (Fan and Chi, 2016).

The chromatin restructuring plays a role in the expression of certain genomic regions by changing the physical access of the DNA sequence by making it active when open or inactive because inaccessible when closed (Cairns, 2007). Chromatin remodelling deregulations have been shown to be connected to cancer (Nair and Kumar, 2012) and some conformations have allowed the identification of possible pharmacological targets in breast cancer (Baxter et al., 2018). The chromatin accessibility state can be measured with techniques such as ATAC-Seq (Buenrostro et al., 2015) and the technique of Chromatin immunoprecipitation Sequencing (ChIP-Seq) (Johnson et al., 2007) allows to measure the abundance of target proteins in a certain genomic location.

Chromosome conformation is the 3D organisation of the whole genome. Its structure allows a proximity and interaction between regions distant on the genome in term of

sequence length and even between regions on different chromosomes. Chromosomes have shown to be really plastic and implicated in cancer progressions (Jia et al., 2017). The Hi-C technique is the most up-to-date way to detect interactions between all possible genomic loci pairs (van Berkum et al., 2010).

One of the main problem of epigenomics is the limitation of sequencing types. Indeed, studies using WGBS are only possible for methylation and it is difficult to target other types of modifications. Most DNA methylation detection techniques are also very times consuming, may require a large amount of DNA and might be of low resolution by only assessing a limited number of CpD residues (Lindsey et al., 2005).

1.2.3 Transcriptomics

Transcriptomics is a *portmanteau* of the words transcripts and genomics and is therefore related to the measure of abundance of different types of RNA sequences (i.e. transcripts) in a given context. The analysis of these molecules can be quantitative (how much of each transcript is expressed) or qualitative (which transcripts are present and what are their characteristics such as their methylation state for example). Most analyses concentrate on the study of messenger RNA (mRNA) which can have a highly variable concentration compared to its resulting proteins (Wegler et al., 2020) and can also differer a lot depending on the tissue (Koch et al., 2002). This variability can come either from stochastic processes of the cell machinery or from upstream of the synthesis of mRNA (Satija and Shalek, 2014). Analysing the transcripts levels can help overcome some limitations of cancer mutation analyses and allow findings of new biomarker targets in cancer drug discoveries (Jeong et al., 2017). Technologies allowing to measure the abundance of transcripts can be either based on probe-microarrays (Duggan et al., 1999; Schulze and Downward, 2001) or RNA-sequencing (Mortazavi et al., 2008; Ozsolak and Milos, 2011; Sultan et al., 2008).

Apart from the common gene expression analysis, transcriptomics englobe also others molecules such as non-coding RNA (ncRNA), micro RNA (miRNA), long non-coding RNA (lncRNA), enhancer RNA (eRNA) and others (Jiang et al., 2015; Kaikkonen and Adelman, 2018). While these RNA transcripts do not encode proteins, they still actively participate in the cell machinery. To give an example, non-coding RNA were shown to regulate brown adipocyte development (Alvarez-Dominguez et al., 2015) and can promote cancer metastasis (Gupta et al., 2010).

There are some complications related to transcriptomics. For instance, the level of a transcript doesn't always correspond to the translation into a protein due to degradation or other post-transcriptional modifications (Gygi et al., 1999). Also, a single transcript can give rise to many different proteins due to alternative splicing or post-transcriptional modifications, making it hard to ensure the correct function of a mRNA sequence.

1.2.4 Proteomics

Proteomics studies involve the quantification and identification of the protein content in a given organism or biological system (Aslam et al., 2017). Studying proteins directly allows us to get rid of the intermediate states and helps focusing directly on the elements from which the actual functions arise. Proteomics have been used intensively to investigate and identify biomarkers and therapeutic targets in cancer (Yakkioui et al., 2017). It has also shown promising results in breast cancer classification (Tyanova et al., 2016; Yanovich et al., 2018).

On top of identifying the protein content, a set of studies focus on the different states of proteins that govern their function. These states includes post-translation modifications such as proteolysis, glycosylation, methylation, acetylation, phosphorylation, nitrosylation, oxidation and ubiquitination (Mann and Jensen, 2003).

The protein content can be measured via techniques like microarrays (Sutandy et al., 2013) but the biggest breakthrough has been achieved through the use of Mass Spectrometry (MS) (Domon and Aebersold, 2006; Macklin et al., 2020).

There are however certain limitations to proteomics. As mentioned above, many proteins experience post-translational modification that affect their activity. It is difficult to get the exact state of these modification since they are in constant evolution and can switch states as time progresses. Because of this, we face problems of reproducibility when comparing or integrating data obtained from different studies. It is also important to point out that most of proteomic studies are limited to available antibodies, thus restricting the target of a study.

1.2.5 Metabolomics

Metabolomics data quantifies the number of small molecules resulting from metabolic processes. The non-exhaustive list of these molecules contains fatty acids, amino-acids sugars, lipids and other metabolic products (Silva et al., 2019). Even more than proteomics, metabolomics can be seen as direct "functional readout of the physiological state of an organism" (Hollywood et al., 2006). And just as proteomics, metabolomics analysis techniques take advantage of Mass Spectrometry. Alterations of the metabolism can contribute to the development of cancer (Vazquez et al., 2016) as well as being correlated with the proliferation of breast cancer cells (Jerby et al., 2012). Moreover, metabolomics paired with modelling has been used to study metabolite fluxes that can help diagnose certain diseases (Heirendt et al., 2019).

With the capacity to generate high throughput profiles, the amount of data becomes overwhelming and it is important to keep in mind certain challenges of metabolomics analyses. One problem inherent to biological systems is the fact that the metabolome is sensitive to various genetic and environment stimuli (Johnson and Gonzalez, 2012). The other limitation is technical and related to liquid crystallography (LC)-MS-based metabolomics, making it difficult to assign an identity to biomarkers (Meier et al., 2017).

1.2.6 Multi-omics

On its own, each of the previously described omics data can be used to extract signatures and markers of a disease states and processes. They can be analysed independently and the results from different types of omics can then be compared between each other to see if the observations are validated across the different levels. Still, Hasin et al. correctly pointed that the *"analysis of only one data is limited to correlations, mostly reflecting reactive processes rather than causative ones"* (Hasin et al., 2017). Hence, a multi-omics approach that makes use of more than one biomolecular technique have emerged.

In the case of a complex pathology such as cancer, researchers are increasingly adopting a systems biology approach by combining these multi-levelled analyses that contribute to the creation of development a malignant state (Du and Elemento, 2015). Such combinations can be of two sorts: (i) late or early integration when combining

results obtained from single omics or (ii) intermediate integration when genuinely combining multiomics data for a common analysis (Gligorijević and Pržulj, 2015). This approach is backed up by the simple argument that a biological phenomenon is not encompassed by a set of independent layers but of complementary mechanisms. Multi-omics integrations have already been been proved successful when combining for example somatic mutations, RNA expression, DNA methylation and ex vivo drug responses by finding novel markers predictive of clinical outcome (Argelaguet et al., 2018). Many tools have since been developed with different goals in mind (Table 1)

Name	Category	Method	Example (cancer type)	Results of data integration	Data type	Programming Language	References
Joint NMF	unsupervised	matrix factorization	ovarian cancer	cancer subtyping	Multi-data	Python	(Zhang et al., 2012)
iCluster+	unsupervised	matrix factorization	colorectal carcinoma	cancer subtyping	Multi-data	R	(Mo et al., 2013)
iClusterBayes	unsupervised	matrix factorization	glioblastoma, kidney cancer	cancer subtyping, disease drivers	Multi-data	R	(Mo et al., 2018)
moCluster	unsupervised	matrix factorization	colorectal carcinoma	cancer subtyping	Multi-data	R	(Meng et al., 2016)
JIVE	unsupervised	matrix factorization	glioblastoma	cancer subtyping	Multi-data	MATLAB	(Lock et al., 2013)
MOFA	unsupervised	PCA	chronic lymphocytic leukemia	novel disease drivers	Multi-data	R/Python	(Argelaguet et al., 2018)
rMKL-LPP	unsupervised	multiple kernel learning, similarity-based	glioblastoma	cancer subtyping	Multi-data	available on request	(Speicher and Pfeifer, 2015)
NetICS	unsupervised	network-based	multiple cancers	disease drivers	Multi-data	MATLAB	(Dimitrakopoulos et al., 2018)
BCC	unsupervised	Bayesian	breast cancer	cancer subtyping	EXP, MET, miRNA, proteomics	R	(Lock and Dunson, 2013)
MDI	unsupervised	Bayesian	glioblastoma	cancer subtyping	Multi-data	MATLAB	(Kirk et al., 2012; Savage et al., 2013)
PARADIGM	unsupervised	pathway networks, Bayesian	glioblastoma, ovarian cancer	cancer subtyping, therapeutic opportunities	Multi-data	Python	(Vaske et al., 2010)
iBAG	supervised	multi-step analysis	glioblastoma	potential biomarkers of survival	Multi-data	R	(Wang et al., 2013)
SNF	unsupervised	network-based, similarity-based	glioblastoma	cancer subtyping	Multi-data	R/MATLAB	(Wang et al., 2014)
iOmicsPASS	supervised	network-based	breast cancer	cancer subtyping, disease drivers	Multi-data	R	(Koh et al., 2019)
NEMO		similarity-based clustering	acute myeloid leukemia	cancer subtyping	Multi-data	R	(Rappoport and Shamir, 2018)
PFA	unsupervised	fusion-based integration	clear cell carcinoma, lung squamous cell carcinoma, glioblastoma	cancer subtyping	Multi-data	MATLAB	(Shi et al., 2017)
CCA	unsupervised	correlation based	kidney renal clear cell carcinoma	mechanisms of carcinogenesis	CNV, methylation, gene expression	R	(El-Manzalawy, 2018; Lin et al., 2013; Zhou et al., 2015)

Table 1. List of bioinformatics tools and machine learning methods capable of performing multi-omics data integrations.
Table adapted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Menyhárt and Győrffy, 2021).

However, as with other methods, multi-omics integration suffers from certain limitations too. The first challenge to arise comes from the need to deal with omics obtained from different studies. Indeed, many tools require omics data to be profiled from the same samples and do not allow the existence of missing data points (Cantini et al., 2021) but hopefully, a standardisation of sample processing and data treatment might help alleviate this problem. Another important consideration to take into account is with what focus to approach the integration. In (Hasin et al., 2017), the authors grouped these approaches in 3 categories: “genomes first", “phenotype first” and "environment first” and described them as follows:

> *“Thus, the genome first approach seeks to determine the mechanisms by which GWAS loci contribute to disease. The phenotype first approach seeks to understand the pathways contributing to disease without centering the investigation on a particular locus. And the environment first approach examines the environment as a primary variable, asking how it perturbs pathways or interacts with genetic variation.”*

This question of focus is not only a food for thought or a set of constraints applied to specific types of studies. As it was shown in (Cantini et al., 2021), methods that seek omics specific factors as a goal often showed better performances than methods designed to find common or mixed factors.

1.2.7 Single-Cell omics

We have seen that each of the previously described types of data are capable of retrieving a specific part of biological processes happening in an organism and each one of them is capable of describing a particular cellular context. However, it is known that in the context of tumours, the environment can be highly heterogeneous, thus leading to a biological heterogeneity between cells. This creates a situation where by analysing a sample, we end up with an aggregation of cells bringing their own cellular context. The resulting data can only be seen then as an averaged pool of cell. But there is so much to be gained by recovering the diversity inside the sample by looking at its different constituents individually.

This need has led to the development of the single cell analysis technology, which allows to get an individual view of each cell to deal with the cell-to-cell variability (Ren et al., 2018) present in bulk data (Figure 1.7). Indeed, by observing the entirety of the cellular content as a single averaged value for each gene, the single-cell technology expands the possibilities to investigate the cellular heterogeneity and understand the cellular content of tumors with precision. Coupled with the discovery of specific cell type markers, techniques such as Fluorescence-activated Cell Sorting (FACS) or immunochemistry can be used to even further annotate single cells.

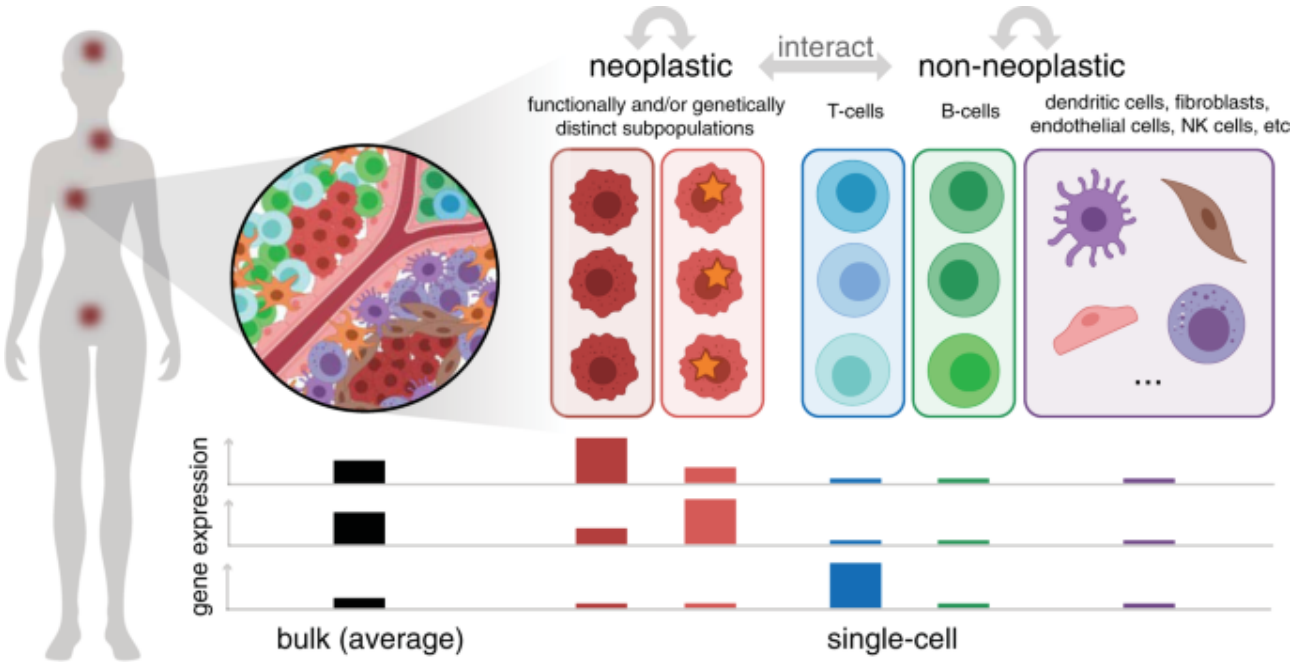


Figure 1.7. Difference of information availability between bulk and single cell data. On top: list of different cell types present at a site of interest. On the bottom: level of expression of 3 genes. While Single-cell data allows to differentiate between the two neoplastic cell as well as detecting a high expression in T-cells, Bulk data obscures this information where the gene highly expressed in T-cells shows a low expression because of the proportionally low abundance due to the average observation. Reprinted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Fan et al., 2020).

However, having such level of precision comes with new challenges, as stated in (Hu et al., 2016). Because of the size of the elements' sequences, this techniques requires a sufficient amount of DNA or RNA for gene expression analyses. During the sorting and separation of cells, a certain amount of material can be lost due to tubes absorptions. Another problem is the difficulty to replicate secondary structures of DNA. For proteomics studies, we are faced with a lack of amplification methods or high affinity probes that are needed to detect low abundance proteins because of the dilution happening after the lysis of cells. As for a methylation state analysis, the commonly used bisulfites sequencing requires a harsh treatment that fragments and degrades DNA, which makes the amount already available even more exacerbated.

1.3 Quantifying and qualifying the tumoral immune infiltration

1.3.1 IMMUCAN project: Integrated IMMUnoprofiling of large adaptive CANcer patients cohorts

Immunotherapy treatments are at the heart of cancer research with the rising of the new paradigm focusing on using the intrinsic mechanisms of defence against cancer instead of introducing foreign killing agents. Among the existing immuno-therapeutic treatments, immune checkpoint inhibitors (ICI) have shown great results

as I have talked about earlier. However, our knowledge of molecular and cellular components of the TME remains limited (Hui and Chen, 2015) and fails to explain why some treatments remain inefficient in certain patients (Roma-Rodrigues et al., 2019). It is to address these challenges that a European project funded for 40M€ jointly by EU and by industrials and pharmaceutic companies has been put in place under the name of IMMUCAN, standing for “Integrated IMMUnoprofiling of large adaptive CANcer patients cohorts”. This is a large scale project aiming at gathering more than 3.000 patients molecular and cellular tumor profiles across multiple cancer types and the generated data will be shared and analysed by 29 participants among which are 10 expert clinical centres.

Through the joint effort of its members, this project aims to improve our understanding of interactions between the tumor cells and various components of TME on a cellular and molecular level in the presence or absence of therapeutic intervention (Figure 1.8). For this, patients with different types of cancer (lung, colorectal, head and neck, breast and renal) are recruited and followed by the program. Some of these patients will receive ICI treatments while others will be considered as non-ICI and will follow a standard of care. The project will then perform an in-depth immuno-profilling by analysing cancer samples from these patients using bulk RNA-seq, exome sequencing, immunofluorescence and imaging analyses cytometry. For a selection of tumors, in-depth molecular analysis such as scRNA sequencing or whole genome sequencing will be applied. This data will then be integrated and analysed by an interdisciplinary team of experts to test several hypotheses with the goal of finding predictive markers for immunotherapy treatments. These results will be integrated into a research platform that will be first shared among the project participants and later can be made publicly accessible for the research community.

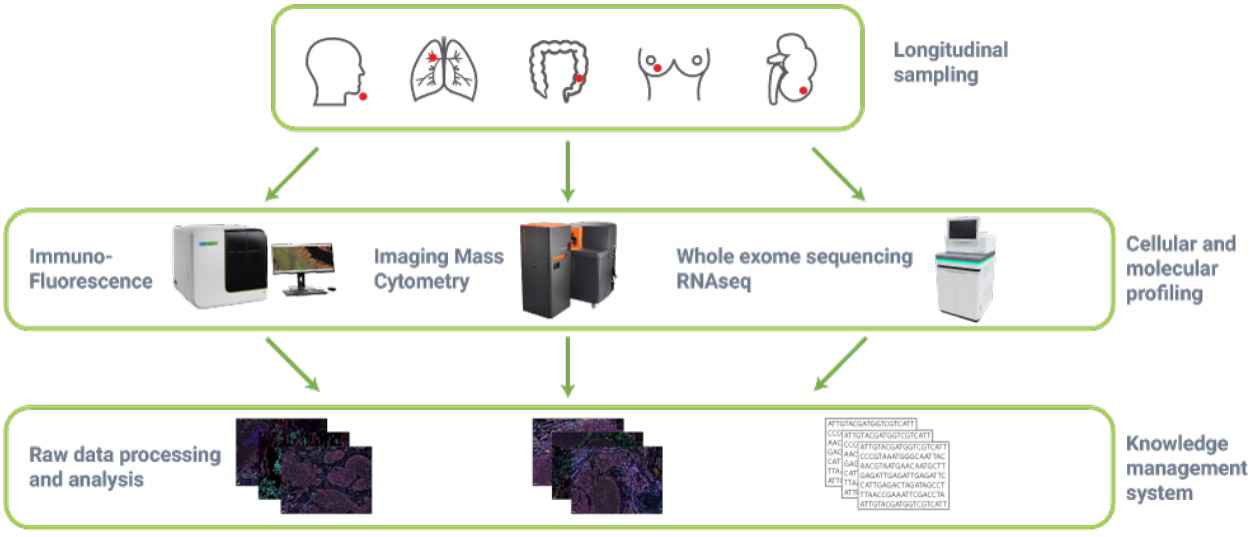


Figure 1.8. General workflow of the IMMUCan project.
Patients with different types of cancer are recruited and followed. Samples are collected and deep immunoprofiling data is generated. This data is then analysed on different levels by all the members of the project to understand tumor host interactions in the hope to identify potential predictive markers for immuno-therapeutic treatments.

It is in the specific task in the bioinformatics work package of IMMUCAN dedicated to unsupervised deconvolution that my PhD project fits and is funded by. This particular task aims at providing an in-depth analysis of whole tumor RNA-seq data with the goal of quantifying and qualifying the tumoral immune infiltration through the use of deconvolution methods. And it is the exact task that I will talk about in the next section of this manuscript.

1.3.2 What is Deconvolution?

I have explained previously in Section 1.1.3 that tumors are convoluted systems where a great number of different cells and cell types are constantly interplaying. I have also described how these interactions can be beneficial for the organism but also detrimental in certain cases and how various treatment types are being introduced to take advantage of the immune effects to cure cancer. Consequently, it is of uttermost importance that we understand the cellular composition of a bulk tumoral sample. The problem of quantifying this mixture of cell that is the TME and identifying how they contribute to the final measured molecular profile is called Deconvolution (Figure 1.9). It is believed that a detailed quantification of bulk tumoral samples can give us clues to the success determinants of the application of cancer therapies, as it sheds more light on the possible modulation of the intrinsic immune system response to therapies.

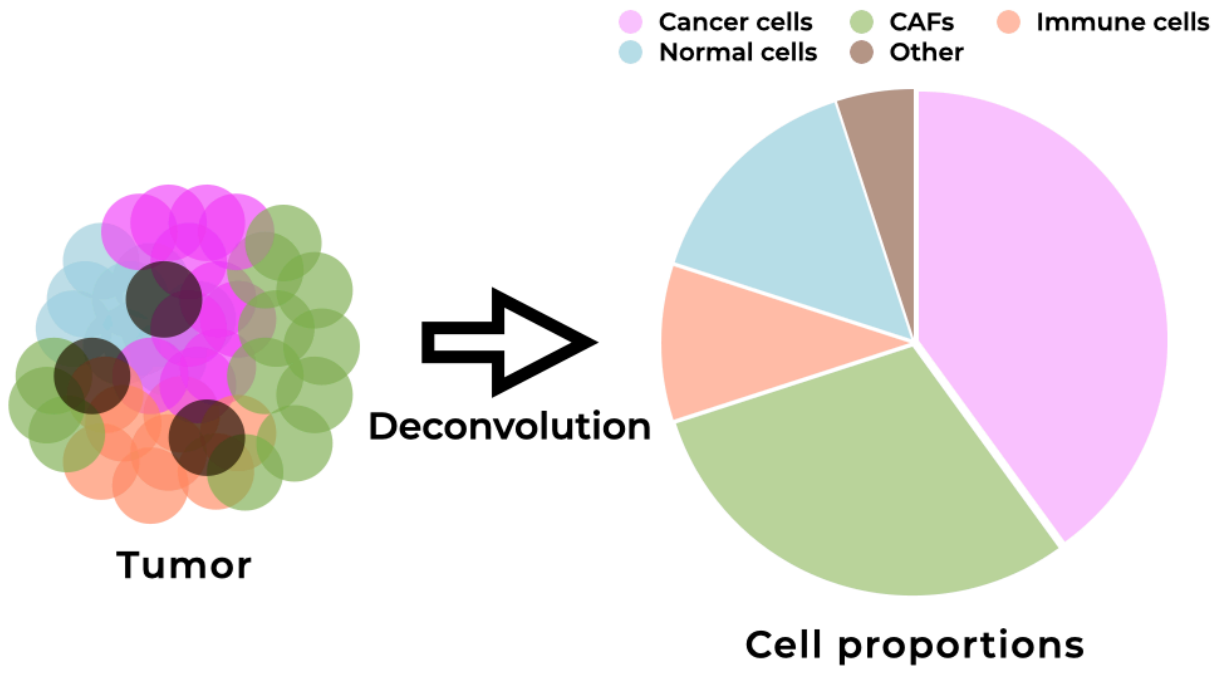


Figure 1.9. Schematic vision of tumor deconvolution for cell proportion identification. On the left, the tumor seen as a bulk of different types of cells that will be sequenced together. The deconvolution step consists in identifying the cellular content of the sequenced tumor and estimate the proportions of each cells present in the bulk. Cell type convolution can benefit from both experimental and computational approaches.

1.3.3 Computational cell-type deconvolution approaches

A similar problem of deconvolution from a mixture of signals can be seen in other fields such as sound processing, with the famous case of the “cocktail party problem” described first by Colin Cherry in 1953 (Cherry, 1953) detailed in (Bee and Micheyl, 2008) and expanded in (Bronkhorst, 2015). In this problem, multiple people are gathered in the same room with music playing in the background while the sound is recorded by several microphones set across the room. The task is then to separate the voice of each people from each other as well as from the musical background. This problem can also be imagined in the context of an orchestra (Figure 1.10) by trying to separate each group of instrument from each other based on their specific sound spectrum even when they play the same notes (Benetos et al., 2006).

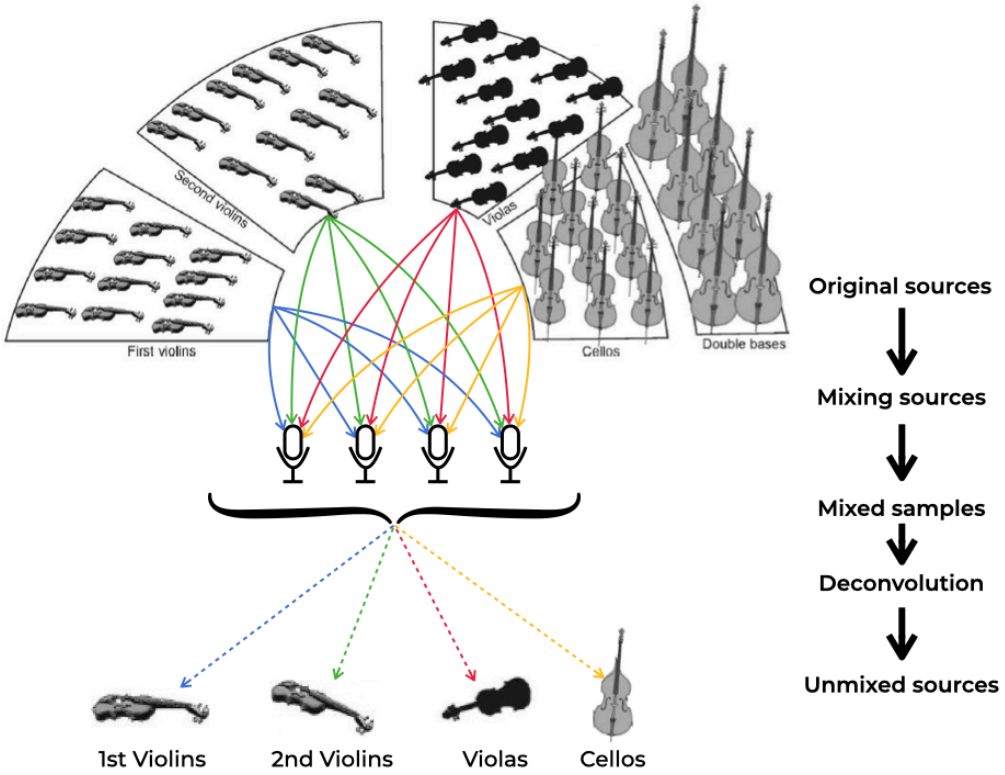


Figure 1.10. Illustration of the cocktail party problem applied to orchestra instruments.
During the play, individual groups of instruments can be recorded with a set of microphones and later recovered through deconvolution. In this illustration, I propose a simple case of 4 types of chord instruments but in real-life applications, the number of sound sources can be much larger.

Cell-type deconvolution methods can be divided into five different categories: probabilistic, enrichment-based, regression, matrix factorisation and convex-hull (Figure 1.11).

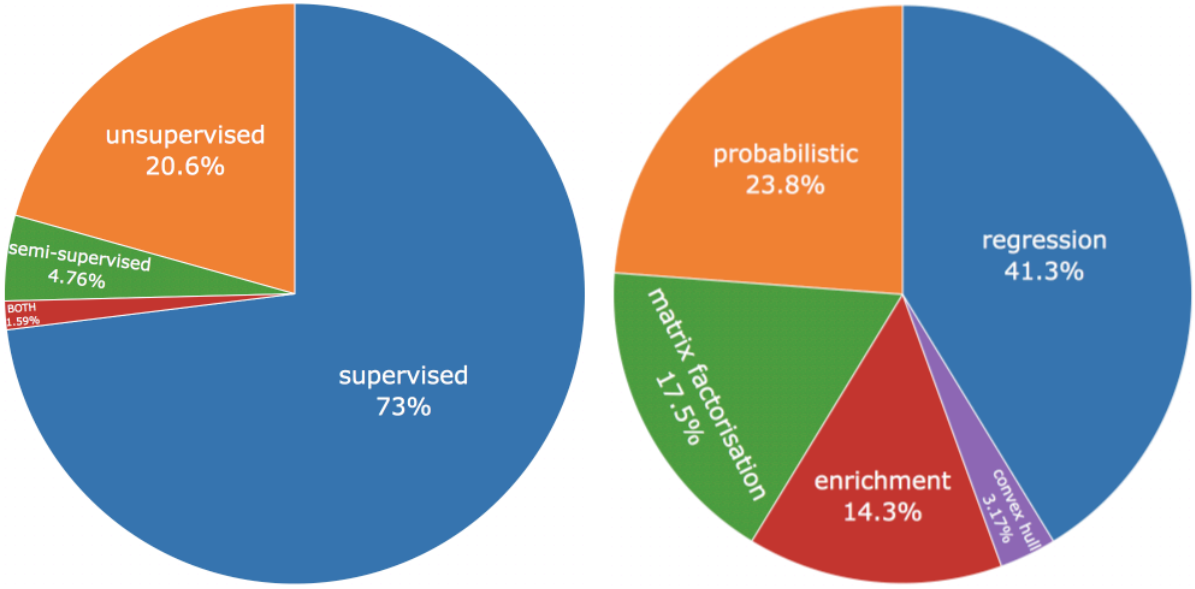


Figure 1.11. Overview of the number and categories of cell-type deconvolution methods.
On the left: proportions of approaches used to solve deconvolution problems, with supervised methods being a majority. On the right: proportions of mathematical approaches used by supervised and unsupervised methods, with regression algorithms being the most popular ones. Reprinted with permission from Urszula Czerwinska's PhD thesis (Czerwińska, 2018).

Probabilistic methods used for deconvolution are centred around the same goal: estimate an unknown joint density of a p-dimensional multivariate random variable. These methods are based on functions related to the Bayes theorem. Models built around this approach can be used to estimate the tumor purity with tools such as ISOpure (Quon et al., 2013) or cell-type proportions like the model named DSection (Erkkilä et al., 2010).

Enrichment based methods are designed to quantify the activity of a set of genes by calculating a score based on a list of reference genes. One of the most popular method to compute such score is known as Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). Tools such as SPEC (Bolen et al., 2011) and xCell (Aran et al., 2017) are based on enrichment approaches and calculate both scores directly related to the immune infiltration.

Regression models are a type of predictive technique which investigates the relationship between dependent variables as targets and independent variables as predictors. In the case of cell-type estimation, methods use already estimated signature genes as dependent variables. There are many different types of regression techniques which are mostly driven by the choice of the number of independent variables, their type and the shape of the regression line we want to fit. To find the best fit to the regression, the deviation of the datapoints is minimised in accord to the corresponding regression fitting shape. Among the existing regression types, we can mention the following: linear, polynomial, stepwise, Ridge, Lasso, ElasticNet and Support Vector Regression (SVR). EPIC for example is a method using a regression approach without assuming the distribution for the gene expression when calculating cell fractions (Racle et al., 2017).

Matrix factorisation will not be detailed here and will be explained later in the manuscript in Section 2.1.

Convex-hull regroup geometry-based methods that try to find sources of mixed signals by fitting specific markers in vertices of a complex plane (Figure 1.12). The goal of this method is to fit the data points into a small convex polygon with its vertices corresponding to the sources of signals present in the data. A convex-hull based approach called Complex Analysis Mixtures was proposed by Zhu et al. (Zhu et al., 2016) and applied to cancer transcriptomics by Wang et al. (Wang et al., 2016) to find subpopulation specific marker genes in various tissues through blind source separation.

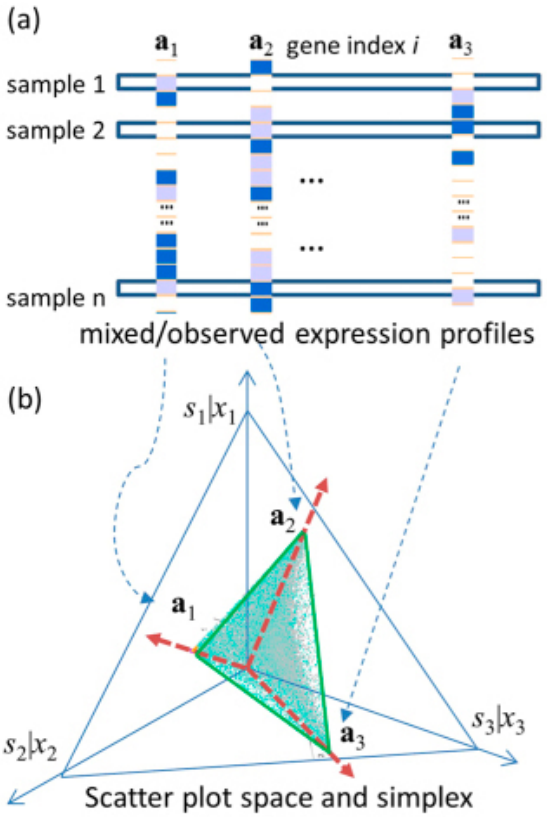


Figure 1.12. Complex Analysis Mixtures (CAM) principle.
Marker genes enriched in specific cell-type subpopulations are identified (here a_1, a_2 and a_3 genes) and then projected onto rotated scatter simplexes whose vertices have subpopulations specific marker genes. Adapted and reprinted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Wang et al., 2016).

All these mathematical methods, be they applied for classification, clustering or regression purposes have an additional distinction between each other depending on the prior knowledge they use. Some algorithm can rely only on a single dataset such as patient's gene expression profiles while other can use additional inputs to help guide the algorithm using *a priori* knowledge. It is this distinction between data agnostic methods called unsupervised and supervised methods requiring additional reference data that I will now present.

1.3.4 Supervised deconvolution approaches

Supervised approach to deconvolution is a mean to predict an output using already known and labeled cellular reference profiles. More generally, supervised learning is performed in two steps. Before the learning process is initiated, the data is separated into 2 sets: one for training and the other for testing. Then, the model is trained in the first step using the training set. In the second step, the model performance is evaluated on the testing set.

In the context of tumor deconvolution, models have to be trained on labeled data to extract cell type specific signatures. Unfortunately, there is still a lack of gold standards and agreed reference data by the community. But one possible way is to use what could be called “silver standards” by using purified single-cell sequences. Many supervised methods of cell type deconvolution use this approach and integrate RNA profiles from single cell sequencing to generate signatures matrices that can later be used for bulk tumor cell type estimations (Figure 1.13). Some methods like contamDE (Shen et al., 2016) focus on estimating the purity of tumors (percentage of cancer cells in the tumor) while others like CIBERSORT try to estimate the cell-type proportions (Newman et al., 2019).

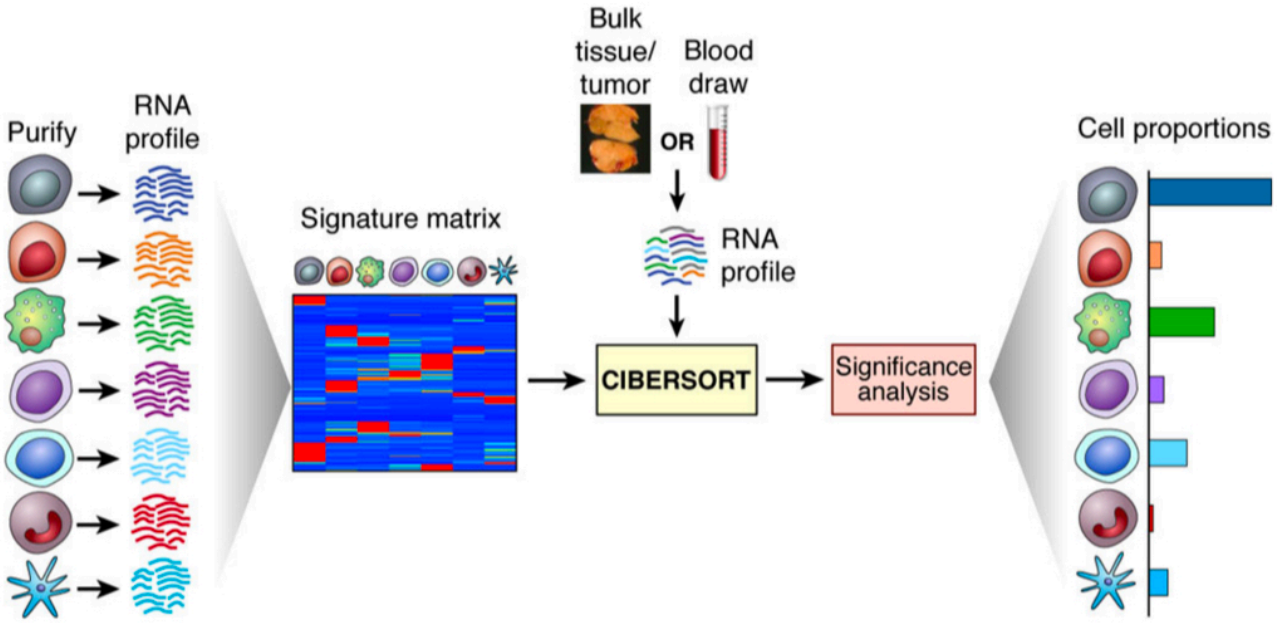


Figure 1.13. Example of a supervised deconvolution approach used by CIBERSORT to determine immune cell proportions.
Cell type specific signatures are first obtained from single-cell sequencing of purified immune cell that are compiled in a reference signature matrix. This matrix can later be used to estimate the immune cell proportions from a bulk tumor by using a Support Vector Regression (SVR) approach. Adapted with permissions from Springer Nature from (Newman et al., 2015).

Supervised methods have a major advantage of being accompanied by insights from domain experts as each signature is usually verified and annotated by them. The obtained signature matrices can then be considered as references and used across other studies. However, as with any supervised learning approaches, a problem may arise, which is the bias of the training data. Indeed, the data used for training might present certain characteristics that could make the learned model unfit for other studies. Gregor Sturm et al. have for example analysed the various mistakes made by supervised models for immune cell-type estimations (Sturm et al., 2019). One of the mentioned problem is that learned signatures are not always specific to their assigned cell-type and can lead to the mislabeling of certain cell-types via what is called the spillover effect (Figure 1.14). The cause for this effect has been justified by filtering out certain genes from the signatures, which improved the classification of previously mislabeled types (Sturm et al., 2019).

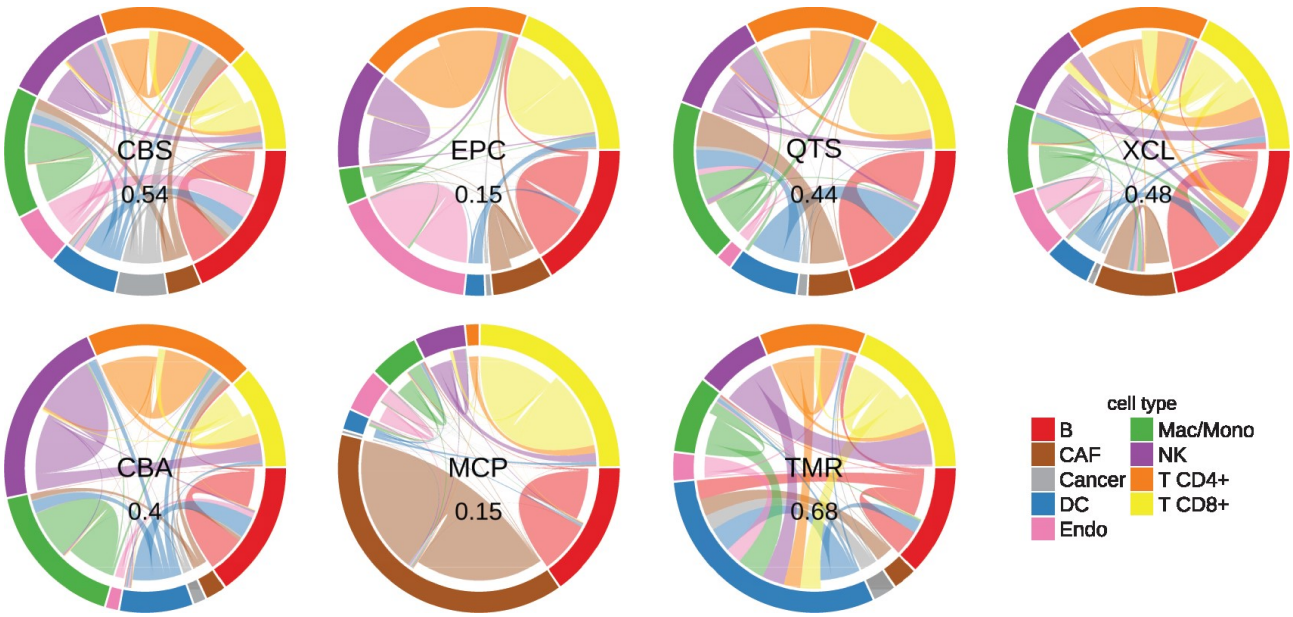


Figure 1.14. Spillover analysis of supervised deconvolution methods.
The spillover effect consists in a method predicting erroneously the presence of a cell type different than actually present in the data. This effect is mainly attributed to a low specificity of signature genes. Reprinted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Sturm et al., 2019).

Another problem that we have to deal with when using single-cell sequences is the scarcity of the data. It is often hard to account for the reason a gene is lowly expressed since it can be due to biological reasons or technical problems. This type of issue can lead to incomplete signatures and could be considered as an inductive bias since the model might have troubles assigning a label to a cell-type that has an expression profile different from the one expected from the reference signature.

While this is not always a problem, the fact remains that some model are sensitive to overfitting. Most models are trained on a restricted set of data and since we saw that tumors can be extremely heterogenous, a certain expression pattern may not be reproducible in all tumors of the same cancer type. Sometimes, some models are trained only on a particular type of cancer and are limited to studies focused on similar type of data. One example can be given with the method EpiDISH which at its

release was only applicable to whole blood, generic epithelial tissue and breast tissue data (Teschendorff et al., 2017). Nevertheless, models trained on a particular type of cancer type can be seen as highly specific and often give better results than more general models.

Nevertheless, sometimes surprisingly, the method can be more important than the reference signatures. In 2019, I have participated in the DREAM Challenge which proposed a challenge dedicated to the prediction of immune types (Decamps et al., 2021). Since only the cancer type was known and the number of different cells and their type in the data used in the challenge were unknown in advance, supervised methods proved to be of limited success. Yet, the two most competitive approaches that gave the best results were the unsupervised method of ICA and a reference-based method called EpiDISH (Teschendorff et al., 2017). Both methods resulted in a very precise estimation of the number of cell types and their proportions. While the type of cells detected was unknown with ICA at first, the annotation were added after the deconvolution step. But EpiDISH necessitated a reference matrix to be able to estimate the presence and proportion of cells in the tumor. Surprisingly and paradoxically, the best proportion estimates were achieved using a reference matrix intended to be used to estimate fractions of epithelial cells, fibroblasts, fat cells and total immune cells in breast tissue (*centEpiFibFatIC.m* matrix from the EpiDISH R package available from (Teschendorff and Zheng, 2021)) while regressing the signals of fibroblasts and fat cells. This result was surprising since the true composition of the data used in the challenge was from Pancreatic Adenocarcinoma (PDAC) and contained only 5 types of cells, namely Classical and Basal tumor cells, Immune cells, Fibroblasts and Healthy pancreatic cells. It is indeed an illustration of the statement that: "***If you torture data long enough, it will always confess***"...

1.3.5 Unsupervised deconvolution

I will not develop much about unsupervised learning as the next chapter will be devoted to a particular set of unsupervised methods called Matrix Factorisation. I will however briefly present what is meant by unsupervised learning and the advantages it can confer by using this type of learning instead of supervised approaches.

Unsupervised learning, contrary to supervised models does not require any prior knowledge to work or the knowledge is required only at the step of result interpretation and, therefore, have less strong impact on the data analysis itself. It is a model that works on its own by segmenting the data and fitting it accordingly to a given constraint. This approach is often used when some parts of the data are unknown or too difficult and expensive to label properly by the user. In this case, the model is required to learn the inherent latent structure and its variables. Unsupervised learning is then most commonly used for explanatory analyses such as clustering and dimensionality reduction (Gorban et al., 2008; Xu and Wunsch, 2009).

Dimensionality reduction is used when dealing with data that is too large to be comprehended or visualised in an easy and straightforward way. To solve this, it is possible to reduce the amount of variable to a limited set that is representative of the whole distribution. Working in a sparser latent structure may be more interesting than in the original space since it may allow to eliminate redundant features and ease consequent data processing.

Clustering consists in grouping together elements sharing similar characteristics into entities called clusters. These characteristics may not be known in advance so an

unsupervised approach is advised in such circumstances. This is typically done with the hope of obtaining clusters corresponding to different phenomena.

Clustering however is not to be confused with dimensionality reduction. While it may be tempting to categorise it as a dimensionality reduction method because of the simplified representation of the data into a lower dimension, clustering is used to reveal a certain structure of the data. Dimensionality reduction, however, is often performed to avoid the “curse of dimensionality” and the problems coming with it such as computational complications or visualisation difficulties. Another important role of dimensionality reduction in the analysis of omics data is improving the signal/ noise ratio.

1.4 Knowledge maps and models for formalised TME description

I have discussed previously in Section 1.1 about the complex system that is cancer and the need to have a multi-level approach to correctly comprehend its functioning. Various methods and types of data are available for researchers. They can either be applied independently or integrated together for a larger view of possible interplaying functions. But even if integrations are possible, the means to detect relations and interactions between the different studied elements remains a challenge. To tackle the problem of identifying existing cross-talks and coordination between molecular functions and pathways or discovering the impact that a particular deregulation can have on other systems, new fields of cancer research have emerged. Knowledge scattered across the diverse publications have been gathered and regrouped in comprehensive and functional representations which can take the form of molecular maps or models (Mazein et al., 2018). I will describe here what these representations are, their content and how they can be used to help to describe the interplay between tumors and their environment.

1.4.1 Molecular networks and maps

Molecular maps are similar to geographical maps in the sense that they represent a top-down view of interacting paths following the biological processes in a living cell. This representation, often in the form of depicted molecular pathways, have three major goals: to make a resource containing a formalised summary of biological knowledge from many research groups, to supply a platform for sharing and discussing biological mechanisms and finally, to create an analytical tool for high throughput data integration and analysis. Organising the available knowledge in comprehensive and structured networks allows to capture non-trivial interactions and regulatory circuits between all molecular components. Networks can come in different types (Farber and Mesner, 2016) but I will focus here mainly on signalling and metabolic networks since those are the type of networks I have the most experience with.

Signaling pathways are a representation of the information flow that passes through biochemical reactions or molecular interactions. By having access to a comprehensive network in a graphical form, we are able to follow the flow of signals with a clearer view and have a better idea of the particular contexts in situations where a deregulation happens. Moreover, by connecting together various elements from different pathways, signaling network can gain new properties that wouldn’t be noticeable when looking at individual components. Azeloglu and Iyengar (Azeloglu

and Iyengar, 2015) have noted 4 different emerging properties which are ultra-sensitivity (a small modification or stimulus can give rise to large change in downstream effectors), bi-stability (appearance of regulating feedback loops), redundancy and robustness (a single input can be connected to an output through multiple pathways) and finally oscillatory behaviour (coupling of positive and negative feedback loops). These properties reflect the complexity of biological processes and help understand the cellular machinery under a new light.

During my research experience, before starting my PhD project, I have participated in the integration and update (Kondratova et al., 2018) of the most comprehensive cancer specific resource named Atlas of Cancer Signalling Network 2 (ACSN2). This atlas represents different interconnected maps of signalling, metabolic and tumor microenvironment networks. Each map is manually curated and covers hundreds of molecular reactions. Compared to the previous version of ACSN (Kuperstein et al., 2015), 10 new maps have been added, with many of them related to immune response or specific immune cell types, leading to a resource composed of over 8.000 reactions integrating 3.000 proteins and 800 genes.
This network is integrated on two web-based platforms, MINERVA (Gawron et al., 2016) and NaviCell (Bonnet et al., 2015), which allows to import and visualise heterogeneous omics data on top of the maps either by using the direct profiles, by using enrichment techniques such as GSEA or a method developed in my team, called ROMA (Martignetti et al., 2016). This data visualisation feature has been applied with subsets of melanoma with annotation on clinical stages and results of visualisations showed distinct patterns in cell cycle, regulated cell death and cell survival pathways as well as active immune response related pathways in long term survival patients (results not yet published). ACSN has also been used to visualise the evolution of pathway activities of Ewing sarcoma cell lines sequenced at different time points (Monraz Gomez et al., 2021).

I have mentioned in Section 1.1.2 that metabolic reprogramming is a hallmark of cancer, making the analysis of metabolic pathways a major point of interest with recent studies such as (Gatto et al., 2020) showing that the activity of certain reactions are specific to certain cancers. The first most complete reconstruction of the human metabolism was achieved with the map of RECON (Duarte et al., 2007) and has allowed the discovery of alternative targets of known drugs.
This network was later upgraded to RECON2 (Thiele et al., 2013) which doubled the number of metabolic reactions and incorporated almost the same amount of new metabolites. This new version could also be used as a predictive model using techniques such as Flux Balance Analysis which analyse the flow of metabolites through a metabolic networks and can impose specific constraints that can be associated to certain diseases (Orth et al., 2010). This possibility transformed a metabolic network into a powerful tool able to process different types of data to predict possible outcomes like cancer drug targets (Folger et al., 2011).
Since then, a new version of this network was released again called RECON3D (Brunk et al., 2018), for which I had the chance to participate in. This map was also integrated into a database called Virtual Metabolic Human (VMH) which encapsulates the current knowledge of human metabolism interlinked resources such as “Human metabolism”, “Gut microbiome”, “Disease”, “Nutrition” and “ReconMaps”. One of the new additions to the metabolic network with this version was the inclusion of gene-protein-reaction associations with the possibility to visualise changes in protein due to genetic mutations and map these changes to related metabolic function modifications.

Thanks to these two types of networks, researchers can have an easier access to a broader range of knowledge. Having said all that, the interplay between metabolic processes and signalling pathways remains poorly understood despite the existence of such complete molecular maps. To fill this gap in knowledge, my work before my PhD led me to the project of integrating the resources of ACNS and RECON2 together (Sompairac et al., 2019). Interconnecting these networks allowed to expand our vision of interactions between signalling and metabolic pathways (Figure 1.15).

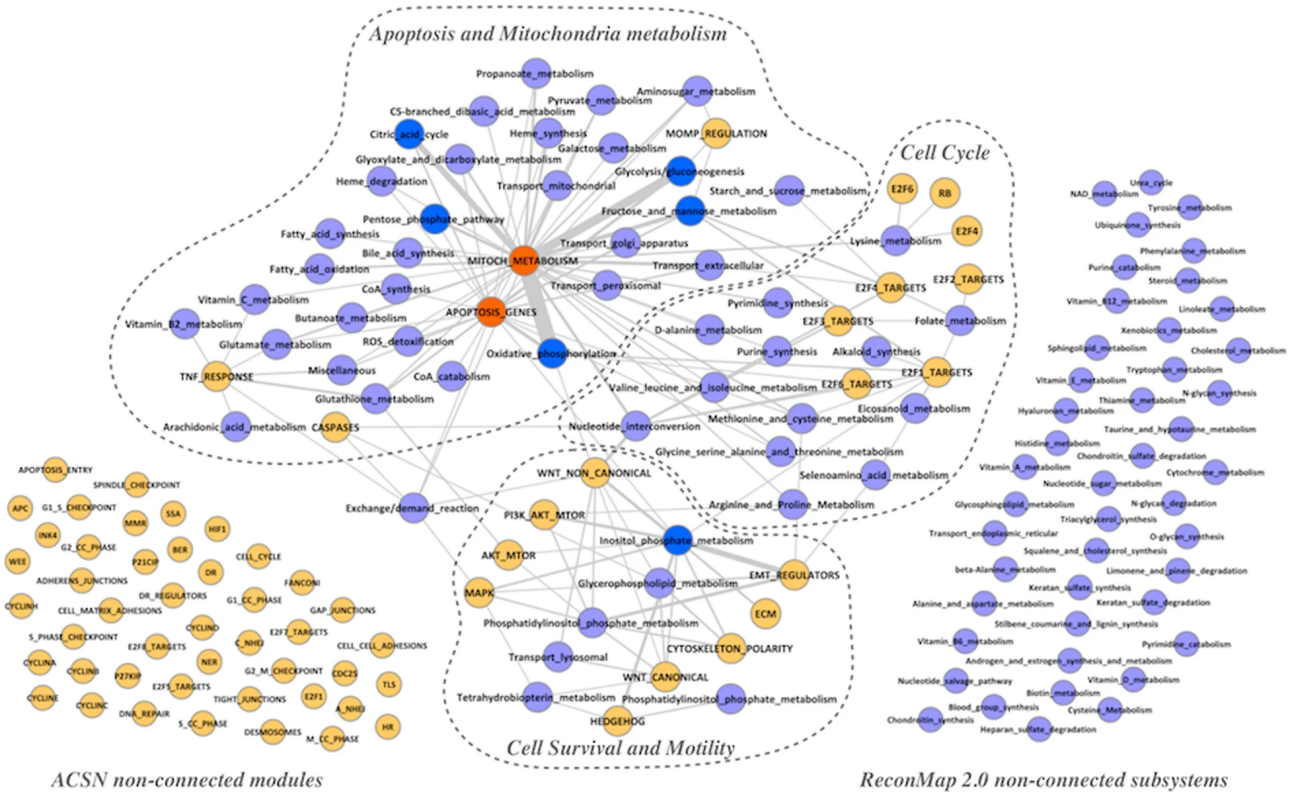


Figure 1.15. Crosstalk between signalling pathways of ACSN and metabolic processes from RECON2.
Nodes representing ACSN pathways are coloured in Orange and RECON2 metabolic reactions are coloured in Light Blue. Reprinted with permission under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Sompairac et al., 2019).

Apart from such large maps, there also exist many others that focus on more specific aspects or a particular mechanism. One of such maps is the map of Regulated Cell Death (RCD) which gathers information about all known modes of mechanisms related (Ravel JM et al., 2020). This map has been created in my team and I have participated in its integration in NaviCell as well as its comparison with other similar pathway ressources. Its modular structure contains molecular informations on all major regulated death pathways, facilitating the visualisation of their cross-talks. And compared to other existing resources presenting RCD pathways, having such a map specifically designed for it makes it stand out by its higher number of recent and uniquely annotated elements participating in RCD. Its application examples of data integration and analysis showed that this resource could not only be used to see the deregulated RCD processes in a disease or sample but also highlight the most contributing players which could be the focus of future therapeutic targets. In particular, the map was used to study the mechanisms of molecular comorbidity

between lung cancer and Alzheimer disease. In addition, relying on the definition of ovarian cancer subtypes reported in (Bell, D et al., 2011), the analysis of ovarian cancer data from TCGA demonstrated the capacity of this resource to identify in greater detail the immunoreactive subtype through the specific up-regulations of various modules in the map.

1.4.2 Models in cancer bioinformatics

Mathematical modelling of biological processes involved in cancer is a vast topic. Therefore, I will only be able to skim over this complex field of study. As Robert Constanza stated in (Costanza et al., 1993):

> *"Models are analogous to maps. Like maps, they have many possible purposes and uses, and no one map or model is right for the entire range of uses. It is inappropriate to think of models or maps as anything but crude, although in many cases absolutely essential, abstract representations of complex territory. Their usefulness can best be judged by their ability to help solve the navigational problems faced."*

This definition applies to the field of bioinformatics, where models can be seen as artificial systems mimicking particular functions of an organism. They often focus on identifying essential elements and their interactions required to solve a distinct problem.

This subject being relatively recent, we can find few successful applications of mathematical modelling to TME. We could mention some modelling of the interaction between certain immune cell-types such as macrophages and T-cells with the TME.

Mahlbacher et al. in (Mahlbacher et al., 2018) created a framework capable of evaluating macrophage interactions with the TME and assessing how it may affect tumor growth. They showed that Tie2 expressing macrophages which are a target of immunotherapy inhibition may fail with the presence of M2 macrophages that continue to exert a tumor growth effect.

Cess and Finley in (Cess and Finley, 2020) have further studied the interaction between macrophages and T cells in regard with tumor proliferation. Their model managed to capture important interactions between M1 macrophages and T-cells that can improve the result of multiple immunotherapies. They also observed how macrophages displayed an M2 phenotype when no treatment was given, leading to an equilibrium where T-cell are able to slow the tumor growth but are unable to remove the tumor.

Finally, Li et al. in (Li et al., 2019) focused on the interaction between macrophages and the TME and the role it played on the transition of cancer cells to epithelial or mesenchymal state associated with metastasis and immune evasion. Through their model, they showed that treatments should focus on the maintenance of an M1 dominated system and the inducing of mesenchymal to epithelial transition can help to limit tumor progression.

In my group, a mathematical model of immune checkpoint network was created with the aim of explaining the synergistic effects of combined immune checkpoint inhibitor therapy and the impact of cytokines in patient response (Kondratova et al., 2020).

1.5 Summary

Cancer is a very old disease that is still omnipresent in our era and touches almost 20 million people in the world with 10 millions cases of cancer death reported in 2020 (Sung et al., 2021). Our growing understanding of tumors led us to the discovery of the importance of their surrounding micro-environment and the role it plays in cancer progression.

New technologies have emerged to help us study cancer at different levels and made possible for researchers to better understand the different states of such complex disease. Cancer diagnosis was thus improved and new treatments were formulated based on the discoveries of TME implications.

But with the advent of these new technologies and an ever increasing sight of interactions between tumor and their TME, it was required to quantify and qualify these interactions in greater detail, not only to ensure a proper interpretation of such mechanisms but also to better guide treatments discovery and applications. For this, many bioinformatics methods were developed to help deconvoluting various signals contained in the TME and interpret them.

In the next chapter I will present in detail a particular set of mathematical methods used to solve the problem of understanding the cellular heterogeneity of tumors and extracting specific signals related to functions that could be promising targets for TME focused therapies.

2. Independent Component Analysis: a matrix factorisation method to solve the deconvolution problem

2.1 Introduction to Matrix factorisation

2.1.1 Matrix factorisation principles

Matrix factorisation or matrix decomposition can be seen as a way to reduce a matrix into its constituent parts. This approach is mainly used to approximate a matrix of full rank by another matrix, having a lower rank. We can use as an analogy the factoring of numbers, such as factoring “15” into “3 x 5”. However, just as there are various ways of factoring numbers, there are also various ways of decomposing matrices.

Matrix factorisation methods are considered as unsupervised approaches for dimensionality reduction so when trying to compute

$$X = A \bullet S$$

only X is known and A and S have to be estimated simultaneously without any *a priori* knowledge. There exist a multitude of various matrix factorisation methods (Theodoridis, 2020) with the most representative ones being Principal component analysis (PCA), Canonical correlation analysis (CCA), Independent component analysis (ICA) and Non-negative matrix factorisation (NMF). I will now briefly introduce these methods in the following sections.

2.1.2 Principal Component Analysis

Principal Component Analysis (PCA) is the oldest and most popular method of matrix factorisation used for dimensionality reduction (Pearson, 1901). The goal of PCA is to project a dataset from many correlated coordinates onto fewer uncorrelated coordinates, called principal components (PCs), while still trying to retain most of the variability present in the data. To achieve that, PCA can be computed via the eigendecomposition of the covariance matrix.

If the data is centred by subtracting the mean μ from each data vector x_i of a matrix X of size $n \times p$ of rank $r \leq min\{n, p\}$, the covariance matrix C is calculated by:

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \frac{1}{n-1} X^T X$$

Then, the eigenvalue decomposition of C can be obtained as follows:

$$C = VDV^T = \sum_{i=1}^{r} \lambda_i v_i v_i^T$$

Where D is the diagonal matrix of eigenvectors and V the matrix containing the eigenvectors of C. Here v_i is the i-th PC and λ_i is the i-th eigenvalue of C and equals to the variance of the data along the i-th PC.

Taking this matrix X, it is also possible to use another method called the Singular Value Decomposition (SVD) and this time factorise X:

$$X = UDV^T$$

Here U is an orthogonal matrix $n \times r$ ($U^T U = I_r$, with I_r the identity matrix $r \times r$) whose columns are called left singular vectors of X; D is a diagonal matrix $r \times r$ whose diagonal elements are called singular values; V is an orthogonal matrix $p \times r$ ($V^T V = I_r$, with I_r the identity matrix $r \times r$) whose columns are called right singular vectors of X. The columns of UD are what we call Principal Components (PC) and the variance of these PCs is given by the square of the singular values of X divided by $n - 1$.

PCA therefore provides an r rank matrix factorisation of X by imposing an orthogonality between them. There exist many different algorithms of PCA that adapt it to achieve modified goals to analyse different types of data. To name a few, Functional PCA, Simplified PCA, Robust PCA and Symbolic PCA have been used to analyse chemical spectroscopy, atmospheric sciences, image processing or histograms (Jolliffe and Cadima, 2016).

2.1.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a technique developed by (Hotelling, 1936) to analyse two datasets in a joint way. The goal behind it is to find a linear transformation for both datasets such that after the transformation, the pair of obtained variables from both datasets are maximally correlated. This allows to find closely related signals coming from different sources.

Suppose we have two sets of variables $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_q)$. We can then define a set of linear combinations U and V such as:

$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ U_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \\ U_p &= b_{p1}X_1 + b_{p2}X_2 + \ldots + b_{pp}X_p \end{aligned}$$

$$\begin{aligned} V_1 &= b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1q}Y_q \\ V_2 &= b_{21}Y_1 + b_{22}Y_2 + \ldots + b_{2q}Y_q \\ &\vdots \\ V_q &= b_{q1}Y_1 + b_{q2}Y_2 + \ldots + b_{qq}Y_q \end{aligned}$$

Thus defining (U_i, V_i) as the i^{th} canonical variate pair.

We can then compute the variance of U_i with the following expression:

$$var(U_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} a_{ik} a_{il} cov(X_k, X_l)$$

The variance of V_j can be similarly computed with the following expression:

$$var(V_j) = \sum_{k=1}^{q} \sum_{l=1}^{q} a_{jk} a_{jl} cov(Y_k, Y_l)$$

The covariance between U_i and V_j is then:

$$cov(U_i, V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} a_{ik} a_{jl} cov(X_k, Y_l)$$

Finally the canonical correlation for the i^{th} canonical variate pair is simply the correlation between U_i and V_j:

$$\rho_i^* = \frac{cov(U_i, V_j)}{\sqrt{var(U_i)var(V_i)}}$$

ρ_i^* is the quantity maximised by CCA algorithms by finding linear combinations of X and Y maximising the above correlation.

2.1.4 Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) is a modified approach to PCA where a constraint is applied to guarantee the non-negativity of the elements of the resulting factors. This constraint is forced in applications where a negativity doesn't make sense, such as imagery where pixels can't have a negative intensity or genomics where a gene can't have a negative expression.

Given a matrix X of size $l \times m$, the NMF tries to find an approximate factorisation of X such that:

$$X \approx AZ$$

A is a matrix of size $l \times N$ and Z a matrix of size $N \times m$ with $N \leq min\{l, m\}$ with all the matrix elements being non-negative.

To obtain a good approximation, a common cost function called Frobenius norm can be used for the error matrix. For this application, the NMF task can be described as follows:

$$\min_{A,Z} \quad \|X - AZ\|_F^2 := \sum_{i=1}^{l} \sum_{j=1}^{m} \Big(X(i,j) - [AZ](i,j)\Big)^2,$$
$$s.t. \quad A(i,k) \geq 0, Z(k,j) \geq 0,$$

Where $[AZ](i,j)$ is the (i,j) elements of matrix AZ and i, j, k run across all the possible values. Many other algorithms have been proposed beside the Frobenius norm (Sra and Dhillon, 2006) but I will not enter into such details here.

NMF has been successfully applied in various fields such as document clustering (Xu et al., 2003), molecular pattern discovery (Brunet et al., 2004), image analysis (Lee and Seung, 1999), clustering (Szymkowiak-Have et al., 2006), music transcription (Smaragdis and Brown, 2003) and music instrument classification (Benetos et al., 2006), face verification (Zafeiriou et al., 2006) as well as immune cell activation signature identification (Davis-Marcisak et al., 2021).

2.1.5 Independent Component Analysis

The Independent Component Analysis (ICA) was first formulated by (Herault and Jutten, 1986) and can be described by the following equation:

$$X = S \bullet A$$

Given a matrix X of size $l \times m$, this method tries to maximise the statistical independence and non-gaussianity of latent variables of the matrix S, called independent components. The matrix A is known as the mixing matrix and its elements as the mixing coefficients.

Let us denote Z as the estimate of the latent variable matrix S as:

$$Z := A^{-1}X$$

With $Z_i, i = 1,...,l$ as the latent variables, we will refer to them as independent components. It is possible to maximise the independence by measuring and minimising the mutual information I of latent variables:

$$I(Z) = -H(Z) + \sum_{i=1}^{l} H(Z_i)$$

Where $H(Z_i)$ is the associated entropy of Z_i and $I(Z)$ is equal to the Kullback-Leiber (KL) divergence. If the KL divergence, and thus the mutual information $I(Z)$ becomes zero, then the independent components Z_i become statistically independent.

There exist multiple methods for computing ICA such as Infomax (Bell and Sejnowski, 1995) and JADE (Rutledge and Jouan-Rimbaud Bouveresse, 2013) but among them, FastICA is probably the most popular (Hyvärinen and Oja, 2000). FastICA has indeed many interesting properties among which are a fast convergence, non-necessity for a prior estimation of a probability distribution function allowing it to find independent components directly and its capacity to estimate independent component one by one.

This algorithm tries to find the direction of the column vector $\mathbf{w} \in \mathbb{R}^N$ that maximises the non-Gaussianity of $\mathbf{w}^T X$ projection distribution.

To measure this non-Gaussianity, Hyvärinen and Oja (Hyvärinen and Oja, 2000) state that FastICA relies on non-quadratic non-linear function $f(x)$, its first derivative $g(x)$ and its second derivative $g'(x)$. Several functions have been suggested for this purpose, each of which highlights certain aspects of non-Gaussian distributions, and serves as a surrogate way to estimate negentropy. For example, one of the popular choices for $f(x)$ is:

$$f(x) = -e^{-x^2/2}, \quad g(x) = xe^{-x^2/2}, \quad and \quad g'(x) = (1 - x^2)e^{-x^2/2}$$

The FastICA algorithm (Figure 2.1) can then be described as follows:

Input: C Number of desired components.
Input: $X \in \mathbb{R}^{N \times M}$ Prewhitened matrix, where each column represents an N-dimensional sample, where $C \leq N$.
Output: $A \in \mathbb{R}^{N \times C}$ Unmixing matrix where each column projects X onto independent component.
Output: $S \in \mathbb{R}^{C \times M}$ Independent components matrix, with M columns representing a sample with K dimensions.

for p **in** 1 *to* C :
  $\mathbf{w_p} \leftarrow Random\ vector\ of\ length\ N$
  while $\mathbf{w_p}$ *changes*

$$\mathbf{w_p} \leftarrow \frac{1}{M}\mathbf{X}g(\mathbf{w_p}^T\mathbf{X})^T - \frac{1}{M}g'(\mathbf{w_p}^T\mathbf{X})\mathbf{1_M}\mathbf{w_p}$$

$$\mathbf{w_p} \leftarrow \mathbf{w_p} - \sum_{j=1}^{p-1}(\mathbf{w_p}^T\mathbf{w_j})\mathbf{w_j}$$

$$\mathbf{w_p} \leftarrow \frac{\mathbf{w_p}}{\|\mathbf{w_p}\|}$$

output $\mathbf{A} \leftarrow [\mathbf{w_1}, \ldots, \mathbf{w_C}]$
output $\mathbf{S} \leftarrow \mathbf{W^T X}$

Figure 2.1. FastICA algorithm.

One of the most famous application examples of ICA is the classical *cocktail party problem* that I already described previously as a deconvolution problem. But ICA can also be used in image and video processing, even when the number of mixed signals is unknown (Isomura and Toyoizumi, 2016). We will see later in Section 2.3 how ICA is also successfully applied to various cancer omics data.

2.2 Standard workflow applied for ICA

While many different matrix factorisation methods have been mentioned that appeared to be useful in various omics data analyses, in my thesis, I will focus mainly on the Independent Component Analysis as a deconvolution method. ICA as a method can be applied in a straightforward way without any a priori knowledge and is known to work with many different types of omics data as we will see in the next section. However, I deemed necessary to briefly but explicitly describe the standard workflow applied to clarify all future mentions of ICA applications.

2.2.1 Preprocessing

One of the first and sometimes disregarded steps of deconvolution is the data pre-processing. Apart from cleaning the data from any noise or defects, it is important to take into account the normalisation applied to the data prior to deconvolution. As observed in (Avila Cobos et al., 2020), data transformation and scaling/normalisation can have a big impact on the resulting analysis. Through multiple analyses using ICA in my team, it was observed that ICA works best in the sense that it gives more biologically interpretable results when the data is transformed using the $log(x+1)$ transformation when dealing with RNA-seq data.

The most popular implementation of ICA is fastICA (Hyvärinen and Oja, 2000) for reasons already mentioned before. And as stated by Hyvärinen and Oja in this article, to make the estimation of ICA simpler and better conditioned, we can perform a preprocessing by centering and whitening the data.

Centering the data consists in removing the mean from each row of the input matrix X containing N columns:

$$x_{ij} \leftarrow x_{ij} - \frac{1}{N}\sum_{j'} x_{ij'}$$

Whitening, also called sphering is a linear transformation where we impose a unit variance along each axis:

$$X \leftarrow ED^{-\frac{1}{2}}E^{T}X$$

Where X is the centred input matrix, E is the eigenvectors matrix and D is the diagonal matrix of eigenvalues.

2.2.2 Decomposition

Once the data is preprocessed, comes the step of decomposition. This is the most crucial step where we have to select the correct algorithm, solver and parameters best suited for our analysis.

To ensure the best result accuracy, a stabilisation can be applied by performing multiple runs to correct for possible outliers as described for the ICASSO method (Himberg and Hyvarinen, 2003). This stabilisation can also be accompanied by a bootstrapping of the data for an additional stability correction of the estimated independent components.

While parameters for the stabilisation quality can be chosen “loosely”, a master parameter has to be chosen regarding the number of signals we want to extract using ICA. This is one of the most problematic steps since there exist many different methods to select it but no way to ensure that the chosen one is optimal. Among these methods, we can use Akaike information criterion (AIC) (Akaike, 1998), Bayesian information criterion (BIC) (Ben-Hur et al., 2002), cross validation (CV) (Wang, 2010) or more specifically the Most Stable Transcriptome Dimension (MSTD) (Kairov et al., 2017) when using transcriptomic data.

2.2.3 Component selection and usage

Once the decomposition has been performed, we can extract components of interest. To do so, we can either use the components from the matrix A or S. For an initial matrix X containing samples as rows and omics variables as columns each component of the matrix S contains a set of vectors called **metagenes** presented as weighted lists of genes usable for enrichment analyses to assign a biological function. It is also possible to use the mixing matrix A that contains a corresponding set of vectors called **metasamples** in the form of lists of weighted samples that can be statistically compared to clinical features if there are any (detailed in section **Biological content and specificity of the components** of (Cantini et al., 2019).

Once an independent component has been identified and given a biological significance, it is possible to use it in various ways.
Since ICA assigns scores to genes based on their driving quality for the component function, we can extract top contributing genes to use as markers of the corresponding functions or cell types in other analyses.
It is also possible to use components related to a particular clinical feature to assign a score that would inform on the state of each patient based on the given clinical status.
If components related to specific cell types are found, we can use them instead to quantify the cellular content of a given tumor sample (Decamps et al., 2021). After the extraction of components related to specific cell types, scores assigned by the method can be used as a basis to impute cell type proportions using the expression profiles observed in different samples (Figure 2.2).

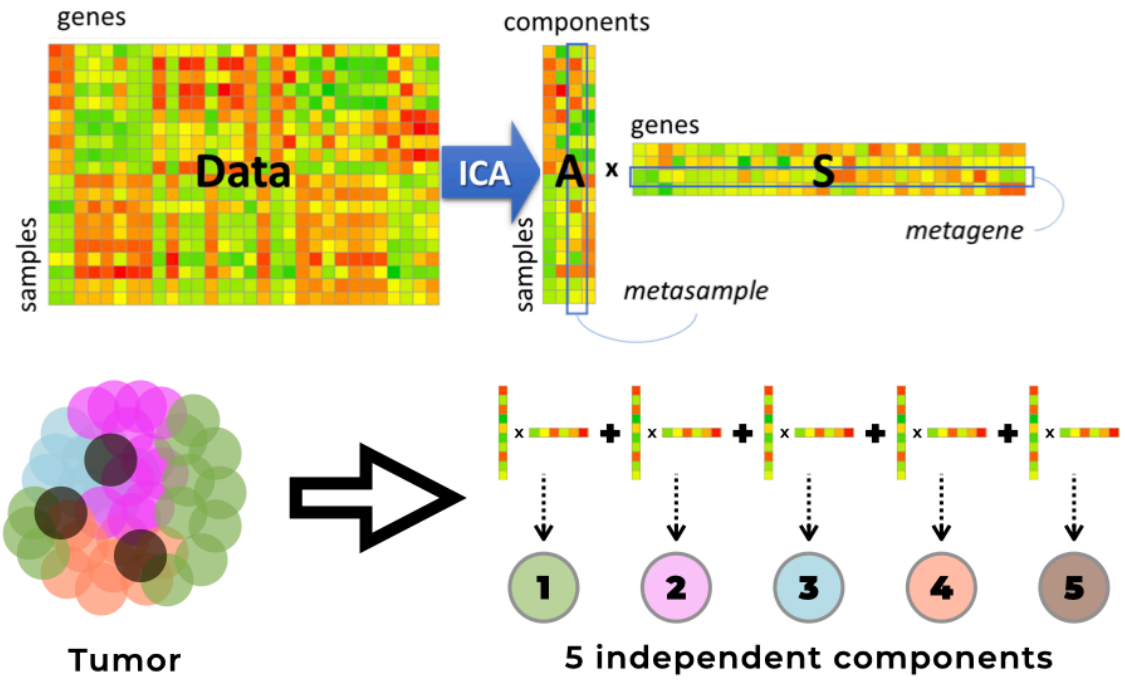


Figure 2.2. Illustration of a cell-type estimation using ICA.
Top: simplified illustration of how the input data is processed by ICA and what output is given as a result. From an initial matrix of samples containing a set of gene expression values, ICA is able to decompose the sources of different biological signals in the form of 2 matrices, the metasamples and metagenes. **Bottom**: a tumor can be visualised as a set of different cell types which contribute as sources of different signals in the data. Once these sources have been detected and separated, it becomes possible to interpret their origin and thus estimate the cellular content of the studied tumor.

2.3 Article: Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets

Nicolas Sompairac, Petr V. Nazarov, Urszula Czerwinska, Laura Cantini, Anne Biton, Askhat Molkenov, Zhaxybay Zhumadilov, Emmanuel Barillot, Francois Radvanyi, Alexander Gorban, Ulykbek Kairov and Andrei Zinovyev.

Published in the International Journal of Molecular Sciences, 7th of September 2019.

In the previous sections 2.1 and 2.2, I have described popular matrix factorisation methods and detailed the workflow applied to the particular method of Independent Component Analysis. In this article, I will review the different applications of ICA for cancer omics data, among which is the unsupervised cell type deconvolution. Finally, I will also mention some of the strengths and limitations of ICA as an unsupervised deconvolution method.

International Journal of
Molecular Sciences

MDPI

Review

Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets

Nicolas Sompairac [1,2,3,4], Petr V. Nazarov [5], Urszula Czerwinska [1,2,3], Laura Cantini [6], Anne Biton [7], Askhat Molkenov [8], Zhaxybay Zhumadilov [8,9], Emmanuel Barillot [1,2,3], Francois Radvanyi [1,10], Alexander Gorban [11,12], Ulykbek Kairov [8] and Andrei Zinovyev [1,2,3,*]

1 Institut Curie, PSL Research University, 75005 Paris, France
2 INSERM U900, 75248 Paris, France
3 CBIO-Centre for Computational Biology, Mines ParisTech, PSL Research University, 75006 Paris, France
4 Centre de Recherches Interdisciplinaires, Université Paris Descartes, 75004 Paris, France
5 Multiomics Data Science Research Group, Quantitative Biology Unit, Luxembourg Institute of Health (LIH), L-1445 Strassen, Luxembourg
6 Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, PSL Research University, 75005 Paris, France
7 Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI, USR 3756 Institut Pasteur et CNRS), 75015 Paris, France
8 Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, 010000 Nur-Sultan, Kazakhstan
9 University Medical Center, Nazarbayev University, 010000 Nur-Sultan, Kazakhstan
10 CNRS, UMR 144, 75248 Paris, France
11 Center for Mathematical Modeling, University of Leicester, Leicester LE1 7RH, UK
12 Lobachevsky University, 603022 Nizhny Novgorod, Russia
* Correspondence: andrei.zinovyev@curie.fr

Received: 3 August 2019; Accepted: 4 September 2019; Published: 7 September 2019

Abstract: Independent component analysis (ICA) is a matrix factorization approach where the signals captured by each individual matrix factors are optimized to become as mutually independent as possible. Initially suggested for solving source blind separation problems in various fields, ICA was shown to be successful in analyzing functional magnetic resonance imaging (fMRI) and other types of biomedical data. In the last twenty years, ICA became a part of the standard machine learning toolbox, together with other matrix factorization methods such as principal component analysis (PCA) and non-negative matrix factorization (NMF). Here, we review a number of recent works where ICA was shown to be a useful tool for unraveling the complexity of cancer biology from the analysis of different types of omics data, mainly collected for tumoral samples. Such works highlight the use of ICA in dimensionality reduction, deconvolution, data pre-processing, meta-analysis, and others applied to different data types (transcriptome, methylome, proteome, single-cell data). We particularly focus on the technical aspects of ICA application in omics studies such as using different protocols, determining the optimal number of components, assessing and improving reproducibility of the ICA results, and comparison with other popular matrix factorization techniques. We discuss the emerging ICA applications to the integrative analysis of multi-level omics datasets and introduce a conceptual view on ICA as a tool for defining functional subsystems of a complex biological system and their interactions under various conditions. Our review is accompanied by a Jupyter notebook which illustrates the discussed concepts and provides a practical tool for applying ICA to the analysis of cancer omics datasets.

Keywords: independent component analysis; cancer; omics data; dimension reduction; data analysis; data integration

1. Introduction

Cancer research is one of the most important providers of large-scale molecular profiling data, which help in understanding not only the state of human cells in disease but also shed light on the normal physiological processes measurable and detectable in various kinds of omics datasets. Determining robust and biologically meaningful ways of quantifying cellular and organismal and normal and pathological physiology using high-throughput molecular data remains a major challenge (making biology a quantitative science). Different kinds of biological processes leave characteristic traces at different levels of genome-wide measurements depending on their nature and timescales: some significantly affect transcriptomes, some rather modify DNA methylation programs or mutational spectrum, others are measurable only at the level of proteome and phosphoproteome. In order to reliably quantify some of these biological mechanisms, one will need to design multi-omics signatures spanning several levels of molecular data descriptions. On top of this, various technical factors interplay with biological ones, frequently in a way which makes it difficult to clearly distinguish both.

Rarely does molecular data "speak for themselves": they need to be properly pre-processed, analyzed in the light of mathematical modeling, statistical assumptions, and prior biological knowledge and, finally, should be represented at some pre-defined level of abstraction. In this sense, one of the simplest paradigms of *linear mixture of signals* plays a pivotal role in the modern molecular data analysis. In this framework, one assumes that a measurable elementary quantity such as expression of a single gene is a result of weighted summation of some latent, and not always directly observable, factor activities which should have associated numerical values. The nature, the number of factors and the way they are represented numerically can be known or unknown in advance. A toolbox of existing mathematical approaches provides concrete scenarios in which the additive factors can be determined and quantified, under acceptance of certain assumptions about the statistical properties of their numerical values or the weights connecting them to the measurements.

One of the standard methods in such a toolbox is independent component analysis (ICA) having a long standing history of application to biological data, including the analysis of molecular profiles (mainly, transcriptomic). Formally, ICA belongs to a family of methods called matrix factorizations (Figure 1), the most popular other representatives of which are principal component analysis (PCA) or the very similar singular value decomposition (SVD), and non-negative matrix factorization (NMF).

The first applications of ICA in biology contrasted it to PCA and standard clustering methods and found that the factors determined through ICA are easier to interpret biologically [1,2]. This raised an increase in interest of ICA and its applications in various contexts, and, in particular, in cancer biology [3,4]. The success of ICA can be connected to the nature of the statistical assumptions which are used to define the method, that match well the underlying high-dimensional distributions of omics datasets. The principles of ICA are briefly introduced in Section 2.1.

Independent component analysis and matrix factorization approaches are standard methods in the rapidly growing arsenal of machine learning methods applied to the molecular biology and medical data. At the same time, remarkable success has recently been achieved in applying deep learning techniques in certain fields of cancer biology such as clinical imaging of various kinds [5–9]. Deep learning has been successfully used in automating the diagnosis and prognosis of several cancer types, claiming to be competitive with human pathologists [10,11]. Successful applications of deep learning methods to multi-omics data have been recently reported, such as in Reference [12]. One should also notice that there exists a certain level of controversy in assessing the actual success of this rapidly growing area [13] and an important methodological discussion on the "deep" versus "shallow" methods in real applications [14]. Reviewing any statistical method today should necessarily take into account the existing intrinsic competition between this relatively recent trend and more "classical" areas of machine learning, even though many of them, including ICA, are rooted in the artificial neural network theory [15].

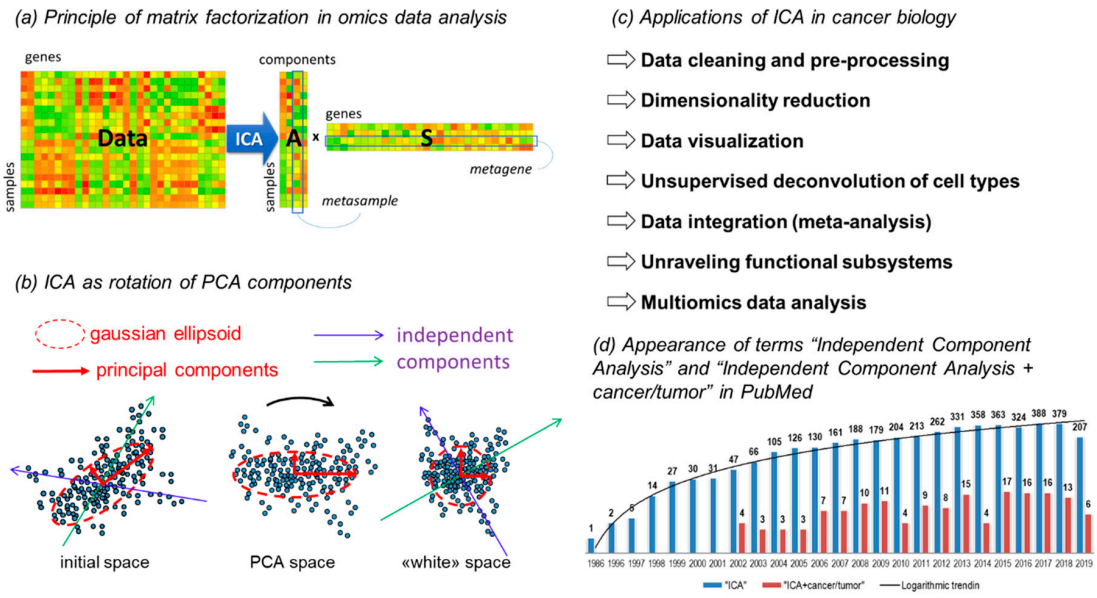


Figure 1. Independent component analysis (ICA) is a standard tool for reducing the complexity of omics datasets in cancer biology. (**a**) ICA belongs to the family of matrix factorization methods, approximating a 2D matrix by a product of two much smaller matrices, containing metagenes and metasamples, in the case of omics data. (**b**) ICA can be considered as a rotation of PCA axes, after data "whitening" (i.e., orienting the Gaussian ellipsoid along the coordinate axes and scaling them to unit variance). (**c**) The major types of applications of ICA in cancer biology. (**d**) The number of publications in PubMed mentioning ICA and the number of publications simultaneously mentioning ICA and "tumor" or "cancer".

Over the last decade, significant experience in applying ICA to different kinds of omics data for addressing various problems has been obtained, including data pre-processing, task of cell type deconvolution, and meta-analysis of multiple omics datasets (Figure 1c). In this paper, we reviewed most of the recent achievements in computational cancer biology research where ICA was used as the main data analysis tool. We also discussed the practices of ICA applications which appeared to be successful in various contexts.

This review is accompanied by interactive Jupyter notebook located at https://github.com/sysbio-curie/ICA-in-Cancer-research-review-materials.

2. Methodology of ICA Application to Cancer Omics Data

2.1. Brief Introduction into Matrix Factorization Applied to Omics Data

Independent component analysis belongs to a family of matrix factorization methods. Each of these methods takes a rectangular matrix $\boldsymbol{X} \in \mathbb{R}^N_m$ of measurements (in sufficiently a large number of observed samples, N, and with number of observed features, m) as an input and approximates it as a sum of products of p pairs of vectors of size N and m. The fundamental equation for all matrix factorization methods states (note that the product of $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vectors gives a one-rank matrix of the same dimension as $\boldsymbol{X}$):

$$\boldsymbol{X} \approx \sum_{k=1}^{p} \boldsymbol{a_k} \times \boldsymbol{s_k} \ (*) \tag{1}$$

and the problem of matrix factorization is to find a set of $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ such that:

$$||\boldsymbol{X} - \sum_{k=1}^{p} \boldsymbol{a_k} \times \boldsymbol{s_k}||^2 \rightarrow \boldsymbol{min} \ (**) \tag{2}$$

where ||..|| is a suitable matrix norm which is most frequently the sum of the Euclidean norms of the columns of the matrix.

Each vector pair $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ will be called a component throughout this review. Therefore, a component is represented by a vector $\boldsymbol{s_k}$ of size m containing weights of omics variables (genes, proteins, CpG sites, etc.). At the same time a component is associated to a vector $\boldsymbol{a_k}$ of size N, containing contributions of the component to measured samples. We will use these notations and meaning of $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vectors throughout the whole review.

In the matrix factorization literature, various terms are used to denote the elements of the vectors $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$. For example, the terms "loadings", "activations", "factor strength" or "sample-associated weights" have been used to denote the elements of $\boldsymbol{a_k}$ vectors. The matrix composed from the $\boldsymbol{a_k}$ vectors is sometimes called the "mixing matrix" and denoted as $\boldsymbol{A}$. The elements of $\boldsymbol{s_k}$ vectors have been called "weights of the component" or "signals" and the matrix composed of them (denoted as $\boldsymbol{S}$) is sometimes called the "signal matrix". Moreover, $\boldsymbol{s_k}$ vectors themselves are frequently referred to as "components" or "factors".

In the context of transcriptomic data analysis, the $\mathbf{s_k}$ vector is frequently named a metagene [16]; in the case of other data types one can use similar naming, e.g., a metaCpG for the analysis of DNA methylation profiles. Further we will use the term metagene (or metagene weights for the individual elements) to refer to vector $\mathbf{s_k}$ even when describing application of ICA to various data types. Similarly, the $\mathbf{a_k}$ vectors are sometimes called metasamples, and we will adopt this term in the text (referring to the individual vector elements as metasample weights), see Figure 1a.

Intuitively, a transcriptome of a biological sample is described as a combined action of p metagenes. Each metagene abstractly represents a molecular program (called a functional subsystem further in the text) by assigning a numerical weight to each gene of the organismal genome. The activity of metagenes in a sample is combined additively, and each metagene acts on a sample with a sample-specific strength or activity. Activities of the same metagene over all measured samples is called a metasample. A metasample is the profile of the corresponding metagene activity similarly to a gene expression profile across samples.

In the equation (*), only the $\boldsymbol{X}$ matrix is known; the $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vectors are unknown. As such, the problem of matrix factorization (**) is heavily underdetermined, and additional constraints need to be introduced on $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vectors in order to find its solution. First of all, it can be required that the all $\boldsymbol{a_k}$ vectors would have length one.

Furthermore, one can require orthogonality of the $\boldsymbol{a_k}$ vectors: $(\boldsymbol{a_i},\boldsymbol{a_j}) = 0$, for $i \neq j$ and that the solution of (**) should give the same result for different orders of matrix decomposition p, i.e., $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vectors computed for the order $p = p'$ would be the same as for the decomposition of order $p'' > p'$. In this case, solving (**) is equivalent to computing the singular value decomposition (SVD) of $\boldsymbol{X}$ and gives a set of principal components. There exist several ways to introduce PCA, as reviewed in Reference [17].

Alternatively, one can require that all elements of $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vectors would be non-negative. This constrains the problem (**) and leads to NMF. The simplest approach to solve (**) with these constraints is to repetitively apply the non-negative least squares regression method, considering $\boldsymbol{a_k}$ as unknown at one iteration and $\boldsymbol{s_k}$ as unknown at the next iteration, until convergence to a local minimum.

When computing ICA, the resulting components are required to be *as mutually independent as possible*. More precisely, the elements of vectors $\boldsymbol{s_k}$ (or sometimes, vectors $\boldsymbol{a_k}$) have to represent maximally mutually independent distributions, for different k. The perfect independence would mean that the joint probability distribution $P(s_1, \ldots, s_p)$ can be factorized as $P(s_1, \ldots, s_p) = P_1(s_1) \times P_2(s_2) \ldots \times P_p(s_p)$. Here, one assumes that the elements of vectors $\boldsymbol{s_k}$ are i.i.d. samples of the underlying probability distributions $P_k(s_k)$.

From the different nature of the constraints follow different properties of matrix factorization algorithms (see Reference [18] and Figure 2a). The PCA solves a convex quadratic optimization problem, which has a unique global minimum. The principal components are orthogonal and can be naturally ranked by the amount of explained variance. The NMF and ICA problems are not convex;

therefore, the algorithms used to solve the optimization problem provide solutions depending on the component initialization. By construction, NMF and ICA do not lead to an orthogonal set of $\boldsymbol{a_k}$ vectors and the components cannot be naturally ranked. The NMF components contain only non-negative elements, which makes the intuitive picture of the additive action of metagenes simpler to interpret, while in PCA and ICA some metagenes can cancel the action of other metagenes if they are summed up with different signs.

2.2. ICA Algorithms

One of the historically first and still popular practical algorithms for solving ICA problem is based on the general Infomax (or maximum entropy) principle [19]. Indeed, the problem of ICA consists in minimizing the mutual information among individual components (represented by finite $\boldsymbol{s_k}$ vectors). It can be shown that maximizing entropy of joint distributions of pairs of $\boldsymbol{s_k}$ leads to minimizing their mutual information.

It appeared also that under some assumptions, minimizing the mutual information is equivalent to maximizing the non-Gaussianity of the individual $\boldsymbol{s_k}$ distributions [20]. Quantification of non-Gaussianity for continuous distributions involves negentropy (or Gibbs free energy, in physics). Negentropy measures the departure from Gaussianity of a random vector of density $P(u)$ by comparing its entropy to the entropy of a normal distribution with same mean and variance. The entropy is defined with a negative sign ($S = -\int P(u) log P(u) du$) and the negentropy is, therefore, a non-negative function reaching zero only for the standardized normal distribution. For the mathematical details, we refer the reader to the classical works [19,20].

Since the length of the $\boldsymbol{s_k}$ or $\boldsymbol{a_k}$ vectors is always finite in real-life applications, one needs to introduce the way to effectively approximate it from the finite samples. For this purpose, various surrogate functions (called non-linearity functions) have been proposed, one the most popular of which involving the kurtosis. Empirically, kurtosis was found to be an appropriate choice of non-linearity in the analysis of transcriptomic data. Other types of non-linearity functions have been suggested; however, the appropriate choice of non-linearity for applying ICA to different kinds of omics measurements remains an open question. The two most popular ICA algorithms based on non-Gaussianity maximization are fastICA [20] and joint approximation diagonalization of Eigen-matrices (JADE) [21]. Most of the recent applications of ICA to omics data were based on fastICA, utilizing approximate Newton iterations to optimize a non-Gaussianity measure. However, other approaches to computing independent components have been used such as the product density estimation-based method (ProDenICA), claimed to have higher sensitivity to a wider range of source distributions than fastICA [22,23].

A typical preprocessing step used before application of ICA algorithms is the so-called data whitening or sphering (see Figure 1b). Whitening imposes unit variance along each axis. It consists of choosing a number of significant principal components, thus defining the resulting number of factors and then rotating the data to the basis of principal Gaussian ellipsoid axes and scaling along the principal axes to the unit variance. In the geometrical language, the Mahalanobis metrics are introduced into the data space instead of the usual Euclidean. Therefore, after whitening, the covariance matrix of the reduced dataset becomes the identity matrix and PCA becomes inapplicable, since all Gaussian signals have been erased from the data. This makes the use of higher-order moments for finding a rotation of the orthonormal coordinate basis easier, which would maximize the non-Gaussianity of the data point projection distributions along each axis. After such a rotation, in the whitened space, the vectors corresponding to the new axes remain orthogonal while in the original data space they can be strongly correlated (see Figure 2a). Because of the use of whitening as a preprocessing procedure, ICA is frequently considered as a step on top of PCA, consisting in rotating the coordinate system, by exploiting the information contained in higher than second moments of the multivariate data distribution (Figure 1c).

Various flavors of ICA have been suggested and some of them were tried on omics data. Bayesian ICA with prior constraints have been suggested and tried on the metabolomics data [24]. The prior constraints can be non-negativity of the $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ vector elements. This allows combining the nice properties of non-negative mixture problem and the requirement for mutual independence of the components. A kernel version of ICA have been developed [25] and sparse ICA was proposed in Reference [26], but both have not yet found wide applications in omics data analysis (though kernel ICA was exploited in Reference [27]). Finally, tensorial ICA was recently developed in References [28,29] and recently applied to the joint analysis of gene expression, copy number changes, and DNA methylation data from colon cancer with some promising results (see more in Section 3.5).

Some flavors of ICA seems interesting to explore more in biological applications, in the view of the concept of the integration of functional subsystems (see Section 3.6), such as tree-dependent component analysis (TCA) [30]. This variant of ICA allows clustering of the components such that they remain independent between the clusters and dependent within them. It was tested on fMRI data [31], but not yet on large-scale omics datasets.

2.3. Various Ways to Apply ICA to Omics Data

Besides the choice of ICA algorithm (which is frequently fastICA), there are several choices to be made when ICA is applied to omics datasets.

The first evident but non-trivial choice concerns a necessity for data log-transformation, which is especially important in the case of gene expression and protein expression data. On one hand, it is strongly desirable in the case of, for example, RNA-Seq data, since empirically they are found to be characterized by log-normal distribution. When ICA is applied to non-transformed data, the resulting components are frequently dominated by single genes or single samples (e.g., each sample acts as an independent component), which contradicts the initial concept of linear mixture (nothing or almost nothing is mixed in this case). Simple log-transformation usually fixes this issue. However, log-transformation makes the direct interpretation of the ICA model difficult, since, formally speaking, one deals with a multiplicative rather than an additive model of signal mixture. This is particularly important for the applications of ICA in the field of cell type deconvolution where the linearity assumption is explicitly made for mixing transcriptomes of different cell types (see Reference [32] which cites a number of references studying the issue of data log transformation). Another aspect is that log-transformation can amplify small values, sometimes creating a heavy tail of negative values, characterized by strong non-Gaussianity and affecting the ICA determination. In practice, log-transformation can be recommended after adding a small value (e.g., 1 sequence count) to all data matrix entries, before taking the log. This is especially true in the case of sparse single cell RNA-Seq data, where the majority of matrix entries can equal to zero. On the other hand, choosing a threshold for small expression values looks like an arbitrary choice, especially if the RNA-Seq data have been normalized beforehand. Despite these difficulties, in most of the applications of ICA to RNA-Seq data analysis, the so called "log(x+1)" transformation can be advised: empirically, it is found to lead to more stable and biologically interpretable components. The problem of log transformation became more relevant after introducing sequencing technologies such as RNA-Seq; for microarray-based methods, the gene expression measurements were frequently provided in log scale, after some standard normalizations such as robust multichip average.

Another choice in applying ICA to a matrix of omics measurements is the choice between what distribution independence (or non-Gaussianity) is maximized [18]. One can maximize the independence of metagenes (vectors $\boldsymbol{s_k}$) or metasamples (vectors $\boldsymbol{a_k}$). Technically, the first case corresponds to the application of ICA algorithm to the initial matrix $\boldsymbol{X}$ containing samples as rows and omics variables as columns, and the second case corresponds to the application of ICA to the transposed matrix $\boldsymbol{X}$. Surprisingly, both ways of applying ICA to omics data are wide-spread, and sometimes it makes an effort to figure out in which way ICA was applied. Some studies aim at maximizing the non-Gaussianity of metagenes [2,33–35], while others maximize non-Gaussianity of

metasamples [36,37]. Empirically it was shown that maximizing the non-Gaussianity of metagenes is clearly preferable in gene expression analyses to maximizing the non-Gaussianity of metasamples [38]. This choice leads to much better reproducibility of metagenes in independent datasets as well as to better interpretability of the components computed within the same dataset.

Furthermore, in several studies it was found that stabilized or consensus independent components have better characteristics in terms of generalization and interpretation [34,38–41]. By stabilization one usually means re-computing ICA using multiple random initialization with subsequent clustering of the resulting components [40,41]. Alternatively, stabilization can be performed through sub-sampling, i.e., computing ICA multiple times after removing a certain percentage of samples. Applying stabilization can characterize computed independent components in terms of their stability that can be further used for ranking them. For example, it was demonstrated that such ranking is usually more meaningful in the case of transcriptomic data analysis compared to other methods of component ranking (e.g., by the measure of non-Gaussianity or by the explained variance) [34]. One of the first and most popular approaches to ICA stabilization is the *icasso* method, introduced by the creators of fastICA [41]. Interestingly, in the case of transcriptomics data, the most stable independent component frequently strongly correlates with the first principal component.

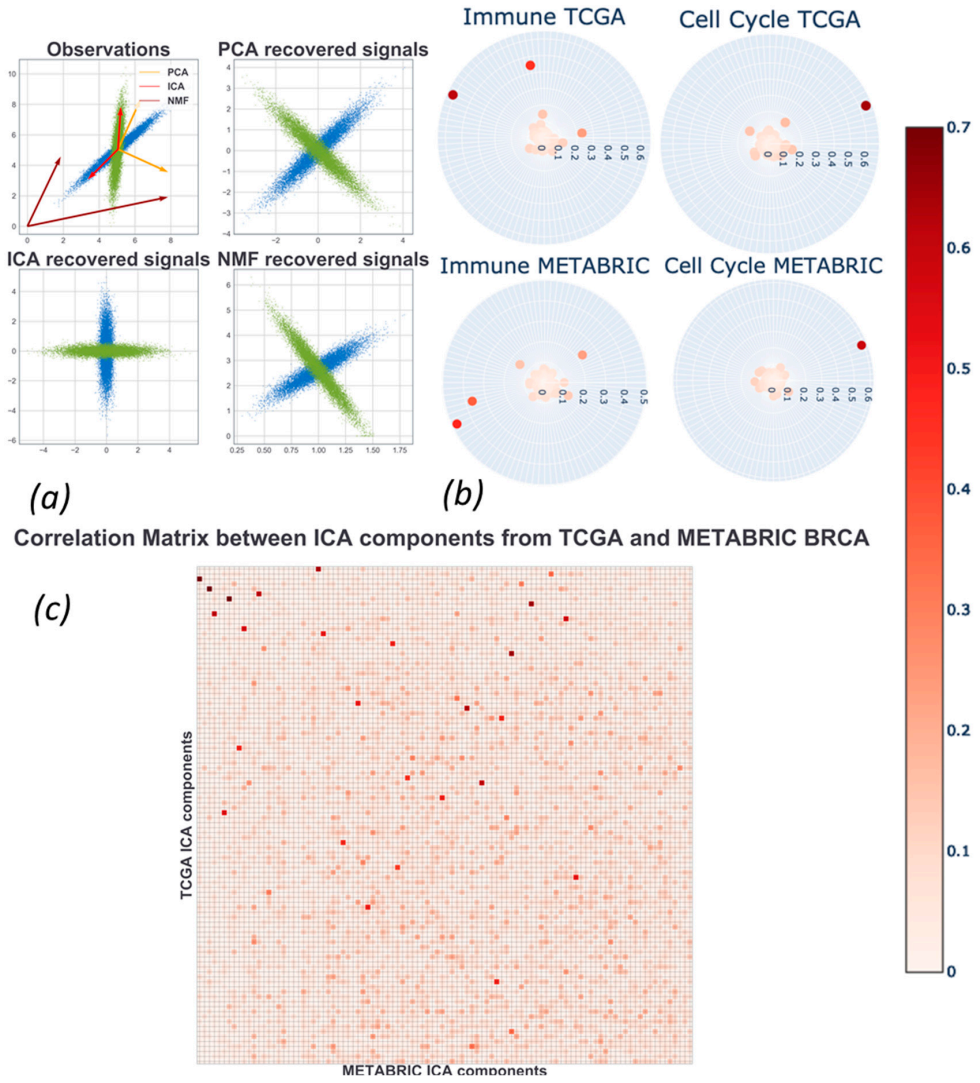


Figure 2. Features of ICA applied to a synthetic (**a**) and two real-life datasets (breast cancer The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) transcriptomic datasets) (**b**,**c**). (**a**) Independent Component Analysis is able to disentangle

(or deconvolute) two intersecting Gaussian distributions with coinciding means and whose principal axes form a sharp angle; (**b**) 100 order ICA decomposition of the TCGA and METABRIC datasets. Each component represented as a metagene was correlated to either immune infiltration-related or proliferation-related meta-metagenes derived from Reference [33]. This analysis shows that only one of the components was strongly correlated to the cell-cycle, while several can be associated with the presence of an immune-infiltrated ICA-derived signature (this, probably, signifies the ability of ICA to deconvolute the major immune cell types in an unsupervised manner (see, Reference [42]); (**c**) correlations matrix between the metagenes of independent components extracted from the TCGA and METABRIC separately. It shows that, for some components computed for different datasets, there exists a strong and unique association between them, indicating the high reproducibility of the ICA results (e.g., see Reference [38]).

Lastly, in some applications of ICA (e.g., cell-type deconvolution), it is desirable to fix the orientation of the independent components. We remind that in PCA and ICA, the signs of the elements in the vectors $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$ can be inverted simultaneously without changing the definition of the component. Some methods (such as BIODICA or DeconICA) avoid this ambiguity by assuming that the heaviest tail of the $\boldsymbol{s_k}$ distribution should correspond to positive values, which usually gives satisfactory results. In Reference [43], each ICA component was characterized by two sets of top contributing genes, from the negative and the positive side of the metagene weight distribution. The largest such set was called a dominating module and the final orientation of the component was chosen to make the weights of the dominating module positive. In other cases, labeling of samples can be used in order to select one of the two possible signs of $\boldsymbol{a_k}$ and $\boldsymbol{s_k}$. In this case, the orientation was chosen based on the values of $\boldsymbol{a_k}$ vectors. For example, in a disease study, one can require that any component would be oriented towards aggravation of the disease condition (e.g., from normal samples to more aggressive cancer stages). This approach was recently used for quantifying disease comorbidity using ICA [44].

2.4. Assessment and Comparison with Other Matrix Factorization Methods

In several recent studies, ICA was systematically compared with the other most used matrix factorization methods such as PCA and NMF, using large collections of cancer omics measurements.

In Reference [38], it was tested which matrix factorization method could produce the most reproducible (i.e., generalizable) definitions of metagenes. In order to achieve this, a notion of a reciprocal best hit (RBH) graph was borrowed from evolutionary bioinformatics. Reciprocal best hit between two metagenes in two ICA decompositions of different datasets defined "orthologous" metagenes. Several criteria have been used in order to evaluate the modular structure of the RBH graphs resulting from application of various ways of applying ICA, PCA, and NMF to the transcriptomic data. In particular, the total number of RBH relations among the components, average clustering coefficient and modularity of the RBH graph, and the number and typical sizes of the identified graph communities have been assessed. The conclusion was that the stabilized version of ICA, where the non-Gaussianity of metagenes (and not metasamples) was maximized, is superior to other matrix factorization methods with respect to these measurements.

Three major matrix factorization approaches were systematically discussed in a recent review for their ability to discover functional subsystems or tissue-type specific signals [45]. The main conclusion was that it might be advantageous to use several matrix factorizations simultaneously. The same authors further suggested using the BioBombe approach [46], where three matrix factorization methods (PCA, ICA, and NMF) and two autoencoder-based dimension reduction techniques were systematically compared based on the pancancer TCGA datasets comprising 11,069 tumoral samples. Indeed, each data decomposition method showed its own advantages with respect to different tests and tasks. For example, the ICA method outperformed other approaches when the extracted metagenes were tested for gene set coverage of specific gene set collections representing transcription factor targets, Reactome pathways, and cancer modules. Higher gene set coverage in this study meant the proportion

of gene sets in a reference collection, which could be significantly associated with one of the metagenes in the decomposition.

2.5. Estimating the Number of Independent Components

The most important parameter in the application of any matrix factorization method is the number of components to determine. This question is less crucial in the case of PCA due to the orthogonality constraint and that computing higher-order components does not affect the definition of the lower-order ones. However, this is not the case with ICA and NMF: choosing the order of decomposition affects the definition of *all* computed components. In the case of ICA, which geometrically only rotates the PCA axes, choosing the number of independent components can rely either on the methods for determining the number of relevant principal components or it can use some features of the independent components themselves in order to determine the optimal decomposition order.

In the first case, the effective global dimensionality of the data can be determined through the standard Kaiser rule, use of broken stick distribution, Horn's parallel analysis or estimating the conditional number of the covariance matrix [47]. One can also use more advanced methods for determining the effective data dimensionality such as the ones using concentration of measure phenomena [48] or data point cloud linear separability statistics [49].

However, the second case appears to be more consistent in applications, even being computationally more challenging. Thus, in Reference [24], Bayesian information criterion (BIC) was exploited to determine the optimal number of independent components in the analysis of a metabolome dataset comprising 1764 samples and 218 measured metabolites. The optimal number of components according to this estimate appeared to be quite small (eight).

In Reference [34] stability indices of independent components were used in order to define so-called maximally stable transcriptomic dimensionality (MSTD) measure, in case of transcriptomic data. The MSTD defines an order of transcriptomic matrix decomposition such that the distribution of stability indices for independent components is not yet dominated by highly unstable ones. It was demonstrated that the independent components within the MSTD range are characterized by better reproducibility and interpretability. Based on the analysis of a large volume of cancer transcriptomic data, several observations were made. Firstly, unstable higher order components are frequently driven by very few (frequently, only one) genes. In other words, their $\boldsymbol{s_k}$ distributions are characterized by the presence of one or few weights with exceptionally large values, separated by a significant gap from the other values. Secondly, it was shown that a certain level of *over-decomposition* of transcriptomic datasets, i.e., choosing the number of independent components several times larger than MSTD, does not drastically change the definition of most of the components within the MSTD range. At the same time, it was observed that increasing the number of independent components over the MSTD value sometimes leads to biologically meaningful splitting of the components. For example, a component within the MSTD range which was associated with the total level of immune infiltrate in tumoral microenvironment splits into three components in higher-order decompositions which can be associated with the presence of T-cells, B-cells, and macrophages [34,42].

In Reference [46], a range of decomposition orders have been tested using various criteria for several matrix factorization methods. The general conclusion was that it can be advantageous to use multiple-order decompositions if the aim is signature discovery. Just as in Reference [34], it was shown that higher-order matrix factorizations with at least 40–50 components provide more precise interpretation with respect to associating the components to the clinical information such as patient gender or to the mutation status of cancer driver genes.

2.6. Methods for Interpretation of Independent Components

Assigning a meaning to the extracted independent factors remains a major problem in exploiting ICA in biological research. Standard practice consists of applying various kinds of functional enrichment analyses to $\boldsymbol{s_k}$ vectors (e.g., applying hypergeometric test or overrepresentation analysis (Webgestalt

2017) to the set of most contributing genes, or Gene Set Enrichment Analysis to the whole ranking defined by s_k), using large-scale collections of reference gene sets. The distribution of gene weights from s_k vectors can be projected on top of genome-wide biological network reconstructions where the network edges represent different types of interactions or regulations between genes and/or proteins. This can be further used for various types of network-based analyses, leading to the determination of biological network "hotspot" areas and eliminating the need of having a reference gene set collection [50]. The s_k vectors (resulting from the analysis of transcriptomic or methylome data) can be projected onto genome and be a subject of peak-calling analysis, which can sometimes lead to associating a component to genomic alterations [33].

Metasample weights a_k are used to associate components to sample annotations such as clinical data (tumor stage, molecular classification, time label, sample processing data, etc.). Metasamples can be also associated with some clinically relevant molecular data, such as mutations in known cancer drivers for a particular cancer type. Metasamples can be also associated with known labels for molecular tumor subtype.

In parallel to rigorous statistical testing, insightful visualizations of the results of ICA application can be of great help. For example, gene weights from s_k vector can be projected on cancer-specific biological network maps such as the Atlas of Cancer Signaling Network (ACSN) using user-friendly Google Maps-based online platforms such as NaviCell and MINERVA [51,52]. Functional enrichment analysis results of ICA metagenes can be visualized using maps representing functional redundancy between reference gene sets, such as InfoSigMap or enrichment maps [53].

There exist integrated solutions allowing the computation of ICA components for omics datasets and containing a built-in set of tools for their interpretation. For example, in the BIODICA package (Available online: https://github.com/LabBandSB/BIODICA), a set of tools is provided for performing hypergeometric tests of the metagenes, automated feeding of Gene Set Enrichment Analysis with ICA metagenes, projecting metagenes onto biological network maps, correlating computed metagenes with a reference database of previously annotated metagenes, associating components with categorical and numerical sample annotations, and tools for meta-analysis of ICA decompositions.

A particular interest represents joint analysis of omics profiles together with histopathological imaging data. A simple analysis was made in Reference [33], where the independent components computed from the transcriptomic data were used to rank the matched histopathological images according to the contribution of the corresponding tumor sample to the component. This simple approach was used in order to confirm the biological meaning of some of the components (see Figure 3). Today this approach can be further elaborated and automated by applying machine learning-based methods for extracting features from medical images and correlating them to the patterns identified from the omics data (such as ICA metagenes), which can lead to getting new insights into cancer biology [54].

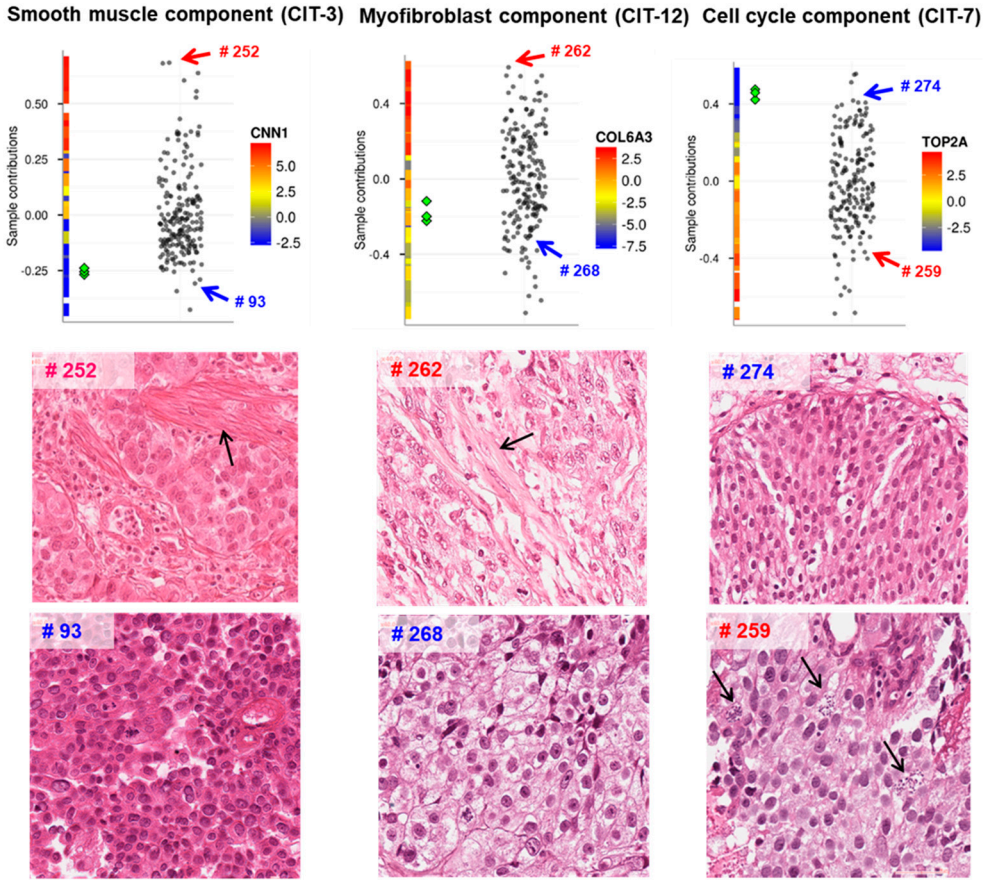


Figure 3. Interpretation of ICA components using histopathology imaging of bladder tumor cross-sections. Each metasample produced by ICA defined a ranking, which was used to sort the images. Visual inspection determines a clear trend in the images towards the increase of certain elements (presence of smooth muscle cells, myofibroblasts (cancer-associated fibroblasts), dividing cells). Two example images per component selected from the top and the bottom of the rankings are shown here. Green rhombuses designate normal samples. Black circles designate cells of interest: muscle cell (left), myofibroblast (middle), cells in mitosis (right). The figure is reproduced from the Supplementary Materials of Reference [33] with permission.

3. Applications of ICA in Cancer Research

3.1. Applications to Data Preprocessing, Classification, Dimensionality Reduction, and Clustering

In multiple studies, ICA was shown to be efficient in disentangling biological and technical factors affecting molecular profiles. This supports the idea to use ICA as a powerful data preprocessing and/or feature engineering method for further application of machine learning methods. The general approach is to apply ICA as an unsupervised machine learning method, to decide on the biological meaning or the technical origin of individual components and then focus on a subset of them containing the relevant signal. This can be achieved either by directly using the relevant subset of components as features or by constructing a modified matrix of molecular measurements which would be free of the influence of those components which are identified as non-relevant or of technical origin.

Frequently, each one-dimensional s_k distribution is analyzed for determining a set of the most contributing genes (e.g., characterized by the most extreme absolute values in s_k). The simplest idea is to select the variables (e.g., genes) bypassing the threshold in p standard deviations, with some choice for p (typically, $p \geq 3$). A combined set of the most contributing to different ICs genes can be used to define a subset of data for further analysis.

Interestingly, ICA decomposition can be used to identify and disregard technical biases among omics datasets produced by different platforms. For example, in the study of 198 bladder cancers in Reference [33], one of the most stable components was found to be associated with a complex time-dependent batch effect. The nature of this batch was not known in advance and was only discovered by correlating the corresponding $\boldsymbol{a_k}$ vector to the dates of sample preparation. Another component frequently identified in the analysis of transcriptomic data is related to GC-content, which might reflect the influence of GC-content on the RNA amplification step common for both microarray-based and sequencing-based methodologies. In Reference [39], a small dataset of three primary melanoma tumors and two matched controls, characterized at the level of transcriptome and miRNA, were merged together with a large reference melanoma dataset from the Cancer Genome Atlas. The ICA decomposition was performed for the merged transcriptomic and miRNA data separately. For both molecular data types, it was possible to identify those independent components capturing technical differences among platforms while focusing the analysis on biologically meaningful factors whose quantification was comparable among platforms.

Interestingly, ICA-based analysis sometimes can lead to identification of the factors whose origin is intermediate between technical and biological. For example, in Reference [33] one of the factors reproducible in several bladder cancer datasets was strongly associated with the surgery type (transurethral resection of the bladder tumor versus cystectomy) and at the same time was enriched with early response genes. This suggests that different ways of tissue processing might leave characteristic patterns in the transcriptome which can be discovered using ICA.

Some components identified through ICA could describe various cell populations present in the sample in addition to cells of direct interest. Typically, this was the case for the stroma-related signals in the ICA-based analysis of tumor bulk samples (see Figure 2). ICA can efficiently deconvolute the contribution from the cells of different types to the bulk transcriptome, which allows studying the properties of tumor cells more directly. In the aforementioned study of bladder cancer, decomposition of bulk tumors into 20 components allowed for the clear distinction of the signals reflecting the presence of immune cells (with the main signal coming from the multiple types of lymphocytes, adipocytes, fibroblasts [33].

Another frequently employed idea is to use the results of ICA decomposition in order to define a set of variables for further application of various machine learning methods. Zhang et al. [55] were among the first who applied ICA as a data-preprocessing step for classification of cancer patients. They used ICA independently on normal and cancer datasets and identified top gene markers able to discriminate between these conditions. Their approach was quite indirect but showed the ability of ICA to prioritize genes. In a study by Huang et al. [56], ICA was followed by a penalized discriminant method, and the authors showed high accuracy of ICA-based approach on several datasets. In both mentioned papers, the authors segregated cancer and normal tissues, which is now considered a trivial task, taking into account the large effect of cancer on cell transcriptome. Later, Zheng et al. [57] proposed a consensus ICA, robust to initial estimations. They showed the applicability of the approach on three datasets, in two of which they classified subtypes of tumors. Support vector machine (SVM) was used to predict classes based on the metasamples, and the authors needed to perform preliminary feature selection to improve their classification accuracy.

Recently ICA was used to engineer features for further use in cancer-related classification tasks, using naïve Bayes classifier [58]. In Reference [59], ICA was used as a data pre-processing step in order to improve the clustering of temporal RNA-Seq data. It was suggested to use ICA in combination with wavelet-based data transformation in order to engineer transcriptomic features at "multiple resolution" [60] and use them to improve tumor classification and biomarker discovery. In Reference [22], it was shown that a set of 139 features built by systematically applying ICA to a large cohort of transcriptomic profiles, can be directly used in machine learning for classification tasks and have advantageous characteristics in small sample studies, compared to the classical differential

expression-based feature selection. It was noticed also that using ICA-based features reduced to some extent the batch effects when clustering the transcriptomic data.

Any matrix factorization method can be used for dimensionality reduction. The specifics of ICA are in that it is usually performed in an already reduced space and only defines a new coordinate basis in the principal linear manifold. Therefore, ICA itself does not reduce the data dimension more that it is done by PCA. Nevertheless, it is a frequent practice to consider the coordinate basis defined by few independent components as a subspace to further application of various data analyses. For example, this approach is used for a standard pipeline of single cell RNA-Seq data analysis [61]. Similar notice can be made with respect to using ICA as a data visualization tool. Selecting a couple of independent components with clearly identified biological meaning can lead to a biologically meaningful 2D data display. For example, in Reference [62], visualizing a single cell dataset in the plane of two independent components associated with proliferative genes clearly revealed the 2D dynamics of tumor cell progression through the cell cycle. The difference with PCA-based data visualization is that, in the case of ICA, there exists no principal pair of components (such as PC1 and PC2) which can be considered as the most representative for visualizing the multi-dimensional distances. This remark should be taken with care since, frequently, the first PCs are affected by technical artifacts and are to be neglected in further analysis. In the case of ICA, any pair of ICs in no particular order can be used for data visualization taking into account their tentative interpretation. Examples of contrasting PCA and ICA approaches for data visualization can be found in References [37].

3.2. ICA for Unraveling Functional Subsystems of a Living Cell or a Cell Ecosystem

One of the strongest motivations behind applications of ICA to omics data is in that it can help identifying functional subsystems (or functional modules and complex biological processes) which are the building blocks determining response to perturbation of a tumoral cell or a whole cellular ecosystem such as tumor microenvironment (TME) composed of different cell types. The underlying principle is that genes or proteins do not react to an external stimulus individually but always integrate into a (sub-)system with more or less defined limits. Importantly, it is biologically feasible to assume the phenomenon of plurifunctionality, i.e., potential participation of an elementary entity (such as gene or protein) into several functional subsystems.

The composition of a functional subsystem is defined by a matrix factorization method in the form of the $\boldsymbol{s_k}$ vector (weights associated with the omics variables) or metagene. The level of activation (or inhibition) of an identified functional subsystem $\boldsymbol{s_k}$ can be read in the corresponding metasample vector $\boldsymbol{a_k}$. The same is relevant for an independent component associated with a technical factor intensity.

If no explicit sparsity constraint is imposed when computing the vectors, then each omics variable (gene, protein) has a non-zero contribution (estimated by its weight in $\boldsymbol{s_k}$) to the definition of the subsystem, which can be positive or negative. However, those variables having close to zero weights can be neglected from the subsystem definition. An important characteristic of a metagene is the set of the most contributing genes (see discussion in the previous section). The most contributing genes are useful to characterize the functional subsystem and to identify if this subsystem corresponds to an existing known one. After determination of the sets of the most contributing genes per each metagene (functional subsystem), one can check if a gene is associated with the subsystem exclusively or contributes to several ones. This analysis can be used to identify potential coupling between the subsystems and their concrete mechanisms (see further discussion). Sometimes it is convenient to distinguish two gene sets per metagene, having the largest and the smallest set of weights, from the positive and the negative sides of the $\boldsymbol{s_k}$ distribution.

We can distinguish two types of functional subsystem response. One is due to the mechanistic downstream effect of a stimulus, i.e., through an induction of a transcription factor downstream of a signaling pathway. Another type is a more systemic one, indirect and related to a longer time scale, caused by an adaptation of the whole system to the presence of potentially harmful factors (such as hypoxia or active immune response) [63,64]. If a studied system's response (e.g., a tumor cell) is

measured in a sufficiently variable number of conditions or perturbation types, one can hope to identify the composition of the most relevant/responsive functional subsystems by applying an appropriate machine learning methodology.

Identification of functional subsystems (modules) from cancer omics data was first historically approached with hierarchical clustering of genes [65]. Matrix factorization in this sense seems to be a more suitable mathematical formalism since it naturally allows taking into account the gene plurifunctionality. This is a simple consequence of a gene that can significantly contribute (i.e., be in the list of the most contributing genes) to the definition of several functional subsystems. ICA is a powerful approach here, because the requirement of maximally possible statistical independence seems to be well suited for the task of subsystem identification. Even if the activity of a pair of functional subsystems is correlated in the most of observed conditions, ICA can still distinguish them based on a smaller number of conditions when they de-synchronize (see a discussion of this aspect in the methodological part of the review). This ability of ICA is also powerful in disentangling the technical biases from biologically relevant signals (as discussed in Section 3.1), making the identification of the functional subsystems less prone to technical biases. Last but not least, ICA allows taking into account the case when the activation of a functional subsystem is connected to inhibition of some of the genes or proteins. One of the simple examples of such a situation is when a transcription factor has a role of an activator for some genes and an inhibitor for other genes.

Functional subsystems identified by ICA can reveal an important coupling of several known biological mechanisms and relate it to the biological phenotype such as cancer patient outcome. In the case of breast cancer, this phenomenon was described in Reference [27] through so-called ICA-based association networks.

One important characteristics of the weight distribution composing $\boldsymbol{a_k}$ is the unimodal or bi- or multi-modal character of the distribution. In the case of well-defined bimodality of a metasample, one can stratify the distribution of samples into two groups, with respect to the nature of the functional subsystem identified. A typical example of this kind is the identification of the functional subsystem of proliferation in single-cell RNA-Seq data, where the corresponding metasamples frequently have two modes, corresponding to proliferative and non-proliferative cell states.

Functional subsystems have been systematically identified using ICA from a large cohort of transcriptomic profiles in Reference [43], where 298 Gene Expression Omnibus (GEO) datasets profiling 9395 human samples (from various conditions including cancer samples) were used to identify 423 "fundamental components of human biology". As an example of their use, the authors characterized the molecular mechanisms of parthenolide anti-cancer drug action. Recently, similar large-scale analysis has been applied to a larger dataset, containing 2753 datasets and 97,049 samples [22]. Compared to the earlier study, the authors improved the methodology in order to avoid redundant and correlated transcriptional component definitions, applying Horn's parallel analysis in order to select the optimal number of components and systematically evaluating the components' reproducibility after resampling. This analysis resulted in defining 139 reproducible and informative transcriptional modules whose value for the downstream analysis was explicitly demonstrated.

Identification of the functional subsystems and distinguishing them from potentially technology driven factors can be strongly improved by the application of ICA analysis to multiple similar datasets independently (without merging them). In this scenario, the ICA results from several datasets were compared with each other in terms of the correlation or other suitable similarity measure among metagenes (Figure 4). In the case of cancer, one of the first applications of this approach was done in Reference [27] for 800 breast cancer samples from four independently profiled cohorts with a conclusion that independent components matched well the underlying cancer mechanisms. This type of meta-analysis was further upscaled in Reference [33], where 22 non-redundant cancer transcriptomic datasets were analyzed. Some of the datasets were related to the same cancer type, i.e., eight of them were collecting samples of bladder cancer and six were from breast cancer. Because the datasets used in this study were produced using different technological platforms, this analysis identified the technical

biases captured by individual components in specific datasets and not reproduced among others. It also distinguished cancer type-specific functional subsystems (such as differentiation program of urothelial tissue) and generic and potential pancancer-wise important functional subsystems (such as the transcriptional program of proliferation or oxphos). Interestingly, one of the bladder cancer-specific components associated with differentiation of urothelial tissue was also associated with amplification of a genomic region, containing a particular transcription factor (PPARG). This led to a conclusion about the role of PPARG in differentiated bladder tumors which was validated experimentally. In Reference [38], 14 non-redundant colon cancer transcriptomic datasets were analyzed by ICA, and the resulting s_k vectors were matched with each other through correlation in order to reveal the functional modules implicated in colon cancer tumor cells' and the variability of tumoral microenvironment.

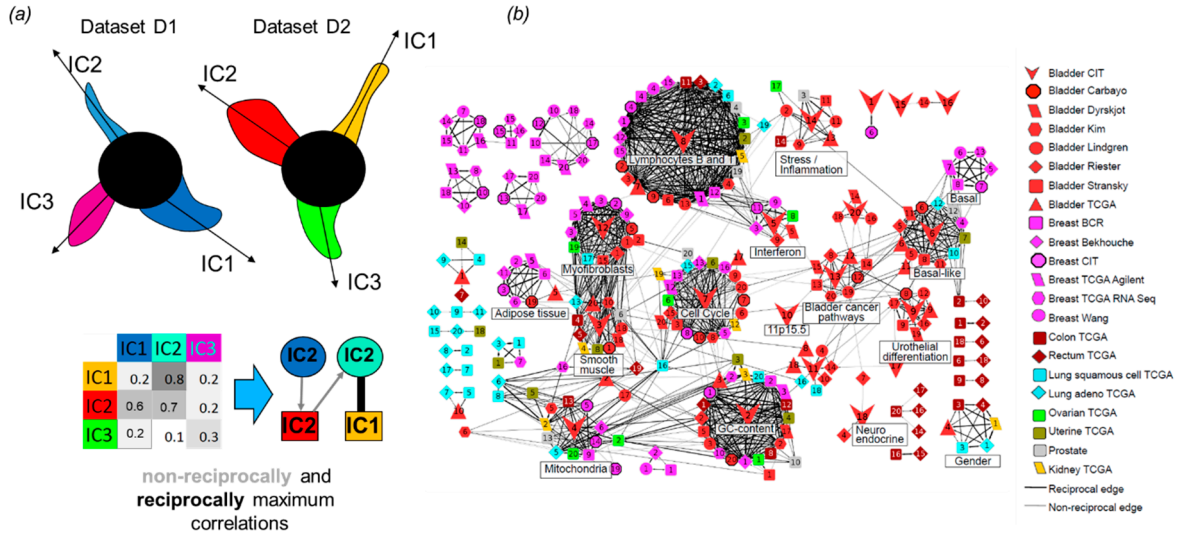


Figure 4. Use of ICA components in meta-analysis of multiple omics datasets. (**a**) Pairwise comparison of two sets of ICA metagenes led to an asymmetric correlation matrix (same as in Figure 2c) which can be converted to a graph using some threshold and selecting the *maximal* correlations. If two components are maximally correlated with each other, then such a correlation defines reciprocal best hit (RBH). (**b**) Graph of maximal correlations (reciprocal and not) exceeding certain threshold among components computed for 22 cancer transcriptomic dataset. Each node is a component, and an edge denotes a correlation. Color reflects the cancer type (e.g., red is bladder cancer). Communities in this graph define highly reproducible cancer type-specific and universal latent factors The figure is reproduced with permission from Reference [33].

In the case of a very good match between ICA-based metagene definitions from several independent datasets, one can define a consensus metagene definition from the meta-analysis (a meta-metagene). An exemplary set of such reference metagenes was built in Reference [33] and used in other studies to facilitate the interpretation of the ICA results. This set included (a) ICA-derived and universal for many cancer types of proliferation-, oxphos-, immune infiltration-, interferon signaling-associated metagenes; (b) consensus metagenes associated with the presence of non-tumor cells of several types in TME; and (c) bladder cancer-specific transcriptional modules (such as differentiation program of urothelial tissue). A comprehensive catalogue of ICs identified in the pan-cancer TCGA dataset containing 32 cancer types was produced in Reference [34]. It appears to be a useful effort to extend the collection of reference consensus metagenes, since they seem to be highly generalizable (reproducible in independent datasets) [22,38].

3.3. Applications to Unsupervised Cell Type Deconvolution

In cancer biology, bulk omics data (especially transcriptome and methylome) represent heterogeneous samples such as peripheral blood mononuclear cell (PBMC) and tumor biopsies, in which the expression

profiles of distinct cell types are mixed in each sample at a priori unknown proportions. The tumor microenvironment is composed of many different cells including a plethora of immune cells, stromal cells, and blood and lymphatic vessels [66]. The quantities and the nature of the TME compartments change with the cancer type and cancer stage. Recent works showed that immune cells could influence tumor cells in different ways and that the immune therapies take advantage of the protective function of the immune system and aim to activate patients' immune defense.

Therefore, one of the major challenges for computational analysis of bulk samples is evaluating the proportions and the properties of individual cell types composing the sample, frequently called deconvolution problem in this context [32]. In general terms, deconvolution stands for unmixing a mixture, which makes it close to the blind source separation methods, including ICA (Figure 5a).

Deconvolution of cancer bulk transcriptomes gained a lot of popularity in the last several years due to the abundance of data sources. Several methods were proposed to estimate the abundance of immune infiltration in cancers at different levels of granularity [67–70] using a pre-defined set of genes, usually generated from pure blood cell population gene expression data [67,68], from single-cell RNA-Seq measurements [70] or mixed [69]. They were proven to correctly estimate the cell-type abundance in silico simulated datasets, in vitro cell mixtures, and blood or PBMC transcriptomes coupled with fluorescence-activated cell sorting (FACS) estimations. However, it remains unclear how many cell types or cell states can be quantified from bulk transcriptomes as each tool comes with own definition of cell-types (e.g., T-cell) and subtypes (e.g., CD4-activated T-cell) or cell states (e.g., cytotoxic T-cell).

In response to this problem, reference-free (also called unsupervised) approaches propose a more data-driven way of performing the deconvolution. This group of approaches is able to discover the cell types and their markers as well as approximate profiles of those cell types (perform "complete deconvolution"). Different types of matrix factorization are suitable for solving this problem. Even though these deconvolution methods are called reference-free, known reference profiles are used to interpret and select the cell type-related components. Different possible benefits of reference-free approaches can be listed as (a) flexibility—discovering the context-dependent cell-type markers, (b) discovery of new cell types or cell types that are specific to a certain context, (c) determining deconvoluted profiles of cell types that can be used to remove the immune-related signal or to better understand the cell type features, and (d) ability to characterize biological processes (such as cell cycle activity) simultaneously with the cell types.

The reference-free approaches were already applied for deconvolution of cell types in blood using semi-supervised NMF [71]. They were used to study brain [72], tumoral single cells [73], and cell-cycle in yeast [74].

In Reference [42] icasso-stabilized fastICA was applied to a set of six large breast cancer patient cohorts profiled for gene expression. It was demonstrated that the immune-related factors, especially the signal of T-cell, B-cell, and macrophages were highly reproducible in independent datasets (Figure 5b). In Reference [75], the DeconICA R package (Available online: https://github.com/UrszulaCzerwinska/DeconICA) was developed with the objective to apply ICA to the task of cell type deconvolution. It was shown that ICA is able to efficiently estimate the cell type proportions with better accuracy than leading supervised algorithms even though it can identify less cell sub-types than most of the published solutions. It suggests that ICA-based deconvolution is less prone to overfitting and enables discovery and quantification of strong and stable signals (not necessary the most abundant but rather the most specific). DeconICA was applied to a big corpus of data containing more than 100 transcriptomic datasets composed of over 28,000 samples of 40 tumor types generated by different technologies and processed independently. In addition, the ICA-derived metagenes were used as context-specific signatures in order to study the characteristics of immune cells in different tumor types. The analysis revealed a large diversity and plasticity of immune cells dependent and independent on tumor type. Some conclusions of the study can be helpful in the identification of new drug targets or biomarkers for immunotherapy of cancer.

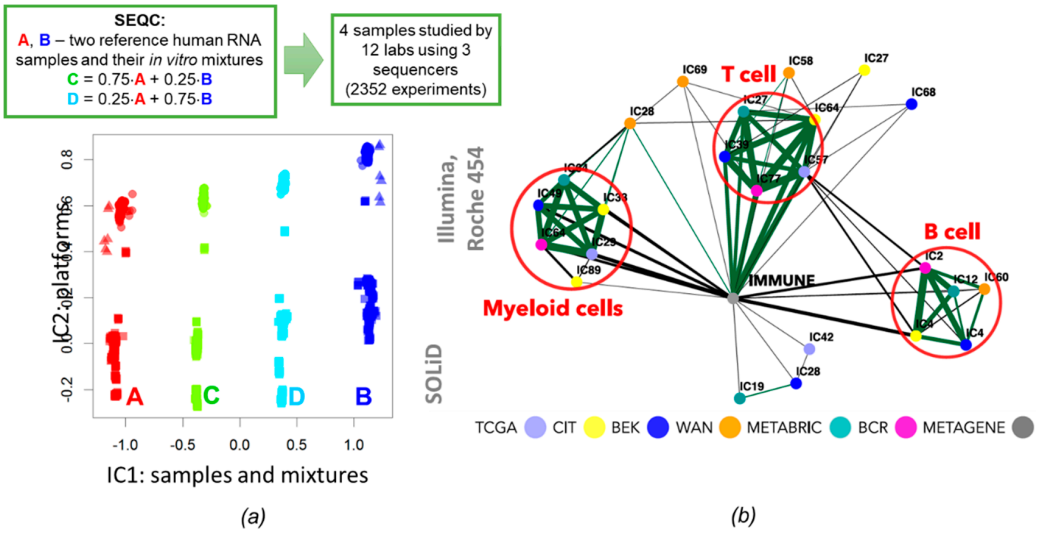


Figure 5. Examples of utility of ICA for unsupervised deconvolution of cell types. (**a**) Application of ICA to the Sequencing Quality Control consortium (SEQC) dataset [76] containing measurements of two references transcriptomic profiles of cell lines and their mixtures at known proportions. The first two ICs identify the types and the effect of the platform. (**b**) Correlation graph among selected components from ICA applied to six non-redundant breast cancer transcriptomic datasets. Three cliques formed in the graph correspond to major immune cell types. The thickness of the edges reflects the absolute correlation value. "Immune" meta-metagene was defined in Reference [33] as the one associated with the presence of immune infiltrate in a tumor. This figure was reproduced with permission from Reference [42].

Cell-type composition can be also computed from DNA methylation data. In the EWAS (Epigenome Wide Association Studies), the variation origination from cell types is considered as an important confounding factor that should be removed before comparing cases and controls and defining Differentially Methylated Positions (DMPs). For example, in Reference [77] ten tools for epigenome deconvolution were reviewed. The authors described six reference-free methods, three regression-based, and one semi-supervised. Some of these methods use approaches close to ICA, such as independent surrogate variable analysis (ISVA) [78], where the goal is to adjust the data for any type of confounder (be it cell-type composition or not). Clear superiority of ICA over PCA in methylome deconvolution has not been yet demonstrated. Most of the existing tools for unsupervised methylome deconvolution assume cell composition as the most contributing to the methylome variability. According to Reference [78], this assumption was not proven to hold true in solid tissues, normal or pathological. It appears to be interesting to test different approaches to ICA coupled with improved reference profiles to check if it cannot open new perspectives in methylome deconvolution.

3.4. ICA Applications to Single-Cell Omics Data Analysis

Statistical properties of ICA seem to be very attractive to justify its application to the emerging wealth of single-cell omics data profiles. ICA can serve here to improve the data analysis regarding dimensionality reduction, removing technical biases, integrating datasets. ICA also looks promising and represents an alternative to the standard dimensionality reduction followed by clustering methodology for identifying cell types or states, suggesting a more continuous way of considering them, with a possibility of existence of intermediate or mixed cell populations.

Similar to bulk RNA-seq, technical biases and batch effects are limiting factors for single-cell RNA-Seq and should be either removed or taken into account. One example of ICA application to normalize batch effect was recently reported by Dirkse et al. [79]. The authors observed a strong difference between the original patient-derived cell line and its two subpopulations measured in the second batch (all cells undergo the same protocol of cell growth and sorting, so biological differences

were excluded). The difference among batches was comparable to the difference among different cell lines. The ICA identified and isolated the batch effect in one of the components. By removing this component and recalculating expression matrix, the authors corrected for this batch effect (see Figure 6a,b). A similar approach was exploited in Reference [80] in order to pre-process the single cell data following the trans-differentiation process of murine pre-B cells into macrophages and their reprogramming into induced pluripotent stem cells. In this study, 15 out of 35 independent components were considered to be connected with technical artifacts such as sample batch effects and cell position in the plate and filtered out from the downstream analysis.

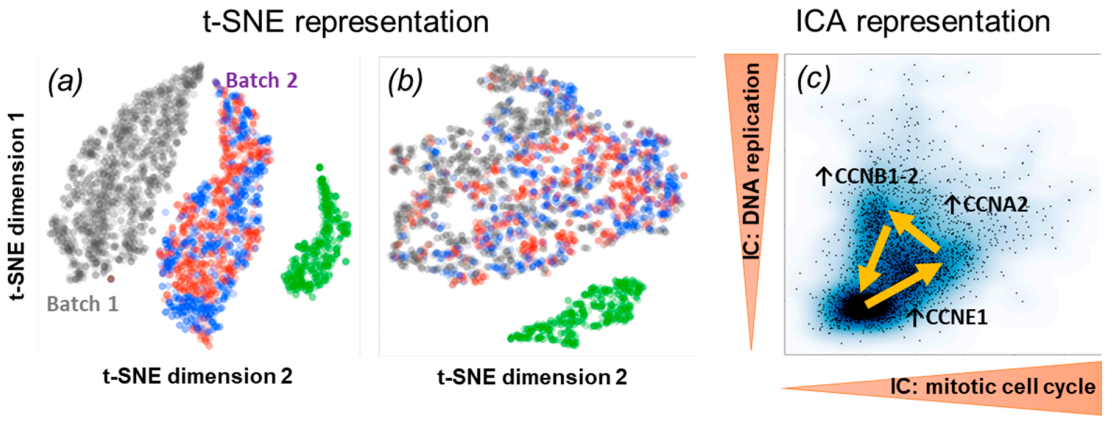


Figure 6. Application of ICA in single cell data analysis of tumors (study of glioblastoma from Reference [79]). (**a**) t-distributed stochastic neighbor embedding (t-SNE) visualization of the data reveals a strong batch effect. Grey and red/blue dots represent cells from the same cell line, analyzed in two batches (batch 1—grey dots, batch 2—red and blue cells). The green dots show a cell population from a different cell line added to the dataset for the reason of comparison. (**b**) t-SNE visualization of the data after eliminating signals contained in one IC associated with batch effect. (**c**) In ICA decomposition of single cell scRNA-Seq data from cancer studies, usually there exist two components associated with phases of the cell cycle (G1/S, DNA replication, and G2/M, mitosis). Here the loadings of such two components are visualized. Black arrows show the regions when the labeled genes are highly expressed. Yellow arrows show assumed direction of the progression through the cell cycle.

The ICA-based dimensionality reduction is a standard step in the most popular packages for analyzing single cell RNA-Seq data. In MONOCLE [61,81], ICA is optionally used for the initial step of dimensionality reduction to 2D, before inferring cellular trajectories. For example, this option was used in order to derive the cellular trajectory of individual MCF-7 breast cancer cells after stimulating them with estrogen [82]. It is also part of the popular toolbox Seurat [83] as one of the standard choices for dimensionality reduction, data visualization, and feature selection. ICA can be exploited, together with other low-dimensional projections, in various recently developed packages for biologically meaningful single-cell data visualization [84].

In Reference [85], ICA was applied in order to define subtypes of the immune-related cells present in the TME of melanoma (with original data from Reference [86]) and relate them to the mechanisms of innate immune response. ICA was used to define the continuous spectrum of differentiation in hematopoietic cells from scRNA-Seq data in Reference [87]; several latent factors were associated to the underlying biological mechanisms of differentiation. In Reference [80] three independent components computed for a scRNA-Seq dataset were matched with transcriptional programs specific to B-cells, macrophages, and monocytes and used to provide an interpreTable 3D data visualization. Interestingly, in order to establish the biological origin of these components, they were correlated to the ICA decomposition of the transcriptomic atlas of murine cell types from which 120 independent components were extracted.

ICA served as a principal machine learning method for discovering functional subsystems involved in the response of Ewing sarcoma cells to the induction of the chimeric oncogenic transcription factor EWSR-FLI1 [62]. In this case, ICA was applied to the temporally resolved single cell RNA-Seq dataset and revealed the existence of few tens of transcriptional programs activated or inhibited after the controlled induction of the oncogene. Quite remarkably, one of the independent components was clearly associated with the functional subsystem composed of the direct targets of EWSR-FLI1, and it was distinguished from its indirect downstream effects such as cell cycle induction (see Figure 6c). Other functional subsystems reacting to the variations of the experimental conditions such as hypoxia or regulation of glucogenesis, were recapitulated in individual ICs. Identification of the functional subsystems from the cell line experiments were further used in order to characterize the patient-derived xenografts (PDXs) of Ewing sarcoma, at single cell level.

In principle, ICA is the methodology able to exploit strong non-Gaussianity in the multidimensional distributions formed by single cells in the space of omics profiles. However, in order to optimally use this potential, one probably needs to identify the most suitable non-linearity functions, for each particular type of single cell measurements, and take into account the nature of the multivariate distribution of points in data space. Recently, a matrix factorization-based method ZINB-WaVE was adapted to the single cell RNA-Seq measurements, using the model of zero-inflated negative binomial distribution (ZINB) [88]. In principle, ICA approach can be applied on top of ZINB-WaVE instead of PCA; however, this approach needs to be tested in practice.

3.5. Multi-Omics ICA Applications in Cancer Research

The majority of published works on applying ICA in cancer research deals with transcriptomic data. This is connected in part to the relatively high abundance of such data type from collections of bulk tumors, and in part to the availability of bioinformatics tools helping to interpret the obtained components (such as Gene Set Enrichment Analysis). Yet another aspect is that transcriptomic data are better connected so far to the clinical questions such as defining molecular subtypes of tumors.

However, applying ICA should not be limited to only one level of omics profiling, and there is a lot of potential in applying it to several levels of molecular description. The multi-level datasets become increasingly available in the cancer biology. The levels of molecular description can be gene copy number profiles, binary mutation profiles, measured mutational signatures, measured total expression of genes or spliced mRNA isoforms or non-coding genes such as microRNAs, DNA methylation or histone mark modification profiling, protein or protein phosphoforms relative abundances or some other less frequently used omics types. Identification of functional subsystems can be facilitated through the use of several data types, since the adaptation process is frequently expected to span several levels. As a good example of such a multi-omics dataset, one can cite recent work on comprehensive characterization of medulloblastoma [89].

Ideally, several levels of omics profiling should be collected for the same and sufficiently large set of samples. Independent components can be then computed for each data type separately and then the identified components can be compared by computing correlations between the corresponding $\boldsymbol{a_k}$ vectors (metasamples). Such an approach was recently applied in Reference [39] to a set of melanoma bulk samples, profiled at the level of transcriptome and microRNA expression. Similarly, in a recent study [90], 77 breast and 84 ovarian cancer samples, profiled simultaneously at transcriptome and proteome level, were analyzed using stabilized ICA, followed by integrating the discovered associations with clinical data and molecular pathways.

An alternative and somewhat more powerful idea consists in stacking several matrices corresponding to the different levels of omics profiling into a tensor (multi-dimensional array), in order to apply the tensorial version of ICA. In this case, ICA will be able to learn and jointly optimize the signals which can involve variables from several levels of molecular description. This requires making at least two dimensions of the data common, while the third matrix dimension indicates the

data type. Typically, all molecular measurements are mapped onto the genes through application of procedures that can be non-trivial (e.g., in the case of Chip-Seq experiments).

The resulting three-dimensional measurement tensor X_{ijk} has dimensions "number of samples × number of genes × number of data types". For example, $Xi = 4$, $j = 5$, $k = 2$ element in the tensor can indicate DNA methylation level of the promoter of the gene 5 in the sample 4. $Xi = 4$, $j = 5$, $k = 1$ could indicate expression of the same gene in the same sample. In the case of tensor factorization, the resulting components represent matrices rather than vectors having dimensions "number of genes versus number of data types" (for metagenes) and "number of data types versus number of samples" (for metasamples). The existence of correlations among different data types within the same matrix-component indicates coupling among several levels of molecular descriptions captured by tensorial ICA.

Tensorial ICA was recently applied in Reference [91] to colon cancer dataset from The Cancer Genome Atlas (TCGA) composed of a matched subset of copy-number variation (CNV), DNAm, and RNA-seq data. A specific implementation of tensorial ICA called tWFOBI, standing for tensorial fourth-order blind identification, accompanied by a tensorial version of whitening (W), using tensor PCA, was used to compute 37 independent components. Most of these components can be associated with the differences between normal and cancer samples, while only four components capturing correlations between CNV and gene expression, and one among them was also characterized by concomitant correlation among all three data types. Of note, the tWFOBI method showed several orders of magnitude better computational performance compared to the state-of-the-art methods developed for multi-level omics data integration (such as iCluster).

Applications of ICA to data types other than transcriptomic or to several data types simultaneously remain limited; however, first applications of this approach in cancer biology are rather promising [91,92]. Multiple issues still remain to be solved for how to define the best practice of ICA application to, for example, DNA methylation profiles and how to interpret the obtained results. For example, in Reference [93], a "spatio-temporal" version of ICA was suggested in order to take into account certain specificity of DNA methylation profiles such as a high level of correlation among probes located close in the genome. Also, in the case of methylation data, ICA should be carefully benchmarked with other machine-learning methods exploiting the non-Gaussian nature of signals [94].

3.6. Correlations and Interactions among Functional Subsystems Defined by ICA

Functional subsystems identified through ICA and fixed in the form of metagenes can be studied for their statistical relationships within a dataset, among datasets of the same kind, or among datasets that are not closely related in terms of the nature of the biological samples profiled.

In the latter case, one can use ICA for studying disease–disease relationships. An example of such a relation is the phenomenon of inverse comorbidity between cancer and some other diseases, in terms of the anti-correlated activation pattern of the common functional subsystems. For example, the ICA method was exploited in Reference [95] in order to identify inversely associated transcriptional modules common in breast cancer and Alzheimer's disease. In a more extensive study [44], 17 transcriptomic datasets (11 collected for the post-mortal brain samples of patients suffering from Alzheimer's disease and six collected for the lung cancer samples) have been analyzed using ICA. The notion of reciprocal best hit (RBH, see the methodological section of this review) was used in order to match the ICA components and define their communities. In order to detect the anti-correlation patterns among the matched components, a specific method was developed to assign an orientation of the components and, hence, the weight signs in the metagene, based on the analysis of the subset of normal control samples, in the $\boldsymbol{a_k}$ vectors. This analysis confirmed previously identified comorbidity patterns based on the analysis of individual gene expression profiles (related to the role of immune system and mitochondrial metabolism) and suggested new molecular mechanisms of comorbidity between lung cancer and Alzheimer's disease such as estrogen receptor signaling pathway or the involvement of cadherins.

Another possibility for exploiting the ICA-based definitions of modules is to study the phenomenon of functional subsystem integration as a result of adaptation to stress or harmful conditions [63]. It was shown in many studies that the correlations among the activation patterns of different functional subsystems can be more informative than the patterns themselves [96]. ICA can deconvolute even strongly correlated signals (see Figure 2A and the Section 2.2 of this review). Also, it computes components which are as mutually independent as possible, but the level of dependence can be different even for subsets of samples within a single dataset. For example, one can expect that in the normal subset of samples, some of the functional subsystems will be less coupled with each other than in stressful conditions caused by more aggressive stages of tumorigenesis. This coupling can be caused by, for example, the shortage of essential metabolic resources making them a common limiting factor for multiple functional subsystems. If the level of mutual information between two signals increases above the ability of ICA to discriminate components, then these signals will be captured by one independent component. This phenomenon of *independent component splitting and merging* might depend on the order of ICA decomposition (and it was empirically studied in Reference [34]), on the specific biases in the composition of samples, or on the number of samples.

The theoretical principles of functional subsystems integration have been developed [63,64]. However, it remains an interesting problem to verify and apply them to the concrete modules identifiable from the (multi-)-omics profiles. Independent Component Analysis represents an interesting option for achieving this objective.

4. Discussion

In recent decades, independent component analysis has become a standard tool for the analysis of tables of omics measurements in cancer biology. In certain applications, it was shown to have advantages, especially in terms of reproducibility or generalizability and biological interpretability, compared to other popular matrix factorization methods. Despite ICA being shown to be a useful tool, it seems to be under-appreciated partly due to the fact of historical reasons and partly due to the presence of existing confusions in the underlying assumptions and/or interpretation of the resulting matrix decompositions. For example, it is frequently commented that biological processes are not perfectly independent and that they are expected to be correlated in some conditions. Even though this is true, ICA can distinguish signals coupled to some extent by making the corresponding components as independent as possible.

In this review, we made a comprehensive effort to mention most of the recent studies in cancer research where ICA served as the essential data analysis tool. We classified them into several common topics: data preprocessing, data dimension reduction and visualization, identification of functional subsystems and their correlations or interactions, deconvolution of cell types, data integration and meta-analysis.

We also reviewed the methodological works aimed at defining the best practices of applying ICA to concrete types of omics data. Compared to the early times of applying ICA to omics datasets, today there exists a variety of implementations and improved methodologies allowing us to use the valuable idea behind ICA (exploiting the concept of statistical independence and use of higher moments of multivariate data distributions) in the best possible way. Certain progress has been made in clarifying such important questions such as determining the optimal number of components to retain or establishing the biological significance of the extracted components.

Most of the existing applications of ICA have been done so far for transcriptomic data even if the interpretation of the components used other types of molecular data (such as mutations, copy number alterations). Since recently, ICA started to be applied to other types of omics profiles, including methylation profiles and proteomics datasets. It seems interesting to determine, using ICA, independent sources of variance in newly emerging omics data types, such as systematic Chip-Seq datasets mapping the state of histone modifications or mutational signature profiles. More experience, standardization and assessment are required to use ICA in the most optimal way for the analysis of

single-cell and multi-omics datasets. Moreover, in this review we did not even mention other fields of ICA application in cancer biology, including the analysis of imaging data (e.g., [97]), clinical records, and other non-omics data types, for which the ICA data model might be of interest.

Matrix factorization represents an alternative approach to the standard clustering methods, being more flexible in terms of taking into account gene plurifunctionality and ability for unsupervised deconvolution of factors whose activity can be correlated. It is worth noticing that some ICA algorithms (such as fastICA) are computationally performant when properly implemented and potentially able to deal with large amounts of molecular measurements. In this sense, ICA remains competitive vis-a-vis many other approaches (e.g., based on likelihood maximization or representation of the data in the form of multilayered networks).

We believe there are interesting directions to further explore and more deeply use the concepts behind ICA in the context of cancer biology data analyses. It would be interesting to reconsider the roots of independent component analysis into artificial neural network methodology, suggesting novel scalable autoencoding-based techniques in order to solve the problem of blind source separation adapted to the nature of the biological data. Assessing the value of supervised learning of the features extracted by ICA from omics and other data types and comparing them to "hand-crafted" or convolutional neural network-based features can lead to designing performant hybrid learning approaches, as in Reference [98]. It appears promising to take advantage of the wealth of recently emerged formalized knowledge on biological mechanisms of cancer and develop methods to inject this knowledge into the component learning process. The biological factors or functional subsystems in cells or cellular ecosystems are organized in complex hierarchies, and we need new approaches to explicitly take this into account, in order to improve the subsystems identifiably.

To conclude, as a team of authors all having extensive experience in applying independent component analysis as a tool in computational cancer biology, we advocate for its wider use in making sense of the growing amount of omics data in this and other fields.

Funding: This work was partially supported by the grant research projects "Pan-cancer deconvolution of omics data using Independent Component Analysis" (IRN: AP05135430) and "Investigation of esophageal cancer tissue gene expression derived from Kazakhstan patients by next-generation sequencing technology" (IRN: AP05134722) of the Ministry of Education and Science of the Republic of Kazakhstan, by the Ministry of Science and Higher Education of the Russian Federation (Project No. 14.Y26.31.0022), the European Union's Horizon 2020 program (grant No. 826121, iPC project), by the European IMI IMMUCAN project, and by Luxembourg National Research Fund (C17/BM/11664971/DEMICS).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ICA	Independent Component Analysis
PCA	Principal Component Analysis
NMF	Non-Negative Matrix Factorization
MSTD	Maximally Stable Transcriptomic Dimension
fMRI	functional Magnetic Resonance Imaging
TCGA	The Cancer Genome Atlas
BIC	Bayesian Information Criterion
FOBI	Fourth-Order Blind Identification
SEQC	Sequencing Quality Control consortium
t-SNE	t-Distributed Stochastic Neighbor Embedding

References

1. Liebermeister, W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **2002**, *18*, 51–60. [CrossRef] [PubMed]
2. Lee, S.-I.; Batzoglou, S. Application of independent component analysis to microarrays. *Genome Biol.* **2003**, *4*, R76. [CrossRef] [PubMed]

3. Saidi, S.A.; Holland, C.M.; Kreil, D.P.; MacKay, D.J.C.; Charnock-Jones, D.S.; Print, C.G.; Smith, S.K. Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* **2004**, *23*, 6677–6683. [CrossRef] [PubMed]
4. Frigyesi, A.; Veerla, S.; Lindgren, D.; Höglund, M. Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinform.* **2006**, *7*, 290. [CrossRef] [PubMed]
5. Wang, Q.; Li, Q.; Mi, R.; Ye, H.; Zhang, H.; Chen, B.; Li, Y.; Huang, G.; Xia, J. Radiomics nomogram building from multiparametric MRI to predict grade in patients with glioma: A cohort study. *J. Magn. Reson. Imaging* **2019**, *49*, 825–833. [CrossRef]
6. Levine, A.B.; Schlosser, C.; Grewal, J.; Coope, R.; Jones, S.J.M.; Yip, S. Rise of the machines: Advances in deep learning for cancer diagnosis. *Trends Cancer* **2019**, *5*, 157–169. [CrossRef]
7. Tandel, G.S.; Biswas, M.G.; Kakde, O.; Tiwari, A.S.; Suri, H.; Turk, M.; Laird, J.R.; Asare, C.K.; Ankrah, A.N.; Khanna, N.; et al. A Review on a deep learning perspective in brain cancer classification. *Cancers (Basel)* **2019**, *11*, 111. [CrossRef]
8. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [CrossRef]
9. Gao, Z.; Wu, S.; Liu, Z.; Luo, J.; Zhang, H.; Gong, M.; Li, S. Learning the implicit strain reconstruction in ultrasound elastography using privileged information. *Med. Image Anal.* **2019**, *58*, 101534. [CrossRef]
10. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
11. Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van der Laak, J.A.W.M.; Hermsen, M.; Manson, Q.F.; Balkenhol, M.; et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **2017**, *318*, 2199–2210. [CrossRef] [PubMed]
12. Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **2018**, *24*, 1248–1259. [CrossRef] [PubMed]
13. Schmidt, C.M.D. Anderson breaks with IBM Watson, raising questions about artificial intelligence in oncology. *J. Natl. Cancer Inst.* **2017**, *109*, 5. [CrossRef] [PubMed]
14. Gorban, A.N.; Mirkes, E.M.; Tyukin, I.Y. How Deep should be the depth of convolutional neural networks: A backyard dog case study. *Cognit. Comput.* **2019**, 1–10. [CrossRef]
15. Karhunen, J.; Oja, E.; Wang, L.; Vigario, R.; Joutsensalo, J. A class of neural networks for independent component analysis. *IEEE Trans. Neural Netw.* **1997**, *8*, 486–504. [CrossRef] [PubMed]
16. Brunet, J.P.; Tamayo, P.; Golub, T.R.; Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4164–4169. [CrossRef] [PubMed]
17. Gorban, A.N.; Zinovyev, A.Y. Principal graphs and manifolds. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*; IGI Global: Hershey, PA, USA, 2008; ISBN 9781605667669.
18. 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*, 1182–1187. [CrossRef]
19. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159. [CrossRef]
20. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
21. Cardoso, J.-F. High-order contrasts for independent component analysis. *Neural Comput.* **1999**, *11*, 157–192. [CrossRef]
22. Zhou, W.; Altman, R.B. Data-driven human transcriptomic modules determined by independent component analysis. *BMC Bioinform.* **2018**, *19*, 327. [CrossRef] [PubMed]
23. Risk, B.B.; Matteson, D.S.; Ruppert, D.; Eloyan, A.; Caffo, B.S. An evaluation of independent component analyses with an application to resting-state fMRI. *Biometrics* **2014**, *70*, 224–236. [CrossRef] [PubMed]
24. Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F.J. Bayesian Independent Component Analysis Recovers Pathway Signatures from Blood Metabolomics Data. *J. Proteome Res.* **2012**, *11*, 4120–4131. [CrossRef] [PubMed]
25. Bach, F.R. Kernel independent component analysis. *J. Mach. Learn. Res.* **2002**, *3*, 1–48.

26. Zibulevsky, M.; Pearlmutter, B.A. Blind source separation by sparse decomposition in a signal dictionary. *Neural Comput.* **2001**, *13*, 863–882. [CrossRef] [PubMed]
27. Teschendorff, A.E.; Journée, M.; Absil, P.A.; Sepulchre, R.; Caldas, C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.* **2007**, *3*, e161. [CrossRef] [PubMed]
28. Virta, J.; Taskinen, S.; Nordhausen, K. Applying fully tensorial ICA to fMRI data. In Proceedings of the 2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 6 December 2016; pp. 1–6.
29. Virta, J.; Li, B.; Nordhausen, K.; Oja, H. Independent component analysis for tensor-valued data. *J. Multivar. Anal.* **2017**, *162*, 172–192. [CrossRef]
30. Bach, F.R.; Jordan, M.I. Beyond independent components: Trees and clusters. *J. Mach. Learn. Res.* **2003**, *4*, 1205–1233.
31. Meyer-Bäse, A.; Theis, F.J.; Lange, O.; Puntonet, C.G. Tree-Dependent and topographic independent component analysis for fMRI analysis. In *International Conference on Independent Component Analysis and Signal Separation*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 782–789.
32. Avila Cobos, F.; Vandesompele, J.; Mestdagh, P.; De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **2018**, *34*, 1969–1979. [CrossRef]
33. Biton, A.; Bernard-Pierrot, I.; Lou, Y.; Krucker, C.; Chapeaublanc, E.; Rubio-Pérez, C.; López-Bigas, N.; Kamoun, A.; Neuzillet, Y.; Gestraud, P.; 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*, 1235–1245. [CrossRef]
34. Kairov, U.; Cantini, L.; Greco, A.; Molkenov, A.; Czerwinska, U.; Barillot, E.; Zinovyev, A. Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genom.* **2017**, *18*, 712. [CrossRef]
35. Kong, W.; Vanderburg, C.R.; Gunshin, H.; Rogers, J.T.; Huang, X. A review of independent component analysis application to microarray gene expression data. *Biotechniques* **2008**, *45*, 501–520. [CrossRef] [PubMed]
36. Meng, C.; Zeleznik, O.A.; Thallinger, G.G.; Kuster, B.; Gholami, A.M.; Culhane, A.C. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **2016**, *17*, 628–641. [CrossRef] [PubMed]
37. Barillot, E.; Calzone, L.; Hupe, P.; Vert, J.-P.; Zinovyev, A. *Computational Systems Biology of Cancer*; Taylor & Francis: Abington, UK, 2012; ISBN 9781439831441.
38. Cantini, L.; Kairov, U.; de Reyniès, A.; Barillot, E.; Radvanyi, F.; Zinovyev, A. Assessing reproducibility of matrix factorization methods in independent transcriptomes. *Bioinformatics* **2019**. [CrossRef] [PubMed]
39. Nazarov, P.V.; Wienecke-Baldacchino, A.K.; Zinovyev, A.; Czerwińska, U.; Muller, A.; Nashan, D.; Dittmar, G.; Azuaje, F.; Kreis, S. Independent component analysis provides clinically relevant insights into the biology of melanoma patients. *BMC Med. Genom.* **2019**, 395145. [CrossRef]
40. Chiappetta, P.; Roubaud, M.C.; Torrésani, B. Blind source separation and the analysis of microarray data. *J. Comput. Biol.* **2005**, *11*, 1090–1109. [CrossRef] [PubMed]
41. Himberg, J.; Hyvarinen, A. Icasso: Software for investigating the reliability of ICA estimates by clustering and visualization. In Proceedings of the 2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), Toulouse, France, 17–19 September 2003; pp. 259–268.
42. Czerwinska, U.; Cantini, L.; Kairov, U.; Barillot, E.; Zinovyev, A. Application of independent component analysis to tumor transcriptomes reveals specific and reproducible immune-related signals. In *Proceedings of the Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2018; Volume 10891LNCS, pp. 501–513.
43. Engreitz, J.M.; Daigle, B.J.; Marshall, J.J.; Altman, R.B. Independent component analysis: Mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* **2010**, *43*, 932–944. [CrossRef]
44. Greco, A.; Sanchez Valle, J.; Pancaldi, V.; Baudot, A.; Barillot, E.; Caselle, M.; Valencia, A.; Zinovyev, A.; Cantini, L. Molecular inverse comorbidity between Alzheimer's disease and lung cancer: New insights from matrix factorization. *Int. J. Mol. Sci.* **2019**, *20*, 3114. [CrossRef]
45. Stein-O'Brien, G.L.; Arora, R.; Culhane, A.C.; Favorov, A.V.; Garmire, L.X.; Greene, C.S.; Goff, L.A.; Li, Y.; Ngom, A.; Ochs, M.F.; et al. Enter the matrix: Factorization uncovers knowledge from omics. *Trends Genet.* **2018**, *34*, 790–805. [CrossRef]

46. Way, G.P.; Zietz, M.; Himmelstein, D.S.; Greene, C.S. Sequential compression across latent space dimensions enhances gene expression signatures. *bioRxiv* **2019**. *bioRxiv*:573782. [CrossRef]
47. Cangelosi, R.; Goriely, A. Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* **2007**, *2*, 2. [CrossRef] [PubMed]
48. Ceruti, C.; Bassis, S.; Rozza, A.; Lombardi, G.; Casiraghi, E.; Campadelli, P. DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognit.* **2014**, *47*, 2569–2581. [CrossRef]
49. Albergante, L.; Bac, J.; Zinovyev, A. Estimating the effective dimension of large biological datasets using Fisher separability analysis. In Proceedings of the International Joint Conference on Neural Networks, Hungary, Budapest, 14–17 July 2019.
50. Kuperstein, I.; Grieco, L.; Cohen, D.P.A.; Thieffry, D.; Zinovyev, A.; Barillot, E. The shortest path is not the one you know: Application of biological network resources in precision oncology research. *Mutagenesis* **2015**, *30*, 191–204. [CrossRef] [PubMed]
51. Bonnet, E.; Viara, E.; Kuperstein, I.; Calzone, L.; Cohen, D.P.A.; Barillot, E.; Zinovyev, A. NaviCell Web Service for network-based data visualization. *Nucleic Acids Res.* **2015**, *43*, W560–W565. [CrossRef] [PubMed]
52. Gawron, P.; Ostaszewski, M.; Satagopam, V.; Gebel, S.; Mazein, A.; Kuzma, M.; Zorzan, S.; McGee, F.; Otjacques, B.; Balling, R.; et al. MINERVA-a platform for visualization and curation of molecular interaction networks. *NPJ Syst. Biol. Appl.* **2016**, *2*, 16020. [CrossRef] [PubMed]
53. Cantini, L.; Calzone, L.; Martignetti, L.; Rydenfelt, M.; Blüthgen, N.; Barillot, E.; Zinovyev, A. Classification of gene signatures for their information value and functional redundancy. *NPJ Syst. Biol. Appl.* **2018**, *4*, 2. [CrossRef] [PubMed]
54. Grossmann, P.; Stringfield, O.; El-Hachem, N.; Bui, M.M.; Rios Velazquez, E.; Parmar, C.; Leijenaar, R.T.; Haibe-Kains, B.; Lambin, P.; Gillies, R.J.; et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife* **2017**, *6*, e23421. [CrossRef] [PubMed]
55. Zhang, X.W.; Yap, Y.L.; Wei, D.; Chen, F.; Danchin, A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *Eur. J. Hum. Genet.* **2005**, *13*, 1303–1311. [CrossRef] [PubMed]
56. Huang, D.-S.; Zheng, C.-H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **2006**, *22*, 1855–1862. [CrossRef] [PubMed]
57. Zheng, C.H.; Huang, D.S.; Kong, X.Z.; Zhao, X.M. Gene Expression Data Classification Using Consensus Independent Component Analysis. *Genom. Proteomics Bioinform.* **2008**, *6*, 74–82. [CrossRef]
58. Aziz, R.; Verma, C.K.; Srivastava, N. A novel approach for dimension reduction of microarray. *Comput. Biol. Chem.* **2017**, *71*, 161–169. [CrossRef] [PubMed]
59. Nascimento, M.; Silva, F.F.E.; Sáfadi, T.; Nascimento, A.C.C.; Ferreira, T.E.M.; Barroso, L.M.A.; Ferreira Azevedo, C.; Guimarães, S.E.F.; Serão, N.V.L. Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data. *PLoS ONE* **2017**, *12*, e0181195. [CrossRef] [PubMed]
60. Han, H.; Li, X.L. Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. *BMC Bioinform.* **2011**, *12*, S7. [CrossRef]
61. Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Lennon, N.J.; Livak, K.J.; Mikkelsen, T.S.; Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **2014**, *32*, 381–386. [CrossRef] [PubMed]
62. Aynaud, M.-M.; Mirabeau, O.; Gruel, N.; Grossetete-Lalami, S.; Boeva, V.; Durand, S.; Surdez, D.; Saulnier, O.; Zaidi, S.; Gribkova, S.; et al. Transcriptional programs define intratumoral heterogeneity of Ewing sarcoma at single cell resolution. *bioRxiv* **2019**. *bioRxiv*:623710. [CrossRef]
63. Gorban, A.N.; Pokidysheva, L.I.; Smirnova, E.V.; Tyukina, T.A. Law of the minimum paradoxes. *Bull. Math. Biol.* **2011**, *73*, 2013–2044. [CrossRef] [PubMed]
64. Gorban, A.N.; Tyukina, T.A.; Smirnova, E.V.; Pokidysheva, L.I. Evolution of adaptation mechanisms: Adaptation energy, stress, and oscillating death. *J. Theor. Biol.* **2016**, *405*, 127–139. [CrossRef]
65. Segal, E.; Friedman, N.; Koller, D.; Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* **2004**, *36*, 1090–1098. [CrossRef]
66. Galon, J.; Mlecnik, B.; Bindea, G.; Angell, H.K.; Berger, A.; Lagorce, C.; Lugli, A.; Zlobec, I.; Hartmann, A.; Bifulco, C.; et al. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J. Pathol.* **2014**, *232*, 199–209. [CrossRef]

67. Becht, E.; Giraldo, N.A.; Lacroix, L.; Buttard, B.; Elarouci, N.; Petitprez, F.; Selves, J.; Laurent-Puig, P.; Sautès-Fridman, C.; Fridman, W.H.; et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **2016**, *17*, 218. [CrossRef]
68. Newman, A.M.; Liu, C.L.; Green, M.R.; Gentles, A.J.; Feng, W.; Xu, Y.; Hoang, C.D.; Diehn, M.; Alizadeh, A.A. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **2015**, *12*, 453–457. [CrossRef] [PubMed]
69. Aran, D.; Hu, Z.; Butte, A.J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **2017**, *18*, 220. [CrossRef] [PubMed]
70. Racle, J.; de Jonge, K.; Baumgaertner, P.; Speiser, D.E.; Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **2017**, *6*, e26476. [CrossRef] [PubMed]
71. Gaujoux, R.; Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **2012**, *12*, 913–921. [CrossRef] [PubMed]
72. Nelms, B.D.; Waldron, L.; Barrera, L.A.; Weflen, A.W.; Goettel, J.A.; Guo, G.; Montgomery, R.K.; Neutra, M.R.; Breault, D.T.; Snapper, S.B.; et al. CellMapper: Rapid and accurate inference of gene expression in difficult-to-isolate cell types. *Genome Biol.* **2016**, *17*, 201. [CrossRef] [PubMed]
73. Kotliar, D.; Veres, A.; Nagy, M.A.; Tabrizi, S.; Hodis, E.; Melton, D.A.; Sabeti, P.C. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **2019**, *8*, e43803. [CrossRef] [PubMed]
74. Wang, N.; Hoffman, E.P.; Chen, L.; Chen, L.; Zhang, Z.; Liu, C.; Yu, G.; Herrington, D.M.; Clarke, R.; Wang, Y. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **2016**, *6*, 18909. [CrossRef] [PubMed]
75. Czerwinska, U. Unsupervised deconvolution of bulk omics profiles: Methodology and application to characterize the immune landscape in tumors. Ph.D. Thesis, University Paris Decartes, Paris, France, 2018.
76. Su, Z.; Łabaj, P.P.; Li, S.; Thierry-Mieg, J.; Thierry-Mieg, D.; Shi, W.; Wang, C.; Schroth, G.P.; Setterquist, R.A.; Thompson, J.F.; et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **2014**, *32*, 903–914.
77. Teschendorff, A.E.; Zheng, S.C. Cell-type deconvolution in epigenome-wide association studies: A review and recommendations. *Epigenomics* **2017**, *9*, 757–768. [CrossRef]
78. Teschendorff, A.E.; Zhuang, J.; Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **2011**, *27*, 1496–1505. [CrossRef]
79. Dirkse, A.; Golebiewska, A.; Buder, T.; Nazarov, P.V.; Muller, A.; Poovathingal, S.; Brons, N.H.C.; Leite, S.; Sauvageot, N.; Sarkisjan, D.; et al. Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment. *Nat. Commun.* **2019**, *10*, 1787. [CrossRef] [PubMed]
80. Francesconi, M.; Di Stefano, B.; Berenguer, C.; de Andrés-Aguayo, L.; Plana-Carmona, M.; Mendez-Lago, M.; Guillaumet-Adkins, A.; Rodriguez-Esteban, G.; Gut, M.; Gut, I.G.; et al. Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. *Elife* **2019**, *8*, e41627. [CrossRef] [PubMed]
81. Qiu, X.; Mao, Q.; Tang, Y.; Wang, L.; Chawla, R.; Pliner, H.A.; Trapnell, C. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **2017**, *14*, 979–982. [CrossRef] [PubMed]
82. Zhu, D.; Zhao, Z.; Cui, G.; Chang, S.; Hu, L.; See, Y.X.; Lim, M.G.L.; Guo, D.; Chen, X.; Poudel, B.; et al. Single-Cell Transcriptome Analysis Reveals Estrogen Signaling Coordinately Augments One-Carbon, Polyamine, and Purine Synthesis in Breast Cancer. *Cell Rep.* **2018**, *25*, 2285–2298. [CrossRef] [PubMed]
83. Butler, A.; Hoffman, P.; Smibert, P.; Papalexi, E.; Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **2018**, *36*, 411–420. [CrossRef] [PubMed]
84. DeTomaso, D.; Yosef, N. FastProject: A tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinform.* **2016**, *17*, 315. [CrossRef] [PubMed]
85. Kondratova, M.; Czerwińska, U.; Sompairac, N.; Amigorena, S.D.; Soumelis, V.; Barillot, E.; Zinovyev, A.; Kuperstein, I. A multiscale signalling network map of innate immune response in cancer reveals signatures of cell heterogeneity and functional polarization. *Nat. Commun.* **2019**. In Press.
86. Tirosh, I.; Izar, B.; Prakadan, S.M.; Wadsworth, M.H.; Treacy, D.; Trombetta, J.J.; Rotem, A.; Rodman, C.; Lian, C.; Murphy, G.; et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **2016**, *352*, 189–196. [CrossRef]

87. Macaulay, I.C.; Svensson, V.; Labalette, C.; Ferreira, L.; Hamey, F.; Voet, T.; Teichmann, S.A.; Cvejic, A. Single-Cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.* **2016**, *14*, 966–977. [CrossRef]
88. Risso, D.; Perraudeau, F.; Gribkova, S.; Dudoit, S.; Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **2018**, *9*, 284. [CrossRef]
89. Forget, A.; Martignetti, L.; Puget, S.; Calzone, L.; Brabetz, S.; Picard, D.; Montagud, A.; Liva, S.; Sta, A.; Dingli, F.; et al. Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. *Cancer Cell* **2018**, *34*, 379–395. [CrossRef] [PubMed]
90. Liu, W.; Payne, S.H.; Ma, S.; Fenyö, D. Extracting pathway-level signatures from proteogenomic data in breast cancer using independent component analysis. *Mol. Cell. Proteom.* **2019**, *18*, S169–S182. [CrossRef] [PubMed]
91. Teschendorff, A.E.; Jing, H.; Paul, D.S.; Virta, J.; Nordhausen, K. Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biol.* **2018**, *19*, 76. [CrossRef] [PubMed]
92. Sefta, M. Comprehensive Molecular and Clinical Characterization of Retinoblastoma. Ph.D. Thesis, Université Paris-Saclay, 2015.
93. Renard, E.; Teschendorff, A.E.; Absil, P.-A. Capturing confounding sources of variation in DNA methylation data by spatiotemporal independent component analysis. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 23–25 April 2014; pp. 195–200.
94. Ma, Z.; Teschendorff, A.; Yu, H.; Taghia, J.; Guo, J. Comparisons of non-gaussian statistical models in DNA methylation analysis. *Int. J. Mol. Sci.* **2014**, *15*, 10835–10854. [CrossRef] [PubMed]
95. Kong, W.; Mou, X.; Deng, J.; Di, B.; Zhong, R.; Wang, S.; Yang, Y.; Zeng, W. Differences of immune disorders between Alzheimer's disease and breast cancer based on transcriptional regulation. *PLoS ONE* **2017**, *12*, e0180337. [CrossRef] [PubMed]
96. Scheffer, M.; Carpenter, S.R.; Lenton, T.M.; Bascompte, J.; Brock, W.; Dakos, V.; van de Koppel, J.; van de Leemput, I.A.; Levin, S.A.; van Nes, E.H.; et al. Anticipating Critical Transitions. *Science* **2012**, *338*, 344–348. [CrossRef]
97. Mesleh, A.M. Lung cancer detection using multi-layer neural networks with independent component analysis: A comparative study of training algorithms. *Jordan J. Biol. Sci.* **2017**, *10*, 239–249.
98. Han, G.; Liu, X.; Zhang, H.; Zheng, G.; Soomro, N.Q.; Wang, M.; Liu, W. Hybrid resampling and multi-feature fusion for automatic recognition of cavity imaging sign in lung CT. *Futur. Gener. Comput. Syst.* **2019**, *99*, 558–570. [CrossRef]

3. Problems statement

The use of unsupervised methods present certain difficulties and the method of ICA has also some limits associated with the deconvolution problem. During my thesis, I tried to clarify what those problems are and attempted to solve them. It is those problems that I will list explicitly in this chapter.

3.1 Assessing relations between functional subsystems

Once a decomposition is performed with ICA, the estimated metagenes are decorrelated, and, hence, we can not use their definitions directly in order to estimate interaction (coupling) between functional subsystems identified with the use of ICA.

For example, components that capture the signal of specific immune cell types may be linked together due to the hierarchical links between their differentiation history, such as the case of T-cells, B-cells et NK cells coming from a common progenitor.

The corresponding question to this problem is: **How to assess hierarchical relations between functional subsystems represented by extracted components?**

I will try to answer this question in Chapter 4 of the results.

3.2 Decomposition's dimension choices

One of the most important and problematic parameter in unsupervised matrix factorisations is the choice of the number of components we want the method to extract from the data. This parameter is a crucial step as it assumes the number of signals present in the data and wrongly estimating it might lead to overlooking the existence of some important signals if we underestimate it and when overestimated it could cause the disruption of previously found signals. As described previously, certain methods were developed to estimate this number but they can still disagree with each other, as shown in Figure 4 of (Bac et al., 2021).

The corresponding question to this problem is: **How to correctly estimate the amount of components to extract from the data such that we are sure we aren't missing signals of interest?**

I propose an answer to this question in Chapter 4 of the results.

3.3 Limits of the deconvolution granularity level

Unsupervised deconvolution methods cannot hope to comprehensively characterise such a complex system as tumors. There is a limit to the amount of details that a deconvolution (both supervised and unsupervised) can hope to achieve, either due to the quality of the data itself or due to the mathematical approaches used to extract the signals. When performing exploratory data analyses, it is interesting to extract as much details as possible but due to the unsupervised property, it is often difficult to ensure that the obtained signals are of good quality and correspond to real biological signals instead of being the data analysis artefacts.

The corresponding question to this problem is: **How can we know the limit of the maximal number of signals an unsupervised deconvolution method can hope to extract?**

I try to answer this question in Chapter 4 of the results.

3.4 Interpretation of unsupervised components

Signals found through unsupervised deconvolution do not offer any direct biological interpretation. It is therefore an important step to assign to each component the biological functions (cellular functional subsystems) they capture. Methods used to label components often use definitions of biological functions represented as a set of functionally related genes. But in reality, genes are known to participate in diverse functions, making annotation of components difficult. Moreover, assigning a function does not always provides information about its state (activity) or its effect for tumor development.

The corresponding question to this problem is: **How can we interpret unsupervised components to understand the function they mirror in a given context?**

A solution to this problem will be proposed in Chapter 5 of the results.

II - Results

4. HACK: Hierarchical Analysis of Component linKs, a tool for multilevel latent factor exploration in omics data

Nicolas Sompairac, Maria Kondratova, Nicolas Captier and Andrei Zinovyev.

To be soon submitted to a peer-reviewed journal.

4.1 Ideas on how to solve some limitations attached to unsupervised deconvolution

4.1.1 Tree-dependent Component Analysis (TCA)

One of the inherent strengths of ICA is its capacity to extract independent signatures from the data, thus allowing each component to be free of eventual noise. However, this is also a weakness when trying to find interactions between these signals. As a metaphor, if we were to analyse a symphonic orchestra, while each type of instrument can be extracted and analysed independently of the others, their activities are in fact not mutually independent. Each instrument may play its own music score, but some instruments still interact with each other through similar patterns. The existence of such interactions is also expected to be found in tumours, where groups of cells can interact and participate in tandem to respond to the same signal but still have different expression profiles. This can be observed for example in the sequential gene activation of the different phases during the cell cycle (Whitfield et al., 2002). While these clusters of genes have their own expression pattern, their activation and inactivation patterns are related between each other.

To recreate such relations between independent components from ICA, (Bach and Jordan, 2003) proposed an algorithm that aimed to arrange components in a structure organised in clusters of components that are dependent within a given cluster but independent between clusters. The proposed method named Tree-dependent Component Analysis (TCA) tried to achieve this by relaxing the independence assumption of ICA and trying to fit a tree-structured graphical model (Chow and Liu, 1968).

The ICA model can be seen as:

$$s = Wx = (s_1, \ldots, s_m)^T$$

where we search for a linear transform W such that each of its components are as independent as possible.
The TCA model is a direct equivalent but W has now to be constrained such that components can be modelled as a tree-structured graph.

The TCA algorithm is shown in Figure 4.1 and can be simplified as follows:

1. Initialise the W matrix using ICA
2. Alternate minimisations:
    a. Minimising with respect to T using the Chow-Liu algorithm (maximum spanning tree problem)
    b. Minimising with respect to W using a steepest descent with line search

---

Input **:** $data\{x\} = \{x^1, \ldots, x^N\}, \forall n, x^n \in \mathbb{R}^m$

Algorithm **:**

1. $Initialization : W\ random$
2. $While\ G(W) = min_T F(W,T)\ is\ decreasing$

$$for\ i = 2\ to\ m, for\ j = 1\ to\ i-1, W \leftarrow arg \min_{V \in L_{ij}(W)} \{\min_T F(V,T)\}$$

$$where\ L_{ij}(W)\ is\ the\ set\ of\ matrices\ V \in M\ such\ that$$

$$(a) \forall k \notin \{i,j\}, V_k = W_k$$

$$(b)\ span(V_i, V_j) = span(W_i, W_j)$$

3. $Compute\ T = argmin F(W,T)$

Output **:** $demixing\ matrix\ W, tree\ T$

Figure 4.1. TCA algorithm.
$F(W,T)$ is the contrast function and M is the search space for W.

By alternating the minimisation steps, we can ensure that we keep a certain amount of independence between each component while still being able to reconstruct the relations between them.

The TCA algorithm has been successfully applied to various types of data such as signals extraction from cosmology images (Kuruoglu, 2010) as well as genetic data (Kim and Choi, 2006).

4.1.2 BIOBOMBE

While the previously described method of TCA tried to solve the limitation of missing relation links between components, we were still left with the problem of finding the optimal number of latent dimensions *k* in the data. However, trying to find a single "optimal" dimension might not be the correct solution since it may happen that different biological signals can be captured at different dimensions. To test this, a methodology called BioBombe (Way et al., 2020) was developed and consisted in decomposition the data across a series of dimensions going from the order 2 up to 200 (Figure 4.2).

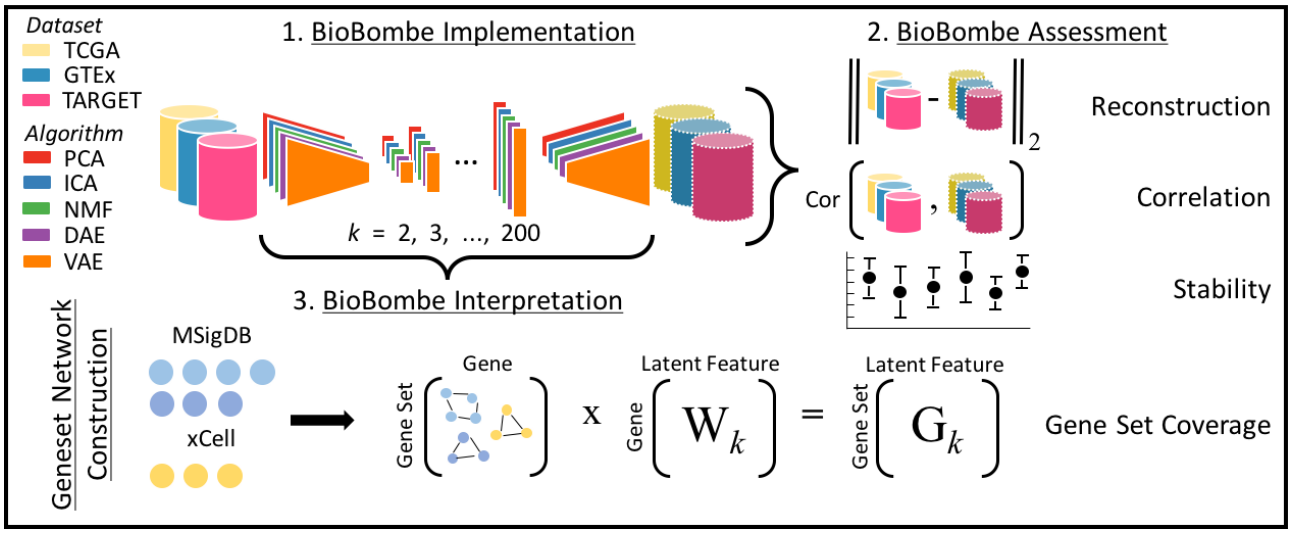


Figure 4.2. Overview of the BioBombe approach.
Step 1 consists in decomposing datasets in an increasing number of components using various methods. Step 2 is used to assess the quality of each component by filtering unstable ones and grouping similar components as a single one. Step 3 is the biological interpretation of found components using a pathway projection approach. Reprinted with permissions under the terms of the Creative Commons Attribution License 4.0 (CC-BY) from (Way et al., 2020).

The conclusions of the BioBombe article were that: (i) different features were found at different dimensions and (ii) the optimal dimension for a given feature differed between decomposition methods.

The first point is crucial because it implies that certain features can only be found at a given decomposition order and may be missed if choosing a dimension too high or too low.
The second point implies that even if one wanted to use one of the many available heuristics to choose a single best dimension, this dimension wouldn't be generalisable to all decomposition methods and would have to be fit specifically.

4.2 Context behind the Hierarchical Analysis of Component linKs (HACK) tool

By looking at the literature, it becomes clear that a simple decomposition for a given number of components might prove too limited in its scope and the user might lose sight of some interesting features. To avoid this, we can instead look over a range of dimensions but by doing so we are losing the structure between found features and may also end up with some redundancy.

In the context of an immune infiltration analysis in a tumor, to make sure that a certain cell type is present or not is of great importance for treatment recommendations. Missing an immune related feature is therefore not acceptable so finding the right parameter for the number of signals is a priority.
However, if we were to increase the order of decomposition, we could expect a better feature granularity. While this is most interesting, since we are using unsupervised methods, the interpretation step might prove difficult because of the lack of clear markers to define those specific features. This is why having an idea of the relations

and structure between different components might be handy when trying to understand their biological meaning.

One way to achieve this is to follow the becoming of each component across the different decomposition order. In fact, as observed in (Nielsen et al., 2005), we can expect 4 different component behaviour when increasing the decomposition order (Figure 4.3):

- A: a component may stay the same and continue to be reproduced with each increasing order.
- B: a new component may appear when a new feature is found at a higher order
- C: a component may be split in two. These new "daughter" components will therefore be related to their "mother".
- D: two components may be split in two. These type of behaviour is mostly expected in cases where we reach a limit in feature extraction and the method starts to create artificial new components by forcefully mixing existing ones.

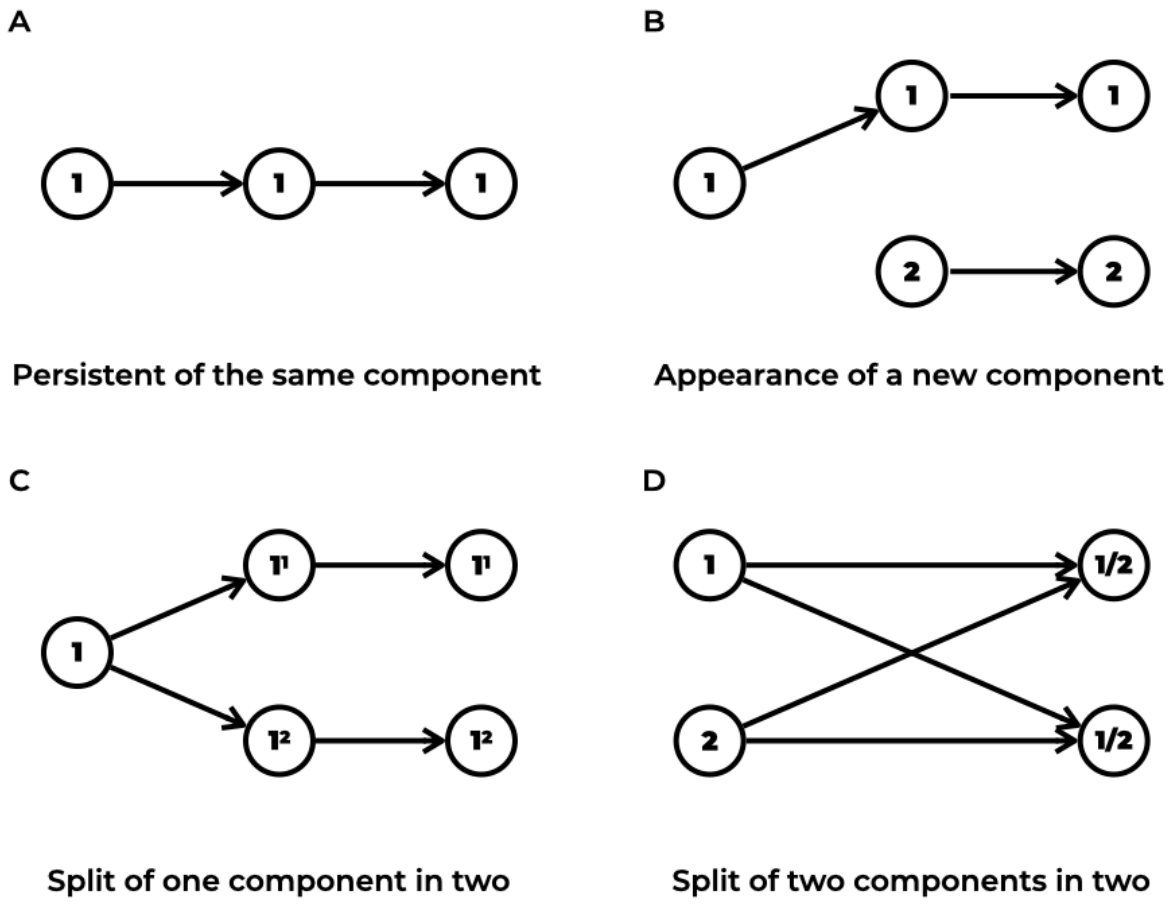


Figure 4.3. The 4 possible behaviours of components when increasing the decomposition order.
The numbers denote the origin component and its possible change in content.

These types of behaviour make sense mathematically but they also make sense biologically. It is expected when using deconvolution methods, that by increasing the number of components we are increasing the granularity of found features. Therefore, by reconstructing such links, we are able not only to reconstruct the history of a given component (meaning at which decomposition order it emerges and where it disappears), we are also able to infer its meaning based on the component of origin.
By focusing on the type of events happening with each increasing order, we are also able to judge the limit of the decomposition method. If after a certain point we don't see the appearance of new components or if there is a majority of events such as the Figure 4.3D, we can assume that we have reached the capacity of the method to extract relevant features.

It is with these intentions in mind that I have developed a method named Hierarchical Analysis of Components linKs (HACK). It is a "hierarchical analysis" because the method follows the fate of each component while increasing the decomposition order step by step. The mention of "components" is related to the fact that the HACK approach can be applied with any matrix factorisation methods. Finally, by checking the similarity between each component with each step, we are able to keep track of the "links" between them. These links, represented by a similarity score, can give more insight into the used decomposition method mechanics

4.3 Manuscript

HACK: Hierarchical Analysis of Component linKs, a tool for multilevel latent factor exploration in omics data

Nicolas Sompairac, Maria Kondratova, Nicolas Captier and Andrei Zinovyev.

Introduction

Recent progress in biotechnologies allows us to characterize biological systems by molecular (omics) profiling of increasing complexity. Multidimensional molecular profiles of biological samples such as tumors require application of data analysis methodology aimed at reducing data dimensionality and granularity. One common analytical tool frequently applied for this purpose is clustering. Clustering aims at grouping the biological observations according to their multidimensional measurements, which is standardly used for reducing the omics dataset complexity. An alternative approach consists in identifying latent low-dimensional representations of individual objects, such that proximity in the latent space would reflect the similarity in high-dimensional space of complete molecular descriptors. One of the commonly used approaches having this objective is related to matrix factorization and the model of linear mixture of signals. Accordingly to this model, we assume that the observable multidimensional variables are weighted sums of relatively few intrinsic factors characterized by their activities in a given biological specimen. Clustering and matrix factorization approaches are complementary methods since the reduced low-dimensional latent space can embed an arbitrary number of clusters.

A tempting idea is to interpret the identified latent factors as molecular programs or functional subsystems whose activity determines any biological systems state [1]. The nature of these subsystems can be related to the actions of important transcription factors, systems reacting to environmental conditions, actions of various extrinsic perturbations such as drugs, adaptation to genome modifications, existence of heterogeneous cell types or states in a biological specimen. In reality, the intrinsic factors can be also related to various biases, such as batch effects, imposed by the measurement technology, adding extra dimensionality to the latent representation.

This way of tackling the complexity and dimensionality of molecular measurements appears to have advantages compared to clustering, establishing the most informative combinations of initial data variables according to which the measured samples can be further grouped or ranked [1]. To mention few successes, a list of molecular programs was established in a large-scale analysis of transcriptomic data [2,3], or reproducible latent factors shaping the molecular heterogeneity of certain cancer types have been identified [4,5,6,7,8,9]. Matrix factorization was constructive in characterizing the gene expression heterogeneity of rare and genetically stable cancer types at single cell level [10] or studying the molecular basis of comorbidity between different diseases [11]. Matrix factorization was an insightful tool to study the epigenetic or multi-modal heterogeneity of biological samples [12,13,14]. In a recent large-scale study, a concept of recurrent patterns of heterogeneity (RHP) has been

established using the large- scale single cell transcriptome data analysis in a large collection of cancer cells, through systematic application of matrix factorization and cross-validation of the resulting molecular program definitions [15].

In all of these studies, the most important parameter of data analysis is the number of the molecular profiles to be identified, just as the number of clusters is the most fundamental parameter of the clustering approach. Here one usually has to deal with finding a balance between the level of granularity which should be adequate to represent the system complexity and the computational robustness of the definitions of functional subsystems [16]. Since determining the true intrinsic dimensionality of a dataset is a challenging task [17,18], one has to rely more on the utility of the identified factors (such as prognostic value in independent measurements) or the possibility of biological interpretation [5,6]. Various approaches have been suggested to determine the optimal number of matrix factors, based on various principles [18,19].

Despite all these efforts, in the practice of data analysis using matrix factorization, it has been noticed that the most informative factors might emerge at some decomposition orders and do not exist at others [18,1]. A typical example of such a scenario are latent linear factors related to the presence of immune cells in the tumor microenvironment. Lower order decompositions reveal the existence of immune infiltrate, combining many immune cells types, and allows one to quantify its presence [5,18], while higher order decompositions give rise to the latent factors which reflect the presence of more specific immune cell sub-populations (such as T-cells, B-cells, myeloid cells, etc.).

Since it is not clear a priori which description granularity will appear the most useful in a particular task, it is tempting to use several decomposition orders simultaneously. In a recent study, BioBombe tool was suggested which exploited decomposition of a dataset across a range of component numbers (e.g., from two to a hundred), using several matrix factorization methods (NMF, PCA, ICA, variational autoencoders) and analyze the entire set of generated latent factors with respect to their clinical utility or biological interpretability [20]. Even if being pragmatic and powerful, such an approach ignores the redundancy existing across factors identified at multiple decomposition orders and creates a problem of multiple testing which is especially difficult since the identified factor set contains strongly dependent components.

In the current study we suggest an approach named Hierarchical Analysis of Component linKs (HACK) which aims at improving the selection of informative linear latent factors resulting from application of matrix factorization at different decomposition orders. HACK relaxes the problem of choice of the right number of components by considering multiple levels of decompositions simultaneously and carefully reconstructing the redundancy structure between them. As a result, HACK generates a relatively small number of so-called persistent components which are the factors that can be reproduced across a sufficiently large range of consecutive decomposition orders. Importantly, when applied to gene-based molecular profiles, the persistent component represents a generalization of the metagene notion, such that each gene in it is characterized not only by a weight but also by a built-in uncertainty estimate. This uncertainty is quantified as a measure of the robustness of a gene to significantly contribute to the definition of a factor despite certain variations in the decomposition order.

HACK is packaged in Python as a user-friendly tool, accompanied by a graphical interface such that a user can interact with the hierarchy of analyzed intrinsic factors. HACK can be readily used with any matrix factorization method. We demonstrate the advantage, compared

to the fixed order decompositions, of its use in applications to bulk and single cell RNASeq profiles.

Materials and Methods

HACK workflow general outline

HACK approach performs the following sequence of data analysis steps, in order to determine the persistent components (Figure 1).

1. Performing matrix factorizations across a range of decomposition orders
2. Construction of the graph of relations between components extracted at different orders
3. Filtering of the full relation graph using the Mutual Nearest Neighbours algorithm
4. Selection of persistent components
5. Simplification of persistent components and splits reconstruction
6. Extraction of average persistent components
7. Performing an “uncertainty” analysis to assign “certainty” gene scores

Terminology

Component: vector obtained as output of the matrix factorisation step. In the transcriptomic context, the matrix factorisation can be performed on genes or samples and the components obtained will be called **metagenes** or **metasamples** and will correspond to vectors of weights associated to each gene or sample.

Persistent component: during the hierarchical analysis, some components are found consistently across multiple decomposition levels as shown by their high correlation maintained with each step.

Split: event during the reconstruction of a component history when the information contained in the previous step is separated into two new but related components.

Branch: in the graphical representation of the hierarchical analysis, a persistent component will be called a branch due to its tree-like structure. In the graph, a branch corresponds to a set of persistent components.

MNN: Mutual Nearest Neighbour algorithm used to find only reciprocal best common factors. The algorithm consists in checking all corresponding correlations between two sets of factors and keeping a given number K of links if the correlation is maximal between a factor from the first set and another from the second set in both directions.

Algorithm for hierarchical analysis of the multilevel matrix decomposition

Performing matrix factorizations across a range of decomposition orders

The first step of the algorithm consists in performing multiple decompositions across a range of dimensions using the chosen algorithm. In this study, we applied the ICA algorithm on an increasing range of orders, going from 2 to 100 by increment of 1. ICA is used but any other matrix factorisation method could be applied instead. To counteract the stochastic behaviour of ICA, we used a stabilisation step where multiple runs of ICA are performed with different initialisations and an average result is obtained by clustering the resulting components. We used the implementation of stabilized ICA in Python [https://github.com/sysbio-curie/Stabilized_ICA], which is based on the original MATLAB version of ICASSO [21].

We also used PCA and NMF methods implemented in the scikit-learn Python package [https://scikit-learn.org].

Construction of the graph of relations between components extracted at different orders

The second step is the construction of a complete relation graph between components of adjacent order (i.e. between order K and K+1). For this, we used the Pearson Correlation as a relation score between each pair of components of adjacent order. As a result, we will obtain a graph with nodes being component vectors connected by edges with the Pearson Correlation as a score. This full graph will be the basis of further analyses.

When looking at the behaviour of components along the decompositions orders, we can identify 4 types of states: (i) persistent high correlation with each step, (ii) appearance of a new component, (iii) “split” of a component in two, (iv) “cross over” between two pairs of components (behaviour mainly due to the stochasticity or ICA). Components presenting a persistent and unique high correlation will be further addressed as “Persistent Components”.

Filtering of the full relation graph using the Mutual Nearest Neighbours algorithm

The whole relation graph is heavily encumbered and therefore can’t be directly used for visualisation purposes. This is why a filtering step has to be performed. Since we are interested mainly in persistent and split behaviours, the filtering will be performed using the Mutual Nearest Neighbours (MNN) algorithm with minimal number of outgoing edges and a minimum value for the correlation threshold. The MNN algorithm forces kept edges to be maximally correlated in a reciprocal manner. Also, to avoid recreating splits by keeping low correlation edges, a MNN_Gap parameter can be used to decide if the second edge is to be kept if it has a correlation difference of, by default, less than 1.5 times the value of the first correlation. This step allows not only to remove most of the noise from the complete relation graph but also to keep eventual splits between components to retrace their origin history.

Selection of persistent components

For this step, all splits will be temporarily removed to extract only persistent components and allow a check of their persistence.

The persistence of each set of components can be determined by the number of decomposition orders they are found consistently across. This value can also be viewed as a stability or reproducibility value of the set of components. To remove the least stable components, a threshold must be set so that all components that haven't been found consistently for that amount of decomposition orders in succession will be pruned. The user can play with this length threshold to keep a more or less sparse graph.

Simplification of persistent components and splits reconstruction

To ease up the visualisation of the resulting graph, each persistent component is simplified by their starting and ending component for the set they are part of. The resulting edge between those components will have the average values between the starting and ending components. By doing so, we can directly see if the final component hasn’t diverged too much from the initial one.

Splits that have been previously deleted will have to be reconstructed. Just as the filtering step, to avoid reconstructing low relevance links, parameters for the MMN algorithm such as correlation threshold, K outgoing edges and a gap will be applied here as well. However, to avoid adding links between persistent components that are far apart, a penalty is applied on the number of decomposition orders separating corresponding components, thus penalising long distance edges and favouring closer links. This penalty factor is mainly used to avoid encumbering the graph with superfluous edges that would make the visualisation of results more complicated.

Extraction of average persistent components

As a result of this algorithm, we obtain a forest-like graph represented by multiple trees containing one or several branches. The whole graph can be visualised as is or it can also be exported as a matrix of metagenes where each column will be a persistent branch and each row will contain genes with associated scores chosen by the user. The average metagenes can be constructed by taking the standard scores given by the decomposition method or by taking the Z-scores. It is also possible to use the average ranks of genes by taking their absolute position in each component of the branch and calculating the average position along the whole branch. This matrix allows to arrange genes in a manner that makes visible genes that are always top contributing in the corresponding signal and moving those that are fluctuating in the middle.

Performing an “uncertainty” analysis to assign “certainty” scores

A persistent component being a set of individual components, it becomes possible to perform additional analyses to check how a certain gene's score behaves along the decomposition dimensions. When looking at how certain gene's scores behave, we have noted that a large portion of genes had large score variations, making them unfit to use as reference drivers of a given component.

To leverage this behaviour, we have implemented an algorithm to take into account the erratic variation of certain genes and assign a corresponding "confidence" score. The algorithm is as follows:

- Take non-simplified persistent components from the filtering step of the HACK algorithm containing a list of each intermediate component and merge them into a single matrix per persistent component
- Transform genes scores of each intermediate component into ascending position rank values
- Normalise ranks to values between -1 and 1
- Simplify normalised ranks based on a threshold: -1 if below the negative threshold, +1 if above the positive threshold and 0 if between the negative and positive threshold.
- Collapse each intermediate component into a single one. If a gene always presents positive or negative value only, calculate the average position rank value. If not, assign a score of 0, symbolising that the gene is uncertain.

This algorithm results in a matrix similar to the one obtained in step 8 but contains instead all persistent components with an assigned gene value between -1 and +1. The closer a gene is to the value of +1, the more certain we can be that this gene is a positive driver of the component along the decompositions orders. And inversely, if a gene is close to the value of -1, the more certain it is that he is in the negative part of the component. Each gene having a value of 0 can be considered as uncertain and shouldn't be used in further analyses because of its unstable behaviour.

Parameters of the HACK algorithm

HACK algorithm contains 10 parameters having the following meaning and default values:

1) **Minimum_decomposition** and: minimum order of decomposition the hierarchical analysis will be performed from. Default value is 2.
2) **Maximum_decomposition**: maximum order of decomposition the hierarchical analysis will be performed from. Default value is 2.
3) **MNN_K**: maximum number of reciprocal components' links to keep between two decomposition orders for the MNN algorithm. Default value is 2.
4) **MNN_Gap**: maximum difference factor between kept links. If the difference in the similarity scores is higher than the value multiplied by the gap, the link is not kept. Default value is 1.5.
5) **MNN_Correlation_Threshold**: minimal correlation value between components to keep. Default value is 0.3.
6) **Minimal_Persistent_Length**: minimal number of consecutive decomposition orders a particular component is found in to be considered as persistent. Default value is 10.

7) **Split_K**: maximum number of outgoing splits from a persistent component to keep. The default value is 2.
8) **Split_Gap**: maximum difference factor between outgoing splits similarity scores. Default value is 1.5.
9) **Split_Correlation_Threshold**: minimal correlation value between components to keep. Default value is 0.3
10) **Split_Penalty**: Penalty score applied to a similarity score based on the difference of decomposition orders that separates two persistent components to avoid distant split reconstructions. Default value is 0.05.

HACK Python package and interactive tool

We implemented the Hack algorithm introduced above as a Python package openly available at [https://github.com/sysbio-curie/HACK]. The package contains a Jupyter notebook with a possibility to construct and interact with the graph of links between the persistent components.

The data to be used in this package can be Bulk or Single-cell and can be composed of RNA counts, microarray measurements or any other omics type.

A preprocessing is however required before it can be used in the workflow. Some basic preprocessing functions are integrated in the package that allows to log transform the data, mean center the rows as well as remove duplicates. These are the required preprocessing steps for an optimal application of the HACK algorithm but other preprocessing methods can be applied if needed.

As a result, the HACK workflow will generate decomposition matrices across a range of decomposition orders. These matrices will be saved for further analyses during the workflow allowing a faster and easier exploratory analysis in case the user wants to change the default parameters. Some intermediate steps such as the fully connected and the MNN filtered hierarchical graphs will be saved in files and if all steps are performed, the final hierarchical graph of persistent components will be exported as a file as well. These files can be easily opened in network visualisation applications such as Cytoscape.

For the last two steps of “Extraction of average persistent components” and “Performing an uncertainty analysis to assign certainty scores”, matrix files will be created containing persistent components as columns and genes as rows with their corresponding assigned scores based on the chosen method.

Visualisation and analysis of the resulting graph of relations between persistent components

Once the graph has been generated and filtered accordingly, it becomes possible to query it for various biological signals. Since each branch can be viewed as a set of components, they can be simplified by an average vector. This average vector can then be used for any required analysis.

The obtained graph can be visualised with the help of an interactive Jupyter Notebook. The generated figure represents the persistent components segments from the bottom-up where each branch is represented by the starting node located at the Y-axis corresponding to the decomposition order this component was first found and an end node for the decomposition order it was stably found last accordingly. Plain vertical edges correspond to persistent components while dashed edges correspond to reconstructed splits.

With ICA, to interpret a component, it is standard procedure to compare the correlation of a component with a reference metagene. For reference, Biton's metagenes [5] have already been implemented in the method and can help identify 11 different signals such as Immune infiltration, Cellular stress or Cell cycle. The user can visualise either the Pearson correlation of each component with the selected Biton's metagene or can select to colour edges accordingly to the Pearson correlation between each component. As an example, we can filter persistent components related to Immune infiltration as seen in Figure 2C.

Another usual procedure is to extract top contributing genes of a certain component and perform enrichment analyses. This can be done on the exported matrices from step 8 but it is also possible to query genes directly on the graph and select branches with components containing particular genes in their top contributing ones. Genes considered as top contributing can be set by using a Standard Deviation threshold in a component's score. The nodes will be coloured accordingly to the fraction of found query genes in the top contributing gene list of a component.

Results

Comparing HACK algorithm results for various decomposition methods (PCA, ICA, NMF)

The proposed HACK algorithm is technically applicable to any matrix factorisation method able to give weighted vectors (aka components) as an output. To test this, we applied the Hierarchical Component Analysis algorithm to PCA, ICA and NMF decomposition methods on the TCGA BRCA RNA-seq dataset.

As described previously, the first step is common to all methods and requires decomposing the dataset in gradually increasing numbers of components. We chose decomposition orders from 2 to 100, incrementing by 1 with each step. It is known that NMF components tend to capture the average gene expression signal [6]. One simple and straightforward step is to "regress-out" this signal by performing a linear regression between each component and the average gene expression vector and then taking the residual as a "cleaned" new component. This additional step can be performed for NMF decompositions only to get rid of the superfluous correlations between each and every NMF component. However, this step was ignored in our analysis because it didn't produce any meaningful difference in the result and the Split_K parameter of HACK was used instead to compensate for this.

To add some challenge for a PCA analysis, we removed 5% of random samples for each decomposition order to generate some instability in principal components which would have been deterministic otherwise. Once this decomposition step has been performed, we can then

apply the HACK algorithm to visualise the behaviour of components obtained from different methods.
When looking at the resulting graph in Figure 2A, we can see that PCA on resampled data presents some instabilities in its principal components (PC) but some are still reconstructed with the HACK algorithm. We can see that the first PCs are the most stable ones and can be consistently found throughout the whole over-decomposition process, as we can see with two PCs related to an Immune Infiltration. With each increasing order, new components tend to become less stable and their link reconstruction becomes more difficult. However, certain PCs can capture certain signals more consistently, even at higher orders of decomposition.

The ICA graph lets us see various behaviours. First, we can see that the graph is quite furnished and contains a large number of persistent components. Also, the first components found tend to be the most stable/persistent. Just as with the PCA, we can also find the two ICs related to an immune infiltration as seen in Figure 2C

For the NMF, we noticed a striking difference with the previous two: the graph is much sparser (Figure S1). This could either be explained by the instability of NMF components or by a latent inter-correlation between components that disrupts persistency calculations during steps 4 and 5 of the HACK algorithm. Since we expected this inter-correlation between components to be removed by regressing the average gene expression, this shouldn't be the cause of such a resulting graph.
However, to test if this is true, we can play with the K parameter of the MNN in step 3 of the HACK algorithm. By setting it to 1, we don't allow splits to be taken into account during the filtering of the initial full correlation graph, thus eliminating all possible inter-correlations between components of adjacent orders. The obtained graph from this parameter change can be seen in Figure 2B . This graph resembles much more the one we found with ICA and contains many persistent components and behaviours common with ICA components. It is also in this case that we are able to find the two Immune components from which one was absent from the initially obtained graph with the parameter K=2.
This difference from the initial graph could imply that even though we removed the average gene expression signal from each component, another signal is still present across most NMF components. In NMF, all components relate to each other. The reciprocity constraint of the HACK algorithm helps counter it to some extent although not entirely. But if this constraint is relaxed, this inter-correlation feature emerges again.

Application to transcriptomic datasets

Reproducibility analysis of persistent components extracted with HACK

To make sure that persistent components found through the HACK algorithm are robust and reproducible across different datasets, we performed a meta-analysis in a similar fashion as [6] using the same CRC datasets obtained from [22]. Only datasets with the number of samples above 100 were kept, leaving us with 12 CRC datasets in total. Through this analysis, we were able to demonstrate that not only persistent components were consistently reproducible, they were even more so than "simple" ICA components. Indeed, as seen in

Figure 3A, the network of persistent components is arranged in a more compact and clustered fashion compared to networks of simple 50 and 100 components.

To confirm this in a meaningful way, we have computed persistent components across the 12 CRC datasets and have generated MNN graphs using all the persistent components from all the datasets. We also performed a simple stabilised ICA decomposition of order 50 and 100 using these datasets and have generated MNN graphs as well. We then compared the obtained networks.

For comparison scores, we counted the total number of reproducible components as those that are connected to a minimum number of K other components. We will refer to this K as the connectivity measure. To check for the strength of the reproducibility, the reciprocal correlation value S between components was used a threshold. By looking across the range of K and S values, we were able to observe a clear difference in reproducibility between persistent and simple components.

In Figure 3B, we set the connectivity K to 2, thus keeping only components that were found in at least 3 different datasets, and looked at how strong the reproducibility of those components was for more stringent values of S. Starting from a correlation threshold of 0.3, the number of reproducible components falls sharply but the number of reproducible persistent components manages to stay above the other two simple reproducible components as shown by the ratio of the number of reproducible persistent and simple components.

For the 3 MNN networks, for a set value of S of 0.6 corresponding to a strong correlation coefficient, we computed across a range of K from 1 to 12: (i) the scores for the total number of reproducible components, (ii) the fraction of reproducible components, (iii) the mean reproducible score corresponding to the means correlation threshold of outgoing edges from a single component across all components and (iv) the sum of reproducible score. As seen in Figure 3C, persistent components have scores significatively above simple components and can therefore be considered as more reproducible.

Meta-analysis of immune-related persistent components in colorectal cancer bulk transcriptomes

One set of components is of particular interest, these are immune-related components. In the context of cancer-immunotherapy the development of immune-classification methods for tumours is an important practical task. Many supervised and unsupervised methods of immune tumor classification focus on determining the cellular composition of the tumor [23,24]. However, this approach has its drawbacks, since there is no unambiguous correspondence between the leukocyte formula of the tumor and its immunogenicity. It is known from the literature that the presence of the same cell types, for example, Th17 or B-cells, can make both a positive and a negative contribution to the success of anti-cancer therapy [25,26]. We believe that a hierarchical approach of ICA is capable of capturing more complex, integral characteristics of the tumor microenvironment and can be useful for solving this problem.

To test this hypothesis we have applied HACK methodology to the 12 colon cancer dataset described before from [22]. Each dataset was decomposed up to 100 components, hierarchical trees were generated and persistent components were filtered from “uncertain” genes.

Observing trees obtained with the HACK approach really increases the resolution of a simple ICA methodology. For example, the low level of component deconvolution TCGA dataset represents just one stable component enriched by immune-related genes, and when the number of components increases we can already see several immune related components representing different aspects of tumor immunity (Figure 4A). And if the metagene from one immune component could provide only “one-dimensional” tumor classification (“cold” or “hot”), then several immune related metagenes potentially give us a chance to develop “multidimensional” immune characteristics of the tumor. Similar results - one “root” immune component divided into several immune “branches” - were also observed in other colon-cancer datasets.

When we project the metagenes scores extracted by HACK from bulk data (GSE39582 dataset used for the illustration) to colon-cancer related single-cell data [27], we can demonstrate a correspondence between immune “branches” and cell types. As we can see, the projection of the “root” metagene highlights all immune cells in the single-cell data, and the “branches” projection highlights particular cell type groups. For instance, projection of branch “a” highlights mostly B-cells, branch “d” - cytotoxic CD8+ T-cells, etc (Figure 4B).

Now we should find a way to characterise the nature of difference between these components and propose possible biological interpretations. To do this, we have used a set of specific knowledge based signatures representing different sides of immune response (professional antigen presentation, cytotoxicity, phagocytosis) as well as cell type specific signatures (T-cells, B-cells, NK, DC etc). We have applied these signatures for clustering of top-contributing genes from “roots” and “branches” of immune components in 12 colon-cancer data-sets (Figure 4C). As expected, all “root” components have almost the same pattern of regulation of different functional modules, then “branches” form several clusters with different groups of functional activations. The clusters have the following functional characteristics:

1. Rather high presence of innate immune-cells, low presence of both T-cells and B-cells.
    a. Upregulated pathways: NO-ROS production, Phagocytosis, Activated immune-checkpoints.
    b. Downregulated pathways: Cytotoxicity, Antigen presentation.
2. High presence of B-cells T-cells as well as basophils and mast cells, low presence of the rest innate immune-cells.
    a. Upregulated pathways: Inhibiting immune checkpoints.
    b. Downregulated pathways: NO-ROS production, Phagocytosis, Activated immune-checkpoints.
3. High presence of T-cells and innate immune cells, low presence of B-cells and basophils and mast cells.
    a. Upregulated pathways: Th1, Cytotoxicity, Antigen presentation.
    b. Downregulated pathways: Th17, Phagocytosis, Activated immune-checkpoints.
4. High presence of T-cells and NK, low presence of B-cells and innate immune cells.
    a. Upregulated pathways: Th1, Cytotoxicity, Antigen presentation.

b. Downregulated pathways: Treg, NO-ROS production, Phagocytosis, Activated immune-checkpoints.

We assume that cluster 4 corresponds to a cytotoxic immune response in tumors and cluster 2 is related to humoral immune response in tumors. The balance between these two sides of adaptive immune response could be a predictive factor of tumor immunogenicity, but testing this hypothesis requires additional studies. However, even these preliminary results allow us to assert that HACK is a powerful method of tumor classification for immunological and immunotherapeutic studies

Using HACK for meta-analysis of the single cell scRNA-Seq profiles from Cancer Cell Line Encyclopedia

The paper of [15] showed that through a matrix factorisation such as NMF, it was possible to extract and identify specific Recurrent Heterogeneous Programs (RHP) within multiple cancer cell lines. We decided to test a similar approach using the HACK method to see if we can obtain similar results.

We took the datasets used in the [15] paper and performed the HACK workflow on all cancer cell lines containing at least 100 cells (125 cell lines were selected). The obtained persistent components were then merged together in a MNN network and clustered using the MCL clustering tool [28] on Cytoscape [29].

We can see in Figure 5 the obtained clusters across all cell lines. Most clusters were composed of a mix of cancer types but some were still more enriched in certain types such as cluster 16 being mostly composed of Head and Neck cancer cell lines and cluster 13 being composed of only Melanoma cell lines.

To compare our finding with the RHPs, we took the 20 most significant genes of each cluster and counted how many were in common. By taking a threshold of a minimum 50% intersection, we found 8/10 clusters corresponding to the RHPs, with the two not found programs being p53-dependent senescence and EMT-I (melanoma).

It is also interesting to note that compared to the 2 programs related to cell cycle found in the reference paper, we were able to find 3 more (Clusters 1, 2, 4, 8 and 17), showing that HACK applied with ICA is able to extract more specific signals of the cell cycle than NMF.

Discussions and Conclusions

Compared to a standard matrix factorisation method, a hierarchical approach performs multiple decomposition of increasing order and creates a relationship graph between each adjacent order. This graph can later be used to have additional information about independent components such as their stability or their evolution history along increasing decomposition orders.

While usual single-shot matrix factorisation applications focus on the selection of a single order of decomposition, with a hierarchical analysis, multiple orders are visualised, allowing to see which components appear or disappear after a certain dimension and can help capture relevant signals more easily. As noted with the BioBombe approach, different signals are

better captured in different dimensions and no single decomposition order is enough to capture everything efficiently [20].

By applying a hierarchical approach, we are able to generate a graph in a forest fashion. This forest consists of various trees where each branch corresponds to a series of individual components, giving us a more detailed view and new possibilities that were impossible before by simply looking at individual components.
For instance, we can judge the robustness of a certain signal by its length, i.e. number of decomposition orders it was found in. By setting the length of a branch, we can ensure that we keep only the most robust ones and exclude the noise from the graph due to eventual stochastic behaviour of a decomposition algorithm.

By taking this over-decomposition approach, we also remove the need to select a single "optimal" decomposition order and risk losing interesting signals. As seen in Figure 2C, if we took a single decomposition corresponding to the earliest order we found the first immune component, we wouldn't have been able to find the second immune signal. And in some other cases, if we overshot by choosing a decomposition order too high, a signal wouldn't be stable enough to be found reliably and would therefore be missed entirely.

Thanks to this over-decomposition view, one could also use the resulting graph as a guide to see the limits of the extractable signals via the applied decomposition method. If after a certain decomposition order, no new branches are observable, this means that the method has reached its limits and probably won't be able to extract more signals from the data, even if we were to increase the order of decomposition.

Additionally, instead of having a single metagene related to a single component, it becomes possible to compute an average metagene by taking into account all individual components along the branch. This results in a cleaner metagene, corrected for eventual biased local results. The cleaned metagene can then be used for usual post-decomposition analyses such as interpretations, enrichments or marker extractions.

Taking a branch as a list of individual components leaves us also with the opportunity to add an "uncertainty" score to each gene. Based on their behaviour along the persistent component segment, genes that have a steady behaviour and stay as top contributing, can be noted as "stable", whereas those who can have slight variations can be marked as "unstable". This can help choose genes in each signal with a level of certainty without risking selecting those that could be present only in a single order of decomposition.

Data availability

For this paper, several publicly available datasets were used:

- Breast cancer RNA-seq data is available on the TCGA portal or can be directly downloaded from the corresponding Zenodo repository alongside all the stabilised ICA deconvolution matrices.
- CRC RNA-seq datasets are available via the Synapse platform with the id syn2623706.
- Raw and processed scRNA-seq data of pan-cancer cell lines are available at the Gene Expression Omnibus (GEO) with the accession number GSE157220.

Funding

This work was supported by IMMUcan. The IMMUcan project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 821558. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. IMI.europa.eu

Conflict of Interest Disclosure

The authors declare no competing interests.

References

1. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, Xu Y, Fertig EJ. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. 2018 Oct;34(10):790-805.

2. Engreitz JM, Daigle BJ Jr, Marshall JJ, Altman RB. Independent component analysis: mining microarray data for fundamental human gene expression modules. J Biomed Inform. 2010 Dec;43(6):932-44

3. Zhou W, Altman RB. Data-driven human transcriptomic modules determined by independent component analysis. BMC Bioinformatics. 2018 Sep 17;19(1):327

4. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol. 2007 Aug;3(8):e161.

5. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, López-Bigas N, Kamoun A, Neuzillet Y, Gestraud P, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 2014 Nov 20;9(4):1235-45.

6. Cantini L, Kairov U, de Reyniès A, Barillot E, Radvanyi F, Zinovyev A. Assessing reproducibility of matrix factorization methods in independent transcriptomes. Bioinformatics. 2019 Nov 1;35(21):4307-4313.

7. Meunier L, Hirsch TZ, Caruso S, Imbeaud S, Bayard Q, Roehrig A, Couchy G, Nault JC, Llovet JM, Blanc JF, et al. DNA Methylation Signatures Reveal the Diversity of Processes Remodeling Hepatocellular Carcinoma Methylomes. Hepatology. 2021 Aug;74(2):816-834

8. PMID:34376232 Davis-Marcisak EF, Fitzgerald AA, Kessler MD, et al. Transfer learning between preclinical models and human tumors identifies a conserved NK cell activation signature in anti-CTLA-4 responsive tumors. Genome Med. 2021;13(1):129. Published 2021 Aug 11. doi:10.1186/s13073-021-00944-5

9. Stein-O'Brien GL, Clark BS, Sherman T, et al. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. Cell Syst. 2021;12(2):203. doi:10.1016/j.cels.2021.01.005

10. Aynaud MM, Mirabeau O, Gruel N, Grossetête S, Boeva V, Durand S, Surdez D, Saulnier O, Zaïdi S, Gribkova S, Fouché A, et al. Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. Cell Rep. 2020 Feb 11;30(6):1767-1779.e6.

11. Greco A, Sanchez Valle J, Pancaldi V, Baudot A, Barillot E, Caselle M, Valencia A, Zinovyev A, Cantini L. Molecular Inverse Comorbidity between Alzheimer's Disease and Lung Cancer: New Insights from Matrix Factorization. Int J Mol Sci. 2019 Jun 26;20(13):3114.

12. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun 20;14(6):e8124.

13. Teschendorff AE, Jing H, Paul DS, Virta J, Nordhausen K. Tensorial blind source separation for improved analysis of multi-omic data. Genome Biol. 2018 Jun 8;19(1):76.

14. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021 Jan 5;12(1):124.

15. Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, Warren A, Rodman C, Roth JA, Bender SA, et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. Nat Genet. 2020 Nov;52(11):1208-1218.

16. Sompairac N, Nazarov PV, Czerwinska U, Cantini L, Biton A, Molkenov A, Zhumadilov Z, Barillot E, Radvanyi F, Gorban A, Kairov U, Zinovyev A. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int J Mol Sci. 2019 Sep 7;20(18):4414.

17. Bac J, Mirkes EM, Gorban AN, Tyukin I, Zinovyev A. Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation. Entropy. 2021; 23(10):1368. https://doi.org/10.3390/e23101368

18. Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, Zinovyev A. Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics. 2017 Sep 11;18(1):712.

19. 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 Aug 3;11(8):4120-31.

20. Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol. 2020 May 11;21(1):109.

21. Himberg, J., Hyvarinen, A.. ICASSO: Software for investigating the reliability of ICA estimates by clustering and visualization. Proceedings of the 13th Workshop on Neural Networks for Signal Processing, 2003, NNSP'03. 259 - 268.

22. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015 Nov;21(11):1350-6.

23. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M, Alizadeh AA. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019 Jul;37(7):773-782.

24. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res. 2017 Nov 1;77(21):e108-e110.

25. Asadzadeh Z, Mohammadi H, Safarzadeh E, Hemmatzadeh M, Mahdian-Shakib A, Jadidi-Niaragh F, Azizi G, Baradaran B. The paradox of Th17 cell functions in tumor immunity. Cell Immunol. 2017 Dec;322:15-25.

26. Largeot A, Pagano G, Gonder S, Moussay E, Paggetti J. The B-side of Cancer Immunity: The Underrated Tune. Cells. 2019 May 13;8(5):449.

27. Zhang L, Li Z, Skrzypczynska KM, Fang Q, Zhang W, O'Brien SA, He Y, Wang L, Zhang Q, Kim A, et al. Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell. 2020 Apr 16;181(2):442-459.e29.

28. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics. 2011 Nov 9;12:436. doi: 10.1186/1471-2105-12-436.

29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498-504.

Figures

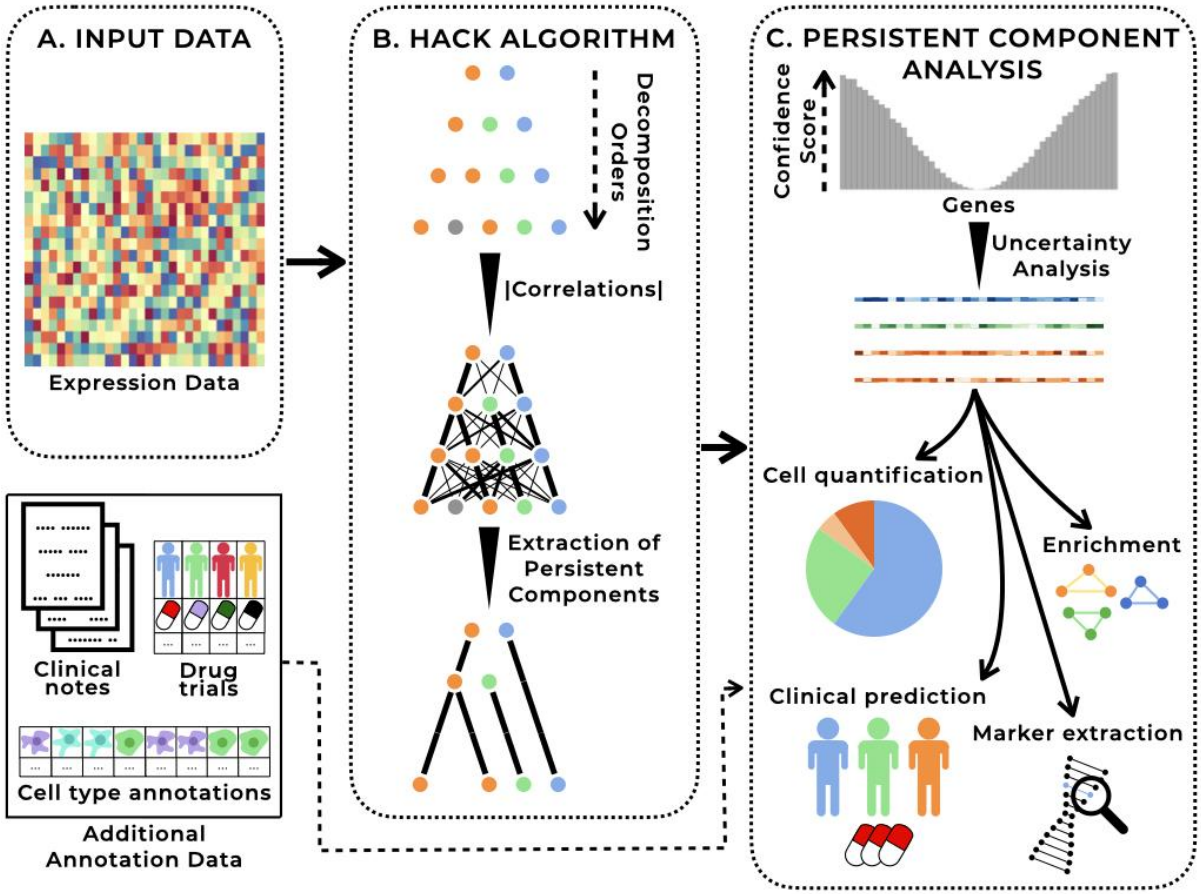


Figure 1. Simplified overview of the applications of the Hierarchical Analysis of Component linKs (HACK). **A.** Input data in the form of a preprocessed expression matrix obtained from bulk or single-cell RNAseq. Additional annotations can be used for final analysis steps. **B.** Main steps of the HACK algorithm, consisting in over-decomposing the input data, constructing the correlation graph between components and selecting the most stable and reproducible components, called persistent components. **C.** Once the tree is obtained, we can extract the signals as weighted vectors. Since a signal in the tree consists of multiple components, it becomes possible to add an uncertainty scoring to each feature to extract only the most significant ones. This cleaned component can later be used for various applications using additional annotation data.

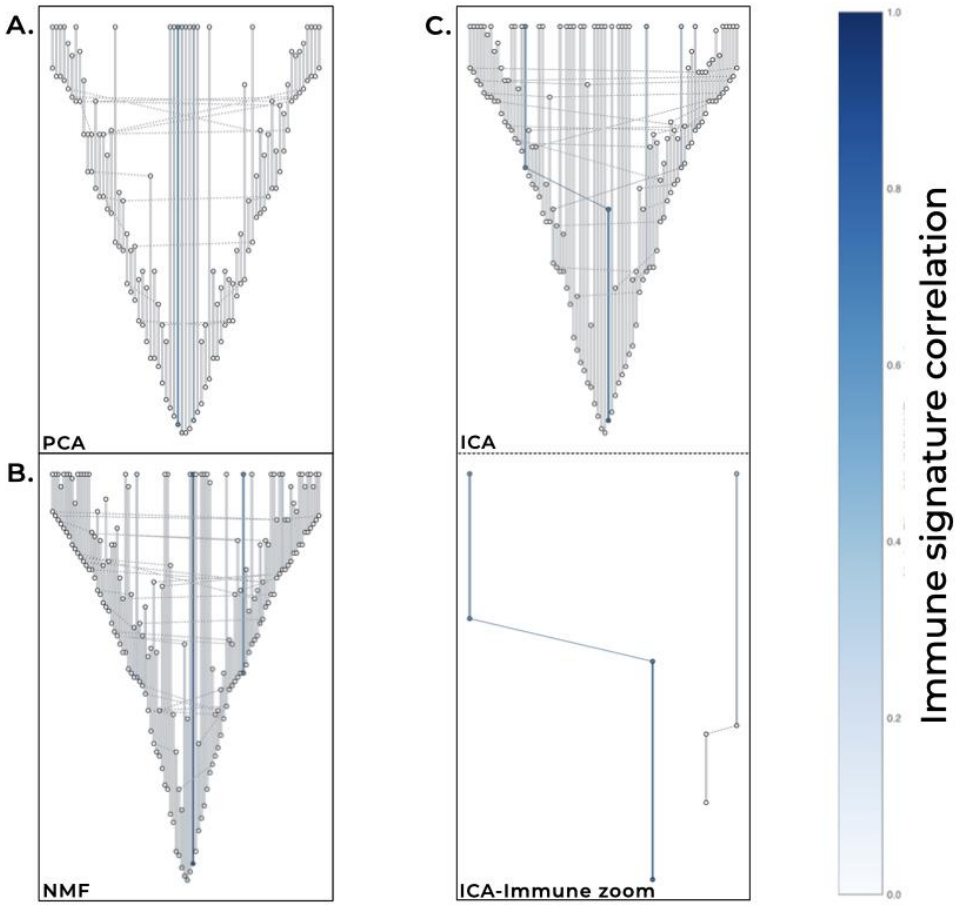


Figure 2. Hierarchical graphs obtained using three different types of Matrix Factorisation methods (PCA, ICA and NMF) on TCGA BRCA RNAseq data. Persistent components are highlighted if correlated with Biton's Immune Infiltration signature. **A.** Persistent component graph obtained using PCA with default parameters and 5% deleted random samples for each decomposition order. **B.** Persistent component graph obtained using NMF with modified default parameter of K=1 during the MNN split filtering in Step 3 of the HACK algorithm. **C.** Persistent component graph obtained using Stabilised ICA with default parameters. Top: full hierarchical graph. Bottom: Zoom on only persistent components having a minimal Pearson correlation of 0.3 with Biton's Immune Infiltration metagene.

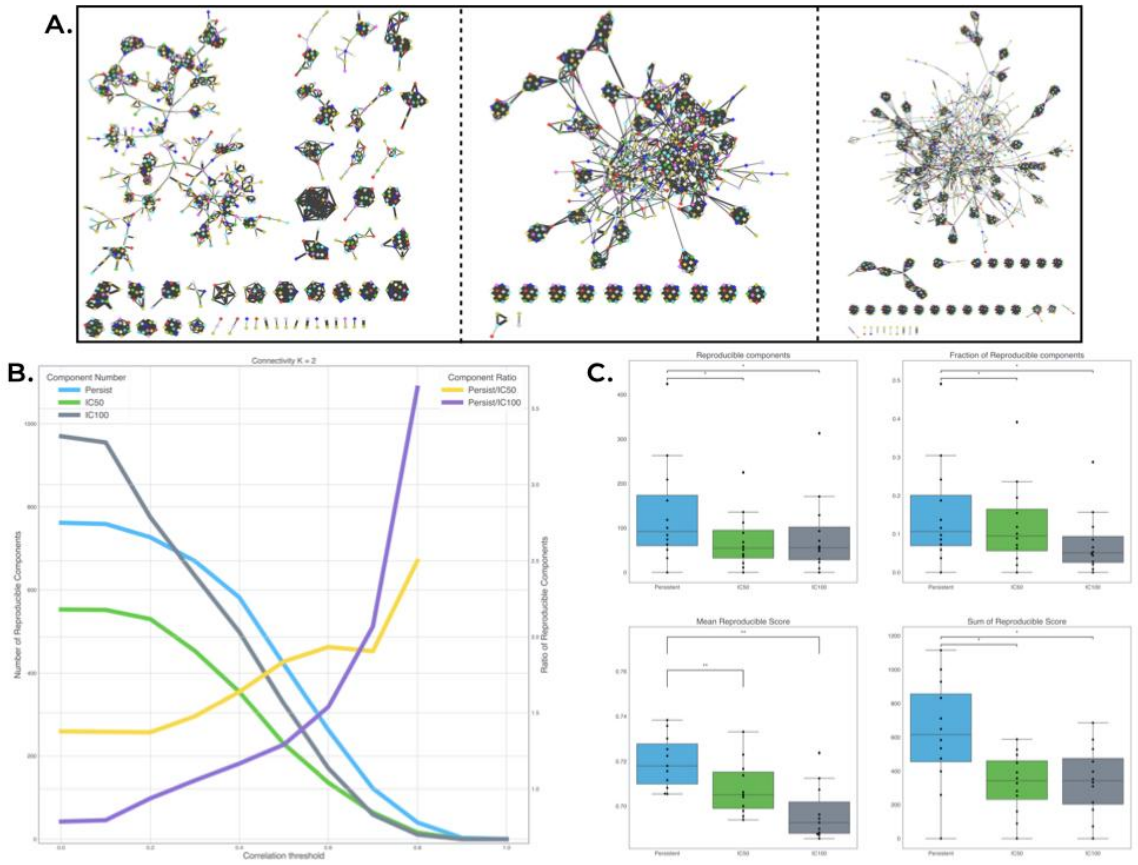


Figure 3. Reproducibility analysis of persistent component and simple stabilised ICA components of order 50 and 100 on 12 CRC RNA-seq datasets. A. View of a MNN correlation graph for K=1. Each node corresponds to a single component with the width of outgoing edges is proportional to the correlation between components. Each colour represents a dataset from which the component was extracted. Left: MNN graph of persistent components. Middle: MNN graph of single ICA components for a decomposition order of 50. Right: MNN graph of single ICA components for a decomposition order 100. **B.** Dependency graph between the correlation threshold S and the number of reproducible components for a connectivity value K=2. Left Y-axis: Green, Gray and Blue curves correspond to the total number of connected components. Right Y-axis: Yellow and Purple curves correspond to the ratio of the total number of connected components between Persistent and single ICA decomposition of order 50 or 100. **C.** Comparison of reproducible components for a range of connectivity from 1 to 12 with a threshold of minimum correlation of 0.6. Significance level of difference is * is P-value < 0.005 and ** if P-value < 0.001 of a Wilcoxon test.

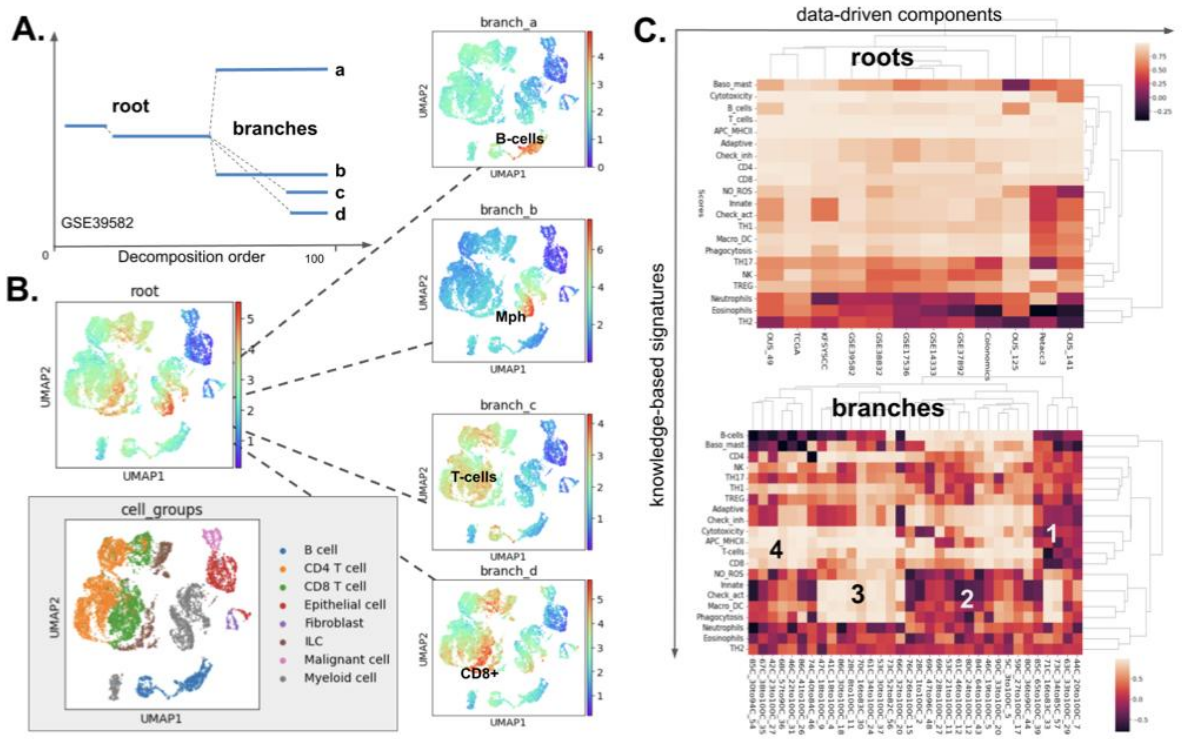
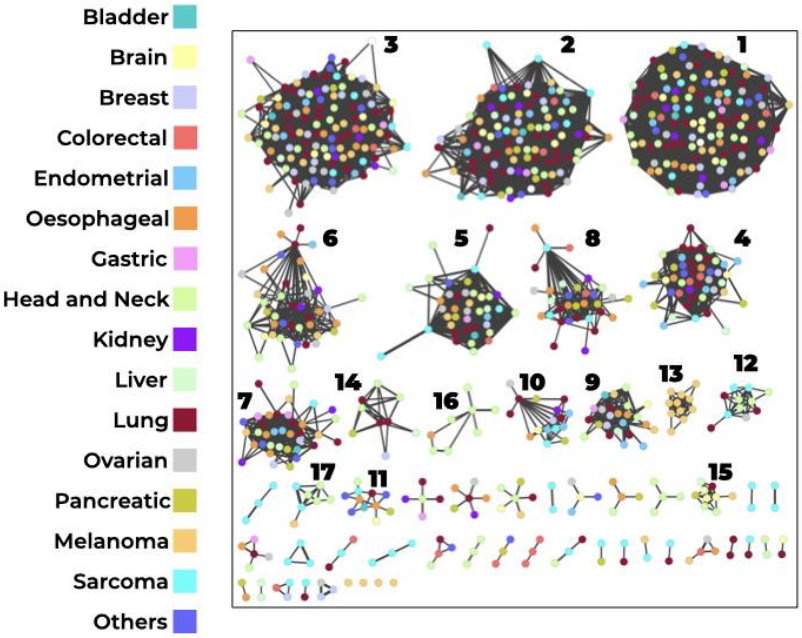


Figure 4. Deconvolution of immune signal in colon-cancer data. A. Typical deconvolution of immune signal in colon-cancer data (GSE39582) **B.** Activity of "root" and "branches" metagenes projected onto immune single-cell data. **C.** Hierarchical clustering and functional interpretations of immune metagenes from 12 colon-cancer datasets

A.



B.

Clusters	Top 5 in common genes
1	AURKA, TOP2A, CENPF, ASPM, PLK1
2	CLSPN, HIST2H2AC, HIST1H4C, HIST1H1E, E2F1
3	PSMC4, PSMB3, MRPL13, MCTS1, EIF4A3
4	C1orf56, CTNNB1, CTC-425F1.4, RP11-538P18.2, HNRNPH1
5	GOLGB1, HIST1H1E, DST, PCM1, SMARCA5
6	CTNNB1, HNRNPH1, C1orf56, B4GALT1, PPP1CB
7	PRDX4, MALSU1, VIMP, ITM2B, PSMG1
8	HIST1H1E, HIST2H2AC, HIST1H1D, HIST1H1C, HIST1H1B
9	MAGED2, EIF4A3, PSMC2, LGALS3BP, HSPA5
10	DDIT3, ATF3, SNHG12, CDKN3, JUN
11	CTNNB1, C1orf56, B4GALT1, HNRNPH1, PTPRS
12	NOP16, GPATCH4, LYAR, DDX21, TUBA1A
13	MLANA, CAPN3, PHACTR1, DCT, PMEL
14	CD81, TIMP2, EIF3B, PDXK, FASN
15	IFIT3, OASL, IFIT2, ISG20, RARRES3
16	S100A9, SPRR1B, SAA2, C15orf48, KRT16
17	BIRC5, CDKN3, CENPW, PTTG1, CDC20

Figure 5. Meta-analysis of scRNA-Seq profiles from Cancer Cell Lines Encyclopedia showing the Recurrent Heterogeneous Programs (RHP) obtained with the HACK method using stabilised ICA. A. Clustered MNN graph of persistent components obtained from 125 pan-cancer cell lines with each cluster corresponding to a specific biological program. Each node corresponds to a persistent ICA component connected via and edges with a width proportional to the Pearson correlation value between both components. **B.** List of the 5 most significant genes extracted from the 17 biggest clusters.

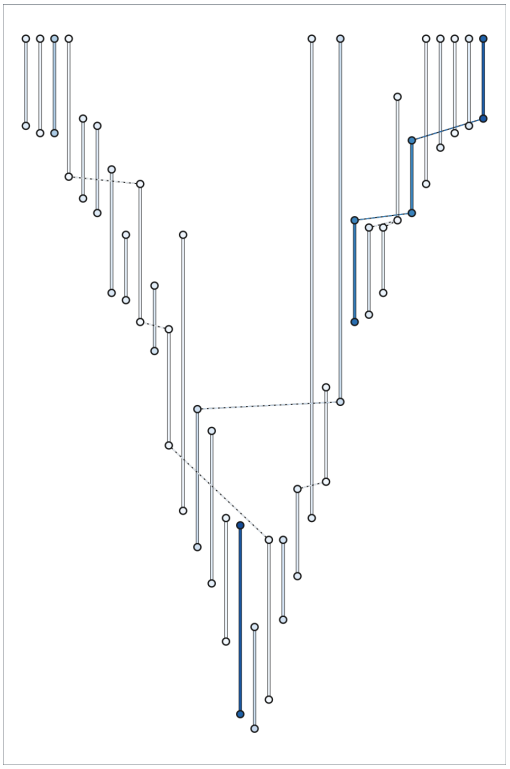


Figure S1. Hierarchical graph obtained using NMF with the default parameters of HACK.

4.4 Additional results and observations

4.4.1 Over-decomposition: To boldly go where no one has gone before!

As stated in the paper of (Kairov et al., 2017), although in general the expected number of signals in a bulk RNA-seq dataset is around 30-40, it is a common procedure to take a higher number of decompositions to still keep the opportunity of extracting additional signals. In the worse case, the additional signals will not be stable and will be discarded from the analysis. It is for that reason that in this paper, datasets were decomposed up to 100 components. By doing so and looking at the resulting hierarchical graph, we are able to follow the number of persistent components along the range of decomposition orders.

As seen in the Figure 4.4, for the 12 CRC datasets, there is a fluctuating but constant increase of the total number of persistent components followed by a drop towards the end. Peaks for the number of persistent components differ between datasets and are also tied to the length of persistent components. But even if we were to increase the length of persistence, the trend stays the same and only the total number diminishes. This behaviour is expected from ICA since the amount of signals in the data is finite and by increasing the number of components above this amount, we are starting to force the appearance of artifactual signals that don't have any biological relevance attached to them. However, when going for high decomposition orders, we begin to see the appearance of components that could be called "gene shaving components". Said components often contain a very small number of significant genes and these genes are often "shaved" from previously found stable components.
This is usually why we stop decompositions around the order given by the MSTD measure since any new signals generated after this order is expected to be either not stable or relevant, or it could just contain a small amount of already identified genes in existing stable components.

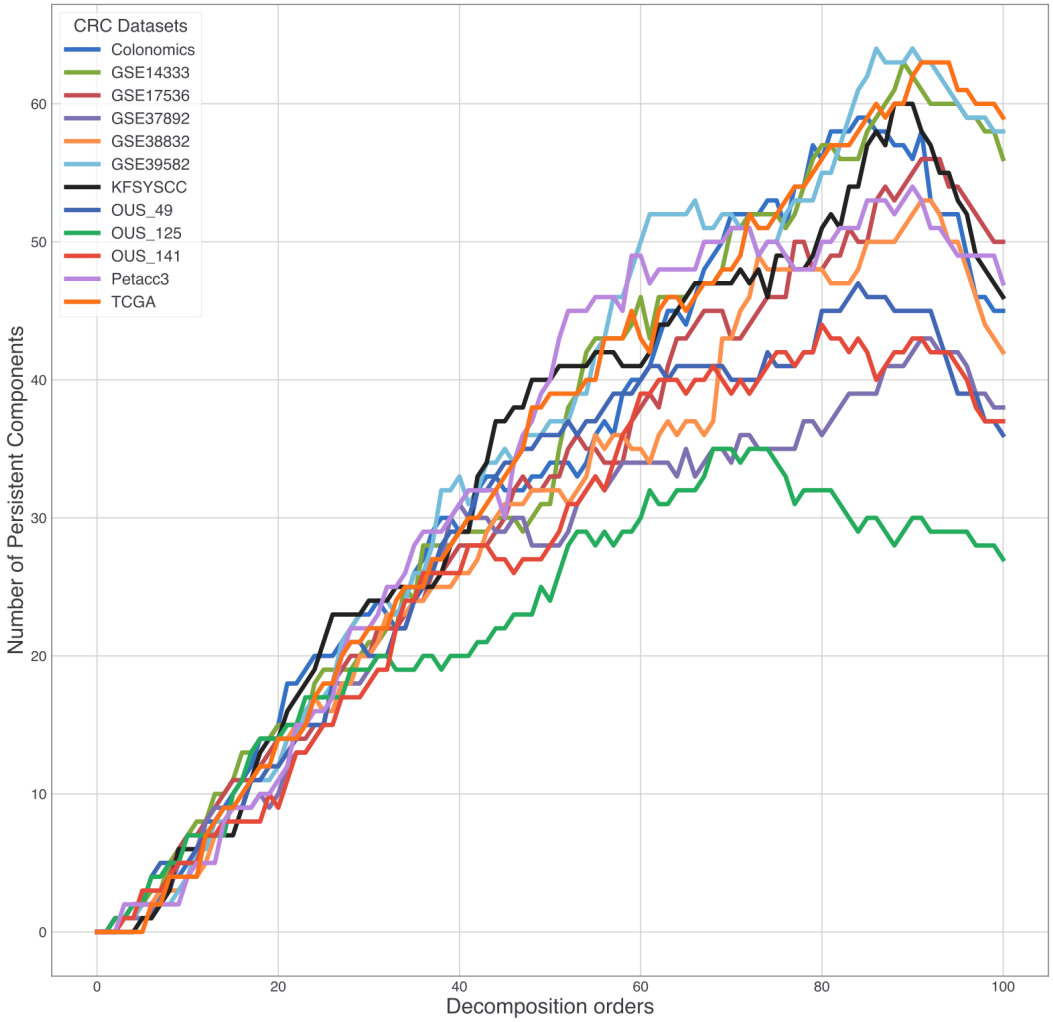


Figure 4.4. Evolution of the number of persistent components across decomposition orders for 12 CRC RNA-seq datasets.
Each persistent component was obtained using a stabilised ICA decomposition.

However, since it has become possible to follow the fate of each component individually with the HACK method, I could test this by going above the usual 100 components. It is important to note that going above the MSTD level, I began encountering convergence problems of ICA algorithms. It is for this reason that the analysis was executed on only one dataset of TCGA CRC.

However, instead of witnessing a stop in the growth of the number of persistent components or even a decline, the number continued to increase instead (Figure 4.5). While at first by setting a limit of 100 components, we could observe a peak of approximately 60 persistent components, by going beyond that, we manage to reach a peak around 90 persistent components. Even more surprisingly, most of these newly found components weren't coming from previously existing ones and contained new specific groups of genes. The majority of previously found persistent components stayed persistent to the end and didn't show any splitting events.

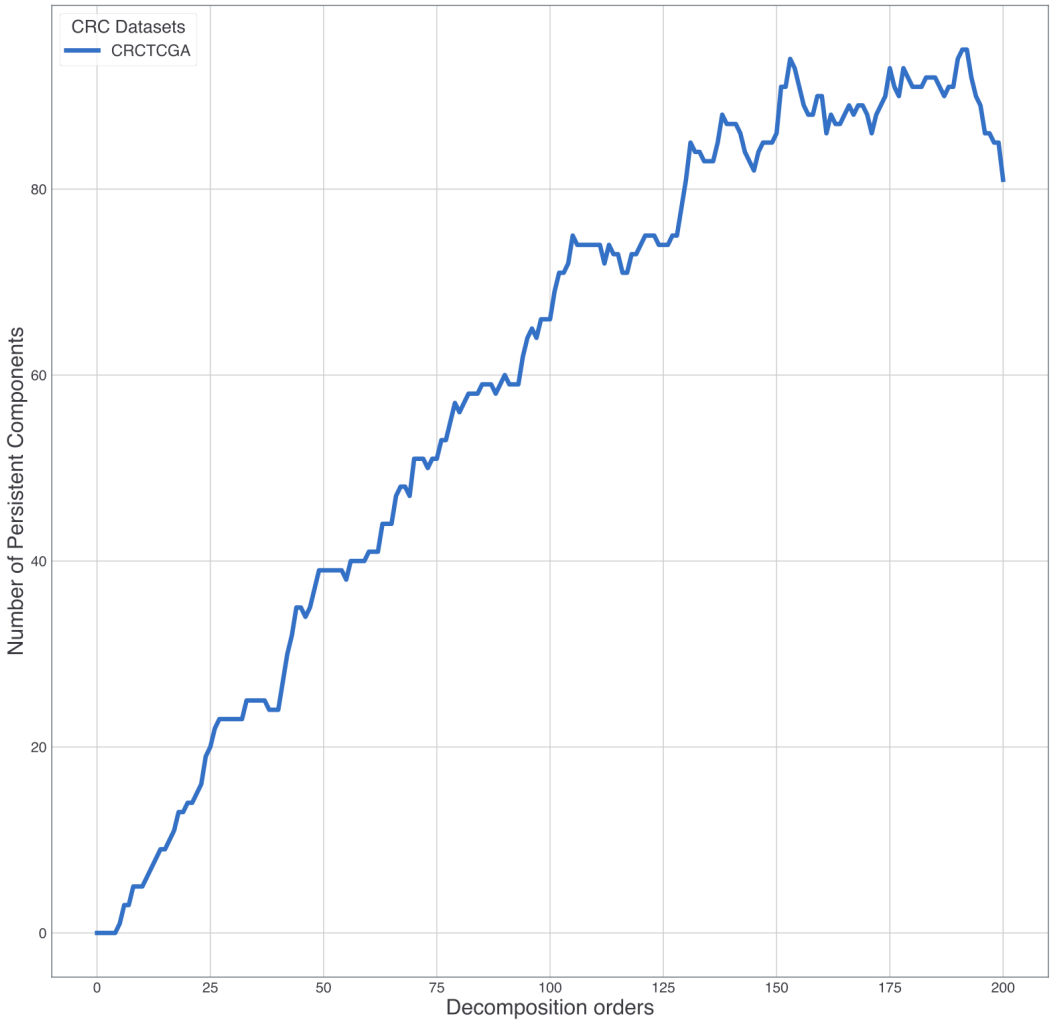


Figure 4.5. Evolution of the number of persistent components across decomposition orders for the CRC TCGA RNA-seq dataset.
Each persistent components were obtained using a stabilised ICA decomposition.

To answer the question if these new components were indeed resulting from gene shavings or not, I tried to look at their number of very stable and significant genes, i.e. with a Z-score above 5 standard deviation (5SD). While some persistent components were indeed containing a very small number of top-contributing genes, there were still a decent number of entirely new components. This shows that ICA is still able to find specific components with specific genes that have good biological associations at high orders of decomposition.
To interpret those high order persistent components, I have extracted the top 50 genes with a good stability value and a Z-score above 5SD and used them for an enrichment analysis with the TOPPGENE tool (Chen et al., 2009). Many of these enrichment returned functions that were not discovered previously in lower dimensions. To cite only the most significant result, a persistent component appearing at the order 136 and staying persistent all the way to 200, contained genes responsible for “Nucleosome assembly and organisation” with an enrichment P-value of 10*-77. These genes were almost all histones, had functions related to chromatin organisation or were implicated in CRC and no other component contained them before. Meaning that the only way of extracting this signal is to decompose the data into at least 136 components using ICA.

This observation shows us that although the average number of expected signals via methods such as the MSTD score makes a good reference for the minimal number of components, it makes total sense to go above and beyond the standard number of decomposition levels, especially if the analysis is focused on specific signature discovery.

It is with the help of such analyses that we can hope to answer the problems stated in Sections 3.2 and 3.3.

4.4.2 Variance of component weights

With the existence of persistent components as a set of related individual components, it becomes now possible to compare all the components of the same set with a focus on the behaviour of metagene scores. By analysing the distribution of the genes' Z-scores or their rank in the metagene, we can see an interesting information that wasn't available before. In Figure 4.6, even if we take a really stable and persistent component present across 92 decomposition levels and containing ~20.000 genes, we can still observe a heavy fluctuation of those scores between each component. This distribution is mainly attributed to the stochastic behaviour of ICA. In fact, only a small number of genes are consistently found in the top and bottom parts of a metagene while the rest have a large standard deviation value, making them unfit to be considered as significant genes. This effect is observed on short persistent components as well and seems to reflect the general score distribution assigned by ICA.

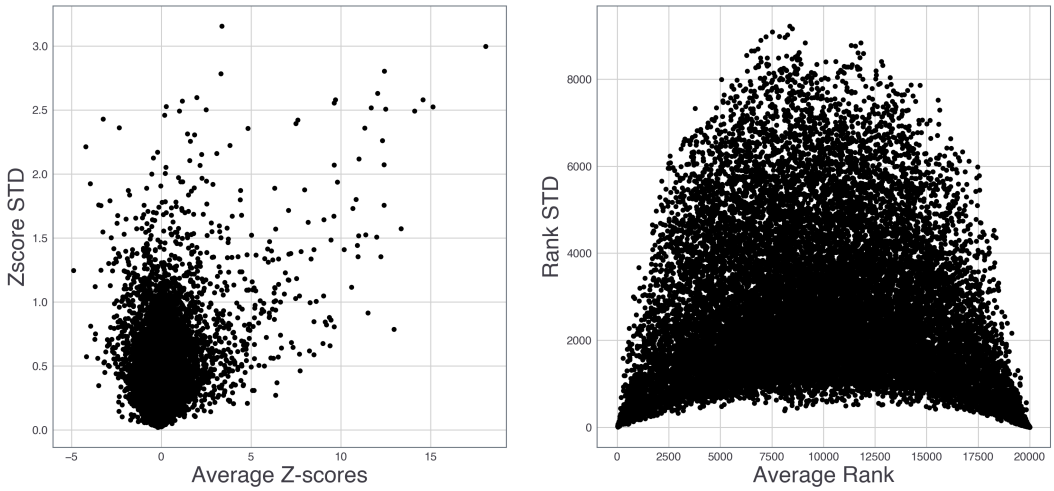


Figure 4.6. Standard deviation analysis of metagene weights distribution from a persistent component of length 92.
Persistent components were obtained using the BRCA TCGA RNA-seq dataset. Each point corresponds to a score assigned to a gene.

In fact, after a filtering of highly variable genes, we can on average safely consider only the top 50 genes with certitude as stable significant genes of a given component. This is important especially in cases if one wants to apply only a single shot ICA decomposition to extract genes from a component.

Thanks to this observation, it gives us an additional insight on the problems of extracting markers from a component using an arbitrary threshold. By doing that, we might extract unrelated genes that were present in the top just by random chance, which might give wrong enrichment results afterwards.

4.4.3 Difficulty of quantifying the biological relevance of unsupervised components

One of the most difficult tasks, that remains unsolved to this day, is the justification that a certain component is biologically significant. When trying to relate the persistence of a certain component with its biological meaning or a certain confidence score, I was struggling to find a correct way of achieving this. In general, what is at our disposal are biological annotation databases that can be used to perform statistical tests such as enrichments to get an idea of how specific a certain component's genes are to an annotated set of genes. However, such databases can present either huge biases or can be entirely unsuitable for certain analyses.

Let's set aside the possibility that certain databases are biased towards certain diseases, pathways or biological functions and consider that we have at our disposal a good reference database with fully annotated genes.

Right from the beginning I was faced with statistical limitations such as the number of genes used as an input for the analysis. As mentioned in the previous part, only a few genes in an independent component can be considered as robust and specific, leaving us with a poor statistical power if I were to take only those. However, if I increase the number of genes as inputs, there is also the problem of the size of the gene set of a certain biological function or pathway. By taking a large number of genes, we increase the chances of hitting low level biological signals. This creates a problem of trust in results as well as adding an additional requirement to set thresholds to the size of reference pathways for example.

However, if we do obtain strong associations between our input genes and a certain biological function, I now have to select a score to reflect this. Usually, the P-value of an appropriate test is the first choice that comes to mind. However, what happens when a highly persistent component doesn't have a good P-value? We would be confronted with the problem of reconciling a good statistical strength from a mathematical model with a poor biological enrichment. In other words, how can a robust and reproducible component not have a good biological significance as well?

It is also important to note the fact that methods such as ICA help extracting and grouping genes that have a similar expression behaviour but often don't participate in a single biological mechanism. Therefore, when using such groups of genes in an enrichment analysis, we won't be able to clearly see a single enriched pathway. Instead we will often observe various pathways poorly enriched due to the low number of genes matching with these specific pathways. As a simple example, it happened for one CRC dataset that a persistent component didn't give any enrichment results at all, despite being the most persistent one. It was only by looking at the individual top genes that it was possible to discern that they were related to male specific mechanisms. Some genes were located on the Y chromosome but not enough to show a chromatically location enrichment. Most of them participated in such different pathways and functions that it wasn't possible to hit a global meaning such as "Male specific function genes".

It was when confronted with these exact problems that I decided to prioritise statistical characteristics such as robustness, reproducibility and clustering capacity. Using enrichment analyses to assign a biological score proved impossible to do on a level that could be compared equally between any given set of components.

5. A multiscale signalling network map of innate immune response in cancer reveals signatures of cell heterogeneity and functional polarization

Maria Kondratova, Urszula Czerwinska, **Nicolas Sompairac**, Sebastian D Amigorena, Vassili Soumelis, Emmanuel Barillot, Andrei Zinovyev, Inna Kuperstein.

Published in Nature communications, 22nd of October 2019.

5.1 Description

Given the importance of TME and more particularly its immune constituents, to follow tumor development and its response to various therapies, the creation of TME related knowledge maps is essential to shed more light in this field.

In this publication, which I am a co-author of, we propose an integrated resource of innate immune response related maps containing molecular mechanisms in the form of signalling networks. This resource is formed of multiple cell-type specific maps of macrophages and myeloid derived suppressor cells, DC and NK. In the first part, the structure and content of this new resource will be described and then a demonstration of data visualisation and analysis possibilities will be shown. My role in this article was mainly to help structure and integrate the data in NaviCell, a web-based platform developed in my team. It is the integration of this resource in NaviCell which allows it to be queryable for data analysis methods.

In this work, ICA was used to extract latent variables from single-cell RNA-seq data of macrophages and NK cells obtained from metastatic melanoma samples. Components that represented a diversity among these cells were selected and an activity score was computed. These activity scores were then projected on the innate immune map to visualise the functional phenotypes of these 2 groups of immune cells in Figure 5 of (Kondratova et al., 2019).

Two groups of biological functions related to NK cells activity were identified this way with pathways from the first group showing increased functions of NK recruitment and activation, coinciding with tumor killing functions, while the second group represented resting or suppressed NK cells.
When applied to macrophages, it was possible to identify the pro and anti-tumor properties of these cells. Pathways of antigen presentation, immuno-suppressive checkpoints and immuno-stimulatory miRNA and TF were shown to be up-regulated in the first group of macrophage expressing anti-tumor phenotypes. The second group related to pro-tumor activity could be explained by an up-regulation of immuno-suppressive cytokines expression.

These two results demonstrate that ICA applied to single cell data is capable of going deeper than with bulk data by differentiating the functional states of immune cell types and that the use of knowledge maps can greatly enhance our interpretation of these functional phenotypes.

An additional study to see if this resource could be used patient survival prediction was performed. Several modules allowing correlating patient survival positively and negatively were found with a strong predominance of positively correlated genes.

5.2 Article

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12270-x OPEN

A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures

Maria Kondratova[1,4], Urszula Czerwinska[1,2,4], Nicolas Sompairac[1,2], Sebastian D. Amigorena[3], Vassili Soumelis[3], Emmanuel Barillot[1], Andrei Zinovyev[1] & Inna Kuperstein[1*]

The lack of integrated resources depicting the complexity of the innate immune response in cancer represents a bottleneck for high-throughput data interpretation. To address this challenge, we perform a systematic manual literature mining of molecular mechanisms governing the innate immune response in cancer and represent it as a signalling network map. The cell-type specific signalling maps of macrophages, dendritic cells, myeloid-derived suppressor cells and natural killers are constructed and integrated into a comprehensive meta map of the innate immune response in cancer. The meta-map contains 1466 chemical species as nodes connected by 1084 biochemical reactions, and it is supported by information from 820 articles. The resource helps to interpret single cell RNA-Seq data from macrophages and natural killer cells in metastatic melanoma that reveal different anti- or pro-tumor sub-populations within each cell type. Here, we report a new open source analytic platform that supports data visualisation and interpretation of tumour microenvironment activity in cancer.

[1] Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, 75005 Paris, France. [2] Université Paris Descartes, Centre de Recherches Interdisciplinaires, Paris, France. [3] Institut Curie, PSL Research University, Inserm, U932, 75005 Paris, France. [4] These authors contributed equally: Maria Kondratova, Urszula Czerwinska. *email: inna.kuperstein@curie.fr

Tumors are engulfed in a complex microenvironment (TME) that critically impacts disease progression and response to therapy. TME includes immune and non-immune interconnected components that exchange multiple signals and are influenced by molecules secreted by cancer cells. The behavior of the tumor and its TME as a whole critically depends on the organization of these different players and their ability to regulate each other in a dynamic manner[1]. The innate immune part of the TME plays important, but sometimes opposite roles in tumor evolution. Innate immune cells can contribute to eliminate the tumor, e.g. through phagocytosis and T cell priming and by induction of adaptive immune response. However, they can also favor tumor escape from immunological control, by a production of immunosuppressive molecules such as transforming growth factor beta (TGFB) or interleukin 10 (IL10)[2]. An additional level of complexity in the TME is that various stimuli can lead to a range of innate immune cells' phenotypes. This results in very heterogeneous subpopulations within each innate immune cell type coexisting in TME[3,4].

Depending on the set of stimuli from TME and tumor, immune cells are able to change their phenotype or polarization status from anti-tumor to pro-tumor[5,6]. Such functional dichotomy was first evidenced for one of the components of innate immunity in TME, the tumor-associated macrophages (TAM) and led to a description of M1 and M2 polarized TAM classes[7]. The same tendency was later documented for other components of innate immunity as neutrophils[8], dendritic cells[9] and natural killers[10]. Therefore, the term "polarization" can be applied for the innate immunity system in TME in general[11] that represents the major focus of current works. The balance between anti-tumor and pro-tumor activity of innate immune cells has an impact on tumor growth, patient response to therapy, and survival[12].

The correct evaluation of the polarization status within the subtle innate immune cell subpopulations in TME is essential for immunotherapy improvement. Nevertheless, the primary activation of adaptive immune response requires innate immune players, the antigen-presenting cells (APC) such as dendritic cells[13] or macrophages[14,15] Therefore, an efficient immune checkpoint therapy depends directly on the proper innate immune activation[16]. In addition, there are studies showing that innate immunity can restrict tumor growth even when the adaptive immune system is inactivated[17]. This indicates that detailed study of potential innate immune-related targets should be performed to identify new types of immunotherapy[18] that could function in synergy with the current T cell-targeted therapies or act independently[19,20].

There is a massive amount of information in the literature about molecular mechanisms implicated in innate immune cells polarization in TME. However, most of the studies are focused on individual molecular components and pathways. They do not integrate the complexity of multiple crosstalks between innate immune cells and tumor cells. To create a holistic picture of the diversity and integrity of innate immune system in TME, the knowledge about molecular circuits should be gathered together and systematically represented[21].

To address these challenges, a systems biology approach is needed[22]. Formalization of biological knowledge in a form of comprehensive signaling maps, both at the intra- and intercellular levels, helps to integrate information from multiple research papers[23]. There are numerous public databases containing signaling pathways related to innate-immune response such as KEGG[24] and REACTOME[25], which are quiet comprehensive, but contain mostly generic mechanisms. Furthermore, there are resources dedicated to different types of innate immune cells such as macrophages[26] or dendritic cells[27]. Finally, there are resources depicting the innate immune system in general as InnateDB[28] and ImmuNet[29], Virtually Immune[30]. However, these repositories are rather pathogen response-oriented than cancer-specific and often represent a catalog of disconnected pathways. Thus, there is a need to create an integrated resource on molecular mechanisms of innate immune response in cancer.

To fill the gap, we construct and present here a system of cell-type-specific maps and an integrated meta-map of innate immune signaling in cancer based on the information retrieved from the literature (Fig. 1). These maps together represent an open source analytic platform for data visualization and interpretation of TME activity in cancer.

Results

Principles of innate immune response in cancer. The molecular mechanisms regulating six major innate immune cell types found in the TME were gathered and depicted in the form of network maps. To cope with a massive body of literature on innate immune response in cancer we followed a systematic procedure of literature selection, knowledge organization, and integration of information in a visual and understandable manner (Fig. 1). The network maps were constructed as two-dimensional maps to facilitate the graphical representation of molecular mechanisms that drive biological processes. The maps possess a particular layout that reflects the accepted vision of spatial organization and propagation of biological processes. The information about molecular mechanisms was manually retrieved by the researchers from the scientific literature along with the information presented in general pathway databases or in the immune system-specialized resources. The information was classified by specificity to the cell types in cancer and organized into three cell-type-specific signaling network maps, namely map of macrophages and myeloid-derived suppressor cells (MDSC), dendritic cells and natural killer (NK) cells (Fig. 2). These maps, enriched by the information on additional cell types as neutrophils and mast cells, were integrated into the meta-map of innate immune response in cancer (Fig. 3).

The molecular mechanisms were depicted in the maps in the form of biochemical reaction network using a well-established methodology[31,32]. The maps were described using Systems Biology Graphical Notation language (SBGN)[33] and drawn using the CellDesigner tool[34] that ensures compatibility of the maps with various tools for network analysis, data integration, and network modeling (Fig. 3b). Each molecular player and reaction in the maps was annotated in the NaviCell format. The NaviCell annotations include PubMed references, cross-references with other databases, and notes of the map manager. In addition, molecular complexes and reactions were assigned with confidence scores and tags indicating their involvement in different biological processes on the maps. Finally, the correspondence of each molecular player on the map to different cell types is also indicated, indicated by cell-type-specific tags (Supplementary Fig. 1)[35]. The principles and procedure for map construction are provided in the Methods.

Content and structure of the innate immune maps. Macrophages are the major immune component of leukocyte infiltration in the tumor. The anti-tumor polarization of macrophages is related to their ability to recognize and to reject tumor cells by phagocytosis, represent tumor antigens on the cell surface and induce a T cell response and attract immune cells into the TME. However, TAMs can also act as pro-tumor agents, expressing tumor-stimulating growth factors, producing immunosuppressive molecules induce angiogenesis and matrix remodeling in TME and consequently facilitate metastatic process[36,37].

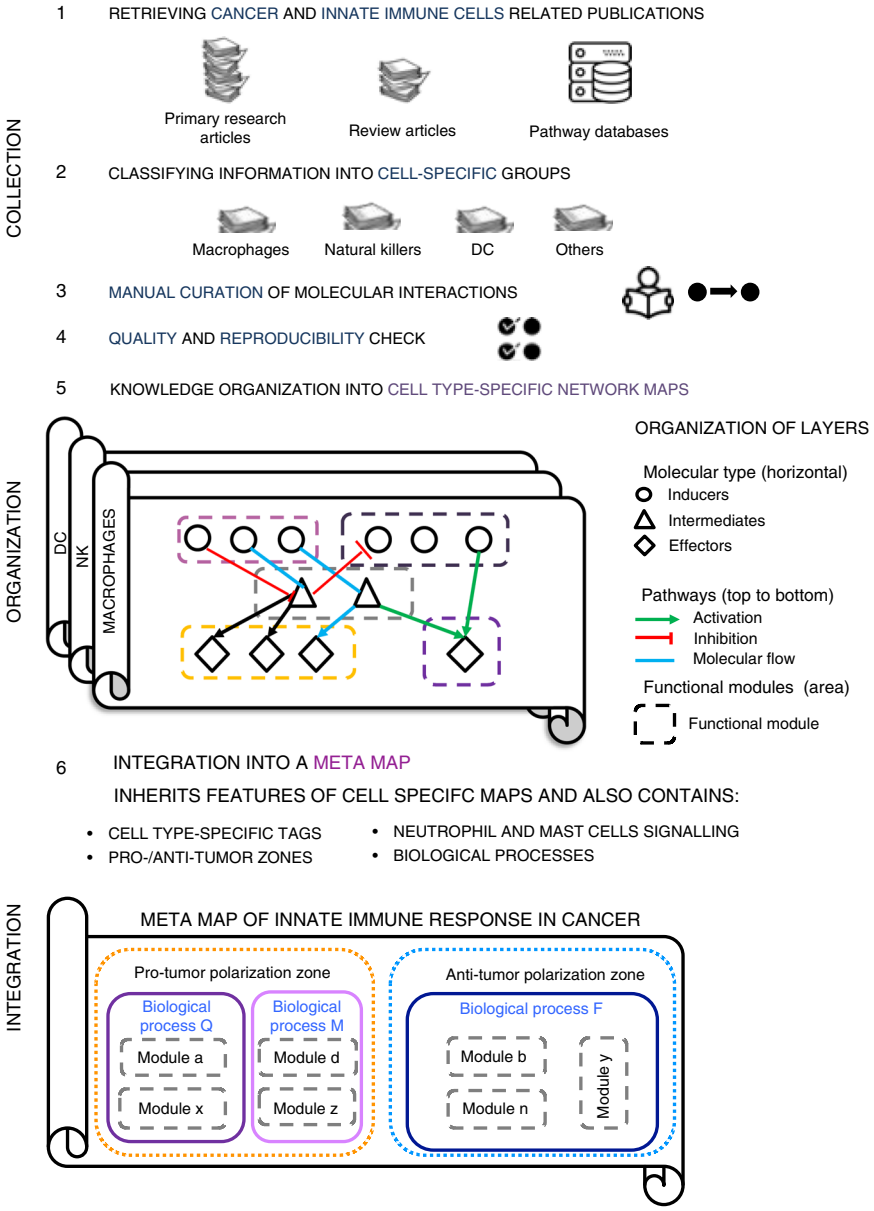


Fig. 1 Map construction workflow and map structure. The scheme demonstrates the steps of meta-map construction starting from collection of cancer-specific and innate-immune specific information about individual molecular interactions from scientific publications and databases, manual annotation and curation of this information (steps 1–4), then organization of this formalized knowledge in form of cell-type specific maps (step 5), and finally integration the cell-type specific networks in one global meta-map of innate immune response in cancer with areas corresponding to biological processes, modules, pro- and anti-tumor polarization (step 6)

MDSC represent a heterogeneous population of myeloid cells. In general, the role of MDSC in TME is similar to TAMs. MDSC suppress T cell response and NKs' activity in TME. In addition, MDSCs induce EMT and angiogenesis and participate in matrix remodeling. MDSC mostly show a pro-tumor activity; therefore, their presence in the tumor is correlated with a poor clinical prognosis[38,39]. The MDSC signaling is included into the macrophage cell-type-specific map.

The macrophage and MDSC cell-type-specific map contains 588 objects and 7 modules representing both pro-tumor and anti-tumor polarization of myeloid cells (Fig. 2a, Supplementary Table 1).

The map is available at https://navicell.curie.fr/navicell/newtest/maps/macrophages_mdsc_cells/master/index.html.

Dendritic cells are innate immune cells that can have both myeloid and lymphoid origin. As with macrophages, dendritic cells have phagocytic abilities and can produce inflammatory cytokines. But the major role of dendritic cells in anti-tumor response is antigen presentation and further T cell activation[40]. The dendritic cell map contains 491 objects and 8 modules (Fig. 2b, Supplementary Table 1).

The map is available at https://navicell.curie.fr/navicell/newtest/maps/dendritic_cell/master/index.html.

NKs are big granular lymphocytes that can be cytotoxic to tumor cells. The main role of NK cells in innate immunity is an elimination of cells lacking MHC1 molecules that therefore cannot be recognized by T cells. The activity of NK cells is stimulated by the target cells expressing NK receptors activating ligands and modulated by inflammatory cytokines, produced by macrophages and dendritic cells. NK cells secrete granules contains lytic enzymes and express the apoptosis inducers. Presence of active NK cells in cancer is correlated with good

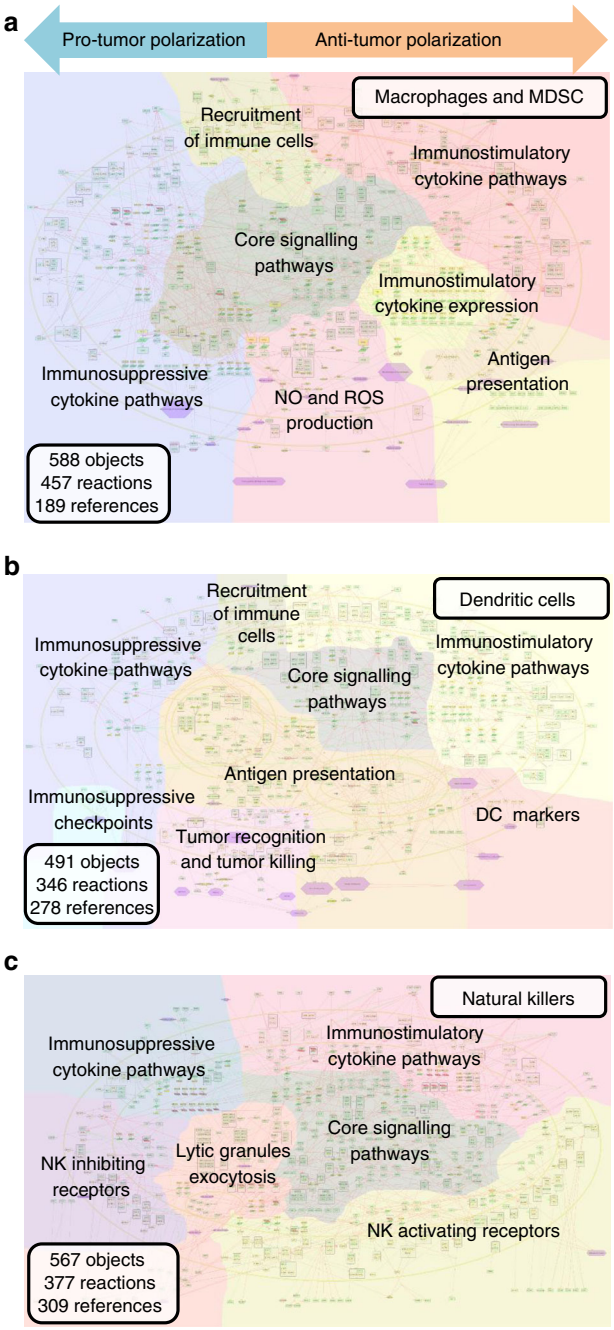


Fig. 2 Cell-type-specific maps. Cell-type-specific networks are visualized at the top-level view, the colorful background indicates boundaries of functional modules of the maps. **a** The maps of macrophages and MDSC. **b** The map of dendritic cells. **c** The map of natural killer cells

prognosis. To escape NK control, tumor cells express immunosuppressive cytokines and downregulate NK ligands expression that collectively inhibit cytotoxic activity of NK cells[41]. A pro-tumor polarization of NK cells is not described in the literature. However, suppressed NK cells are incapable to reject tumor cells and, therefore, indirectly promote cancer progression. The NK map contains 567 objects and 6 modules (Fig. 2c, Supplementary Table 1).

The map is available at https://navicell.curie.fr/navicell/newtest/maps/natural_killer_cell/master/index.html.

Neutrophils form a subtype of granulocytic leukocytes. The role of neutrophils in the tumor microenvironment is not well documented, but it is known that they can produce ROS, inflammatory cytokines and demonstrated tumoricidal activity. Although, in other conditions, neutrophils act as pro-tumor agents via stimulation of matrix remodeling, angiogenesis, and metastasis, therefore these cells have both pro- and anti-tumor polarization potential[8,42]. The signaling on neutrophils is included into the innate immune meta-map (Fig. 3, Table 1).

Mast cells resemble blood basophils and contain granules rich in histamine and heparin. The experimental data about the influence of mast cells on tumor microenvironment is contradictory. It is known that mast cells can produce inflammatory cytokines and secrete Chondroitin sulfate which acts as a decoy for tumor cells and blocks the metastatic process. However, mast cells also secrete molecules stimulating tumor growth, angiogenesis and local immunosuppression[43,44]. Probably the polarization of mast cells in TME is context-dependent. The signaling on mast cells is included into the innate immune meta-map (Fig. 3, Table 1).

The aforementioned cell-type-specific maps gathered together and enriched by additional information gave rise to the global, seamless meta-map of innate immunity in cancer. The meta-map contains 1466 chemical species as nodes connected by 1084 biochemical reactions, and it is supported by information from 820 cell-type specific and cancer-related articles (Table 1).

The layout design of the meta-map reflects the current understanding of signaling propagation in cells. To cope with the complexity of the signaling network and to make it understandable and navigable, the meta-map has a hierarchical structure (Figs. 1 and 3). The meta-map possesses two major structuring dimensions: the internal organization of the map (layers, zones, meta-module, modules, and pathways) and the external organization represented by zoom levels (see explanation below).

The internal organization of the meta-map is provided in a form of three layers entitles Inducers, Core Signaling, and Effectors (Fig. 3a, Table 1). The top part of the meta-map is the Inducers layer that depicts inducer molecules frequently present in TME. The inducers interact through specific receptors and adaptor proteins that propagate the signal via limited number of transmitters, also called hub molecules as NF-kB, PLCG, PI3K, etc. These molecules are located in the middle parts of the meta-map in the Core Signaling layer. The signaling is further propagated to the Effectors layer, located in the lower part of the meta-map, which actually executes the biological activity and therefore defines the outcome phenotype, namely, the positive or negative influence of the innate immunity system on the tumor growth and invasion (Fig. 3b, Table 1).

Further, the whole meta-map is divided into multiple signaling pathways, running through the aforementioned layers (Fig. 3b). A signaling pathway on the meta-map represents a sequence of molecular interaction which transforms extracellular signals into intracellular activity or into single or multiple cell phenotypes. For instance, the TGFB pathway in innate immune cell upregulates the expression of immune-suppressive ligands, inhibits expression of immune-activating molecules and NO production, and modulates migration of immune cells (Supplementary Fig. 2A).

The meta-map is composed of 98 signaling pathways, 30 of which contain more than 10 molecules in the sequence (Supplementary Data 2). It is worth highlighting that there are many crosstalks between different signaling pathways (Fig. 3b).

The signaling pathways of the meta-map form together 25 functional modules. A module on the meta-map represents a group of signaling pathways collectively executing a phenotype, e.g. the functional module NO and ROS production contains several signaling pathways implicated in a single biological function (Supplementary Fig. 2B).

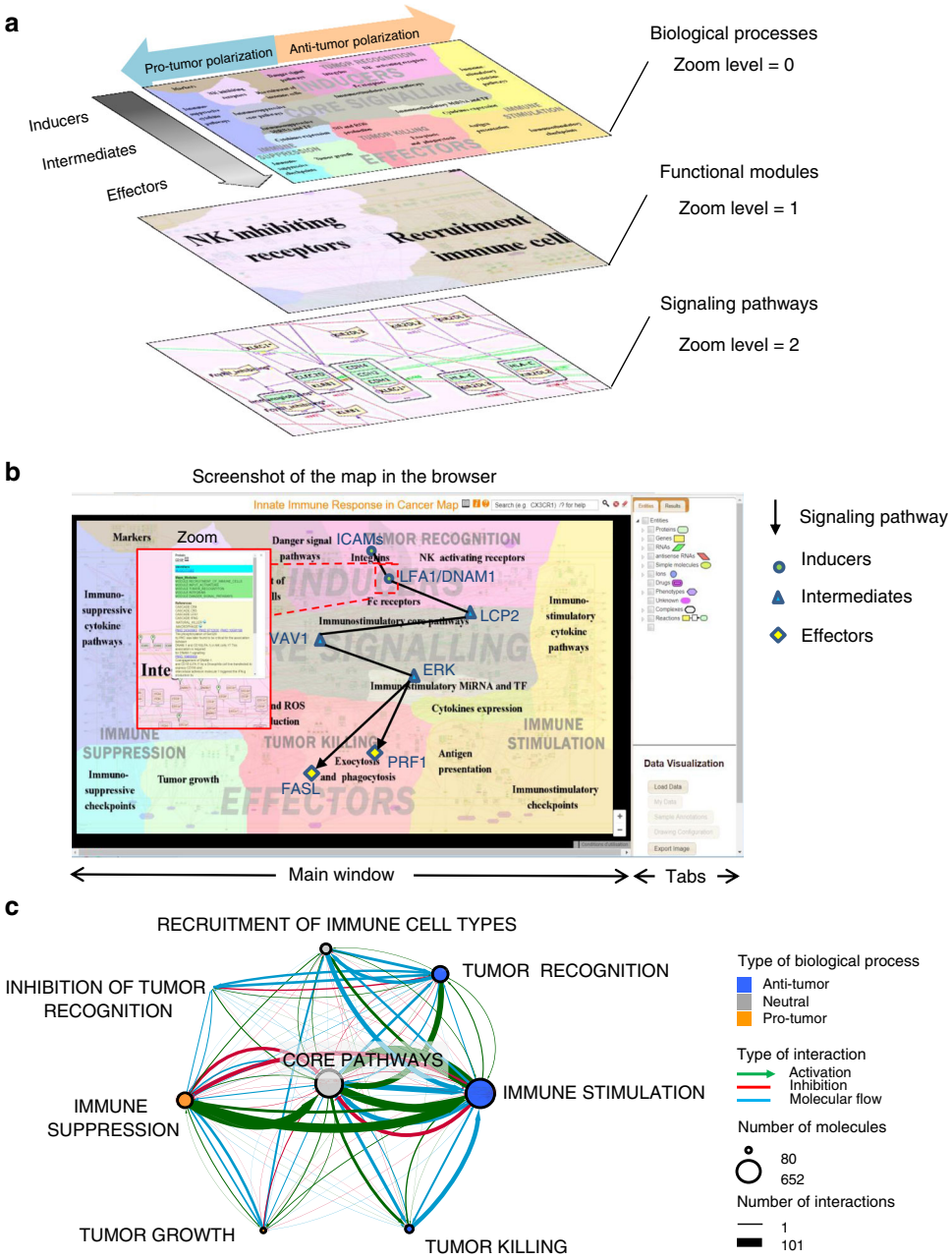


Fig. 3 Structure of meta-map of innate immune response in cancer. **a** Top view layout of the innate immune meta-map. Functional modules represent key processes involved in pro-tumor and anti-tumor activity of innate immunity in cancer, showed at different zoom levels (0—polarization and biological processes, 1—functional modules, 2— signaling pathways, molecules, interaction types, and annotation details. **b** Signaling pathways in the meta-map structure in a browser window. **c** The network of modules demonstrating crosstalks between biological processes represented on the meta-map. Nodes represent biological processes with the size associated to number of molecules in a process, color of the node is related to pro/anti-tumor polarization (see legend), interactions reflect cross-talk between the biological processes, the thickness of the edge is related to number of interactions and the color to the nature of interactions

These functional modules are assembled into the structures of higher level, namely nine biological processes (meta-modules), reflecting the major biological activities of the innate immune system with respect to a tumor, i.e. Tumor recognition, Inhibition of Tumor Recognition, Tumor Growth, Tumor Killing, Immune Stimulation, Immune Suppression, Recruitment of Immune Cells, Core Activation, and Core Inhibition.

Finally, at the highest level, all biological processes (meta-modules) are grouped into two zones representing the concept of innate immune system polarization into anti- or pro-tumor mode. The Anti-Tumor zone covers the meta-modules named Tumor Recognition, Immune Activation, Tumor Killing, and Core Activation, whereas the Pro-Tumor zone is composed of Inhibition of Tumor Recognition, Immune Suppression, Tumor Growth and Core Inhibition meta-modules (Figs. 1, 3a and Table 1). The list of map nodes per signaling pathways, modules, biological processes (meta-modules), and zones is available in the Supplementary Data 3 and downloadable form the resource website (https://navicell.curie.fr/pages/maps_innateimmune.html).

The various map levels are interconnected and cross-talk to each other. The crosstalks between the biological processes (meta-modules) are represented as an interaction network shown

Table 1 Hierarchical modular structure of innate immune response meta-map

Zones metamodule module	Chemical species as entities	Proteins	Genes	RNAs	asRNAs	Reactions	References
Zone: Pro-tumor polarization							
Inhibition of Tumor Recognition							
NK inhibiting receptors	35	23	1	1	0	14	57
Immune Suppression							
Immunosuppressive cytokine pathways	109	46	10	11	3	67	114
Immunosuppressive cytokine expression	55	19	14	14	0	36	75
Immunosuppressive chekpoints	8	7	0	0	0	8	13
Core Signaling Pathways							
Immunosupppressive core pathways	43	23	5	5	1	25	54
MIRNA TF Immunosuppressive	77	20	23	14	12	48	62
Tumor Growth							
Tumor growth	60	42	8	8	0	71	58
Zone: Anti-tumor polarization							
Tumor Recogntiton							
NK activating receptors	114	45	16	14	6	72	115
Danger signal pathways	60	30	2	1	0	36	66
FC receptors	18	12	0	0	0	8	37
Integrins	38	24	0	0	0	21	56
Immune Stimulation							
Immunostimulatory cytokine pathways	152	74	18	18	3	92	193
Immunostimulatory cytokine expression	43	17	12	11	1	27	109
Antigen presentation and immunostimulatory checkpoints	99	65	6	6	0	91	152
Core Signaling Pathways							
Immunostimulatory core pathways	184	93	6	6	114		244
MIRNA TF immunostimulatory	50	17	12	10	5	33	60
Tumor Killing							
Lytic granules exocytosis and phagocytosis	73	39	6	6	5	50	75
No ROS production	33	10	4	4	0	23	44
Cell-type specific markers							
Markers							
Markers macrophage	22	10	6	6	0	0	8
Markers NK	10	10	0	0	0	0	36
Markers mast	6	6	0	0	0	0	9
Markers DC	16	14	0	2	0	0	14
Markers neutrophil	11	11	0	0	0	0	15
Markers MDSC	9	9	0	0	0	0	9
Recruitment							
Recruitment of immune cells							
Recruitment of immune cells	103	48	17	17	0	93	83
Meta-map	1466	582	162	152	20	1084	820

Structure and content of innate immune meta-map

in the Fig. 3c. The interaction network demonstrates different types of links between meta-modules of the map, including activation, inhibition, molecular flow. The Core Signaling meta-module is a network "hub" where most signaling pathways converge. In addition, it is notable that there are numerous positive and negative crosstalks between Immune Stimulation and Immune Suppression meta-modules on the map (Fig. 3c).

The external organization of the meta-map is reflected in the hierarchical structure of zoom levels, similar to geographical maps, where only limited information is displayed on each zoom level (Fig. 3a). This hierarchical structure facilitates Google Maps-like navigation of the map.

Access, navigation, and maintenance of the resource. The cell-type-specific and the integrated meta-map are open source, can be browsed online, and are available at https://navicell.curie.fr/pages/maps_innateimmune.html. Each map is presented under three independent platforms, namely NaviCell, MINERVA, and NDEx. All map components are clickable, making the map interactive. The extended annotations of map components contain rich tagging system converted to links and confidence scores. This allows tracing the involvement of molecules into different map sub-structures as pathways, modules, and biological processes (meta-modules) (Fig. 3). Tagging system also allows to use the meta-map as a source of annotated signatures (Supplementary Fig. 1).

The semantic zooming feature of NaviCell[35] simplifies the navigation through large maps of molecular interactions, showing readable amount of details at each zoom level.

Comparison of meta-map with existing pathway databases. The meta-map content (Supplementary Fig. 3) and the coverage of literature used to annotate the entities (Supplementary Fig. 4) were compared to a sub-set of pathways related to the innate immune system from the existing molecular interaction databases (Supplementary Table 2). The InnateDB database contains a detailed description of the innate-immune signaling, even though more general databases as KEGG and REACTOME also include immune pathways. A description of comparison procedure is provided in the Methods.

We further compared the major features of innate immune response representation in different pathway databases. The

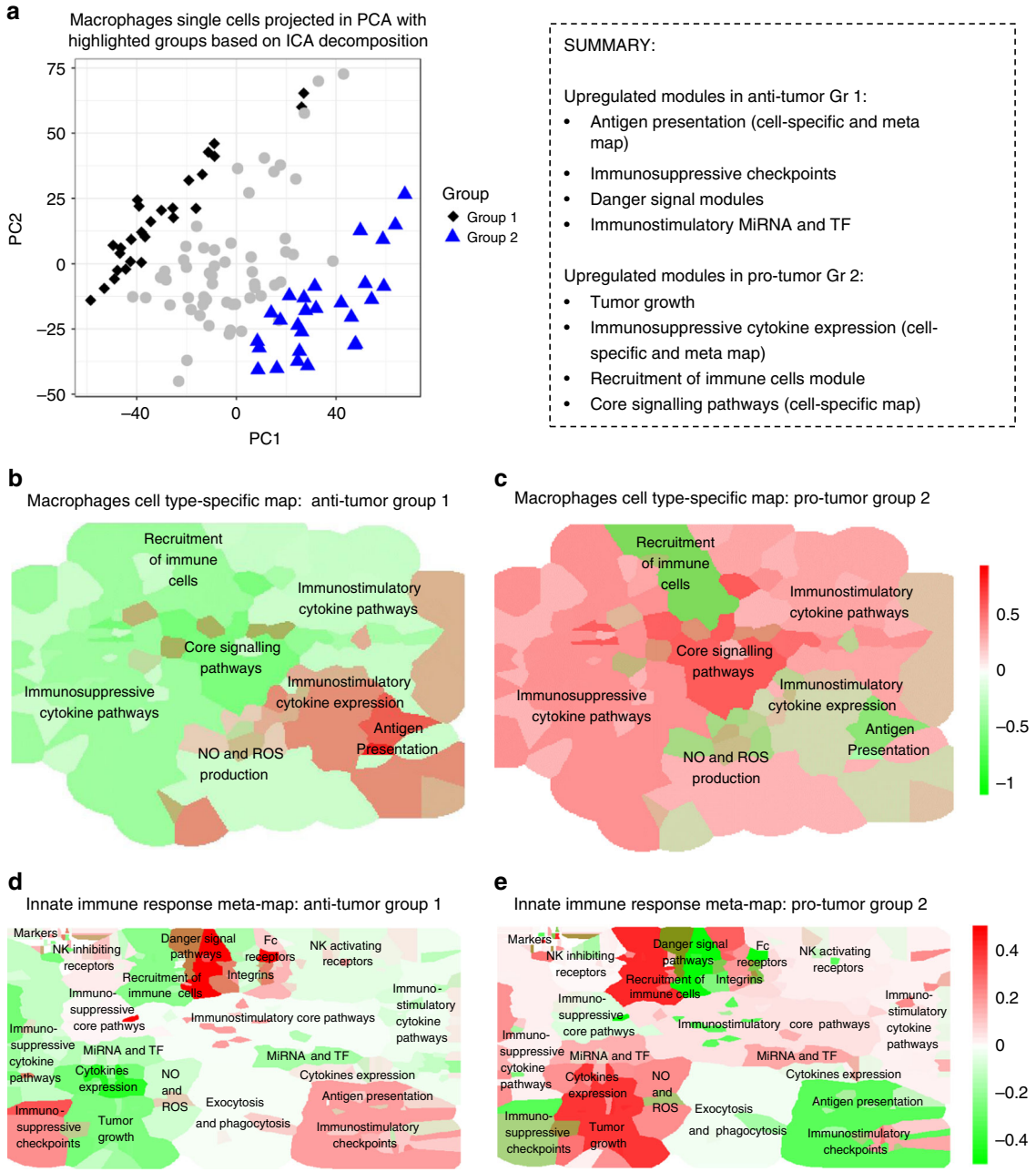


Fig. 4 Visualization of modules activity scores using expression data from melanoma macrophages. **a** Macrophages single cells in PC1 and PC2 coordinates space. Two groups, the first and the fourth quartiles of distribution along the IC1 axis, are colored distinctly in blue and black. Staining of the macrophage cell-type-specific map with modules activity scores calculated from single-cell RNAseq expression data for **b** Macrophages group 1 (Anti-tumor) and **c** Macrophages Group 2 (pro-tumor) cells. Staining of the innate immune meta-map with modules activity scores calculated from single-cell RNAseq expression data for **d**. Macrophages Group 1 (Anti-tumor) and **e** Macrophages Group 2 (Pro-tumor) cells. Color code: red—upregulated, green—downregulated module activity

innate immune response in cancer resource contains cell-type-specific maps in contrast to other databases. The comparison indicates that the cross-talk between the pathways is visually represented at the maps of immune response in cancer resource. Finally, the combination of hierarchical organization of knowledge and possibility of navigation through the layers of the maps thanks to semantic zooming feature makes the innate immune resource more suitable for meaningful data visualization. The visualization tool box is built into the NaviCell environment which allows easy data integration and visualization in the context of the innate immune maps (Figs. 4 and 5).

Taken together, the results of database comparisons indicate that the innate immune response in cancer resource is topic-specific, and describes immune-related and cancer-relevant signaling processes based on the latest publications about innate immune component in TME. The thoughtful layout and visual organization of the biological knowledge on the maps makes it a distinguished resource for data analysis and interpretation.

Application of the maps for data visualization and analysis. The cell-type-specific maps and the meta-map were applied to explore the heterogeneity of innate immune cell types in cancer.

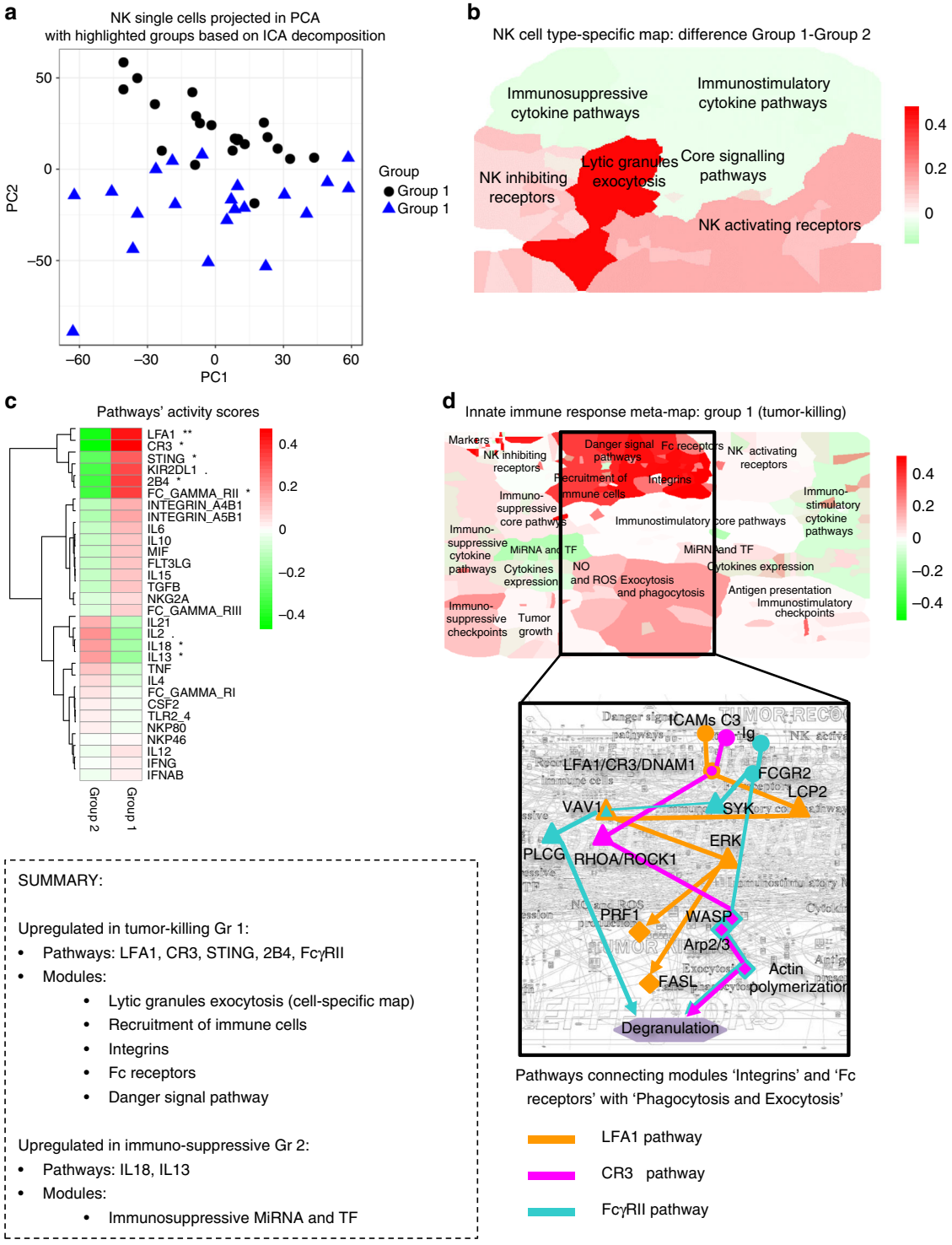


Fig. 5 Visualization of modules activity scores using expression data from melanoma natural killers (NK). NK single cells in PC1 and PC2 coordinates space. Two groups are colored distinctly in blue and black. **a** Map staining of the NK cell-type-specific map with modules activity scores calculated from single-cell RNAseq expression data for **b** NK Group 1. **c** Heatmap of activity scores in signaling pathways of NK groups. Map staining of the innate immune response meta-map with modules activity scores for **d** NK Group 1 ("tumor killing") with a zoom into three signaling pathways relating the two upregulated modules: "Danger signal pathways" and "Exocytosis and phagocytosis" with main molecular players named. Color code: red–upregulated, green—downregulated module activity. The *p* values of the *t* test were reported in the heatmaps with the standard code of significance (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, < 0.1)

The single-cell RNA-Seq data for macrophages and NK cells from metastatic melanoma samples were used[45].

A matrix factorization technique, independent components analysis (ICA)[46] allows ranking genes or samples along data-driven axes. The independent components instead of detecting highest variability axes as PCA, extract independent and non-Gaussian signals called components. The most stable component was used as a way to order the cells based on some latent process that we aim to interpret using innate immune maps. In order to better understand the differences in the cell ranking, the cells with

extreme rank values were selected, which resulted in Groups 1 and 2. When projected in the PCA space (Fig. 4a), those macrophage cell groups are lying on the borders of the cloud of points.

Furthermore, the activity scores were computed for each macrophage cell group (as defined in the Methods) for functional modules at different levels: pro- and anti-tumor general classification, innate map modules, and macrophage-specific map modules.

First, the analysis of potential pro- and anti-tumor properties of the macrophage cell groups was examined in the context of the innate immunity meta-map. Group 1 has significantly higher anti-tumor score (*t*-test *p* value: 0.02) and Group 2 is the pro-tumor one (*t*-test *p* value: 0.003). Second, the expression profile differences of the cells from the two groups were interpreted in the context of the Macrophage cell-type-specific map and the innate immune response meta-map. The results of the enrichment study for the two Macrophage groups were also represented as heatmaps with a significance level of *p* value for Student's *t*-test (see Methods) (Supplementary Fig. 5). The module activity values were plotted on the maps using BiNoM plugin of Cytoscape[47].

Visualization of the module activity scores in the context of macrophage cell-type-specific demonstrates that the module Antigen Presentation is upregulated in Macrophage Group 1 (Fig. 4b) comparing to Macrophage Group 2 (Fig. 4c). Whereas, Macrophage Group 2 (Fig. 4c) shows upregulated modules Core Signaling Pathways and Immunosuppressive Cytokines Pathways comparing to Macrophage Group 1 (Fig. 4b).

Then, the module activity scores for the two Macrophage cell groups were analyzed in the context of the meta-map that allowed to detect several additional modules differentially regulated between the two groups. The four modules Antigen Presentation, Immunosuppressive Checkpoints, Danger Signal Module, and Immunostimulatory MiRNA and TF were significantly over-expressed in Anti-tumor Macrophage Group 1 (*t*-test *p* values, respectively: $<10^{-4}$, 0.009, $<10^{-8}$, $<10^{-5}$, Fig. 4d) compared to Pro-tumor Macrophage Group 2 (Fig. 4e). On the contrary, the three modules Recruitment of Immune Cells Module, Tumor Growth, and Immunosuppressive Cytokine Expression were strongly upregulated in Pro-tumor Macrophage Group 2 (*t*-test *p* values, respectively: $<10^{-6,}$ $<10^{-6}$, $<10^{-5}$, Fig. 5d). in comparison to Anti-tumor Macrophage Group 1 (Fig. 4d, e).

From these results, it can be concluded that the Macrophage Group 1 has a tendency to express an anti-tumor phenotype, because it is characterized by the expression of inflammatory cytokines that are able to induce local adaptive immunity via antigen presentation process. Interestingly, the most typical modules responsible for tumor elimination as Exocytosis and Phagocytosis and Immunostimulatory Cytokine Pathways are not over-activated in this cell sub-set. In contrary, Macrophage Group 2 demonstrated a pro-tumor phenotype, characterized by expression of immunosuppressive cytokines restricting local immune response and growth factors supporting tumor growth.

Alike macrophages, NK cells were ranked along a latent variable obtained with ICA algorithm. Due to low cell number available, the 42 single NK cells were split in half according to the ICA ranks. Subsequently, the module activity scores were computed of each group and then a *t*-test was applied to evaluate the difference in module activity between the two NK subpopulations (Group 1 referred to as Tumor Killing and Group 2 referred to as Immunosuppressed) (Fig. 5a, Supplementary Fig. 6).

First, the comparison and visualization of the module activity between the two NK cells groups demonstrated the activation of Lytic Granules Exocytosis module in NK Group 1 compared to NK Group 2 (*t*-test *p* value: 0.006), on the NK cell-type-specific map (Fig. 5b). The activity of this module is directly responsible of tumor killing capacity of NK Group 1 cells that most probably exposes stronger anti-tumor abilities compared to Group 2 (Supplementary Fig. 6A).

Next, the two NK cells groups were analyzed in the context of the meta-map that allowed detection of five differentially regulated modules between the two groups of NK cells (Fig. 5d). The four modules Recruitment of Immune Cells, Integrins, Fc Receptors, and Danger Signal Pathway were significantly upregulated in the NK Group 1 comparing to the NK Group 2 (*t*-test *p* values, respectively: 0.0001, $<10^{-4}$, 0.004, $<10^{-5}$). In contrary, the module Immunosuppressive MiRNA and TF was inhibited in the NK Group 1 comparing to the NK Group 2 (*t*-test *p* value: 0.001). Finally, although the activity of Phagocytosis and Exocytosis module is not significantly different between the two groups, this module is rather activated in the NK Group 1 compared to the NK Group 2 (Supplementary Fig. 6B).

Collectively these results demonstrate that the NK Group 1 is characterized by upregulation of biological functions related to NK cell recruitment and activation, coinciding with upregulation of the mechanisms responsible for tumor killing. Thus, the NK Group 1 can be interpreted as newly recruited, actively migrating NKs with strong anti-tumor polarization. In contrary, most probably, NK Group 2 contains resting or suppressed NK cells that do not expose a well-defined phenotype.

The activation of upstream map zones and downstream effector zones in NK Group 1 is notable (Fig. 5d). However, which mechanisms coordinate this co-activation is not clear. The structure of the network was analyzed to address this question and the signaling pathways connecting the two activated zones were retrieved. The activation state of 30 signaling pathways from the meta-map was assessed for the cell from Group 1 and Group 2 (Fig. 5c). There are all together seven differentially regulated pathways between the two cell groups. Five are upregulated pathways in Group 1 (LFA1, CR3, STING, 2B4, FcγRII) and two upregulated pathways in Group 2 (IL13, IL18) (*t*-test *p* values <0.05).

Within the pathways activated in the Group 1 there are three pathways regulated through receptors LFA1, CR3, and FcγRII. The key players of the pathways are presented schematically in Fig. 5d. The meta-map described difference between NK subtypes both on the level of functional modules and signaling pathways. It allows us to draw the conclusion that tumor recognition via these pathways plays an even more important role for NK-activation than well studied activation via classical NK receptors.

Meta-map as a source of patient survival signatures. To study whether the innate immune response meta-map can be used for assessment of processes contributing to patient survival, the list of genes from the map was used to find correlation with prognosis of patient survival using data published elsewhere[48] (see Methods). First, the presence of the genes on the innate immune response meta-map correlating with the patient survival from the aforementioned study was verified. It was detected that out of 627 proteins and protein coding genes depicted on the meta-map, 295 are significantly correlated with patient survival (*z*-score *p* value < 0.05), that represents 47% of the map content vs. 27% in the whole genome study[48] (Supplementary Data 1).

The genes enriched on the meta-map can be divided into two groups, positively and negatively correlated with the patient survival, which confirms the observation that innate immune system can play a dual role in cancer disease. Interestingly, from the whole genome analysis in the original study by Gentles et al. (2015)[48], it emerges that there is quasi equal proportion of positively and negatively correlated genes. However, in the innate immune response meta-map, there is a strong predominance of genes positively correlated with patient survival (Table 2).

Table 2 Distribution of genes with positive ($z < 0$) and negative ($z > 0$) correlation with patient survival across functional meta-modules in innate immune response meta-map

Innate immune map meta-module	Mean z-score	Positive correlation with patient survival	Negative correlation with patient survival
Tumor Growth	1.3	12	26
Inhibition of Tumor Recognition	−1.86	18	6
Tumor Recogntiton	−1.56	67	28
Recruitment of Immune Cells	−0.94	29	14
Immune Stimulation	−0.53	122	87
Tumor Killing	−0.5	25	29
Core Signaling Pathways	−0.46	114	84
Immune Suppression	−0.33	39	24

Values indicate number of genes

In order to highlight what biological functions on the innate immune response in cancer meta-map are associated to positive or negative patient survival, mean values of gene *z*-scores per meta-modules were calculated and visualized in the context of the meta-map (see Methods). As a general trend, the meta-map layers Inducers and Core Signaling are more significantly correlated with patient survival, compared to the layer Effectors. Furthermore, the meta-modules with biological functions related to anti-tumor activity as Immune Response Stimulation and Tumor Recognition, Recruitment of Immune Cells, etc. are positively correlated with patient survival. Interestingly the meta-module Tumor Killing is also positively correlated with patient survival, though not reaching the statistical significance (Table 2, Supplementary Fig. 7). The minority of meta-modules related to pro-tumor activity as Tumor Growth, Immunosuppressive Core Pathways, Immunosuppressive MiRNA and TF correlated negatively with patient survival (Table 2, Supplementary Fig. 7). The described analysis demonstrates that the meta-map can serve for evaluation of innate immune response signatures associated with patient survival in cancer.

Discussion

One of the challenges of cancer biology today is understanding the phenomena of tumor heterogeneity. It consists of two relatively independent parts: first, heterogeneity of the tumor cells themselves, as a result of their clonal divergence or action of epigenetic mechanisms; second, heterogeneity of tumor microenvironment (TME). Recent years discoveries have shown that understanding how the components of this multicellular TME system interact with each other is very important for effective drug design. Actually, the attempt to modulate the interactions within the tumor microenvironment lies on the basis of new anti-cancer immune checkpoint inhibition therapy.

The analysis of large amounts of scientific information and the creation of optimal forms of its representation, require the development of new approaches for network map construction and annotation. Our first goal was to preserve the natural multidimensionality of the biological knowledge available for the different cell types in the innate component of the TME. Indeed, different cells types in innate immune system are studied from different angles. Some signaling pathways are described in detail for the macrophages and others for natural killer cells and so on. It is clear that the molecular knowledge described for one cell-type cannot always be extrapolated to another. This motivated us to create two complementary representations of innate immune system in cancer, one in the form of cell-type-specific maps and the second as an integrated meta-map of innate immune response in cancer. To be able to trace the correspondence of molecular entities and processes to a particular cell type, we introduced a system of cell-type-specific tags, included into the annotation of all entities on the maps.

Our second goal was to provide a complete and not controversial picture on the processes occurring in the TME. The generation of an integrated meta-map of innate immunity immediately exposed a problem of map complexity. We coped with the complexity problem by introducing the hierarchical structure into the integrated meta-map, respecting the biological functions. The general layout of the integrated meta-map is based on the idea of immune cells polarization in TME, reflected in the representation of both, pro-tumor and anti-tumor signaling mechanisms. In accordance with the literature, all functional modules and meta-modules on the map are grouped into pro-tumor and anti-tumor zones. There two types of signaling modes lead to the corresponding phenotypes. In addition, the mechanism responsible for a switch in the polarization state is also represented.

The modular hierarchical map structure and complex tagging system of maps entities facilitated the production of geographical-like easily browsable open source repository. Taking an advantage of NaviCell platform, which provides Google Maps-engine and map navigation features, the innate immune maps can be explored in an intuitive way, allowing the shuttling between the cell-type-specific maps to the integrated meta-map.

NaviCell-based representation of the maps facilitates visualization of various types of omics data. Analysis of data in the context of both, cell-type-specific and integrated maps, can help in the formalization of biological hypotheses for the processes and interactions that are studied in some cell types, but unexplored in others. In addition, thanks to the rich system of tags, the maps content can be used as a source of knowledge-based gene signatures of innate immune cell type. Finally, hierarchical organization of the map provides a basis for structural network analysis, complexity reduction, and eventual transformation of the map into executable mathematical models.

The integration of the innate immune response in cancer resource into additional platforms allows broader exposure and use of the valuable maps. Therefore, in addition to NaviCell platform, the resource is also exposed in the MINERVA platform and integrated into the NDEx repository and platform. In the future, the resource will be also integrated into larger pathway collections. These moves will allow a deeper involvement of the scientific community into the maintenance and update of the maps with the latest discoveries.

The resource of innate immune maps is useful for computing network-based molecular signatures of innate immune cells polarization. These signatures will help to characterize the overall status of the signaling dictating pro-tumor and anti-tumor states of TME in cell lines and tumoral samples. It will also help to stratify cancer patients according to the status of the TME and

potentially predict patient survival and response to immunotherapies. In addition, the resource might potentially provide new immunotherapy targets, among innate immunity components of TME in tumor infiltrates. These targets can be complementary or synergistic to the well-known immune checkpoint inhibitors.

As other studies show, similar resources are used for omics data visualization in the context maps that can provide network-based molecular portraits of studied cases. Comprehensive maps are rich in molecular details carefully compiled together, therefore structural analysis of the maps can explain particular phenotypes, redundancies, and robustness[49,50]. Such analysis together with omics data can guide to design of complex druggable interventions[51]. Further, complex maps contain modules that correspond to particular biological processes; therefore, the content of these modules are used as signatures of the corresponding biological functions[52]. These lists of genes are frequently used for enrichment studies[53].

Construction of the innate immune response map is the first step in the attempt to build a global network describing the molecular interactions in the TME. The next perspective is to represent the knowledge on adaptive immune response and non-immune components in the tumor environment, including fibroblasts and endothelial cells. The final goal is to build a complete map of signaling in cancer representing both intracellular interactions of tumor cells and each component in the TME and their intracellular interactions, and describing the coordination among the components of this multicellular system.

In addition, being included into a broader Disease Maps project, the meta-map of innate immune response will be helpful, together with maps or other diseases, in the study of disease comorbidities and drug repositioning[54,55].

Methods

Map and model. The maps sere drawn in CellDesigner diagram editor[34] using Process Description (PD) dialect of Systems Biology Graphical Notation (SBGN) syntax which is based on the Systems Biology Markup Language (SBML)[33]. The data model used includes the following molecular objects: proteins, genes, RNAs, antisense RNAs, simple molecules, ions, drugs, phenotypes, complexes. These objects can play the role of reactants, products, and regulators in a connected reaction network. The objects phenotypes play a role biological process outcome or readout (e.g. Migration, Tumor killing, ROS production, etc). Edges on the maps represent biochemical reactions or reaction regulations of various types. Different reaction types represent post-translational modifications, translation, transcription, complex formation or dissociation, transport, degradation and so on. Reaction regulations include catalysis, inhibition, modulation, trigger and physical stimulation. The naming system of the maps is based on HUGO identifiers for genes, proteins, RNAs and antisense RNAs and CAS identifiers for drugs, small molecules, and ions.

Manual literature mining. The molecular interactions reported in the scientific articles were manually curated and the information extracted from the papers was used for reconstruction and annotation of the maps. Three types of articles were used for map annotation: (i) experimental innate-immunity specific articles directly or indirectly confirming molecular interactions based on mammalian experimental data; (ii) review articles; (iii) experimental articles from non-immune cells that helped to complement the mechanisms present in immune cells (3% of the literature used for the map). In addition, pathway databases were used to retrieve information of the canonical pathways reported for the innate immune signaling general pathway databases (e.g. KEGG, REACTOME, SPIKE SignaLink, EndoNET) or in the immune system-specialized resources such as VirtuallyImmune (http://www.virtuallyimmune.org) and InnateDB (www.innatedb.com).

Map structure and tag×ging system. The annotation of each molecular object on the maps (protein, gene, RNA, small molecule, etc.) includes several tags indicating participation of the object in signaling pathways (tag PATHWAY:NAME), functional modules (tag MODULE:NAME), and cell-type-specific map (tag: MAP:NAME). Each PATHWAY obtains the name of the initiating ligand or receptor, in case when several ligands are acting through the same receptor. The tags are converted into the links by the NaviCell factory in the process of online map version generation. The links allow to trace participation of entities in different cell-type-specific maps and the sub-structure of the same map (pathway, module, biological process) and also facilitate shuttling between these structures.

Reaction and protein complex confidence scores. To provide information on the reliability of the depicted molecular interactions, two confidence scores have been introduced. Both scores represent integer numbers varying from 0 (undefined confidence) to 5 (high confidence). The reference score (REF) indicates both the number and the "weight" associated with publications found in the annotation of a given reaction. The functional proximity score (FUNC) is computed based on the external network of protein–protein interactions (PPI), InnateDB, which contains both experimental and literature-based curated interaction data[28]. The score reflects an average distance in the PPI graph between all proteins participating in the reaction (reactants, products, or regulators).

Map entity annotation in NaviCell format. The annotation panel followed the NaviCell annotation format of each entity of the maps includes sections Identifiers, Maps_Modules, References, and Confidence as detailed in ref. [32]. Identifiers section provides standard identifiers and links to the corresponding entity descriptions in HGNC, UniProt, Entrez, SBO, GeneCards, and cross-references in REACTOME, KEGG, Wiki Pathways, and other databases. Maps_Modules section includes tags of modules, meta-modules, and cell-type-specific maps in which the entity is implicated (see above). References section contains links to related publications. Each entity annotation is represented as a post with extended information on the entity.

Generation of NaviCell map with NaviCell factory. CellDesigner map annotated in the NaviCell format is converted into the NaviCell web-based front-end, which is a set of html pages with integrated JavaScript code that can be launched in a web browser for online use. HUGO identifiers in the annotation form allow using NaviCell tool for visualization of omics data. A detailed guide of using the NaviCell factory embedded in the BiNoM Cytoscape plugin[47] is provided at https://navicell.curie.fr/doc/NaviCellMapperAdminGuide.pdf.

Depositing maps at several web-based platforms. Cell-type specific maps and the meta-map of innate immune response in cancer were made available at other platforms such as MINERVA and NDEx. To integrate maps within NDEx, CellDesigner maps were first loaded in Cytoscape using the BiNoM Cytoscape plugin and then uploaded on NDEx using the CyNDEx Cytoscape plugin.

Databases content comparison. Pathways related to the human innate immune system were selected from the InnateDB 5.4 version, except Complement Cascade (Human), NOD-like Receptor Signaling Pathway, Regulation of Autophagy (Human), and RIG-I-Like Receptor Signaling Pathway (Human). The excluded pathways represent virus and bacterial infection-specific pathways that do not correspond to TME signaling. The innate immune-related pathways from KEGG 84.1 version were retrieved from the list 5.1-Immune System. The pathways obtained from REACTOME 63rd version cover Class I MHC Mediated Antigen Processing & Presentation, MHC Class II Antigen Presentation from Adaptive Immune Branch, and all pathways from Innate Immune Branch. All together 666 gene names from InnateDB 5.4, 563 gene names from KEGG 84.1, and 2156 gene names from REACTOME 63rd were selected. These lists were compared with the innate immune response meta-map that contains 683 gene names. The complete list of selected pathways with gene names is available in the Supplementary Data 2).

The selected InnateDB pathways contain altogether, nearly the same number of objects as the innate immune response meta-map (Supplementary Data 4). The content of selected KEGG or REACTOME pathways is richer than in the innate immune response meta-map, due to the fact that KEGG and REACTOME are generic databases, describing all innate immune-related interactions, whereas the meta-maps is rather oriented to cancer signaling. The overlap between the meta-map and the three selected databases represents 61% for InnateDB, 58% for KEGG, and 30% for REACTOME. It is important to note that there are 188 genes that present exclusively at the innate immune response meta-map (Supplementary Fig. 4A, Supplementary Data 2). These unique genes are relatively homogeneously distributed across the meta-map, indicating that the depicted processes are described in more depth on the meta-map compared to the other three databases (Supplementary Fig. 3A). Several modules are significantly enriched by unique genes on the meta-map (Supplementary Fig. 3B). Thus, the modules Tumor Growth and Immunosuppressive Checkpoints contain signaling that are very well studied in cancer cells and therefore represented in great details on the meta-map. Two additional modules, entitled MIRNA TF Immunostimulatory and MIRNA TF Immunosuppressive, contain the latest information of miRNA involvement in the innate immune system control in cancer and unique for the meta-map, compared to other databases. It was concluded that the content of the meta-map is not redundant with the other pathway databases and that several functional modules directly related to TME functions are unique to the meta-map.

Databases annotation literature comparison. In addition, the sets of publications used to annotate the InnateDB resource and aforementioned preselected

pathway from REACTOME resource were compared to the set of publications used in the meta-map. The overlap of the literature body from the meta-map with references from InnateDB and REACTOME databases is relatively small, because 785 papers out of 820 papers that were used to annotate the meta-map are unique (Supplementary Fig. 4B). It confirms that the meta-map is not a mechanical compilation of existing databases, but rather an independent resource. It formalizes the part of biological knowledge which was not annotated before and highlights the difference between reconstruction of generic and cell-type specific pathways in terms of literature sources.

Although the median age of the literature references in the meta-map is only one year-younger compared to InnateDB and REACTOME, there is a 27% of papers dating 2010–2017 in the literature body annotating the meta-map. The literature set in the meta-map contains more papers published after year 2010 than in InnateDB and REACTOME, indicating that the meta-map represents the most recent discoveries in the corresponding fields (Supplementary Fig. 4C).

Finally, the journal types represented in the three databases were also compared. The choice of the journals used for annotating the meta-map and the other two databases is similar; however, the distribution of the papers from different types of journals is not even. The annotations of meta-map mainly contain papers from immunological journals such as *Journal of Immunology*, *Immunity*, *Nature Immunology*, and cancer-specific journals, such as *Cancer Research* and *Oncogene*, comparing to the other two databases. The annotations of InnateDB and REACTOME are rather oriented towards more generic molecular biology journals as *JBC*, *MCB*, *Nature,* and *PNAS* (Supplementary Fig. 4C and D).

High-throughput data analytical pipeline. Normalized melanoma data sets from GEO (GSE72056)[45] were transformed into log expression levels and mean centered. The exploratory analysis and statistical testing was performed and visualized using R packages (ggplot2, stats, pheatmap)[56–58] then MATLAB ICA implementation of FastICA algorithm[46] and icasso package[59] to improve the stability. Colored map images were obtained using function "Stain CellDesigner map" from BiNoM Cytoscape plugin[47] using .xml map files and the mean expression from the analysis described below.

Analytical pipeline. The single-cell molecular profiles are characterized by high variability that have both biological and technical origin. A common practice is to group single cells in order to make an aggregated representative profile that minimizes the technical biases but still represents finer level of granularity than a bulk sample. In order to define cell groupings that would lead to functional interpretation we used a matrix factorization technique called ICA

ICA is a matrix factorization-based technique aiming at defining statistically independent hidden factors shaping gene expression. Stability-based analysis revealed only one sufficiently stable independent component in the case of both Macrophage and NK data subsets. Therefore, first independent component was used to rank the individual cells. We grouped the NK single cells depending on the first independent component (IC1) projection score such that Group 1 had positive projection scores and the Group 2 has negative projection scores. For macrophage single cells we selected the first and the last quartiles of the macrophage scores of IC1 projection. In order to best interpret the "extreme" tendencies of the cells placed on the opposite side of IC. The distinction of the groups plotted in first and the second principal components space (PC1 and PC2) can be seen in Figs. 4a and 5a.

For cell groups defined as described above, the following procedure was applied in order to define the map module scores. For each module, 50% of most variant genes were retained in order to select genes over the median variability. The module score was defined as the mean of the selected genes.

Standard *t*-test was used to assess statistical differences between single-cell groups for each module. The *p* values of the *t*-test were reported in the heatmaps with the standard code of significance (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, <0.1).

The data on pan-cancer meta-analysis of expression signatures from ~18,000 human tumors across 39 malignancies accompanied by survival clinical data were used[48]. In total, 6323 genes with significant *z*-scores (*p* value <0.05) indicating correlation to patient survival were retrieved[48] and overlapped with the gene lists from the innate immune response meta-map. Enrichment of the meta-map with the genes significantly positively or negatively correlated with patient survival was assessed using the χ^2 test with *p* value threshold 0.001.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The cell-type-specific maps and meta-map of innate immune response in cancer are freely available at the web page (https://navicell.curie.fr/pages/maps_innateimmune.html). The meta-map and cell-type-specific maps are provided in three platforms, NaviCell, MINEVRA and integrated into the repository NDEx. The maps exist and can be downloaded in several exchange formats (CellDesigner SBML level 2 version 4, SBGN-ML 0.2, SBML level 3 version 1, Cytoscape CX version 3.4.0). In addition, the composition of map signaling pathways, modules, and meta-modules is provided in a form of GMT files (Supplementary Tables 2 and 3, respectively) suitable for further functional data analysis. A network of binary relations between proteins generated from the meta-map and the complete list of references annotating the maps are also available.

Code availability

The documentation and the scripts for module activity calculation and generation of life example is provided at GitHub (https://github.com/sysbio-curie/NaviCell/tree/master/auxiliary_scripts). The step-by-step procedure on modular hierarchical maps construction is also provided at https://github.com/sysbio-curie/NaviCell.

Received: 24 May 2018; Accepted: 2 August 2019;
Published online: 22 October 2019

References

1. Becht, E., Giraldo, N. A., Dieu-Nosjean, M.-C., Sautès-Fridman, C. & Fridman, W. H. Cancer immune contexture and immunotherapy. *Curr. Opin. Immunol.* **39**, 7–13 (2016).
2. Calì, B., Molon, B. & Viola, A. Tuning cancer fate: the unremitting role of host immunity. *Open Biol.* **7**, 170006 (2017).
3. Laoui, D. et al. Mononuclear phagocyte heterogeneity in cancer: different subsets and activation states reaching out at the tumor site. *Immunobiology* **216**, 1192–1202 (2011).
4. Van Overmeire, E., Laoui, D., Keirsse, J., Van Ginderachter, J. A. & Sarukhan, A. Mechanisms driving macrophage diversity and specialization in distinct tumor microenvironments and parallelisms with other tissues. *Front. Immunol.* **5**, 127 (2014).
5. Chávez-Galán, Leslie, Olleros, M. L., Vesin, D. & Garcia, I. Much more than M1 and M2 macrophages, there are also CD169+ and TCR+ macrophages. *Front. Immunol.* **6**, 263 (2015).
6. Vesely, M. D., Kershaw, M. H., Schreiber, R. D. & Smyth, M. J. Natural innate and adaptive immunity to cancer. *Annu. Rev. Immunol.* **29**, 235–271 (2011).
7. Goswami, K. K. et al. Tumor promoting role of anti-tumor macrophages in tumor microenvironment. *Cell Immunol.* **316**, 1–10 (2017).
8. Fridlender, Z. G. & Albelda, S. M. Tumor-associated neutrophils: friend or foe? *Carcinogenesis* **33**, 949–955 (2012).
9. Gordon, J. R., Ma, Y., Churchman, L., Gordon, S. A. & Dawicki, W. Regulatory dendritic cells for immunotherapy in immunologic diseases. *Front. Immunol.* **5**, 7 (2014).
10. Cooper, M. A., Fehniger, T. A. & Caligiuri, M. A. The biology of human natural killer-cell subsets. *Trends Immunol.* **22**, 633–640 (2001).
11. Mittal, D., Gubin, M. M., Schreiber, R. D. & Smyth, M. J. New insights into cancer immunoediting and its three component phases–elimination, equilibrium and escape. *Curr. Opin. Immunol.* **27**, 16–25 (2014).
12. Marvel, D. & Gabrilovich, D. I. Myeloid-derived suppressor cells in the tumor microenvironment: expect the unexpected. *J. Clin. Invest.* **125**, 3356–3364 (2015).
13. Vo, M.-C. et al. Combination therapy with dendritic cells and lenalidomide is an effective approach to enhance antitumor immunity in a mouse colon cancer model. *Oncotarget* **8**, 27252–27262 (2017).
14. Mantovani, A., Marchesi, F., Malesci, A., Laghi, L. & Allavena, P. Tumour-associated macrophages as treatment targets in oncology. *Nat. Rev. Clin. Oncol.* **14**, 399–416 (2017).
15. Bonelli, S., et al. Beyond the M-CSF receptor—novel therapeutic targets in tumor-associated macrophages. *FEBS J.* **285**, 777–787 (2018).
16. Moynihan, K. D. & Irvine, D. J. Roles for Innate Immunity in Combination Immunotherapies. *Cancer Res.* **77**, 5215–5221 (2017).
17. O'Sullivan, T. et al. Cancer immunoediting by the innate immune system in the absence of adaptive immunity. *J. Exp. Med.* **209**, 1869–1882 (2012).
18. Tokunaga, R. et al. CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation—a target for novel cancer therapy. *Cancer Treat. Rev.* **63**, 40–47 (2018).
19. Bellora, F. et al. Imatinib and Nilotinib off-target effects on human NK cells, monocytes, and M2 macrophages. *J. Immunol.* **199**, 1516–1525 (2017).
20. Gebremeskel, S. et al. Natural killer T-cell immunotherapy in combination with chemotherapy-induced immunogenic cell death targets metastatic breast cancer. *Cancer Immunol. Res.* **5**, 1086–1097 (2017).
21. Kreuzinger, C., et al. A complex network of tumor microenvironment in human high-grade serous ovarian cancer. *Clin. Cancer Res.* **23**, 7621–7632 (2017)
22. Bhinder B., Elemento O. Towards a better cancer precision medicine: systems biology meets immunotherapy. *Curr. Opin. Syst. Biol.* **2**, 67–73 (2017).
23. Dorel, M., Barillot, E., Zinovyev, A. & Kuperstein, I. Network-based approaches for drug response prediction and targeted therapy development in cancer. *Biochem. Biophys. Res. Commun.* **464**, 386–391 (2015).
24. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).

25. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477 (2014).
26. Raza, S. et al. A logic-based diagram of signalling pathways central to macrophage activation. *BMC Syst. Biol.* **2**, 36 (2008).
27. Cavalieri, D. et al. DC-ATLAS: a systems biology resource to dissect receptor specific signal transduction in dendritic cells. *Immunome Res.* **6**, 10 (2010).
28. Breuer, K. et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228–D1233 (2013).
29. Gorenshteyn, D. et al. Interactive big data resource to elucidate human immune pathways and diseases. *Immunity* **43**, 605–614 (2015).
30. O'Hara, L. et al. Modelling the structure and dynamics of biological pathways. *PLoS Biol.* **14**, e1002530 (2016).
31. Kondratova, M., Sompairac, N., Barillot, E., Zinovyev, A. & Kuperstein, I. Signalling maps in cancer research: construction and data analysis. *Database* **2018**, bay036 (2018).
32. Kuperstein, I. et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis.* **4**, e160 (2015).
33. Le Novère, N. et al. The systems biology graphical notation. *Nat. Biotechnol.* **27**, 735–741 (2009).
34. Kitano, H., Funahashi, A., Matsuoka, Y. & Oda, K. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* **23**, 961–966 (2005).
35. Kuperstein, I. et al. NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Syst. Biol.* **7**, 100 (2013).
36. Biswas, S. K. & Mantovani, A. Macrophage plasticity and interaction with lymphocyte subsets: cancer as a paradigm. *Nat. Immunol.* **11**, 889–896 (2010).
37. Murray, P. J. & Wynn, T. A. Obstacles and opportunities for understanding macrophage polarization. *J. Leukoc. Biol.* **89**, 557–563 (2011).
38. Gabrilovich, D. I. & Nagaraj, S. Myeloid-derived suppressor cells as regulators of the immune system. *Nat. Rev. Immunol.* **9**, 162–174(2009).
39. Ostrand-Rosenberg S. & Sinha P. Myeloid-derived suppressor cells: linking inflammation and cancer. *J. Immunol.* **182**, 4499–4506 (2009).
40. Palucka, K. & Banchereau, J. Cancer immunotherapy via dendritic cells. *Nat. Rev. Cancer.* **12**, 265–277 (2012).
41. Vivier, E., Ugolini, S., Blaise, D., Chabannon, C. & Brossay, L. Targeting natural killer cells and natural killer T cells in cancer. *Nat. Rev. Immunol.* **12**, 239–252 (2012).
42. Fridlender, Z. G. et al. Polarization of tumor-associated neutrophil phenotype by TGF-beta: "N1" versus "N2" TAN. *Cancer Cell* **16**, 183–194 (2009).
43. Marichal, T., Tsai, M. & Galli, S. J. Mast cells: potential positive and negative roles in tumor biology. *Cancer Immunol. Res.* **1**, 269–279 (2013).
44. Theoharides, T. C. & Conti, P. Mast cells: the Jekyll and Hyde of tumor growth. *Trends Immunol.* **25**, 235–241 (2004).
45. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science.* **352**, 189–196 (2016).
46. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430 (2000).
47. Bonnet, E., et al. BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Syst. Biol.* **7**, 18 (2013).
48. Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
49. Chanrion, M. et al. Concomitant Notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut. *Nat. Commun.* **5**, 5005 (2014).
50. Grieco, L. et al. Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput. Biol.* **9**, e1003286 (2013).
51. Jdey, W. et al. Drug-driven synthetic lethality: bypassing tumor cell genetics with a combination of AsiDNA and PARP inhibitors. *Clin. Cancer Res.* **23**, 1001–1011 (2017)
52. Cantini, L. et al. Classification of gene signatures for their information value and functional redundancy. *NPJ Syst. Biol. Appl.* **4**, 2 (2018).
53. Monraz Gomez, L. C., et al. Application of Atlas of Cancer Signalling Network in preclinical studies. *Brief Bioinform.* **20**, 701–716 (2018)
54. Mazein, A. et al. Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ Syst. Biol. Appl.* **4**, 21 (2018).
55. Ostaszewski, M., et al. Community-driven roadmap for integrated disease maps. *Brief Bioinform.* **20**, 659–670 (2018).
56. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013).
57. Wickham H. *ggplot2 Elegant Graphics for Data Analysis*. Media. 211 (2009).
58. Perry, M. *Flexible Heatmaps for Functional Genomics and Sequence Features*. R package version 1.8.0. (Bioconductor, 2019).
59. Himberg J. & Hyvärinen A. ICASSO: Software for investigating the reliability of ICA estimates by clustering and visualization. *Neural Networks for Signal Processing—Proc. IEEE Workshop* 259–268 (IEEE, Toulouse, France, 2003).

Acknowledgements

We thank Daniel Rovera for help with network structure analysis and L. Cristobal Monraz Gomez for help with data visualization and critical reading of the paper. We thank Marek Ostaszewski and Piotr Gawron for integration of the resource into the MINERVA platform. This work has been funded by INSERM Plan Cancer No. BIO2014-08 COMET grant under ITMO Cancer BioSys program. This work received support from MASTODON program by CNRS (project APLIGOOGLE), COLOSYS grant ANR-15-CMED-0001-04, provided by the Agence Nationale de la Recherche under the frame of ERACoSysMed-1, the ERA-Net for Systems Medicine in clinical research and medical practice and by IMI2-IMMUcan grant. ITMO cancer (AVIESAN) provided 3-year PhD grant and foundation Bettencourt Schueller and Center for Interdisciplinary Research supported the training of the Ph.D. student.

Author Contributions

M.K. constructed signaling networks, performed data visualization, and wrote the paper; U.C. performed data analysis and enrichment calculations and wrote the paper; N.S. performed statistical analysis of maps content, integration of the resource into browsable platforms, and wrote the paper; S.D.A. and E.B. advised during the project and revised the paper; V.S. advised during the project and critically revised and restructured the paper; A.Z. supervised the data analysis, advised during the project, and revised the paper; and I.K. led the project and wrote the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-12270-x.

Correspondence and requests for materials should be addressed to I.K.

Peer review information *Nature Communications* thanks Tomáš Helikar and other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

III - Discussion, perspectives and conclusion

6. Discussion

I have presented in this work a series of methods aimed at enhancing our understanding of cancer functions through a combination of mathematical approaches and biological knowledge. My team has long expertise in the construction and usage of unsupervised approaches; and it is in this direction that I focused my efforts during my PhD. And since the data analysis method of Independent Component Analysis had a long history among the team successful applications to cancer data, it was with the vision of improving our usage and understanding of ICA applications that I embarked on this PhD project.

6.1 ICA vs other methods

With the constant increase of biological data, many researchers tried to look for existing mathematical approaches that could help solve various biological problems since many foundations were already established a long time before even computers existed (Gauthier et al., 2019). Some focused on the usage of methods that could be coupled with expert knowledge to ensure a certain curation of results. Some others however went on an exploratory trip with the sole hope of discovering new unexpected phenomena.

Although ICA wasn't a method initially developed for biological applications, it was still taken and applied to various types of omics and proved successful. Being part of the big family of matrix factorisation methods, one might ask why I would use ICA above others methods such as the popular NMF. Indeed, both of the methods are systematically compared in numerous other fields than biology. The main argument for applying NMF to biological data such as gene expression is its inherent constraint of non-negativity, which better reflects living systems. On the other hand, ICA has the advantage of recovering statistically independent signals from data, which is often a direct assumption for source separation. However, in the field of cancer biology, while some argue about the effectiveness of one method above the other, others claim similar performances (Kim et al., 2011). Most still agree that their efficiency is highly context specific when it comes to blind source separation problems (Mirzal, 2017).

6.2 The good and the bad of unsupervised deconvolution methods

In Section 1.3, I have introduced the problem of quantifying and qualifying the TME content and state. While a majority of methods to solve the deconvolution problems rely on supervised approaches, unsupervised methods still present many advantages.

While supervised methods are more precise than unsupervised ones, they still have a limited focus by concentrating only on the task they are made for. This eliminates the possibility to make new discoveries about the TME unless new knowledge is constantly added onto the method. Unsupervised approaches do not suffer from this problem and are suitable approaches to uncover new elements of the tumoral microcosm. In Chapter 4, I showed that through the use of HACK, I was able to find components related to different immune functional states. This type of information can give us more insight on the possible states of the TME than a simple estimation of cell-type proportions.

One other implication of unsupervised approaches is their (almost) total agnostic state when analysing data. As Gregor Sturm et al. (Sturm et al., 2019) compared supervised deconvolution methods, they realised that they all in a sort “found what they were looking for” even when there was nothing to be found. Unsupervised methods on the other hand simply cannot be susceptible to this behaviour since they are not told anything about the type of information expected. Nevertheless, to pay for that price, unsupervised approaches may fall blind and in the end miss existing information. Some teams, such as Shi et al. (Shi and Zhang, 2011) try to overcome these issues by using semi-supervised methods that help guide unsupervised learning in the right direction.

However, when talking about exploratory analyses, we are always affected by the issue of data interpretation. Since unsupervised methods are given freedom to explore and present the data as they see fit, users are often left with the task to label the results to make sense of the information extracted. Because of that, no unsupervised analysis is truly so since an expert knowledge input is required at some step of the study. Unfortunately, with the lack of golden standards and references, this interpretation step remains one of the most difficult problems of unsupervised methods.

One last point about unsupervised analyses is at an unknown state and could be considered as a possible strength or a weakness if properly analysed. I am speaking here about the unknown limits of signal extraction, also known as granularity, as stated in Section 3.3. Some authors claim to be able to distinguish dozens of cell-types in the TME (Aran et al., 2017) but much debate still exists in the definition of certain cell-types. This non agreement comes from the fact that cell-types are described through many different characteristics such as molecular data, morphology, phenotype and function and a clear definition for each distinct cell type is still lacking. Since I mentioned above the trouble of interpreting components resulting from an unsupervised analysis, it is this incertitude that creates an unknown characteristic of unsupervised deconvolution. I tried to alleviate this problem in Chapter 4 with the HACK method by offering a possibility to visualise the limits of signal extraction. This capacity of being able to follow how the method decompose the data with each step should allow to study further the limits of various decomposition methods and get an idea of the true amount of information contained in the data

6.3 Validity of the HACK method

As the main focus of my PhD work, the development of the HACK method was supposed to get rid of certain limitations of unsupervised deconvolution while still maintaining the advantages such approaches procured. It was therefore important to justify the hierarchical approach and the use of its new possibilities to correct for previously missed errors to create what I called “persistent components”.

Because ICA is the preferred matrix factorisation method used in my team, I had at my disposition a panel of tests to compare with. Thanks to that, I was able to confirm the robustness and reproducibility of produced persistent components by performing pan-cancer analyses. I was also able to show how new additions to the processing of independent components improved their overall quality. Unfortunately, I was unable to justify these improvements in relation to possible improvement for biological interpretation. Despite my efforts, I was unable to find a good measure to use as a score of the biological relevance of a component. In the end, only statistical scores helped me confirm the advantages of the method I developed.

7. Perspectives

A researcher's work is a never ending journey and it is also the case for a PhD project. Although results were obtained and observations were made, there are still many questions that remain to be answered. I will try here to describe some ideas that could support future studies using the work accomplished in this PhD project.

7.1 The future of HACK

The HACK method is young and it would be a shame to let it in its current state without exploring all the opportunities it provides. I have ended Chapter 4 by describing some interesting observations made by "playing around" with the method. Among them, the one that caught my eye was the graph I obtained by trying to decompose the data by forcing the extraction of 200 signals from a bulk CRC RNA-seq dataset. From my researches of the literature, decompositions applied to genetic data rarely go that far and some applications can even prove successful with only a few components as it is the case in (Quintero et al., 2020) where only 8 components were deemed enough to extract features specific of exactly 8 immune related cell-types. From observations made by users of ICA in my team, it was hypothesised that components found after an order of decomposition corresponding the intrinsic dimensionalities of the data weren't robust enough to be accepted for further analyses. However, by inspecting the graph obtained with HACK and looking at components of high order of decomposition, I realised that this initial hypothesis might be wrong. In fact, many components found at high orders presented a good number of stable genes that showed significant enrichments for specific functions. The next step of such analysis would be to push various decomposition methods towards their ultimate limit and inspect what level of decomposition is reachable and what precision of signals we can hope to extract.

Another interesting point of the method is its capacity to project decomposition events and component relations in a graphical form that could be assimilated to a network. For now, graphs of this sort were only created from individual datasets. It would be of great interest to use this graphical representation to integrate different decompositions together and see how they can relate to each other and if additional information can be reconstructed. An example of such an integration can be seen in Figure 7.1. In this case, we take two persistent components that have a common origin function and merge them together. By doing so, we are able to enhance our comprehension of possible splits found by the method. And by extent, since components represent particular biological signals and functions, we would be able to discover new links between functions.

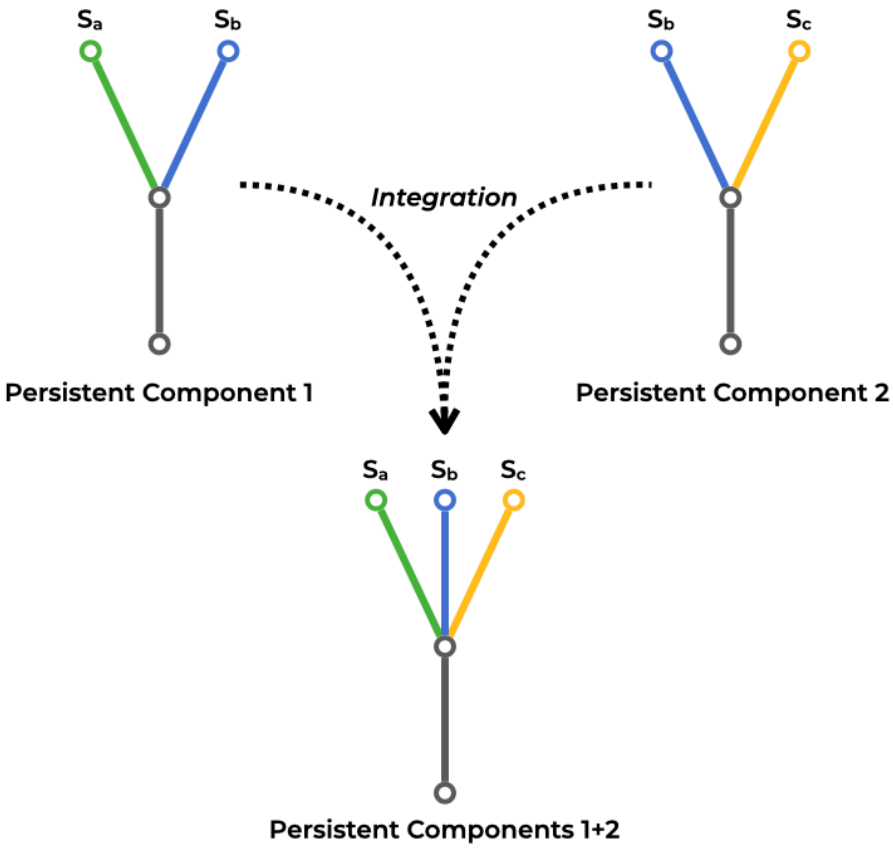


Figure 7.1. Illustration of the possible integration of a set of persistent component graphs.
From two different sets of persistent components obtained on different datasets and their corresponding graphical representation, it may be possible to integrate them together into a composite persistent component. In this example, the integration corresponds to finding the common component between the two (in this case the grey portion) and using this as a scaffold for merging the eventual emergent components. In this case, the integration leads to a more detailed visualisation of the possible splits of an origin signal (in gray) into 3 different signals S_a (in green), S_b (in blue) and S_c (in yellow).

When developing this method, I tried to answer the question raised in Section 3.1 by trying to take advantage of the hierarchical approach to reconstruct the functional links between unsupervised components. This was achieved in part as I have demonstrated it with components related to the immune infiltration but other relations were still scarce. However, the parameters for the case studies have been chosen to balance the amount of information displayed with the difficulty to visualise it. A more detailed representation could be obtained by relaxing some parameters and maybe by sacrificing visual interpretability, more functional links could be retrieved.

7.2 Comprehensive molecular maps and their applications for interpreting the data analysis results

In this thesis, I have reviewed two types of knowledge maps, namely signalling and metabolic pathways networks. I have also proposed the possibility to integrate such networks together through common players such as genes. At the current time, such integrations aren't common and most maps remain independent. But with the rising popularity of disease related maps and the joint effort of researcher communities such as the Disease Maps Project (Mazein et al., 2018), integrations of diverse resources are expected to increase in the following years. The integration of metabolism and signaling pathways is already a reality with the help of ACSN and RECON and since these resources are continually updated and upgraded with the addition of new maps, our understanding becomes more and more systemic with each step.

Integrations of this sort also open entirely new possibilities with the usage of techniques restricted to specific types of omics that can now be applied out of their initial context. If we take for example the method of Flux Balance Analysis (Heirendt et al., 2019) that I described for metabolic networks, if target metabolic pathways detected through such method are connected to other maps, it becomes possible to export these observations onto another levels of information and impute novel hypothesis about existing interactions and perturbation implications.

7.3 New omics and their interests

My applications of deconvolution stayed limited to gene expression data but this type of analysis could be extended to other omics to broaden the vision of perturbations related to cancer and their impact on other functions. Many methods using matrix factorisation as a basis have been proposed to deal with multi-omics integration (Cantini et al., 2021) but little is known on how to deal with the inherent problem of integration. Indeed, when trying to group different types of data, each one of them presents limits for which resolution may not overlap, thus requiring an extensive knowledge on how to deal with them so that it wouldn't affect other types. Indeed, each dataset and each type of data contain their particular noise. Usually, this is taken into account into the protocol of their corresponding analysis. But we can easily imagine the underlying difficulty of compensating for every possible defect of each data type without these modifications cascading into other types.

I mentioned at several occasions the lack of existing golden standards that could be used to either guide methods or to ensure their correctness. Hope may be brought by the development and improvement of single-cell technologies. Since deconvolution relies on the existence of specific signatures, by isolating specific cell-types, we can hope to extract unique signals that could be used as references for future studies. Although in the current state, single-cell technology is costly and presents some challenges related to data generation and analysis, progress continues to be made and computation methods are keeping up day by day. And it is through such efforts that we can hope to reach an immune cell census which will profit the progress of cancer therapies (Stubbington et al., 2017).

8. Conclusions

My conclusions will be based on the opinion I have formed during my PhD work.

During these 3 years, I have focused on the analysis of transcriptomic data using unsupervised methods with a main focus on ICA. I have tried to stay informed on other methods that could answer my problem and tried to compare them. I will not say that the method I ended up choosing and working on is the best but I still have a strong belief that unsupervised approaches are very promising in the field of cancer research.

While supervised methods can excel at predicting certain outcomes based on expert knowledge, the only way to progress our knowledge is through unsupervised procedures. Research is all about expanding our knowledge and to do so, we have to explore and go beyond predictable behaviours to bring new and unexpected results. And with the emergence of new types of data and the incredible quantity of it, it becomes even more crucial to continue in this direction.

However, going in blindly isn't a good idea either so there is a need to set some rules and know the limits. With the method I developed during my PhD, I hoped to accomplish something which might give us a new way of looking at unsupervised deconvolution. I tried to shape mathematical results to be visualisable and understandable by non-experts and make this black box that is unsupervised deconvolution more transparent. With this, I also wanted to make more accessible the studies of deconvolution methods' limits, with an interest in how deep they can go and what level of granularity of signal extraction can be achieved. I tried to go beyond suspected limits and when I did so I realised that I didn't probably even reach them yet.

It is with the hope that my work has shed some light on unsupervised deconvolution that I end my thesis. And it is with the desire to achieve more that I will continue my work as a researcher in the future.

9. Résumé de la thèse en Français

Les tumeurs solides sont caractérisées par une organisation complexe de l'écosystème cellulaire, dans lequel les cellules tumorales résident et évoluent. Cet écosystème est appelé le Micro-Environnement Tumoral (TME). Il est constitué de nombreux types de cellules qui interagissent et échangent des informations de manière constante. Les constituants majeurs qui jouent un rôle important dans l'établissement et la progression d'une tumeur sont d'ordre immunitaire. Cependant, bien que leur fonctions immunitaires permettent dans la majorité des cas de luter contre les cellules cancéreuses, il arrive parfois que les tumeurs s'adaptent et deviennent résistantes et dans certains cas les tumeurs peuvent même reprogrammer le système immunitaire en place pour le tourner à leur avantage et ainsi favoriser le développement du cancer.
Le TME est donc la cible privilégiée de l'immunothérapie qui vise a impacter fortement la croissance de la tumeur ou son potentiel invasif et métastatique. Les traitement d'immunothérapie se focalisent sur un maintient ou une amélioration du système immunitaire qui rencontre souvent des difficultés à repérer et détruire les tumeurs. C'est pourquoi la caractérisation de l'état et du contenu du TME d'un patient atteint du cancer est une priorité.

Mon travail doctoral s'insert dans le cadre du projet Européen IMMUCan. Ce projet rassemble des dizaines de partenaires pour tenter de répondre aux questions liées à l'immunothérapie. En effet, même si les traitements immunothérapeutiques démontrent une efficacité pour un grand nombre de cancers, nous observons encore beaucoup de situations où le traitement reste inefficace ou entraine des effets secondaires graves chez certains patients. Dans le but d'éclaircir ces phénomènes, IMMUCan propose de recruter plus de 3.000 patients atteint de divers types de cancers et de réaliser un suivi de ces patients qui seront traités ou non avec l'immunothérapie. Un grand nombre de données sera collecté avec des technologies de pointes telle le séquençage Single-Cell ou la cytométrie de flux.
Ma responsabilité dans ce projet est liée à une tâche particulière qui est de réaliser la déconvolution des tumeurs et de leur micro-environnement afin d'en extraire les signaux spécifiques à chaque type cellulaire présent chez les patients. Cependant, dû à la forte variabilité du TME et à sa complexité moléculaire, il est difficile de choisir des cibles précises par avance pour une telle analyse. C'est pour cette raison qu'il devient raisonnable d'appliquer des approches non-supervisées qui ne requièrent pas de connaissances à priori.

Pour atteindre cet objectif, c'est au travers de l'utilisation d'une méthode de factorisation matricielle appelée l'Analyse par Composante Indépendantes (ICA) que nous pouvons commencer à disséquer les données d'expression génique et extraire les signaux liés à l'infiltration immunitaire. ICA est la méthode de prédilection employée dans mon équipe et lorsqu'appliquée aux données d'expression, elle a pour but d'extraire des signaux biologiques indépendants entre eux sous la forme de vecteur de poids associés à chaque gène. Ces vecteurs, aussi appelés composantes, obtenus par cette méthode de factorisation matricielle sont appelés "metagènes" et peuvent ensuite être interprétés par des méthodes d'enrichissement pour déterminer à quelles fonctions biologiques ils sont assimilés. En effet, l'ICA est une méthode capable d'extraire de données tumorales d'expression de gènes sous forme de signaux biologiques correspondants à des fonctions biologiques. La capacité de cette méthode de pouvoir extraire des signaux biologiques sans avoir besoin connaissances

a priori permet son application sur un large panel de données et de cancers, augmentant ainsi la portée de découvertes possibles.

Pourtant, même si cette méthode a fait ses preuves et a démontré son efficacité pour les taches de déconvolution, dû à son aspect non-supervisé, elle comporte tout de même quelques complications lorsqu'il s'agit de choisir le nombre de signaux attendus dans les données ou bien lors de l'étape d'interprétation de ces signaux. L'un des paramètres clef obligatoire de l'ICA est l'indication du nombre de signaux que l'on veut extraire des données. Cependant, cette information n'est jamais connue à l'avance et doit donc être estimée au mieux de manière mathématique.

Pour compenser ce problème du choix d'une dimension spécifique pour la décomposition des données, j'ai développé lors de ma thèse une nouvelle méthode pour permettre de projeter les signaux sur un large champ de dimensions en retraçant leur évolution et comportement le long d'une l'analyse hiérarchique. Cette approche permet aussi d'avoir une idée sur la qualité des signaux récupérés tout en aidant à reconstruire les relations entre certains de ces signaux.

J'ai nommé cette méthode "Hierarchical Analysis of Component linKs" (HACK) et son principe peut se résumer de la manière suivante:

- Une méthode de décomposition des données par factorisation matricielle doit être choisie par l'utilisateur. Parmi les méthodes possibles, la PCA, ICA et NMF ont été testées avec succès.
- Les données sont ensuite décomposées en nombre croissant de composantes.
- Les composantes obtenues sont arrangées le long d'un axe de manière hiérarchique, puis des liens entre ces composantes sont calculés et ajoutés en utilisant un score de similarité telle la corrélation de Pearson par exemple. Ceci permet d'obtenir une structure organisée sous forme d'un graphe, qui donne lieu à la possibilité de suivre l'état des composantes individuelles au cours des décompositions croissantes ainsi que retracer leurs "liens de parentés".
- Les liens de ce graphe sont filtrés pour ne garder que les liens les plus forts en enlevant le bruit de fond. De cette façon, le graphe devient plus facilement interprétable visuellement et seules les vraies relations de parenté éventuelle sont conservées.
- Pour s'assurer que seuls les signaux stables sont présents, nous devons procéder à l'élimination des composantes qui sont retrouvées de manière consécutive lors des multiples décompositions. Cette étape est cruciale car elle aide à structurer le graphe tout en garantissant que les composantes conservées sont robustes et correspondent à de vrais signaux biologiques et non du bruit. Ces composantes stables seront appelées des composantes persistantes.
- Une fois les vérifications de stabilité terminées, le résultat de la méthode peut être visualisé sous forme d'un graphe interactif où les utilisateurs peuvent repérer les signaux biologiques d'intérêt en soumettant une liste de gènes correspondants. La méthode génère aussi une matrice contenant les composantes persistantes avec pour chaque composante la liste de gènes avec leurs poids associés, ainsi que leur score de stabilité.

La méthode HACK a été appliquée à 12 jeux de données RNA-seq de cancer colorectal pour estimer sa robustesse et reproductibilité. Les résultats de HACK ont été comparés aux résultats utilisant les approches standard de l'ICA. L'utilisation de composantes persistantes par approche hiérarchique a été démontré comme plus reproductible parmi les différents jeux de données comparés à l'utilisation de composantes d'ordre fixe.
Une étude supplémentaire a été réalisée sur des données single-cell de RNA-seq sur 125 lignées cellulaires représentants 22 types de cancers. Les fonctions biologiques extraites par les composantes de HACK ont non seulement retrouvé les fonctions observées à l'aide d'autres méthodes de déconvolution mais ont aussi distingué des fonctions plus spécifiques non observées auparavant, qui avaient notamment attrait au cycle cellulaire.

Une analyse plus approfondie des signaux liés à l'infiltration immunitaire a aussi été réalisée. L'observation des graphes obtenus à l'aide de la méthodologie HACK a démontré une capacité de séparation de signaux plus détaillée qu'une méthodologie utilisant l'ICA avec une dimension unique fixée. Les composantes liées à l'immunité ont été extraites et leur fonctions biologiques étudiées plus en détail. Nous avons pu observer que les composantes ainsi que leur structure dans le graphe permettaient de repérer des fonctions liées à l'état de l'activité immunitaire. Il était de ce fait possible de séparer les fonctions de immunité cellulaire et de l'immunité humorale. Cette capacité de détection pourrait donc s'avérer extrêmement utile pour établir l'état du système immunitaire d'un patient atteint de cancer pour pouvoir améliorer la recherche du traitement le plus approprié.

Cependant, l'une des limites de ce type d'analyse reste l'interprétation correcte des signaux détectés. Pour améliorer cette étape, l'une des possibilités est d'utiliser des reconstructions approfondies de voies de signalisation moléculaires pour tirer des conclusions sur leur sens biologique mais aussi pour récupérer des information supplémentaires sur un niveau plus systémique. Lors de mon travail de thèse, j'ai participé à l'élaboration d'une carte détaillée des voies de signalisation spécifiques au cancer du système immunitaire. Cette carte contient les réseaux d'interactions moléculaire de multiples types de cellulaire immunitaire telles que les macrophages, les cellulaires myéloïdes, les DC ainsi que les NK.
Mon rôle a été d'intégrer ces cartes dans la plateforme de NaviCell qui permet d'utiliser ces réseaux comme des outils de visualisation de données en donnant un contexte clair et détaillé. Comme démonstration de ces capacités, une décomposition de données de métastases de mélanome a été réalisée à l'aide de l'ICA. Ces données étant annotées avec la survie des patients, il était possible de comparer nos observations en reliant les résultats à un pronostique de survie. Une projection des données d'expression des patients a en effet pu démontrer une capacité de prédiction de survie des patients. Ces prédictions étaient accompagnées de groupes de gènes d'intérêt qui corrélaient positivement ou négativement avec le taux de survie. De plus, en sélectionnant les composantes de l'ICA spécifiques des cellules NK et des macrophages, nous avons pu clairement différencier des phénotypes de ces cellules en rapport avec leur activités immunitaire pro ou anti tumorales. Grâce à cette analyse, il a été possible de repérer les voies de signalisations liées au recrutement et l'activation des cellules NK ainsi que celles représentant des cellules réprimées. L'activité des macrophages quand à elle a pu être représentée sous deux formes: pro-tumorale de part ses voies de cytokines immuno-réprimantes et anti-tumorale grâce aux voies spécifiques à la présentation aux antigènes ainsi que d'une immuno-stimulation de miRNA et de facteurs de transcriptions.

Lors de ma thèse, mon travail s'est limité à l'utilisation de données d'expression mais il est tout à fait possible d'étendre la méthode développée à d'autres types de données. En ajoutant de nouveaux types de données, cela nous permettrait d'obtenir de nouvelles informations qui viendraient compléter les hypothèse générées.
Qui plus est, le développement de la méthode HACK a donné l'occasion de pouvoir directement observer les capacités de déconvolution des méthodes non-supervisées. Cela ouvre donc la possibilité de réaliser des analyses plus poussées tout en suivant les limites des méthodes utilisées. En réalité, il est risqué de pousser les analyses non-supervisée trop loin de peur de dégrader les signaux d'intérêt ou bien de forcer la méthode à générer des signaux artificiels qui risquent d'être confondus avec de vrais signaux biologiques.

En définitive, c'est au travers de l'utilisation d'analyses non-supervisées, couplées à une description détaillée des interactions moléculaires, que nous pouvons élucider la complexité du micro-environnement tumoral. Les approches non supervisées d'analyse de données connaissant un essor dans le monde de la recherche sur le cancer et les possibilité exploratoires qu'elles offrent font d'elles des acteurs majeurs de découvertes de nouveaux traitements contre le cancer.

Bibliographie

Akaike, H. (1998). A New Look at the Statistical Model Identification. In Selected Papers of Hirotugu Akaike, E. Parzen, K. Tanabe, and G. Kitagawa, eds. (New York, NY: Springer), pp. 215–222.

Albertson, D.G. (2006). Gene amplification in cancer. Trends Genet. TIG *22*, 447–455.

Alvarez-Dominguez, J.R., Bai, Z., Xu, D., Yuan, B., Lo, K.A., Yoon, M.J., Lim, Y.C., Knoll, M., Slavov, N., Chen, S., et al. (2015). De Novo Reconstruction of Adipose Tissue Transcriptomes Reveals Long Non-coding RNA Regulators of Brown Adipocyte Development. Cell Metab. *21*, 764–776.

de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Computational Oncology in the Multi-Omics Era: State of the Art. Front. Oncol. *10*, 423.

Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. *18*, 220.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124.

Aslam, B., Basit, M., Nisar, M.A., Khurshid, M., and Rasool, M.H. (2017). Proteomics: Technologies and Their Applications. J. Chromatogr. Sci. *55*, 182–196.

Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat. Commun. *11*, 5650.

Azeloglu, E.U., and Iyengar, R. (2015). Signaling networks: information flow, computation, and decision making. Cold Spring Harb. Perspect. Biol. *7*, a005934.

Bac, J., Mirkes, E.M., Gorban, A.N., Tyukin, I., and Zinovyev, A. (2021). Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation. Entropy *23*, 1368.

Bach, F.R., and Jordan, M.I. (2003). Beyond independent components: trees and clusters. J. Mach. Learn. Res. *4*, 1205–1233.

Balkwill, F.R., Capasso, M., and Hagemann, T. (2012). The tumor microenvironment at a glance. J. Cell Sci. *125*, 5591–5596.

Barillot, E., Calzone, L., Hupe, P., Vert, J.-P., and Zinovyev, A. (2020). Computational Systems Biology of Cancer (S.l.: CRC Press).

Baxter, J.S., Leavy, O.C., Dryden, N.H., Maguire, S., Johnson, N., Fedele, V., Simigdala, N., Martin, L.-A., Andrews, S., Wingett, S.W., et al. (2018). Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nat. Commun. *9*, 1028.

Bee, M.A., and Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? J. Comp. Psychol. Wash. DC 1983 *122*, 235–251.

Bell, A.J., and Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Comput. *7*, 1129–1159.

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D.W., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., et al. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Benetos, E., Kotti, M., and Kotropoulos, C. (2006). Applying Supervised Classifiers Based on Non-negative Matrix Factorization to Musical Instrument Classification. In 2006 IEEE International Conference on Multimedia and Expo, pp. 2105–2108.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 6–17.

van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. J. Vis. Exp. JoVE 1869.

Bernstein, B.E., Meissner, A., and Lander, E.S. (2007). The mammalian epigenome. Cell *128*, 669–681.

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., et al. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. *9*, 1235–1245.

Bolen, C.R., Uduman, M., and Kleinstein, S.H. (2011). Cell subset prediction for blood genomic studies. BMC Bioinformatics *12*, 258.

Bonnet, E., Viara, E., Kuperstein, I., Calzone, L., Cohen, D.P.A., Barillot, E., and Zinovyev, A. (2015). NaviCell Web Service for network-based data visualization. Nucleic Acids Res. *43*, W560-565.

Bos, J.L. (1989). ras oncogenes in human cancer: a review. Cancer Res. *49*, 4682–4689.

Breasted, J.H. (1991). The Edwin Smith Surgical Papyrus, Volume 1: Hieroglyphic Transliteration, Translation, and Commentary | The Oriental Institute of the University of Chicago (Oriental Institute Publications).

Bromberg, J.F., Horvath, C.M., Wen, Z., Schreiber, R.D., and Darnell, J.E. (1996). Transcriptionally active Stat1 is required for the antiproliferative effects of both interferon alpha and interferon gamma. Proc. Natl. Acad. Sci. U. S. A. *93*, 7673–7678.

Bronkhorst, A.W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten. Percept. Psychophys. *77*, 1465–1487.

Brudno, J.N., and Kochenderfer, J.N. (2016). Toxicities of chimeric antigen receptor T cells: recognition and management. Blood *127*, 3321–3330.

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

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. *36*, 272–281.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1-21.29.9.

Cairns, B.R. (2007). Chromatin remodeling: insights and intrigue from single-molecule studies. Nat. Struct. Mol. Biol. *14*, 989–996.

Campbell, D.J., and Koch, M.A. (2011). Treg cells: patrolling a dangerous neighborhood. Nat. Med. *17*, 929–930.

Cantini, L., Kairov, U., de Reyniès, A., Barillot, E., Radvanyi, F., and Zinovyev, A. (2019). Assessing reproducibility of matrix factorization methods in independent transcriptomes. Bioinforma. Oxf. Engl. *35*, 4307–4313.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat. Commun. *12*, 124.

Cess, C.G., and Finley, S.D. (2020). Multi-scale modeling of macrophage-T cell interactions within the tumor microenvironment. PLoS Comput. Biol. *16*, e1008519.

Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305-311.

Cherry, E.C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. J. Acoust. Soc. Am. *25*, 975–979.

Chow, C., and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory *14*, 462–467.

Costanza, R., Wainger, L., Folke, C., and Mäler, K.-G. (1993). Modeling complex ecological economic systems: toward an evolutionary, dynamic understanding of people and nature. Inst. Sustain. Solut. Publ. Present.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Croce, C.M. (2008). Oncogenes and cancer. N. Engl. J. Med. *358*, 502–511.

Czerwińska, U. (2018). Unsupervised deconvolution of bulk omics profiles : methodology and application to characterize the immune landscape in tumors. PhD Thesis.

Davis-Marcisak, E.F., Fitzgerald, A.A., Kessler, M.D., Danilova, L., Jaffee, E.M., Zaidi, N., Weiner, L.M., and Fertig, E.J. (2021). Transfer learning between preclinical models and human tumors identifies a conserved NK cell activation signature in anti-CTLA-4 responsive tumors. Genome Med. *13*, 129.

Decamps, C., Arnaud, A., Petitprez, F., Ayadi, M., Baurès, A., Armenoult, L., HADACA consortium, Escalera, S., Guyon, I., Nicolle, R., et al. (2021). DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. BMC Bioinformatics *22*, 473.

Dimitrakopoulos, C., Hindupur, S.K., Häfliger, L., Behr, J., Montazeri, H., Hall, M.N., and Beerenwinkel, N. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. Bioinforma. Oxf. Engl. *34*, 2441–2448.

Domon, B., and Aebersold, R. (2006). Mass spectrometry and protein analysis. Science *312*, 212–217.

Du, W., and Elemento, O. (2015). Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene *34*, 3215–3225.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. U. S. A. *104*, 1777–1782.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M. (1999). Expression profiling using cDNA microarrays. Nat. Genet. *21*, 10–14.

Dunn, G.P., Bruce, A.T., Ikeda, H., Old, L.J., and Schreiber, R.D. (2002). Cancer immunoediting: from immunosurveillance to tumor escape. Nat. Immunol. *3*, 991–998.

El-Kenawi, A.E., and El-Remessy, A.B. (2013). Angiogenesis inhibitors in cancer therapy: mechanistic perspective on classification and treatment rationales. Br. J. Pharmacol. *170*, 712–729.

El-Manzalawy, Y. (2018). CCA based multi-view feature selection for multi-omics data integration. In 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–8.

Erez, N., Truitt, M., Olson, P., Arron, S.T., and Hanahan, D. (2010). Cancer-Associated Fibroblasts Are Activated in Incipient Neoplasia to Orchestrate Tumor-Promoting Inflammation in an NF-kappaB-Dependent Manner. Cancer Cell *17*, 135–147.

Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., and Lähdesmäki, H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinforma. Oxf. Engl. *26*, 2571–2577.

Faguet, G.B. (2015). A brief history of cancer: age-old milestones underlying our current knowledge database. Int. J. Cancer *136*, 2022–2036.

Fan, S., and Chi, W. (2016). Methods for genome-wide DNA methylation analysis in human cancer. Brief. Funct. Genomics *15*, 432–442.

Fan, J., Slowikowski, K., and Zhang, F. (2020). Single-cell transcriptomics in cancer: computational challenges and opportunities. Exp. Mol. Med. *52*, 1452–1465.

Fan, Y., Wang, W., Ma, G., Liang, L., Shi, Q., and Tao, S. (2007). Patterns of insertion and deletion in Mammalian genomes. Curr. Genomics *8*, 370–378.

Farber, C.R., and Mesner, L.D. (2016). Chapter 3 - A Systems-Level Understanding of Cardiovascular Disease through Networks. In Translational Cardiometabolic Genomic Medicine, A. Rodriguez-Oquendo, ed. (Boston: Academic Press), pp. 59–81.

Fife, B.T., and Bluestone, J.A. (2008). Control of peripheral T-cell tolerance and autoimmunity via the CTLA-4 and PD-1 pathways. Immunol. Rev. *224*, 166–182.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. Mol. Syst. Biol. *7*, 501.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805-811.

Fu, C., and Jiang, A. (2018). Dendritic Cells and CD8 T Cell Immunity in Tumor Microenvironment. Front. Immunol. *9*, 3059.

Gatto, F., Ferreira, R., and Nielsen, J. (2020). Pan-cancer analysis of the metabolic reaction network. Metab. Eng. *57*, 51–62.

Gauthier, J., Vincent, A.T., Charette, S.J., and Derome, N. (2019). A brief history of bioinformatics. Brief. Bioinform. *20*, 1981–1996.

Gawron, P., Ostaszewski, M., Satagopam, V., Gebel, S., Mazein, A., Kuzma, M., Zorzan, S., McGee, F., Otjacques, B., Balling, R., et al. (2016). MINERVA-a platform for visualization and curation of molecular interaction networks. NPJ Syst. Biol. Appl. *2*, 16020.

Gligorijević, V., and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. J. R. Soc. Interface *12*, 20150571.

Gorban, A.N., Kégl, B., Wunsch, D.C., and Zinovyev, A. (2008). Principal Manifolds for Data Visualization and Dimension Reduction (Berlin Heidelberg: Springer-Verlag).

Guo, C., Manjili, M.H., Subjeck, J.R., Sarkar, D., Fisher, P.B., and Wang, X.-Y. (2013). Therapeutic cancer vaccines: past, present, and future. Adv. Cancer Res. *119*, 421–475.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19*, 1720–1730.

Hajdu, S.I. (2011). A note from history: landmarks in history of cancer, part 1. Cancer *117*, 1097–1102.

Hanahan, D., and Folkman, J. (1996). Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. Cell *86*, 353–364.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol. *18*, 83.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. *14*, 639–702.

Herault, J., and Jutten, C. (1986). Space or time adaptive signal processing by neural network models. AIP Conf. Proc. *151*, 206–211.

Himberg, J., and Hyvarinen, A. (2003). Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. In 2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), pp. 259–268.

Hollywood, K., Brison, D.R., and Goodacre, R. (2006). Metabolomics: current technologies and future trends. Proteomics *6*, 4716–4723.

Hotelling, H. (1936). Relations Between Two Sets of Variates. Biometrika *28*, 321–377.

Hu, P., Zhang, W., Xin, H., and Deng, G. (2016). Single Cell Isolation and Analysis. Front. Cell Dev. Biol. *4*, 116.

Hui, L., and Chen, Y. (2015). Tumor microenvironment: Sanctuary of the devil. Cancer Lett. *368*, 7–13.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Netw. Off. J. Int. Neural Netw. Soc. *13*, 411–430.

Isomura, T., and Toyoizumi, T. (2016). A Local Learning Rule for Independent Component Analysis. Sci. Rep. *6*, 28073.

Jeong, E., Moon, S.U., Song, M., and Yoon, S. (2017). Transcriptome modeling and phenotypic assays for cancer precision medicine. Arch. Pharm. Res. *40*, 906–914.

Jerby, L., Wolf, L., Denkert, C., Stein, G.Y., Hilvo, M., Oresic, M., Geiger, T., and Ruppin, E. (2012). Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. Cancer Res. *72*, 5712–5720.

Jeschke, J., Bizet, M., Desmedt, C., Calonne, E., Dedeurwaerder, S., Garaud, S., Koch, A., Larsimont, D., Salgado, R., Van den Eynden, G., et al. (2017). DNA methylation-based immune response signature improves patient diagnosis in multiple cancers. J. Clin. Invest. *127*, 3090–3102.

Jia, R., Chai, P., Zhang, H., and Fan, X. (2017). Novel insights into chromosomal conformations in cancer. Mol. Cancer *16*, 173.

Jiang, Z., Zhou, X., Li, R., Michal, J.J., Zhang, S., Dodson, M.V., Zhang, Z., and Harland, R.M. (2015). Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. Cell. Mol. Life Sci. CMLS *72*, 3425–3439.

Johnson, C.H., and Gonzalez, F.J. (2012). Challenges and opportunities of metabolomics. J. Cell. Physiol. *227*, 2975–2981.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Jolliffe, I.T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philos. Transact. A Math. Phys. Eng. Sci. *374*, 20150202.

Juliano, R.L. (2020). Addressing cancer signal transduction pathways with antisense and siRNA oligonucleotides. NAR Cancer *2*, zcaa025.

Kaikkonen, M.U., and Adelman, K. (2018). Emerging Roles of Non-Coding RNA Transcription. Trends Biochem. Sci. *43*, 654–667.

Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., and Zinovyev, A. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics *18*, 712.

Kim, J.K., and Choi, S. (2006). Tree-Dependent Components of Gene Expression Data for Clustering. In Artificial Neural Networks – ICANN 2006, S. Kollias, A. Stafylopatis, W. Duch, and E. Oja, eds. (Berlin, Heidelberg: Springer), pp. 837–846.

Kim, M.H., Seo, H.J., Joung, J.-G., and Kim, J.H. (2011). Comprehensive evaluation of matrix factorization methods for the analysis of DNA microarray gene expression data. BMC Bioinformatics *12 Suppl 13*, S8.

Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., and Wild, D.L. (2012). Bayesian correlated clustering to integrate multiple datasets. Bioinforma. Oxf. Engl. *28*, 3290–3297.

Klug, A. (2004). The discovery of the DNA double helix. J. Mol. Biol. *335*, 3–26.

Knox, S.S. (2010). From “omics” to complex disease: a systems biology approach to gene-environment interactions in cancer. Cancer Cell Int. *10*, 11.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013). The next-generation sequencing revolution and its impact on genomics. Cell *155*, 27–38.

Koch, I., Weil, R., Wolbold, R., Brockmöller, J., Hustert, E., Burk, O., Nuessler, A., Neuhaus, P., Eichelbaum, M., Zanger, U., et al. (2002). Interindividual variability and tissue-specificity in the expression of cytochrome P450 3A mRNA. Drug Metab. Dispos. Biol. Fate Chem. *30*, 1108–1114.

Koh, H.W.L., Fermin, D., Vogel, C., Choi, K.P., Ewing, R.M., and Choi, H. (2019). iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. NPJ Syst. Biol. Appl. *5*, 22.

Kondratova, M., Sompairac, N., Barillot, E., Zinovyev, A., and Kuperstein, I. (2018). Signalling maps in cancer research: construction and data analysis. Database J. Biol. Databases Curation *2018*.

Kondratova, M., Czerwinska, U., Sompairac, N., Amigorena, S.D., Soumelis, V., Barillot, E., Zinovyev, A., and Kuperstein, I. (2019). A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures. Nat. Commun. *10*, 4808.

Kondratova, M., Barillot, E., Zinovyev, A., and Calzone, L. (2020). Modelling of Immune Checkpoint Network Explains Synergistic Effects of Combined Immune Checkpoint Inhibitor Therapy and the Impact of Cytokines in Patient Response. Cancers *12*, E3600.

Krisher, R.L., and Prather, R.S. (2012). A role for the Warburg effect in preimplantation embryo development: metabolic modification to support rapid cell proliferation. Mol. Reprod. Dev. *79*, 311–320.

Kuperstein, I., Bonnet, E., Nguyen, H.-A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., et al. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis *4*, e160.

Kuruoglu, E.E. (2010). Dependent Component Analysis for Cosmology: A Case Study. In Latent Variable Analysis and Signal Separation, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, eds. (Berlin, Heidelberg: Springer), pp. 538–545.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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

Lehmann-Che, J., Poirot, B., Boyer, J.-C., and Evrard, A. (2017). Cancer genomics guide clinical practice in personalized medicine. Therapie *72*, 439–451.

Li, X., Jolly, M.K., George, J.T., Pienta, K.J., and Levine, H. (2019). Computational Modeling of the Crosstalk Between Macrophage Polarization and Tumor Cell Plasticity in the Tumor Microenvironment. Front. Oncol. *9*, 10.

Lin, D., Zhang, J., Li, J., Calhoun, V.D., Deng, H.-W., and Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. BMC Bioinformatics *14*, 245.

Lindsey, J.C., Anderton, J.A., Lusher, M.E., and Clifford, S.C. (2005). Epigenetic events in medulloblastoma development. Neurosurg. Focus *19*, E10.

Lock, E.F., and Dunson, D.B. (2013). Bayesian consensus clustering. Bioinforma. Oxf. Engl. *29*, 2610–2616.

Lock, E.F., Hoadley, K.A., Marron, J.S., and Nobel, A.B. (2013). JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann. Appl. Stat. *7*, 523–542.

Macklin, A., Khan, S., and Kislinger, T. (2020). Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. Clin. Proteomics *17*, 17.

Mahlbacher, G., Curtis, L.T., Lowengrub, J., and Frieboes, H.B. (2018). Mathematical modeling of tumor-associated macrophage interactions with the cancer microenvironment. J. Immunother. Cancer *6*, 10.

Mann, M., and Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. Nat. Biotechnol. *21*, 255–261.

Martignetti, L., Calzone, L., Bonnet, E., Barillot, E., and Zinovyev, A. (2016). ROMA: Representation and Quantification of Module Activity from Target Expression Data. Front. Genet. *7*, 18.

Mazein, A., Ostaszewski, M., Kuperstein, I., Watterson, S., Le Novère, N., Lefaudeux, D., De Meulder, B., Pellet, J., Balaur, I., Saqi, M., et al. (2018). Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. NPJ Syst. Biol. Appl. *4*, 21.

McCarroll, S.A., and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. Nat. Genet. *39*, S37-42.

Meier, R., Ruttkies, C., Treutler, H., and Neumann, S. (2017). Bioinformatics can boost metabolomics research. J. Biotechnol. *261*, 137–141.

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J. Proteome Res. *15*, 755–765.

Menyhárt, O., and Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput. Struct. Biotechnol. J. *19*, 949–960.

Michot, J.M., Bigenwald, C., Champiat, S., Collins, M., Carbonnel, F., Postel-Vinay, S., Berdelou, A., Varga, A., Bahleda, R., Hollebecque, A., et al. (2016). Immune-related adverse events with immune checkpoint blockade: a comprehensive review. Eur. J. Cancer Oxf. Engl. 1990 *54*, 139–148.

Milne, K., Köbel, M., Kalloger, S.E., Barnes, R.O., Gao, D., Gilks, C.B., Watson, P.H., and Nelson, B.H. (2009). Systematic analysis of immune infiltrates in high-grade serous ovarian cancer reveals CD20, FoxP3 and TIA-1 as positive prognostic factors. PloS One *4*, e6412.

Mirzal, A. (2017). NMF versus ICA for blind source separation. Adv. Data Anal. Classif. *11*, 25–48.

Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. U. S. A. *110*, 4245–4250.

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K.S., and Hilsenbeck, S.G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostat. Oxf. Engl. *19*, 71–86.

Monraz Gomez, L.C., Kondratova, M., Sompairac, N., Lonjou, C., Ravel, J.-M., Barillot, E., Zinovyev, A., and Kuperstein, I. (2021). Atlas of Cancer Signaling Network: A Resource of Multi-Scale Biological Maps to Study Disease Mechanisms. In Systems Medicine, O. Wolkenhauer, ed. (Oxford: Academic Press), pp. 490–506.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Murdoch, C., Muthana, M., Coffelt, S.B., and Lewis, C.E. (2008). The role of myeloid cells in the promotion of tumour angiogenesis. Nat. Rev. Cancer *8*, 618–631.

Nair, S.S., and Kumar, R. (2012). Chromatin remodeling in cancer: a gateway to regulate gene transcription. Mol. Oncol. *6*, 611–619.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.

Nielsen, F.A., Balslev, D., and Hansen, L.K. (2005). Mining the posterior cingulate: segregation between memory and pain components. NeuroImage *27*, 520–532.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. *2*, 559–572.

Pelon, F., Bourachot, B., Kieffer, Y., Magagna, I., Mermet-Meillon, F., Bonnet, I., Costa, A., Givel, A.-M., Attieh, Y., Barbazan, J., et al. (2020). Cancer-associated fibroblast heterogeneity in axillary lymph nodes drives metastases in breast cancer through complementary mechanisms. Nat. Commun. *11*, 404.

Qin, Z., and Blankenstein, T. (2000). CD4+ T cell--mediated tumor rejection involves inhibition of angiogenesis that is dependent on IFN gamma receptor expression by nonhematopoietic cells. Immunity *12*, 677–686.

Quail, D.F., and Joyce, J.A. (2013). Microenvironmental regulation of tumor progression and metastasis. Nat. Med. *19*, 1423–1437.

Quintero, A., Hübschmann, D., Kurzawa, N., Steinhauser, S., Rentzsch, P., Krämer, S., Andresen, C., Park, J., Eils, R., Schlesner, M., et al. (2020). ShinyButchR: Interactive NMF-based decomposition workflow of genome-scale datasets. Biol. Methods Protoc. *5*, bpaa022.

Quon, G., Haider, S., Deshwar, A.G., Cui, A., Boutros, P.C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. *5*, 29.

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D.E., and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. ELife *6*, e26476.

Ragoussis, J. (2009). Genotyping technologies for genetic research. Annu. Rev. Genomics Hum. Genet. *10*, 117–133.

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. *46*, 10546–10562.

Ravel, J.-M., Monraz Gomez, L.C., Sompairac, N., Calzone, L., Zhivotovsky, B., Kroemer, G., Barillot, E., Zinovyev, A., and Kuperstein, I. (2020). Comprehensive Map of the Regulated Cell Death Signaling Network: A Powerful Analytical Tool for Studying Diseases. Cancers 12, E990.

Ren, X., Kang, B., and Zhang, Z. (2018). Understanding tumor ecosystems by single-cell sequencing: promises and limitations. Genome Biol. *19*, 211.

Roma-Rodrigues, C., Mendes, R., Baptista, P.V., and Fernandes, A.R. (2019). Targeting Tumor Microenvironment for Cancer Therapy. Int. J. Mol. Sci. *20*, E840.

Rosenberg, S.A., Yannelli, J.R., Yang, J.C., Topalian, S.L., Schwartzentruber, D.J., Weber, J.S., Parkinson, D.R., Seipp, C.A., Einhorn, J.H., and White, D.E. (1994). Treatment of patients with metastatic melanoma with autologous tumor-infiltrating lymphocytes and interleukin 2. J. Natl. Cancer Inst. *86*, 1159–1166.

Rothschild, B.M., Tanke, D.H., Helbling, M., and Martin, L.D. (2003). Epidemiologic study of tumors in dinosaurs. Naturwissenschaften *90*, 495–500.

Rutledge, D.N., and Jouan-Rimbaud Bouveresse, D. (2013). Independent Components Analysis with the JADE algorithm. TrAC Trends Anal. Chem. *50*, 22–32.

Satija, R., and Shalek, A.K. (2014). Heterogeneity in immune responses: from populations to single cells. Trends Immunol. *35*, 219–229.

Savage, R.S., Ghahramani, Z., Griffin, J.E., Kirk, P., and Wild, D.L. (2013). Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. ArXiv13043577 Q-Bio Stat.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays--a technology review. Nat. Cell Biol. *3*, E190-195.

Shankaran, V., Ikeda, H., Bruce, A.T., White, J.M., Swanson, P.E., Old, L.J., and Schreiber, R.D. (2001). IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. Nature *410*, 1107–1111.

Shaulian, E., and Karin, M. (2001). AP-1 in cell proliferation and survival. Oncogene *20*, 2390–2400.

Shen, Q., Hu, J., Jiang, N., Hu, X., Luo, Z., and Zhang, H. (2016). contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples. Bioinforma. Oxf. Engl. *32*, 705–712.

Shi, M., and Zhang, B. (2011). Semi-supervised learning improves gene expression-based prediction of cancer recurrence. Bioinforma. Oxf. Engl. *27*, 3017–3023.

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., and Chen, L. (2017). Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. Bioinforma. Oxf. Engl. *33*, 2706–2714.

Silva, C., Perestrelo, R., Silva, P., Tomás, H., and Câmara, J.S. (2019). Breast Cancer Metabolomics: From Analytical Platforms to Multivariate Data Analysis. A Review. Metabolites *9*, E102.

Smaragdis, P., and Brown, J.C. (2003). Non-negative matrix factorization for polyphonic music transcription. In 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684), pp. 177–180.

Smyth, M.J., Thia, K.Y., Street, S.E., Cretney, E., Trapani, J.A., Taniguchi, M., Kawano, T., Pelikan, S.B., Crowe, N.Y., and Godfrey, D.I. (2000). Differential tumor surveillance by natural killer (NK) and NKT cells. J. Exp. Med. *191*, 661–668.

Sompairac, N., Modamio, J., Barillot, E., Fleming, R.M.T., Zinovyev, A., and Kuperstein, I. (2019). Metabolic and signalling network maps integration: application to cross-talk studies and omics data analysis in cancer. BMC Bioinformatics *20*, 140.

Spaeth, E.L., Dembinski, J.L., Sasser, A.K., Watson, K., Klopp, A., Hall, B., Andreeff, M., and Marini, F. (2009). Mesenchymal stem cell transition to tumor-associated fibroblasts contributes to fibrovascular network expansion and tumor progression. PloS One *4*, e4992.

Speicher, N.K., and Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinforma. Oxf. Engl. *31*, i268-275.

Sra, S., and Dhillon, I.S. (2006). Nonnegative Matrix Approximation: Algorithms and Applications (The University of Texas at Austin).

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Stubbington, M.J.T., Rozenblatt-Rosen, O., Regev, A., and Teichmann, S.A. (2017). Single-cell transcriptomics to explore the immune system in health and disease. Science *358*, 58–63.

Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., List, M., and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinforma. Oxf. Engl. *35*, i436–i445.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956–960.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. *71*, 209–249.

Sutandy, F.X.R., Qian, J., Chen, C.-S., and Zhu, H. (2013). Overview of protein microarrays. Curr. Protoc. Protein Sci. *Chapter 27*, Unit 27.1.

Swartz, M.A., and Lund, A.W. (2012). Lymphatic and interstitial flow in the tumour microenvironment: linking mechanobiology with immunity. Nat. Rev. Cancer *12*, 210–219.

Szymkowiak-Have, A., Girolami, M.A., and Larsen, J. (2006). Clustering via kernel decomposition. IEEE Trans. Neural Netw. *17*, 256–264.

Tachibana, T., Onodera, H., Tsuruyama, T., Mori, A., Nagayama, S., Hiai, H., and Imamura, M. (2005). Increased intratumor Valpha24-positive natural killer T cells: a prognostic factor for primary colorectal carcinomas. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *11*, 7322–7327.

Tammela, T., and Alitalo, K. (2010). Lymphangiogenesis: Molecular mechanisms and future promise. Cell *140*, 460–476.

Teschendorff, A.E., and Zheng, S.C. (2021). EpiDISH: Epigenetic Dissection of Intra-Sample-Heterogeneity (Bioconductor version: Release (3.13)).

Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics *18*, 105.

Theodoridis, S. (2020). Chapter 19 - Dimensionality Reduction and Latent Variable Modeling. In Machine Learning (Second Edition), S. Theodoridis, ed. (Academic Press), pp. 1039–1115.

Thiele, I., Swainston, N., Fleming, R.M.T., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., et al. (2013). A community-driven global reconstruction of human metabolism. Nat. Biotechnol. *31*, 419–425.

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. *18*, 1009–1021.

Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., and Geiger, T. (2016). Proteomic maps of breast cancer subtypes. Nat. Commun. *7*, 10259.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinforma. Oxf. Engl. *26*, i237-245.

Vazquez, A., Kamphorst, J.J., Markert, E.K., Schug, Z.T., Tardito, S., and Gottlieb, E. (2016). Cancer metabolism at a glance. J. Cell Sci. *129*, 3367–3373.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. Biometrika *97*, 893–904.

Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods *11*, 333–337.

Wang, N., Hoffman, E.P., Chen, L., Chen, L., Zhang, Z., Liu, C., Yu, G., Herrington, D.M., Clarke, R., and Wang, Y. (2016). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. Sci. Rep. *6*, 18909.

Wang, W., Baladandayuthapani, V., Morris, J.S., Broom, B.M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinforma. Oxf. Engl. *29*, 149–159.

Way, G.P., Zietz, M., Rubinetti, V., Himmelstein, D.S., and Greene, C.S. (2020). Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol. *21*, 109.

Wegler, C., Ölander, M., Wiśniewski, J.R., Lundquist, P., Zettl, K., Åsberg, A., Hjelmesæth, J., Andersson, T.B., and Artursson, P. (2020). Global variability analysis of mRNA and protein concentrations across and within human tissues. NAR Genomics Bioinforma. *2*, lqz010.

Whiteside, T.L. (2008). The tumor microenvironment and its role in promoting tumor growth. Oncogene *27*, 5904–5912.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell *13*, 1977–2000.

Wishart, D.S. (2015). Is Cancer a Genetic Disease or a Metabolic Disease? EBioMedicine *2*, 478–479.

Xu, R., and Wunsch, D. (2009). Clustering (Wiley).

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, (New York, NY, USA: Association for Computing Machinery), pp. 267–273.

Yakkioui, Y., Temel, Y., Chevet, E., and Negroni, L. (2017). Integrated and Quantitative Proteomics of Human Tumors. Methods Enzymol. *586*, 229–246.

Yanovich, G., Agmon, H., Harel, M., Sonnenblick, A., Peretz, T., and Geiger, T. (2018). Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. Cancer Res. *78*, 6001–6010.

Ye, K., Wang, J., Jayasinghe, R., Lameijer, E.-W., McMichael, J.F., Ning, J., McLellan, M.D., Xie, M., Cao, S., Yellapantula, V., et al. (2016). Systematic discovery of complex insertions and deletions in human cancers. Nat. Med. *22*, 97–104.

Zafeiriou, S., Tefas, A., Buciu, I., and Pitas, I. (2006). Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Trans. Neural Netw. *17*, 683–695.

Zhang, J.-Y., Looi, K.S., and Tan, E.M. (2009). Identification of tumor-associated antigens as diagnostic and predictive biomarkers in cancer. Methods Mol. Biol. Clifton NJ *520*, 1–10.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P.W., and Zhou, X.J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. *40*, 9379–9391.

Zhou, Y., Liu, Y., Li, K., Zhang, R., Qiu, F., Zhao, N., and Xu, Y. (2015). ICan: an integrated co-alteration network to identify ovarian cancer-related genes. PloS One *10*, e0116095.

Zhu, Y., Wang, N., Miller, D.J., and Wang, Y. (2016). Convex Analysis of Mixtures for Separating Non-negative Well-grounded Sources. Sci. Rep. *6*, 38350.

Zumsteg, A., and Christofori, G. (2009). Corrupt policemen: inflammatory cells promote tumor angiogenesis. Curr. Opin. Oncol. *21*, 60–70.

List of scientific work published during the PhD project

Peer-reviewed articles:

- **Sompairac N**, Nazarov PV, Czerwinska U, et al. (2019). Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int J Mol Sci. 20(18):4414. (https://doi.org/10.3390/ijms20184414)
- Kondratova M, Czerwinska U, **Sompairac N**, et al. (2019). A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures. Nat Commun. 10(1):4808. (https://doi.org/10.1038/s41467-019-12270-x)
- Ravel JM, Monraz Gomez LC, **Sompairac N**, et al. (2020). Comprehensive Map of the Regulated Cell Death Signaling Network: A Powerful Analytical Tool for Studying Diseases. Cancers (Basel). 12(4):990. (https://doi.org/10.3390/cancers12040990)
- Decamps C, Arnaud A, Petitprez F, et al. (2021). DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. BMC Bioinformatics. 22(1):473. (https://doi.org/10.1186/s12859-021-04381-4)

Chapters:

- Ravel JM, Monraz Gomez LC, Kondratova M, **Sompairac N**, Kuperstein I. (2019). Atlas of Cancer Signalling Network: An Encyclopedia of Knowledge of Cancer Molecular Mechanisms. Kuperstein I & Barillot E (ed.), Computational Systems Biology Approaches in Cancer Research. Academic Press, 1st Edition, pp. 17-24. (http://dx.doi.org/10.1201/9780429330179)
- Czerwinska U, **Sompairac N**, Zinovyev A. (2019). Deconvolution of Heterogeneous Cancer Omics Data. Kuperstein I & Barillot E (ed.), Computational Systems Biology Approaches in Cancer Research. Academic Press, 1st Edition, pp. 62-69. (http://dx.doi.org/10.1201/9780429330179)
- Monraz Gomez LC, Kondratova M, **Sompairac N**, Lonjou C, Ravel JM, Barillot E, Zinovyev A, Kuperstein I. (2021). Atlas of Cancer Signaling Network: A Resource of Multi-Scale Biological Maps to Study Disease Mechanisms. Wolkenhauer O (ed.), Systems Medicine. Academic Press, Vol. III, pp. 490-506. (https://doi.org/10.1016/B978-0-12-801238-3.11683-6)