UNIVERSITÉ PARIS DESCARTES

ED 474 Frontières du vivant

Institut Curie, PSL Research University, Mines Paris Tech, Inserm U900
Centre de Recherches Interdisciplinaires
Paris, France

Unsupervised deconvolution of bulk omics profiles: methodology and application to characterize the immune landscape in tumors

par Urszula Czerwińska

Thèse de doctorat Interdisciplinaire

Thèse dirigée par Andrei Zinovyev et Vassili Soumelis

Présentée et soutenue publiquement le 2 octobre 2018

Devant un jury composé de :

Andrei ZINOVYEV	directeur de thèse - Paris 5 Descartes
Vassili SOUMELIS	directeur de thèse - Paris 7 Diderot
Christophe AMBROISE	rapporteur - Université d'Evry Val d'Essone
Aurélien DE REYNIÈS	rapporteur - Université Paris 6 Pierre et Marie Curie
Jean-Yves BLAY	examinateur - Université Lyon 1
Marielle CHIRON	examinatrice - Sanofi
Marie-Caroline DIEU-NOSJEAN	examinatrice - Université Paris 6 Pierre et Marie Curie
Daniel GAUTHERET	examinateur - Université Paris Sud

Title: Déconvolution non supervisée des profils omiques de masse: méthodologie et application à la caractérisation du paysage immunitaire des tumeurs

Résumé (français) :

Les tumeurs sont entourées d'un microenvironnement complexe comprenant des cellules tumorales, des fibroblastes et une diversité de cellules immunitaires. Avec le développement actuel des immunothérapies, la compréhension de la composition du microenvironnement tumoral est d'une importance critique pour effectuer un pronostic sur la progression tumorale et sa réponse au traitement. Cependant, nous manquons d'approches quantitatives fiables et validées pour caractériser le microenvironnement tumoral, facilitant ainsi le choix de la meilleure thérapie.

Une partie de ce défi consiste à quantifier la composition cellulaire d'un échantillon tumoral (appelé problème de déconvolution dans ce contexte), en utilisant son profil omique de masse (le profil quantitatif global de certains types de molécules, tels que l'ARNm ou les marqueurs épigénétiques). La plupart des méthodes existantes utilisent des signatures prédéfinies de types cellulaires et ensuite extrapolent cette information à des nouveaux contextes. Cela peut introduire un biais dans la quantification de microenvironnement tumoral dans les situations où le contexte étudié est significativement différent de la référence.

Sous certaines conditions, il est possible de séparer des mélanges de signaux complexes, en utilisant des méthodes de séparation de sources et de réduction des dimensions, sans définitions de sources préexistantes. Si une telle approche (déconvolution non supervisée) peut être appliquée à des profils omiques de masse de tumeurs, cela permettrait d'éviter les biais contextuels mentionnés précédemment et fournirait un aperçu des signatures cellulaires spécifiques au contexte.

Dans ce travail, j'ai développé une nouvelle méthode appelée DeconICA (Déconvolution de données omiques de masse par l'analyse en composantes immunitaires), basée sur la méthodologie de séparation aveugle de source. DeconICA a pour but l'interprétation et la quantification des signaux biologiques, façonnant les profils omiques d'échantillons tumoraux ou de tissus normaux, en mettant l'accent sur les signaux liés au système immunitaire et la découverte de nouvelles signatures.

Afin de rendre mon travail plus accessible, j'ai implémenté la méthode DeconICA en tant que librairie R. En appliquant ce logiciel aux jeux de données de référence, j'ai démontré qu'il est possible de quantifier les cellules immunitaires avec une précision comparable aux méthodes de pointe publiées, sans définir a priori des gènes spécifiques au type cellulaire. DeconICA peut fonctionner avec des techniques de factorisation matricielle telles que l'analyse indépendante des composants (ICA) ou la factorisation matricielle

non négative (NMF).

Enfin, j'ai appliqué DeconICA à un grand volume de données : plus de 100 jeux de données, contenant au total plus de 28 000 échantillons de 40 types de tumeurs, générés par différentes technologies et traités indépendamment. Cette analyse a démontré que les signaux immunitaires basés sur l'ICA sont reproductibles entre les différents jeux de données. D'autre part, nous avons montré que les trois principaux types de cellules immunitaires, à savoir les lymphocytes T, les lymphocytes B et les cellules myéloïdes, peuvent y être identifiés et quantifiés.

Enfin, les métagènes dérivés de l'ICA, c'est-à-dire les valeurs de projection associées à une source, ont été utilisés comme des signatures spécifiques permettant d'étudier les caractéristiques des cellules immunitaires dans différents types de tumeurs. L'analyse a révélé une grande diversité de phénotypes cellulaires identifiés ainsi que la plasticité des cellules immunitaires, qu'elle soit dépendante ou indépendante du type de tumeur. Ces résultats pourraient être utilisés pour identifier des cibles médicamenteuses ou des biomarqueurs pour l'immunothérapie du cancer.

Title: Unsupervised deconvolution of bulk omics profiles: methodology and application to characterize the immune landscape in tumors

Abstract: Tumors are engulfed in a complex microenvironment (TME) including tumor cells, fibroblasts, and a diversity of immune cells. Currently, a new generation of cancer therapies based on modulation of the immune system response is in active clinical development with first promising results. Therefore, understanding the composition of TME in each tumor case is critically important to make a prognosis on the tumor progression and its response to treatment. However, we lack reliable and validated quantitative approaches to characterize the TME in order to facilitate the choice of the best existing therapy.

One part of this challenge is to be able to quantify the cellular composition of a tumor sample (called deconvolution problem in this context), using its bulk omics profile (global quantitative profiling of certain types of molecules, such as mRNA or epigenetic markers). In recent years, there was a remarkable explosion in the number of methods approaching this problem in several different ways. Most of them use pre-defined molecular signatures of specific cell types and extrapolate this information to previously unseen contexts. This can bias the TME quantification in those situations where the context under study is significantly different from the reference.

In theory, under certain assumptions, it is possible to separate complex signal mixtures,

using classical and advanced methods of source separation and dimension reduction, without pre-existing source definitions. If such an approach (unsupervised deconvolution) is feasible to apply for bulk omic profiles of tumor samples, then this would make it possible to avoid the above mentioned contextual biases and provide insights into the context-specific signatures of cell types.

In this work, I developed a new method called DeconICA (Deconvolution of bulk omics datasets through Immune Component Analysis), based on the blind source separation methodology. DeconICA has an aim to decipher and quantify the biological signals shaping omics profiles of tumor samples or normal tissues. A particular focus of my study was on the immune system-related signals and discovering new signatures of immune cell types.

In order to make my work more accessible, I implemented the DeconICA method as an R package named “DeconICA”. By applying this software to the standard benchmark datasets, I demonstrated that DeconICA is able to quantify immune cells with accuracy comparable to published state-of-the-art methods but without a priori defining a cell type-specific signature genes. The implementation can work with existing deconvolution methods based on matrix factorization techniques such as Independent Component Analysis (ICA) or Non-Negative Matrix Factorization (NMF).

Finally, I applied DeconICA to a big corpus of data containing more than 100 transcriptomic datasets composed of, in total, over 28000 samples of 40 tumor types generated by different technologies and processed independently. This analysis demonstrated that ICA-based immune signals are reproducible between datasets and three major immune cell types: T-cells, B-cells and Myeloid cells can be reliably identified and quantified.

Additionally, I used the ICA-derived metagenes 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 identification of new drug targets or biomarkers for immunotherapy of cancer.

Mots-clés (français) : microenvironnement tumoral, biologie des systèmes de cancer, analyse de données omiques, analyse de données monocellulaires, bioinformatique, hétérogénéité, séparation aveugle de source, apprentissage non supervisé, cancer, oncologie, immunologie

Keywords: tumor microenvironment, cancer systems biology, omic data analysis, single cell data analysis, bioinformatics, heterogeneity, blind sources separation, unsupervised learning, cancer, oncology, immunology

Dédicace

À Richard

Avertissement

Remerciments

Merci tout le monde

Motto

And now, let's repeat the Non-Conformist Oath!
I promise to be different!
I promise to be unique!
I promise not to repeat things other people say!
— Steve Martin, *A Wild and Crazy Guy (1978)*

Contents

List of Tables

List of Figures

Abbreviations

TME

DNA

RNA (mRNA, miRNA)

FACS

scRNA-seq

RNA-seq

CAF

TIL

DGE

BSS

CRI

ML

AI

TMA

CNV

CNA

Preamble about Interdisciplinary Research

> *We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline.* — Karl Popper

The piece of work you are reading should harvest fruit of an interdisciplinary research conceived in an interdisciplinary environment of Center for Interdisciplinary Research in Paris (CRI) in École doctorale *Frontières du Vivant* (FdV) and Institut Curie in groups Computational Systems Biology of Cancer and Integrative Biology of Human Dendritic Cells and T-cells. CRI's main mission can be formulated as follows:

> *to empower the students to take initiative and develop their own research projects* ***at the crossroads of life, learning, and digital sciences****.* [1]

Interdisciplinarity has many definitions and meanings. According to the book *Facilitating Interdisciplinary Research* [2]

> *Interdisciplinary research and education are inspired by the drive to solve* ***complex questions*** *and problems, whether generated by scientific curiosity or by society, and lead researchers in different disciplines to meet at the* ***interfaces*** *and* ***frontiers*** *of those disciplines and even to* ***cross frontiers*** *to* ***form new disciplines****.*

For me, the essence of interdisciplinarity is the need to solve a complex problem, whatever expertise would be necessary to solve it. I consider that fighting cancer disease, deciphering cancer heterogeneity and interactions of immune system are causes worth an interdisciplinary effort. This is even more true in the era of big data, when the demand for quantitative tools is exponentially growing, in order to extract information and knowledge.

Though this preamble I would like not only praise the interdisciplinary research but also underline possible limitations and constraints that come with it and which could affect

this thesis.

What does interdisciplinarity in science mean in XXI century?

In the ancient history, being formed and practice multiple disciplines was not anything unusual which is strongly reflected in Greek philosophy initiating the dispute about the division and hierarchical classification of knowledge. [194]. Figures as Aristotle and Leonardo Da Vinci, that can be called *homo universals* served different disciplines from arts through history, natural sciences to mathematics. With time human knowledge about the word, i.e. natural sciences got bigger and bigger, to the point that it became hard to master all the disciplines. The specialisation would allow to study in deep a certain subject and make possible discoveries about it. And even if, interdisciplinary efforts never stopped, for long time they were not mainstream in scientific communities divided into academies, chairs and specialization.

Different fields differ in term of concept, method, tools, processes and theories [194]. Thanks to division into scientific disciplines certain order is conserved across space and time. Hierarchical classification of knowledge comes with human nature.

It can be observed that there is an increasing gap between disciplines along with specialization.

> *advancing specialisation leads to gaps in the level of comprehension between individual disciplines and eventually gives rise to the demand for interdisciplinarity - in order to close the gaps between disciplines.*[194]

It is not really clear why this gap must happen. Would it somehow reflect a human nature, the strong need to divide things into discrete categories rather than to see a continuum?

Nowadays, the knowledge is accessible, we can profit from achievements of different disciplines thanks to easy means of communication. Two different terms can be defined to describe initiatives that use the knowledge of different specialities: multidisciplinarity which is a sum of efforts of different disciplines and interdisciplinarity that allows to profit from synergy of multiple disciplines (Fig. 1). With interdisciplinary research and education come flexibility, creativity and novelty but also limit of depth on ingested knowledge and possibilities of cross-interactions between disciplines.

Why not all the labs are interdisciplinary?

> *Scientists tend to resist interdisciplinary inquiries into their own territory. In many instances, such parochialism is founded on the fear that intru-*

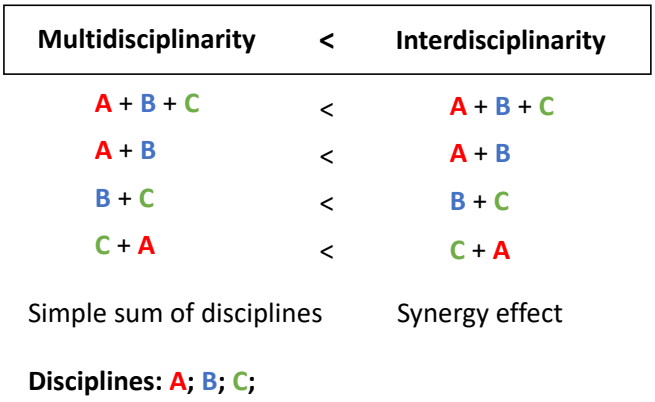


Figure 1: Symbolic illustration of sum (multidisciplinarity) versus synergy (interdisciplinarity), in an interdisciplinary project sum of thee disciplines A, B, C should have more value than a simple sum od disciplines: an interdisciplinary project should have an added value compared to a multidisciplinary one.

> *sion from other disciplines would compete unfairly for limited financial resources and thus diminish their own opportunity for research* — Hannes Alfvén

Crossing frontiers is not an easy task, and it was quite difficult in the beginnings of modern interdisciplinarity. Some examples of early interdisciplinary efforts of 20th century are nicely described by Ledford et al. [112] in *Nature* special issue on Interdisciplinarity. It illustrates Theodore Brown in 1980s, while trying to organise a new interdisciplinary research project and reorganise university space to engage exchange between students of different faculties, he encounter a lot of reluctance.

> *And then there was the stigma. "Interdisciplinary research is for people who aren't good enough to make it in their own field," an illustrious physicist chided* [112].

The story seems to end up with a happy ending of 40-milion US dollars grant and foundation of Beckman Institute for Advanced Science and Technology. However, recruiting open-minded director to lead this unconventional organisation was a struggle. Soon, the organisation became a model for others and met a great scientific and technological success.

Even though, since then the idea of interdisciplinary research spread around the world. Still, not all problem got overcome.

> *"There's a huge push to call your work interdisciplinary," says David Wood, a bioengineer at the University of Minnesota in Minneapolis. "But there's still*

resistance to doing actual interdisciplinary science".

First, the institutions, universities where research is performed should equip scientist with a passport to other disciplines, facilitate exchange, funding the interdisciplinary research, accepting fusion of disciplines as new ones. Then, a proper communication between disciplines is necessary. Finally, forming interdisciplinary researches is extremely challenging as it often requires extra effort from an apprentice.

Are all the disciplines idependent units nowadays?

Can we do molecular biology without technical, mathematical and computational support? Can we study cognitive science without knowledge of biology, physics and psychology? Can we advance medecine without basic research in biology, physiology, electronics?

Bioinformatics and/or computational biology is an interesting case. Working in this field being between biology, medecine, computer science, mathematics and statistics, the role of a computational biologist is sometimes reduced to a service. A biological lab may need a computational biologist to perform an analysis, restructure the data, that is needed for the biological discovery. Often, there is not enough space for research in computational biology itself, where the discovery does not depend on the original data but on tools and approaches to complex, data-intensive biological problems. It may happen also the other way round, when a computational biologist ask a bench researcher to perform an experiment to prove his theoretical model. In both cases, the long-term interdisciplinary partnership would probably fail. A wet and dry researchers should collaborate as equal with important research advances on both side to assure a long term equilibrium.

How interdisciplinarity changed over years? Are all disciplines affected equally?

From the chart (Fig. 2) we can see that Social Studies of Medicine seems to be the most interdisciplinary field. In general Biology, Health and Biomedical Sciences seem to be more open into flow of knowledge from other fields than humanities. On the extreme opposite of health, Clinical Medecine appears to be very conservative field.

Strengths, Weaknesses Opportunities, Threats (SWOT) of an interdisciplinary PhD - personal perspective

> *I'm not good enough to do well something I dislike. In fact, I find it hard enough to do well something that I like* — Jim Watson, Succeeding In Science: Some Rules Of Thumb [45]

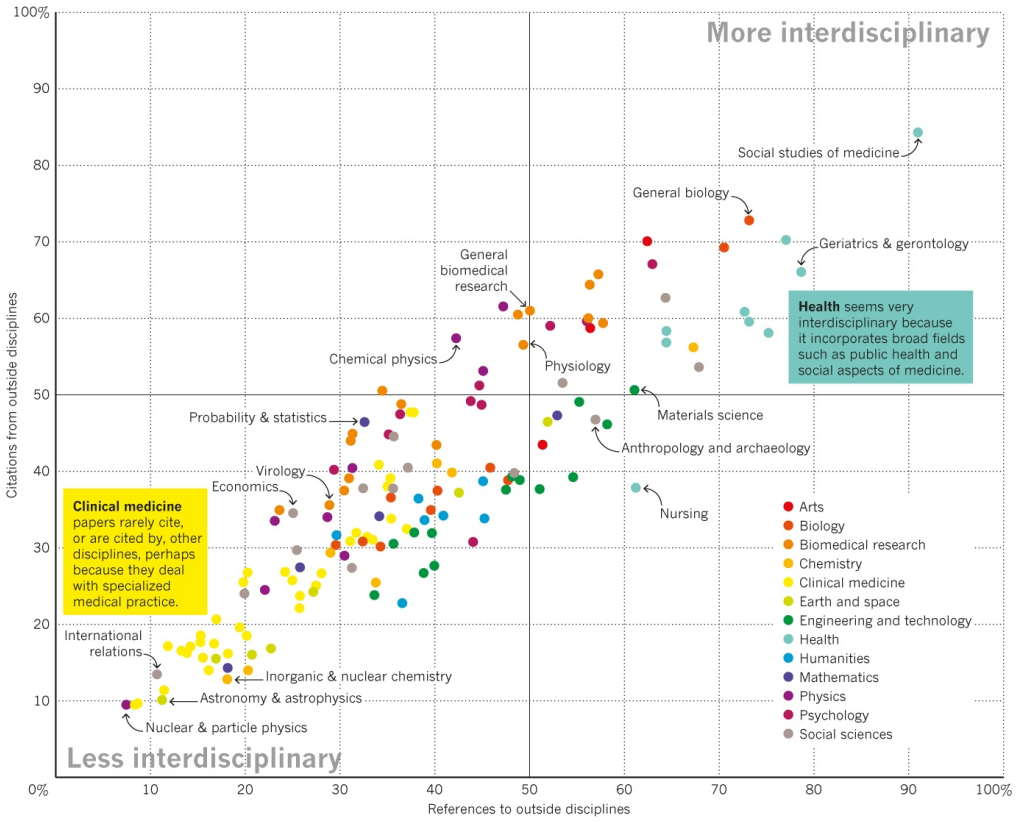


Figure 2: Interdisciplnarity of different fields. "From 1950-2014, a field's position is determined by how much its papers cite outside disciplines (x-axis), and by how much outside disciplines subsequently cite its papers (y-axis). (Some years, certain fields have too few references to be plotted.)". Reprinted by permission from Springer Nature [213] © 2015 Nature America, Inc. All rights reserved.

Being formed first in double major in biology and mathematics, then participating in interdisciplinary research projects during my master studies, I can witness that the learning curve of multiple disciplines can be steep. It is also often associated with frustration of not going deep enough in all of disciplines or the feeling of being overwhelmed by the amount of knowledge.

Coming with expertise of biology and mathematics, I got fascinated by complex biological systems. One way of study high-dimensional data is to reduce them into smaller interpretable units. This is what I tempted to achieve in this thesis in order to enrich our knowledge about tumor microenvironment and possible contribute to orienting future research on immunotherapies.

However, being an interdisciplinary researcher was not always a privilege. *To which category do I belong? To whom should I present my work?* I often asked myself these questions. I also often encountered lack of understanding where my methodological results were not bringing enough of *biological insights*. Or the constraints of my biological application seemed very obscured and complicated for mathematicians and my work often lacked *important methodological advances*.

Does it mean that my work is not accurate, useless? Probably, for many, it is not enough. However, I still hope that our findings will be interesting to some. I enjoy working with data and statistics that serve an actual purpose. The Tab. 1 summarizes Strengths, Weaknesses, Opportunities and Threats (SWOT analysis) of an interdisciplinary projects, in the way I see it.

Besides conducting research that crosses the boundaries of one discipline, I also could meet and work with inspiring people coping like me with filling the gap in understanding of an interdisciplinary work, multiple supervisors and reporting to many institutions. I gained (even if only superficial) understanding of many topics in mathematics, statistics, data science, immunology, cancer but also oral and written presentation skills, time and work management

Is my thesis really interdisciplinary? Does biology profits from mathematics and mathematics from biology? I will let you judge it.

What impact had biology on the statistical/mathematical modelling ? The practical problems, systems that go beyond theoretical formulations challenge the theoretical tools. In my work, I did my best to fuse theory and practice that should serve a biological application. I can image the project more complete if the results of my work would inspire changes in biological experiments, uncover new paths to follow for experimental biologists or translational researchers.

Table 1: SWOT analysis of Interdisciplinary research. In SWOT analysis, Strengths, Weakenesses, Opportunities and Threats are enumerated. Strengths and Wearknesses are internal and Opportunities and Threats are external factors.

Strengths (internal, positive)	Weaknesses (internal, negative)	Opportunities (external, positive)	Threats (external, negative)
Having a holistic view of the problem	Not seeing details of the problem	Mulitple possibilities to convey research	Spending too much time filling knowledge gap
Being supervised by multiple experts	Following multiple, sometimes contradictory, advice on the same problem	Take advantage of synergistic effect of fields	Inhibiting effect of oppinions from different fields
Joining expertises of different fields	Not covering in details all the disciplines	Doing a new discovery	Obtaining too generic results
Using new/non standard approach	Experiencing steep learning curve	Raising interest in different expert domains	Not mastering the specific vocabulary of different fields
Having better understanding of complex processes	Being in constant need of help of domain experts	Making progress	Not being understood
Higher creativity		Creating a new field	Being hard to classify/ fall into a category
Having great flexibility		Sovling many problems impossible to solve with traditional approach	Being considered as superficial
Feeling a thrill of adventure			
Being open			

The origins of the PhD topic

> *The universe will lead me where I need to go. I am like a leaf in the stream of creation* — Dirk Gently, Holistic detective

When finishing my master I was looking for an interdisciplinary subject where I could deepen my quantitative skills and apply to a real-life healthcare problem. I came across a project proposed by Andrei Zinovyev in close collaboration with Vassili Soumelis. I was quite anxious that my knowledge of cancer immunology would not be sufficient to lead the project to a success. I recognise that the complexity and heterogeneity of immune systems are is very complex and dynamic system and many years of expertise are needed to really grasp the understanding of it. I had a great chance to work hand in hand with domain experts that would suggest me the direction I should take in my research.

The project started by causal exploration of different blind source separation or dimension reduction techniques and their ability to dissect bulk transcriptomic data into cell type-related units. We also faced an important problem of lack of gold standard data that would define efficiency and accuracy of different methods.

I have spent void efforts working on a bulk transcriptomic data simulation framework, important statistical issues come into our way and probably another Ph.D would be necessary to solve them. In the meantime, many tools dissecting tumor bulk transcriptome were published. Serving a similar purpose, they used different means and assumptions, which left a space for my project to continue. In my third year, I am finally publishing a tool that performs the analysis I developed together with the Sysbio team members, and I can apply it to a corpus of publicly available data to learn about actual question: the immune system infiltrating cancers and the context-dependent signatures (see Chapters 4 & 5).

In a parallel project, I worked on exploration of brand new data type: single cell transcriptomic (RNAseq) in the context of tumor microenvironment (see Chapter 6).

We have also participated in Dream Idea Challenge, a project that aimed to put closer experimental and theoretical researchers [12]Annexe1.

I have collaborated in numerous projects within and outside my team. Some of the projects resulted in publications, such as my work on analysing pDC subsets of X cancer Annexe2. Others are in still preparation.

Alongside with pursuing the compelling scientific research, I completed a wide variety of courses. Thanks to this extensive (>300 hours of training over 3 years), I am equipped with soft skills that not only helped me to shape my thesis project on the go but also, I hope, will help me to succeed in my future career path.

Organisation of the dissertation

As it is a fruit of an interdisciplinary work, I decided to introduce the topic from two perspectives: describe the biological and biomedical dimension of the topic (see Chapter 1), as well as, the mathematical dimension of the problem of separation of sources in complex mixtures (see Chapter 2). I hope, it will make the subject of my thesis easy to understand also for non-biologists or non-mathematicians. In the results part, I compare reproducibility of blind source separation methods NMF and ICA (see Chapter 4), I cite our results on the estimating the Most Reproductible Transciptomic Dimension (MSTD). I also apply ICA-based deconvolution to Breast cancer transcriptomes to to prove its reproducibility Chapter 3. Then I introduce the DeconICA R package (see Chapter 5) and finally present results of application of DeconICA and other tools to >100 transcriptomic datasets (see Chapter 6). A second part of results is dedicated to my work on cell type heterogeneity (see Chapter 7). The manuscript finishes with Chapter 8 that contains discussion, conclussions and perspectives. In annexes you can find publications to which I contributed during my doctorate that are not strictly linked with the topic of this thesis.

INTRODUCTION

- Chapter 1: introduction to cancer biology and immunity, challenges in cancer immunotherapies and cancer immune phenotyping as well as data sources most commonly used to face the topic.

- Chapter 2: introduction to a problem of mixed sources in biological samples, overview of blind source separation methods and supervised deconvolution methods, with focus on those applied to bulk transcriptome to uncover and quantify immune compartments

RESULTS

- Chapter 3: Most Reproducible Transcriptome Dimension (MSTD)
- Chapter 4: application of ICA-based deconvolution to six breast transcriptomes
- Chapter 5: comparison of reproducibility of NMF and ICA methods
- Chapter 6: DeconICA R package

- Chapter 7: application of DeconICA R package and other tools to analyse >100 transcriptome datasets of bulk cancer transcriptomes
- Chapter 8: study of immune cell types heterogeneity in tumor microenvironment using innate immune map and scRNAseq data

DISCUSSION

- Chapter 9: discussion, conclussions and perspectives

ANNEXES

- Other publicatons:
  - pDC subsets
  - Idea Dream Challenge

Part I

Introduction

Chapter 1

Immuno-biology of cancer

This chapter will first introduce a short history of cancer with a focus on discoveries linking cancer and its environment. It will also describe participation of TME in cancer development, progression and response to treatment. Most important types of data used to study cancer microenvironment will be discussed. I also introduce a link between tumor immune-biology and cancer phenotyping for development of immunotherapies.

1.1 Cancer disease

According to GLOBOCAN study [62], 14.1 million cancer cases was estimated to happen around the world in 2012. It touched 7.4 million men and 6.7 million women. It is estimated that the cancer cases will increase almost two-fold to 24 million by 2035.

In France only, in 2012 there were 349426 cases of cancer, of which leading is Prostate cancer (16,3%) followed by Breast (14%) and Lung (11,5%).

For a long time studying tumor was focused on tumor cells, their reprogramming, mutations. Cancer was seen as disease of uncontrolled cells by the mainstream research. At the same time, the idea of importance of the impact of other cells and structures on cancer cells was present but often not believed. Recent success of immunotherapies moved research focus to tumor cells in their context: tumor microenvironment. We will describe here what is the composition and role of the TME in tumor progression, diagnosis and response to treatment.

1.1.1 Historical understanding of cancer

Cancer was historically described by a physician Hippocrates (460–370 B.C) [199]. Even though there exist even earlier evidence of the disease. Hippocrates stated that the body contained 4 humors (body fluids) : blood, phlegm, yellow bile and black bile. Any imbalance of these fluids will result in disease. Particularly the excess of black bile in an organ was meant to provoke cancer. For years, it was not known what factors cause cancer and it was easily confounded with other diseases. In the middle ages in the Renaissance Period it was believed cancer is a punishment for the sins they committed against their god, that they deserved it to some extend

Until 18th century it was believed that cancer is contagious and is spread by parasites.

In the 19th century, tumor cells started to be analysed by pathologists. They were strike with their ability to proliferate uncontrollably, ability to spread and destroy the original tissue [143]. Around the same time leukocytes from the blood was first described by Gabriel Andra and William Addison. Just a few years later, in 1845 Bennett and Virchow described blood cells in leukaemia (Fig. 1.1). Virchow is also a father of Chronic irritation theory (nowadays calledl chronic inflammation) that says that cancer is caused by local "irritation" and, incorrectly, that cancer cells spread like liquid resulting in metastasis.

In 1889, Stephen Paget introduced *soil and seed* hypothesis of metastases [157]. He formulates it as follows

> *When a plant goes to seed, its seeds are carried in all directions, but they can only live and grow if they fall on congenial soil.*

Which is a parallel to cancer cells disseminated by body fluids, and they can grow only tissues - "soil" that is predisposed to host the cancer cell - "the seed". He focused on the importance of tissue characteristics that favorise tumor development as opposed to most researchers of his time that were focusing on the "seed" itself.

In the 20th century, molecular causes started to be investigated. It was discovered that cancer could be caused by environmental factors, i.e. chemicals (carcinogens), radiation, viruses and also inherited from ancestors. Those factors would damage but contrary to a healthy condition they would not die.

Also in 1909, Paul Ehrlich, called one of fathers of immunology and Nobel Prize laureate, indicated a link between immune system and tumor suppression [53]. One of remarkable first immontherapy attempts can be attributed to William Coley, that practiced injecting streptococcus bacteria directly into patients after cancer surgery in 1891, later called "Colley vaccine". However, the impact of this procedure on patients recovery was judged by scientific community as "unclear".

In 1968, Melvin Greenblatt and Philippe Shubik showed that tumour transplants secete

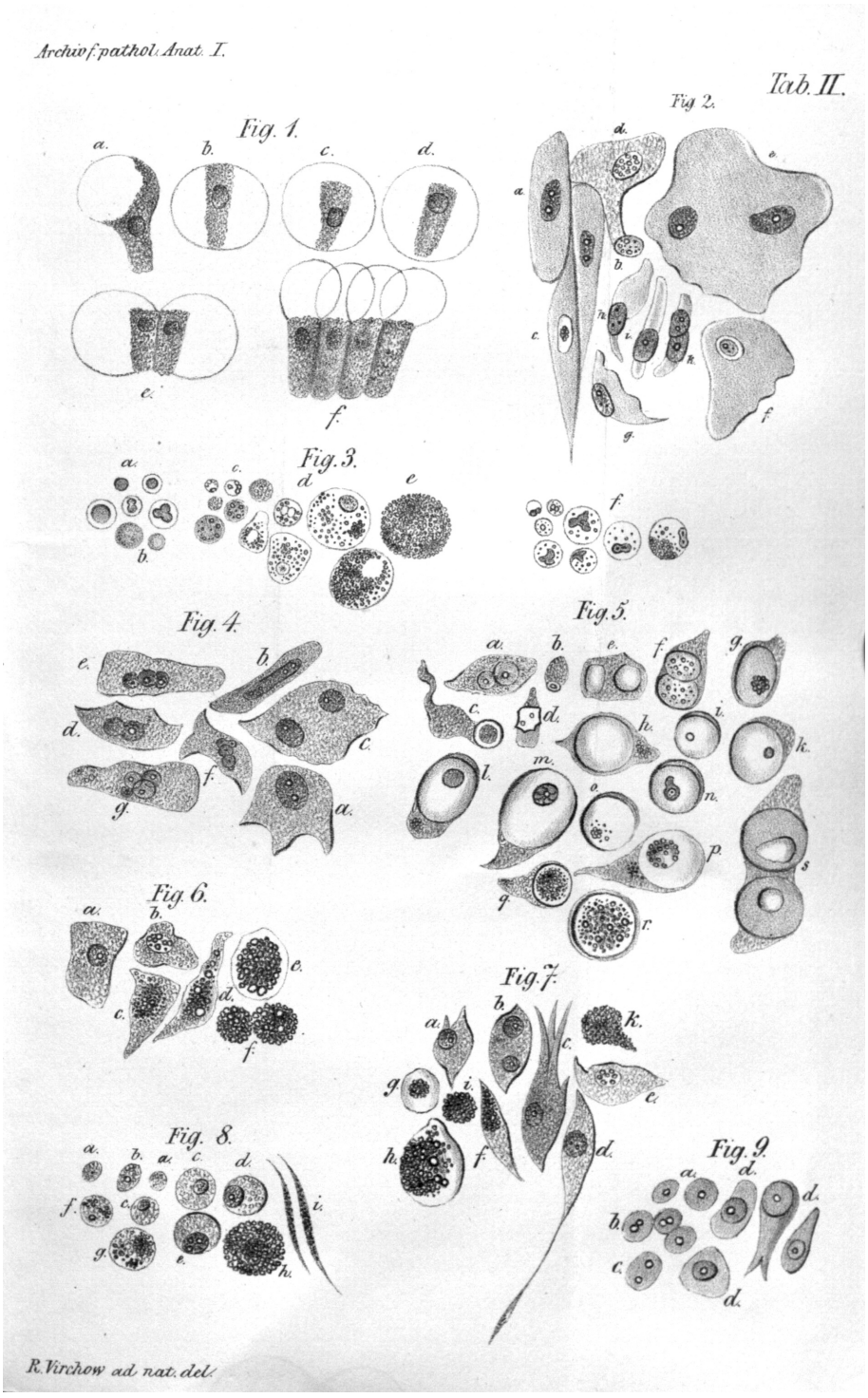


Figure 1.1: Illustration of Virchow's cell theory. Virchow depicted different cells transformation due to irritation. [219]

a substance stimulating the growth of blood vessels [81], later identified as “tumour angiogenic factor (TAF)” by Judah Folkman in 1971 [61]. Folkman also suggested that TAF can be a target of a therapy itself. This was a revolutionary idea, at the time, as it did not target the tumor cells directly acted on thier environment.

During the 1970s, oncogenes and tumor suppressor genes were discovered. Oncogenes are genes that allow a cell to become cancer cell, while the tumor suppressor genes would repair DNA or execute cell death of a damaged cell. A new dimension to cancer studies was added in the 1980s, epigenetic changes was proven to occur to both oncogenes and tumour suppressors [59, 82], which are presently known as epigenetic markers used for diagnostics and therapeutic targets for cancer.

In 1982, Aline van Pel and Thierry Boon [214] discovered that a specific immunity to spontanious tumor cells could be induced by vaccinating mice with mutagenized tumour cells. This araised an inspiration for many years of immune therapy developement.

In Napoleone Ferrara and colleagues identified gene encoding vascular endothelial growth factor (VEGF) that was shown to stimulate growth of enothelial cells proliferation *in vitro* and angiogenesis (blood vessels formation) *in vivo* [115].

In 1999 for the first time, gene-expression was used to study cancer (leukeamia) by Todd Golub, Donna Slonim and colleagues [76].

Since the end of the 20th century, cancer screens are developed along with multiple strategies to fight tumor. Most classical ones are based on the idea of removing tumor cells (surgery), killing tumor cells with DNA-blocking drugs (chemotherapy), radiation, inhibit cancer growth (hormonal therapy, adjuvant therapy and immunotherapy). As non of those methods is fully efficient, often a combination of treatments is proposed. Nowadays, science is aming in the direction of trageted therapies and personalized treatment.

The recent success of immunotherapies (discussed in Immunotherapies section attracted the attention the scientific community again to the context in which tumor cells are found. This context called Tumor Microenvironment, as well as the communication that happens within it between different agents, nowadays studied differently with available knowledge of molecular biology, have become a popular scientific topics of 21st century (Fig. 1.2).

1.1.2 Tumor Microenvironment as a complex system

Tumor Microenvironment is a complex tissue that surrounds tumor cells. It is composed of different compartments (in solid tumors):

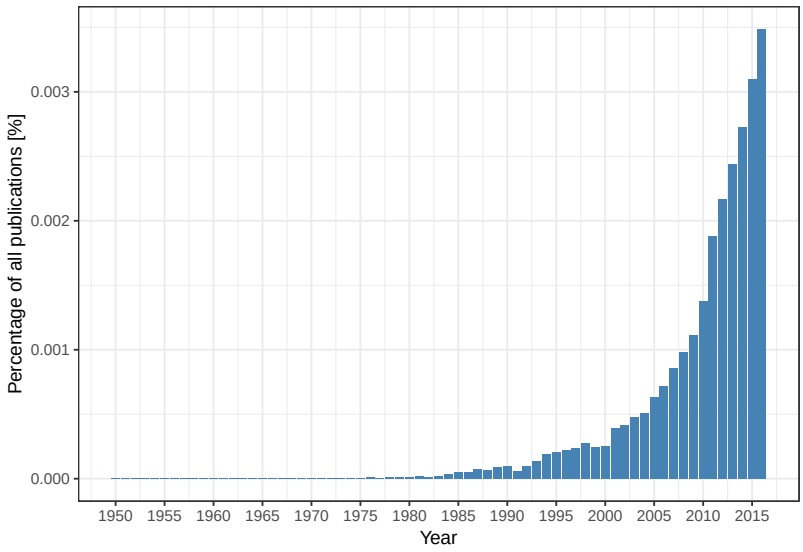


Figure 1.2: Percentage of publications containg phrase "tumor immunotherapy" is growing, numbers retreived on 17.01.2018 from Medline Trends [43]

- Stroma: blood and lymphatics vessels, epithelial cells, mesenchymal stem cells, fibroblasts, adipocytes supported by extracellular matrix (EM)
- Immune cells: T cells, B cells, NK cells, Dendritic cells, Macorphages, Monocytes etc.

Their proportion and specific roles vary significantly with tumor type and stage. Communication between the environmental cells and the tumor is critical for tumor development and its impact on patient's response to treatment. These communication between different compartments is bidirectional and all the players can influence each other. Depending on the nature and prevailing direction of those interactions different destiny is possible for each of the compartments, i.e. immune cells can be recruited to protect tumor cells or they can kill them directly. Many of the signals can be contradictory, many can suppress each other. Then is it possible to tilt this complex ecosystem into patients' favour? Can we decipher the most important factors of this molecular knot and manipulate it?

Next section describes different scenarios of interaction within TME in order to illustrate the complexity of TME and possible targets for cancer therapies.

1.1.2.1 Interactions between TME and Tumor

Three scenarios can be considered to describe the relationship between TME and tumor cells:

1. TME stimulates tumor growth and/or progression and/or impact negatively the response to treatment
2. TME has no impact on tumor cells and disease development
3. TME has a tumor suppressive role and impact positively the response to treatment

As can be seen partly in Historical understanding of cancer these three hypothesis were gaining and loosing popularity in scientific and medical community over the decades.

1.1.2.1.1 TME as a foe: inflammation

In 1863 Rudolf Virchow observed a link between chronic inflammation and tumorigenesis. According to Virchov theory, genetic damage would be the "match that lights the fire" of cancer, and the inflammation or cytokines produced by immune cells should be the "fuel that feeds the flames" [13]. Therefore lymphocyte infiltration was confirmed by subsequent studies as a hallmark o cancer. The question one may ask is why our immune system is not enough to defend the organism from tumor cells as it does efficiently in a range of bacterial and viral infections? It is mainly because of the ability of tumor cells to inhibit immune response through activation of negative regulatory pathways (so called immune checkpoints).

Many examples can be cited on how TME facilitates tumor development (Fig. 1.3). For instance, in the early stages of tumorigenesis some macrophage phenotypes support tumor growth and mobility through TGF-beta signaling. Also, it was shown that NK cells and myeloid-derived suppressor cells (MDSCs) have an ability to suppress immune defence i.e. immunosurveillance by dendritic cells (DCs), T cell activation and macrophage polarisation and they promote tumor vascularisation as well. [201, 65] They create so-called niches that facilitates tumor colonization. Tregs and myeloid-derived suppressor cells can negatively impact natural immune defence and by these means allow growth and invasion of tumor cells [203]. Another cell type, a part of ECM, fibroblast, or more precisely Cancer Associated Fibroblasts (CAFs) have proven pro-tumor functions in breast cancer where they enhance metastasis [50]. The blood and lymphatic vessels maintain tumor growth providing necessary nutritive compound to malignant cells.

According to [85] immune and stroma cells participate in almost all of Cancer Hallmarks [84, 85]. Most of the hallmarks of cancer are enabled and sustained to varying degrees through contributions from repertoires of stromal cell types and distinctive subcell types.

1.1.2.1.2 TME seen as neutral

In front of lack of definitive proof that TME can positively or negatively impact on tumor development, many scientist, in a long time, ignored the importance of this factor. Un-

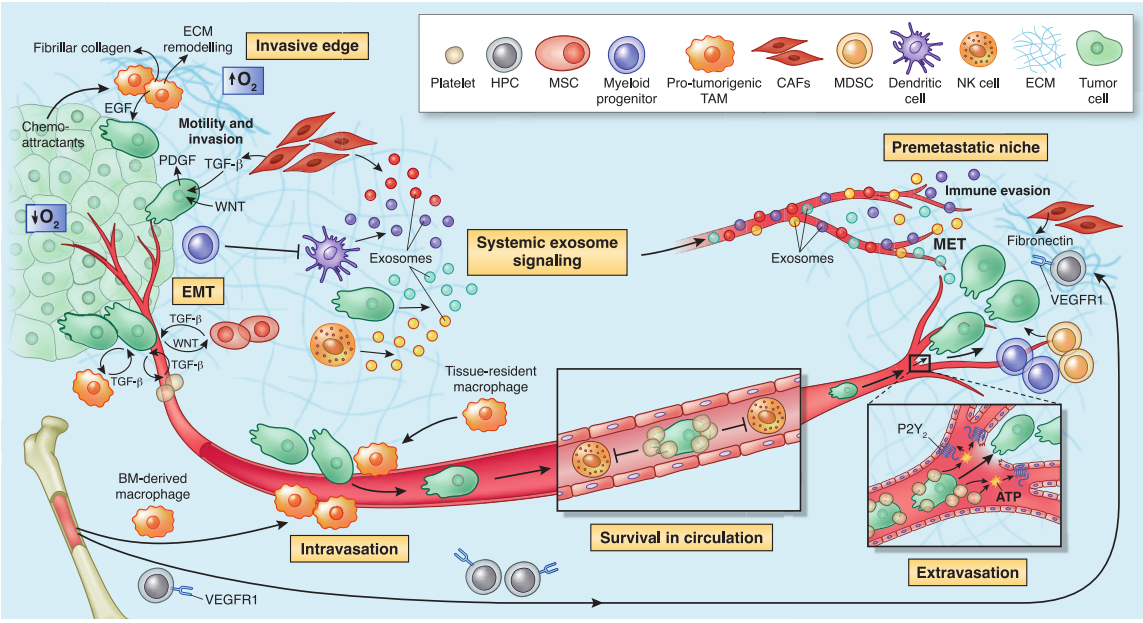


Figure 1.3: The microenvironment supports metastatic dissemination and colonization at secondary sites. Different tumor sites can communicate through exosomes realized by tumor cells and also immune and stromal cells such as NK cells, CAFs and DCs. Reprinted by permission from Springer Nature [171] © 2013 Nature America, Inc. All rights reserved.

til the early-mid eighties, the TME research was mostly limited to angiogenesis and immune environment and most areas that are now driving the field were not represented.

From early 70. until the end of the 90. the most accepted statement was that genetic alterations in oncogenes and tumor suppressor genes are both necessary and sufficient to initiate tumorigenesis and drive tumor progression. Therefore TME was not seen as an important element of the puzzle.

The cancer geneticists, at the time had a lot of influence on scientific community diminishing the work of made on TME which were considered as "uninteresting" and definitely not "mainstream".

After 90s, with discovery of signalling molecules involved in communication of TME like VEGF general opinion started to change. Also discoveries made by developmental biology field supported the hypothesis that microenvironment plays an important role in development which was later shown for tumorigenesis. Also success of immune vaccines starting with the tuberculosis vaccine Bacille Calmette-Guérin (BCG) in 1976 and finishing, at the moment with checkpoint inhibitors did not leave the scientific community indifferent.

1.1.2.1.3 TME as a friend: immunosurveillance

As mentioned in Section 1.1.1 Paget proposed a hypothesis of "seed and soil" where the TME in a certain tissue (the soil) can either stimulate or suppress the metastasis (the seed). William Coley tested a possibility to trigger tumorsuppressive effect via stimulation of the immune system with bacteria. In the 1960s, the immune surveillance theory hypothesized "the ability to identify and destroy nascent tumors as a central asset of the immune system" [187, 28]. Thus, the hypothesis that TME can have a positive role in tumor prognosis is not new.

In modern immuno-oncology, the term *immune-eiditing* was introduced by Dunn et al. [51] in 2002, to describe the relation between the tumor cells and the immune system. The immunosurveillance through immune-editing can be summarized in three processes: elimination, equilibrium, and escape [51].

The elimination is direct killing of cancer cells or growth inhibition by immune system. The adoptive T cells and NK are actively involved in tumor killing and stimulate other immune cells. The CD8 + cytotoxic lymphocytes (CTLs) directly recognize tumor cells. Employing perforin- and granzyme-dependent mechanisms they can lyze tumor cells. The CD4 + T cells release factors to induce proliferation of B cells and to promote their differentiation to antibody (Ab)-secreting plasma cells, activate macrophages. Macrophages use phagocytosis to eliminate cancer cells [218].

The tumor-infiltrating lymphocytes (TILs) have been associated with an overall good

prognosis and better survival in different cancer studies. Also, abundance of CD3 + and CD8 + T cells, NK cells, and $\gamma\delta$T cells correlates with improved outcomes in epithelial ovarian cancers [128]. Several studies report that the presence of the abundant immune infiltrate is correlated with good prognosis or better survival [105, 15, 138, 155]. Spontaneous regression of human tumors has been reported in cutaneous melanoma, retinoblastoma, osteosarcoma, etc. [10].

The equilibrium is the phase when cancer and immune cells coexist and their crosstalk is preventing metastasis.

T cells are the main actor maintaining the equilibrium. Progressively, the tumor cells become more immunogenic as they are not edited by the immune system [21]. The state of tumor cells is then identified as “dormant” and active scientific reports investigate the possible molecular pathways that maintain dormancy or lead to escape [204].

The immune escape is the final process when tumor cells impair the immune response.

1.1.2.2 Two-faced nature of immune cells: context-dependent functional plasticity

Modern vision of TME-tumor interactions assumes that tumor can be directed to several molecular pathways. This direction is decided by signals that are native of tumor cell and/or coming from the microenvironment.

Recent studies unveil ambivalent nature of immune cells in TME. While some as cytotoxic T cells, B cells and macrophages can manage to eliminate tumor cells. Treg cells role is to regulate expansion and activation of T and B cells. Depending on cancer type, they can be either pro- or anti-tumor. For example, as it has been shown for T-regs, usually associated with bad prognosis, they can be associated with improved survival (i.e. in colorectal cancer [63]). For innate immunity, there are widely accepted M1 (ani-tumor) and M2 (pro-tumor) extreme macrophages phenotypes in TME [169]. Most of the statements seem to be context dependent and not valid universally across all cancer types. We already mentioned Macrophages phenotypic plasticity as well as different behaviour of EMC depending on tumor stage.

From more general point of view, it has been observed that immunodeficiency can correlate with high cancer incidence. Results of analysis based on observations of 25,914 female immunosuppressed organ transplant recipients, the tumor incidence was higher than predicted for multiple cancers. However, the number of breast cancer cases decreased which can be really disturbing if we need to decide on the role of immune defence in tumor progression [197]. This indicates that immune microenvironment can be cancer stimulating or inhibiting depending on the type of cancer and/or other factors.

1.1.2.3 Immune cell (sub)types in TME

We are taught that a cell is the basic structural, functional, and biological unit of all known living organisms. Human body contains around 10^{14} which is three order of magnitude more than number of stars in the Milky Way. This ensemble of cells are traditionally classified into cell types based on their phenotypical variety.

> *for their immense number, the variety of cells is much smaller: only about 200 different cell types are represented in the collection of about 10^{14} cells that make up our bodies. These cells have diverse capabilities and, superficially, have remarkably different shapes....* Boal [25]

In the description of TME, I have referred to cell types of immune cells as well-established entities of immune system. However, the definition of cell types remains controversial and there is no consensus among researchers how exactly a cell type should be defined. The notion of the cell-subtypes is even more vague. The problem does not only concerns immune cells, most of cell types of our organism, classified initially according to their morphology, seem to fulfil multiple functions. One can also relate cell-type problem to species problem where scientist also debate about where to draw the borders between species. This problem is widely generalized as "theory of types" [193] in many disciplines as philosophy, linguistics, mathematics.

In this chapter I will limit the description to immune cell types.

An immune cell can be described nowadays along many axes:

- Phenotype /surface markers
- Stability
- Morphology (expressed proteins)
- Ultrastructure (electron microscopy)
- Molecular data (gene expresson, genotype, epigenome)
- Cell fate
- Cell of origin
- Function

Depending how well a cell is different from all other cells along those axes, it will (or not) be defined as a distinct cell type. However, this comes with more or less subjective threshold on where the cells become *significantly different*. This thresholds can be established computationally or by an expert. Usual practice is a mix of both methods.

Since the beginning of immunology, there were disagreement between pre-defined cell types and cell functions.

> *Cette espèce de leucocytes a une grande ressemblance avec certains éléments fixes du tissu conjonctif, ainsi qu'avec des cellules endothéliales*

> *et des cellules de la pulpe splénique. On est donc souvent embarrassé, surtout lorsqu'on trouve ces leucocytes mononucléaires en dehors des vaisseaux, pour les distinguer des autres espèces de cellules mentionnées.*
> — Elie Metchnikoff, Leçons sur la pathologie comparée de l'inflammation, 1891

The definition of cell types and subtypes is widely discussed today with arrival of single cell technologies that allow a change of paradigm in cell classifications. Up to now, the top-down approach was mostly used. Pre-defined set of parameters describing a cell was fixed in order to select cells and then other parameters were measured. Now, it is possible to practice bottom-up approach where all (or some) parameters are measured for a single cell and then, depending on its distance from other cells, cell types are defined [183].

> *The concept of "cell type" is poorly defined and incredibly useful*
> — Allon Klein, Harvard Medical School

Researchers agree that the concept of cell type is artificial and a continuum of cell types is closer to the reality. According to Susanne Rafelski,

> *A useful way to classify cells might thus be a multiscale and multi- parameter cell-type space that includes vectors for key intracellular organizational, dynamic, and functional features as well as tissue location, gene expression etc.*

Some, as Allon Klein, propose to introduce a concept of *cell states* which would better describe a cell depending on its context and function. However, an emerging challenge would be to connect *cell states* with historical *cell types*. [52] .

Another aspect of cells, that I am not approaching in this thesis is time. Cells are shaped by thier environment, intrinsic and extrinsic events and can change states, functions etc. Can one cell belong to different cell types depending on its trajectory? How to include the dynamic aspect of the cells into the classification?

Thus, most scientist agree that used convention of cell types is not ideal and it is more matter of convenience than biological reality. This leaves a room to study cells and challenge existing classification. Describing cell types or cell states in tumor micorenvironment is extremely interesting as still little is know about the diversity of cell infiltrated in solid tissues.

1.1.2.4 Summary

Cancer is a disease concerning milliards of people with a long history. Scientific community recognises role of the environment where the tumor cells find themselves as an

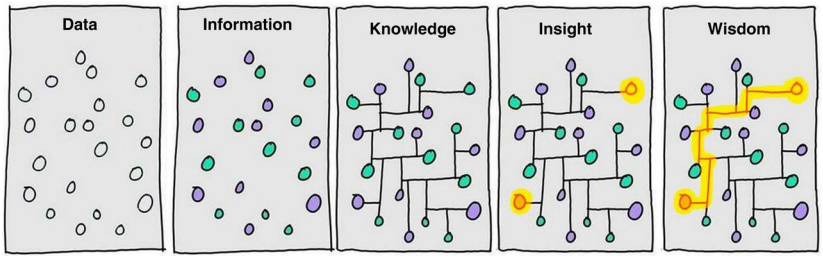


Figure 1.4: From *Data* to *Wisdom*. Illustration of different steps that it takes to go from *Data* to generating *Wisdom*. It highlights that generating data is not equal to understanding it and additional efforts are needed to generate value. Image authored by Clifford Stoll and Gary Schubert published by Portland Press Limited on behalf of the Biochemical Society and the Royal Society of Biology and distributed under the Creative Commons Attribution License 4.0 (CC-BY) in [165].

important factor influencing tumor development, prognosis and response to treatment. TME is a complex environment that constantly interacts with tumor cells, where both tumor and TME influence and shape each other.

Over the years, many interactions are being discovered and cell types re-defined and described in their context. However, lots of mechanisms and interactions of TME remains unknown due to very heterogeneous nature of this micro environment. This leaves room to more extensive investigation of TME.

A therapeutic goal are target interactions that would be able to pivot the essential processes in tumorigenesis or tumor escape in order to put the cells “back on track” and facilitate anti-tumor therapies.

These goals can be meet thanks to improvement of investigation methods, data quality and abundance. I will discuss the most important data types used in this project to investigate the TME.

1.2 Quantifying and qualifying immune infiltration (data)

Nowadays, more and more biological data is produced. However, this proliferation of accessible resources is not proportional to generated insights and wisdom. In this thesis, I aim to generate *Knowledge* and *Insights* and we hope to generate some *Wisdom* (Fig. 1.4). In this section, we will introduce the foundation of our analysis: different data types that will be further discussed and explored in chapters that follow.

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

1.2.1 Cell sorting

1.2.1.1 Flow cytometry

Flow cytometry is a laser-based technology. It uses marker genes: cell surface proteins to sort cells in different compartments. Nowadays, it permits quantification of the abundance of up to 17 cell surface proteins using fluorescently labelled antibodies [159]. However this techniques is not free from bias, our knowledge about cell markers is limited and several markers may not be relevant in some context. Moreover, the scientific community did not clearly agree on the marker choice even for popular and well studied cell types which introduced additional heterogeneity when independent studies are compared. Also the quality of antibodies may influence the results of the FACs analysis. Besides those limitations FACs remains quite popular method for analysing cells in complex tissues. It was among first methods that allowed molecular phenotyping of immune cells, a discovery of numerous subsets and thier further functional interpretation.

1.2.1.2 Mass cytometry

Mass cytometry (also known as CyTOF allows for the quantification of cellular protein levels by using isotopes. It allows to quantify up to 40 proteins per cell [159]. It also demands lower starting number of cells (1000 - 1000000), a realistic number that can be extracted from patient biopsy [126].

1.2.2 Microscope Staining

Using microscope technics, histopathological cuts are analysed. The number of cells per a unit of area (i.e. mm^2) is defined either manually by human or though diverse image analysis algorithms.

Current pathology practice utilises chromogenic immunohistochemistry (IHC) [144]. Multiplexed approaches allow to identify multiple markers in the same histopathology cut. Modern techniques as imaging mass cytometry using FFPE tissue samples uses fluorescence and mass cytometry to identify and quantify marker proteins [73].

The main advantage of aforementioned technics the number of cells that can be analysed and the information about spatial distribution of the different cell types. The limiting factor, as for cell sorting methods, is the number of markers (~10-100) and consequently number of cell types that can be identified [184].

The cell sorting methods and microscope staining are usually considered as a gold standard for multidimensional data techniques. The reason why they are not applied at large

scale is the cost but also quite laborious and time consuming sample preparation demanding a fresh sample. In contrast, the -omics methods propose more scalable way to measure tumor micro environment.

1.2.2.1 Tissue Microarrays

Tissue Microarrays aim to automatize "staining" techniques. A large number of small tissue segments can be organised in a single paraffin block where 100 tissue samples can be easily examined on one slide. A variety of molecular or microscopic method can be then applied to FFPE tissue including immunohistochemistry, FISH, and in situ hybridization [226]. It is a technique in between traditional imaging and omic high-throughput.

1.2.3 omics

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

1.2.3.1 Transcriptome

Transcriptomics measures the number of counts of mRNA molecules using high-throughput techniques. mRNA is the part of genetic information that should be translated to proteins. It reflects the activity of ongoing processes in a cell. In contrast to DNA, mRNA concentration can be highly variable [216]. This variability can be either "intrinsic" that reflect the stochastic process of cell machinery or "extrinsic" reflecting impact of factors upstream to mRNA synthesis [183].

Transcriptome can be measured by microarrays or RNA-seq NGS technology. Microarrays remain cost-efficient and popular technique designed in 90. There exist two and one color fluorescent probes, both representing different challenges in experimental design for batch effect removal. RNA-seq, in contrast, uses sequenced RNA to quantify the

expression. As not only selected genes (probes) are quantified it can be used to study unknown parts of the genome. RNA-seq is also characterised by lower background noise than microarrays.

Bulk transcriptome data are quite accessible nowadays. They can be obtained from either flash-frozen or formalin-fixed, paraffin-embedded (FFPE) tissue samples, including both surgically resected material and core needle biopsies [184].

The main flaw of transcriptomic data is that the reproducibility between different platforms is limited. As a result, direct comparison (direct merging, statistical difference tests) between two datasets produced by different platforms is not advised. There are 12 thousands genes that are matching between four sequencing platforms. Through gene names conversions a lot of information is lost and bias is introduced.

Different strategies can be adapted to analyse bulk transcriptome.

Cieślik and Chinnaiyan [40] describes five groups of most popular approaches that can be applied to study transcriptome (Fig. 1.5). Despite a diversity of bioinformatic and statistical tools, the most popular differential approaches, mainly differential gene expression (DGE) based on difference between two experimental conditions.

RNA-seq data was proven to be a useful indicator for clinical applications [134, 145, 177]. Its utility for immune profiling was demonstrated in may studies through a use of transcriptomic signatures to predict immunotherapy response or survival [34].

In this work transcriptome data analysis fails into multiple categories: Compositional, Relative and aims to construct a Global-level conclusions.

1.2.3.2 Single cell RNA-seq

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

Continuous discovery of new immune subtypes has proven that cell surface markers that are used for phenotyping by techniques like FACS and immunohistochelistry cannot capture the full complexity. ScRNA-seq methods allow to cluster known cell types in subpopulations based on their genetic features. ScRNA-seq is also able to capture particularly rare cell types as it requires much less of RNA material (1 ng isolated from 100-1000 cells) compared to ‘bulk’ RNA-seq (~ 1 µg of total mRNA transcripts). It also allows to study cells at high resolution capturing the phenotypes in much more refined scale than previously [159].

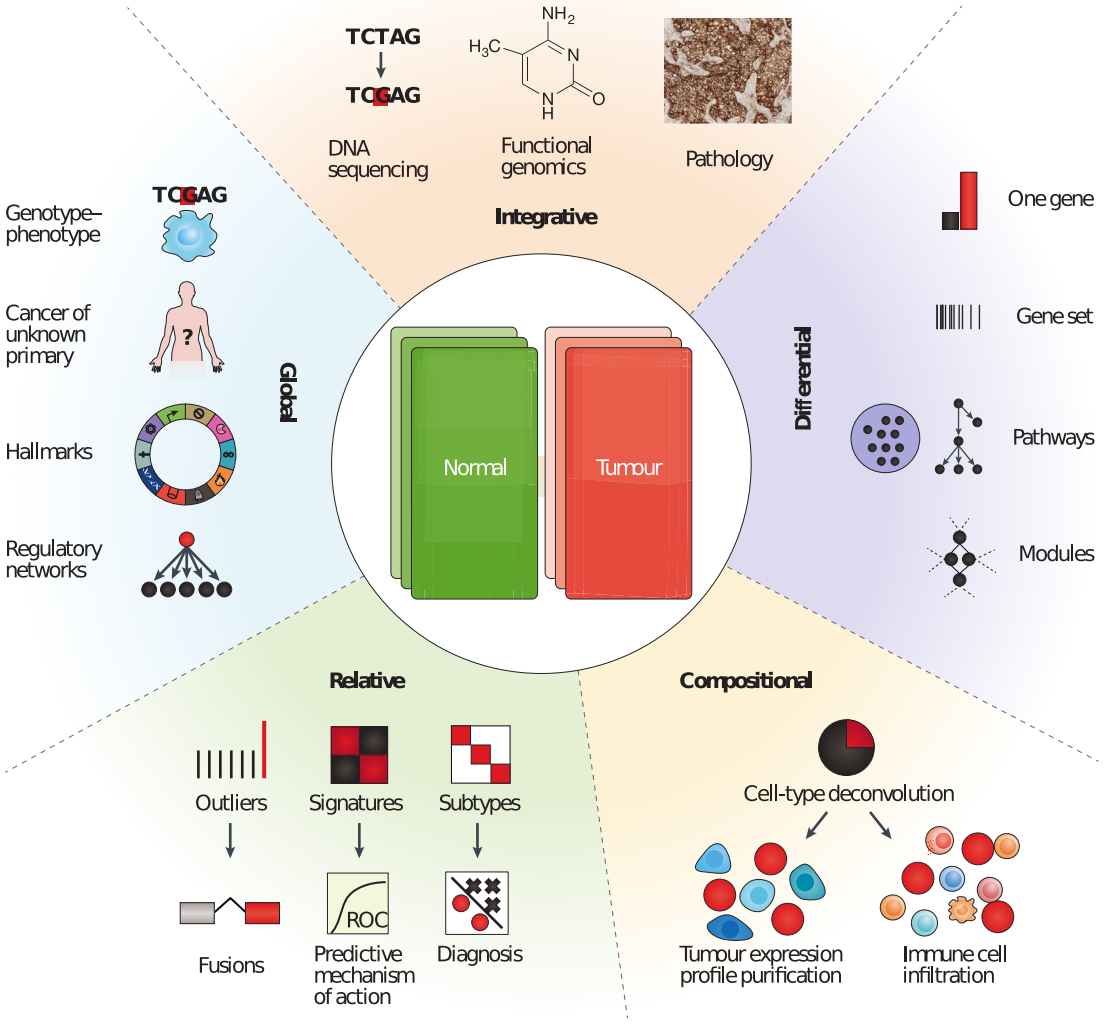


Figure 1.5: Five categories of RNA-seq data analysis. Differential analyses: comparing two (or more) conditions, Relative analyses: comparing to an internal refernce (average, base level), Compositional analyses: inferring cell types or groups of cell types (i.e. tumor purity), Global analyses: pan-tissue and pan-cancer analyses and Integrative analyses: compiling heterogneious data types. Reprinted by permission from Springer Nature [40] © 2018 Macmillian Publishers Limited, part of Springler Nature. All rights reserved.

This new data type also brings into the field new challenges related to data processing due to the volume, distribution, noise, and biases. Experts highlight as the most "batch effect", "noise" and "dropout effect" [162]. So far, there are no official standards that can be applied which makes data comparison and post-processing even more challenging. Up to date, there are around 70 reported tools and resources for single cell data processing [47] . A limited number of single-cell datasets of tumors are made publicly available and more are to come.

One can ask why then developing computational deconvolution of bulk transcriptome if we can learn relevant information from single-cell data. Firstly, that single cell data do not provide a straightforward answer to the estimation of cell proportions. The coverage is not full and sequenced single cells are not fully representative of the true population. For instance, neutrophiles are not found in scRNA-seq data because of they are "difficult to isolate, highly labile *ex vivo* and therefore difficult to preserve with current single-cell methods" [184]. In addition, a number of patients included in published studies of range <100 cannot be compared to thousand people cohorts sequenced with bulk transcriptome methods. This is mostly because single cell experiments are challenging to perform, especially in clinical setting as fresh samples are needed [184]. Today, single cell technology brings very interesting "zoom in" perspective, but it would be incautious to make fundings from a restricted group of individuals universal to the whole population. Major brake to the use of single cell technology more broadly might be as well the price that is nearly 10x higher for single cell sample compared to bulk [42].

In this work, we are using single cell data in two ways. Firstly, in Chapter 5 we compare immune cell profiles defined by scRNA-seq, blood and blind deconvolution (problem introduced in Immune signatures section). Secondly, in Chapter 6 we use single call data of Metastatic melanoma generated by Tirosh et al. [209] to demonstrate heterogeneity of subpopulations of Macrophages and NK cells.

1.2.3.3 Epigenome

An epigenome can be defined as a record of the chemical changes to the **DNA and histone proteins** of an organism. Changes to the epigenome can provoke changes to the structure of chromatin and changes to the function of the genome [20]. Epigenome data usually contains information about methylation **CpG island changes**. In cancer, global genomic hypomethylation, CpG island promoter hypermethylation of tumor suppressor genes, an altered histone code for critical genes, a global loss of monoacetylated and trimethylated histone H4 were observed. Methylome profiles can be also use as molecular signature of disease and potential diagnostic or predictive biomarker [101].

1.2.3.4 Copy number variation (CNV) and Copy number aberration (CNA)

The differences between human genome comes in majority from **Copy Number Variation** [129]. CNV regions constitute 4.8–9.7% of the whole human genome [232]. They can be reflected in structural variation that are duplication or a deletion of DNA bases. CNV can affects a lot of base pairs of DNA code (deletion of more than 100 genes) and result in a phenotype change.

In addition, there can be distinguished, **Copy number alterations/aberrations (CNAs)** that are changes in copy number that have arisen in **somatic** tissue (for example, just in a tumor), in contrast to CNV that originated from changes in copy number in **germline** cells (and are thus in all cells of the organism) [129]. CNV and CNA profiles can be associated with diseases or cancer subtypes.

There exist disease-related exome panels that focus on regions with high copy variation or the full exome can be sequenced using whole-exome sequencing (WES) [228].

1.2.3.5 Spatial transcriptomics

> Spatial transcriptomics provides quantitative gene expression data and visualization of the distribution of mRNAs within tissue sections and enables novel types of bioinformatics analyses, valuable in research and diagnostics [195]

It combines RNA-seq technology with spatial labelling which allows to have a bulk gene expression of 10-20 cells with given space coordinates within the sample. It allows to localize regions of highest gene expression and perform *Spatially Variable Genes* (Svensson et al. [200]). Some attempts were already made to combine Spatial Transcriptomics and scRNA-seq [137]. It remains an early-stage technique and so far it is not widely used but it might be a future of omics to add a spatial information as it can be essential for many research problems.

1.3 From cancer phenotyping to immune therapies

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

1.3.1 Cancer immune phenotypes

Since 20s century physicians decided on common nomenclature that classify tumors into distinct groups that are relatively homogenous or that share common characteristic important for treatment and prognosis. Tumor typing should help to better predict prognosis, to adapt a therapy to the clinical situation, to enable therapeutic studies which are essential in proving any therapeutic progress.

Most of the classifications are based on clinical data. Most common factors taken into account are: the degree of local invasion, the degree of remote invasion, histological types of cancer with specific grading for each type of cancer, possibly various tumour markers, general status of the patient.

However, cancers with similar morphological and histopathological features reveal very distinct patterns of progression and response to therapy [68]. In the era of gene sequencing, gene and protein expression as well as epigenome can provide an important complementary information. Therefore gene markers or proteomic abnormalities can integrated into classification panel. One popular example is a gene signature *PAM50* [160] used for prediction of patients' prognosis in breast cancer, patented as a tumor profiling test.

Since the increase of importance of the immunotherapies, researches proposed several ways to classify tumors based on their microenvironment. Given different parameters describing TME, cancers can be sorted into groups that show similar characteristics. We will discuss most common frameworks that allow to phenotype cancers based on the TME.

The localisation of the immune cells can be an indicator of the state and response to the therapy [22].

The most standard approach is to convey an analysis of histopathological cuts to asses the number of infiltrating lymphocytes (TILs). Two typical patterns are usually identified: "hot" - immune inflamed and "cold" - no active immune response [19].

Chen and Mellman [33] describe classification into inflamed and non-inflamed tumors, where non-inlamed phenotypes: can be further split into the immune-desert phenotype and the immune-excluded phenotype (Fig. 1.6). The inflamed phenotype is characterised by rich presence of immune cells : T cells, myeloid cells, monocytes in tumor margin. Along with the immune cels, due to their communication, a high expression of cytokines is characteristic for this phenotype. According to Chen and Mellman [33], this is a mark that an anti-tumor response was arrested by tumor. The inflamed phenotype has shown to be most responsive to immunotherapies. In the immune-excluded phenotype, the immune cells are present as well but located in the stroma [91], sometimes penetrating inside tumor. However, when exposed to check point immuotherapy, T cells

does not gain the ability to infiltrate the tumor, therefore the treatment is inefficient. The immune-desert main features is little or no presence of immune cells, especially T cells. Surprisingly, this tumors have been to proven to rarely respond to the checkpoint therapy [91]. In non-inflamed tumours cytokines associated with immune suppression or tolerance are expressed.

A presence of immune phenotypes was confirmed by for example by Becht et al. [16] in colorectal cancer, where after deconvolution of bulk tumor profiles, pattern of immune and stroma cells abundance was matching four cancer subtypes. The good prognosis was related to cytotoxic response and bad prognosis to lymphocytes and cells of monocytic origin.

According to Gajewski et al. [66], the immunogenicity of the tumors can be explained by tumor-intrinsic factors and tumor-extrinsic factors. Tumor-intrinsic factors are: the neoantigen load and frequency, the mutational load, the expression of immunoinhibitors and immunostimulators (e.i. PD-L1), and alteration of HLA class I molecules. Tumor-extrinsic factors include chemokines regulating T cell trafficking, infiltration of effector TILs and immunosuppressive TILs, and soluble immunomodulatory factors (cytokines).

1.3.2 Scoring the immune infiltration

Experimental techniques and computational tools enabled us to characterize and classify TME with multi-omics data. Here I present a short list of most influencing and complete analysis aiming to redefine tumor phenotypes based on the immune infiltration, with a focus on computational techniques.

1.3.2.1 Immunoscore

One of the most recognised scoring method, based on fluorescent images is authored by Jerôme Galon lab in Paris and names Immunoscore. The Immunoscore ranges from 0 to 4 and it is based on the density of lymphocyte populations CD3/CD45RO, CD3/CD8, or CD8/ CD45RO. It also takes into account the spacial position of the cells: the tumor core and margins [67]. It was successfully applied to colorectal cancer to predict patients' survival [7] . Since it resulted in numerous application to many cancer types. Immunoscore has been recently validated in big cohort international independent study (14 centres in 13 countries) as a relevant prognostic score of time to recurrence, defined as time from surgery to disease recurrence [156].

The immunoscore is an interesting indicator, especially in the scope of clinical applications, although it does not tell us a lot about underlying biology. It is also limited to a

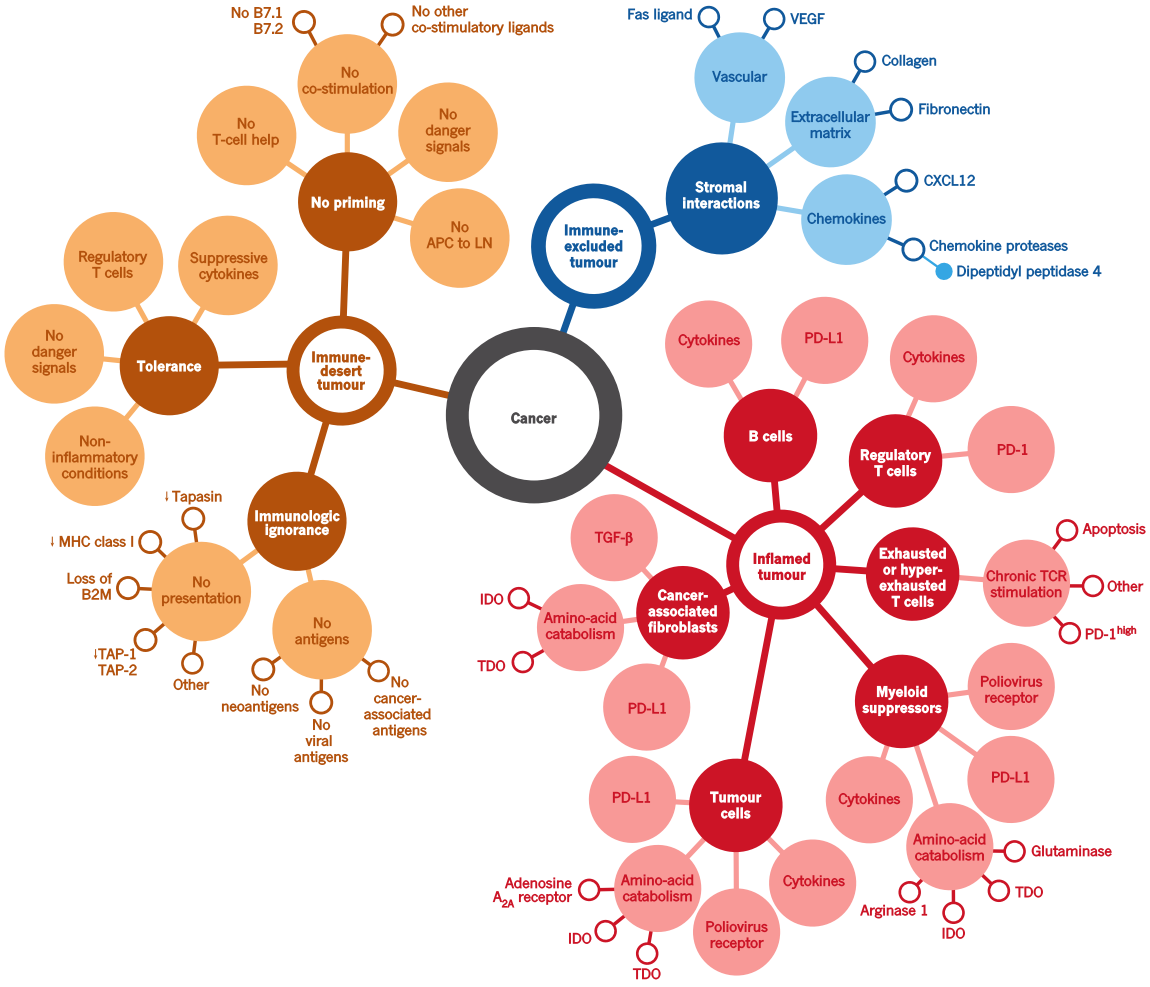


Figure 1.6: Cancer-immune phenotypes: the immune-desert phenotype (brown), the immune-excluded phenotype (blue) and the inflamed phenotype (red). The immune-desert phenotype is characterised by paucity of immune cells and cytokines. In the immune-excluded phenotypes the T cells are often present but trapped in stroma, enabled to migrate to the tumor site. The immune-inflamed phenotype is rich in immune cells and the most responsive to the immune check point therapies. Reprinted by permission from Springer Nature [33] © 2017 Macmillian Publishers Limited, part of Springler Nature. All rights reserved.

few cell types while it may be that in some cancer types or patients, the system requires more detailed or rich analysis of larger panel of cells.

1.3.2.2 Spatiotemporal dynamics of Intratumoral Immune Cells of Colorectal Cancer

Bindea et al. [22] published very complete, and supported with strong experimental evidence, immune landscape of colorectal cancer. Authors introduced *the immunome compendium* containing 577 cell-type specific genes, derived from analysis of big corpus of publicly available data. They used it to analyse CRC large transcriptomic data (105 patients). Using qPCR (more sensitive technique than microarray) expression of 81 "representative" genes from the compendium was investigated in 153 CRC patients. This study validated correlation of markers of the same type and also revealed correlation of different cell-type markers (i.e. Tcells and NK or Th and macrophages). The data matrix was grouped into 3 clusters which were corresponding to 1) tumor 2) adaptive 3) innate imune repsonses. In addition spatial positioning of markers was visualized thanks to Tissue Microarray technology in samples from 107 CRC patients distinguishing marker densities in tumor center and tumor margin areas. This was followed by in deep study of chemokines expression and genomic alterations. In addition, authors validated potential prognostic biomarkers in murine orthotopic CRC models.

In summary, using marker genes measured and visualized with different data types of CRC, a high inter-patient heterogeneity was observed. It was confirmed that the immune landscape evolves over time (tumor stages). Adaptive immunity cells were associated with core of the tumor and the innate ones with the tumor margin. A mechanism involving CXCL13, Tfh cells, B cells and IL-21 was identified as associated with good prognosis.

1.3.2.3 Immunophenoscore

Different approaches, sub-typing oriented, are based essentially on gene expression patterns. Most commonly, machine learning supervised algorithms are trained to match known phenotype (established with microscopy or with clinical features) to genetic patterns or an unsupervised clustering is used to discover new classification.

An example of well-formulated classification framework is Immunophenoscore [32], based on publication of Angelova et al. [6], where methylome, transcriptome and mutation of TCGA CRC dataset (n = 598) was used to describe *immunophenotypes*. Later on, it was reduced to gene expression indicator and summarised in a form of a score. This scoring scheme is based on the data of 20 solid tumors, using expression of marker genes selected by a machine learning algorithm (random forest) for best prediction in each cancer. These indicators can be grouped into four categories:

Table 1.1: Six immunological subtypes of cancer. General characteristic of subtypes generated by Thorsson et al. [208] as described in the original publication.

Cluster	Features	Macrophage..lymphocyte	Th1.Th2	Proliferation	Intratumoral.heterogeneity	Other
C1	Wound healing	Balanced	Low	High	High	
C2	IFN-γ dominant	Lowest	Lowest	High	Highest	Highest M1 and CD8 T cells
C3	Inflammatory	Balanced	High	Low	Lowest	Highest Th17
C4	Lymphocyte depleted	High	Minimal Th	Moderate	Moderate	
C5	Immunologically quiet	Highest	Minimal Th	Low	Low	Highest M2
C6	TGF-β dominant	High	Balanced	Moderate	Moderate	Highest TGF-□ signature

- MHC molecules (MHC)
- Immunomodulators (CP)
- Effector cells (EC)
- Suppressor cells (SC)

The immunophenscore (IPS) is calculated on a 0-10 scale based on the expression of genes in each category. Stimulatory factors (cell types) impact the score positively and inhibitory factors (cell types) negatively. Z-scores ≥ 3 were designated as IPS10 and z-scores ≤ 0 are designated as IPS0. A similar conceptual framework called *cancer immunogram* was proposed by Blank et al. [24] included seven parameters: tumor foreignness (Mutational load), general immune status (Lymphocyte count), immune cell infiltration (Intratumoral T cells), absence of checkpoints (PD-L1), absence of soluble inhibitors (IL–6, CRP), absence of inhibitory tumor metabolism (LDH, glucose utilisation), tumor sensitivity to immune effectors (MHC expression, IFN-□ sensitivity). Charoentong et al. [32] claim that the immunophenoscore can predict response to CTLA-4 and anti-PD-1.

Nonetheless, the details of the use of *cancer immunogram* in practice remain unclear and result could be sensitive to patients' and data heterogeneity as no standardisation was proposed. It should be also validated in a systematic independent study.

1.3.2.4 The immune landscape of cancer

Thorsson et al. [208] performed a multi-omic analysis of TCGA datasets that allowed them to define 6 subtypes that are valid across cancer types (see Tab. 1.1).

Authors selected eight indicators to define these six phenotypes:

1. differences in macrophage or lymphocyte signatures
2. Th1:Th2 cell ratio
3. extent of intratumoral heterogeneity
4. aneuploidy
5. extent of neoantigen load
6. overall cell proliferation
7. expression of immunomodulatory genes

8. prognosis

These indicators were selected among many other indicators though machine learning (elastic net regression) for the best predictive power of survival.

All the data and computed parameters can be accessed at CRI iAtlas Portal. Among the six phenotypes C3 (Inflammatory) has the best associated prognosis while C1 (wound healing) and C2 (IFN-γ dominant), much less favourable outcome. This again illustrates the ambivalent nature of the immune system as the best and the worst prognosis are associated with immunologically active tumors. C4 (lymphocyte depleted) and C6 (TGF-β dominant) subtypes had the worst prognosis. The content of immune cells was determined using different tools and data types (expression, DNA methylation, images etc.) We can learn a lot from the study, however, it seems difficult to integrate the methods to an ordinary practice because different data levels are necessary for the same samples to compute all the indicators.

1.3.2.5 A pan-cancer landscape of immune-cancer interactions in solid tumors

A different classification was proposed by Tamborero et al. [202], also using TCGA data. They distinguished 17 immune infiltration patters based on the immune cell proportions and 6 different clusters based on cytotoxicity measure across all cancer types (named immune-phenotypes) that were finally summarized in three groups: cytotoxic immune infiltrate, infiltrate with more immune-suppressive component and poor immune infiltrate. According to the analysis, one of the most important factors is cytotoxicity. Tumors with high cytotoxicity were characterized by low clonal heterogeneity, with gene alterations regulating epigenetic, antigen presentation and cell-cell communication. The medium-level cytotoxic tumors had activated invasion and remodelling of adjacent tissue, probably favourable to immune-suppressive cells. The low cytotoxicity subgroup of tumors had altered: cell-cycle, hedgehog, β- catenin and TGF-β pathways. This result roughly overlaps with the one of Thorsson et al. [208]. The survival analysis based on the 6 immune-phenotypes revealed that for most cancer types, high cytotoxic tumors are associated with better survival. To evaluate tumor environment cells authors used gene set variation analysis [86] with a set of pre-defined cell-type markers. Another important conclusion of Tamborero et al. [202] is that tissue of origin is not the only important factor shaping cell-type patterns in tumors. However, the least infiltrated tumors were lung, uterine and bladder cancers, while the most infiltrated were pancreatic, kidney, skin cancers and glioblastoma. They also analysed cancer cell pathways after computational purification of tumor samples (subtraction of the immune signal) in order to better understand cancer signalling.

A different approach, is to characterize tumors based on signaling pathways organized in functional modules.

1.3.2.6 Immune maps

Another way to summarize tumor phenotype can be though use of molecular maps. Atlas of Cancer Signaling Network (ACSN) [108, 107] is a pathway database that contains a collection of interconnected cancer-related signalling network maps. An additional feature is ACSN web-based Google-maps-like visualisation of the database. User data can be projected on the molecular map (for example gene/protein expression from user data can be paired with entities on the map). ACSN 2.0 contains Cancer cell map and TME map (at the time: angiogenesis, innate immune map, T-cell signaling maps). All separate maps are available in Navicell website. Through projection of the data on the innate immunity map, one can see if a tumor sample is characterised by pro- or anti-tumor activated pathways due to the organisation of the map layout. Also, different CAF subtypes were characterised with the CAF specific map in [44]. Kondratova and colleagues (including myself) used innate immune map to characterize NK and Macrophages subtypes (see Chapter Z).

1.3.2.7 Summary

Despite all scientific efforts, the gene expression based classifications are not yet used in clinics. The measured multi-panel mRNA expression, that can be included into category of In Vitro Diagnostic Multivariate Index Assay (IVDMIA) [83, 178], may be a future of TME-based cancer classification, diagnosis and treatment recommendation [74]. For this best tools need to be used to properly evaluate the state of TME and tumor-stroma-immune cells communication.

1.3.3 Immune signatures - biological perspective

A gene signature is

> *a single or combined group of genes in with a uniquely characteristic pattern of gene expression that occurs as a result of an altered or unaltered biological process or pathogenic medical condition* [99, 122].

They can be classified based on their form:

- metagene
- gene list
- weighted gene list

A term **metagene** or *eigen gene* describes an aggregated pattern of gene expression. The aggregation can correspond to simple mean of samples or can be obtained though

matrix factorisation or source separation techniques, clustering. A metagene usually provides values for all measured genes (all probes) in contrast to a wighted gene list where weights are associated with selected genes.

Gene lists are simple enumeration of transcripts names or gene identifiers. Application of gene list is often limited to gene enrichment analysis tools or gene selection from the data.

An alternative is a **weighted gene list** or ranked gene list, where genes are ranked according to their importance. Often the ranks is obtained though comparison between two conditions or test/control. They can be also based on absolute gene expression values[126]. One possible problem with this weighted gene list can be platform dependence.

There exist a big choice of databases storing collections of signatures. They contain gene expression and other genomic data such as genotype, DNA methylation, and protein expression data attributed to some condition of reference. A big collection of immune signatures are regrouped by Immunological Genome Project (IGP, ImmGen) [89]. Gene expression of protein coding genes measure in mice immune cells, ex vivo, in different conditions (drug treatment, perturbations) were regrouped in this ressource. A different ressource Immuno-navigator [215] that stores information about human and murine immune genes and co-expression networks. ImmuneSigDB is a collection of gene-sets that describe immunity and inflammation in transcriptomic data [75] and a part of popular MSigDB ressource used commonly for gene set enrichment analysis (GSEA) [198].

They can also be classified based on their use:

- prognostic signatures
- predictive signatures
- diagnostic signatures
- specific signatures

The *prognostic* signatures can distinguish between patients with a good or from patients with bad prognosis when deciding to assign a patient to a therapy.

The *predictive* signatures are able to predict treatment benefit between experimental and/or nontraditional treatment groups vs. control, i.e. in clinical trials [131].

The *diagnostic* signature, also called *biomarkers* can be used for detection of a disease in a patient, like for example in blood tests.

The *specific* signatures should describe with robustness and reproducibility the same group of cells, or patients, or condition with respect to other considered groups. For instance, in the context of cell-types, among studied cell-types a specific signature will distinguish only one cell type. In the context of cancer subtypes, it will indicate clearly

one subtype among others.

Examples of predictive and prognostic gene signatures, used in clinical practice are Oncotype DX, EndoPredict, PAM50, and Breast Cancer Index for breast cancer [87].

Studies discussed in this Chapter showed plausible importance of immune-related signals in cancer therapy. However, there is no no immune-related gene signatures used in clinical practice currently. This can be because of the lack of consistency of genes, both within the same tumor type and among different tumors that can be found in the signatures [38]. Difference in gene expression of different cell populations were found even intra- and interlabs. This difference can be due to confounding factors like stress or to contamination [89].

In many studies *specifc* signatures of cell types are used. They seem to be good in discriminating between broad lineages of cell type, such as lymphoid and myeloid. Although thier capacity to describe cell states and cell subtypes is more discutable [38]. Another matter is that cell type signatures are often obtained in model organisms or extracted from different tissue (i.e. blood-derived signatures vs cancer-derived signatures).

> *the gene expression profiles of tumour-associated immune cells differ considerably from those of blood derived immune cells* [184]

With emergence of single-cell signatures, there are new horizons of gene signatures to be discovered. Especially signatures of rare cell types in solid tissues. Yet, it is up to researches to cross validate single cell signatures with different types of data as scRNA-seq is not free of platform and post-processing bias.

Immune signatures will be also discussed as a part of deconvolution pipeline in the Chapter 2 under the section about *basis matrix* in mathematical terms.

1.3.4 Cancer therapies

Cancer is a complex disease. Up to date, no uniform and fully effective treatment was proposed and usually different strategies are tested to kill tumor cells. **Surgery** is one of the oldest methods. The cancer is removed from the patient body. There are different ways, more or less invasive, that it can be performed. it is usually applied for solid tumor contained in a small area. **Radiation Therapy** uses high doses of radiation to eliminate tumor cells and shrink tumor mass. It can be applied externally or internally. **Chemotherapy** uses a drug (or a combination of drugs) that kill cancer cells, usually altering cell proliferation and growth. The drawback of radiotherapy and chemotherapy are strong side effects. **Hormone therapy** modulate hormone levels in the body in order to inhibit tumor growth in breast and prostate cancers. In leukemia and lymphoma, can be applied

stem cell transplants that restore blood-forming stem cells destroyed by the very high doses of chemotherapy or radiation therapy that are used to treat certain cancers.

Alternatively, **targeted therapies** represent more focused strategy that aims to be more effective and cause less side effects than systematic therapies. Two main types of targeted therapies are small-molecule drugs and monoclonal antibodies. Targeted therapies usually aim to stimulate/inhibit a selected molecular function. A special types of targeted therapies are **Immunotherapies**. Through adtivation/inhibition of immune regulatory pathways, it stimulates immune system to destroy malignant cells. A continuation of targeted therapies is **precision medecine approach**. It is based on genetic information to specify patient's profile and find adapted treatment. A number of innovative treatments targeting a specific change in tumor ecosystem are being tested presently in precision medicine clinical trials [98].

1.3.5 Recent progress in immuno-therapies

The immunotherapies, in contrast with other types of cancers therapies discussed in the previous section, aim to trigger or restart the immune system to defend the organism and attack the malignant cells without provoking persisting inflammation state [166]

The idea of stimulating immune system to fight malignant cell was not born recently. Since a long time a possibility of development of an anti-cancer vaccine has been investigated. Unfortunately, this idea faced two important limitations 1) lack of knowledge of antigens that should be used in vaccine to successfully stimulate cytotoxic T cells 2) the ability of cancer to block the immune response also called *immunostat*. Despite those impediments works on anti-tumor vaccines do not cease [158]. A very recent promising an in-situ ani-tumor vaccine was proposed by Sagiv-Barfi et al. [182]. The therapy tested in mice, would be based on local injections of the combination of "unmethylated CG-enriched oligodeoxynucleotide (Cp-G) - a Toll-like receptor 9 (TLR9) ligand and anti-OX40 antibody. Low doses of CpG injected into a tumor induce the expression of OX40 on CD4+ T cells in the microenvironment in mouse or human tumors. An agonistic anti-OX40 antibody can then trigger a T cell immune response, which is specific to the antigens of the injected tumor". Sagiv-Barfi et al. claim this therapy could be applied to all tumor types, as long as they are leucocyte-infiltrated. As a local therapy, in situ vaccination should have less side-effects than systematic administration. It is now undergoing clinical trials to test its efficiency in human patients.

Another idea involving using immune system as a weapon to fight cancer, would be the use of genetically modified patient's T-cells, carrying *CARs* (chimeric antigen receptors) [100]. After a long period of small unsuccessful trials, recently in 2017, two CAR T-cell therapies were accepted, one to "treat adults with certain type of large B-cell lymphoma"

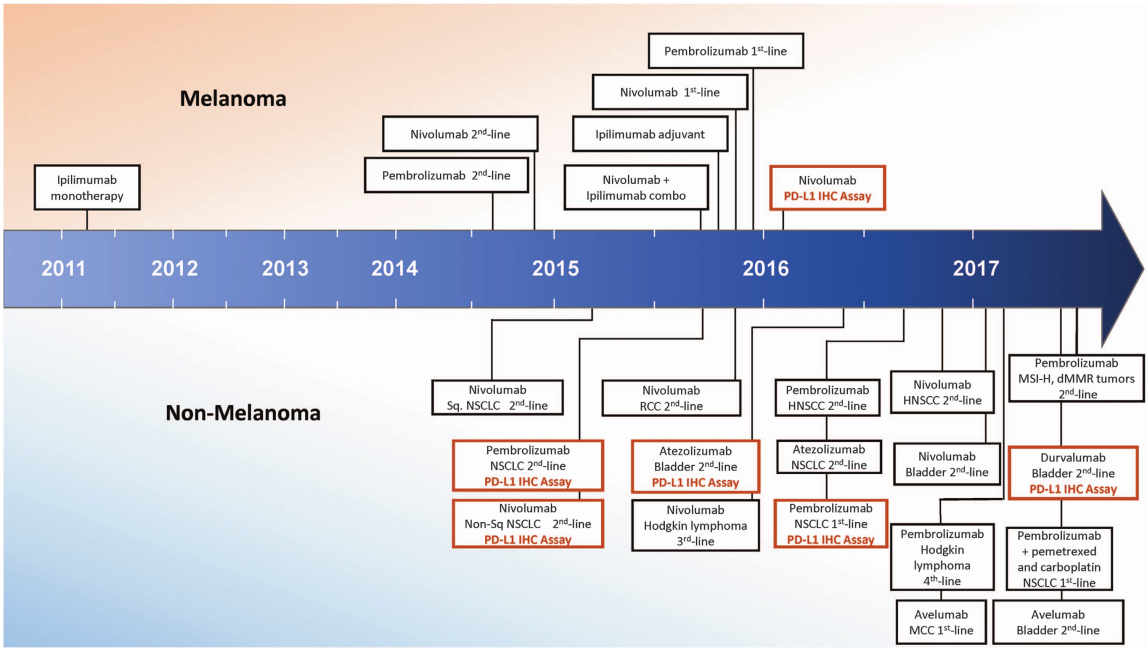


Figure 1.7: This timeline describes short history of FDA approval of checkpoint blocking immunotherapies up to 2017. Reprinted by permission from Springer Nature [203] Macmillan Publishers Limited, part of Springer Nature. All Rights Reserved.

[151], other to treat "children with acute lymphoblastic leukemia (ALL)" [150] , which are, at the same time, the first two gene therapies accepted by FDA.

However, the two most promising immuno-related strategies with proven clinical efficiency are based on blocking so called immune check point inhibitors: cytotoxic T-lymphocyte protein 4 (CTLA4) and programmed cell death protein 1 (PD-1). The anti-CLTA4 antibodies blocks repressive action of CLTA4 on T-cells and they become therefore activated. It was shown efficient in melanoma patients and accepted by FDA in 2015 as adjuvant therapy for stage III metastatic melanoma patients [147]. PD-1 is a cell surface receptor of T cells, that binds to PD-L1/PD-L2. After binding, an immunosuppressive pathway is activated and T cells activity is dampened. An action of an anti-PD-L1 antibody is to prevent this immune exhaustion [33]. A stepping stone for anti-PD-L1 therapies was approval of Tecentriq (atezolizumab) for Bladder cancer [148] and anit-PD1 Keytruda (pembrolizumab) initially accepted for NSCLC and further extended to head and neck cancer, Hodgkin's lymphoma, gastric cancer and microsatellite instability-high cancer [149]. Since other anti-PD-L1 or anti-PD1 antibodies were accepted or entered advanced stages of clinical trials [227]. A short history of immunotherapy FDA-accepted treatments can be found in Fig. 1.7

The main drawback of immunotherapies is a heterogeneity of response rate, which can

vary i.e. from 10–40% in case of PD-L1blocking [238], suggesting that some patients can have more chances than others to respond to an immune therapy. So far, it has been shown that anti PD-L1 therapies works more effectively in T cell infiltrated tumors with exclusion of Tregs because of lack of difference in expression of FOXP3 in responding and non-responding group of patients [91]. Also some light has been shade by Rizvi et al. [176] who connected mutational rate of cancer cells to the chances of response to an immunotherapy.

Despite those fundings, the precise qualifications of patients that should be sensitive to an immunotherapy are not defined [164]. As most patients do not answer to immunotherapies, it stimulates researches to look for better biomarkers and patient stratifications, and pharmaceutical industries to discover new immune checkpoints based therapies.

1.4 Summary of the chapter

Cancer remains an important health problem of our era that touches many people. Tumor cells are interacting with their mircoenvironement (called Tumor Microenvironment (TME)) including normal cell, stroma cells and a variety of immune cells. These cells can have a role in disease progression and response to treatment. A modern approach to modulate TME was proposed through application of immune therapies.

Many researches aim to link the composition and state of TME with patients clinical features and survival. It has been shown that in some cases certain cell types are beneficial to tumor development and some are not. However a case by case approach of personalized medecine may be necessary to fully understand the inter-patient heterogeneity.

A new way to classify cancers based on their TME is called immunophenotyping. There is no well establish procedures to perform it yet, but it can be integrated in the near future into the clinical classification of tumors. It is a subject of active research to find new biomarkers and signatures of different factors gouvening the TME. Lot of interest is directed nowadays towards immune cells. Given that traditional cell-type definitions can be questioned in the cancer context, new cell states and functional subtypes are being redefined by researchers.

There are many experimental techniques, as for instance immune staining and FACS, that allow the in deep study of the immune system. They require an important preparation steps and fresh samples. Also, a limited number of variables can be observed through this techniques and knowledge-based hypotheses are necessary. On the other hand, high-throughput omic data allow to measure of the all system at the same time as they can measure all referenced units (genes/methylaton sites/ copy number aberra-

tions) from FFPE samples that can be stocked for a long time. A discovery-type studies are then favoured and new biomarkers can be discovered. Particularly suited for studying complex biological systems is scRNA-seq as it provides gene expression profile of each cell without a compulsory use of marker genes, indispensable in other techniques to define cell-types. However, this technique remains quite costly and it is not yet optimized.

Therefore to have very detailed system-level view of the TME with traditional experimental techniques an uncountable amount of work and resources would be necessary. Using omic techniques system approach is possible, however to embrace fully the data complexity a computational tools are needed. From data generation to analyzis different statistical and mathematical challenges need to faced before arriving to valid biological results and interpretations.

As I will present in the next chapter, in order to solve the problem of extraction of cell-type heterogeneity from cancer bulk omic data, a number of approaches was developed.

Chapter 2

Mathematical foundation of cell-type deconvolution of biological data

In the previous chapter I presented state-of-art of the current immuno-oncology research that has to embrace great complexity of cancer disease and the immune system. One part of this complexity can be explained by the presence and quantities of tumor-infiltrating immune cells, their interactions with each other and the tumor.

In this chapter, I will discuss how mathematical models can be used to extract information about different cell-types from 'bulk' omics data or how to de-mix mixed sources composing the bulk samples. To start with, I will introduce you to basic concepts of machine learning. Then I will focus on approaches adapted for cell-type deconvolution. In a literature overview, I will depict the evolution of the field as well as discuss the particularities of different tools for estimating presence and proportion of immune cells within cancer bulk omic data.

2.1 Introduction to supervised and unsupervised learning

Machine learning (ML) is a filed of computer science where system is able to learn and improve given an objective function and the data.

A popular definition of machine learning has been given by Mitchell in 1997:

> Machine learning: *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E*.
>
> — Mitchell in 1997 [133]

Term *Artificial intelligence* (AI) is often used by the media or general public to describe machine learning. Indeed ML can be considered as a branch of AI, together with computer vision and deep neural networks applications. However, commonly ML and AI are used interchangeably by the wide public.

ML is applied commonly in many fields of science and industry. I will not discuss here a subtle differences between machine learning, statistical learning, computational statistics and mathematical optimisation.

In general, algorithms can be divided into groups given the application:

- classification - aims to assign observations to a group (discrete variable)
- regression - aims to predict a continuous response of an input (continuous variable)
- clustering - aims to divide data into groups that are related to each other based on a distance

Another important distinction can be made given the inputs to the algorithm. Here, I present the differences between supervised and unsupervised learning.

2.1.1 Supervised learning

Supervised learning can be described as “the analysis of data via a focused structure” [163]. The main task is to predict an output given the inputs. In the statistical language, the inputs are often called the predictors or the independent variables. In the pattern recognition literature the term features is preferred. The outputs are called the responses, or the dependent variables. [88]

The initial data is divided into two sets: training and test. First the model is trained with correct answers on the training data (learning to minimise the error) , and then its performance is evaluated on the test data.

Among widely used classifiers there are Support Vector Machines (SVM), partition trees (and their extension random forests), and neural networks. For regression it is common to encounter linear regression, boosted trees regression,

2.1.2 Unsupervised learning

In Unsupervised learning is given the data and is asked to divide the data given a certain constraint. However, the correct division of the data is not known. Therefore an unsupervised algorithms aims to unveil the “hidden structure” of the data, or latent variables.

One group of unsupervised learning are descriptive statistic methods, such as: principal components, multidimensional scaling, self-organizing maps, and principal curves. This methods aim to represent to the data most adequately in low-dimensional space [88].

Another group are clustering algorithms. Clustering is the way to create groups (multiple convex regions) based on the intrinsic architecture of the data. These groups are not necessarily known beforehand, but can validated with the domain knowledge. Popular clustering algorithms are knn, k-means, hierarchical clustering.

In both descriptive statistics and clustering, one important parameter (often called k) is number to which we want to decompose the data (number of factors, variables, clusters). Different algorithms and applications can propose an automatic choice of k based on formal indexes or previous knowledge, in others, user need to provide the k.

2.1.3 Low-dimensional embedding for visualization

There is a common confusion, often seen in computational biology, between dimension reduction and clustering. This confusion is highly pronounced with, a popular in biology, algorithm: T-distributed Stochastic Neighbor Embedding (t-SNE) [212]. t-SNE works in 2 main steps: (1) a probability distribution over pairs of high-dimensional objects is computed in such a way that similar objects have a high probability of being picked, whilst dissimilar points have an extremely small probability of being picked, (2) t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map. It is not reliable to use t-SNE for clustering as it does not preserve distances. It can also easily overfit the data and uncover 'fake' or 'forced' patterns. Therefore, a clustering should not be applied to t-sne reduced data. An alternative to t-SNE method is recently published Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [130]– that is based on Laplacian eigenmaps, highly scalable, reproducible and recently applied to biological data [17]. Older used alternatives are ISOMAPS (non linear dimension reduction) or PCA (Principal components analysis). For any non-linear dimension reduction method, it is not recommended to use clustering *a posteriori*. Clusters should be computed on original data and then the cluster labels can be visualized in low-dimentional embedding.

2.2 Types of deconvolution

One specific application of mathematical/statistical tools is deconvolution of mixed signals.

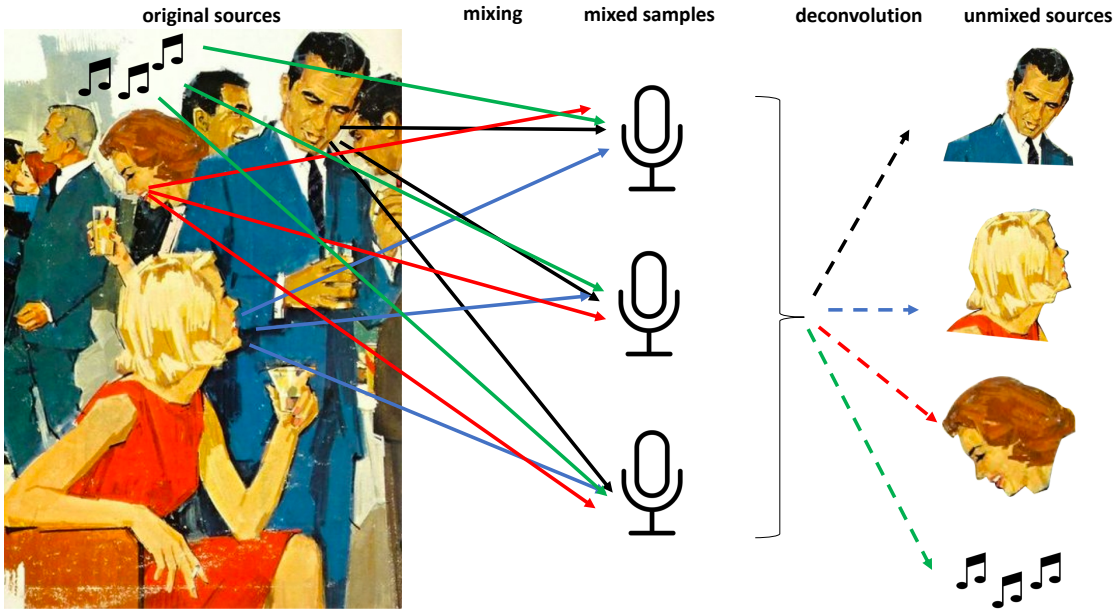


Figure 2.1: Illutration of the cocktail party problem. During a cocktail party voices of participalnts can be recorded with a set of microphones and then recovered though blind source separation. Fot the illustration purposes only four sources are mixed with three microphones, in reality the analysis can be performed with many sources. However, number of samples (microphones) should be higher than numer of sources (contrary to the illustration).

According to mathematical definition:

> Deconvolution : *the resolution of a convolution function into the functions from which it was formed in order to separate their effects*

Or in plain English:

> *a process of resolving something into its constituent elements or removing complication*

The similar problem of mixed sources can be encountered in other fields, i.e. signal processing, known also under the name of "**cocktail party problem**". In the cocktail party problem, at a party with many people and music, sound in recorded with several microphones. Through blind source separation, it is possible to separate the voices of different people and the musical background (Fig. 2.1) [37].

The same concept can be transposed to the bulk omic data, each biological species (like gene) is a cocktail party where each sample is a microphone that gathers mixed signals of different nature. The signals that form the mixtures can be different depending on the data type and scientific question asked.

In general, the total bulk data can be split into three abundance components [190]:

1. sample characteristic (disease, clinical features)
2. individual variation, genotype-specific or technical variation
3. presence and abundance of different cell types expressing set of characteristic genes

Many scientists invested their efforts in order to dissect the bulk omic data into interpretable biological components.

In scientific literature, there can be encountered three main understanding of tumor deconvolution:

- **estimating clonality**: using genomic data is it possible to trace tumor phylogeny raised from mutations and aberrations in tumor cells; therefore it is dissecting *intra*-tumor heterogeneity (i.e. using transcriptomic data [186], or more often CNA data (see Section 2.4.2)
- **estimating purity**: deconvolution into tumor and immune/stroma compartments, often aiming to "remove" not-tumor signal from the expression data, can be performed with different data types, the most reliable estimations are obtained usually from CNA data (see Section 2.4)
- **estimating cell-type** proportions and/or profiles from bulk omics data, most of works were performed on transcriptome data (see Section 2.3) and some on the methylome data (see Section 2.4.1)

These three types of deconvolution can be performed on the bulk omics data. Here we will focus on cell-type deconvolution models using bulk transcriptome. I will also briefly introduce deconvolution models applied to other data types (methylome and CNA).

2.3 Cell-type deconvolution of bulk transcriptomes

The idea of un-mixing the bulk omic profiles is documented to first appear in an article of Venet et al. [217] as a way to

> *infer the gene expression profile of the various cellular types (...) directly from the measurements taken on the whole sample*

In the basic hypothesis [3], mixture of signals from TME in transcriptomic samples can be described as a linear mixture.

$$X = SA \tag{2.1}$$

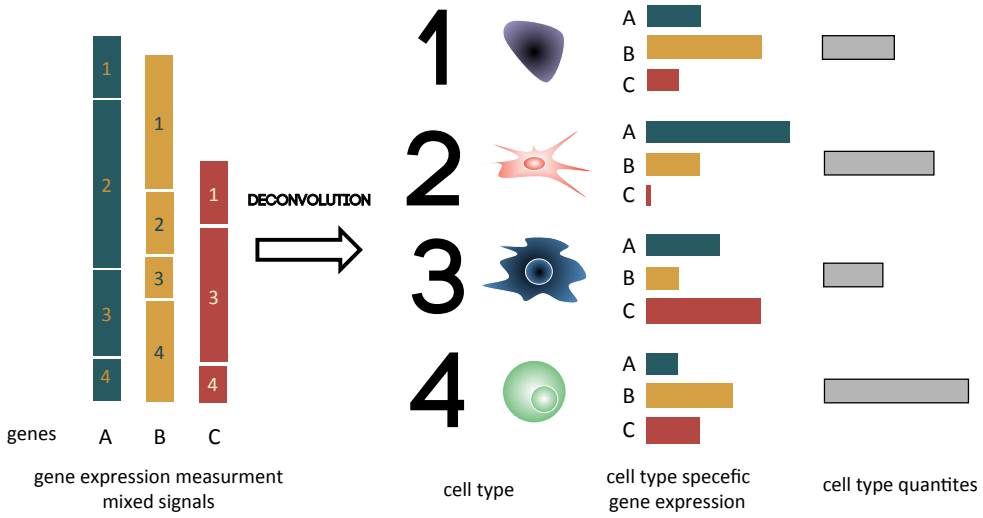


Figure 2.2: Principle of the deconvolution applied to transcriptome Graphical illustration of the deconvolution of mixed samples. Starting from the left, gene expression of genes A B C is a sum of expression of cell types 1, 2, 3, 4. After deconvolution, cell types are separated and gene expression of each cell type is estimated taking into account cell type proportions.

Where in Equaltion (2.1) X is microarray data matrix of one biological sample, A are mixing proportions and S is the matrix of expression of genes in each cell type.

Algebraically the same problem can be formalized as latent variable model:

$$\begin{gathered} \forall i \in \{1, M\}, \forall j \in \{1, N\} \\ x_{ij} = \sum_{k=1}^{K} a_{kj} * s_{ik} + e_{ij} \end{gathered} \tag{2.2}$$

Where x_{ij} is expression of gene i in sample j, a_{kj} is the proportion of cell type k in sample j and s_{ik} is the expression of the gene i in the cell type k, K total number of cell types, N total number of samples, M total number of genes. The error term e_{ij} cannot be directly measured.

The goal of deconvolution is to reverse these equations and starting from the mixture infer the A (or a_{kj}) and S (or s_{ik}).

Graphically the deconvolution of bulk gene expression can be depicted as in Fig. 2.2.

However, in this model, either the mixing proportions, number of mixing sources or an array of specific genes need to be know. While, in the real-life case, only X is truly known. Therefore, developed models proposed various manners for estimating number of mixing sources and their proportions, or the specific cell type expression.

Why there is a need for cell-type deconvolution approaches?

- for differential gene expression analysis, to avoid confusion between a feature and cell-type abundance
- difference in gene expression in one cell type can be blurred by presence of other cells expressing the gene
- to obtain information about a fraction of given component in the sample
- to infer context-specific profile or signature

2.3.1 Literature overview

In order to answer general and specific need for cell-type deconvolution of bulk transcriptomes researches produced a large collection of tools. I have collected all (to my knowledge) articles published in journals or as a pre-print (up to May 2018) that propose original models/tools of cell-type deconvolution of bulk **transcriptomes** (Tab. 2.1). Therefore clonal deconvolution methods are not included in this overview. The transcriptome-based purity estimation methods are included as many of them proposed an initial 2-sources model that could be, at least in theory, extended to multiple sources model. Also, I did not included cell-type deconvolution methods of other data types (such as methylome). A separate (section X)[#otherDecon] is dedicated to non-transcriptome methods.

The Table 2.1 contains 64 (including mine) deconvolution methods. It can be observed (Fig. 2.3) that since the begging of my thesis (2015) the number of publications has doubled (64 publications in 2018 vs. 33 in 2014). Also, since 2014 more methods are published every year. In Fig. 2.3 *hallmark* publications are indicated in red above their year of publication. The three most popular methods (based of number of citations/number of years since publication) are CIBERSORT [141] (2015, total number of citations: 343 and 88.75 citations per year), ESTIMATE [230] (2013, total number of citations: 266 and 44.33 citations per year), and csSAM [191] (2010, total number of citations: 286 and 31.77 citations per year). It can be noticed that the high impact of the journal plays a role, the top 3 cited methods were published in *Nature Methods* and *Nature Communications* followed by Virtual Microdissection method [135] (2015) published in *Nature Genetics*. However, the fifth most cited publication Abbas et al. [3] (2009, total of 207 citations) appeared in *PLOS ONE*. As the index is a bit penalizing for recent publications, among commonly cited tools after

2015 are MCPcounter with 42 citations (2016, 32 without self-citations) and xCell with 14 citations (2017, 11 without self-citations). A big number of publications with low number of citations were published in *Oxford Bioinformatics* or *BMC Bioinformatics* which underlines importance of publishing a computational tool along with an important biological message rather than in a technical journal in order to increase a chance to be used by other researchers.

Another important aspect is availability of the tool. One-third (in total 21) methods do not provide source code or a user-interface tool to reproduce their results. Among those articles, 13 was published before 2015. Therefore, it can be concluded that the pressure of publishers and research community on reproducibility and accessibility of bioinformatic tools gives positive results. Shen-Orr and Gaujoux [190], authors of semi-supervised NMF method [71], published *CellMIx: a comprehensive toolbox for gene expression deconvolution* where he implements most of previously published tools in R language and group them in the same R-package. This work tremendously increased the usability of previously published deconvolution methods. The CellMix package is one of the state-of-the-art work on deconvolution that regroups algorithms, signatures and benchmark datasets up to 2013.

Table 2.1: Summary of methods for cell-type deconvolution of bulk transcriptome. Data gathered based on pubmed and google scholar search in May 2018.

name	data	type	doi	year	application	availability	out.profiles	out.proportions	category	language	citations	pop.index	previously.covered
GSVA scores	RNA-seq	supervised	https://doi.org/10.1158/1078-0432.CCR-17-3509	2018	Cancer transcriptome	NA	FALSE	TRUE	enrichment	unknown	1	1.00	FALSE
MySort	MA	supervised	https://doi.org/10.1186/s12859-018-2069-6	2018	Blood	https://testtoolshed.g2.bx.psu.edu/repository?repository_id=6e9a9abf63e578e0&changeset_revision=e3afa097e80a	FALSE	TRUE	regression	R, web tool	0	0.00	FALSE
ADVOCATE	RNA-seq	supervised	https://doi.org/10.1101/288779	2018	Cancer transcriptome	NA	TRUE	TRUE	probabilistic	R	0	0.00	FALSE
DTD	scRNA-seq	supervised	https://arxiv.org/abs/1801.08447v1	2018	Cancer transcriptome	NA	FALSE	TRUE	regression	unknown	0	0.00	FALSE
CellDistinguisher	MA + RNA-seq	unsupervised	https://doi.org/10.1371/journal.pone.0193067	2018	yeast cell cycle	https:// github.com/GeneralElectric/CcellDdistinguisher	TRUE	TRUE	convex hull	R	0	0.00	FALSE
dtangle	MA + RNA-seq	supervised	https://doi.org/10.1101/290262	2018	Blood	https://cran.r-project.org/package=dtangle	FALSE	TRUE	regression	R	0	0.00	FALSE
DeconICA	MA + RNA-seq	unsupervised	https://doi.org/10.5281/zenodo.1250069	2018	Cancer transcriptome	https://urszulaczerwinska.github.io/DeconICA/	TRUE	TRUE	matrix factorisation	R, matlab	0	0.00	FALSE
xCell	MA + RNA-seq	supervised	https://dx.doi.org/10.1186%2Fs13059-017-1349-1	2017	Cancer transcriptome	http://xcell.ucsf.edu/; https://github.com/dviraran/xCell	FALSE	TRUE	enrichment	R, web tool	15	7.50	FALSE
BioQC	MA + RNA-seq	supervised	https://doi.org/10.1186/s12864-017-3661-2	2017	Gene expression	https://www.bioconductor.org/packages/release/bioc/html/BioQC.html	FALSE	FALSE	enrichment	R	6	3.00	TRUE
EPIC	RNA-seq	supervised	https://dx.doi.org/10.7554%2FeLife.26476	2017	Cancer transcriptome	https://github.com/GfellerLab/EPIC	FALSE	TRUE	regression	R	4	2.00	FALSE
Estimation of immune cell content	scRNA-seq	supervised	https://doi.org/10.1038/s41467-017-02289-3	2017	Cancer transcriptome	NA	FALSE	TRUE	regression	unknown	3	1.50	FALSE
Enumerateblood	MA	supervised	https://doi.org/10.1186/s12864-016-3460-1	2017	Blood gene expression	https://github.com/cashoes/enumerateblood	TRUE	TRUE	probabilistic	R	2	1.00	TRUE
ImmunoStates	MA	supervised	https://doi.org/10.1101/206466	2017	Blood, solid tissue, disease	NA	FALSE	TRUE	regression	R	1	0.50	FALSE
quanTIseq	RNA-seq + Images	supervised	https://doi.org/10.1101/223180	2017	Cancer transcriptome	http://icbi.at/software/quantiseq/doc/index.html	FALSE	TRUE	regression	web tool	1	0.50	FALSE
SMC	MA	unsupervised	https://doi.org/10.1371/journal.pone.0186167	2017	Tissue mixtures	https://github.com/moyamine/smcgenedeconv	TRUE	TRUE	probabilistic	matlab	1	0.50	FALSE
Modular discrimination index	MA + RNA-seq	supervised	https://doi.org/10.1371/journal.pone.0169271	2017	Skin tuberculosis	https://github.com/MJMurray1/MDIScoring	FALSE	TRUE	enrichment	R	1	0.50	FALSE
DemixT	MA + RNA-seq	supervised	https://doi.org/10.1101/146795	2017	Cancer transcriptome	https://github.com/wwylab/DeMixT	TRUE	TRUE	probabilistic	R	0	0.00	FALSE
Post-modified non-negative matrix factorization	RNA-seq	unsupervised	https://doi.org/10.3002/bem.2929	2017	Cancer transcriptome	NA	TRUE	TRUE	matrix factorisation	matlab	0	0.00	FALSE
Infino	RNA-seq	supervised	https://doi.org/10.1101/221671	2016	Cancer transcriptome	https://github.com/hammerlab/infino	FALSE	TRUE	probabilistic	Stan	0	0.00	FALSE
MCPcounter	MA	supervised	https://doi.org/10.1186/s13059-016-1070-5	2016	Cancer transcriptome	https://github.com/ebecht/MCPcounter	FALSE	TRUE	enrichment	R	42	14.00	TRUE
ssGSEA applied to renal cell carcinoma	RNA-seq	supervised	https://doi.org/10.1186/s13059-016-1092-z	2016	Cancer transcriptome	NA	FALSE	TRUE	enrichment	R	40	13.33	TRUE
CAM	MA	unsupervised	https://doi.org/10.1038/srep18909	2016	yeast cell cycle	http://mloss.org/software/view/437/	TRUE	TRUE	convex hull	R-java	12	4.00	TRUE
Immune Quant	undefined	supervised	https://doi.org/10.1093/bioinformatics/btw535	2016	Human tissues	http://csgi.tau.ac.il/ImmQuant/	FALSE	TRUE	regression	web tool	5	1.67	TRUE
VoCAL	MA, GWAS	supervised	https://doi.org/10.1371/journal.pcbi.1004856	2016	Lung tissue	https://cran.r-project.org/web/packages/ComICS/index.html	FALSE	TRUE	regression	R	5	1.67	TRUE
CellMapper	MA	semi-supervised	https://doi.org/10.1186/s13059-016-1062-5	2016	Brain tissue	http://bioconductor.org/packages/release/bioc/html/CellMapper.html	TRUE	FALSE	matrix factorisation	R	5	1.67	TRUE
contamDE	RNA-seq	supervised	https://doi.org/10.1093/bioinformatics/btv657	2016	Tumor purity	https://github.com/zhanghfd/contamDE/	TRUE	TRUE	probabilistic	R	4	1.33	TRUE
ImSig	RNA-seq	supervised	https://doi.org/10.1101/077487	2016	Cancer transcriptome	NA	FALSE	TRUE	enrichment	unknown	0	0.00	FALSE
CIBERSORT	MA	supervised	https://doi.org/10.1038/nmeth.3337	2015	Cancer transcriptome	http://cibersort.stanford.edu/	FALSE	TRUE	regression	R, web tool	343	85.75	TRUE
Virtual Microdissection	MA	unsupervised	https://doi.org/10.1038/ng.3398	2015	detection of cancer and stroma in PDAC (TCGA)	NA	TRUE	TRUE	matrix factorisation	matlab	86	21.50	TRUE
CellCODE	MA	semi-supervised	https://doi.org/10.1093/bioinformatics/btv015	2015	Blood	http://www.pitt.edu/~mchikina/CellCODE/	TRUE	TRUE	matrix factorisation	R-C-C++- Fortran	28	7.00	TRUE
CoD	RNA-seq	supervised	https://doi.org/10.1093/bioinformatics/btv498	2015	Mice diseased tissues	http://www.csgi.tau.ac.il/CoD/	FALSE	TRUE	regression	web tool	4	1.00	TRUE
DCQ	RNA-seq	supervised	https://dx.doi.org/10.1002%2Fmsb.134947	2014	Mice blood under flu infection	http://www.dcq.tau.ac.il/	FALSE	TRUE	regression	web tool	32	6.40	TRUE
UNDO	MA	unsupervised	https://doi.org/10.1093/bioinformatics/btu607	2014	Cancer transcriptome	https://www.bioconductor.org/packages/release/bioc/html/UNDO.html	TRUE	TRUE	matrix factorisation	R	18	3.60	TRUE
ESTIMATE	MA + RNA-seq	supervised	https://doi.org/10.1038/ncomms3612	2013	Cancer transcriptome	https://sourceforge.net/projects/estimateproject/	FALSE	TRUE	enrichment	R	266	44.33	TRUE
DeconRNASeq	RNA-seq	supervised	https://doi.org/10.1093/bioinformatics/btt090	2013	Tissue mixtures	https://www.bioconductor.org/packages/release/bioc/html/DeconRNASeq.html	FALSE	TRUE	regression	R	52	8.67	TRUE
DSA	MA	supervised	https://dx.doi.org/10.1186%2F1471-2105-14-89	2013	Cancer transcriptome	https://github.com/shandong/DSA	TRUE	TRUE	regression	R	52	8.67	TRUE
ISOpure	MA	supervised	https://doi.org/10.1186/gm433	2013	Cancer transcriptome	https://qlab.faculty.ucdavis.edu/isopure/	TRUE	TRUE	probabilistic	matlab, R	44	7.33	TRUE
DeMix	MA	supervised	https://doi.org/10.1093/bioinformatics/btt301	2013	Cancer purity	http://bdm.mdacc.tmc.edu/bewang/11/DeMix.html.	TRUE	TRUE	probabilistic	C, R	38	6.33	TRUE
Nanodissection	MA	supervised	https://doi.org/10.1101/gr.155697.113	2013	Chronic kidney disease (Cell lineages)	http://nano.princeton.edu/	FALSE	TRUE	regression	web tool	33	5.50	TRUE
TIMER	MA + RNA-seq	supervised	https://doi.org/10.1186/s13059-016-1028-7	2013	Cancer transcriptome	http://cistrome.org/TIMER/	FALSE	TRUE	regression	web tool	33	5.50	TRUE
Self-directed Method for Cell-Type Identification	MA	unsupervised	https://doi.org/10.1371/journal.pcbi.1003189	2013	Cancer transcriptome	NA	TRUE	TRUE	matrix factorisation	matlab	18	3.00	TRUE
MMAD	MA	BIOTH	https://doi.org/10.1093/bioinformatics/btt566	2013	in vitro tissue mixtures	http://sourceforge.net/projects/mmad/	TRUE	TRUE	regression	matlab	11	1.83	TRUE
Statical mechanics approach	undefined	unsupervised	https://arxiv.org/abs/1210.7508v1	2013	Udefined	NA	TRUE	TRUE	probabilistic	unknown	2	0.33	FALSE
TEMT	RNAseq	supervised	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S5-S11	2013	in vitro cell mixtures	https://github.com/uci-cbcl/TEMT	FALSE	TRUE	probabilistic	python	0	0.00	TRUE
Semi-supervised Nonnegative Matrix Factorization	MA	semi-supervised	https://doi.org/10.1016/j.miegid.2011.08.014	2012	Blood	https://web.cbio.uct.ac.za/~renaud/CIMA/web/CellMix/	TRUE	TRUE	matrix factorisation	R	61	8.71	TRUE
PERT	MA	supervised	https://doi.org/10.1371/journal.pcbi.1002838	2012	blood	https://github.com/jguon/PERT	TRUE	TRUE	probabilistic	octave	33	4.71	TRUE
CTen	MA	supervised	https://doi.org/10.1186/1471-2164-13-460	2012	Infacted lung tissue	http://www.influenza-x.org/~jshoemaker/cten/	TRUE	FALSE	enrichment	web tool	31	4.43	TRUE
PSEA	MA	supervised	https://doi.org/10.1038/nmeth.1710	2011	Brain tissue	https://bioconductor.org/packages/release/bioc/html/PSEA.html	FALSE	TRUE	regression	R	96	12.00	TRUE
Quadratic programming	MA	supervised	https://doi.org/10.1371/journal.pone.0027156	2011	Blood	NA	FALSE	TRUE	regression	unknown	76	9.50	TRUE
SPEC	MA	supervised	https://doi.org/10.1186/1471-2105-12-258	2011	Blood	http://clip.med.yale.edu/SPEC/	FALSE	TRUE	enrichment	R	39	4.88	TRUE
csSAM	MA	supervised	https://doi.org/10.1038/nmeth.1439	2010	Blood	https://github.com/shenorrLab/csSAM	TRUE	FALSE	regression	R	286	31.78	TRUE
Statistical expression deconvolution	MA	supervised	https://doi.org/10.1093/bioinformatics/btq097	2010	Cancer xenografts	NA	FALSE	TRUE	regression	unknown	53	5.89	TRUE
DSection	MA	supervised	https://doi.org/10.1093/bioinformatics/btq406	2010	Tissue mixtures	http://informatics.systemsbiology.net/DSection	TRUE	FALSE	probabilistic	matlab	52	5.78	TRUE
deconf	MA	unsupervised	https://doi.org/10.1186/1471-2105-11-27	2010	Blood	https://static-content.springer.com/esm/art%3A10.1186%2F1471-2105-11-27/MediaObjects/12859_2009_3484_MOESM1_ESM.ZIP	TRUE	TRUE	matrix factorisation	R	41	4.56	TRUE
Abbas regression	MA	supervised	https://doi.org/10.1371/journal.pone.0006098	2009	Blood	NA	TRUE	TRUE	regression	R	207	20.70	TRUE
ISOLATE	MA	supervised	https://dx.doi.org/10.1093%2Fbioinformatics%2Fbtp378	2009	Cancer transcriptome	https://qlab.faculty.ucdavis.edu/isolate/	TRUE	TRUE	matrix factorisation	matlab	37	3.70	TRUE
Electronical substraction	MA	supervised	https://doi.org/10.1093/bioinformatics/btm508	2007	Infected macrophages	NA	TRUE	TRUE	regression	unknown	30	2.50	TRUE
Computational expression deconvolution	MA	supervised	https://doi.org/10.1186/1471-2105-7-328	2006	Murine mammary gland	NA	FALSE	TRUE	regression	unknown	36	2.77	TRUE
Robust Computational Reconstitution	MA	supervised	https://doi.org/10.1186/1471-2105-7-369	2006	Synovial tissue (cell types in silico)	NA	FALSE	TRUE	regression	unknown	6	0.46	TRUE
MHMM	MA	unsupervised	https://doi.org/10.1089/cmb.2006.13.1749	2006	Yeast cell cycle	NA	TRUE	TRUE	probabilistic	unknown	4	0.31	TRUE
In silico microdissection	MA	unsupervised	https://doi.org/10.1186/1471-2105-6-54	2005	In vitro tissue mixtures	NA	TRUE	TRUE	probabilistic	unknown	45	3.21	TRUE
Mixture models	MA	supervised	https://doi.org/10.1093/bioinformatics/bth139	2004	Cancer transcriptome	broken link	TRUE	TRUE	probabilistic	R	66	4.40	TRUE
DECONVOLUTE	MA	supervised	https://doi.org/10.1073/pnas.1832361100	2003	yeast cell cycle	broken link	FALSE	TRUE	regression	Java 2	135	8.44	TRUE
Direct method	MA	unsupervised	https://www.ncbi.nlm.nih.gov/pubmed/11473019	2001	cander and normal tissue	NA	TRUE	TRUE	matrix factorisation	unknown	96	5.33	TRUE

The Table 2.1 contains 64 (including mine) deconvolution methods. It can be observed (Fig. 2.3) that since the begging of my thesis (2015) the number of publications has doubled (64 publications in 2018 vs. 33 in 2014). Also, since 2014 more methods are published every year. In Fig. 2.3 *hallmark* publications are indicated in red above their year of publication. The three most popular methods (based of number of citations/number of years since publication) are CIBERSORT [141] (2015, total number of citations: 343 and 88.75 citations per year), ESTIMATE [230] (2013, total number of citations: 266 and 44.33 citations per year), and csSAM [191] (2010, total number of citations: 286 and 31.77 citations per year). It can be noticed that the high impact of the journal plays a role, the top 3 cited methods were published in *Nature Methods* and *Nature Communications* followed by Virtual Microdissection method [135] (2015) published in *Nature Genetics*. However, the fifth most cited publication Abbas et al. [3] (2009, total of 207 citations) appeared in *PLOS ONE*. As the index is a bit penalizing for recent publications, among commonly cited tools after 2015 are MCPcounter with 42 citations (2016, 32 without self-citations) and xCell with 14 citations (2017, 11 without self-citations). A big number of publications with low number of citations were published in *Oxford Bioinformatics* or *BMC Bioinformatics* which underlines importance of publishing a computational tool along with an important biological message rather than in a technical journal in order to increase a chance to be used by other researchers.

Another important aspect is availability of the tool. One-third (in total 21) methods do not provide source code or a user-interface tool to reproduce their results. Among those articles, 13 was published before 2015. Therefore, it can be concluded that the pressure of publishers and research community on reproducibility and accessibility of bioinformatic tools gives positive results. Shen-Orr and Gaujoux [190], authors of semi-supervised NMF method [71], published *CellMIx: a comprehensive toolbox for gene expression deconvolution* where he implements most of previously published tools in R language and group them in the same R-package. This work tremendously increased the usability of previously published deconvolution methods. The CellMix package is one of the state-of-the-art work on deconvolution that regroups algorithms, signatures and benchmark datasets up to 2013.

The most popular language of implementation of published methods is R (49.2 %), followed by Matlab (11.11%), only one tool so far was published in Python.

Also, most of methods were designed to work with microarray data. There is a high chance that some of them are adaptable to RNA-seq. However, little number of older methods was tested in a different setup. For some method, as CIBERSORT, demonstrated to work with microarray and applied commonly to RNA-seq by other researchers, the validity of results remains unclear as some studies claim that CIBERSORT performs accurately applied to RNA-seq [208] and other opt agains it (Li et al. [118]; Tamborero et al. [202]). Most of newer methods (i.e. EPIC [172], quanTIseq [60] or Infino [233]) are specif-

ically designed for RNA-seq TPM-normalized data. Some methods, mostly enrichment-based methods, are applicable to both technologies (i.e. xCell [9]).

It is remarkable that the general aim of the cell-type deconvolution changed with time. The earlier methods aimed to improve the power of differential expression analysis though *purification* of the gene expression. For example, to compare differentially expressed genes (DEG) in T-cell from the blood in two conditions. However, the obtained purified profiles from complex mixtures were often uncertain [152]. Recently, the most mentioned goal of deconvolution is quantification of proportions of different cell types, especially in the context of cancer transcriptomes motivated by redefinition of immunophenotypes discussed in the previous chapter. The most popular tissue of interest for deconvolution algorithms are cancer tissues and blood, other applications are cell-cycle time dependent fluctuations of yeast, brain cells and glands .

Mathematically speaking, I have divided methods into four categories: probabilistic, regression, matrix factorisation and convex hull depending on the nature of approach. Most of the methods (48 - 74.6%) are working within a supervised framework and only 20% (14) are unsupervised. The approaches will be described in detail in the following section.

There are numerous practical differences between the methods. Shen-Orr and Gaujoux [190] in his review of deconvolution tools, grouped the tools depending on their inputs and outputs. Given type of outputs deconvolution can be considered as complete (proportions and cell profiles) or partial (one of those). Moreover, the inputs of the algorithms can be important to evaluate how practical the tool is. The most popular tools and the most recent tools ask minimal input from the user: the bulk gene expression matrix, or even raw sequencing data [60]. Older methods usually request either at least approximative proportions of mixed cells or purified profiles to be provided. The newer methods include the reference profiles in the tool if necessary. Some tools, including most of purity estimation tools, demand an additional data input as normal samples or another data type such as CNA data (Timer [119], VoCAL [196]] or image data (quanTIseq [60]). An important parameter is also a number of sources (k) to which the algorithm deconvolutes the mixture. In many methods it should be provided by the user, which can be difficult in a case of complex mixtures of human tissues. In addition, type of method can also limit the number of sources, for example, a probabilistic framework privilege lower number of sources (2-3) due to the theoretical constraints. In regression depending on provided reference the output number of estimated sources is imposed. Because of the problem of collinearity and similarity of immune cell profiles, it is hard to distinguish between cell sub-types, deconvolution into fine-grain cell subtypes is often called often deep deconvolution. Some methods (i.e. CIBERSORT, Infino, xCell) give a specific attention to deconvolution of cell-subtypes. An absolute presence of a cell type in the mixture can be also an important factor, if it is too low it can reach a detection limit, Electronic

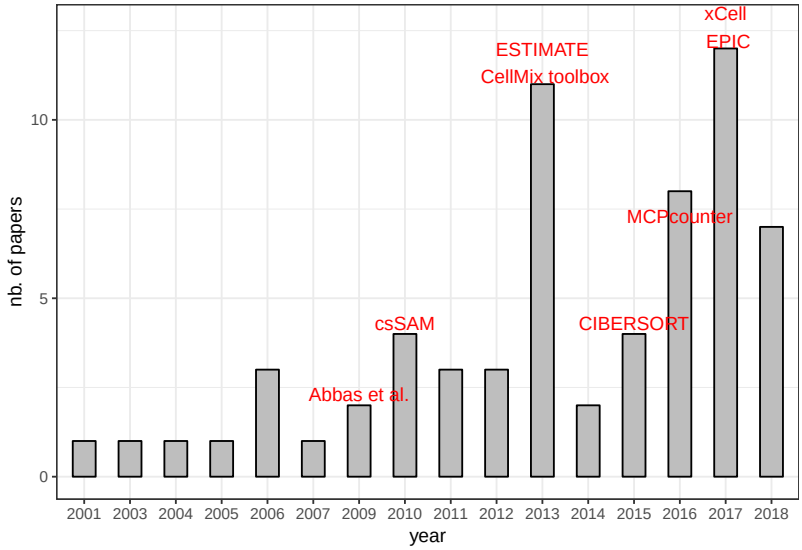


Figure 2.3: Distribution of publications of cell-type deconvolution of bulk transcriptome over the years. In red: hallmark publications. Data gathered based on pubmed and google scholar search in May 2018.

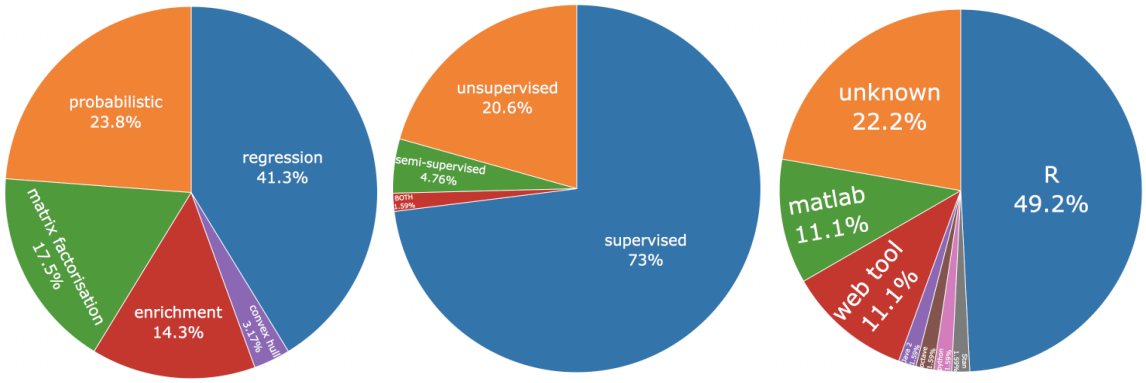
subtraction [80] discuss specifically the detection of rare cel-types.

Running time and necessary infrastructure are another way to characterize the methods. Although it is hard to compare objectively the running time simultaneously of all the tools because of the heterogeneity of methods and different datasets analysed, some tendencies can be observed. If one thinks about applying deconvolution methods to big cohorts, regression and enrichment-based methods should be well suited. As far as matrix factorisation is concerned, it is depending on implementation (i.e. R vs Matlab) and if number of sources needs to be estimated (multiple runs for different k parameter) or if a stabilisation needs to applied (multiple runs for the same k parameter). Finally, probabilistic tools seem to be difficult to scale, i.e. authors of Infino admit that their pipeline is not yet applicable at high-throughput.

In order to let user better understand the differences between different mathematical approaches, I will introduce shortly the types of approaches used for cell-type deconvolution of transcriptomes as well as their strong and weak points.

2.3.2 Regression-based methods

Regression models are the most popular methods for bulk gene expression deconvolution. They use estimated pure cell profiles as depending variables (or selected signature genes) that should explain the mixed profiles choosing best β parameters ((2.3)) that can



(a) approach type **(b)** supervised/unsupervised **(c)** programming language

Figure 2.4: Simple statistics illustrating characteristics of published cell-type deconvolution tools: 2.4a - Percentage of used approach type, 2.4b - Percentage of supervised/unsupervised tools, 2.4c - Percentage of the programming languages of implementation. Data gathered based on pubmed and google scholar search in May 2018.

be interpreted as cell proportions.

A standard type of regression is called linear regression. It reflects linear dependence between independent and dependent variables. The linear regression was developed in the *precoputer age of statistics* [88].

In linear regression, we want to predict a real-valued output Y, given a vector $X^T = (X_1, X_2, , X_p)$. The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.3}$$

Where the β_js are unknown parameters or coefficients, and X_js are the explaining variables. Given pairs of (x_1,y_1)...(x_N ,y_N), one can estimate coefficients β with an optimization of an objective function (also called cost function).

The most popular estimation method is **least squares**, the coefficients $\beta = (\beta_0, \beta_1, ..., \beta_n)$ are computed to minimize the residual sum of squares (RSS):

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_i \beta_j)^2 \tag{2.4}$$

Ordinary least squares regression is using Eq.(2.4) to compute β.

Ridge regression (Equation (2.5)) adds a regularizer (called $L2$ norm) to shrink the coefficients ($\lambda \geq 1$) through imposing a penalty on their size.

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}}\{\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \tag{2.5}$$

Similarly **Lasso regression** (Equation (2.6)) adds a regularization term to RSS (called $L1$ norm), it may set coefficients to 0 and therefore perform feature selection.

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}}\{\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} |\beta_j|\} \tag{2.6}$$

In **Elastic net regression** both penalties are applied.

Support Vector Regression (SVR) is regression using **Supported Vector Machines (SVM)**. In SVR β can be estimated as follows:

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i f(x_i)) + \frac{}{2}\|\beta\|^2 \tag{2.7}$$

where error is measured as follows:

$$V_\epsilon(r) = \begin{cases} 0, & \text{if } |r| < \epsilon, \\ |r| - \epsilon, & \text{otherwise.} \end{cases} \tag{2.8}$$

with ϵ being the the limit of error measure, meaning errors of size less than ϵ are ignored.

In the SVM vocabulary, a subset of the input data that determine hyperplane boundaries are called the **support vectors** (Fig.2.5). SVR discovers a hyperplane that fits maximal possible number of points within a constant distance, ϵ, thus performing a regression.

In brief, in SVR, RSS is replaced by a linear ϵ-insensitive loss function and uses $L2$-norm penalty function. There exist variants of SVR algorithm, i.e. ϵ-SVR [49] and ν-SVR [185]. ϵ-SVR allows to control the error, this favorizes more complex models. In the ν-SVR the distance of the ϵ margin can be controlled and therefore number of data points used for regression can be controlled. Ju et al. [103] used SVM-based method to define cell type specific genes. A model using ν-SVR with linear kernel was used by Newman et al. [141] in CIBERSORT.

As unconstrained optimization of the objective function can result in negative coefficients. In the context of cell-type deconvolution, authors often aim to avoid as it com-

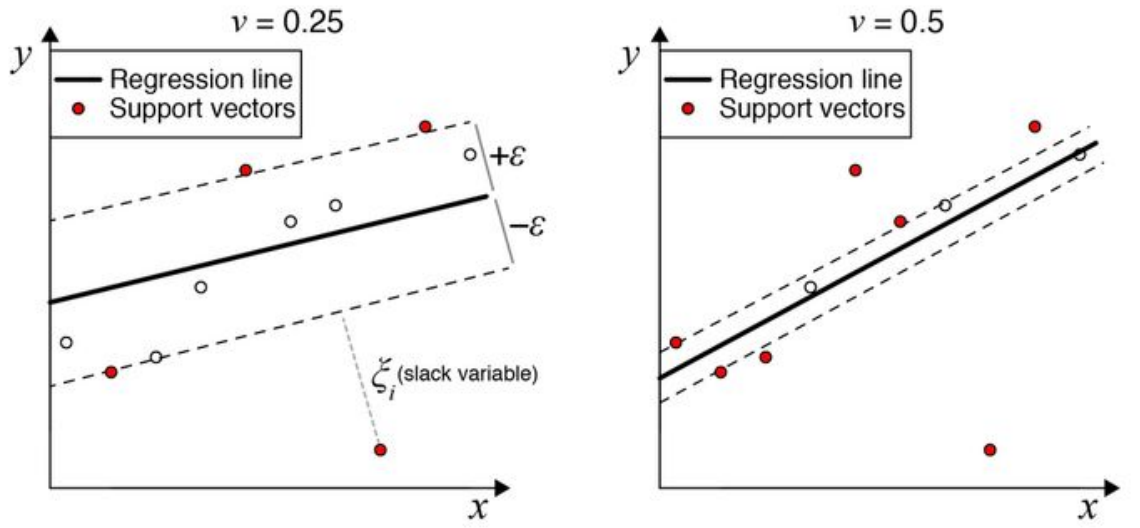


Figure 2.5: Priciple of the SVR regression. In SVR regression ϵ represents limit of error measure, input data points higher than $+\epsilon$ or lower than $-\epsilon$ are called support vectors. The ν parameter in ν-SVR regression controls the distance of training error bonds: left - lower ν value larger bound, right - higher v margin, smaller bound. Reprinted by permission from Springer Nature [141] © 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

plicates the interpretation. Therefore, different constraints can be imposed to the β coefficients. The most common conditions are $\beta_0 + \beta_1 + ... + \beta_n = 1$ and $\forall \beta_i \geq 0$. Solution respecting the non-negativity condition is also called non-negative least squares (NNLS) to contrast with ordinary least squares (OLS). NNLS was adapted by many authors [217, 3, 175, 239, 223].

The task can be also solved differently from computational perspective. Lu et al. [124] and Wang et al. [220] propose to use simulated annealing to minimize the cost function. Gong et al. [77] proposed to solve task using quadratic programming.

An extensive review on optimisation of the objective function for regression methods in cell-type deconvolution was published by Mohammadi et al. [136]. Authors carefully consider different possibilities of parameter choice in the loss and regularization fomulations and its performance. They present as well recommendations for construction of basis matrix and data pre- and post-processing. Digital tissue deconvolution (DTD) [79] aims to train the loss function with *in silico* mixtures of single cell profiles resulting in improved performance of rare cell types (present in small proportion). However, the training is computationally heavy and the proper training data for bulk transcriptomes are not available.

Since the publication of CIBERSORT [141] some authors [35, 184] used the Newman et al.

[141] implementation directly with pre/post modifications or different signature matrix or re-implemented the SVR regression [35].

Another recent method EPIC [172] introduced weights related to gene variability. In their constrained regression, they add it explicitly in the cost function modifying RSS (Eq.(2.4)):

$$RSS^{weighted}(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - w_i \sum_{j=1}^{p} x_i\beta_j)^2 \tag{2.9}$$

with the negativity and sum constraints we discussed above. The w_i weights are corresponding to variance of the given gene measure in the same cell type. It aims to give less importance to the variant genes. EPIC also allows a cell type that is not a referenced in the signature matrix with an assumption that the non-referenced cell type is equal to 1- sum of proportions of other cell types (Eq.(2.10)). This non-referenced cell type is interpreted by authors as the tumor fraction:

$$\beta_m = 1 - \sum_{j=1}^{m-1} \beta_j \tag{2.10}$$

An additional feature of EPIC is advanced data normalisation and an estimation of mRNA produced by each cell to adjust cell proportions, which was previously proposed by Liebner et al. [120] in the context of microarray data:

$$p_j = \alpha \frac{\beta_j}{r_j} \tag{2.11}$$

where p_j are actual cell proportions that are 'normalized' with empirically derived coefficient α and measured r_j is the amount of RNA nucleotides in cell type j.

Recently CIBERSORT proposed an *absolute mode* where the proportions are not relative to the leucocyte infiltration but to the sample. It can be obtained with assumption that the estimation of proportion of all genes in CIBERSORT matrix is corresponding to sample purity. This functionality was not yet officially published and it is still in experimental phase [142].

Regression methods combined with pre- and post-processing of data can result in estimation of proportions that can be interpreted directly as a percentage of cells in mixed sample. It is an important feature hard to achieve with other methods. Some methods provide relative proportions of the immune infiltrate [141] and other aim to provide absolute abundance [172]. The absolute proportions are easily comparable between data sets and cell types. Regression based methods are usually quite fast and can process big

Table 2.2: Contangency table is the count of overlap of genes present in a certain codition (Y) vs not present (Y-Z) and association to a pathway X (in X or not in X). Contangency table is used in frequency based test as Fisher exact test.

	Y	Z-Y
in X	a	b
not in X	c	d

transcriptomic cohorts. However, as I will discuss in Validation section, they pose on the hypothesis that the reference profiles available in some context (i.e. blood) are valid in a different one (i.e. tumor) or that profiles extracted from one data type (scRNA-seq) are adapted to deconvolute bulk RNAseq. Most of recent regression methods focused on estimating proportions and do not estimate context specific profiles and can process as little as one sample.

2.3.3 Enrichment-based methods

Enrichment-based methods aim to evaluate an amount of activity of a given list of genes within the data. This can be obtained though calculating a score based on gene expression. Traditionally enrichment methods were used to analyse set of DEG. Different statistical approaches were adapted: like fisher exact test giving a p-value that estimated the chance a given list of genes is over/under present in the input list of DEGs and therefore characterise the condition vs. control expressed genes.

Let's take an example, if one wants to compute enrichment in pathway X of the list of DEG genes Y with total number of tested genes Z, a contingency table need to be constructed (Tab. 2.2).

In the Fisher exact test formula (Eq. (2.12)) the a, b,c and d are the individual frequencies, i.e. number of genes in of the 2X2 contingency table, and N is the total frequency ($a + b + c + d$).

$$p = \frac{((a+b)!(c+d)!(a+c)!(b+d)!)}{a!b!c!d!N!} \tag{2.12}$$

Another famous (>14000 citations) algorithm computing such a score (enrichment score ES) is named gene set enrichment analysis (GSEA) [198] uses sum-statistics. The list of genes user wants to test for enrichment is usually ranked by fold change odd or p-value of DGE analysis.

The high score indicated high activity of genes included in the list. GSEA can also indicate

an anti-activity of correlation. A variant of GSEA, single sample GSEA (ssGSEA) [14] was used by Şenbabaoğlu et al. [240], Yoshihara et al. [230] and Aran et al. [9] to compute infiltration scores. In the ssGSEA genes are ranked by their absolute expression. Variance-based variant of GSEA - GSVA [86] was used by Tamborero et al. [202] in the same purpose. MCPcounter [16] uses an arithmetic mean of gene expression of highly specific signature genes to compute a score.

In this way obtained scores, are not comparable between different cell types and datasets. Therefore some authors propose normalization procedures that make the score more comparable. For instance xCell, uses a platform-specific transformation of enrichment scores. Similarly, Estimate transforms scores for TCGA though an empirically derived formula. MCPcounter authors use z-scoring to minimise platform-related differences. Unfortunately, the normalization is not directly included in the R package

Even though enrichment methods do not try to fit the linear model and derived scores are not mathematically conditioned to represent cell proportions, usually there can be observed a strong linear dependence. An advantage of the enrichment-based methods is the speed and possibility to include diverse signatures that can characterise cell-types and cell-states of different pathways.

2.3.4 Probabilistic methods

The probabilistic methods share a common denominator: they aim to minimise a likelihood function of Baye's theorem:

$$p(y|\theta) = \frac{p(\theta|y) * p(y)}{p(\theta)} \tag{2.13}$$

In Eq.(2.13) y is our data, θ a parameter, $p(y|\theta)$ *posterior*, $p(\theta|y)$ *likelihood* and $p(\theta)$ *prior*. Prior distribution is what we know about the data before it was generated and combined with a probability distribution of the observed data is called posterior distribution. The likelihood describes how likely it is to observe the data (y) given the parameter θ (probability of y given θ - $p(y|\theta)$). A parameter is characteristic of a chosen model and a hyper-parameter is a parameter of prior distribution.

In the literature, there are mainly different types of probabilisitic models, one that assumes some type of distribution of mixed sources (i.e. gaussian or poisson), other that learn the distribution parameters empirically from a training set, another that try find the parameters of the distribution given the number of given sources. Then in each case, there are different ways of constructing different priors and posteriors functions. Among used techniques are Markov Chain Monte Carlo or Expectation-Maximisation,

which themselves can be implemented in different ways [57, 72, 109, 119, 180, Zaslavsky et al. [233]].

The probabilistic approaches are the most popular for purity estimation (2 components models), that seem to be possible to extend to 3-components model [224]. As far as cell-type decomposition into a number of cells is concerned, a method published on BioRxiv *Infino* uses Bayesian inference with a generative model, trained on cell type pure profiles. Authors claim their method is notably suited for deep deconvolution that is able to build cell type similarities and estimate the confidence of the estimated proportions which help to better interpret the results.

Probabilisitic framework is an attractive approach with solid statistical bases. It can be suited to many specific cases. The pitfalls are (1) the need of prior profiles or correct hypothesis on the distribution parameters (2) reduced performance when applied to high dimensional datasets due to extensive parameters search.

2.3.5 Convex-hull based methods

An emerging family of BSS methods are convex geometry (CG)-based methods. Here, the *sources* are found by searching the facets of the convex hull spanned by the mapped observations solving a classical convex optimization problem [229]. It can be implemented in many ways [167].

Convex hull can be defined as follows [56]:

> *We are given a set P of n points in the plane. We want to compute something called the **convex hull** of P. Intuitively, the convex hull is what you get by driving a nail into the plane at each point and then wrapping a piece of string around the nails. More formally, the convex hull is the smallest convex polygon containing the points:*
>
> - ***polygon***: *A region of the plane bounded by a cycle of line segments, called **edges**, joined end-to-end in a cycle. Points where two successive edges meet are called **vertices**.*
> - ***convex***: *For any two points p, q inside the polygon, the line segment pq is completely inside the polygon.*
> - ***smallest***: *Any convex proper subset of the convex hull excludes at least one point in P. This implies that every vertex of the convex hull is a point in P.*

Convex hull methods have been used in many fields, from economics and engineering, I will discuss it with a focus on biological context to link tightly to cell-type deconvolution.

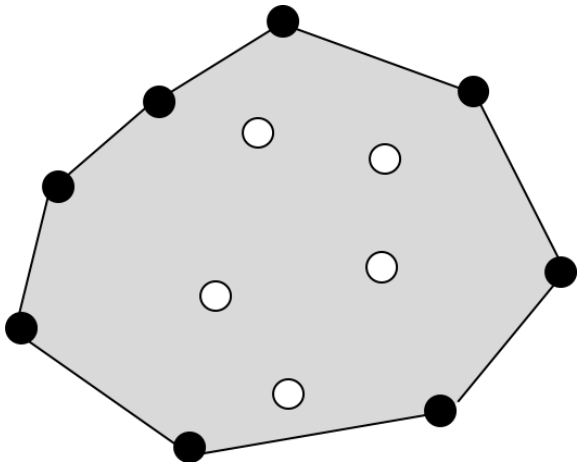


Figure 2.6: Convex hull illustration. A set of points and its convex hull (line). Convex hull vertices are black; interior points are white. Image reproduced after Erickson, Jeff [56].

The main assumptions of Convex hull optimization are that the gene expression of pure cell types is non-negative and that cell type proportions are linearly independent.

The shapes can be fitted to a cloud of points in many ways in order to respond to a given optimality criteria. A popular method introduced by Shoval et al. [192] and applied to gene expression and morphological phenotypes of biological species employ the **Pareto front** concept which aims to find a set of designs that are the best trade-offs between different requirements.

Visually Pareto front correspond to the edge of the convex hull.

Wang et al. [221] proposed Complex Analysis of Mixtures (CAM) method to find the Pareto front (the vertices of X mixed matrix (a convex set)). In the context of the cell-type deconvolution it can be said that "the scatter simplex of pure subpopulation expressions is compressed and rotated to form the scatter simplex of mixed expressions whose vertices coincide with cell proportions"[223]. In respect to the assumptions, under a noise-free scenario, novel *marker genes* can be blindly identified by locating the *vertices* of the mixed expression scatter simplex [58]. In the figure (Fig. 2.7), the a_i's are cell-type proportions of k cell types, s_i pure cell type expression and x_j mixed expression in sample j. Therefore the vertices correspond to the column vectors of the matrix A (Eq. (2.1)). The genes placed in a distance d from the vertices can be interpreted as marker genes.

In the procedure suggested by Wang et al. [221], before performing CAM, clustering (more precisely affinity propagation clustering (APC)) is applied to the matrix in order to select genes representing clusters, called cluster centers g_m and dimention reduction(PCA) is applied to the sample space. Then in order to fit a convex set, a margin-of-error should be minimized. The Eq. (2.14) explains the computation of the error which computes $L2$

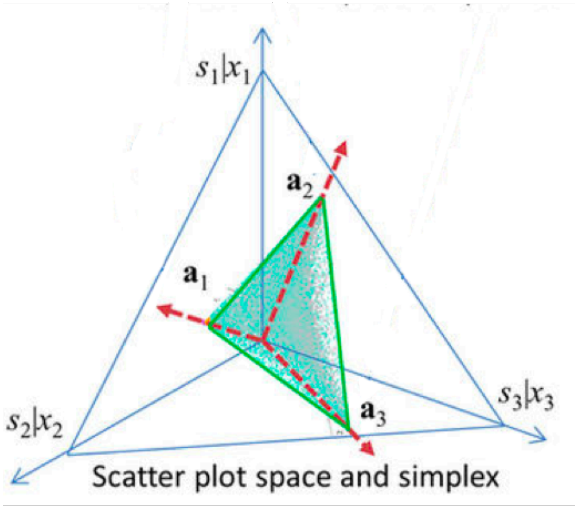


Figure 2.7: Fitting gene expression data of mixed populations to a convex hull shape. Geometry of the mixing operation in scatter space that produces a compressed and rotated scatter simplex whose vertices host subpopulation-specific marker genes and correspond to mixing proportions.

norm of the difference between g_m possible vertices and remaining exterior clusters. All possibilities of combinations drew from C_K^M, M number of clusters and K true vertices, are tested.

$$\begin{gathered} \text{given } \alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1 \\ \delta_{m,\{1,\ldots,K\} \in C_K^M} = \min_{\alpha_k} \sqrt{g_m - \sum_{k=1}^{K} \alpha_k g_k} \end{gathered} \tag{2.14}$$

Once optimal configuration is found, the proportions are computed using standardised averaging:

$$\hat{\alpha_k} = \frac{1}{n_{markers}} \sum_{i \in markers} \frac{x(i)}{\|x(i)\|} \tag{2.15}$$

where $\hat{\alpha_k}$ is proportion of cell type k, $n_{markers}$ number of marker genes (obtained from CAM), and $\|x(i)\|$ is the $L1$ or $L2$ norm of a given marker gene x_i.

Then the cell-type specific profiles are obtained with linear regression. Authors of CAM also propose a minimum description length (MDL) index that determines number of sources in the mixture. It selects the K minimizing the total description code length.

So far, the published R-Java package *CAM* does not allow to extract gene specific signatures and it is not scalable to big cohorts (many samples). In the article, authors apply

important pre-processing steps that are not trivial to reproduce and which are not included in their tool. Authors apply CAM and validate on rather simple mixtures (tissue *in vitro* mixtures and yeast cell cycle).

A slightly different approach was proposed by Newberg et al. [140]. It does not require initial dimension reduction steps or clustering before fitting the convex hull and it is based on a probabilisitic framework. The toll *CellDistinguisher* was inspired by topic modelling algorithm [11]. It first computes Q matrix (Eq. (2.16)). Then each row vector of Q is normalized to 1 giving $\overline{Q}$ matrix. Every row of $\overline{Q}$ lies in the convex hull of the rows indexed by the cell-type specific genes. Then L_2 norm of each row is computes. Genes which rows have the highest norm can be used as *distinguishers* or *marker* genes. Then another runs of selections are applied after recentering the matrix to find more markers.

$$Q = XX^T \tag{2.16}$$

Once the set of possible distinguishers is defined, proportions and cell profiles are computed using Bayesian framework to fit the convex hull. Authors provide a user-friendly R package *CellDistinguisher*. Unfortunately, they do not provide any method for estimation of number of sources, which is critical for source separation of complex tissues. Additionally, quantitative weights are provided only for signature genes which number can vary for different sources, and can be as small as one gene. Authors do not apply their algorithm to complex mixtures as tumor transcriptome, they establish a proof of concept with *in vitro* mixtures of tissues.

The convex hull-based method does not require the independence of cell types assumption, nor the non-correlation assumption which can be interesting in the setup of closely related cell types. In theory, they also allow $k > j$ (more sources than samples). So far, the existing tools are not directly applicable to tumor transcriptomes.

2.3.6 Matrix factorisation methods

Matrix factorisation is a general problem not specific to cell types deconvolution. It has been extensively used for signal processing [236]and extraction of features from images [88]. Matrix factorisation can also be called BSS or dimension reduction. Despite quite simple statistical bases they have been proven to be able to solve quite complex problems. Many matrix factorisation methods can solve the problem of Eq. (2.1). They can solve it in different ways and in respect with different hypotheses.

Naturally matrix factorisation methods estimate simultaneously A and S matrices (cell proportions and profiles) given X rectangualr matrix (genes $\times$ samples) without any additional input.

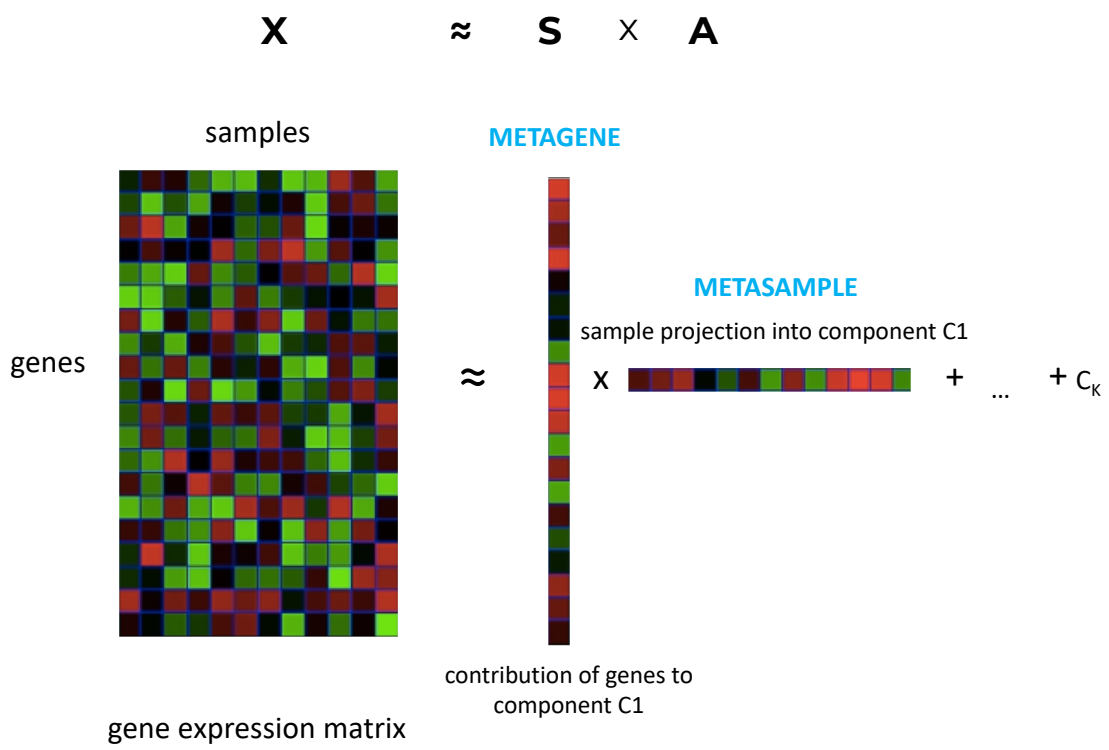


Figure 2.8: Principle of matrix factorisation of gene expression. The gene expression matrix X is decomposed into a set of *metagenes* S matrix and *metasamples* A. Number of components C is defined with parametre k.

2.3.6.1 Principal Components Analysis

One of the most popular methods, **Principal Components Analysis** (PCA). computes projections of the data, mutually uncorrelated and ordered in variance. The principal components provide a sequence of best linear approximations to that data.

Traditionally PCA is computed through eigen decomposition of the covariance matrix. Covariance matrix is computed as follows:

$$\Sigma = \frac{1}{n-1}((X - \bar{x})^T(X - \bar{x})) \tag{2.17}$$

where $\bar{x}$ is mean vector of feature column in the data X.

Then the matrix is decomposed to eigenvalues:

$$\mathbf{V}^{-1}\mathbf{\Sigma}\mathbf{V} = \mathbf{D} \tag{2.18}$$

where $\mathbf{V}$ is matrix of eigenvectors and $\mathbf{D}$ diagonal matrix of eigenvalues.

It can be also computed using **singular value decomposition** (SVD) (computationally more efficient way):

$$X = UDV^T \tag{2.19}$$

Here U is an $N \times p$ orthogonal matrix ($U^TU = I_p$) whose columns u_j are called the left singular vectors; V is a $p \times p$ orthogonal matrix ($V^TV = I_p$) with columns v_j called the right singular vectors, and D is a $p \times p$ diagonal matrix, with diagonal elements $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ known as the singular values. The columns of UD are called the principal components of X.

PCA *sees'* the data as a cloud of points and finds directions in which the samples to define Principal Components. This dispersion is measured with variance, and resulting PCA components are variance-ordered.

As nicely described in [181]

> *The first PC is the vector describing the direction of maximum sample dispersion. Each following PC describes the maximal remaining variability, with the additional constraint that it must be orthogonal to all the earlier PCs to avoid it containing any of the information already extracted from the data matrix. In other words, each PC extracts as much remaining variance from the data as possible. The calculated PCs are weighted sums of the*

> *original variables, the weights being elements of a so-called loadings vector. Inspection of these loadings vectors may help determine which original variables contribute most to this PC direction. However, PCs being mathematical constructs describing the directions of greatest dispersion of the samples, there is no reason for the loadings vectors to correspond to underlying signals in the data set. Most of the time, PCs are combinations of pure source signals, and do not describe physical reality. For this reason their interpretation can be fraught with danger.*

Especially in the context of the cell-type deconvolution, it can be imagine that different cell-types contribute to the variance but one PC could explain joint variance of many cell types.

Wang et al. [222] used SVD to compute matrix inversion in order to separate tumor from the stroma. The method was applied to tumor transcriptomes and gives purity estimation quite different from other popular enrichment-based method ESTIMATE [230].

Nelms et al. [139] in CellMapper uses a semi-supervised approach based on SVD decomposition to dissect human brain bulk transcriptome. Authors define a query gene (a specific known gene) and then they decompose transcriptome into components (eigenvectors) and multiply by weights that are higher for the components correlated with the query gene. Then the matrix is transformed back to gene $\times$ samples matrix but query signal is amplified. The point being to find marker genes that characterise the same cell-type as the query gene. Authors did not aim the identification of cell-type proportions or cell types profiles but identification of cell-type specific markers. They underline applicability of the method to rare cell types where many markers are not available. This approach was proposed by authors to be used in order to prioritize candidate genes in disease susceptibility loci identified by GWAS.

2.3.6.2 Non-negative matrix factorisation

Non-negative matrix factorization [188] is an alternative approach to principal components analysis. It assumes that data and components are non-negative. It finds its application in image analysis and gene expression analysis where data are indeed non-negative. The $N \times p$ data matrix X is approximated by

$$X \approx WH$$

where W is $N \times r$ and H is $r \times p, rmax(N, p)$. We assume that x_{ij} ,w_{ik}, $h_{kj} \geq 0$.

Which is a special case of Eq. (2.1).

The matrices W and H are found by maximizing

$$\mathcal{L}(W,H) = \sum_{i=1}^{N}\sum_{j=1}^{p}[x_{ij}log(WH)_{ij}(WH)_{ij}] \tag{2.20}$$

This is the log-likelihood from a model in which x_{ij} has a Poisson distribution with mean $(WH)_{ij}$.

This formula can be maximized through minimization of divergence:

$$\min_{W,H} f(W,H) = \frac{1}{2}\|X - WH\|_F^2 \tag{2.21}$$

Where $\|.\|_F$ is Forbenius norm, which can be replaced by Kullback-Leibler divergence.

The optimization can be done employing different methods:

- **euclidean** update with multiplicative update rules, it is the classic NMF [188]

$$\begin{aligned} W &\leftarrow W\frac{XH^T}{WHH^T} \\ H &\leftarrow H\frac{W^TX}{W^TWH} \end{aligned} \tag{2.22}$$

- **alternating least squares** [154] where the matrices W and H are fixed alternatively
- **alternating non-negative least squares** using projected gradients (Lin [121])
- **convex-NMF** [48] imposes a constraint that the columns of W must lie within the column space of X, i.e. $W = XA$ (where A is an auxiliary adaptative weight matrix that fully determines W), so that $X = XAH$. In this method only H must be non-negative.

The NMF algorithms can differ in initialization method as well and even in situations where $X = WH$ holds exactly, the decomposition may not be unique. This implies that the solution found by NMF depends on the starting values. The performance of different combinations applied to MRS data from human brain tumours can be found in Ortega-Martorell et al. [153].

Brunet et al. [27] created a NMF matlab toolbox and demonstrated applicability of NMF (using Kullback-Leibler divergence and euclidean multiplicative update [113] to cancer transcriptomes with focus on cancer subtyping (focusing on H matrix). Brunet et al. [27] also proposed a way to evaluate optimal number of factors (sources) to which matrix should be decomposed.

NMF as imposing non-negaivity in the context of decomposition of transcriptomes

seems as an attractive concepts as both cell profiles and cell proportions should be non-negative. It is not suprising then that some authors used NMF to perform cell-type deconvolution.

To my knowledge, *deconf* [175] was the first tool proposing NMF cell-type deconvolution of PBMC transcriptome, of considerable dimensions, 80 samples (40 control and 40 case) of Tuberculosis. Repsilber et al. [175] employed random initialization and alternating non-negative least squares to minimize the model divergence. The complete deconvolution of the transcriptome was used to perform DEG analysis on the deconvoluted profiles.

Shen-Orr and Gaujoux [190], not only presented exhaustive literature review through implementing cell-type deconvolution methods in a R package *CellMix* [69] but also proposed a semi-supervised NMF for cell-type deconvolution and published an R package implementing different NMF methods [70]. The semi-supervised version of NMF proposed by Gaujoux and Seoighe [69], need a set of specific marker genes for each desired cell type. Then at initialization and after each iteration of the chosen NMF algorithm (applies to some versions of NMF [188, 27, 161], "each cell type signature has the values corresponding to markers of other cell types set to zero. The values for its own markers are left free to be updated by the algorithm's own iterative schema". Applying their algorithm to *in vitro* controlled dataset [GSE11058 [3], testing selected NMF implementations and varying number of markers per cell, authors observed the best performance with guided version of *brunet* [27] implementation.

Moffitt et al. [135] applied NMF to separate tumor from stroma in pancreatic ductal carcinoma (PDAC) using multiplicative update NMF. They scaled H matrix rows to 1 so that the values correspond to the proportions. Authors tested different possibilities of number of sources (k), the final number of factors was defined through hierarchical clustering on gene-by-gene consensus matrix of top 50 genes of each component.

Finally Liu et al. [123] proposed post-modified NMF in order to separate *in vitro* mixtures of different tissues.

In brief, NMF is a popular, in biology, algorithm performing source separation with non-negativity constraint. It was applied to *in vitro* cell-mixtures and blood transcriptome showing satisfying accuracy of cell-type *in silico* dissection and evaluating proportions. It was also applied in cancer context, however it did not recover cell-type specific signals but rather groups of signals that could be associated to cancer or stroma.

2.3.6.3 Independent Components Analysis

Independent Components Analysis is written as in Eq. (2.1) with assumption that columns of S: S_i are *independent* and *non-Gaussian*, which adds orthogonality condi-

tion to A, since S also has covariance I. It was first formulated by Herault and Jutten [90]

The independence can be measured with entropy, kurtosis, mutual information or negentropy measure $J(Y_j)$ [97] defined by

$$J(Y_j) = H(Z_j)H(Y_j) \tag{2.23}$$

where Z_j is a Gaussian random variable with the same variance as Y_j . Negentropy is non-negative, and measures the deviation of Y_j from Gaussianity. It is used in a popular implementation of **FastICA** [97]. Other existing implementations of ICA are

- Infomax [18] using Information-Maximization that maximizes the joint entropy
- JADE (Cardoso and Souloumiac [30]) on the construction of a fourth-order cumulants array from the data

However, they are usually a lot slower which limits their application to big corpus of data and Teschendorff et al. [206] demntrated that *FastICA* gives the most interpretable results.

Therefore, I will focus on FastICA implementation as it will be extensively used in the Results part.

FastICA requires *prewhitening* of the data (centering and whitening). Centering is removing mean from each row of X (input data). Whitening is a linear transformation that columns are not correlated and have variance equal to 1.

Prewhitenning

1. Data centering
$$x_{ij} \leftarrow x_{ij} - \frac{1}{M}\sum_{j'} x_{ij'} \tag{2.24}$$
x_{ij}: data point
2. Whitenning
$$\mathbf{X} \leftarrow \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{X} \tag{2.25}$$
Where $\mathbf{X}$ - centered data, $\mathbf{E}$ is the matrix of eigenvectors, $\mathbf{D}$ is the diagonal matrix of eigenvalues

However, the results of this algorithm (Alg. 1. (2.26)) are not deterministic, as the $\mathbf{w}_p$ initial vector of weights is generated at random in the iterations of fastICA. If ICA is run multiple times, one can measure **stability** of a component. Stability of an independent component, in terms of varying the initial starts of the ICA algorithm, is a measure of internal compactness of a cluster of matched independent components produced in multiple

Algorithm 1 FastICA multiple component extraction

Input: K Number of desired components

Input: $X \in \mathbb{R}^{N \times M}$ Prewhitened matrix, where each column represents an N-dimensional sample, where $K \leq N$

Output: $A \in \mathbb{R}^{N \times K}$ Un-mixing matrix where each column projects $\mathbf{X}$ onto independent component.

Output: $S \in \mathbb{R}^{K \times M}$ Independent components matrix, with M columns representing a sample with K dimensions.

(2.26)

for $p \leftarrow 1, K$ **do**
  $\mathbf{w}_p \leftarrow$ *Random vector of length* N
  while $\mathbf{w_p}$ *changes* **do**
    $\mathbf{w}_p \leftarrow \frac{1}{M} X g(\mathbf{w}_p^T X)^T - \frac{1}{M} g'(\mathbf{w}_p^T X) 1 w_p$
    $\mathbf{w}_p \leftarrow \mathbf{w}_p - (\sum_{j=1}^{p-1} \mathbf{w}_p^T \mathbf{w}_j \mathbf{w}_j^T)^T$
    $\mathbf{w}_p \leftarrow \frac{\mathbf{w}_p}{\|\mathbf{w}_p\|}$
  end while
end for
where $\mathbf{1}$ is a column vector of 1's of dimension M

Output: $A = [\mathbf{w}_1, ..., \mathbf{w}_K]$

Output: $S = \mathbf{A}^T \mathbf{X}$

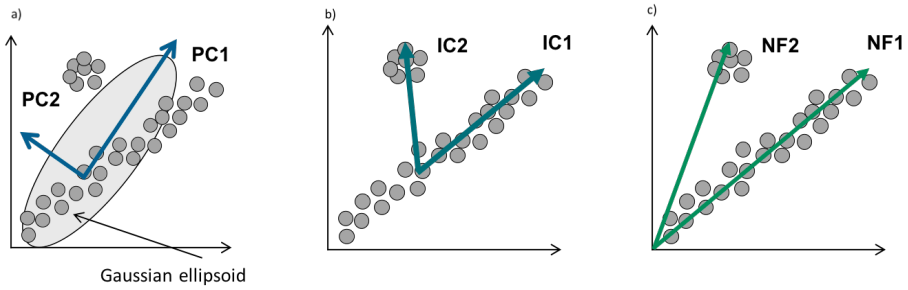


Figure 2.9: Simple illustration of matrix factorisation methods. Adapted with permision from [236]

ICA runs for the same dataset and with the same parameter set but with random initialization [92].

Icasso procedure can be summarized in a few steps :

1. applying multiple runs of ICA with different initializations
2. clustering the resulting components
3. defining the final result as cluster centroids
4. estimating the compactness of the clusters

In brief, ICA looks for a sequence of orthogonal projections such that the projected data look as far from Gaussian as possible. ICA starts from a factor analysis solution, and looks for rotations that lead to independent components.

So far, ICA was used to deconvolute transcriptomes into biological functions [23, 55, 78, 206, 236]. However, it have never been used for cell-type deconvolution.

In theory, ICA outputs: S could be interpreted as sources and A as proportions in the cell-type deconvolution context. In practice, the fact that the ICA allows negative weights of projections, it makes the interpretation less trivial.

To my knowledge my DeconICA R-package (that will be described in the results part) is the first method allowing interpretation of ICA-based signals as cell-type context-specific signatures and quantify their abundance in transcriptome.

Taken together, matrix factorisation methods are able to decompose a gene expression matrix into a weigheted set of genes (metagene)(S) and weigheted set of samples (metasample A). Discussed here PCA, NMF and ICA differ in constraints and starting hypotheses. PCA components are ordered by variance and are orthogonal in the initial space of data (Fig. 2.9). NMF impose non-negativy contraint and ICA independence of sources hypothesis. NMF and ICA do not have a prticular order. For all the matrix factorization methods number of components (or factors) (k) needs to be given to the algorithm. Some authors propose way to estimate optimal number of components

usualy justified in a specific context. NMF and SVD wera applied in the contex of cell-type deconvolution while ICA, so far, was used to dissect transcriptome into factors related to signaling pathwyas, technical biases or clinical feautres. In addition, ICA was proven to find reproducible signals between different datasets [29, 206]. I am going to discuss this aspect in the [Results] section.

2.3.7 Attractor metagenes

A method proposed by Cheng et al. [36], that can be run in semi-supervised or unsupervised mode, is called attractor metagenes. Authors describe their rationale as follows:

> *We can first define a consensus metagene from the average expression levels of all genes in the cluster, and rank all the individual genes in terms of their association (defined numerically by some form of correlation) with that metagene. We can then replace the member genes of the cluster with an equal number of the top-ranked genes. Some of the original genes may naturally remain as members of the cluster, but some may be replaced, as this process will "attract" some other genes that are more strongly correlated with the cluster. We can now define a new metagene defined by the average expression levels of the genes in the newly defined cluster, and re-rank all the individual genes in terms of their association with that new metagene; and so on. It is intuitively reasonable to expect that this iterative process will eventually converge to a cluster that contains precisely the genes that are most associated with the metagene of the same cluster, so that any other individual genes will be less strongly associated with the metagene. We can think of this particular cluster defined by the convergence of this iterative process as an "attractor" i.e., a module of co-expressed genes to which many other gene sets with close but not identical membership will converge using the same computational methodology.*

Which in pseudocode works as described in Algorithm 2 (2.27) and it is implemented in R code is available online in Synapse portal.

The produced signatures' weights are non-negative. In the original paper, the generation of tumor signatures leads to three reproducible signatures among different tumor types, including *leucocyte* metagene. Typically with the essential parmeter $\alpha = 5$, they discovered typically approximately 50 to 150 resulting attractors.

Attractor Metagenes algorithm can be seen as a variant of clustering approach where distance metric is mutual information between genes and metagenes are weighed average of gene expression. This method was further to study breast cancer [4] and to SNP data [54].

Algorithm 2 Attractor metagenes algorithm

Input: α shrinkage parameter
Input: $X \in \mathbb{R}^{N \times M}$ gene expression matrix
Output: m_j metagene of g_{seed}

(2.27)

$g_{seed} \leftarrow$ *a gene from* $1 : N$
$I^{\alpha}(g_{seed}; g_i)$ ▷ compute association beteen g_{seed} and g_i
$w_i = f(I^{\alpha}(g_{seed}; g_i))$ ▷ compute weights for each gene
$m_0 = \frac{\sum^{N} -1_{i=1}(g_i w_i)}{\sum^{N} -1_{i=1} w_i}$ ▷ compute metagene as weighted average of all genes
$I^{\alpha}(m_0; g_i)$ ▷ compute association between metagene m_0 and each gene g_i
repeat
  $w_i = f(I^{\alpha}(m_0; g_i))$
  $m_j = \frac{\sum^{N} -1_{i=1}(m_0 w_i)}{\sum^{N} -1_{i=1} w_i}$
until $m_{j+1} = m_j$

There is a possibility to tune the α parameter in order to obtain more or less metagenes that would be possibly interpretable as cell-type signatures.

2.3.8 Others aspects

Here I will discuss transversal aspects common to most of deconvolution methods. They play critical role in the final results and are often omitted while algorithms are published which impacts significantly the reproducibility.

2.3.8.1 Types of biological reference

Let us consider the most common case of the deconvolution where neither A or S are not known (Eq. (2.1)) and we would like to estimate cell proportions or both cell proportions and cell profiles. No matter if the method is supervised or unsupervised at some point of the protocol the biological knowledge about cell types is necessary in order to either derive the model or interpret the data. I discussed signatures from biological perspective in Section X. Here, I would like to stress the importance of the design of gene signatures which aim is to facilitate cell-type deconvolution.

Depending on chosen solution different type of reference can be used. In regression algorithms a proxy for purified cell profiles are necessary to estimate proportions. How-

ever, the genes that are the most variant between cell types are enough for regression and not all profiles are necessary. The choice of the genes and the number of the genes impacts significantly the outcome [211]. Therefore, most of regression methods come together with a new **basis matrix**, ranging from hundreds to tens of genes. Normally, genes selected for basis matrix should be cell-type specific in respect to other estimated cell types, validated across many biological conditions [93]. Racle et al. [172] adds a weight directly in the regression formula (see Eq. (2.9)) that corresponds to the variability of a signature gene between independent measurements of the same cell type so that the least inter-cell type variable genes have more weight in the model. *CellMix* [69] regroups different indexes to select the most specific genes based on signal-to-noise ratio. However, the most popular method is selection of differentially expressed genes between pure populations. Often criteria for optimal number of genes in the basis matrix are not knowledge-based but data-driven. Abbas et al. [3] uses condition number of basis matrix (*kappa*) in order to select the number of genes. The same approach is followed by CIBERSORT and many other regression methods. Newman et al. [141] also added another step while constructing the basis matrix, it preselects reference profiles having maximal discriminatory power. Some authors [103, 139] propose to find marker genes though correlation with a provided marker gene (a single one or a group of genes).

In enrichment methods, **gene list** can be enough to estimate cell abundance, sometimes (i.e. GSEA) ranked gene list is necessary. The choice of extremely specific markers is crucial for accurate call-type abundance estimation. The choice of markers can also be platform-dependent, this point is strongly underlined in [16]. Interesting possibility is the use of gene list of different *cell states* in order obtain coarse-grain resolution.

The impact of missing gene from a signature in the bulk dataset remains an unanswered question. It would be logical that shorter the gene list for a specific cell, a lack of a gene can have more impact on the result. There is a need of an accurate threshold between robustness and accuracy of the method.

In unsupervised methods, purified cell-profiles, signatures or list of genes can be used ***a posteriori*** to interpret the obtained sources. Even though the choice of reference do not affect the original deconvolution, it affects the interpretation. The advantage of *a posteriori* interpretation is a possibility to use different sources and types of signatures in order to provide the post plausible interpretation. It is common, that the way of interpretation of components is not included in the deconvolution tool [223, Newberg et al. [140], Moffitt et al. [135]], even though it is a crucial part of the analysis.

For the deconvolution of tumoral biopsies, most of reference profiles, up to now, are coming from the blood, which is the most available ressource. Therefore most of methods make a hypothesis that blood cell-type profiles/signatures are correct approximation of cancer infiltrating immune cells. Rare models like PERT [170] or ImmuneStates [211] discuss the perturbation of the blood-derived profiles in diseases.

With availability of single-cell RNA-seq of human cancers [39, 111, 118, 168, 184, 209, 234], we gain more knowledge on immune cells in TME and there is a growing evidence that they differ importantly from blood immune cells. Racle et al. [172] show that lymph node resident immune cells have expression profile closer to blood immune cells than cancer immune cells. Schelker et al. [184] shows, using a synthetic bulk dataset, that using single cell profiles with existing regression methods (CIBERSORT) can improve their performance in the cancer context. However, availability of scRNA-seq remains succinct and probably do not embrace the patient heterogeneity that can be found in big bulk transcriptome cohorts.

2.3.8.2 Data normalization

Data pre- and post-processing can have an important impact on the deconvolution. Many authors apply stong filetering of genes [223], removing probes with low and moderate expression as well as genes with the highest expression (potential outliers). In many cases, data preprocessing is not detailed and therefore impossible to reproduce.

There is also a debate on the data normalization. Most of authors suggest to use counts (not log-space) for estimating cell abundance as log) transformed data violate the linearity assumption [235], some opt against it [189, 41] and some envisage both possiblities [57, 175]. For the RNA-seq data TPM (transcriprs per milion) normalization is paciced or even required by most methods [35, 60, 172].

2.3.8.3 Validation

Most of algorithm validation starts with *in silico* mixtures (Fig. 2.10). In published articles, the bulk transcriptome is simulated in two ways (1) mixing numerically simulated sources at defined proportions of given distribution (i.e. uniform) using linear model (for instance NMF) (2) using sampling (for instance Monte Carlo) to randomly select existing pure profiles and mixing them (additive model) at random proportions. To the obtained bulk samples, noise can be added at different steps of the simulation. Additional parameters can be defined in *in silico* mixtures, for instance, CellMix allows defining number of marker genes (specific to only one source) for each cell type. The simulated benchmark based on single cell data was used in Schelker et al. [184] and Görtler et al. [79]. In this framework simulated data was obtained though summing single cell profiles at known proportions. The main pitfall of those methods is that in the proposed simulations the gene covariance structure is not preserved. In reality, the proportions of cell types are usually not random and some immune cell types can be correlated or anti-correlated. In addition, these simulations create a simple additive model which perfectly agrees with

the linear deconvolution model. This is probably not the case of the real bulk data affected by different factors as cell cycle, technical biases, patients heterogeneity and especially cell-cell interactions.

Naturally, algorithms validated with simulated mixtures are then validated with controlled *in vitro* mixtures of cell types or tissues mixed in known proportions. The most popular benchmark datasets are:

- mix of human cell lines Jurkat, THP-1, IM-9 and Raji in four different concentration in triplicates and the pure cell-line profiles (GSE11058) [3];
- mix of rat tissues: liver, brain, lung mixed in 11 different concentrations in triplicates and the pure tissues expression(GSE19830) [191]

Similar simple mixtures are proposed also by other authors [16, Kuhn et al. [106]]. This type of benchmark adds complexity of possible data processing and experimental noise. However, it still follows an almost perfect additive model as the cell/tissues do not interact and they are only constituents of the mixture.

Several tools performed systematic benchmark using PBMC or whole blood datasets, where for a number of patients (that can be over one hundred) FACS measured proportions of selected cell types and bulk transcriptomes are available. Many of such datasets can be found at IMMPORT database. Aran et al. [9] kindly shared with scientific community two datasets with considerable number of patients (~ 80 and ~ 110) and processed FACS data (actual proportions) on their github repository. It is still important to remember that liquid tissues are easier to deconvolute and for the tools using *a priori* reference, the reference profiles are obtained from the blood. Therefore the context remains consistent.

Fo the cancer solid tissues deconvolution, some of the tools were validated with the stained histopathology cuts using in situ-hybridisation (ISH) [106] or immunohistochemistry (IHC) (Becht et al. [16]). Often this method estimate a limited number of cell types and the measured abundance from pictures can also be biased by the technical issues (image/ staining quality).

Authors of EPIC validated thier tool with paired RNAseq and Lymph node FACS-derived proportions in 4 patients (GSE93722). They also noticed that it is more straightforward to correctly evaluate lymph node immune cell types than cancer infiltrating cell types as lymph node resident cells are more similar to the blood immune cells.

FACS and gene expression of blood cancer (Folicular lymphoma) was also used by Newman et al. [141] for a 14 patients (unpublished data). For solid tissues Newman et al. [141] used paired FACS and expression datasets of lung normal tissues for B-cell and CD8 and CD4 T cels of 11 patients (unpublished data).

Some authors proposed to cross validate estimated proportions with estimated propor-

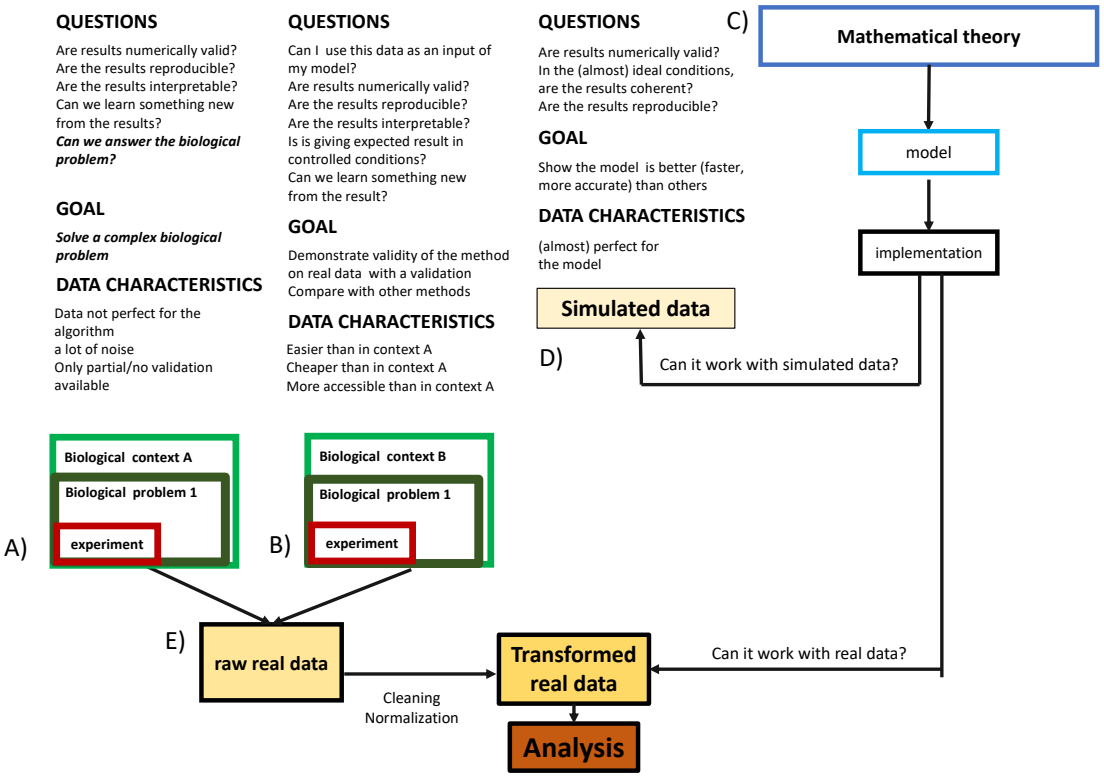


Figure 2.10: From theory to practice: simplified pipeline of model validation. The scheme reflects pipeline of data validation commonly used in transcriptome deconvolution methods validation. Project can be started from biological problem (A) and then a way to solve the problem in mathematical model (C) is tested. Most commonly, the project starts with the model (C) then it is tested on simulated data (D). Next level of difficulty is testing the model with real data, so called, benchmark data (B) that were generated in some biological context different from initial problem, they need to be usually normalized (E) before the model is challenged. B data are widely used as they are easily available and there is some validation available facilitating comparison. Lastly, it is assumed that if the method works fine in the context B, it will work as well in the context A, preferably accompanied by some partial validation. One can replace A by cancer transcriptomics, B by blood data or *in vitro* mixtures, if the focus is TME bulk transcriptomics deconvolution.

tions based on a different data input (i.e. methylome) [117, Şenbabaoğlu et al. [240]] or CNA (Şenbabaoğlu et al. [240]). This type of validation is interesting, even though in many projects only one type of data types is available for the same samples. TCGA data is one of few exceptions.

Finally, a validation of deconvolution of solid cancer tissues remains incomplete as no paried expression and FACS data is available up to date.

2.3.8.4 Statistical significance

Little number of tools propose a statistical significance assessment. CIBERSORT computes empirical p-values using Monte Carlo sampling. Infino authors [233] provide a confidence interval for the proportion estimations. This allows to know which proportion estimation are more trustful than other.

Most tools compare themselves to others measuring accuracy of the prediction, or Pearson correlation, on the benchmark datasets (described above). Often, in the idealized mixtures, methods perform well. Evaluation of their performance in cancer tissues remains unanswered without proper statistical evaluation.

2.3.9 Summary

The field of computational transcriptome deconvolution is constantly growing. Initially used to solve simple in vitro or simulated mixtures of quite distinct ingredients, then to deconvolute blood expression data, finally applied to solid cancer tissues. In cancer research digital quantification of cancer purity become a routine part of big cancer research projects [230]. Cell-type quantification, even though the validation framework and statistical significance of deconvolution tools can still be improved, seems to be considered as a popular part of analytical pipeline of bulk tumor transcriptomes [40]. Different types of approaches try to solve the deconvolution problem, focusing on different aspects of the quantification, or proposing methodologically different approaches. Methods proposing unsupervised solution to the deconvolution problem of transcriptomes are still underrepresented. All the tools assume a linear additive model without explicitly including impact of possible interactions on the cell-type quantification. The tools that met the biggest success were proven by the authors to be easily applicable to a variety of cancer datasets and reusable without an extra effort (through a programming library or web interface). The field is still waiting for a gold standard validation benchmark that would allow a fair comparison of all the tools in solid cancer tissues. It is also remarquable that the recent methods focus on quantification of abundance of average representation of cell-types without aspiring to deconvolute the cell-type context-specific profiles.

Thanks to diverse cancer single cell data and big-scale projects [174], we will be able to improve the existing deconvolution approaches and finally replace the collection of bulk transcriptomes by a collection of scRNA-seq ones.

2.4 Deconvolution of other data types

The transcriptome data is not the unique omic data type that can be used to infer cell type proportions. Genomic and epigenomic data was used in numerous publications to perform cell-type deconvolution or estimate sample purity. I will present a general landscape of the tools and methods used for this purpose.

2.4.1 DNA methylation data

Cell-type composition can be computed from DNA methylation data (described in Section X). In EWAS (Epigenome Wide Assotication Studies) variation origination from cell types is considered as important confounding factor that should be removed before comparing cases and controls and defining Differentialy Methylated Positions (DMPs). Teschendorff and Zheng [205] reviewed 10 tools for epigenome deconvolution. Six of the described methods are identified by authors as reference-free (which I called in this Chapter *unsupervised*), three are regression-based and one is semi-supervised. Another review on this topic was authored by Titus et al. [210].

Unsupervised methods employed in methylome cell-type deconvolution are RefFreeEWAS [96], SVA [114] are based on SVD, ISVA based on ICA [Teschendorff et al. [207]) are more general methods that aim to detect and remove confounders from the data (that do not need to be necessary the cell types). RUVm [127] is semi-supervised method using generalized least squares (GLS) regression with negative reference also used to remove *unwanted variation* from the data and could be potentially adapted to cell-type deconvoltion. EWASher (Zou et al. [237]) is linear mixed model and PCA based method that corrects for cell-type composition. Similarly, ReFACTor [173] use sparse PCA to remove the variation due to cell-type proportions. Houseman et al. [94] proposed RefFreeCellMix: a NMF model with convex constraints to estimate factors representing cell types and cell-type proportions and a likelihood-based method of estimating the underlying dimensionality (k number of factors). A different tool MeDeCom [125] uses alternating non-negative least squares to fit a convex hull.

As far as supervised methods are concerned, EPiDISH (*E*pigenetic *D*issection of *I*ntra-*S*ample-*H*eterogeneity) R-package [205] includes previously published tools: Quadratic programming method using reference specific DNAse hypersensitive sites [Constrained

Projection (CP) [95]), adapted to methylome deconvolution CIBERSORT algorithm (ν-SVR) and robust partial correlations (RPC) method (a form of linear regression). Reference cell-type specific profiles were obtained from the blood.

eFORGE [26] can detect in a list of DMPs if there is a significant cell-type effect.

EDec [152] uses DNAm to infer factors proportions using NMF and then derives factors profiles though linear regression of transcriptome data of cancer datasets. Authors identify tumor and stroma compartments and profiles. However, they admit the error rate for profiles is quite high for most genes.

Wen et al. [225] focused on intra-tumor heterogeneity (clonal evolution) based on DNAm data. Profiles obtained from cell lines were used in a regression model to identify the proportions of sub-clones in breast cancer data. InfiniumPurify [234] and LUMP [8] uses DNAm to estimate sample purity.

Validation framework for methylation deconvolution is very similar to transcriptome ones: in silico mixtures and FACS-measured proportions of the blood. Most of the tools assume the cell composition is a factor the most contributing to the variability and therefore SVD/PCA based approaches are sufficient to correct for the variability. According to Teschendorff and Zheng [205] this assumption was not proven to hold true in solid tissues like cancer. Supervised methods have the same drawback as in the case of transcriptome, they use purified profiles from one context to derive cell proportions in a different context. In overall, it seems that there was no study that proposed cell-type quantification based on methylome profiles in pan-cancer manner.

2.4.2 Copy number aberrations (CNA)

To my knowledge there is no method using CNA data in order to estimate cell-type composition, as CNA occur in tumor tissue and natural distinction can be made between tumor and normal cells and within tumor cells (intra-tumor).

Therefore, copy number aberrations can be used to estimate tumor purity and clonality. BACOM 2.0 [64], ABSOLUTE [31], CloneCNA [231], PureBayes [110], CHAT [116], ThetA [146], SciClone [132], Canopy [102], PyClone [179], EXPANDS [5] estimate tumor purity and quantify true copy numbers. OmicTools website reports 70 tools serving this purpose and their review goes beyond the scope of my work. Most of tools use tumor and normal samples, paired if possible. QuantumClone seem to be the only tool that requires a few samples from the same tymor (in time or space dimension).

Aran et al. [8] published Consensus measurement of purity estimation that combines purity estimations based on different data types (available in cBioportal) using: ESTIMATE [230] (gene expression data), ABSOLUTE [31] (CNA), LUMP [8] (DNAm and IHC of stained

slides. Authors concluded that the estimation based on different data types highly correlate with each other, besides the IHC estimates, which suggests that IHC provides a qualitative estimation of purity.

2.5 Summary of the chapter

A plethora of machine learning solutions have been developed to solve problems of different nature. Supervised and Unsupervised approaches can be distinguished depending if a model is provided set of training data with known response or the algorithm works blindly trying to find patterns in the data. Some of the algorithms found an important application in healthcare and are included in clinical routine.

One of important problems that can, in theory, be solved with ML, is bulk omic data deconvolution. Different types of deconvoluton of cancer samples can be distinguished: clonal, purity and cell-type deconvolution. Here I focused on cell-type deconvolution of transcriptome data. Through an extensive review I presented 64 tools and divided them in categories depending on adapted type of approach. I distinguished probabilistic, enrichment-based, regression-based, convex hull, matrix factorisation and attractor metagene approaches that can be used for cell-type deconvolution. I detailed basis of the different models and highlighted the most important features counting for cell-type deconvolution.

DNAm data were also used to estimate cell-type proportions. However, the heterogeneity found in methylome data resulting from difference in cell type proportions is usually seen as a confounding factor to be removed. CNA data can be used for estimation of tumor purity and clonality.

In brief, for the transcriptome cell-type deconvolution, it can be observed that just a limited number of tools are usable in practice in order to deconvolute big cancer cohorts and without need to provide hard to estimate parameters. Supervised methods appleid to cancer use reference-profiles coming from different context. Unsupervised tools, so far, are rather underrepresented in the field and do not offer a solution directly applicable to cancer transcriptomes of high dimensions. All of the presented methods are still waiting for consistent validation with gold standard benchmark. This could be done if systematic data of bulk transcriptome paired with FACS-measured cell-type proportions information for many cell and in many samples were generated. Another unanswered question is the validity of the linear mixture model in the presence of cell-cell interactions.

There is still a room for improvement in the field in order to provide more user-friendly, accurate and precise cell-type abundance estimations.

A question can be asked, **are cell-type proportions enough to understand tumor immune phenotypes?** Can we extract more valuable information from the bulk omic data that would give use insight to biological functions of the *in silico* dissected cells?

Objectives

In the introduction, I have described two sides of studying TME complexity. I placed in the context of cancer research and cancer therapy the most recent studies of tumor immunity with focus on system-level computational approaches. I have also introduced a wide array of available approaches to address deconvolution of bulk omic data. I reviewed thier strong and weak points and I presented general trends since the field was established.

Answers to important questions on *how TME modulates tumor*, *how to propose better cancer subtyping for immune therapies* and *how to better predict a response to treatment* are perhaps **hidden in already generated bulk omic data**. However, new methodological tools and more overall view is needed to better uncover hidden patterns.

In this thesis I aim to bring new insights into composition and function of TME. It is clear that complex information is necessary to understand the role of different immune cells in cancer and not only presence but also function are to be deciphered from available data. Therefore, this project, on its biological side, has two main aims:

1. fundamental research: understand presence of different cell type, their interactions and functions in TME of different cancers types and how other factors as stress, cell cycle etc. shape them. Thanks to data-driven and discovery nature of the project, I will also hope to understand how signature of cell type evolves in different conditions shaped by other cells and factors.
2. translational research: how immune landscape and its state can help to predict patient survival and better tailor recommendation for therapy. The analysis could also bring to the light possible biomarkers or drug targets for immune therapies.

I aim to explore publicly available data, challenge inter-lab and inter-platform biases. I will use mainly bulk transcriptomic data (because of accessible volume) and cross-validate with other data types: scRNA-seq, FACS, IHC when possible.

On its computational/mathematical side it will face following challenges:

1. Establish state-of-art of existing bulk deconvolution methods, discuss their advan-

tages and limits
2. Propose new unsupervised method that will fill the knowledge gap giving an insight into context-specific signatures of cell types/cell states in cancer
3. Deliver well-documented and user friendly tool that can be used by the scientific community
4. Decompose a big corpus of bulk omic data into interpretable biological functions, with a particular focus on the immune cell types
5. Use different data types (scRNAseq, microarray, RNAseq, FACS etc.) to complete, compare and contrast findings of the analysis.
6. Decompose established immune cells populations from metastatic melanoma in order to better understand cell-type heterogeneity

In order to face these challenges, I have first focused on testing and creating new methods. This is why methods and results are interlaced in this thesis. Reproducing work of other researchers it is not always easy, sometimes it is even impossible. A lot of time was invested to understand and reuse previous publications, part of this effort was reflected in the introduction, some of my thoughts will be expressed in the discussion.

Next important step was improving and testing ideas born in our team. I collaborated to a publication on a topic Chapter 3 and I have authored an extension of this work described in Chapter 4 . I have also compared my tool to other similar method, an overview of the results are in Chapter 5.

Once I have found the most appropriate way to apply my method, that I validated with multiple datasets, I have build a tool to share it with scientific community. The tool is freely available online as an R package. During my work, I have collected many datasets of tumor signatures, tumor metagenes, benchmark datasets some of which are part of my tool. I have also accessed, thanks to curtesy of our collaborators a collection of pan-cancer bulk transcriptomic datasets that I competed with other publicly available datasets. I build my working environment in which I managed and cleaned the data.

Finally, I realized a pan-cancer analysis of over 100 datasets which is the main outcome of my work. I completed results of this work with published scRNAseq data from tumor samples. This analysis is a source of many information, I have, so far, explored only part of possible direction focusing on cancer infiltrating T-cells. This results will be find in a manuscript in preparation in Chapter 6. However, more information can still be extracted in the further work. There is also a possibility to provide an experimental validation to my finding and it will be considered in the discussion part.

In parallel, I used part of methods to study cell-type heterogeneity in an independent project resulted in a submitted publication (Chapter 8).

The remaining time, I have invested into collaborations and contributions to different works within and outside of my team. Published work from those projects will be shortly

described in Annexes.

Part II

Results

Chapter 3

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

Ulykbek Kairov*, Laura Cantini*, Alessandro Greco, Askhat Molkenov, __Urszula Czerwinska__, Emmanuel Barillot and Andrei Zinovyev

Here is text that places article in the context

Background

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

Results

Here we address the question of optimizing the number of statistically independent components in the analysis of transcriptomic data for reproducibility of the components in multiple runs of ICA (within the same or within varying effective dimensions) and in multiple independent datasets. To this end, we introduce ranking of independent components based on their stability in multiple ICA computation runs and define a distinguished number of components (Most Stable Transcriptome Dimension, MSTD) corresponding to the point of the qualitative change of the stability profile. Based on a large

body of data, we demonstrate that a sufficient number of dimensions is required for biological interpretability of the ICA decomposition and that the most stable components with ranks below MSTD have more chances to be reproduced in independent studies compared to the less stable ones. At the same time, we show that a transcriptomics dataset can be reduced to a relatively high number of dimensions without losing the interpretability of ICA, even though higher dimensions give rise to components driven by small gene sets.

Conclusions

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

[104]

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

BMC Genomics

RESEARCH ARTICLE Open Access

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

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

Abstract

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

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

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

Keywords: Transcriptome, Independent component analysis, Reproducibility, Cancer

Background

Independent Component Analysis (ICA) is a matrix factorization method for data dimension reduction [1]. ICA defines a new coordinate system in the multi-dimensional space such that the distributions of the data point projections on the new axes become as mutually independent as possible. To achieve this, the standard approach is maximizing the non-gaussianity of the data point projection distributions [1]. ICA has been widely applied for the analysis of transcriptomic data for blind separation of biological, environmental and technical factors affecting gene expression [2–6].

The interpretation of the results of any matrix factorization-based method applied to transcriptomics data is done by the analysis of the resulting pairs of metagenes and metasamples, associated to each component and represented by sets of weights for all genes and all samples, respectively [6, 7]. Standard statistical tests applied to these vectors can then relate a component to a reference gene set (e.g., cell cycle genes), or to clinical

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

annotations accompanying the transcriptomic study (e.g., tumor grade). The application of ICA to multiple expression datasets has been shown to uncover insightful knowledge about cancer biology [3, 8]. In [3] a large multi-cancer ICA-based metaanalysis of transcriptomic data defined a set of metagenes associated with factors that are universal for many cancer types. Metagenes associated with cell cycle, inflammation, mitochondria function, GC-content, gender, basal-like cancer types reflected the intrinsic cancer cell properties. ICA was also able to unravel the organization of tumor microenvironment such as the presence of lymphocytes B and T, myofibroblasts, adipose tissue, smooth muscle cells and interferon signaling. This analysis shed light on the principles underlying bladder cancer molecular subtyping [3].

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

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

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

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

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

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

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

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

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

Results

Definition of component reproducibility measures used in this study

Stability of an independent component, in terms of varying the initial starts of the ICA algorithm, is a measure of internal compactness of a cluster of matched independent components produced in multiple ICA runs *for the same dataset and with the same parameter set but with random initialization*. The exact index used for quantifying the clustering is documented in the Methods section. Conservation of an independent component in terms of choosing various orders of ICA decomposition is a correlation between matched components computed in two ICA decompositions of different orders (reduced data dimensions) *for the same dataset*. Reproducibility of an independent component is an (average) correlation between the components that can be matched after applying the ICA method using the same parameter set but *for different datasets*. For example, if a component is reproduced between the datasets of the same cancer type, then it can be considered a reliable signal less affected by technical dataset peculiarities. If the component is reproduced in datasets from many cancer types, then it can be assumed to represent a universal cancerogenesis mechanism, such as cell cycle or infiltration by immune cells. The details on computing correlations between components from different datasets are described in Methods.

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

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

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

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

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

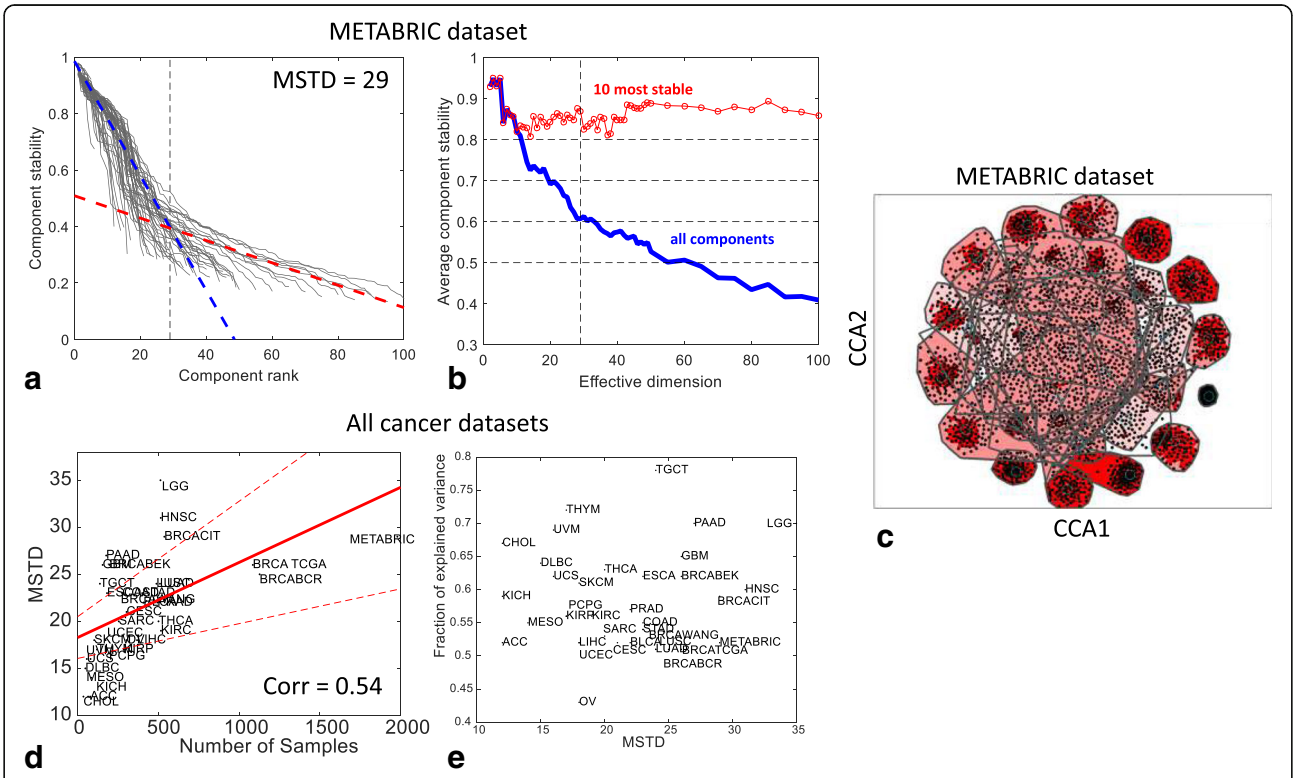


Fig. 1 Defining Maximally Stable Transcriptomic Dimension (MSTD) value in 37 transcriptomics cancer datasets (13,027 samples in total). In **a**-**c**) an example of the analysis is presented for the largest breast cancer dataset METABRIC. **a** stability profiles for ICA decompositions in various dimensions (from 2 to 100) shown by *grey lines*. Two-*line* clustering result is shown by *blue* and *red dashed lines*, with MSTD determined as the point of their intersection (*vertical dashed line*). **b** average stability profile S_M^{total} (*blue line*) and the average stability of 10 most stable components $S_M(10)$ (*red line*). **c** visualizing the results of computing ICA 100 times with MSTD = 29 components in the METABRIC dataset and component clustering (*icasso* package, Canonical Correlation Analysis (CCA) plot). Each *black point* represents a component, *red lines* show significant correlations between them, polygons show the convex hull area of the clusters. **d** Dependence of MSTD on the number of samples for all datasets. **e** Dependence of the fraction of explained variance on MSTD for all datasets

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

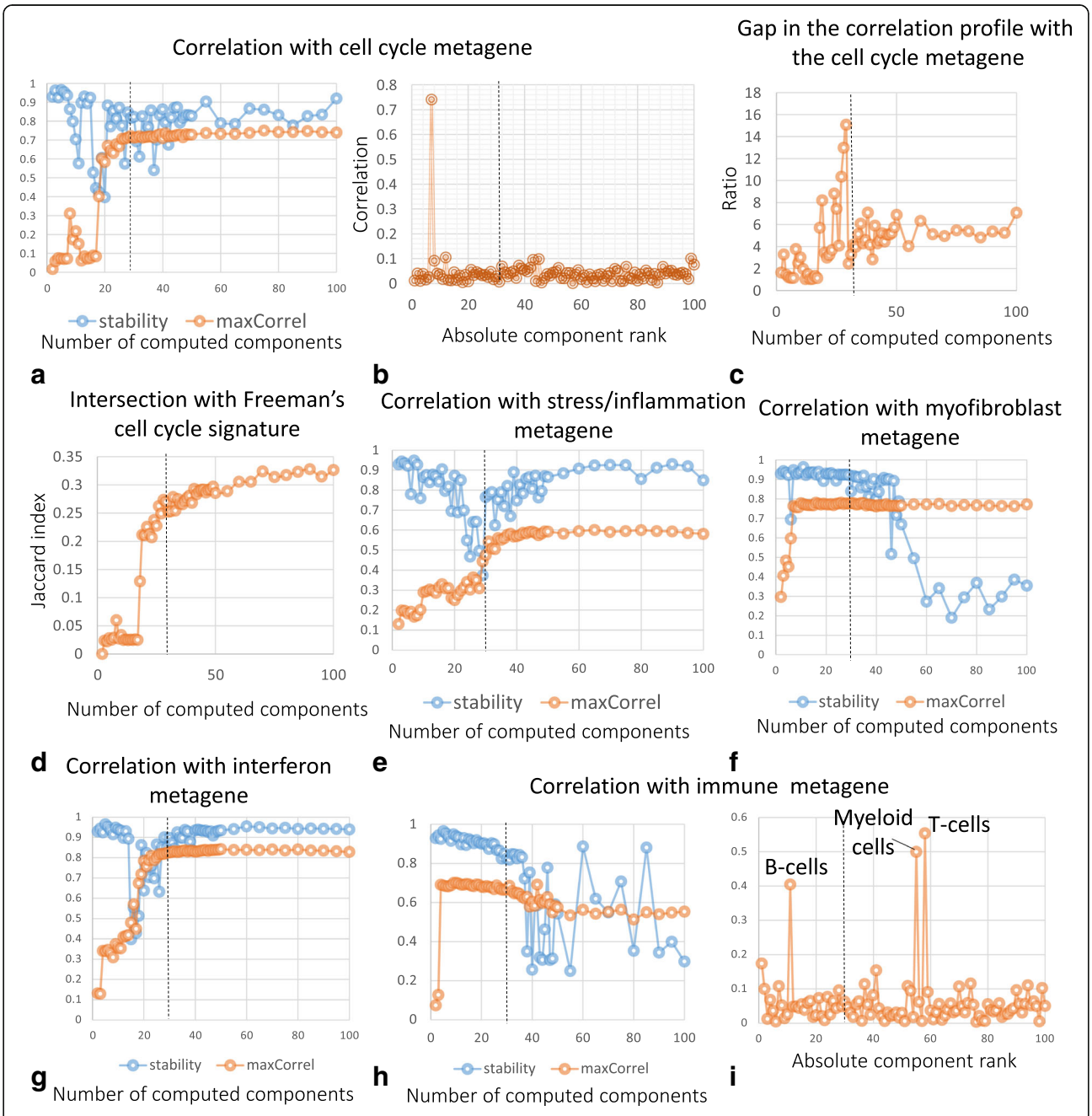


Fig. 2 Analysis of reproducibility of previously identified metagenes in independent components of the METABRIC dataset. **a**,**e**,**f**,**g**,**h** show correlations (see Methods section) with the cell cycle, inflammation, myofibroblast, interferon signaling, immune-related metagenes from [3] as a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

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

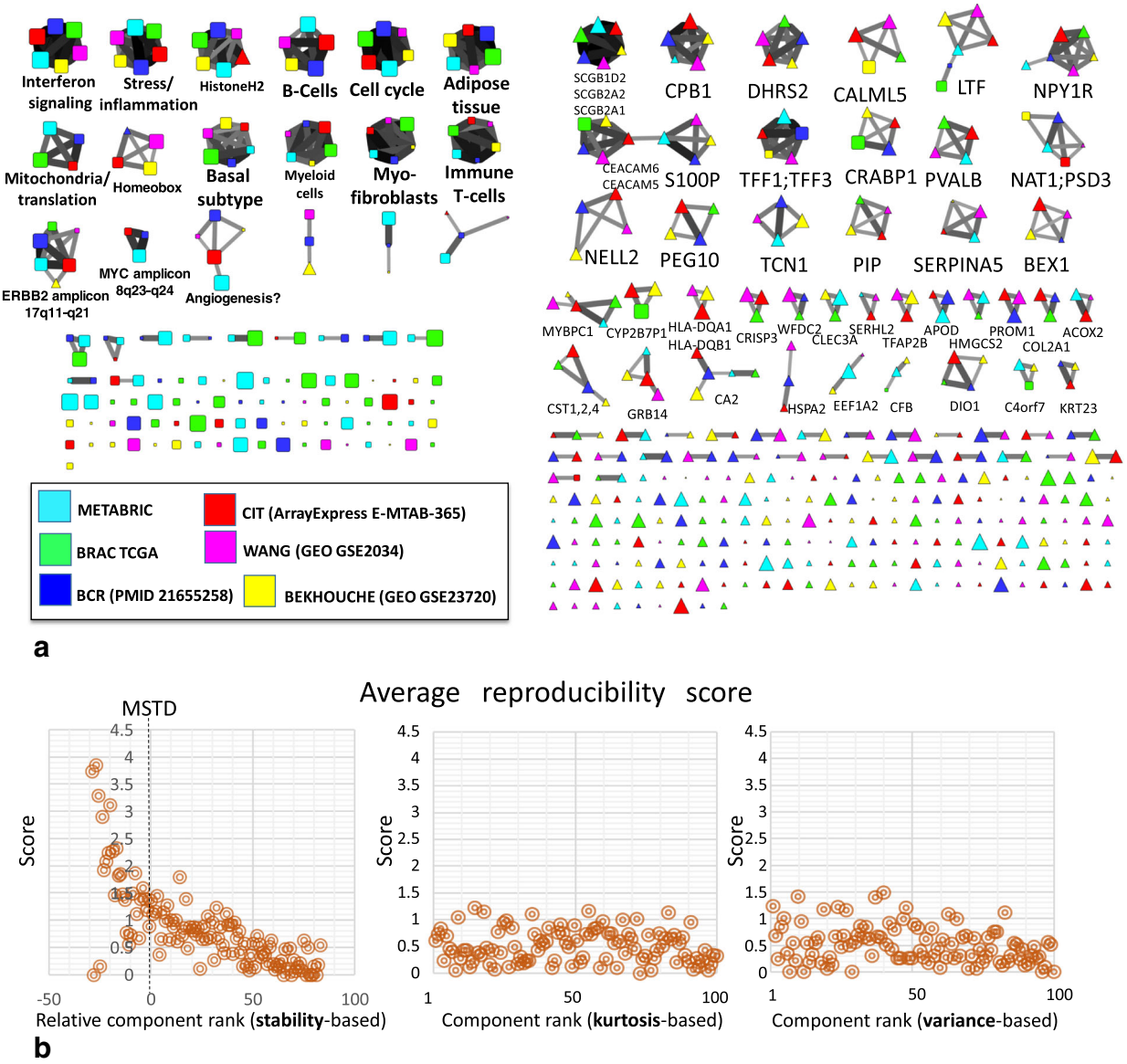


Fig. 3 Analysis of component reproducibility in independent datasets. **a** Graph of reciprocal correlations showing the reproducibility of the metagenes of independent components in 6 independent breast cancer datasets. Each node here is an independent component, represented by a metagene, from an ICA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

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

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

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

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

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

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

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

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

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

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

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

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

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

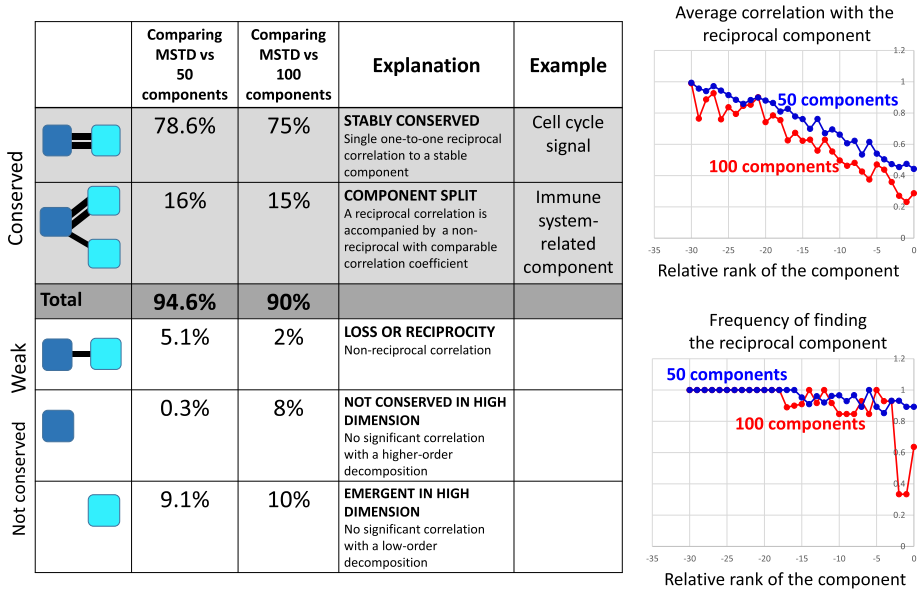


Fig. 4 Conservation or non-conservation of independent components in higher than MSTD dimensions. ICA decomposition in MSTD data dimension is compared with the one of dimension 50 and 100. Table on the left: list of different scenario with the relative frequency of each of them estimated for a pan-cancer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

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

Discussion

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

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

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

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

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

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

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

Conclusion

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

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

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

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

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

Methods

Transcriptomics cancer data used in the analysis

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

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

ICA decompositions computation

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

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

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

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

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

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

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

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

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

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

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

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

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

Comparing metagenes computed for different datasets and in different analyses

Following the methodology developed previously in [3], the metagenes computed in two independent datasets were compared by computing a Pearson correlation coefficient between their corresponding gene weights. Since each dataset can contain a different set of genes, the correlation is computed on the genes which are common for a pair of datasets. Note that this common set of genes can be different for different pairs of datasets. The same correlation-based comparison was done with previously defined and annotated metagenes. We computed the correlation only between those genes having projection value more than 3 standard deviations in the identified component.

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

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

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

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

Determining if a small gene set is driving an independent component

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

Additional files

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

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

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

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

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

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

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

Abbreviations

IC: Independent Component; ICA: Independent Component Analysis

Acknowledgements

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

Funding

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

Availability of data and materials

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

Authors' contribution

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

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Publisher's Note

Author details

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

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

References

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

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

Chapter 4

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

Urszula Czerwinska, Laura Cantini, Ulykbek Kairov, Emmanuel Barillot, Andrei Zinovyev

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

[46]

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

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

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

AQ1

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

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2 Methods

2.1 ICA Overdecompostion Procedure

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

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

2.2 Interpretation of Components

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

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

2.3 Selecting Immune-Related Components

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

2.4 Comparing Independent Components from Different Datasets

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

3 Results

3.1 Most of Known Metagenes Can Be Found in Overdecomposed Datasets

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

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

3.2 Reproducibility of the Signals in Breast Cancer Datasets

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

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

3.3 Three Pseudo-cliques Related to Three Immune Cell Types

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

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

4 Discussion

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

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

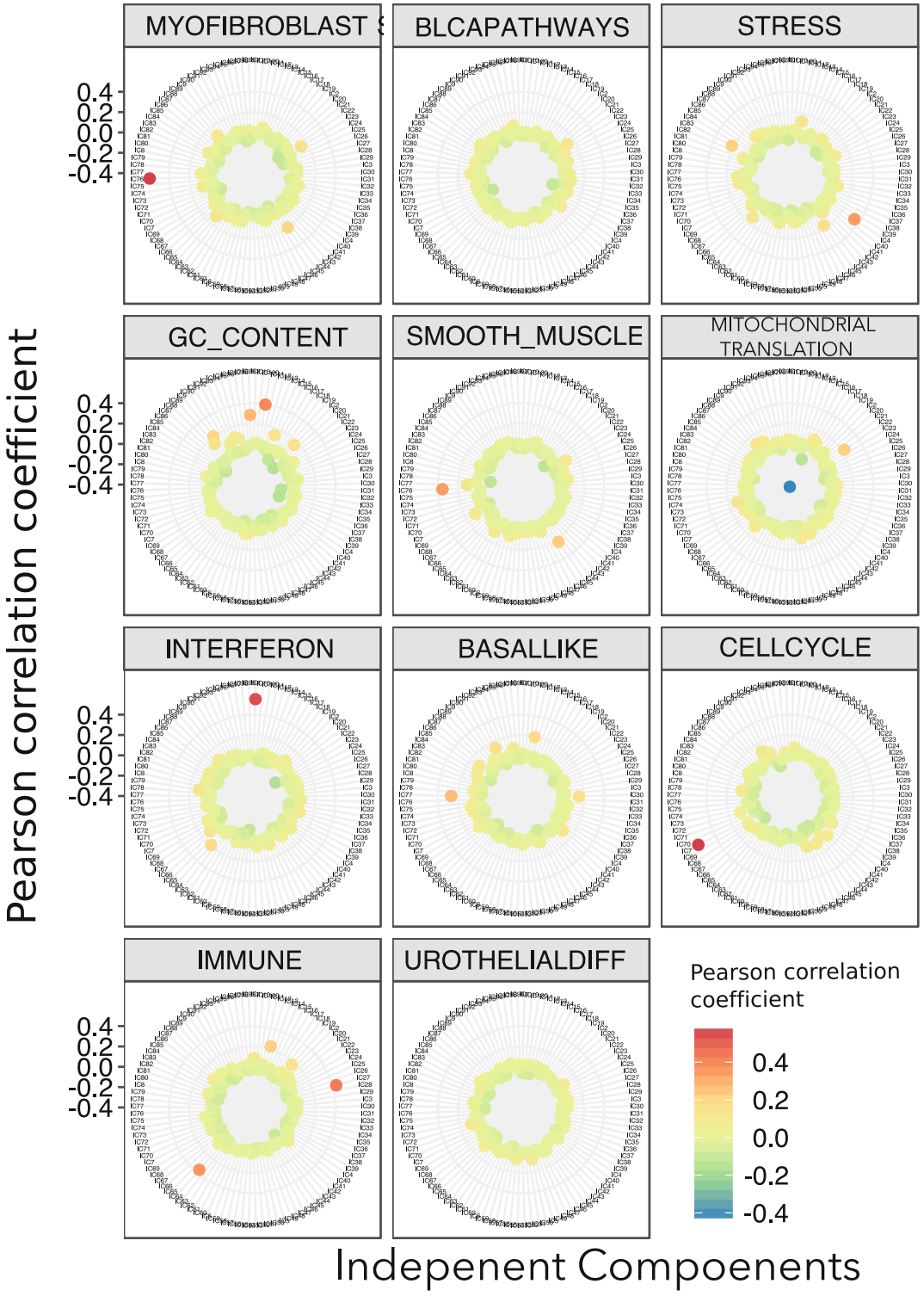


Fig. 1. Correlations between 11 metagenes [11] **and 100 independent components of METABRIC dataset** [17]. Each panel shows correlation coefficients between a given metagene and 100 ICs of METABRIC, the components are ordered in the same manner for all panels from 1 to 100 in a circle. For a high correlation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

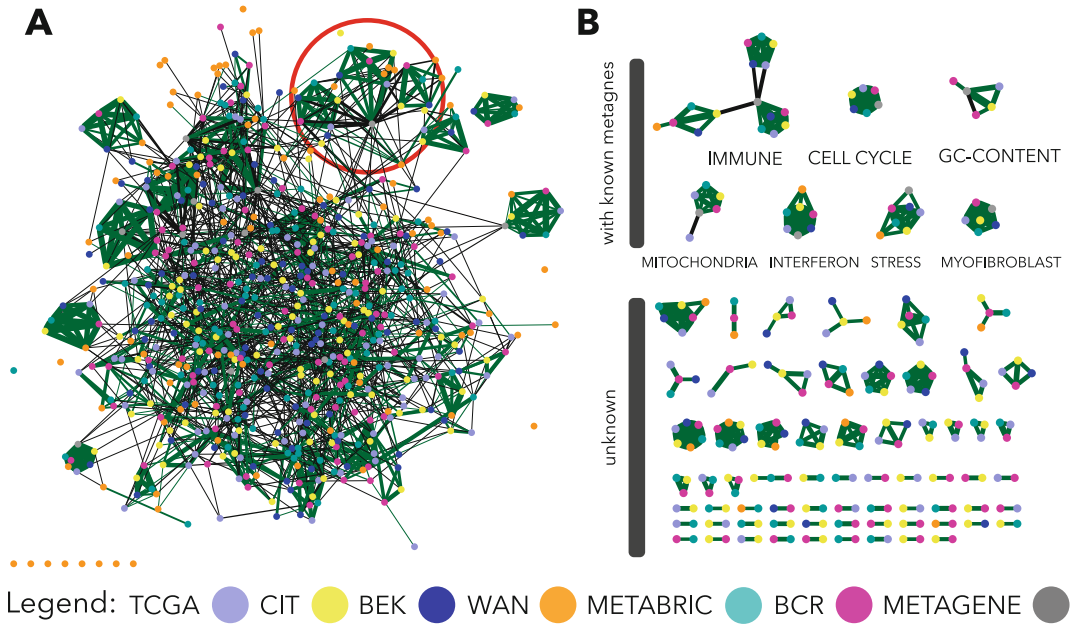


Fig. 2. Correlation plot of six tumor datasets and the reference metagenes [11] A- Correlation graph between decompositions into 100 ICs of the six transcriptomic datasets and the 11 reference metagenes. The IMMUNE metagene and related ICs in encircled; B - collection of pseudo-cliques extracted from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

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

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

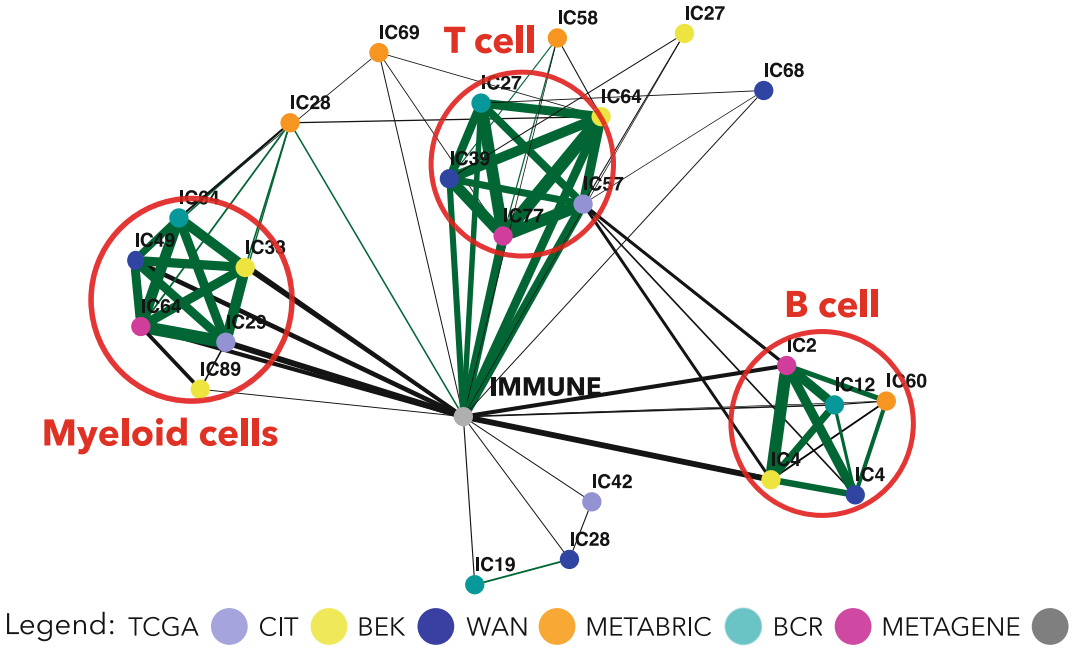


Fig. 3. Correlation graph of ICs correlated with IMMUNE metagene > 0.1. Three pseudo-cliques are encircled and labeled according to the results of Fisher exact test. The thickness of edges is proportional to the Pearson correlation coefficients, green color indicates reciprocal edges, colors of nodes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

5 Conclusions

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

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

References

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

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

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

Chapter 5

Comparison of reproducibility between NMF and ICA

NMF and ICA are both algorithms often applied to solve blind source deconvolution problem. NMF gained a popularity as a tool of transcriptomic analysis mainly thanks to the publications [publicaition_list]. However, the non-negativity contraint, an attractive concept in the case of non-negative transcriptome counts, may be a reason why the reusults of NMF decomposition are not the best candidate for our deconvolution task. We observed that NMF-based metagenes are less reprouctible between different transcriptomic datasets than ICA-based metagenes.

5.0.1 Comparing metagenes obtained with NMF vs ICA.

We compared the reproducibility of NMF and ICA through decomposition of four breast cancer datasets (BRCATCGA, METABRIC, BEK, WAN)[ref]. Those datasets were selected because of their size (number of samples > 50) and because they were available in not centred format necessary for NMF.

For NMF the procedure was following:

- data was transformed into log2
- zero rows were removed
- the algorithm assesing cophentic index was applied to chose optimal number of components
- datasets were decomposed with matlab NMF implementation from Brunet et al. [**?**] into (i) number of components suggested by cophenetic coefficient (ii) MSTD dimension (iii) 50 components (approaching overdecomposition)

- the obtained metagenes were decorellated from the mean using a linear regression model

For ICA, the procedure was following:

- data were transformed into log2
- transformed data were mean-centered by gene
- our implementation of MSTD (most stable transcriptomic dimension) from [104] was used to evaluate most stable dimension
- datasets were demposed into (i) MSTD dimension and (ii) 50 components (approaching overdecomposition) with matlab implementantion of fastICA with icasso stabilisation

We did not decompose ICA into low number of components as we consider it as strong underdecomposition and we suspect signals would not be the most reproducible. We limited the over decomposition higher than 50 with NMF as for our biggest dataset (METABRIC) NMF decomposition into 50 took 30245 minutes (3 weeks).

Then separately for NMF and ICA, we correlated all obtained metagens with each other and with known Biton et al. metagenes (obtain from previous ICA decompostion applied pan-cancer). We represented the results in a form of a correlation graph where nodes are metagenes from different datasets and decompostion levels and edge width corresponds peasron correaltion coefficients (Fig 5.1).

We hoped to observe a subset of components from different datasets (no matter the decomposition level) correlate with each strongly and much less with other components in order to confirm that the signal is reproducible (can be found in several dataset) and specific. We used the Biton et al. componets here to help with eventual identification of signals (labelling). What we observe from ICA-decomposition that indeed, without applying any threshold some emerging clusters can be remarked and after application of >0.4 threshold on the correlation coeffcient pseudo-cliques emerge. While metagenes from NMF-decpmpostion are more tighlty connected globally and when the threshold is applied, remaing metagenes do not form clear clusters but group by data set. In NMF decomposition if it hard to define different signals as the datasets seem to be all related to each other. We can see from (Fig 5.1D) that the IMMUNE signal is correlated >0.4 with a high number of NMF components that are also linked to some other components. In ICA (Fig 5.1C) components related to the IMMUNE metagens form a pseudo-clique that is related with one link to INTERFERON metagene.

This simple analysis illustrates that NMF applied to cancer transcriptomes decomposes them to metagenes that are not highly reproductible between datasets. In practice, it will not always be possible to work with big cohorts and the same processing methods. Using ICA for decomposition gives mor credit that it will be possible to use the obtained metagenes as reference in which new data of similar type could be projected.

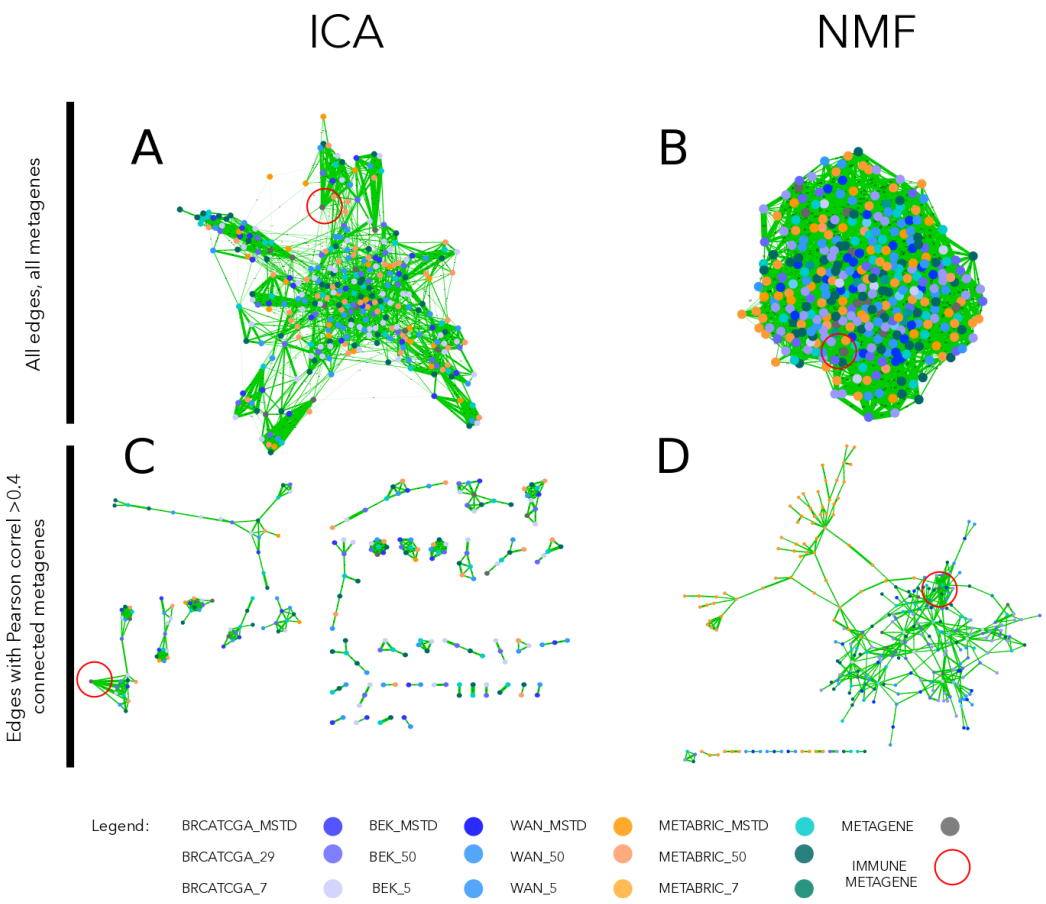


Figure 5.1: Correlation graph of ICA and NMF multiple decompositions. In the upper part of the figure (A,B) we observe the correlation graph of all metagenes (ICA or NMF-based) disposed using edge-weighted bio layout. In the lower part of the figure (C,D) we applied >0.4 thereshold in order to filter the edges. In the case of ICA (C), remaining nodes form pseudo-cliques, immune-related pseudo-clique is highlighted. In the case of NMF (D), components cluster by dataset. Edges' width coressponds to Pearson correlation coefficient. Node colors correspond to dataset from which a metagene was obtained (see legend).

to do:

- *quantify: with clustering coefficient?*
- Explain why ICA is more reproducible

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

Carry on a "sensitivity study":

- remove some % of genes from basis matrix or marker gene list
- evaluate how it changes results

Chapter 6

Deconvolution of transcriptomes and methylomes

We describe our methods in this chapter. The pre-eliminary pipeline and simple results are described in the manuscript submitted to Springer-Verlag's Lecture Notes in Computer Science (LNCS) entitled **Application of Independent Component Analysis to Tumor Transcriptomes Reveals Specific And Reproducible Immune-related Signals** that is placed at the end of this chapter. In the final thesis final pipeline will be split into following structure

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

Few lines describing our idea

Figure?

6.1.1 The ICA-based deconvolution of Transcriptomes

- remind shortly ICA
- describe stabilisation procedure *icasso*
- explain IC-metagene concept

If completed add related section about two other ways of getting metagenes

- attractor metagenes

- k-lines

6.1.2 Interpretation of Independent components

6.1.2.1 Correlation based identification of confounding factors

6.1.2.2 Identification of immune cell types with enrichment test / other

6.1.3 Transforming metagenes into signature matrix

6.1.4 Regression-based estimation of cell-type proportions : solving system of equations

6.2 *DeconICA* R package for ICA-based deconvolution

This part of the chapter will be adapted from package vignettes

It will contain

- technical package description
- user guide
- examples

6.2.1 *Demo*

The package needs to installed and then imported.

```
#import package
library(deconica)
```

Then we can perform our pipeline on sample data available in the package

```
#import sample data
data(BRCA)
#decompose data
fastica.res <- run_fastica (
  BRCA,
  optimal = TRUE,
  row.center = TRUE,
```

```
  with.names = TRUE,
  gene.names = NULL,
  alg.typ = "parallel",
  method = "C",
  n.comp = 100,
  isLog = TRUE,
  R = TRUE
)
#correlate obtained metagenes with Biton et al.
#metagenes (by default)
correlate.res <-
  correlate_metagenes(fastica.res$S, fastica.res$names)
#assign reciprocal components
assign.res <- assign_metagenes(correlate.res$r)
#identify components that are >0.1 correlated with
#immune and are not assigned to any other component
identify.immune <-
  identify_immune_ic(correlate.res$r[, "M8_IMMUNE"], assign.res[, 2])
#test enrichment with fisher test in
#Immgen signatures (by default)
enrichment.res <- gene.enrichment.test(
  fastica.res$S,
  fastica.res$names,
  names(identify.immune),
  gmt = ImmgenHUGO,
  alternative = "greater",
  p.adjust.method = "BH",
  p.value.threshold = 0.05
)
```

The present state of the package is described in Fig 6.1.

Next step will be:

- adding the metagenes selection and transformation into basis matrix for deconvolution
- identifying confounding factors
- estimating purity with an existing tool
- running an equations solver (based on least squares or other type of regression) including basis matrix, confounding factors, purity
- including regularisation factos
- adding graphics

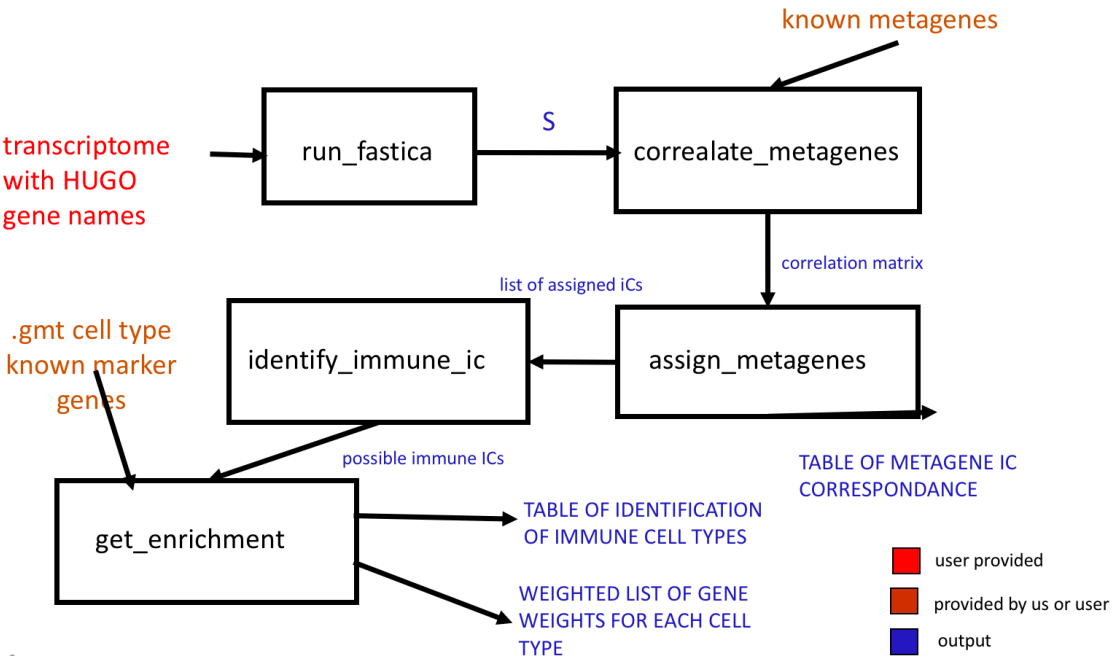


Figure 6.1: State of the deconICA package in January 2018. The flow chart illustrates existing functions in the R package *DeconICA*. Squares represent functions, red are user-provided inputs, brown are inputs we provide but that can be replaced easily by user and in blu we marked outputs.

- adding user interface
- writing a demo (best interactive)

Chapter 7

Comparative analysis of cancer immune infiltration

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

- application to Breast cancer
  - compare metagenes of the same cell type in different datasets
  - compare metagenes of the same cell type in the same dataset (happens sometimes)
  - compare A matrix (sample weights) with clinical metadata
  - compare patients with opposite extreme phenotypes (the gene expression) with DEG ou others
  - run enrichment with more specific list of genes ex. Th1/2/17 cells in T cels etc.

- application pan cancer
  - derivation of meta-metagenes for immune cell types
  - above points are true for pan cancer
- follow up of Biton paper ?
  - *Idea of Vassili from the lab meeting*, personally I am not sure if there is no conflict of interest with other members of the team

Chapter 8

Heterogeneity of immune cell types

We include here an extract of a *ready to submit* article of Kondratova et al. **(co-first authored by Urszula Czerwinska)** - the abstract and figures which are result of work on single cell heterogeneity.

Explication how deconvolution methodology can be used for analysis of heterogeneity of immune cells

- describe the context briefly
- describe more in details my part - data analysis of single cell data

To be defined:

- add CAFS (that will maybe appear in *JBM*)
- add unpublished analysis made for the Nature Immunology *Michea et al.* paper (to be defined)
- The single T-cell study (if done)

Signalling network map of innate immune response in cancer reveals signatures of cell heterogeneity and polarization in tumor microenvironment

SUMMARY (150 words) (now 190)

To describe the contribution of innate immune components to anti- and pro-tumor effect of tumor microenvironment (TME), we collected information on molecular mechanisms governing innate immune response in cancer and represented it in a form of network maps. The signalling maps of macrophages, dendritic cells, myeloid-derived suppressor cells, natural killers were constructed. These cell type-specific maps, integrated together and updated by intra-cellular interactions, gave rise to a seamless comprehensive meta-map of innate immune response in cancer. The meta-map depicts signalling of anti- and pro-tumor activities of innate immunity system as a whole. The cell type-specific maps and the meta-map were used for interpretation of single cell RNA-Seq data from natural killers and macrophages in metastatic melanoma. The analysis demonstrated existence of sub-populations within each cell type that possess different anti- and pro-tumor polarization status. In addition, we used the meta-map for interpretation of pan-cancer patient survival data to retrieve patient survival signature. The cell type-specific signalling maps together with the meta-map of innate immune response in cancer form an open source platform available online that can be applied by wide community for assessment of TME status in cancer and beyond.

Key words

Tumor immunology, tumor microenvironment, innate immunity signalling, cancer systems biology, comprehensive signalling network map, semantic zooming, single cell data analysis, bioinformatics, molecular pathways and networks, intercellular communication, cell reprogramming, polarization, heterogeneity

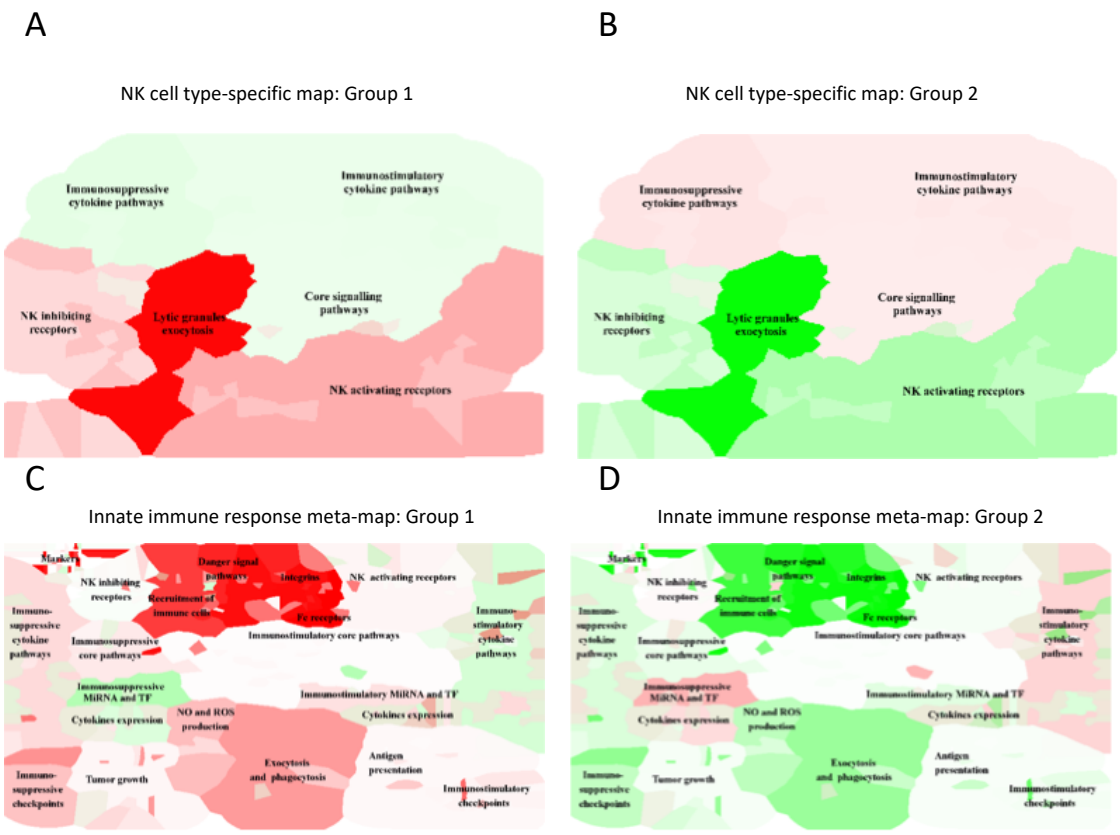


Figure 4. Visualization of modules activity scores using expression data from melanoma natural killers (NK) cells in the context of maps. Staining of the NK cell type-specific map with modules activity scores calculated from single cell RNAseq expression data for (A) NK Groups 1 and (B) NK Groups 2 cells. Staining of the innate immune response meta-map with modules activity scores for (C) NK Groups 1 and (D) NK Groups 2 cells. Red–upregulated, green- downregulated module activity.

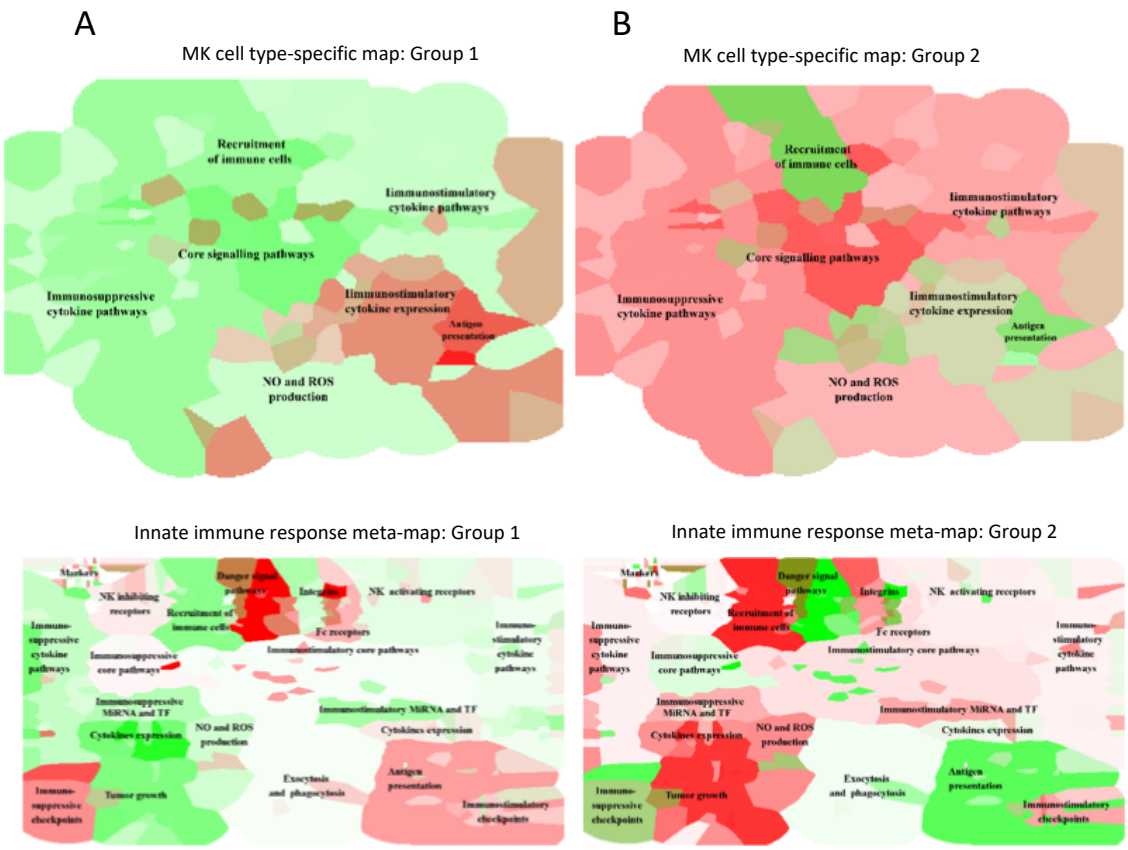
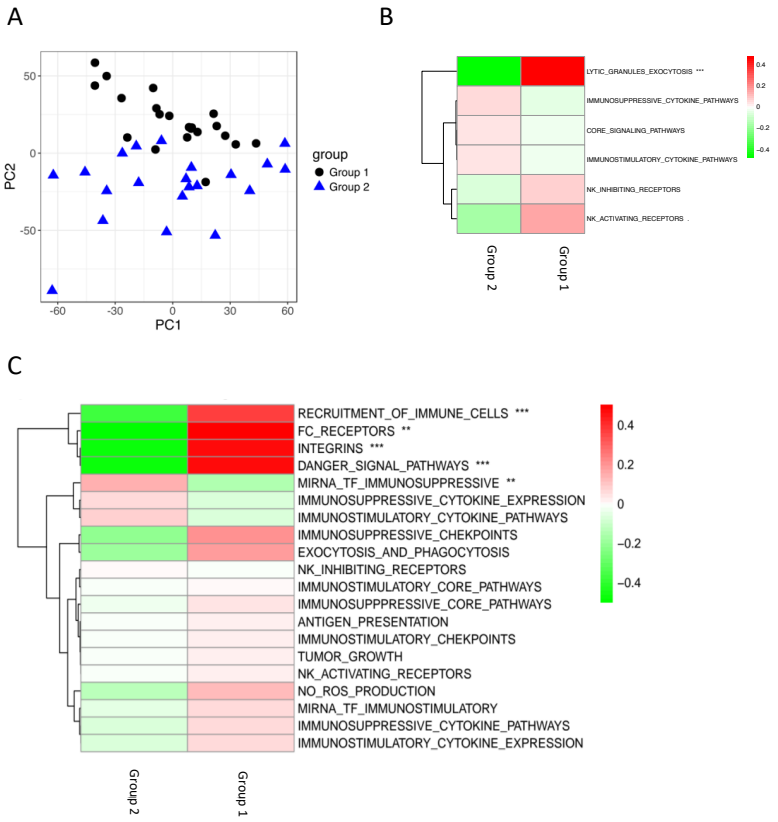
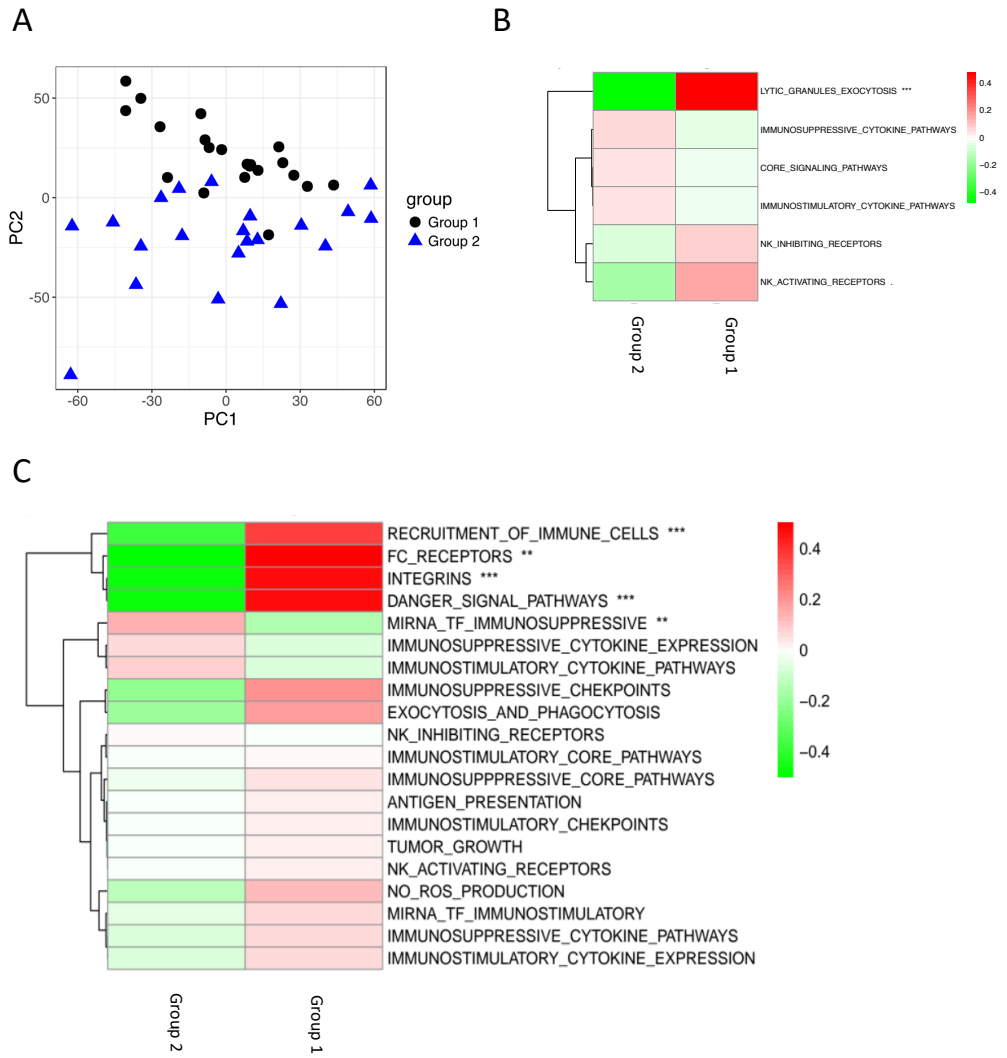


Figure 5. Visualization of modules activity scores using expression data from melanoma macrophages (Mph) cells in the context of maps. Staining of the Mph cell type-specific map with modules activity scores calculated from single cell RNAseq expression data for (A) Mph Groups 1 and (B) Mph Groups 2 cells. Staining of the innate immune response meta-map with modules activity scores for (C) Mph Groups 1 and (D) Mph Groups 2 cells. Red-upregulated, green-downregulated module activity.



Supplemental Figure 6. Sup-populations study and calculation of modules activity scores using expression data from melanoma natural killers (NK) cells. (A) NK single cells in PC1 and PC2 coordinates space. Two groups are colored distinctly in blue and black. Heatmap of mean values of 50% most variant genes divided by group of modules in (B) cell type-specific map and (C) meta-map. The p-value of the t-test between gene expression is reported following the code: *** < 0.001, ** < 0.01 , * < 0.05 , . < 0.1



Supplemental Figure 7. **Sup-populations study and calculation of modules activity scores using expression data from melanoma macrophages (Mph) cells.** (A) Mph single cells in PC1 and PC2 coordinates space. Two groups, the first and the fourth quartile of distribution along the IC1 axis, are colored distinctly in blue and black Heatmap of mean values of 50% most variant genes in groups in modules of (B) cell type-specific map and of(C) meta-map. The p-value of the t-test between gene expression is reported following the code: *** < 0.001, ** < 0.01 , * < 0.05 , . < 0.1

Part III

Discussion

Chapter 9

Discussion

Chapter 10

Conclusions and perspectives

Here we will have some interesting and well-written conclusion that will validate the quality of this thesis.

Problems:

- reproducibility of other tools (code accessing, making work, pre-processing)
- no-spatial dimension
- heterogeneity - analyse sample by samples (need of big dimension)
- time dimention
- validation - no gold statndard
- our solution - more contex specific but less interpretable?

A major part of this thesis has been to reproduce earlier work[7][8], and it has been time consuming to try to reproduce different approaches or scripts. It has been brought up that other scientists have struggled - and many failed - to reproduce another scientists work[48]. The article states that of 1,576 researchers, over 70% have failed to reproduce others work and over 50% have failed to reproduce their own. 52% of the participants in the survey state that is is a “significant crisis”, which indicates that we could call this a “reproducibility crisis”[48]. Such high numbers may suggest in- accurate or poor documentation of the different steps towards achieving the results, or even going as far as suggesting untrustworthy results. The latter is a bold statement, but according to the article, less than 31% believe that struggles to reproduce published results are due to wrong results[48].

Moreover, the immune cells can be situated in different locations, either in the core, in the invasive margin or in the adjacent tertiary lymphoid structures (TLS), ectopic lymphoid

formations found in inflamed, infected, or tumoral tissues exhibiting all the characteristics of structures in the lymph nodes (LN) associated with the generation of an adaptive immune response (Dieu-Nosjean et al. 2014): a correlation has been found between high densities of TLS and prolonged patient's survival in more than 10 different types of cancer (Sautès-Fridman et al. 2016).

For instance, CD8+ T cells can be visible in both the invasive margin and the core of the tumor, while the TLS seem to lack these cells. In addition, the mixture of immune cells can vary differently in relation to tumor types. Some components of the immune contexture, more than others, are helpful in terms of good prognosis: this fact is shared by multiple papers, such as Dave et al. 2004, which paved the way in the early years of the XXI century, while in Parker et al. 2008 and Parker et al. 2009

—-

. In the case of T cells and cancer, although the total frequencies of tumor-specific T cells are difficult to assess (due to uncertainties about the range of targets, see below), reports of blood-derived tumor-specific T cells suggest that these frequencies are low (much less than 1% of CD8þ T cells, which typically make up 10%–20% of peripheral blood mononuclear cells; refs. 50, 51). Therefore, at least in the case of blood, a large part of the signal measured in bulk T-cell profiles should originate from irrelevant cells. In tumor tissues, the problem is certainly less severe, in that tumor-infiltrating T cells are likely enriched for tumor-specific T cells. However, as discussed above, because the immunologic composition of tumor tissues is complex and heterogeneous, similar efforts dedicated to identifying, quantifying, and profiling relevant (tumor-specific) T cells are still needed.

Annexes

Dc subtypes

DreamIdea Challenge

Full list of publications

CV

Post Scriptum: Thesis writing

This Thesis is written in *bookdown*. I have chosen this form as it can easily compile to *LaTeX*, PDF, MS Word, ebook and html. Optimally, the final manuscript will be also published online in a form of an open source gitBook and an ebook including interactive figures and maybe even data demos. Another good reason for using *bookdown* is its simple syntax of markdown and natural integration of code snippets with .Rmd. It reduces formatting time and give multiple outputs.

The template of for this thesis manuscript was adapted from *LaTeX* template provided by University Paris Descartes.

Citations are stocked in Mendeley Desktop and exported to .bib files automatically.

Bibliography

[1] The CRI | Centre for Research and Interdisciplinarity. URL https://cri-paris.org/the-cri/.

[2] *Facilitating Interdisciplinary Research*. National Academies Press, Washington, D.C., 2004. ISBN 978-0-309-09435-1. doi: 10.17226/11153. URL http://www.nap.edu/catalog/11153.

[3] A. R. Abbas et al. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS One*, 4(7):e6098, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006098. URL http://dx.plos.org/10.1371/journal.pone.0006098.

[4] F. Al-Ejeh et al. Meta-analysis of the global gene expression profile of triple-negative breast cancer identifies genes for the prognostication and treatment of aggressive breast cancer. *Oncogenesis*, 3(4):e100–e100, 2014. ISSN 2157-9024. doi: 10.1038/oncsis.2014.14. URL http://www.nature.com/articles/oncsis201414.

[5] N. Andor et al. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60, 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt622. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt622.

[6] M. Angelova et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.*, 16(1):64, dec 2015. ISSN 1465-6906. doi: 10.1186/s13059-015-0620-6. URL http://genomebiology.com/2015/16/1/64.

[7] M. G. Anitei et al. Prognostic and predictive values of the immunoscore in patients with rectal cancer. *Clin Cancer Res*, 20(7):1891–1899, 2014. ISSN 1078-0432. doi: 10.1158/1078-0432.ccr-13-2830. URL http://clincancerres.aacrjournals.org/content/20/7/1891.full.pdf.

[8] D. Aran et al. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, 6:8971, 2015. ISSN 2041-1723. doi: 10.1038/ncomms9971. URL

http://www.ncbi.nlm.nih.gov/pubmed/26634437http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4671203.

[9] D. Aran et al. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, 18(1):220, 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1349-1. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1349-1.

[10] M. Aris et al. Lessons from Cancer Immunoediting in Cutaneous Melanoma. *Clin. Dev. Immunol.*, 2012:1–14, 2012. ISSN 1740-2522. doi: 10.1155/2012/192719. URL http://www.ncbi.nlm.nih.gov/pubmed/22924051http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3424677http://www.hindawi.com/journals/jir/2012/192719/.

[11] S. Arora et al. A practical algorithm for topic modeling with provable guarantees, 2013. URL https://dl.acm.org/citation.cfm?id=3042925.

[12] C.-A. Azencott et al. The inconvenience of data of convenience: computational research beyond post-mortem analyses. *Nat. Methods*, 14(10):937–938, sep 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4457. URL http://www.nature.com/doifinder/10.1038/nmeth.4457.

[13] F. Balkwill, and A. Mantovani. Inflammation and cancer: Back to Virchow?, 2001. ISSN 01406736.

[14] D. A. Barbie et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112, 2009. ISSN 0028-0836. doi: 10.1038/nature08460. URL http://www.nature.com/articles/nature08460.

[15] C. N. Baxevanis et al. Tumor specific cytolysis by tumor infiltrating lymphocytes in breast cancer. *Cancer*, 74(4):1275–82, aug 1994. ISSN 0008-543X. URL http://www.ncbi.nlm.nih.gov/pubmed/7914469.

[16] E. Becht et al. Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clin. Cancer Res.*, 22(16):4057–66, aug 2016. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-15-2879. URL http://www.ncbi.nlm.nih.gov/pubmed/26994146.

[17] E. Becht et al. Evaluation of UMAP as an alternative to t-SNE for single-cell data. *bioRxiv*, page 298430, 2018. doi: 10.1101/298430. URL https://www.biorxiv.org/content/early/2018/04/10/298430.

[18] A. J. Bell, and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. 1995. URL http://www.inf.fu-berlin.de/lehre/WS05/Mustererkennung/infomax/infomax.pdf.

[19] A. Berghoff et al. Genetics and immunology: Tumor-specific genetic alterations as a target for immune modulating therapies. In L. Zitvogel, and G. Kroemer, editors, *Oncoimmunology : a practical guide for cancer immunotherapy*, chapter 13, pages 231–246. 01 2018. ISBN 978-3-319-62430-3.

[20] B. E. Bernstein et al. The mammalian epigenome. *Cell*, 128(4):669–81, feb 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.01.033. URL http://www.ncbi.nlm.nih.gov/pubmed/17320505.

[21] A. Bhatia, and Y. Kumar. Cancer-immune equilibrium: questions unanswered. *Cancer Microenviron.*, 4(2):209–17, aug 2011. ISSN 1875-2284. doi: 10.1007/s12307-011-0065-8. URL http://www.ncbi.nlm.nih.gov/pubmed/21607751http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3170416.

[22] G. Bindea et al. Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity*, 39(4):782–795, oct 2013. ISSN 10747613. doi: 10.1016/j.immuni.2013.10.003. URL http://www.ncbi.nlm.nih.gov/pubmed/24138885http://linkinghub.elsevier.com/retrieve/pii/S1074761313004378.

[23] A. Biton et al. Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes. *Cell Rep.*, 9(4):1235–1245, 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.10.035. URL http://www.ncbi.nlm.nih.gov/pubmed/25456126http://linkinghub.elsevier.com/retrieve/pii/S2211124714009048.

[24] C. U. Blank et al. CANCER IMMUNOLOGY. The "cancer immunogram". *Science*, 352(6286):658–60, may 2016. ISSN 1095-9203. doi: 10.1126/science.aaf2834. URL http://www.ncbi.nlm.nih.gov/pubmed/27151852.

[25] D. H. Boal. *Mechanics of the cell*. Cambridge University Press, 2002. ISBN 9780511810954.

[26] C. E. Breeze et al. eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Rep.*, 17(8):2137–2150, 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.10.059. URL http://www.ncbi.nlm.nih.gov/pubmed/27851974http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5120369http://linkinghub.elsevier.com/retrieve/pii/S2211124716314796.

[27] J.-P. Brunet et al. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, 101(12):4164–9, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308531101. URL http://www.ncbi.nlm.nih.gov/pubmed/15016911http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC384712.

[28] F. M. Burnet. The concept of immunological surveillance. *Prog. Exp. Tumor Res.*, 13: 1–27, 1970. ISSN 0079-6263. URL http://www.ncbi.nlm.nih.gov/pubmed/4921480.

[29] L. Cantini et al. Stabilized Independent Component Analysis outperforms other methods in finding reproducible signals in tumoral transcriptomes. *bioRxiv*, page 318154, 2018. doi: 10.1101/318154. URL https://www.biorxiv.org/content/early/2018/05/09/318154.

[30] J. Cardoso, and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proc. F Radar Signal Process.*, 140(6):362, 1993. ISSN 0956375X. doi: 10.1049/ip-f-2.1993.0054. URL http://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0054.

[31] S. L. Carter et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, 30(5):413–421, 2012. ISSN 1087-0156. doi: 10.1038/nbt.2203. URL http://www.nature.com/articles/nbt.2203.

[32] P. Charoentong et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.*, 18(1):248–262, jan 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.12.019. URL http://www.ncbi.nlm.nih.gov/pubmed/28052254.

[33] D. S. Chen, and I. Mellman. Elements of cancer immunity and the cancer-immune set point, 2017. ISSN 14764687.

[34] P. L. Chen et al. Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov.*, 6(8):827–837, 2016. ISSN 21598290. doi: 10.1158/2159-8290.CD-15-1545.

[35] S.-H. Chen et al. A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. *BMC Bioinformatics*, 19(S4):154, 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2069-6. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2069-6.

[36] W.-Y. Cheng et al. Biomolecular Events in Cancer Revealed by Attractor Metagenes. *PLoS Comput. Biol.*, 9(2):e1002920, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002920. URL http://dx.plos.org/10.1371/journal.pcbi.1002920.

[37] E. C. Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953. ISSN 0001-4966. doi: 10.1121/1.1907229. URL http://asa.scitation.org/doi/10.1121/1.1907229.

[38] J. Chifman et al. Conservation of immune gene signatures in solid tumors and prognostic implications. *BMC Cancer*, 16(1):911, dec 2016. ISSN 1471-2407. doi: 10.1186/s12885-016-2948-z. URL http://bmccancer.biomedcentral.com/articles/10.1186/s12885-016-2948-zhttp://www.ncbi.nlm.nih.gov/pubmed/27871313http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5118876.

[39] W. Chung et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.*, 8:15081, 2017. ISSN 2041-1723. doi: 10.1038/ncomms15081. URL http://www.ncbi.nlm.nih.gov/pubmed/28474673http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5424158.

[40] M. Cieślik, and A. M. Chinnaiyan. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.*, 19(2):93–109, dec 2017. ISSN 1471-0056. doi: 10.1038/nrg.2017.96. URL http://www.nature.com/doifinder/10.1038/nrg.2017.96.

[41] J. Clarke et al. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, 26(8):1043–1049, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq097. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq097.

[42] C.-S. G. Core. Single cell genomics, 2016. URL https://www.cedars-sinai.edu/Research/Research-Cores/Genomics-Core/Documents/website_pricing_jan_2018_final.pdf.

[43] A. D. Corlan. Medline trend: automated yearly statistics of pubmed results for any query, 2004. URL http://dan.corlan.net/medline-trend.html.

[44] A. Costa et al. Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer. *Cancer Cell*, 33(3):463–479.e10, mar 2018. ISSN 1878-3686. doi: 10.1016/j.ccell.2018.01.011. URL http://www.ncbi.nlm.nih.gov/pubmed/29455927.

[45] P. Csermely et al. *Science education : models and networking of student research training under 21*. IOS Press, 2007. ISBN 9781586037215.

[46] U. Czerwinska et al. Application of Independent Component Analysis to Tumor Transcriptomes Reveals Specific and Reproducible Immune-Related Signals. pages 501–513. Springer, Cham, 2018. doi: 10.1007/978-3-319-93764-9_46. URL http://link.springer.com/10.1007/978-3-319-93764-9{_}46.

[47] S. Davis. Github: awesome-single-cell, 2016. URL https://github.com/seandavi/awesome-single-cell.

[48] C. Ding et al. Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):45–55, 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.277. URL http://ieeexplore.ieee.org/document/4685898/.

[49] H. Drucker et al. Support vector regression machines. In *Adv. Neural Inf. Process. Syst.*, pages 155–161, 1997.

[50] N. Dumont et al. Breast Fibroblasts Modulate Early Dissemination, Tumorigenesis, and Metastasis through Alteration of Extracellular Matrix Characteristics. *Neoplasia*, 15(3):249–IN7, 2013. ISSN 14765586. doi: 10.1593/neo.121950. URL http://linkinghub.elsevier.com/retrieve/pii/S1476558613800553.

[51] G. P. Dunn et al. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.*, 3(11):991–998, nov 2002. ISSN 1529-2908. doi: 10.1038/ni1102-991. URL http://www.nature.com/articles/ni1102-991.

[52] Ediorial Cell Systems. What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Syst.*, 4(3):255–259, mar 2017. doi: 10.1016/j.cels.2017.03.006. URL http://www.ncbi.nlm.nih.gov/pubmed/28334573.

[53] P. Ehrlich. Über den jetzigen Stand der Chemotherapie. *Berichte der Dtsch. Chem. Gesellschaft*, 42(1):17–47, jan 1909. ISSN 03659496. doi: 10.1002/cber.19090420105. URL http://doi.wiley.com/10.1002/cber.19090420105.

[54] A. Elmas et al. Discovering Genome-Wide Tag SNPs Based on the Mutual Information of the Variants. *PLoS One*, 11(12):e0167994, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0167994. URL http://dx.plos.org/10.1371/journal.pone.0167994.

[55] J. M. Engreitz et al. Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.*, 43(6):932–44, 2010. ISSN 1532-0480. doi: 10.1016/j.jbi.2010.07.001. URL http://www.ncbi.nlm.nih.gov/pubmed/20619355http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2991480.

[56] Erickson, Jeff. Jeff Erickson's Algorithms, Etc. URL http://jeffe.cs.illinois.edu/teaching/algorithms/.

[57] T. Erkkilä et al. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577, 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq406. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq406.

[58] Fa-Yu Wang et al. Nonnegative Least-Correlated Component Analysis for Separation of Dependent Sources by Volume Maximization. *IEEE Trans. Pattern Anal.*

Mach. Intell., 32(5):875–888, 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.72. URL http://ieeexplore.ieee.org/document/4815260/.

[59] A. P. Feinberg, and B. Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92, jan 1983. ISSN 0028-0836. doi: 10.1038/301089a0. URL http://www.nature.com/doifinder/10.1038/301089a0.

[60] F. Finotello et al. quanTIseq: quantifying immune contexture of human tumors. *bioRxiv*, page 223180, 2017. doi: 10.1101/223180. URL https://www.biorxiv.org/content/early/2017/11/22/223180.

[61] J. Folkman et al. Isolation of a tumor factor responsible for angiogenesis. *J. Exp. Med.*, 133(2):275–88, feb 1971. ISSN 0022-1007. doi: 10.1084/JEM.133.2.275. URL http://www.ncbi.nlm.nih.gov/pubmed/4332371http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2138906.

[62] I. A. for Research in Cancer. Globocan - facts sheets by cancer, 2018. URL http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

[63] D. M. Frey et al. High frequency of tumor-infiltrating FOXP3+ regulatory T cells predicts improved survival in mismatch repair-proficient colorectal cancer patients. *Int. J. Cancer*, 126(11):2635–2643, 2010. ISSN 00207136. doi: 10.1002/ijc.24989.

[64] Y. Fu et al. BACOM2.0 facilitates absolute normalization and quantification of somatic copy number alterations in heterogeneous tumor. *Sci. Rep.*, 5(1):13955, 2015. ISSN 2045-2322. doi: 10.1038/srep13955. URL http://www.nature.com/articles/srep13955.

[65] D. I. Gabrilovich et al. Coordinated regulation of myeloid cells by tumours, 2012. ISSN 14741733.

[66] T. F. Gajewski et al. Immune resistance orchestrated by the tumor microenvironment. *Immunol. Rev.*, 213(1):131–145, oct 2006. ISSN 1600-065X. doi: 10.1111/J.1600-065X.2006.00442.X. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1600-065X.2006.00442.x/abstract.

[67] J. Galon et al. The immune score as a new possible approach for the classification of cancer. *J. Transl. Med.*, 10(1):1, jan 2012. ISSN 1479-5876. doi: 10.1186/1479-5876-10-1. URL http://translational-medicine.biomedcentral.com/articles/10.1186/1479-5876-10-1.

[68] J. Galon et al. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J. Pathol.*, 232(2):199–209, jan 2014. ISSN 00223417. doi:

10.1002/path.4287. URL http://www.ncbi.nlm.nih.gov/pubmed/24122236http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4255306http://doi.wiley.com/10.1002/path.4287.

[69] R. Gaujoux, and C. Seoighe. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29(17):2211–2212, 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt351. URL http://www.ncbi.nlm.nih.gov/pubmed/23825367https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt351.

[70] R. Gaujoux, and C. Seoighe. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, 11(1):367, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-367. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-367.

[71] R. Gaujoux, and C. Seoighe. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.*, 12(5):913–921, 2012. ISSN 1567-1348. doi: 10.1016/J.MEEGID.2011.08.014. URL https://www.sciencedirect.com/science/article/pii/S1567134811002930?via{%}3Dihub.

[72] D. Ghosh. Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*, 20(11):1663–1669, 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth139. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth139.

[73] C. Giesen et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11(4):417–422, apr 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2869. URL http://www.nature.com/articles/nmeth.2869.

[74] S. Gnjatic et al. Identifying baseline immune-related biomarkers to predict clinical outcome of immunotherapy. *J. Immunother. cancer*, 5:44, 2017. ISSN 2051-1426. doi: 10.1186/s40425-017-0243-4. URL http://www.ncbi.nlm.nih.gov/pubmed/28515944http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5432988.

[75] J. Godec et al. Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity*, 44(1):194–206, jan 2016. ISSN 10747613. doi: 10.1016/j.immuni.2015.12.006. URL http://www.ncbi.nlm.nih.gov/pubmed/26795250http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5330663http://linkinghub.elsevier.com/retrieve/pii/S1074761315005324.

[76] T. R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, oct 1999. ISSN

0036-8075. doi: 10.1126/SCIENCE.286.5439.531. URL http://www.ncbi.nlm.nih.gov/pubmed/10521349.

[77] T. Gong et al. Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLoS One*, 6(11):e27156, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0027156. URL http://dx.plos.org/10.1371/journal.pone.0027156.

[78] A. N. A. N. Gorban. *Principal manifolds for data visualization and dimension reduction*. Springer, 2007. ISBN 9783540737506.

[79] F. Görtler et al. Loss-function learning for digital tissue deconvolution. 2018. URL http://arxiv.org/abs/1801.08447.

[80] M. M. Gosink et al. Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics*, 23(24):3328–3334, 2007. ISSN 1460-2059. doi: 10.1093/bioinformatics/btm508. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm508.

[81] M. Greenblatt, and P. Shubi. Tumor angiogenesis: transfilter diffusion studies in the hamster by the transparent chamber technique. *J. Natl. Cancer Inst.*, 41(1):111–24, jul 1968. ISSN 0027-8874. URL http://www.ncbi.nlm.nih.gov/pubmed/5662020.

[82] V. Greger et al. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum. Genet.*, 83(2):155–158, sep 1989. ISSN 0340-6717. doi: 10.1007/BF00286709. URL http://link.springer.com/10.1007/BF00286709.

[83] B. Győrffy et al. Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res.*, 17(1):11, jan 2015. ISSN 1465-542X. doi: 10.1186/s13058-015-0514-2. URL http://www.ncbi.nlm.nih.gov/pubmed/25848861http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4307898.

[84] D. Hanahan, and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, jan 2000. ISSN 0092-8674. doi: 10.1016/S0092-8674(00)81683-9. URL http://www.ncbi.nlm.nih.gov/pubmed/10647931.

[85] D. Hanahan, and L. Coussens. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell*, 21(3):309–322, mar 2012. ISSN 15356108. doi: 10.1016/j.ccr.2012.02.022. URL http://www.ncbi.nlm.nih.gov/pubmed/22439926http://linkinghub.elsevier.com/retrieve/pii/S1535610812000827.

[86] S. Hänzelmann et al. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1):7, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-7. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-7.

[87] L. N. Harris et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. *J. Clin. Oncol.*, 34(10):1134–1150, apr 2016. ISSN 0732-183X. doi: 10.1200/JCO.2015.65.2289. URL http://www.ncbi.nlm.nih.gov/pubmed/26858339http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4933134http://ascopubs.org/doi/10.1200/JCO.2015.65.2289.

[88] T. Hastie et al. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

[89] T. S. P. Heng et al. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.*, 9(10):1091–1094, oct 2008. ISSN 1529-2908. doi: 10.1038/ni1008-1091. URL http://www.ncbi.nlm.nih.gov/pubmed/18800157http://www.nature.com/doifinder/10.1038/ni1008-1091.

[90] J. Herault, and C. Jutten. Space or time adaptive signal processing by neural network models. In *AIP Conf. Proc.*, volume 151, pages 206–211. AIP, 1986. doi: 10.1063/1.36258. URL http://aip.scitation.org/doi/abs/10.1063/1.36258.

[91] R. S. Herbst et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*, 515(7528):563–567, nov 2014. ISSN 0028-0836. doi: 10.1038/nature14011. URL http://www.ncbi.nlm.nih.gov/pubmed/25428504http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4836193http://www.nature.com/articles/nature14011.

[92] J. Himberg, and A. Hyvarinen. Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. In *2003 IEEE XIII Work. Neural Networks Signal Process. (IEEE Cat. No.03TH8718)*, pages 259–268. IEEE. ISBN 0-7803-8177-7. doi: 10.1109/NNSP.2003.1318025. URL http://ieeexplore.ieee.org/document/1318025/.

[93] M. Hoffmann et al. Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics*, 7(1):369, 2006. ISSN 14712105. doi: 10.1186/1471-2105-7-369. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-369.

[94] E. A. Houseman et al. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 17(1):259, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1140-4. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1140-4.

[95] E. Houseman et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012. ISSN 1471-2105. doi: 10.

1186/1471-2105-13-86. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-86.

[96] E. A. Houseman et al. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu029. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu029.

[97] A. Hyvärinen, and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(45):411–430, 2000. URL https://www.cs.helsinki.fi/u/ahyvarin/papers/NN00new.pdf.

[98] N. C. Institute. Types of cancer treatment, 2017. URL https://www.cancer.gov/about-cancer/treatment/types.

[99] H. Itadani et al. Can Systems Biology Understand Pathway Activation? Gene Expression Signatures as Surrogate Markers for Understanding the Complexity of Pathway Activation. *Curr. Genomics*, 9(5):349–360, aug 2008. ISSN 13892029. doi: 10.2174/138920208785133235. URL http://www.ncbi.nlm.nih.gov/pubmed/19517027http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2694555http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=1389-2029{&}volume=9{&}issue=5{&}spage=349.

[100] H. J. Jackson et al. Driving CAR T-cells forward, 2016. ISSN 17594782.

[101] J. Jeschke et al. DNA methylation–based immune response signature improves patient diagnosis in multiple cancers. *J. Clin. Invest.*, 127(8):3090–3102, jul 2017. ISSN 0021-9738. doi: 10.1172/JCI91095. URL http://www.ncbi.nlm.nih.gov/pubmed/28714863http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5531413https://www.jci.org/articles/view/91095.

[102] Y. Jiang et al. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):E5528–37, 2016. ISSN 1091-6490. doi: 10.1073/pnas.1522203113. URL http://www.ncbi.nlm.nih.gov/pubmed/27573852http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5027458.

[103] W. Ju et al. Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.*, 23(11):1862–73, 2013. ISSN 1549-5469. doi: 10.1101/gr.155697.113. URL http://www.ncbi.nlm.nih.gov/pubmed/23950145http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3814886.

[104] U. Kairov et al. Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics*, 18(1):712, 2017. ISSN

1471-2164. doi: 10.1186/s12864-017-4112-9. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4112-9.

[105] M. J. Kornstein et al. Immunoperoxidase localization of lymphocyte subsets in the host response to melanoma and nevi. *Cancer Res.*, 43(6):2749–53, jun 1983. ISSN 0008-5472. URL http://www.ncbi.nlm.nih.gov/pubmed/6342758.

[106] A. Kuhn et al. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, 8(11):945–947, 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1710. URL http://www.nature.com/articles/nmeth.1710.

[107] I. Kuperstein et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4(7):e160–e160, jul 2015. ISSN 2157-9024. doi: 10.1038/oncsis.2015.19. URL http://www.ncbi.nlm.nih.gov/pubmed/26192618http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4521180http://www.nature.com/articles/oncsis201519.

[108] I. Kuperstein et al. NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Syst. Biol.*, 7(1):100, oct 2013. ISSN 1752-0509. doi: 10.1186/1752-0509-7-100. URL http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-100.

[109] H. Lähdesmäki et al. In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, 6(1):54, 2005. ISSN 14712105. doi: 10.1186/1471-2105-6-54. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-54.

[110] N. B. Larson, and B. L. Fridley. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics*, 29 (15):1888–9, 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt293. URL http://www.ncbi.nlm.nih.gov/pubmed/23749958http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3712213.

[111] Y. Lavin et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell*, 169(4):750–765.e17, 2017. ISSN 0092-8674. doi: 10.1016/J.CELL.2017.04.014. URL https://www.sciencedirect.com/science/article/pii/S0092867417304270.

[112] H. Ledford. How to solve the world's biggest problems. *Nature*, 525(7569):308–311, sep 2015. ISSN 0028-0836. doi: 10.1038/525308a. URL http://www.nature.com/doifinder/10.1038/525308a.

[113] D. D. Lee, and H. S. Seung. Algorithms for non-negative matrix factorization, 2000. URL https://dl.acm.org/citation.cfm?id=3008829.

[114] J. T. Leek, and J. D. Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.*, 3(9):e161, 2007. ISSN 1553-7390. doi: 10.1371/journal.pgen.0030161. URL http://www.ncbi.nlm.nih.gov/pubmed/17907809http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1994707http://dx.plos.org/10.1371/journal.pgen.0030161.

[115] D. Leung et al. Vascular endothelial growth factor is a secreted angiogenic mitogen. *Science (80-.).*, 246(4935):1306–1309, 1989. ISSN 0036-8075. doi: 10.1126/science.2479986. URL http://www.sciencemag.org/cgi/doi/10.1126/science.2479986.

[116] B. Li, and J. Z. Li. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.*, 15(9):473, 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0473-4. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0473-4.

[117] B. Li et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.*, 17(1):174, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1028-7. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1028-7.

[118] B. Li et al. Revisit linear regression-based deconvolution methods for tumor gene expression data. *Genome Biol.*, 18(1):127, 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1256-5. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1256-5.

[119] Y. Li, and X. Xie. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinforma. 2013 145*, 14(5):S11, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-s5-s11. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S5-S11.

[120] D. A. Liebner et al. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, 30(5):682–689, 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt566. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt566.

[121] C.-J. Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Comput.*, 19(10):2756–2779, 2007. ISSN 0899-7667. doi: 10.1162/neco.2007.19.10.2756. URL http://www.mitpressjournals.org/doi/10.1162/neco.2007.19.10.2756.

[122] J. Liu et al. Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. *BMC Med. Genomics*, 1(1):39, dec 2008. ISSN 1755-8794. doi: 10.1186/1755-8794-1-39. URL http://www.ncbi.nlm.nih.gov/pubmed/18786252http://www.pubmedcentral.nih.gov/

articlerender.fcgi?artid=PMC2551605http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-1-39.

[123] Y. Liu et al. Post-modified non-negative matrix factorization for deconvoluting the gene expression profiles of specific cell types from heterogeneous clinical samples based on RNA-sequencing data. *J. Chemom.*, page e2929, 2017. ISSN 08869383. doi: 10.1002/cem.2929. URL http://doi.wiley.com/10.1002/cem.2929.

[124] P. Lu et al. Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. U. S. A.*, 100(18):10370–5, 2003. ISSN 0027-8424. doi: 10.1073/pnas.1832361100. URL http://www.ncbi.nlm.nih.gov/pubmed/12934019http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC193568.

[125] P. Lutsik et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.*, 18(1):55, 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1182-6. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1182-6.

[126] Y. A. Lyons et al. Immune cell profiling in cancer: molecular approaches to cell-specific identification. *npj Precis. Oncol.*, 1, 2017. doi: 10.1038/s41698-017-0031-0. URL https://www.nature.com/articles/s41698-017-0031-0.pdf.

[127] J. Maksimovic et al. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Res.*, 43(16): e106–e106, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv526. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv526.

[128] D. Marquez-Medina et al. Role of gamma-delta T-cells in cancer. Another opening door to immunotherapy. *Clin. Transl. Oncol.*, 14(12):891–895, dec 2012. ISSN 1699-048X. doi: 10.1007/s12094-012-0935-7. URL http://www.ncbi.nlm.nih.gov/pubmed/23054752http://link.springer.com/10.1007/s12094-012-0935-7.

[129] S. A. McCarroll, and D. M. Altshuler. Copy-number variation and association studies of human disease. *Nat. Genet.*, 39(7s):S37–S42, jul 2007. ISSN 1061-4036. doi: 10.1038/ng2080. URL http://www.ncbi.nlm.nih.gov/pubmed/17597780http://www.nature.com/doifinder/10.1038/ng2080.

[130] L. Mcinnes, and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. URL https://arxiv.org/pdf/1802.03426.pdf.

[131] S. Michiels et al. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice. *Ann. Oncol.*, 27(12): 2160–2167, dec 2016. ISSN 0923-7534. doi: 10.1093/annonc/mdw307. URL

http://www.ncbi.nlm.nih.gov/pubmed/27634691http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5178139https://academic.oup.com/annonc/article-lookup/doi/10.1093/annonc/mdw307.

[132] C. A. Miller et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput. Biol.*, 10(8):e1003665, 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003665. URL http://dx.plos.org/10.1371/journal.pcbi.1003665.

[133] T. M. T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. ISBN 0070428077.

[134] R. J. Mody et al. Integrative Clinical Sequencing in the Management of Refractory or Relapsed Cancer in Youth. *JAMA*, 314(9):913, sep 2015. ISSN 0098-7484. doi: 10.1001/jama.2015.10080. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.10080.

[135] R. A. Moffitt et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.*, 47(10):1168–1178, 2015. ISSN 1061-4036. doi: 10.1038/ng.3398. URL http://www.nature.com/articles/ng.3398.

[136] S. Mohammadi et al. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. *Proc. IEEE*, 105(2):340–366, 2017. ISSN 0018-9219. doi: 10.1109/JPROC.2016.2607121. URL http://ieeexplore.ieee.org/document/7676285/.

[137] R. Moncada et al. Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. *bioRxiv*, page 254375, mar 2018. doi: 10.1101/254375. URL https://www.biorxiv.org/content/early/2018/03/05/254375.

[138] Y. Naito et al. CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. *Cancer Res.*, 58(16):3491–4, aug 1998. ISSN 0008-5472. URL http://www.ncbi.nlm.nih.gov/pubmed/9721846.

[139] B. D. Nelms et al. CellMapper: rapid and accurate inference of gene expression in difficult-to-isolate cell types. *Genome Biol.*, 17(1):201, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1062-5. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1062-5.

[140] L. A. Newberg et al. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PLoS One*, 13(3):e0193067, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0193067. URL http://dx.plos.org/10.1371/journal.pone.0193067.

[141] A. M. Newman et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3337. URL http://www.nature.com/articles/nmeth.3337.

[142] A. Newman et al. CIBERSORT- Absolute mode. URL https://cibersort.stanford.edu/manual.php.

[143] NPR. Science Diction: The Origin Of The Word 'Cancer' : NPR, 2010. URL https://www.npr.org/templates/story/story.php?storyId=130754101.

[144] NPR. Science Diction: The Origin Of The Word 'Cancer', 2010. URL https://www.npr.org/templates/story/story.php?storyId=130754101.

[145] J. A. Oberg et al. Implementation of next generation sequencing into pediatric hematology-oncology practice: moving beyond actionable alterations. *Genome Med.*, 8(1):133, dec 2016. ISSN 1756-994X. doi: 10.1186/s13073-016-0389-6. URL http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0389-6.

[146] L. Oesper et al. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.*, 14(7):R80, 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-7-r80. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-7-r80.

[147] U. D. of Health, and H. Services. Fda approves yervoy to reduce the risk of melanoma returning after surgery, 2015. URL https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm469944.htm.

[148] U. D. of Health, and H. Services. Fda approves new, targeted treatment for bladder cancer, 2016. URL https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm501762.htm.

[149] U. D. of Health, and H. Services. Pembrolizumab (keytruda) checkpoint inhibitor, 2016. URL https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm526430.htm.

[150] U. D. of Health, and H. Services. Fda approval brings first gene therapy to the united states, 2017. URL https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm574058.htm.

[151] U. D. of Health, and H. Services. Fda approves car-t cell therapy to treat adults with certain types of large b-cell lymphoma, 2017. URL https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm581216.htm.

[152] V. Onuchic et al. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Rep.*, 17(8):2075–2086, 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.10.057. URL http://www.ncbi.nlm.nih.gov/pubmed/27851969http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5115176http://linkinghub.elsevier.com/retrieve/pii/S2211124716314772.

[153] S. Ortega-Martorell et al. Non-negative Matrix Factorisation methods for the spectral decomposition of MRS data from human brain tumours. *BMC Bioinformatics*, 13(1):38, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-38. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-38.

[154] P. Paatero, and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994. ISSN 11804009. doi: 10.1002/env.3170050203. URL http://doi.wiley.com/10.1002/env.3170050203.

[155] F. Pagès et al. Effector Memory T Cells, Early Metastasis, and Survival in Colorectal Cancer. *N. Engl. J. Med.*, 353(25):2654–2666, dec 2005. ISSN 0028-4793. doi: 10.1056/NEJMoa051424. URL http://www.ncbi.nlm.nih.gov/pubmed/16371631http://www.nejm.org/doi/abs/10.1056/NEJMoa051424.

[156] F. Pagès et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet (London, England)*, 391(10135):2128–2139, may 2018. ISSN 1474-547X. doi: 10.1016/S0140-6736(18)30789-X. URL http://www.ncbi.nlm.nih.gov/pubmed/29754777.

[157] S. Paget. THE DISTRIBUTION OF SECONDARY GROWTHS IN CANCER OF THE BREAST. *Lancet*, 133(3421):571–573, mar 1889. ISSN 01406736. doi: 10.1016/S0140-6736(00)49915-0. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673600499150.

[158] K. Palucka, and J. Banchereau. Dendritic-Cell-Based Therapeutic Cancer Vaccines, 2013. ISSN 10747613.

[159] E. Papalexi, and R. Satija. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.*, 18(1):35–45, 2017. ISSN 1474-1733. doi: 10.1038/nri.2017.76. URL http://www.nature.com/doifinder/10.1038/nri.2017.76.

[160] J. S. Parker et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27(8):1160–7, mar 2009. ISSN 1527-7755. doi: 10.1200/JCO.2008.18.1370. URL http://www.ncbi.nlm.nih.gov/pubmed/19204204http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2667820.

[161] A. Pascual-Montano et al. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):403–415, 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.60. URL http://ieeexplore.ieee.org/document/1580485/.

[162] J. M. Perkel. Single-cell sequencing made simple. *Nature*, 547(7661):125–126, jul 2017. ISSN 0028-0836. doi: 10.1038/547125a. URL http://www.nature.com/doifinder/10.1038/547125a.

[163] W. W. Piegorsch. *Statistical data analytics : foundations for data mining, informatics, and knowledge discovery*. ISBN 9781118619650. URL https://www.wiley.com/en-fr/Statistical+Data+Analytics:+Foundations+for+Data+Mining,+Informatics,+and+Knowledge+Discovery-p-9781118619650.

[164] J. M. Pitt et al. Resistance Mechanisms to Immune-Checkpoint Blockade in Cancer: Tumor-Intrinsic and -Extrinsic Factors. *Immunity*, 44(6):1255–69, jun 2016. ISSN 1097-4180. doi: 10.1016/j.immuni.2016.06.001. URL http://www.ncbi.nlm.nih.gov/pubmed/27332730.

[165] C. P. Ponting. Big knowledge from big data in functional genomics. *Emerg. Top. Life Sci.*, 1(3):245–248, nov 2017. ISSN 2397-8554. doi: 10.1042/ETLS20170129. URL http://www.emergtoplifesci.org/lookup/doi/10.1042/ETLS20170129.

[166] J. Predina et al. Changes in the local tumor microenvironment in recurrent cancers may explain the failure of vaccines after surgery. *Proc. Natl. Acad. Sci.*, 110(5):E415–E424, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1211850110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1211850110.

[167] F. P. Preparata, and M. I. Shamos. Convex Hulls: Basic Algorithms. In *Comput. Geom.*, pages 95–149. Springer New York, New York, NY, 1985. doi: 10.1007/978-1-4612-1098-6_3. URL http://link.springer.com/10.1007/978-1-4612-1098-6{_}3.

[168] S. V. Puram et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*, 171(7):1611–1624.e24, 2017. ISSN 0092-8674. doi: 10.1016/J.CELL.2017.10.044. URL https://www.sciencedirect.com/science/article/pii/S0092867417312709.

[169] B. Z. Qian, and J. W. Pollard. Macrophage Diversity Enhances Tumor Progression and Metastasis, 2010. ISSN 00928674.

[170] W. Qiao et al. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Comput. Biol.*, 8(12):e1002838, 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002838. URL http://dx.plos.org/10.1371/journal.pcbi.1002838.

[171] D. F. Quail, and J. A. Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nat Med*, 19(11):1423–1437, 2013. ISSN 1546-170X. doi: 10.1038/nm.3394. URL http://www.ncbi.nlm.nih.gov/pubmed/24202395.

[172] J. Racle et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife*, 6:e26476, 2017. ISSN 2050-084X. doi: 10.7554/eLife.26476. URL https://elifesciences.org/articles/26476.

[173] E. Rahmani et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods*, 13(5):443–445, 2016. ISSN 1548-7091. doi: 10.1038/nmeth.3809. URL http://www.nature.com/articles/nmeth.3809.

[174] A. Regev et al. The Human Cell Atlas. *Elife*, 6:e27041, 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL https://elifesciences.org/articles/27041.

[175] D. Repsilber et al. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, 11(1):27, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-27. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-27.

[176] N. A. Rizvi et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (80-.).*, 348(6230):124–128, 2015. ISSN 10959203. doi: 10.1126/science.aaa1348.

[177] D. R. Robinson et al. Integrative clinical genomics of metastatic cancer. *Nature*, 548(7667):297–303, aug 2017. ISSN 0028-0836. doi: 10.1038/nature23306. URL http://www.nature.com/doifinder/10.1038/nature23306.

[178] J. S. Ross et al. Commercialized Multigene Predictors of Clinical Outcome for Breast Cancer. *Oncologist*, 13(5):477–493, may 2008. ISSN 1083-7159. doi: 10.1634/theoncologist.2007-0248. URL http://www.ncbi.nlm.nih.gov/pubmed/18515733http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.2007-0248.

[179] A. Roth et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, 11(4):396–398, 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2883. URL http://www.nature.com/articles/nmeth.2883.

[180] S. Roy et al. A Hidden-State Markov Model for Cell Population Deconvolution. *J. Comput. Biol.*, 13(10):1749–1774, 2006. ISSN 1066-5277. doi: 10.1089/cmb.2006.13.1749. URL http://www.liebertonline.com/doi/abs/10.1089/cmb.2006.13.1749.

[181] D. Rutledge, and D. Jouan-Rimbaud Bouveresse. Independent Components Analysis with the JADE algorithm. *TrAC Trends Anal. Chem.*, 50:22–32, 2013. ISSN 0165-9936. doi: 10.1016/J.TRAC.2013.03.013. URL https://www-sciencedirect-com.gate2.inist.fr/science/article/pii/S0165993613001222.

[182] I. Sagiv-Barfi et al. Eradication of spontaneous malignancy by local immunotherapy. *Sci. Transl. Med.*, 10(426):eaan4488, jan 2018. ISSN 1946-6234. doi: 10.1126/SCITRANSLMED.AAN4488. URL http://stm.sciencemag.org/content/10/426/eaan4488.

[183] R. Satija, and A. K. Shalek. Heterogeneity in immune responses: from populations to single cells. *Trends Immunol.*, 35(5):219–229, may 2014. ISSN 14714906. doi: 10.1016/j.it.2014.03.004. URL http://www.ncbi.nlm.nih.gov/pubmed/24746883http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4035247http://linkinghub.elsevier.com/retrieve/pii/S1471490614000520.

[184] M. Schelker et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.*, 8(1):2032, dec 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-02289-3. URL http://www.nature.com/articles/s41467-017-02289-3http://dx.doi.org/10.1038/s41467-017-02289-3.

[185] B. Schölkopf et al. New Support Vector Algorithms. *Neural Comput.*, 12(5):1207–1245, 2000. ISSN 0899-7667. doi: 10.1162/089976600300015565. URL http://www.mitpressjournals.org/doi/10.1162/089976600300015565.

[186] R. Schwartz, and S. E. Shackney. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, 11(1):42, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-42. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-42.

[187] T. A. Sebeok. DISCUSSION. *Ann. N. Y. Acad. Sci.*, 280(1 Origins and E):481, oct 1976. ISSN 0077-8923. doi: 10.1111/j.1749-6632.1976.tb25511.x. URL http://doi.wiley.com/10.1111/j.1749-6632.1976.tb25511.x.

[188] H. S. Seung, and D. D. Lee. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. ISSN 00280836. doi: 10.1038/44565. URL http://www.nature.com/doifinder/10.1038/44565.

[189] C. P. Shannon et al. Enumerateblood – an R package to estimate the cellular composition of whole blood from Affymetrix Gene ST gene expression profiles. *BMC Genomics*, 18(1):43, 2017. ISSN 1471-2164. doi: 10.1186/s12864-016-3460-1. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3460-1.

[190] S. S. Shen-Orr, and R. Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.*, 25(5):571–578, 2013. ISSN 09527915. doi: 10.1016/j.coi.2013.09.015. URL http://www.ncbi.nlm.nih.gov/pubmed/24148234http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3874291http://linkinghub.elsevier.com/retrieve/pii/S0952791513001507.

[191] S. S. Shen-Orr et al. Cell type–specific gene expression differences in complex tissues. *Nat. Methods*, 7(4):287–289, 2010. ISSN 1548-7091. doi: 10.1038/nmeth.1439. URL http://www.nature.com/articles/nmeth.1439.

[192] O. Shoval et al. Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. *Science (80-.).*, 336(6085):1157–1160, 2012. ISSN 0036-8075. doi: 10.1126/science.1217405. URL http://www.ncbi.nlm.nih.gov/pubmed/22539553http://www.sciencemag.org/cgi/doi/10.1126/science.1217405.

[193] M. H. Slater. Cell Types as Natural Kinds. *Biol. Theory*, 7(2):170–179, feb 2013. ISSN 1555-5542. doi: 10.1007/s13752-012-0084-9. URL http://link.springer.com/10.1007/s13752-012-0084-9.

[194] G. Slavicek. Interdisciplinary -A Historical Reflection. *Int. J. Humanit. Soc. Sci.*, 2(20), 2012. URL http://www.ijhssnet.com/journals/Vol{_}2{_}No{_}20{_}Special{_}Issue{_}October{_}2012/10.pdf.

[195] P. L. Ståhl et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, jul 2016. ISSN 1095-9203. doi: 10.1126/science.aaf2403. URL http://www.ncbi.nlm.nih.gov/pubmed/27365449.

[196] Y. Steuerman, and I. Gat-Viks. Exploiting Gene-Expression Deconvolution to Probe the Genetics of the Immune System. *PLOS Comput. Biol.*, 12(4):e1004856, 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004856. URL http://dx.plos.org/10.1371/journal.pcbi.1004856.

[197] T. Stewart et al. Incidence of de-novo breast cancer in women chronically immunosuppressed after organ transplantation. *Lancet*, 346(8978):796–798, 1995. ISSN 01406736. doi: 10.1016/S0140-6736(95)91618-0.

[198] A. Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–50, oct 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102. URL http://www.ncbi.nlm.nih.gov/pubmed/16199517http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1239896.

[199] A. Sudhakar. History of Cancer, Ancient and Modern Treatment Methods. *J. Cancer Sci. Ther.*, 1(2):1–4, dec 2009. ISSN 1948-5956. doi: 10.4172/1948-5956.100000e2.

URL http://www.ncbi.nlm.nih.gov/pubmed/20740081http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2927383.

[200] V. Svensson et al. SpatialDE: identification of spatially variable genes. *Nat. Methods*, 15(5):343–346, mar 2018. ISSN 1548-7091. doi: 10.1038/nmeth.4636. URL http://www.nature.com/doifinder/10.1038/nmeth.4636.

[201] J. E. Talmadge, and D. I. Gabrilovich. History of myeloid-derived suppressor cells. *Nat Rev Cancer*, 13(10):739–752, 2013. ISSN 1474-1768. doi: 10.1038/nrc3581. URL http://www.nature.com/nrc/journal/v13/n10/pdf/nrc3581.pdf.

[202] D. Tamborero et al. A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations. *Clin. Cancer Res.*, page clincanres.3509.2017, apr 2018. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-17-3509. URL http://www.ncbi.nlm.nih.gov/pubmed/29666300.

[203] J. M. Taube et al. Implications of the tumor immune microenvironment for staging and therapeutics. *Mod. Pathol.*, 2017. ISSN 0893-3952. doi: 10.1038/modpathol.2017.156. URL http://www.nature.com/doifinder/10.1038/modpathol.2017.156.

[204] M. W. L. Teng et al. Immune-mediated dormancy: an equilibrium with cancer. *J. Leukoc. Biol.*, 84(4):988–993, oct 2008. ISSN 07415400. doi: 10.1189/jlb.1107774. URL http://www.ncbi.nlm.nih.gov/pubmed/18515327http://doi.wiley.com/10.1189/jlb.1107774.

[205] A. E. Teschendorff, and S. C. Zheng. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5):757–768, 2017. ISSN 1750-1911. doi: 10.2217/epi-2016-0153. URL http://www.ncbi.nlm.nih.gov/pubmed/28517979https://www.futuremedicine.com/doi/10.2217/epi-2016-0153.

[206] A. E. Teschendorff et al. Elucidating the Altered Transcriptional Programs in Breast Cancer using Independent Component Analysis. *PLoS Comput. Biol.*, 3(8):e161, 2007. ISSN 1553-734X. doi: 10.1371/journal.pcbi.0030161. URL http://dx.plos.org/10.1371/journal.pcbi.0030161.

[207] A. E. Teschendorff et al. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 2011. ISSN 1460-2059. doi: 10.1093/bioinformatics/btr171. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr171.

[208] V. Thorsson et al. The Immune Landscape of Cancer. *Immunity*, 48(4):812–830.e14, apr 2018. ISSN 1097-4180. doi: 10.1016/j.immuni.2018.03.023. URL http://www.ncbi.nlm.nih.gov/pubmed/29628290.

[209] I. Tirosh et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).*, 352(6282):189–196, 2016. ISSN 10959203. doi: 10.1126/science.aad0501.

[210] A. J. Titus et al. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.*, 26(R2):R216–R224, 2017. ISSN 0964-6906. doi: 10.1093/hmg/ddx275. URL http://www.ncbi.nlm.nih.gov/pubmed/28977446http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5886462https://academic.oup.com/hmg/article/26/R2/R216/3979364.

[211] F. Vallania et al. Leveraging heterogeneity across multiple data sets increases accuracy of cell-mixture deconvolution and reduces biological and technical biases. *bioRxiv*, page 206466, 2017. doi: 10.1101/206466. URL https://www.biorxiv.org/content/early/2017/10/20/206466.

[212] L. Van Der Maaten, and G. Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008. URL https://lvdmaaten.github.io/publications/papers/JMLR{_}2008.pdf.

[213] R. Van Noorden. Interdisciplinary research by the numbers. *Nature*, 525(7569):306–307, sep 2015. ISSN 0028-0836. doi: 10.1038/525306a. URL http://www.nature.com/doifinder/10.1038/525306a.

[214] A. Van Pel, and T. Boon. Protection against a nonimmunogenic mouse leukemia by an immunogenic variant obtained by mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 79 (15):4718–22, aug 1982. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/6981814http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC346748.

[215] A. Vandenbon et al. Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. *Proc. Natl. Acad. Sci. U. S. A.*, 113(17):E2393–402, apr 2016. ISSN 1091-6490. doi: 10.1073/pnas.1604351113. URL http://www.ncbi.nlm.nih.gov/pubmed/27078110http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4855614.

[216] V. E. Velculescu et al. Characterization of the yeast transcriptome. *Cell*, 88(2):243–51, jan 1997. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/9008165.

[217] D. Venet et al. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17 Suppl 1:S279–87, 2001. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/11473019.

[218] M. D. Vesely et al. Natural Innate and Adaptive Immunity to Cancer. *Annu. Rev. Immunol.*, 29(1):235–271, apr 2011. ISSN 0732-0582. doi: 10.1146/annurev-immunol-031210-101324. URL http://www.ncbi.

nlm.nih.gov/pubmed/21219185http://www.annualreviews.org/doi/10.1146/annurev-immunol-031210-101324.

[219] Virchow Rudolf. Archiv für pathologische Anatomie und Physiologie und für klinische Medizin., 1847. URL file://catalog.hathitrust.org/Record/000493993http://hdl.handle.net/2027/njp.32101076036878(Bd.64(1875))http://hdl.handle.net/2027/uc1.31175008645106(v.170{&}indexv.161-170)http://hdl.handle.net/2027/hvd.32044093331494(Bd.111(1888))http://hdl.han.

[220] M. Wang et al. Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, 7(1):328, 2006. ISSN 14712105. doi: 10.1186/1471-2105-7-328. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-328.

[221] N. Wang et al. The CAM Software for Nonnegative Blind Source Separation in R-Java. *J. Mach. Learn. Res.*, 14:2899–2903, 2013. URL http://www.jmlr.org/papers/volume14/wang13d/wang13d.pdf.

[222] N. Wang et al. UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics*, 31(1):137–139, 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu607. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu607.

[223] N. Wang et al. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.*, 6(1):18909, 2016. ISSN 2045-2322. doi: 10.1038/srep18909. URL http://www.nature.com/articles/srep18909.

[224] Z. Wang et al. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *bioRxiv*, page 146795, 2017. doi: 10.1101/146795. URL https://www.biorxiv.org/content/early/2017/10/14/146795.

[225] Y. Wen et al. Cell subpopulation deconvolution reveals breast cancer heterogeneity based on DNA methylation signature. *Brief. Bioinform.*, 18(3):bbw028, 2016. ISSN 1467-5463. doi: 10.1093/bib/bbw028. URL http://www.ncbi.nlm.nih.gov/pubmed/27016391https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw028.

[226] S. P. Wilczynski. Molecular Biology. In *Mod. Surg. Pathol.*, pages 85–120. Elsevier, 2009. ISBN 9781416039662. doi: 10.1016/B978-1-4160-3966-2.00006-0. URL http://linkinghub.elsevier.com/retrieve/pii/B9781416039662000060.

[227] J. D. Wolchok. PD-1 Blockers, 2015. ISSN 10974172.

[228] T. Yamamoto et al. Challenges in detecting genomic copy number aberrations using next-generation sequencing data and the eXome Hidden Markov Model: a clinical exome-first diagnostic approach. *Hum. Genome Var.*, 3(1):16025, dec 2016. ISSN 2054-345X. doi: 10.1038/hgv.2016.25. URL http://www.nature.com/articles/hgv201625.

[229] Z. Yang et al. A Convex Geometry-Based Blind Source Separation Method for Separating Nonnegative Sources. *IEEE Trans. Neural Networks Learn. Syst.*, 26 (8):1635–1644, 2015. ISSN 2162-237X. doi: 10.1109/TNNLS.2014.2350026. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6893008.

[230] K. Yoshihara et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, 4(1):2612, 2013. ISSN 2041-1723. doi: 10.1038/ncomms3612. URL http://www.nature.com/articles/ncomms3612.

[231] Z. Yu et al. CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. *BMC Bioinformatics*, 17(1):310, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1174-7. URL http://www.ncbi.nlm.nih.gov/pubmed/27538789http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4990858http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1174-7.

[232] M. Zarrei et al. A copy number variation map of the human genome. *Nat. Rev. Genet.*, 16(3):172–183, mar 2015. ISSN 1471-0056. doi: 10.1038/nrg3871. URL http://www.ncbi.nlm.nih.gov/pubmed/25645873http://www.nature.com/articles/nrg3871.

[233] M. E. Zaslavsky et al. Infino: a Bayesian hierarchical model improves estimates of immune infiltration into tumor microenvironment. *bioRxiv*, page 221671, 2017. doi: 10.1101/221671. URL https://www.biorxiv.org/content/early/2017/11/21/221671.

[234] C. Zheng et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell*, 169(7):1342–1356.e16, 2017. ISSN 0092-8674. doi: 10.1016/J.CELL.2017.05.035. URL https://www.sciencedirect.com/science/article/pii/S0092867417305962.

[235] Y. Zhong et al. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, 14(1):89, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-89. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-89.

[236] A. Zinovyev et al. Blind source separation methods for deconvolution of complex signals in cancer biology. *Biochem. Biophys. Res. Commun.*, 430(3):1182–1187, 2013. ISSN 0006-291X. doi: 10.1016/J.BBRC.2012.12.043. URL https:

//www-sciencedirect-com.gate2.inist.fr/science/article/pii/S0006291X12023741?via{%}3Dihub.

[237] J. Zou et al. Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods*, 11(3):309–311, 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2815. URL http://www.nature.com/articles/nmeth.2815.

[238] W. Zou et al. PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: Mechanisms, response biomarkers, and combinations. *Sci. Transl. Med.*, 8(328), 2016. ISSN 19466242. doi: 10.1126/scitranslmed.aad7118.

[239] N. S. Zuckerman et al. A Self-Directed Method for Cell-Type Identification and Separation of Gene Expression Microarrays. *PLoS Comput. Biol.*, 9(8):e1003189, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003189. URL http://dx.plos.org/10.1371/journal.pcbi.1003189.

[240] Y. Şenbabaoğlu et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.*, 17(1):231, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1092-z. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1092-z.