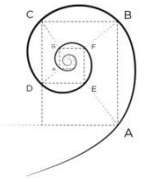
UNIVERSITÀ DEGLI STUDI
DI MILANO



DOTTORATO IN MEDICINA MOLECOLARE E TRASLAZIONALE

CICLO XXXII

Anno Accademico 2018/2019

TESI DI DOTTORATO DI RICERCA

BIO10

Development of a bioinformatics framework to identify cell subpopulations from bulk transcriptional data

Dottorando: Andrea Grilli

Matricola N°: R11684

Tutore :	Prof. Cristina Battaglia
Co-tutore:	Prof. Silvio Bicciato
Coordinatore del dottorato:	Prof. Michele Samaja

Ad Alessandro, Leonardo e Giulia,

perché si può raggiungere qualsiasi traguardo.

Ad Elisa, perché da solo non l'avrei raggiunto.

Science is the attempt to make the chaotic diversity of oursense-experience correspond to a logically uniform system of thought.

A. Einstein

Sommario

I livelli d'espressione genica dei campioni biologici sono influenzati dalla composizione intrinsecamente eterogenea delle cellule e dei tessuti che li compongono. Ciononostante, nell'analisi dei profili trascrizionali, il segnale di ogni campione viene quantificato senza considerare la presenza di sottopopolazioni cellulari trascrizionalmente molto diverse tra loro (analisi in bulk). Questa limitazione può essere estremamente critica quando si analizzano i profili di espressione genica in campioni tumorali, dove definire la composizione cellulare del tessuto rappresenta un aspetto fondamentale per svelare l'eterogeneità intratumorale e i meccanismi molecolari che determinano i diversi comportamenti del cancro. Poiché cambiamenti nella composizione cellulare del tessuto tumorale possono influenzare sia la previsione della sopravvivenza sia la risposta al trattamento del paziente, è estremamente importante poter quantificare in maniera accurata i sottotipi cellulari contenuti nella massa tumorale. A tal fine sono stati sviluppati una serie di metodi computazionali di deconvoluzione che vengono utilizzati per inferire, dall'espressione genica del campione in toto, le quantità relative dei sottotipi cellulari che lo compongono. Storicamente, questi metodi sono stati sviluppati per quantificare le proporzioni delle popolazioni leucocitarie e le loro prestazioni sono state validate soprattutto su profili di cellule purificate.

In questo progetto, abbiamo inizialmente stabilito lo stato dell'arte degli attuali metodi di deconvoluzione del segnale di espressione genica. Dopo la loro valutazione, abbiamo selezionato quattro metodi (CIBERSORT, EPIC, ssGSEA e xCell) per definire una *pipeline* per l'analisi di deconvoluzione. Successivamente, attraverso l'uso di studi indipendenti e selezionati, abbiamo analizzato come questi metodi stimino i diversi tipi cellulari originati da differenti formati di dati. Innanzitutto, abbiamo valutato potenziali errori di stima dei metodi di deconvoluzione usando profili di cellule purificate raccolte da diversi *dataset* pubblici. Quindi, utilizzando tre *dataset* pubblici di *single cell* RNA-seq da diversi tumori (carcinoma mammario, carcinoma polmonare e melanoma), abbiamo valutato la capacità di ogni metodo di stimare diversi tipi cellulari a diverse proporzioni. I risultati di tutte le analisi sono pubblicamente fruibili attraverso l'applicazione web ARDESIA (https://bicciatolab.shinyapps.io/ardesia/).

Nella seconda parte di questo lavoro abbiamo valutato l'applicabilità della *pipeline* di deconvoluzione a dati provenienti da organismi diversi o per la determinazione dei sottotipi molecolari in campioni tumorali. Nel primo caso, abbiamo sfruttato un database murino contenente dati d'espressione generati in condizioni rigorosamente standardizzate (ImmGen) per creare una nuova firma genica in grado di discriminare un ampio numero di popolazioni immunitarie, in particolare della linea mieloide. Testata su campioni purificati murini, questa nuova firma è stata in grado di discriminare cellule strettamente correlate e con profili trascrizionali simili, come lo stesso sottotipo leucocitario estratto da tessuti diversi (ad esempio

macrofagi da tessuto alveolare o peritoneale). Nel secondo caso, abbiamo applicato i diversi metodi di deconvoluzione per studiare l'eterogeneità molecolare del carcinoma mammario (BC). A tal fine, siamo partiti da un *dataset* di carcinoma mammario, i cui campioni erano stati etichettati in diversi sottotipi molecolari sulla base di segnali di immunoistochimica (IHC), per creare una firma genica di 230 geni caratterizzanti i diversi sottotipi molecolari. Successivamente, abbiamo applicato questa firma per la deconvoluzione di 2 coorti di campioni di carcinoma mammario triplo negativo (TNBC). Sebbene i campioni dovessero essere clinicamente omogenei, l'analisi di deconvoluzione ha evidenziato che circa il 40% dei campioni presenta invece un grado variabile di co-presenza di più di un sottotipo molecolare. L'analisi di associazione della frazione TNBC, definita attraverso la deconvoluzione, e della risposta clinica o della sopravvivenza ha evidenziato l'esistenza di un sottogruppo di pazienti caratterizzato da una risposta e una sopravvivenza inferiori, proprio in virtù dell'eterogeneità molecolare della massa tumorale.

In conclusione, abbiamo creato una *pipeline* computazionale per identificare sottopopolazioni cellulari da dati trascrizionali di campioni mediante analisi di deconvoluzione. Inoltre, abbiamo generato due firme molecolari per determinare la presenza di popolazioni immunitarie (umane o murine) e di sottotipi molecolari del carcinoma mammario.

Summary

Expression levels of biological samples are affected by the intrinsic heterogeneity of cells and tissue composition. Nevertheless, in bulk transcriptional profiling, each sample is evaluated without considering the presence of multiple subpopulations. This limitation might be extremely critical when analyzing bulk gene expression profiles of cancer samples, where dissecting the mix of cell populations could shed light on the intratumoral heterogeneity and on the molecular mechanisms shaping different cancer behaviors. Since changes in tumor composition can both impact the prediction of patient survival and therapeutic response, reaching high confidence about the real content within these bulk tissues is extremely significant. For this reason, several deconvolution tools have been developed to infer (deconvolve) the signals of each constituent cell type from bulk gene expression data. Historically, these tools have been mainly developed to define leukocyte proportions, and their performance has been mostly validated on profiles of purified cells.

In this project, we initially established the state-of-art of existing methods for transcriptional deconvolution. After their evaluation, we finally retained four tools (CIBERSORT, EPIC, ssGSEA and xCell) to define a bioinformatics framework for the deconvolution analysis. Next, using independent and selected studies, we investigated how these selected tools perform on different cell types and data format.

First, we assessed presence of potential biases of deconvolution methods using profiles of purified cells from different public datasets. Then, based upon three public single cell RNA-seq datasets from different tumors (breast cancer, lung cancer and melanoma), we evaluated tools capability in estimating different cell types at variable abundances, eventually wrapping these results in an interactive web application named ARDESIA (https://bicciatolab.shinyapps.io/ardesia/).

The second part of this work investigated adaptability of the deconvolution analysis pipeline and its application in different conditions. To this end, we exploited a mouse database containing expression data generated in rigorously standardized conditions (ImmGen) to create a novel gene signature able to discriminate a widespread number of immune cellular populations, in particular from the myeloid lineage. When tested on murine purified samples, this new signature was able to discriminate closely related cells with similar transcriptional profiles, like the same cell type from different tissues (e.g. macrophages from alveolar or peritoneal tissue). Based on this validation, we applied the same approach to further investigate subtype heterogeneity in breast cancer (BC). To this end, we started from a dataset of breast cancer subtypes based on immunohistochemistry (IHC) to create a custom gene signature of 230 genes. Then, we applied this signature to deconvolve 2 cohorts of clinically-defined triple negative breast cancer (TNBC) samples. Although both datasets were clinically uniform, deconvolution analysis highlighted a variable degree of heterogeneity in tumor

subtypes for about 40% of samples. Test of the TNBC fraction identified through deconvolution with either clinical response or survival refined a subgroup of patients characterized by poorer response and survival due to heterogeneous composition of the tumor.

In conclusion, we created a general bioinformatics framework to identify cell subpopulations from bulk transcriptional data by deconvolution analysis. Furthermore, we generated two molecular signatures to addressed bulk heterogeneity either for immune populations in mouse or tumor subtypes in breast tumors.

Index

Abbreviations

aDC	**a**ctivated **D**endritic **C**ells
BC	**B**reast **C**ancer
CD4	**T**-cells **CD4**+
CD8	**T**-cells **CD8**+
cDC	**c**onventional **D**endritic **C**ells
CN	**C**ondition **N**umber
CPM	**C**ount **P**er **M**illion
DEG	**D**ifferentially **E**xpressed **G**enes
EPIC	**E**stimate the **P**roportion of **I**mmune and **C**ancer
ES	**E**nrichment **S**core
GSEA	**G**ene **S**et **E**nrichment **A**nalysis
iDC	**i**mmature **D**endritic **C**ells
IHC	**I**mmuno**H**isto**C**hemistry
MΦ	**M**acrophages
NES	**N**ormalized **E**nrichment **S**core
NK	**N**atural **K**iller
PBMC	**P**eripheral **B**lood **M**ononuclear **C**ell
PCA	**P**rincipal **C**omponent **A**nalysis
pCR	**P**athological **C**omplete **R**esponse
pDC	**p**lasmacytoid **D**endritic **C**ells
RNA-seq	**RNA** sequencing
sc	**s**ingle **c**ell
scRNA-seq	**s**ingle **c**ell **RNA seq**uencing
SVM	**S**upport **V**ector **M**achine
TCGA	**T**he **C**ancer **G**enome **A**tlas
TME	**T**issue **M**icro-**E**nvironment
TPM	**T**ranscript **P**er **M**illion

1. Introduction

Usually, expression levels in biological samples are determined by the heterogeneity of cells and tissue composing the analyzed sample. Indeed, a sample generally comprises the investigated tissue along with a variable presence of several other components, such as fibroblast, blood vessels and, importantly, infiltrating leukocytes. Nevertheless, when analyzing transcriptional profiles, each sample is generally evaluated as a bulk, without considering the presence of multiple subpopulations [1], making sample heterogeneity one of the major confounders in gene expression studies. This is especially true when considering samples derived from pathological biopsies [2], due to the presence of variable proportions of tumor and healthy tissues, along potentially relevant infiltration of multiple components of the hematopoietic system. Recent advances depict a wide repertoire of immune cellular subtypes [3] and an unexpected complex ecosystem of cells in the tumor microenvironment (TME) [4]. Tumor-infiltrating lymphocytes (TIL), tumor-infiltrating myeloid cells (TIM), tumor-associated macrophages (TAM) terms are examples of the close interaction between cancer cells and the surrounding elements of the immune system. However, immune cells can play different roles in cancer and act in distinct or even opposite manners, with either tumor or antitumor activity: cells originating from the same type but characterized by different marker expression landscapes are

considered as separated and sometimes counterpoised entities [5,6] (**Figure 1.1**).

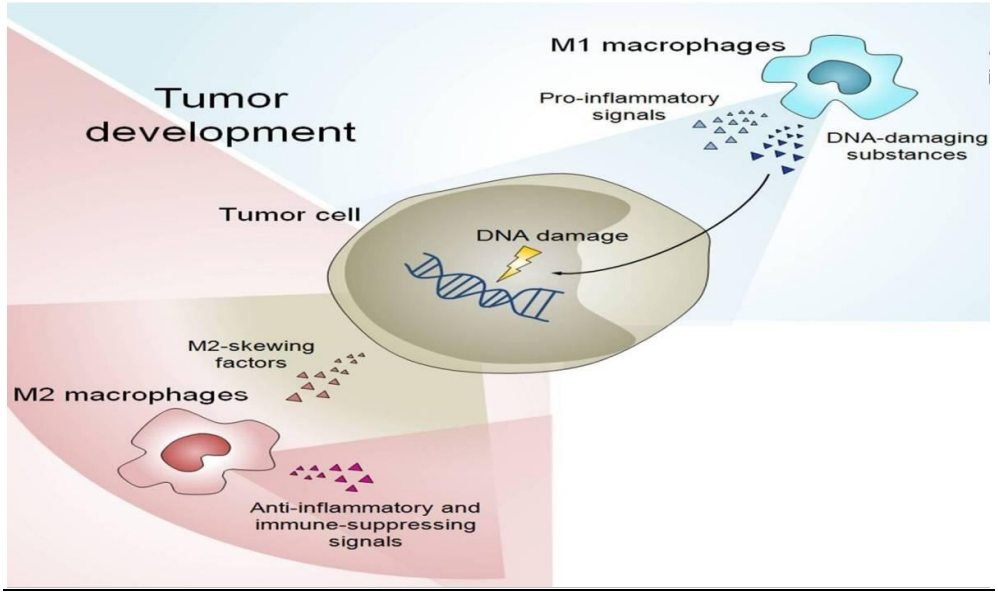


Figure 1.1. A classic example of immune counterpoised roles in cancer: monocytes can be activated in two distinct phenotypes with opposite roles in tumor progression and response, the M1 or M2 macrophages. The first phenotype is a classical activation of monocytes with immunostimulatory activity and production of pro-inflammatory cytokines, which promote an anti-tumor activity; on the contrary, M2 is an alternative activation with an opposite activity by production of anti-inflammatory cytokines and an immunosuppressive action. Image modified from [10].

In cancer, the definition of the non-tumoral proportion in each biological sample significantly improves the accuracy of response prediction by expression data analysis [7]. It is now well established that the immune response has the potential to enhance clinical prediction, thus providing potential candidates for immunotherapy [8]. Also, small differences in immune infiltrate are associated to both prognosis and response to treatment [9]. Thus, detailed identification of infiltrating cells composition is increasingly important both in

pathological and physiological conditions. However, routinely addressing patients to best treatment needs dependable and validated approaches for quantitative estimating of TME that are still missing.

To date, flow cytometry is the reference method for the identification of cell subtypes in the immune system, however, it requires large amount of material and can be sometimes equivocal due to the association of a specific surface marker to multiple cell types. Other dedicated techniques for single-cell isolation are available, e.g. the laser micro-capture dissection or microfluidics, but nowadays they are expensive and require highly specialized resource, further to be low-throughput in some cases [11]. An effective alternative to accurately disclose the fractions of the cellular content within a tumor is the emerging single cell (SC) technology. scRNA-seq is, in fact, a method that allows the analysis of the transcriptome of every single cell present in a sample. This technique has enhanced a considerable progression in the understanding of samples heterogeneity. Also, in cancer studies, it highlighted a complexity and heterogeneity of TME not previously hypothesized [4]. For these reasons, scRNA-seq is a useful tool for the evaluation of the diverse cellular content within the samples, even though with several limitations: (i) the complexity required to analyze this kind of data; (ii) high costs; (iii) recent introduction; (iv) infeasibility on fixed samples, which importantly hampers its use in the clinical setting. All these drawbacks hinder the creation of large case studies with scRNA-seq, which on the contrary are currently available for bulk profiles.

Nowadays, large repositories containing the profile of thousands of bulk samples are available for interrogation and investigation, either of healthy (GTEx) or tumor samples (TCGA). Indeed, gene expression profiling by array or RNA-seq is a consolidated and cheaper technique and expression data from most, if not all cell subpopulations, are already publicly available.

1.1. Define immune infiltration by transcriptional deconvolution

Deconvolution was introduced in the 1950s in seismology, but it was rapidly and extensively deployed to the correction of optical images, especially in telescope observation (**Figure 1.2**). Each space image is actually a mixture of different signals at different wavelengths, which correspond to the different chemical elements observed in the patch of the sky that the instrument is looking at. The application of deconvolution algorithms better clarifies the reflectance of materials at different wavelengths that are mixed according to the material composition of each pixel [12], thus improving the final image definition. In 2001, Venet and colleagues [13] for the first time applied a deconvolution algorithm to transcriptional data analysis. Similarly, to optical images, each gene signal in a bulk is a mix of the expressions of that gene from the constituent cell types of the sample. However, the inherent heterogeneity of tissue composition makes difficult to define the real expression level of each cell type, also considering that further noise factors take over on expression

determination, such as cross-hybridization or library protocol, depending on the technique.

Importantly, differential gene expression analysis can be influenced by unexpected changes in cell types composition rather than real gene expression modifications between conditions. Thus, the composition and variation in cell types are important factors in determining several biological conditions, such as the definition of diagnosis, response to therapy or a specific pathological state. For these reasons, several deconvolution tools have been developed to infer *in silico* (deconvolve) the signals of each constituent cell type from bulk gene expression and reconstitute the cellular composition of the sample.

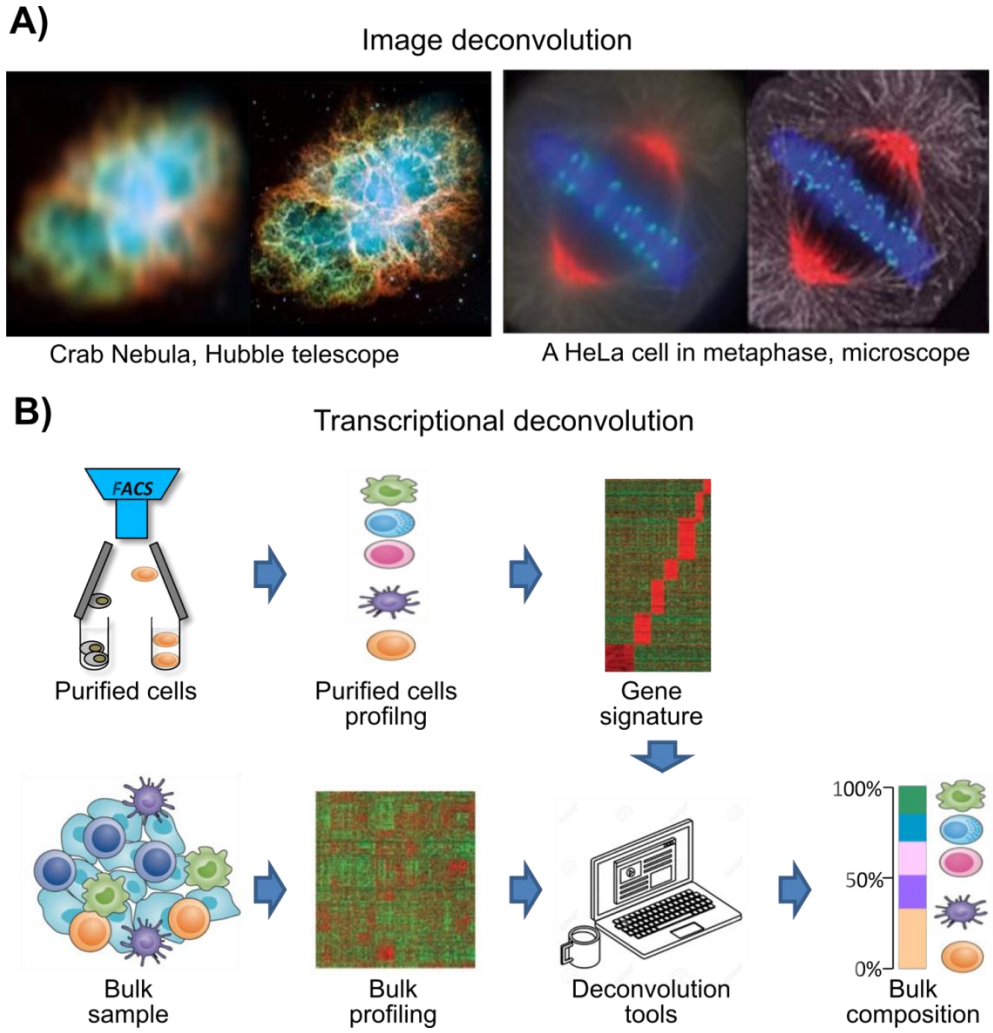


Figure 1.2. **A)** Two examples of optical deconvolution: the Crab Nebula, a remnant after a supernova explosion as acquired by Hubble telescope [14] and a HeLa cell in metaphase (adapted from [15]). In both cases, on raw images (on left) a specific type of deconvolution for optical images, e.g. blind deconvolution, is applied to obtain an improved image (on right) and clarify the observed composition; **B)** Framework for the application of deconvolution to transcriptional analysis. Purified cells (first row) are collected, from FACS or other techniques; a signature matrix or a gene list is derived from purified cell transcriptional profiles to distinguish among the selected cell types. The bulk profile (second row) is a mix of the expression signals from the constituent cell types; nevertheless, by transcriptional deconvolution we can infer a score, e.g., the fraction or enrichment, enlightening sample composition.

There exist three main different purposes for the utilization of deconvolution methods: (i) estimate the proportion of the different cell types in a bulk; (ii) calculate the proportion and gene expression profile of each of the constituent cell types in a bulk; (iii) in tumor analysis, estimate the purity of samples for quantifying both the non-tumoral and tumoral fractions. In this thesis, I'll mainly focus on the first application of deconvolution methods.

1.2. Deconvolution algorithms and gene signatures

Tools for transcriptional deconvolution are based on two components: the algorithm and the gene signature (see Figure 1.3, "*Tools selection*" box for details of the tools used in this work). Mathematically, the deconvolution problem applied to bulk gene expression is generally defined with the formula:

$$M = f \; x \; B$$

where *M* is the bulk gene expression matrix, *f* is the unknown fractions of each cell type in the mixture and *B* is the gene signature matrix (with the constrain of the number of genes in *B* > number of cell types to determine, which is usually respected in gene expression studies). However, in bulk deconvolution M is the only known factor, whereas f and B should be estimated. Different types of algorithms have been applied in deconvolution analysis: the most used are based on regression, enrichment, non-negative matrix factorization and probabilistic methods. Briefly, regression-based

algorithms define the mixture profiles as variables dependent (Y) on the profiles of the reference signature (X). Different models of regression algorithms exist, such as the least-square regression [16] or the Support Vector Machine (SVM) [17]. However, independently of the specific model, the use of regression-based methods results in a determination of the relative fraction of the cell types present in the mixture. Conversely, enrichment algorithms have been initially conceived to define differential enrichment of gene sets between two different conditions, e.g. pathways or gene ontologies. In detail, genes are ranked according to a statistic, i.e. significance or fold change between conditions, and a score, named enrichment score (ES), defines if genes from a signature are significantly enriched in the top or down ranked list. In 2009, Barbie and colleagues [18] proposed an adaptation to calculate the enrichment score within a single sample, wherein this specific case the gene list is ranked by the absolute expression in that sample. The generated enrichment score can be applied without modifications [19], or after correction to make it linearly associated with the mixture fraction [20]. Importantly, the score calculated by enrichment analysis corresponds to the cellular activity of the analyzed gene signature, rather than to the cellular proportion.

The second fundamental component for deconvolution analysis is the gene reference. It is composed of marker genes used by the algorithm to infer either the proportion or enrichment for a specific cell type. Ideally, a marker gene should be expressed uniquely in a specific cell type, but constantly expressed across different biological

conditions, e.g., in response to a stimulus or pathological disease, at satisfactory levels to be detected and platform-independent. If we consider the intrinsic gene expression variability across different tissues or states during differentiation processing, the definition of a marker gene is often challenging. Generally, the selection of markers takes place by differential expression of a purified cell type compared to the other populations that compose the gene reference, and the most significant genes are included in the gene signature. In some cases, highly modulated genes according to fold change are used upon filtering based on significance [17]. Alternatively, additional filtering on genes can be performed by manual selection [16], exclusion of non-hematopoietic genes [17], high expression in cancer cell lines [17,20] or by correlation across all specific immune genes in the same cell type [19].

For enrichment-based tools the list of the genes is sufficient to perform the deconvolution [19], whereas regression methods also require a measure of genes variability across cell types of the reference, either it being the gene expression [17] or the standard deviation [16]. A signature can be composed by tens to hundreds genes, and the definition of the optimal number is often dependent by a mix of biological and technical considerations; a short list can be more easily defined by highly cell-specific genes, but some of them could be absent in the mixture to deconvolve and the impact of missing genes in a signature has not yet been completely addressed. On the other hand, excessively large number of genes in the signature can be source of noise during the deconvolution. To at

least partially address this problem, in some pipeline [17] multiple gene references are generated with an increasing number of marker genes per cell type, and the “most stable” reference is finally used. The reference stability is a mathematical property measured by introducing input variation or noise and it is evaluated by a score called condition number (CN); lower CN corresponds to higher matrix stability. In other cases [20], multiple gene signatures are generated for each cell type, and top signatures, which are reliable also on test set data, are finally chosen. In the gene reference generation, a further condition should be considered: closely related cell types can also have similar gene signatures (multicollinearity), which influence each other and can result in their inaccurate estimates; specifically, high abundance of a cell type can influence quantification of a less abundant population with which it shares common marker genes.

For now, there is no method that can be defined as optimal for the selection of marker genes. The main indication remains the use of marker genes from reference profiles as close as possible to the samples to be examined. Usually, reference profiles used for the generation of gene signature derive from purified cells, mainly isolated from blood: their expression profiles from different datasets profiled using different platforms are then collected to create a meta-dataset with all the cell types that will be included in the gene signature [17,19]. In almost all gene signature generations, hematopoietic cells from a health condition are generally considered an ideal reference also for immune infiltrating cells [17,19,20], with the exception of few tumor-specific cell types, e.g. M2 Macrophages

[17]. In some cases, different gene signatures are specifically available for the evaluation of immune infiltration in either healthy or tumor tissues [16]: signatures for healthy tissues are generated by RNA-Seq from bulk of Peripheral Blood Mononuclear Cell (PBMC) samples and sorted immune cells of either healthy or pathological non-cancer samples, e.g. multiple sclerosis or type one diabetes. Conversely, a second gene reference is created using tumor-infiltrating cells from scRNA-seq [21]. Some authors [20], created a big compendium of gene signatures after collection of a large set of populations from several big data sources, e.g. ENCODE [22] or Blueprint [23] projects, and preserving only signatures reliable on several data sources.

My thesis project is organized as follow (Figure 1.3): I initially established the state-of-art of existing methods for transcriptional deconvolution, searching for tools with a modifiable framework (ch: "*Define deconvolution tools for a customizable framework*"; Figure 1.3., "*Tools selection*" section); afterwards, for a selection of deconvolution tools, I assessed presence of potential biases using profiles of purified cells from different conditions, achieving important indications and drawbacks in the use of deconvolution methods (ch: "*Test deconvolution tools on human leukocyte populations*"). Then, I took advantage of three different single-cell RNA-seq datasets to evaluate the ability of deconvolution tools in estimating different cell types at variable abundances; these results have been thereafter aggregated in an interactive web application (ch: "*Single-cell RNA-*

seq as a gold standard to test tools performance"; Figure 1.3., "*Tools validation*" section). In a later step, to evaluate to which extent deconvolution framework can be modified, I changed the organism for which it has been designed for. By creation of a large murine meta-dataset, I moved the deconvolution framework from human to mouse immune populations detection. (ch: "*Creation of a murine gene signature for immune heterogeneity*"). Finally, I assessed molecular subtype heterogeneity in breast cancer bulk tumors by deconvolution analysis; the detected cellular fractions have been lastly tested for association with clinical response and survival (ch: "*Define tumor heterogeneity in breast cancer*"; Figure 1.3., "*Address heterogeneity*" section).

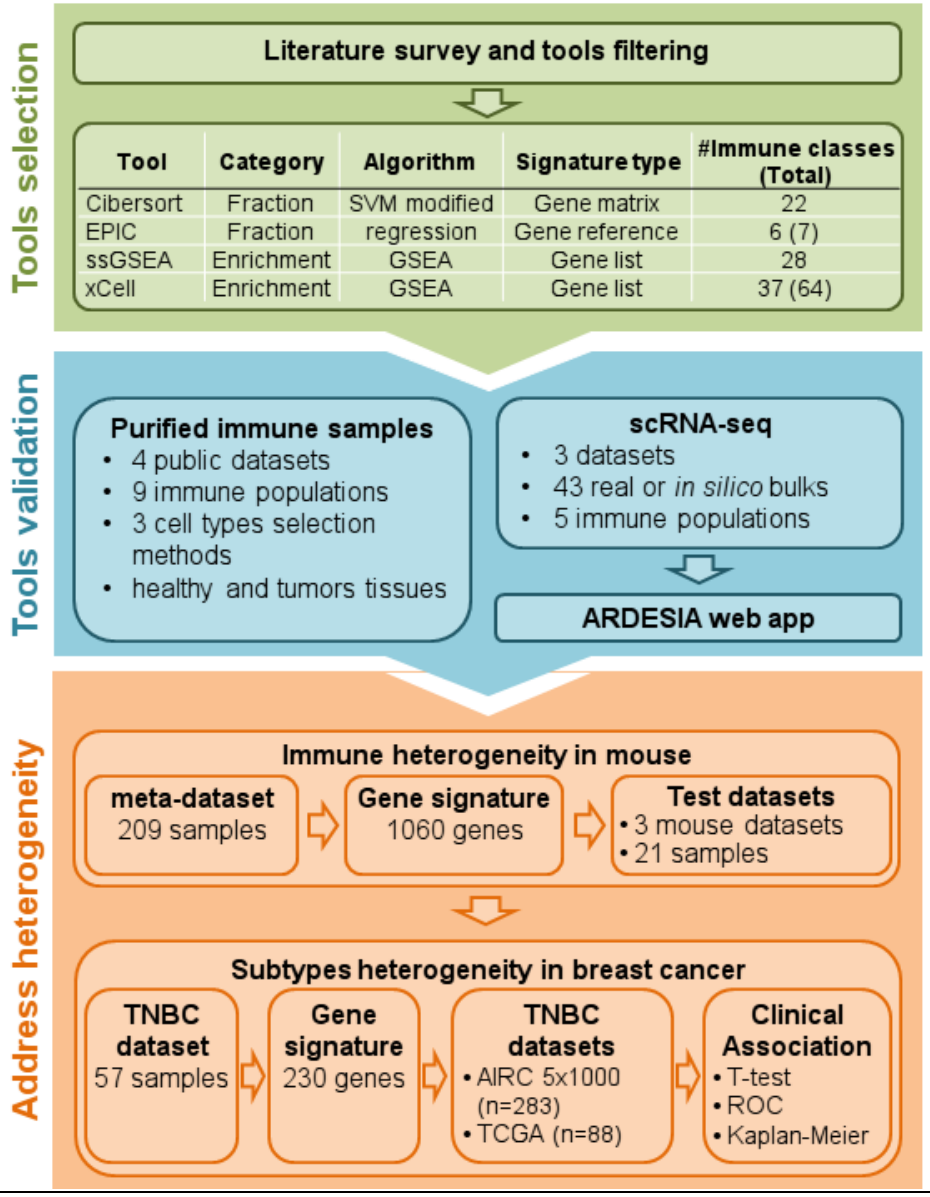


Figure 1.3. The first part (light green) focused on a survey of the available deconvolution methods and selection of the tools with a modifiable framework; the selected tools are listed together with their main features. The second the second part (light blue) performed an assessment of the four selected tools by the use of both purified immune cells and single cell RNA-seq datasets. These steps helped in the definition of the tool, i.e., CIBERSORT, used in the final part of the project (orange) to address transcriptional heterogeneity either for murine immune populations first and for subtypes heterogenity in breast cancer finally.

2. Aim of the thesis

Gene expression levels are partially affected by the intrinsic heterogeneity of cells and tissue composition of biological samples. Nevertheless, when analyzing transcriptional profiles each sample is generally evaluated as a bulk, without considering the presence of multiple subpopulations. For this reason, several deconvolution tools have been developed to infer *in silico* the signals of each constituent cell type from bulk gene expression and reconstruct the cellular composition of the sample, i.e. transcriptional deconvolution. Historically, these tools have been mainly developed to define leukocyte proportions and are still lacking a gold standard protocol for validation in solid tumors.

Given these premises, the main focus of this project has been the definition of a bioinformatics framework aimed at the investigation of heterogeneity in bulk transcriptional data of healthy and tumor samples, using deconvolution methods. A preliminary assessment of these methods, based on independent and selected studies from bulk profiles of purified cells and of scRNA-seq experiments was necessary, due to the lack of a gold standard to evaluate transcriptional deconvolution performances. These results guided the development of an accurate computational deconvolution analysis framework, tested in different conditions and datasets. At first, it was evaluated through the generation of a dedicated murine signature for immune populations detection. Finally, it has been applied to address

intratumoral heterogeneity in breast cancer, testing the detected cellular fractions for significant association with clinical response and survival.

3. Materials and methods

3.1 Tools for transcriptional deconvolution

Selection of deconvolution tools has been based on three criteria: (i) tool usability (ii) availability of a well-defined immunological signature (iii) customizable gene signature at the time of my survey (December 2017).

Selected tools could be classified according to the information they report to the user in fraction-based methods (CIBERSORT[17] and EPIC[16]) and enrichment-based methods (ssGSEA[19] or xCell[20]). In the former, for each sample the output corresponds to fractions of the cell types from the gene signature: fractions can be calculated as either a ratio among all immunological subtypes only (CIBERSORT) or a ratio including both hematopoietic and non-hematopoietic tissues (EPIC). In the latter method, the output for each sample is a score of each cell type (enrichment score) which corresponds to an amount of activity of a given list of genes.

Each tool was implemented with a gene signature able to identify a variable number of immunological populations and, in some case, also for possible surrounding tissues. In these tools, the gene signature can be replaced with a custom, user-defined one, providing the ability to identify a different set of populations; however, each tool

may require a different input setting, being either a list of marker genes or a gene expression matrix.

In my work, all the deconvolution analyses have been performed using the predefined parameters of each tool or, if otherwise, as detailed for each specific analysis. For all tools and analyses, the deconvolution data was subsequently analyzed using different packages and functions in R on the base of the purpose (see below for details).

3.1.1. CIBERSORT

CIBERSORT[17] is a deconvolution tool based on a modification of a largely used classifier algorithm, the Support Vector Machine (SVM). This tool is available via a web page (https://cibersort.stanford.edu/, **Figure. 2.1**) upon free registration. An R version also exists but is available only upon request from authors. CIBERSORT reports the relative proportion, named "*relative mode*", of 22 different human leukocyte populations. Recently, a method called "*absolute mode*" has been made available: it transforms cellular fractions into a score that reflects the absolute proportion of each cell type in the whole sample. This method was not discussed in the original paper and it is still under development, so it was not applied in this thesis. Further to fractions, different statistics are reported for each analyzed sample, e.g. p-value, Pearson's correlation and RMSE: these metrics are generated by the comparison of the mixture profile with the gene signature, meaning that a significant p-value indicates a considerable

presence of some of 22 populations in the bulk. These metrics are better discussed in the chapter below "*Statistics for performance evaluation*".

CIBERSORT is based upon a signature called "LM22", which is defined by a gene expression matrix composed of 547 genes for each of the 22 investigated cell subtypes. This signature was generated and validated using gene expression data of immune purified cells profiled both by Affymetrix and by Illumina microarrays. The populations in the gene signature matrix are:

- Lymphoid lineage: B-cells naïve, B-cells memory, Plasma cells, T cells $CD8^+$, T cells $CD4^+$ naïve, T cells $CD4^+$ memory resting, T cells $CD4^+$ memory activated, T cells follicular helper, T cells regulatory (Tregs), T cells gamma delta, NK cells resting, NK cells activated;

- Myeloid lineage: Monocytes, Macrophages M0, Macrophages M1, Macrophages M2, Dendritic cells resting, Dendritic cells activated, Mast cells resting, Mast cells activated, Eosinophils, Neutrophils.

This tool has been designed to work mainly on array data, but it is also widely used for RNA-seq data [24], for which there is a specific option to set in the analysis. The input data should contain positive values in linear scale; however, the tool automatically detects and converts log2 data into non-log linear space.

In my project, the analyses with CIBERSORT have been performed using the web interface. Expression data have been uploaded in CIBERSORT as either log2 or as CPM (Count Per Million) and TPM (Transcript Per Million) normalized expression data, respectively for datasets profiled by array or by RNA-seq. All the analyses have been performed in *relative mode*, e.g. fractions detection, using 1000 permutations (the maximum available in the tool) and disabling the quantile normalization for RNA-seq data only. Results have been downloaded as tabular data (.txt file).

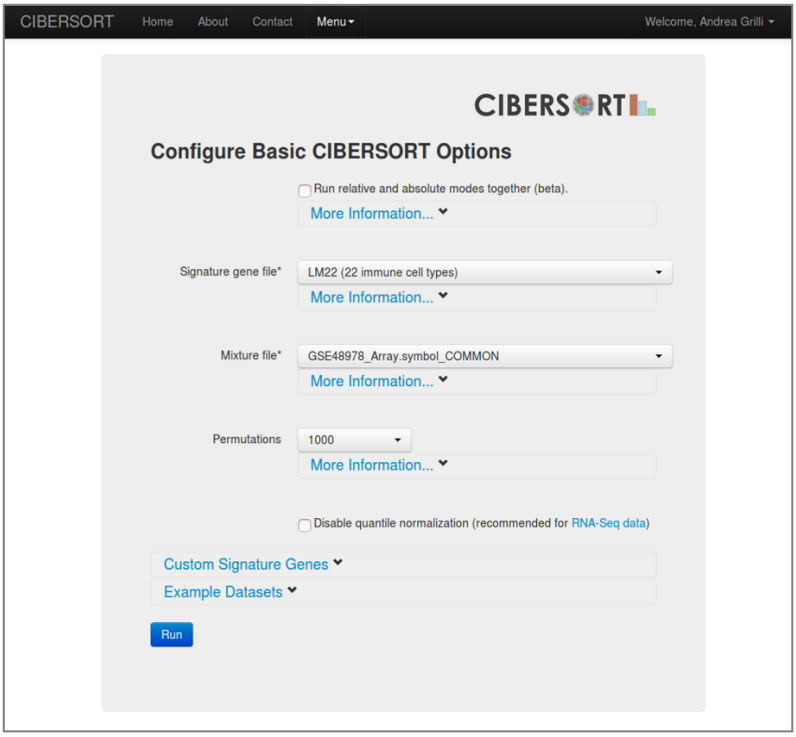


Figure 2.1. The CIBERSORT web page: the deconvolution analysis starts selecting the immunological signature, loading the gene expression data, and setting the appropriate parameters. The tool is accessible only upon free registration.

3.1.2. EPIC

EPIC [16] is a tool based on a linear regression algorithm and it is accessible both as a web application (http://epic.gfellerlab.org/, **Figure 2.2**) and as an R package at github repository (https://github.com/GfellerLab/EPIC). Two different reference profiles are available with this tool: one reference, named “blood circulating immune cells*”,* is used to determine the proportion of immune populations of circulating cells, while a second reference, named “tumor-infiltrating cells”, contains also the signatures for tumor-infiltrating non-malignant cell types (stroma and endothelial cells). A further class, named as "other cells", is automatically generated during the analysis by the tool; this class comprises tumor cells or tissues different from the previous ones. In my thesis, the “blood circulating immune cells” signature has been used and it is composed of:

- Lymphoid lineage: B-cells, CD4 T-cells, CD8 T-cells;
- Myeloid lineage: Monocytes, Neutrophils, NK cells.

This tool was designed specifically for RNA-seq data, but it can be applied on any mixture sample: it requires as input a gene expression data matrix, which should be normalized also for the length of the gene, e.g. TPM or RPKM, for the analysis of RNA-seq. No other assumptions are required about expression distribution (use of log2 or linear data). As for CIBERSORT, the total sum of all population fractions is equal to 1. However, it differentiates from CIBERSORT for both the algorithm (it used the linear regression),

and because it introduces weights related to gene variability per gene per cell type, e.g. the interquartile range of the expression or the standard deviation, either for the circulating or the tumor-infiltrating references, respectively.

All the analyses have been performed using the web application. Expression data have been uploaded in EPIC as either log2 or as CPM and TPM normalized expression data, respectively for array or RNA-seq datasets. No parameters should be specified during the analysis. Results are presented as cells fraction and are available as tabular data.

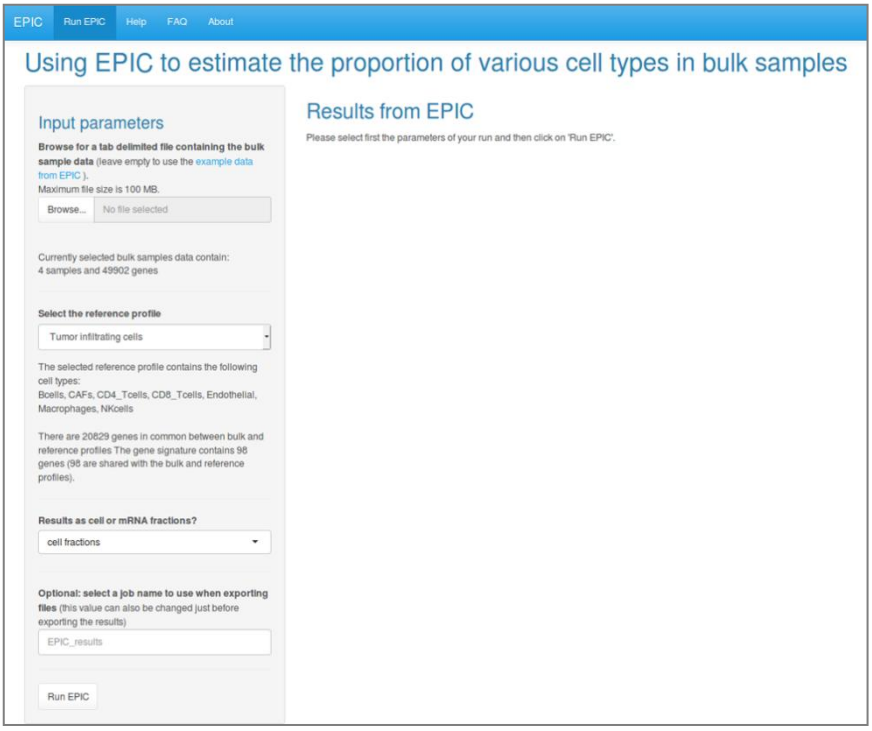


Figure 2.2. The EPIC homepage.

3.1.3. single sample GSEA (ssGSEA)

This algorithm is an adaptation of the most popular GSEA (Gene Set Enrichment Analysis)[25] that, although initially conceived for detecting enrichment among conditions, in 2009 Barbie and colleagues [18] modified for single sample analysis. Later on, this approach has been used by Charoentong and its collaborators [19] for analyzing the immune composition of 20 different tumors from the TCGA. In particular, they created immunological signatures able to distinguish among 28 different leukocyte subtypes to be used with the ssGSEA:

- Leukocyte lineage: activated B-cell, activated CD4 T-cell, activated CD8 T-cell, central memory CD4 T-cell, central memory CD8 T-cell, effector memory CD4 T-cell, effector memory CD8 T-cell, gamma delta T-cell, immature B-cell, memory B-cell, regulatory T-cell, T follicular helper, type 1 T helper, type 17 T helper, type 2 T helper, CD56bright NK cell, CD56dim NK;

- Myeloid lineage: activated dendritic cell (aDC), eosinophil, immature dendritic cell (iDC), macrophage, mast cell, MDSC, monocyte, NK, NK-T, neutrophil, plasmacytoid dendritic cell (pDC).

The signatures comprise a total of 782 genes which have been identified from different studies with purified samples profiled by Affymetrix arrays.

The ssGSEA analyses in this thesis have been carried out with the java file of the GSEA software [25] (v. 3.0) using a bash script on a local machine. For each analyzed sample, the expression has been log2 transformed and not expressed genes (expression=0) have been removed. For RNA-seq data, the log2 values have been normalized by their median expression to obtain a ranked gene list for each sample. The resulting file has been used as input for the GSEA software, using 10.000 permutations and the *weighted* statistic. Resulting files have been parsed by bash scripts to obtain a tabular data file containing all enrichment scores and their adjusted p-values (FDR) for all populations.

3.1.4. xCell

xCell[20] is a deconvolution tool designed for the analysis of RNA-seq data. It uses an adaptation of the ssGSEA algorithm to calculate an enrichment score for each population in the analyzed samples. Differently from the ssGSEA, xCell transforms the original score (*raw*) to an adjusted score linearly associated to the abundance of the cell population through the use of a dedicated pipeline; also, a spillover compensation technique is applied on score calculation to reduce dependencies between closely related cell types. The algorithm requires some heterogeneity across the samples of the dataset to perform at best, therefore the result also partially depends on the dataset each sample is analyzed with. The tool is available through a web interface (http://xcell.ucsf.edu/, **Figure 2.3**) and as an

R package (https://github.com/dviraran/xCell). The analysis can be performed using different built-in signatures, which derive from different publications; however, we tested the new signatures collection published with the paper, a compendium for 64 different tissues, including immune, epithelial, and extracellular matrix cell subsets. Cell profiles for generating gene signatures have been collected by several public data sources: ENCODE [22], FANTOM5 [26], Blueprint [23], IRIS [27], and publicly available datasets from GEO. Expression profiles were generated by either microarray or RNA-seq. A total of 37 out of the 64 populations are hematopoietic tissues:

- Leukocyte lineage: B-cells, Memory B-cells, Class-switched memory B-cells, naive B-cells, pro B-cells, Plasma cells, CD4+ T-cells, CD4+ Tcm, CD4+ Tem, CD4+ memory T-cells, CD4+ naive T-cells, CD8+ T-cells, CD8+ Tcm, CD8+ Tem, CD8+ naive T-cells, Tgd cells, Th1 cells, Th2 cells, Tregs, NK cells, NKT;

- Myeloid lineage: Basophils, Eosinophils, Erythrocytes, Macrophages, Macrophages M1, Macrophages M2, Mast cells, Megakaryocytes, Monocytes, Neutrophils, Platelets, DC, aDC, cDC, iDC, pDC.

A heterogeneous list of non-immune populations or early myelopoiesis progenitors (which can be attributed with difficulty to specific cell types) has been removed from the analysis. Specifically, I filtered out the gene lists of Adipocytes, Astrocytes, CLP (common

lymphoid progenitor), CMP (common myeloid progenitor), Chondrocytes, Endothelial cells, Epithelial cells, Fibroblasts, GMP (granulocyte/macrophage progenitors), HSC (hemopoietic stem cell), Hepatocytes, Keratinocytes, MEP (megakaryocytic and erythroid progenitor), MPP (Multipotent Progenitor), MSC (Mesenchymal stem cells), Mesangial cells, Myocytes, Neurons, Osteoblast, Pericytes, Preadipocytes, Sebocytes, Skeletal muscle, Smooth muscle, ly (Ly-6C) Endothelial cells and mv (Measles virus) Endothelial cells.

In my analyses, expression data have been uploaded using the web interface of xCell as either log2 or as CPM and TPM normalized expression data, for datasets profiled by array or by RNA-seq, respectively. No parameters need to be specified before performing the analysis. Results are presented in two separated tabular files containing respectively the population enrichment scores and their p-value for each analyzed sample. The adjusted score has been used in all the analyses.

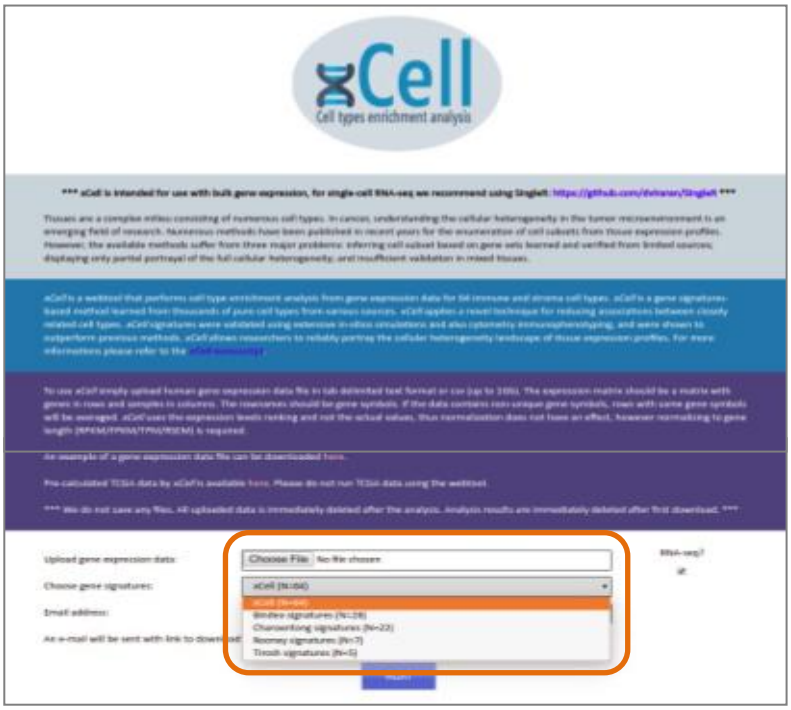


Figure 2.3. The web page of xCell. The drop-down menu showing the different gene signatures available has been highlighted. The preselected gene signature is composed of 64 cell types (1st line of the drop-down menu).

3.2. Purified immune population datasets

Several public datasets containing the profile of different purified immune populations have been used to test the performances of the various tools. Datasets have been mainly downloaded from the GEO repository. The tested GEO datasets contain multiple immunological populations profiled by array, either from healthy or pathological samples:

- GSE28490 [28] is composed by a total of 47 hematopoietic samples from nine cell subsets, CD16+CD66b+ Neutrophils, CD16-CD66b+ Eosinophils, CD14+ Monocytes, CD4+ T cells, CD8+ T cells, CD56+ NK cells, CD19+ B cells, CD123+ pDCs and CD11c+ mDCs. Part of the cells was isolated by positive selection and enrichment kits. For each cell type, from 4 to 10 biological replicates are available. The purity of the isolated cell population was assessed by FACS. Profiles have been generated using Affymetrix HG-U133Plus 2.0 microarrays;

- GSE28491 [28] is composed by 33 samples from 7 cell subsets. This is a subset contained in the same GEO SuperSeries of the previous GSE28490 dataset. As compared to the data of GSE28490, the only difference consists in the laboratory where the cells were isolated. This dataset contains the same populations of GSE28490, except for the Dendritic cells;

- GSE50008 [29] is composed by 50 samples of 4 hematopioetic subtypes, CD4+, CD8+ T cells, B cells and monocytes, profiled using Illumina HumanHT-12 V4.0 expression beadchip arrays. Interestingly, each cell subtype has been isolated with three different methods, e.g. positive or negative selection or FACS, from each of the 5 tested subjects; however, not all combinations of subject and isolation method are available;

- GSE21029 [30] is composed by purified tumor cells from 26 patients affected by CLL (Chronic Lymphocytic Leukemia); interestingly, samples from three different tissues were used as the source of cells for each subject, peripheral blood, bone marrow and Lymph Node. Samples were profiled using Affymetrix HG-U133Plus 2.0 arrays. Mononuclear cells were isolated by centrifugation over lymphocyte separation followed by CD19+ selection;

- GSE48978 [31] is a time-course experiment using purified memory T-cells activated under Th17 condition. Interestingly, the 12 samples have been profiled using both arrays (Affymetrix HT HG-U133) and RNA-seq; this dataset has been used to evaluate CIBERSORT deconvolution differences between the two types of technology for transcriptional profiling.

For all datasets profiled using Affymetrix arrays, expression data were downloaded as raw CEL files with the *GEOquery* package (v. 2.40). The raw intensity signals were extracted from CEL files and normalized using the *justRMA* function of the *affy* package (v. 1.52). Fluorescence intensities were background-adjusted and normalized using quantile normalization; log2 expression values were calculated using the median polish summarization and custom Brain Array chip definition files based on Entrez genes for Human Affymetrix arrays (v. 21.0.0; hgu133plus2hsentrezg for Plus 2 arrays and hthgu133pluspmhsentrezg for HT arrays).

GSE50008 dataset was downloaded as normalized matrices from GEO. For this experiment, the available expression matrix contains only genes with IQR>0.7 (InterQuartile Range), for a total of 4,726 genes.

For GSE48978, array data has been processed as described above for Affymetrix arrays; RNA-seq expression data has been downloaded as RPKM normalized values from the original publication (Suppl. Tab. S6 of *Zhao, Plos One, 2014*). Only common genes (n. 16,016) between arrays and RNA-seq were retained for the comparison of the deconvolution analysis.

3.3. Single-cell RNA-seq datasets

Three different datasets profiled with scRNA-seq were used to evaluate the tool's performance; these datasets are composed of samples of three different tumor types: breast cancer, lung cancer and melanoma. The breast cancer dataset, alias Breast dataset in the text, is composed by 13 samples from 4 different breast tumor subtypes, for which both bulk and single-cell expression data have been performed. The lung cancer and melanoma datasets were among the larger single-cell datasets available in public repositories at the date of my analysis, composed by 4,645 and 52,698 cells, respectively. However, for these two datasets the expression profile of bulks was not available and was reconstructed *in silico*.

Breast cancer, Lung cancer and Melanoma datasets were downloaded from GEO or ArrayExpress repositories: GSE75688 [32], E-MTAB-6149 [4], GSE72056 [21], respectively. Single-cell annotation was downloaded from the corresponding repositories, except for the lung dataset [4], which was kindly provided by the authors of the original publication. For each dataset, single cells classification was used as is provided by the authors. However, the classification for the hematopoietic cells was verified by independent immunological signatures from the Immune Response In Silico (IRIS) collection [27]; IRIS signatures were recovered from the xCell package (v. 1.1) in R, which contains a compendium of multiple hematopoietic signatures. For each single cell, the activity for each signature was calculated as the mean expression of the corresponding genes; then, cells were divided according to the single-cell annotation, e.g. B-cells, T-cells, Myeloid or Macrophages and NK, and the signature activity plotted by violin plots. Accuracy of the annotations was evaluated by visual inspection. Finally, hematopoietic fractions in single-cell data were calculated as the fraction of each cell type respect to either the total number of cells in the sample or on the hematopoietic counterpart, for EPIC or CIBERSORT, ssGSEA and xCell, respectively.

Due to the different available expression data, bulk profiles were generated differently between breast cancer and the melanoma and lung datasets. In the breast cancer dataset, the expression data of the bulk samples consists of TPM (Transcript Per Million) normalized values; then, multiple Ensembl ids matching the same gene symbol

were collapsed to the median of their expressions and genes not expressed in all samples in both the bulk and the single-cell experiment were removed. Differently, expression data in both lung and melanoma datasets consist respectively of either log2(CPM) or log2(TPM) values of each single cell. Duplicated genes were present in the melanoma dataset only; they were collapsed to the median of their expression values. Even though normalized data, we observed a high sparcity of expression matrices in both sc experiments, 18.7% and 4.9% for the melanoma and lung datasets, respectively; for this reason, we finally reconstructed the bulk expression for each sample using the sum of the normalized expression signals of its single cells instead of other statistics, e.g. mean or median. In all three datasets, we filter out samples with less than 5 hematopoietic cells according to scRNA-seq annotation, for a total of 5 and 2 samples from the Breast cancer and Melanoma datasets, respectively.

To evaluate cellular fractions, the deconvolution analyses for each tool were performed through the respective web interface and default immunological signatures, with the exception of the single sample GSEA (ssGSEA), for which we used the method and the immunological signatures as in Charoentong et al [19], as described in detail in the above section "*Tools for transcriptional deconvolution*". To correlate deconvolution results with single-cell fractions, main cell types were generated by merging the numerous cell subtypes according to their hematopoietic lineage. In detail, for tools reporting cell fractions (CIBERSORT, EPIC), the linear sum of the fractions from all subtypes was used as a measure of the main type fraction.

For tools reporting the enrichment (ssGSEA, xCell), since the score they generate corresponds to cellular activity and not to a cellular proportion, we firstly transformed the enrichment scores (ESs) to fractions. So, for each sample, we initially applied 3 preprocessing steps, using a CIBERSORT-like approach. In CIBERSORT, after the calculation of the regression coefficients with the SVM, "*negative values are set to 0 and the remaining regression coefficients are normalized to sum 1*" [17], Similarly, in our approach: (1) negative ESs are set to 0 (2) non-significant populations, e.g. with p-value ≥ 0.05, were considered with null (equal to 0) fraction, then (3) for remaining significant populations, we transformed the score to fractions by dividing each ES by the sum of the ESs within that sample, so to report the total sum to 1. Finally, as for the fraction-based tools, the linear sum of the fractions from all subtypes was used as a measure of the main cell type fraction. For xCell, the analysis was performed filtering out all non-hematopoietic tissues (27 out of 64 populations) from the beginning of the pipeline.

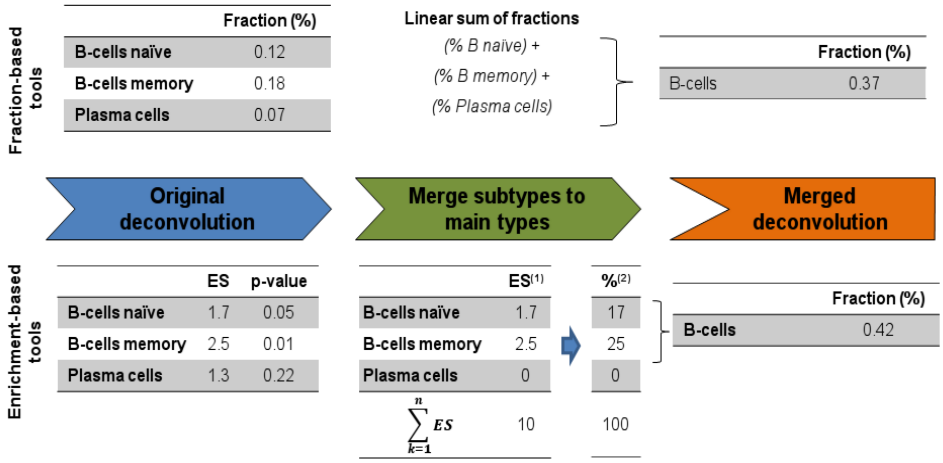


Figure 2.4. The flowchart describes the steps to merge the multiple subtypes from the deconvolution analysis (original deconvolution) into main cell types (merged deconvolution) before the comparison with the single-cell annotation. Both for ssGSEA and for xCell, we applied two intermediate steps: (1) the first step filters out non-significant populations (2) the enrichment scores are transformed into fractions.

3.4. Web application to report deconvolution tools performance

The web application has been developed using *shiny* library (v. 1.3) in Rstudio (v. 1.2). Several R packages have been used to generate the plots and the analysis within the web application: bar plots and bubble plots have been generated using the *ggplot* library. The correlation between fractions from single-cell annotation and from bulks deconvolution was calculated using the Pearson's correlation score with the *cor.test* function from the *stats* package; only correlations with p-value ≤ 0.05 were considered significant. Correlation plots have been generated with the *ggscatter* function of the *ggpubr* package.

3.5. Statistics for performance evaluation

Different measures can be used to estimate the effectiveness of bulk deconvolution respect to actual fractions [12]:

1) the most used measure is the correlation distance, which calculates the linear dependence between two numerical vectors, e.g. the estimated fractions vs. actual fractions. Mathematically, it is defined by the formula (the higher the better):

$$\rho_{p,q} = \frac{cov(p,q)}{\sigma(p)\sigma(q)}$$

where p and q are vectors with fractions from deconvolution analysis or actual fraction, respectively, whereas *cov* and *σ* correspond to covariance and standard variation of vectors, respectively. Correlation ranges from -1 to 1. In all analysis, specifically we used Pearson's correlation method;

2) a largely used measure [17] is the root mean squared distance (RMSD), which outlines the dispersion from the regression line of the estimated fraction vs. actual fraction, so it is a measure of the accuracy of the estimation (the lower the better):

$$RMSD = \sqrt{\frac{1}{p \times q}\sum_{j=1}^{p}\sum_{k=1}^{q} r_{jk}^2}$$

For cell-type k in sample j , with r_{jk}= p_{jk}- q_{jk};

3) a further measure is the mean absolute difference (mAD), which is the average difference for all cell types in all the analyzed mixture samples respect to the real fraction (the lower the better):

$$mAD = \frac{1}{p \times q} \sum_{j=1}^{p} \sum_{k=1}^{q} |r_{jk}|$$

where *r* is the fraction difference between the estimated fraction and actual fraction in the cell type *k* for sample *j*. When calculated on fraction matrices, it ranges from 0 to 1;

4) when considering multiple datasets, we implemented above summation with a further factor, the number of datasets, to have a unique value for each tool in all the analysis (the lower the better):

$$gmAD = \frac{1}{g \times p \times q} \sum_{n=1}^{g} \sum_{j=1}^{p} \sum_{k=1}^{q} |r_{jk}|$$

where g is the dataset. We named this measure throughout the text as gmAD, global mean absolute difference.

No measure overlooks the others since each one is giving different information on the difference between fractions determined by deconvolution tools and actual fraction.

3.6. Generation of a signature for murine immune populations

To generate a gene signature for the definition of immune populations in mouse, we selected gene expression profiles of purified cells from the ImmGen project and further publicly available datasets from GEO. In detail, we collected the samples from following datasets:

- GSE15907 [33]: it has been the main source for profiles of immune populations. This series comprises the samples from the ImmGen project, a large compendium of expression profiles from mouse cell populations generated in rigorously standardized conditions. Cells were isolated from different tissues and purified using different sorting protocol and mAbs depending on the subtype. From this dataset, we selected samples of the macrophages, monocytes, dendritic cells and granulocytes. Samples were profiled using the Affymetrix Mouse Gene 1.0 ST array;

- GSE35435 [34]: this dataset contains macrophages from the bone marrow, either resting or IL-4 treated to differentiate into macrophages M2. Samples were profiled using the Mouse Genome 430 2 Affymetrix array;

- GSE69607 [35]: it contains bone marrow-derived macrophages of wild-type mice (n=2-3 independent mice) treated in M0, M1 or M2 conditions. Samples were profiled using the Mouse Genome 430 2 Affymetrix array;

- GSE53321 [36]: this dataset contains cultured mouse untreated macrophages M0 and treated to differentiate into M1 or M2 macrophages. Samples were profiled using the Mouse Genome 430 2 Affymetrix array.

For all datasets, expression data were downloaded from the GEO database as raw CEL files with the *GEOquery* package (v. 2.40). The raw intensity signals were extracted from CEL files and normalized using the *justRMA* function of the *affy* package (v. 1.52). Fluorescence intensities were background-adjusted and normalized using the quantile normalization; log2 expression values were calculated using the median polish summarization and custom Brain Array chip definition files based on Entrez genes v.21.0.0 (the mogene10stmmentrez and the mouse4302mmentrezg for the ImmGen or Macrophages datasets, respectively). Expression data from the different datasets were merged into a single meta-dataset keeping only the 16,805 common genes between the two platform types. Then, expression data were corrected for batch effect using the *combat* function from the *sva* package (v. 3.22), with the GEO dataset set as batch and the cell subtype as covariate. Effects on batch correction were visually evaluated both by a PCA analysis using the *made4* package (v. 1.48) or by clustering analysis using the *heatmap3* package (v. 1.1), with the Pearson's correlation as distance measure and the average as linkage method.

The new murine reference matrix has been created through the following pseudocode:

- *detect the differentially expressed genes between each population and all other populations;*
- *order the genes by decreasing fold change (FC);*
- *remove genes enriched in non-hematopoietic tissues;*
- *combine top N marker genes in a matrix;*
- *Iterate N from 10 to 100 across all subsets to create M matrices;*
- *test the M matrices stability and retaining the most stable reference matrix.*

In detail, cellular subtypes comparison for the creation of the reference matrix was performed by Welch's t-test; p-values were adjusted by false discovery rate correction (FDR) and only significant genes were further considered (corrected p-value ≤ 0.05). Stability of the signature matrices was evaluated on the base of their relative condition number by the *kappa* function in R. The matrix with the lower condition number was retained. All analyses were performed in R (v.3.3).

The new murine gene signature was tested using samples of purified cells from several publicly available datasets. Treated samples, if available, were excluded from the test. The following datasets have been tested:

- GSE28621 [37] is composed of tissue-resident macrophages and Ly-6B+ bone marrow monocytes from the peritoneal cavity of naive mice;
- GSE339 [38] is composed of normal mouse dendritic cells from spleen and isolated by FACS. Cells were selected by positive selection of either CD4 or CD8 populations or by negative selection of both CD4 and CD8;
- E-MTAB-5012 [39] dataset is composed of monocytes and macrophages from lung tissue and isolated by flow cytometry strategy. Macrophages are isolated from both interstitial and alveolar compartments of the lung.

3.7. Generation of a signature for breast cancer subtyping

Training and validation breast cancer (BC) datasets were collected from formalin-fixed embedded tissue BC samples profiled on Illumina WG DASL HT arrays. Each sample was clinically evaluated for subtype by immunohistochemistry (IHC). Expression profiles were quantile normalized using the *normaliseIllumina* function of the *BeadArray* package in R (version 3.3). Multiple probes for the same gene were collapsed to the median of their expression.

The TCGA dataset was downloaded as log2 lowess normalized ratio of sample to reference signal (cy5/cy3) collapsed by gene with the *TCGAbiolinks* package in R. The ER, PR and Her2 status were used to define triple-negative samples.

The BC subtype genes signature (BCsig) was created using the following pseudocode:

- *detect the differentially expressed genes between each molecular subtype and all other subtypes;*
- *order the genes by decreasing fold change (FC);*
- *combine top N marker genes in a matrix;*
- *Iterate N from 3 to 200 across all subsets to create M matrices;*
- *test the M matrices stability and retaining the most stable reference matrix.*

In detail, each subtype was compared to the rest of samples by Welch's t-test; p-values were adjusted by false discovery rate correction (FDR) and only significant genes were further considered (FDR ≤ 0.1). Then, the top *n* modulated genes of each comparison were merged to create a gene signature; the number of top N genes was iterated from 3 to 200 and the most stable matrix according to the lowest Condition Number was used.

Subtypes fraction of each bulk sample was calculated using the *BCsig* for all four tools with the parameters as described in the above section "*Tools for transcriptional deconvolution*". The correlation between the TNBC fraction and pathological response vs. non-response or alive vs. dead status was investigated using boxplots and 2-sides, 2 sample Welch's t-test; assess of performance was tested by receiver-operating characteristic (ROC) curve and by the

area under the curve (AUC) with the *survcomp* and *ROCR* packages; probability of survival events was tested by Kaplan-Meier curves and log-rank test in the *survival* R package (v. 3.3).

4. Results

4.1 Deconvolution: biases and criticisms in populations detection

We initially searched for deconvolution tools specific for transcriptomic analysis, evaluating their usability in the proposed framework. To define indications and potential biases of deconvolution analysis, we tested a selection of tools using the profile of purified cells from different publicly available datasets. These datasets differentiate for profiling technology, sorting method or tissue from which samples were extracted.

4.1.1. Define deconvolution tools for a customizable framework

At the date of my survey (December 2017), 17 different algorithms were available for deconvolution on bulk transcriptional data (**Table 3.1**). To be inserted in the framework, each tool has to satisfy three different criteria: (i) its usability; (ii) the availability of a well-defined immunological signature; and (iii) the possibility to customize gene signature. Based on these requirements, I selected four tools, i.e., CIBERSORT[17], EPIC[16], ssGSEA (i.e., GSEA[25] with the single sample method from [19]), and xCell[20].

Table 3.1. Deconvolution tools as surveyed in December 2017. Sig.=Signature.

Tool	Environment			Sig.	Custom	Usability	Availability	Ref.
	R	Web	Java					
CIBERSORT	✓(*)	✓	✓(*)	✓	✓	*H*	*https://cibersort.stanford.edu/*	[17]
EPIC	✓	✓		✓	✓	*H*	*http://epic.gfellerlab.org/*	[16]
ssGSEA (TCIA)			✓	✓	✓	*H*	*(https://tcia.at/)*	[18,19]
xCell	✓	✓		✓	✓	*H*	*http://xcell.ucsf.edu/*	[20]
TIMER		✓		✓		*H*	*https://cistrome.shinyapps.io/timer/*	[47]
CellMix	✓					*M*		[52]
CellPred		✓				*M*	*http://webarraydb.org/webarray/index.html*	[48]
COD		✓				*M*	*http://csgi.tau.ac.il/CoD/*	[46]
Dsection		✓				*M*	*http://informatics.systemsbiology.net/DSection*	[43]
ImmQuant			✓		✓	*M*	*http://csgi.tau.ac.il/ImmQuant/*	[44]
DCQ		✓		✓		*L*	*https://rdrr.io/cran/ComICS/man/dcq.html*	[40]
MCP-counter	✓			✓	✓	*L*	*https://cit.ligue-cancer.net/mcp-counter/*	[51]
DeconRNASeq	✓				✓	*L*	*http://bioconductor.org/packages*	[42]
Deconf	✓					*L*	*Repsilber, 2010*	[41]
ESTIMATE		✓				*L*	*https://bioinformatics.mdanderson.org/estimate/index.html*	[49]
quantiSeq				✓	✓	*L*	*https://icbi.i-med.ac.at/software/quantiseq/doc/*	[50]
SPEC	✓					*L*	*http://clip.med.yale.edu/SPEC/*	[45]

(*) available only upon request to the authors; Custom=customizable gene signature.

The majority of tools have been discarded either because (i) they represented implementation of new algorithms without a specific gene signature (this applies to DCQ[40], deconf[41], DeconRNA-seq[42], DSection[43], ImmQuant[44] and SPEC[45]); (ii) their pipeline cannot be modified for including different gene signatures (as for COD[46] and TIMER[47]); (iii) they just provide scores for

tumor purity (as for CellPred[48] and ESTIMATE[49]); (iv) despite multiple trials, the cannot be installed in any OS (this was the case of quantiSeq[50] and MCP-counter[51]). A specific remark should be dedicated to CellMix[52]: this is an important R package which enormously simplified the deconvolution analysis by implementing all previous methods, but it is mainly a wrapper of very old frameworks and algorithms published before 2013 and, for this reason, it has been excluded.

Finally, the four selected tools can be grouped in 2 different categories according to the method they report cell types content in bulk (**Table 3.2**): tools detecting cell type fraction (CIBERSORT and EPIC), and tools detecting cell type enrichment (ssGSEA and xCell). The number of detectable leukocyte populations changes widely, from 5 to 37, thus providing different levels of detail when recapitulating the immune infiltration.

Table 3.2. Tools included in the framework and details about the input data (profiling tech) and algorithms.

Tool	Profiling tech		Immune classes (Total)	Category	Algorithm	Ref
	Array	RNA-seq				
CIBERSORT	✓	✓(*)	22	Fraction	SVM modified	[17]
EPIC	✓	✓	5 (8)	Fraction	regression	[16]
ssGSEA (TCIA)	✓	✓	28	Enrichment	GSEA	[18,19]
xCell	✓	✓	37 (64)	Enrichment	GSEA	[20]

(*) No proof of validation on RNA-seq data has been reported in the original manuscript, but the tool has been subsequently adapted also for this type of data.

4.1.2. Test deconvolution tools on human leukocyte populations

Initially, I tested the performance of the selected tools on many datasets with different populations of leukocytes derived from both healthy and pathological samples. These datasets are composed by profiles of purified cells, providing accurate inputs for the evaluation of deconvolution tools performance.

The first analysis was performed on 2 large datasets (GSE28490 and GSE28491 from Allantaz and collegues [28]) composed by leukocyte samples isolated by flow cytometry from pools of healthy donors and accounting for 9 different populations with a variable number or replicates (from 3 to 15). These are two subsets of a unique GEO SuperSeries and no differences exist in samples processing, with the exception of the laboratory they were isolated from. In general, all algorithms correctly identify enrichment of the cell subtype corresponding to the tagged cellular type. In particular, CIBERSORT detects fractions above 90% for monocyte, eosinophils, neutrophils or NK cells (**Figure S1** and **Figure S2**). Some samples are assigned across multiple subtypes of a unique parental cell, e.g., the naïve and memory subtypes of the B or T cells (**Figure 3.1**). On the contrary, dendritic cells of myeloid (mDC) or plasmacytoid (pDC) origin are not framed in the dendritic cell type: while the former contains variable fractions of monocyte (33.7%-36.7%) and, interestingly, of M2 macrophages (26.0%-34.3%), the latter are fragmented across B cells, Plasma cells and M2 macrophages (from 67.1% to 75.6% in total), and fraction of dendritic cells is almost null.

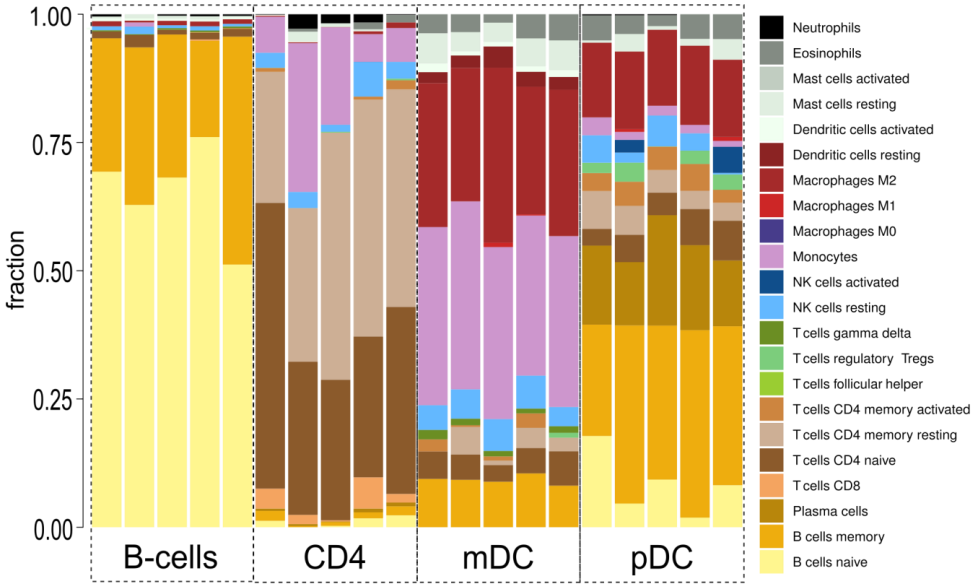


Figure 3.1. Detail of CIBERSORT analysis of the test dataset from Allantaz et al. For each sample (on column) colors correspond to a deconvoluted fraction. The complete image on all 9 cell subtypes is available in Figure S1, first row.

Deconvolution analysis of these datasets with EPIC was initially performed on log2 expression data since no specific signal distribution of the mixture sample is assumed by the tool. However, the considerable low fractions of immune cell types and the low consistency with the annotated cell types (see **Figure 3.2**, plot above) suggested a further test using linear expression values. The clear improvement in subtypes detection (**Figure 3.2**, plot below, and **Figure S1**) lead to the use of linear expression for the following analyses with this tool on array data. Then, in the analysis of the first dataset of Allantaz et al (i.e., GSE28490) [28] transformed to a linear scale, monocytes and T-cell types are correctly detected, despite CD8 T subtype is mainly recognized as CD4 cells (average

CD4=70.0%; average CD8=27.7%). B-cells, Neutrophils and overall the NK cells are detected with fractions lower than deconvolution using CIBERSORT, 61.2%, 65.6% and 39.0% on average, respectively. No signature is available for dendritic cells, which are mainly classified as Monocytes or other cells, depending on the dendritic subtype. Surprisingly, deconvolution on the second dataset from Allantaz et al (i.e., GSE28491) [28] (**Figure S2**) showed lower consistency respect to previous dataset for almost all annotated cell types despite the use of the same protocol. Only CD4 and, partially, Neutrophils subtypes are correctly detected. A feeble decreasing in population classification was noticed in CIBERSORT analysis too, suggesting some batch effect due to the laboratories; however, this effect impacts with a stronger effect on several classes of EPIC deconvolution, with on average 54.9% of the "other" component for B-cells, Eosinophils, NK and Neutrophils subtypes.

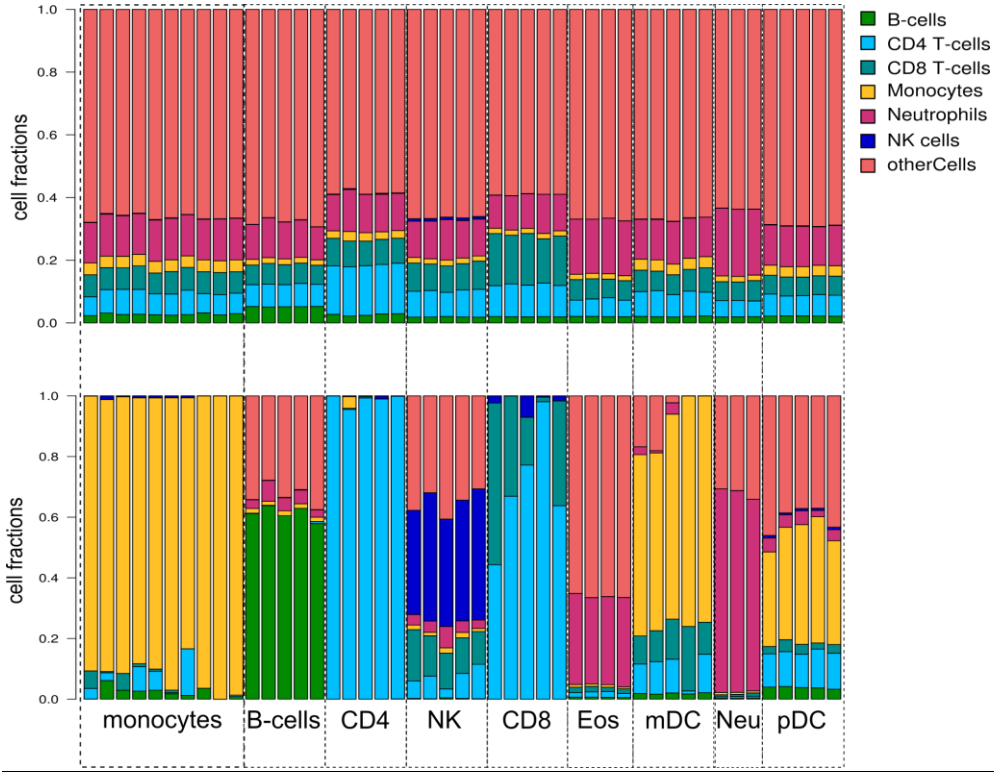


Figure 3.2. Deconvolution analysis with EPIC on either log2 (above) or linear (below) expression data on the second dataset from Allantaz et al. Eos=eosinophils; Neu=neutrophils

The analysis using ssGSEA on both datasets from Allantaz and collegues (detail in **Figure 3.3**, complete deconvolution in **Figure S1, S2**) highlighted a widespread and unspecific enrichment of several populations, with a more evident effect in the first dataset. Some populations, e.g. Central Memory CD4, MDSC, Monocyte and pDC are significantly enriched in all samples from the first dataset; for all these populations, there is a higher but modest enrichment in the tagged cell type respect to the other populations (data not shown).

This widespread and unspecific enrichment was observed also in the second dataset.

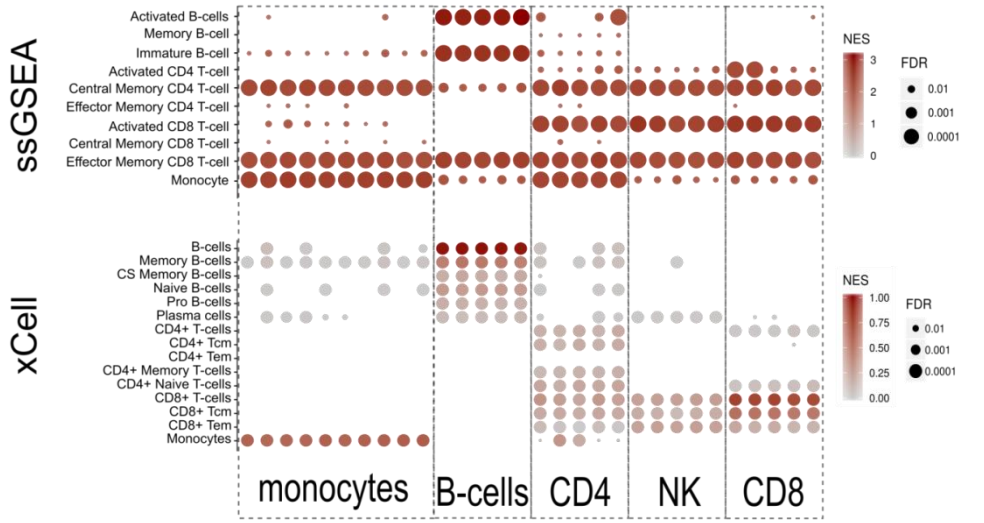


Figure 3.3. Detail of enrichment analysis using ssGSEA and xCell tools on B-cells, CD4, CD8 and monocytes populations of the first dataset from Allantaz et al. ssGSEA detects an unspecific enrichment also among not related cell types, whereas xCells deconvolution reflects cell types annotation with higher consistency.

We supposed some inability of the enrichment algorithm to correctly detect populations from these datasets, but the analysis with xCell showed more accurate classification for several cell types: B-cells CD8, Eosinophils, monocytes, pDC and mDC (classified as cDC) (**Figure S1** and **Figure S2**). Differently, NK and neutrophils populations are detected with low specificity. Of note, CD4 populations are fragmented in CD4 but also CD8 subpopulations in both datasets. **Table 3.3** summarizes the evaluation on the observed deconvolution results for all four tools in these two datasets.

Table 3.3. Synthetic evaluation of tools performance in the deconvolution analysis of multiple purified cell types from both datasets of Allantaz et al.

Dataset	Population	CIBERSORT	EPIC	ssGSEA	xCell
GSE28490	B-cells	+++	++	++	+++
	CD4 T-cells	+++	+++	++	+
	CD8 T-cells	+++	+	++	+++
	Eosinophils	+++	/	+	++
	Monocytes	+++	+++	+	+++
	Neutrophils	+++	+++	+	++
	NK	+++	+	+	++
	mDC	+	/	+	+++ (as cDC)
	pDC	+	/	++	+++
GSE28491	B-cells	+++	+	++	+++
	CD4 T-cells	+++	+++	+++	+
	CD8 T-cells	+++	++	++	+++
	Eosinophils	+++	/	+	++
	Monocytes	+++	++	++	+++
	Neutrophils	+++	++	++	++
	NK	+++	+	ns	++

Legend:
+++: the annotated cell type is recognized with high specificity
++: cell types other than the tagged type are significantly recognized
+: cell types of both myeloid and lymphoid lineage are recognized
ns: the annotated cell type is not significantly recognized.

Subsequently, we tested samples from a study from Beliakova-Bethell et al [29] which investigated the effects on expression profile for different sorting methods in four leukocyte subtypes, CD4, CD8, monocytes and B-cells isolated by FACS, positive and negative selection (**Figure S3**). The analysis with CIBERSORT on expression mixtures reported a high fraction of the correct cell type for all samples. Nonetheless, detected percentage changes depend on the combination between the sorting method and the cell type: this is particularly true for the CD4+ T cells sorted with positive selection

which are detected with lower fraction (sum of T cells CD4+ naïve+ memory: 29.3%-44.1%) respect to FACS or negative selection (61.3%-74.8%; 56.4%-65.0%, respectively). Interestingly, this dataset was profiled on the Illumina beadchip array, highlighting the ability of the algorithm to properly work on profiles from different profiling technologies. Deconvolution with EPIC shows highly variable consistency among the 4 cell types: while CD4 and monocytes are correctly recognized (on average 79.9% and 57.3%, respectively), the CD8 population are mainly fragmented into CD4 and CD8 cells, and the B-cells are poorly recognized (12.5% on average). The ssGSEA algorithm shows a pattern similar to EPIC for T-cells: while CD4 are correctly identified as activated CD4 T-cells or central memory CD4 T-cells, despite with minor enrichment of activated CD8 cells, CD8 are mainly enriched in both CD4 and CD8. B-cells are enriched with variable patterns of activated B-cells, T-cells and unexpectedly iDC, depending on the sample. Strangely, monocytes samples enrich in macrophages, MDSC, mast cell and pDC, but monocyte gene signature is not significant in all samples. xCell analysis was not performed because required the whole gene expression (see "*Purified immune population datasets*" section of Materials and Methods for detail). Interestingly, overall the positive selection of cells influences positively (for monocytes) or negatively (for CD4) deconvolution with EPIC respect to negative or FACS cells selection. A similar effect was also noticed in CIBERSORT and ssGSEA deconvolution, but only for CD4 cells (**Table 3.4**).

Finally, we tested samples enriched for specific cell populations from pathological patients: we analyzed expression data from the dataset of McCoy JP et al [30] of Chronic Lymphocytic Leukemia (CLL) which have been processed by CD19+ positive selection. CIBERSORT correctly identifies the expected clinic expansion of the B-cells, with fractions from 73.1% to 92.7% (**Figure S4**). Interestingly, the CLL dataset comprised multiple samples of the same patient from different origin: Peripheral Blood, Bone Marrow and Lymph Nodes. Even though the fraction of memory or naïve B-cells changes across patients, it seems consistent within the samples of the same subject. The different combinations of memory and naïve cells could be an interesting observation to investigate in further studies. EPIC detected a relevant fraction of B-cells (from 31.5% to 41.9%), but with a prevalence of the "other" component (50.6% on average). ssGSEA analysis is significantly enriched in activated or immature B-cells, despite a lower but constantly significant enrichment was found for several other cell types, both of B-cells or myeloid lineage. Strangely, xCell detects enrichment of populations uncorrelated with the pathology, mainly Tregs, DC or Megakaryocytes, at higher level than cells from B lymphocyte lineage (**Figure 3.4**). Then, as a first step, we investigated the enrichment of all 64 populations, noticing relevant (ES range=0.21-0.52, average ES=0.34) and significant enrichment of multipotent progenitor (MPP) cells in 14 out 24 samples, which was absent in previous healthy datasets (ES=0 for all cells except for DCs, average ES=0.02). As a further step, we verified if low enrichment was caused by poor variability across

samples: xCell requires a heterogeneous dataset to perform at best. No healthy controls were available in the CLL dataset, so we created a meta-dataset composed by samples from the dataset of McCoy et al. [30] with samples from the first dataset of Allantaz et al. [28]. The ES for B-cells strongly increases (average ES=0.895).

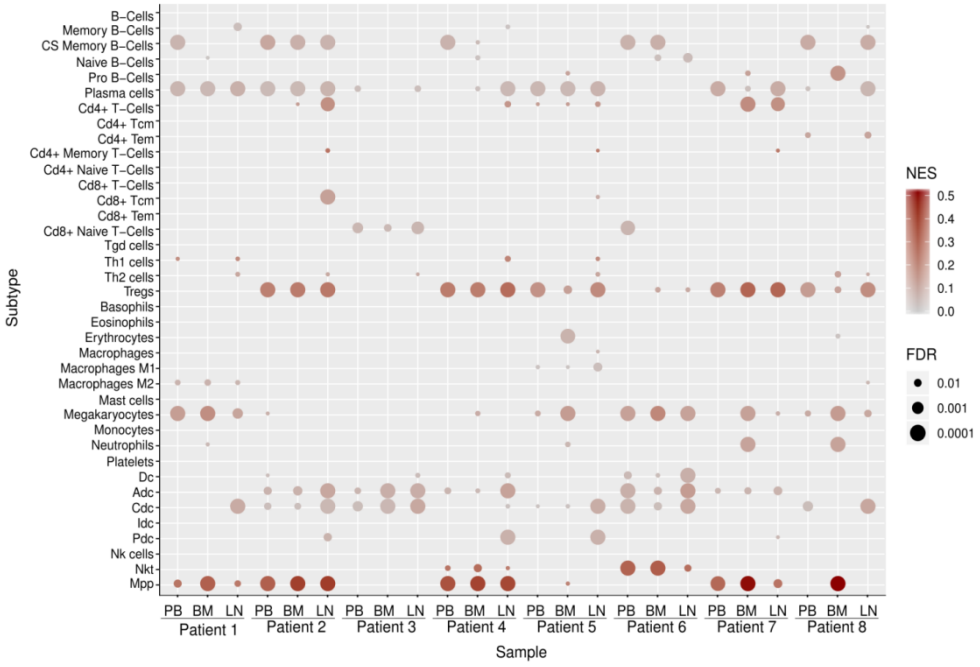


Figure 3.4. xCell deconvolution of pathological samples from CLL [30]. The MPP population is in the last row. Only samples from the first 8 patients are shown for clarity. PB; peripheral blood; BM: bone marrow; LN: lymph node.

Table 3.4 Synthetic evaluation of tools performance in deconvolution of samples from Beliakova-Bethell et al. [29] and McCoy et al [30]. For GSE21029 al. [30], we expected an enrichment of B-cells population because of the CD19+ selection in sample processing. na=not analyzed, rest of legend as in **Table 3.3**.

Dataset	Population	CIBERSORT	EPIC	ssGSEA	xCell
GSE50008	B-cells	++	+	+	na
	CD4 T-cells	++	+++	+++	na
	CD8 T-cells	++	++	++	na
	Monocytes	++	+++	ns	na
GSE21029	B-cells	+++	++	+	+

Legend:
+++: the annotated cell type is recognized with high specificity
++: cell types other than the tagged type are significantly recognized
+: cell types of both myeloid and lymphoid lineage are recognized
ns: the annotated cell type is not significantly recognized.

CIBERSORT has been built and tested on microarray data only, despite the author's state in the tool web page that it can be used also on RNA-seq data. The limiting element seems mainly the gene signature than the algorithm used for deconvolution. Indeed, the authors performed several tests on RNA-seq data, and they are "*in the process of deriving an immune signature matrix for RNA-seq*" (see CIBERSORT web app, FAQ page). However, this is still an open debate, with both papers for [24] and against [53] its application on RNA-seq data. For this reason, we searched for datasets of purified cells where the same sample is profiled by both technologies. We performed the analysis with CIBERSORT on the dataset from Zhao et al. [31], which is composed by 12 samples of memory T-cells stimulated toward activated T-cells. To reduce the possible biases related to different marker genes available by the two profiling technologies types, the analysis has been restricted to

16,016 common genes. There is a high superimposition between the deconvolution results of either the T-cells memory or T-cells activated profiled by the two platforms (Pearson's r correlations: 0.99 and 0.98, respectively) (**Figure 3.5**). After cells activation, there is a clear decrease of the memory component for the benefit of the activated population, as expected. Interestingly, fraction comprising both subtypes is higher when using the RNA-seq expression as input data, suggesting good confidence for using CIBERSORT on this profiling technology.

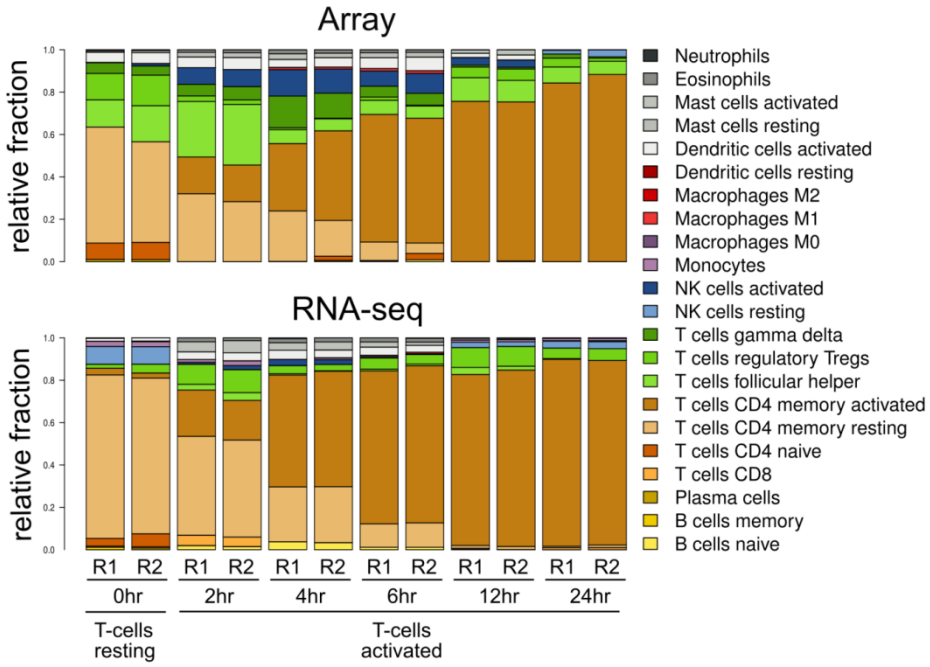


Figure 3.5. The fractions reconstructed by CIBERSORT on samples profiled both by Array and RNA-seq of the Zhao et al. dataset. As expected, there is a clear decreasing portion of memory resting cells (light brown color) for the increasing of activated cells (brown color).

4.1.3. Single-cell RNA-seq as a gold standard to test tools performance

Single-cell technology is currently the gold standard to identify the heterogeneity of populations in bulk samples. For this reason, we took advantage of three different single-cell RNA-seq datasets to evaluate the ability of deconvolution tools in estimating both cell types and cell abundances. These results have been finally aggregated in an interactive web application named ARDESIA.

We initially searched for datasets to be tested, focusing on experiments with matched bulk and single-cell RNA-seq data. Despite an extensive search in public resources as GEO [54], arrayExpress[55], single-cell Expression Atlas[56], scRNASeqDB [57] and SRA [58], we could find only one dataset with matched human bulk and single-cell data, i.e. the study on breast cancer from Chung and collaborators [32]. The restriction in the available datasets is due to the several required constraints for this type of validation: (i) matched bulk and single-cell gene expression from the same sample; (ii) human organism; (iii) no filter for the immune component during sample processing; (iv) a minimal presence of leukocytes infiltration; and (v) availability of both bulk and single-cell expression data.

To cope with this lack of data, we opted to reconstruct synthetic bulks from two of the largest scRNA-seq public cancer datasets, both comprising 19 or 17 samples and 52,698 or 4,645 cells from lung cancer [4] or melanoma [21]), respectively. Importantly, the

melanoma dataset was already analyzed in EPIC manuscript, even though authors reported only an aggregated correlation between bulk deconvolution and single-cell data, considering all available populations at once (i.e., B-cells, CD4, CD8, macrophages, NK, melanoma cells and cancer-associated fibroblast; , Figure 3C of [16]). Nevertheless, we opted to include this dataset in our framework to obtain consistency evaluation at single population level and to by-pass the shortage of single-cell datasets with suitable features for our pipeline.

In all three works, the subset of cells from the immune system was classified into 4 major leukocyte types at most: B-cells, T-cells, Myeloid or Macrophages, and NK (**Table 3.1**). Despite the available annotations, we opted for verifying the transcriptional profiles by independent immunological human signatures from the IRIS collection [27] (**Figure 3.6**, panel **A**), mainly confirming the classifications proposed by the authors. In general, the immune cell types within each dataset covered different immunological admixtures, e.g., from the less abundant NK cells to a high fraction of T-cells, further to a variable presence of B-cells and Myeloid or Macrophages components (**Table 3.5** and **Figure 3.6,** panel **B**).

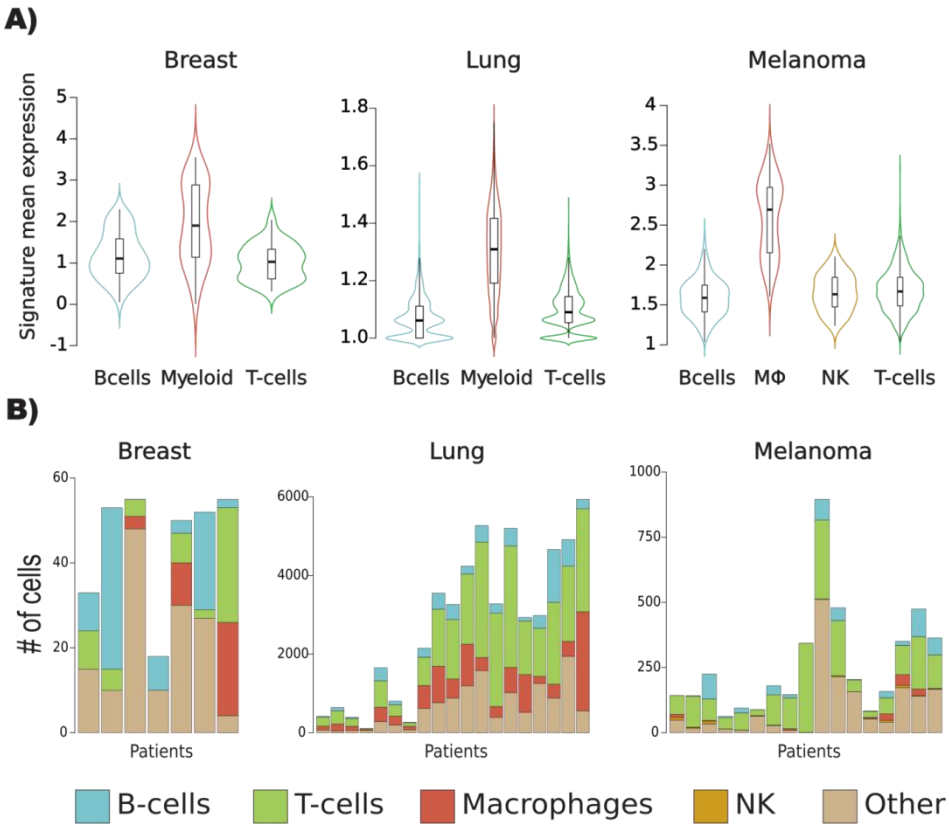


Figure 3.6. **A)** The agreement between the annotated cell type and the transcriptional profile was verified by visual inspection using independent immunological signatures. In this example, the expression of the gene signature for Macrophages is higher in either Myeloid or Macrophages respect to cells from the leukocyte lineage. For each plot, columns correspond to different annotated cell types, while on the secondary axis the expression levels of the signature are shown. The analysis was performed for all combinations of annotated cell types with IRIS signatures, for a total of 39 plots, thus, for clarity, only an example using the Macrophages signature on each dataset is shown; **B)** The immunological content of the three datasets used in the analysis according to the original annotation of the papers. Each column corresponds to a different sample. On the upper row, the total number of cells, whereas on the lower row their corresponding immunological fractions are shown. The higher component is the T-cells, in green, whereas few cells of NKs, in orange, exist in the Melanoma dataset only. “Other”: either the tumoral or non-annotated cells.

Table 3.5. The hematopoietic cell types and their fractions according to the annotations from the papers' authors. Fractions are calculated either as percentages within the leukocyte component only (Leukocyte) or on the total number of cells of the dataset (Total), thus including tumoral, stroma or non-annotated cells. *na*=annotation not used in the dataset.

	Breast		**Lung**		**Melanoma**	
	Leukocyte (%)	**Total (%)**	**Leukocyte (%)**	**Total (%)**	**Leukocyte (%)**	**Total (%)**
B-cells	48.3	26.2	13.9	10.6	18.6	11.1
Macrophages	*na*	*na*	*na*	*na*	4.6	2.7
Myeloid	20.3	11.1	24.2	18.5	*na*	*na*
NK	*na*	*na*	*na*	*na*	1.9	1.1
T-cells	31.4	17.1	61.9	47.3	74.9	44.5
TOT	100	54.4	100	76.4	100	59.4

Then, we performed the deconvolution analysis on the transcriptional profile of both real and *in silico* bulks. The quantification obtained by this analysis can be now compared with the true positive, e.g. the fractions from the single-cell annotation. However, the straightforward use of deconvolution results was puzzling due to two different problems: (i) the comparison between the deconvolution and scRNA-seq quantifications was not immediate, due to the high parceling out of cell types in the deconvolution analysis respect to the less detailed leukocytes classification of the single cell, and (ii) the difficulties to compare the results of the fraction-based and enrichment-based tools, reporting two different biological information, a quantification or an activity, respectively. For these reasons, we carried out two consequential steps: (i) merge the multiple cell subtypes from deconvolution analysis into main cell types, named "Merged deconvolution"; and (ii) for the enrichment-based tools, transform the enrichment scores to fractions, using a CIBERSORT-like approach (see M&M for details). By these steps, we finally

performed the correlation between the bulk fractions defined by deconvolution and the single-cell fractions. All steps described above are summarized in **Figure 3.7**. **Table 3.6**. specifies into which main types the different subtypes have been merged; this merge changes depending on the populations detectable by the deconvolution tool and the cell types annotated in each dataset.

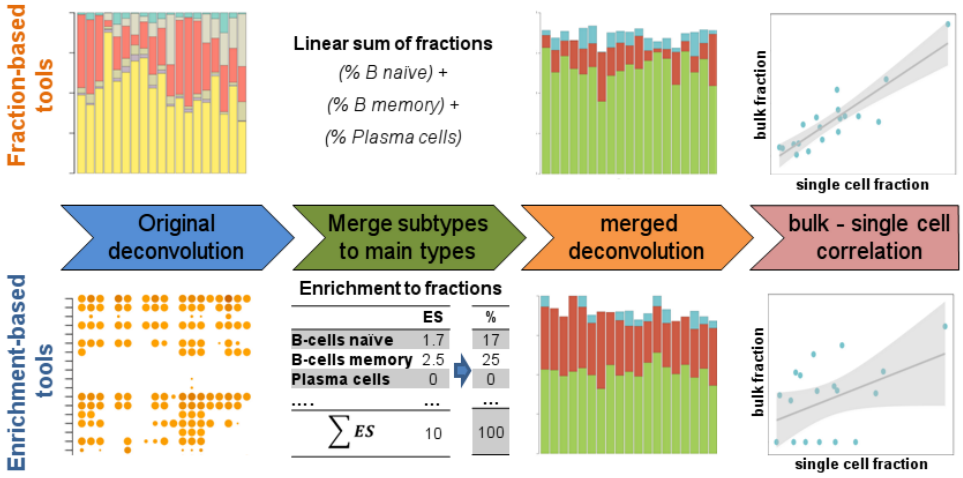


Figure 3.7. Pipeline resuming the steps from the deconvolution data to the final bulk-sc correlation. Above row: for fraction-based tools, e.g. CIBERSORT and EPIC, the linear aggregation (e.g. sum) of the different sub-populations has been applied before comparing with the sc data, while for the enrichment-based tools a more complex approach has been generated.

Table 3.6. The cell subtypes from the deconvolution analysis have been merged into the main subtypes to be compared with the single-cell annotation. For each sub-table, the first column shows the populations from the tool, while the other columns report the cell type each population is merged according to dataset annotation. For xCell, only immune component is shown. *na*=not used in the merged deconvolution calculation.

CIBERSORT Subtype	**Breast**	**Melanoma**	**Lung**
B cells memory	B-cells	B-cells	B-cells
B cells naïve	B-cells	B-cells	B-cells
Dendritic cells activated	*na*	*na*	*na*
Dendritic cells resting	*na*	*na*	*na*
Eosinophils	Myeloid lineage	*na*	Myeloid lineage
Macrophages M0	Myeloid lineage	Macrophages	Myeloid lineage
Macrophages M1	Myeloid lineage	Macrophages	Myeloid lineage
Macrophages M2	Myeloid lineage	Macrophages	Myeloid lineage
Mast cells activated	Myeloid lineage	*na*	Myeloid lineage
Mast cells resting	Myeloid lineage	*na*	Myeloid lineage
Monocytes	Myeloid lineage	*na*	Myeloid lineage
Neutrophils	Myeloid lineage	*na*	Myeloid lineage
NK cells activated	*na*	NK cells	*na*
NK cells resting	*na*	NK cells	*na*
Plasma cells	B-cells	B-cells	B-cells
T cells CD4 memory activated	T-cells	T-cells	T-cells
T cells CD4 memory resting	T-cells	T-cells	T-cells
T cells CD4 naive	T-cells	T-cells	T-cells
T cells CD8	T-cells	T-cells	T-cells
T cells follicular helper	T-cells	T-cells	T-cells
T cells gamma delta	T-cells	T-cells	T-cells
T cells regulatory Tregs	T-cells	T-cells	T-cells

EPIC Subtype	**Breast**	**Melanoma**	**Lung**
B cells	B-cells	B-cells	B-cells
CD4 Tcells	T-cells	T-cells	T-cells
CD8 Tcells	T-cells	T-cells	T-cells
Monocytes	Myeloid lineage	Macrophages	Myeloid lineage
Neutrophils	Myeloid lineage	*na*	Myeloid lineage
NK cells	*na*	NK cells	*na*
otherCells	*na*	*na*	*na*

ssGSEA Subtype	**Breast**	**Melanoma**	**Lung**
Activated B Cells	B-cells	B-cells	B-cells
Activated Cd4 T Cell	T-cells	T-cells	T-cells
Activated Cd8 T Cell	T-cells	T-cells	T-cells
Activated Dendritic Cell	Myeloid lineage	*na*	Myeloid lineage
Cd56Bright Natural Killer Cell	*na*	NK cells	*na*
Cd56Dim Natural Killer Cell	*na*	NK cells	*na*

Central Memory Cd4 T Cell	T-cells	T-cells	T-cells
Central Memory Cd8 T Cell	T-cells	T-cells	T-cells
Effector Memory Cd4 T Cell	T-cells	T-cells	T-cells
Effector Memory Cd8 T Cell	T-cells	T-cells	T-cells
Eosinophil	Myeloid lineage	*na*	Myeloid lineage
Gamma Delta T Cell	T-cells	T-cells	T-cells
Immature B Cell	B-cells	B-cells	B-cells
Immature Dendritic Cell	Myeloid lineage	*na*	Myeloid lineage
Macrophage	Myeloid lineage	Macrophages	Myeloid lineage
Mast Cell	Myeloid lineage	*na*	Myeloid lineage
Mdsc	Myeloid lineage	*na*	Myeloid lineage
Memory B Cell	B-cells	B-cells	B-cells
Monocyte	Myeloid lineage	*na*	Myeloid lineage
Natural Killer Cell	*na*	NK cells	*na*
Natural Killer T Cell	*na*	NK cells	*na*
Neutrophil	Myeloid lineage	*na*	Myeloid lineage
Plasmacytoid Dendritic Cell	Myeloid lineage	*na*	Myeloid lineage
Regulatory T Cell	T-cells	T-cells	T-cells
T Follicular Helper Cell	T-cells	T-cells	T-cells
Type 17 T Helper Cell	T-cells	T-cells	T-cells
Type 1 T Helper Cell	T-cells	T-cells	T-cells
Type 2 T Helper Cell	T-cells	T-cells	T-cells

xCell Subtype	**Breast**	**Melanoma**	**Lung**
aDC	Myeloid lineage	*na*	Myeloid lineage
Basophils	Myeloid lineage	*na*	Myeloid lineage
B-cells	B-cells	B-cells	B-cells
CD4+ memory T-cells	T-cells	T-cells	T-cells
CD4+ naive T-cells	T-cells	T-cells	T-cells
CD4+ T-cells	T-cells	T-cells	T-cells
CD4+ Tcm	T-cells	T-cells	T-cells
CD4+ Tem	T-cells	T-cells	T-cells
CD8+ naive T-cells	T-cells	T-cells	T-cells
CD8+ T-cells	T-cells	T-cells	T-cells
CD8+ Tcm	T-cells	T-cells	T-cells
CD8+ Tem	T-cells	T-cells	T-cells
cDC	Myeloid lineage	*na*	Myeloid lineage
Class-switched memory B-cells	B-cells	B-cells	B-cells
DC	Myeloid lineage	*na*	Myeloid lineage
Eosinophils	Myeloid lineage	*na*	Myeloid lineage
Erythrocytes	Myeloid lineage	*na*	Myeloid lineage
iDC	Myeloid lineage	*na*	Myeloid lineage
Macrophages	Myeloid lineage	Macrophages	Myeloid lineage
Macrophages M1	Myeloid lineage	Macrophages	Myeloid lineage
Macrophages M2	Myeloid lineage	Macrophages	Myeloid lineage
Mast cells	Myeloid lineage	*na*	Myeloid lineage
Megakaryocytes	Myeloid lineage	*na*	Myeloid lineage

Memory B-cells	B-cells	B-cells	B-cells
Monocytes	Myeloid lineage	*na*	Myeloid lineage
naive B-cells	B-cells	B-cells	B-cells
Neutrophils	Myeloid lineage	*na*	Myeloid lineage
NK cells	*na*	NK cells	*na*
NKT	*na*	NK cells	*na*
pDC	Myeloid lineage	*na*	Myeloid lineage
Plasma cells	B-cells	B-cells	B-cells
Platelets	Myeloid lineage	*na*	Myeloid lineage
pro B-cells	B-cells	B-cells	B-cells
Tgd cells	T-cells	T-cells	T-cells
Th1 cells	T-cells	T-cells	T-cells
Th2 cells	T-cells	T-cells	T-cells
Tregs	T-cells	T-cells	T-cells

Then, we performed bulk-sc correlations (**Table 3.7**). Despite all correlation coefficients are all positives, there are considerable variations, depending on the dataset and the tool evaluated. The most evident result is the difficulty in assessing several cell type fractions from the unique real bulk dataset. Indeed, in the Breast dataset, few correlations are significant, though this outcome is at least partially caused by the lower number of samples as compared to the *in-silico* bulk datasets (**Figure 3.8A**). However, EPIC and ssGSEA can't reconstruct any of the cellular fractions. For this reason, we verified if this difficulty in quantifying cell types could be related to significant changes between bulk and sc sample composition. The correlation analysis of bulk and single-cell pairwise transcriptional profiles (average r=0.680, **Figure 3.8B**) indicated that the low correlations could be only partially ascribed to compositional differences. Instead, the poor correlation of the Myeloid component in 3 out of 4 tools could be due to two different reasons: (i) the Myeloid class includes a wide repertoire of transcriptionally different

populations, and (ii) deconvolution tools are assigning anyhow some fractions or enrichment to absent populations, as is for half of the samples of this dataset (see **Figure S.5**).

Table 3.7. Bulk-sc correlation between fractions from sc annotations and fractions from the deconvolution analysis.

		CIBERSORT		EPIC		ssGSEA		xCell	
Dataset	**Cell Line**	**r**	**p**	**r**	**p**	**r**	**p**	**r**	**p**
Breast	B cells	0.713	0.072	0.568	0.183	0.544	0.206	0.732	0.061
	T cells	0.806	0.028	0.523	0.228	0.200	0.666	0.552	0.198
	Myeloid	0.210	0.650	0.601	0.154	0.302	0.510	0.776	0.040
Lung	B cells	0.840	<0.001	0.887	<0.001	0.432	0.064	0.874	<0.001
	T cells	0.681	0.001	0.598	0.007	0.678	0.001	0.908	<0.001
	Myeloid	0.826	<0.001	0.643	0.003	0.737	<0.001	0.868	<0.001
Melanoma	B cells	0.934	<0.001	0.947	<0.001	0.777	0.018	0.975	<0.001
	T cells	0.914	<0.001	0.141	0.588	0.255	0.322	0.854	<0.001
	MΦ	0.827	<0.001	0.504	0.039	na	na	0.771	<0.001
	NK	0.436	0.078	0.676	0.003	0.297	0.245	0.237	0.359
Average correlation		0.719		*0.609*		0.503		0.755	

Legend: r=Pearson's correlation coefficients, p=p-value of Pearson's correlation. na=r calculation not feasible. MΦ=Macrophages.

On the contrary, in the *in-silico* bulk datasets, there is a high agreement between the sc composition and the quantification from the deconvolution tools, where almost all correlations are significant. In the Lung dataset, all populations are correctly quantified by all tools, whereas in the Melanoma dataset we can observe high correlations when using CIBERSORT or xCell and a more heterogeneous result for EPIC or ssGSEA tools. Curiously, EPIC deconvolution correlates very poorly in the T-cell population of the Melanoma dataset, while the same population is significantly detected in the Lung dataset. To investigate the possible causes of this bizarre result, we repeated the whole pipeline for EPIC in the

Melanoma dataset using the “tumour-infiltrating cells” signature available in the tool. The correlation significantly increases (r=0.491, p=0.0327), but to a small extent as compared to CIBERSORT or xCell, suggesting that both the gene signature and the tool could affect to the low consistency in the detection of T-cells from this dataset. Interestingly, the less abundant population of NK cells is wrongly detected by all tools, EPIC included (this correlation can be considered as a false positive, according to **Figure 3.8C**).

All bulk-sc correlation plots and statistics are available at the ARDESIA web application in the “Bulk-Single Cell Correlation” page, either in the “Across Datasets” or in the “Across tools” drop-down menu, where the user can respectively visualize the deconvolution results of one tool across different datasets or across different tools in a single dataset.

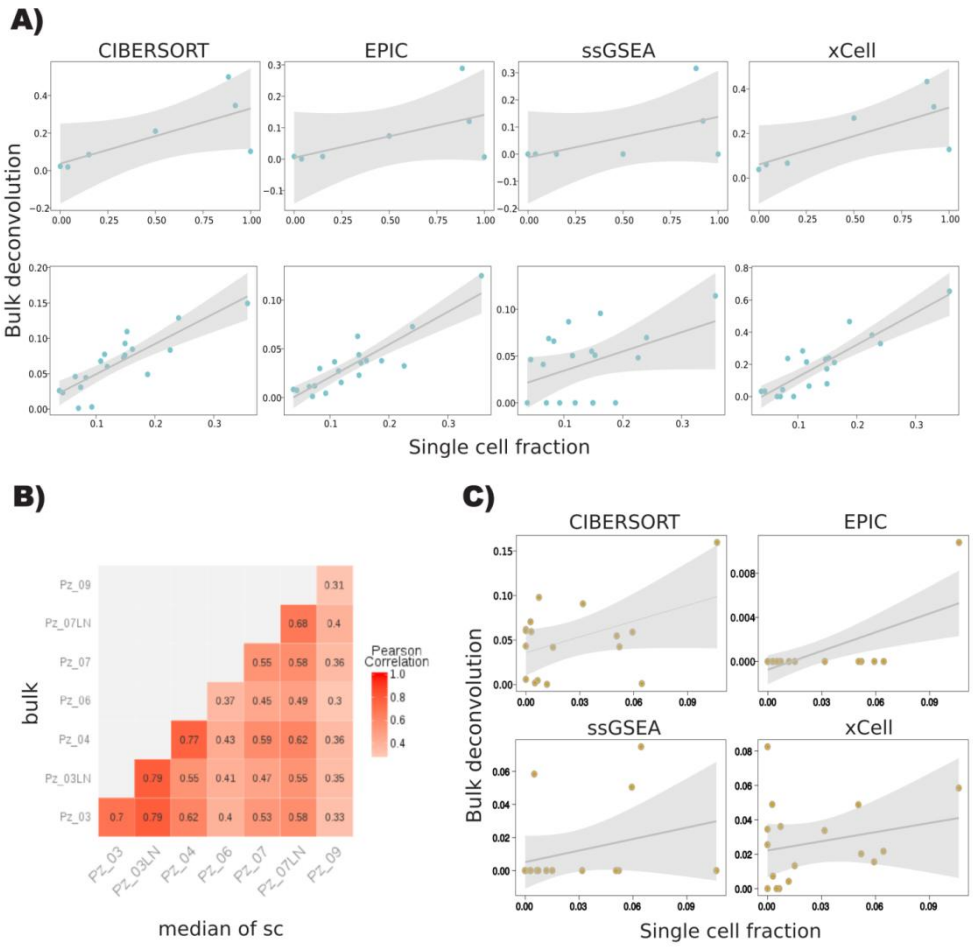
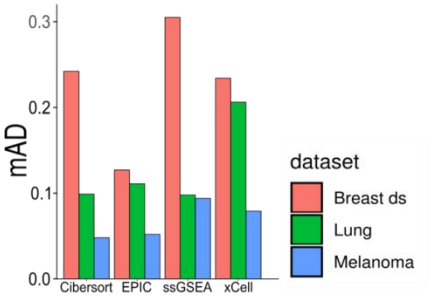


Figure 3.8. **A)** bulk-sc correlation for B-cells from the Breast (first row) or lung dataset; **B)** for the Breast dataset, the heatmap shows the correlation between the transcriptional profiles of bulk vs. *in silico* sc samples, calculated for each gene as the median expression among all sc of the sample. The diagonal corresponds to matched bulk-sc samples from same patient (mean r=0.680±0.19 SD); **C)** the bulk-sc correlation on NK cells from the Melanoma dataset: the deconvolution analysis is not able to correctly detect the low fractions of these cells, also for EPIC (top-right box), for which the significant correlation should be considered as a false positive.

A)

mAD	Cibersort	EPIC	ssGSEA	xCell	gmAD
Breast	0.242	0.127	0.305	0.234	0.227
Lung	0.099	0.111	0.098	0.206	0.129
Melanoma	0.048	0.052	0.094	0.079	0.069
gmAD	0.13	0.097	0.166	0.173	0.141



B)

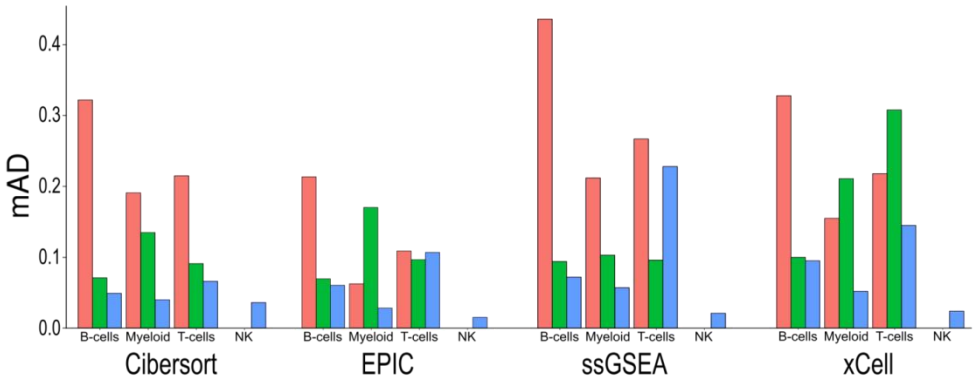


Figure 3.9. mAD estimation between bulk and single-cell fractions calculated **A**) for each dataset or **B**) singularly for each subtype.

Finally, to assess absolute discrepancies between the actual fractions from a single cell and the estimated fractions, we calculated mAD (mean absolute deviation) and gmAD (global mean absolute deviation) estimates for all tools (**Figure 3.9**). In the Breast cancer dataset, we observed a 0.227 gmAD, that is 22.7% of difference between the sc fractions and fractions detected by deconvolution (**Figure 3.9A**). The high gmAD observed in the Breast cancer dataset was expected considering the low Pearson's correlations observed for all tools. However, high mADs were observed also in the analysis of the Lung *in silico* dataset. To better investigate this

result, we calculated the single mAD for each cell type in all three datasets (**Figure 3.9B**). Interestingly, the higher error in fractions estimation in the Breast dataset is for B-cells, which are constantly underestimated (**Figure S.5**). Indeed, there is an estimated fraction of 49.9% of B-cells at most by all tools, while the immune population is almost exclusively composed by this cell type in 3 out of 7 samples (**Figure 3.6B**). In the Lung dataset, xCell showed a mAD of 0.206: this result is caused by the underestimation of T-cells for the benefit of an overestimation in the myeloid counterpart. Finally, deconvolution of the Lung dataset showed high consistency with the sc annotation, but with low performance on T-cell quantification by both enrichment tools.

In conclusion, the use of single-cell data suggested that no tool significantly correlates with all cell types in all datasets. In general, there is higher consistency between the single-cell data and bulk deconvolution when using *in silico* than real bulk transcriptional signals, despite the breast cancer dataset is composed of a considerably lower number of samples. The less abundant population, the NK cells, is poorly detected: however, this limitation is expected when considering the almost null frequency of these cells (1.1% mean on total cells 1.1%, range 0%-6.9%, **Table 3.5**), which is near to or lower than the detection limit declared by several tools (e.g. 0.5% for CIBERSORT). Indeed, the macrophages in the lung dataset account for about the 4% of the immune cells and less than 3% of total cells, suggesting that fractions slightly higher than 1% can

be correctly detected by all tools except the ssGSEA. The analysis by mAD highlighted a better performance for the tools based on fractions determination, where, with the exclusion of the breast dataset, there is a high consistency of the proportions estimated by deconvolution.

4.1.4. ARDESIA: a web app for Automatic Report of DEconvolution tools by SIngle cell Annotation

To better illustrate all the analyses of the bulk-sc correlations, we created a web application, ARDESIA (Automatic Report of DEconvolution tools by SIngle cell Annotation) (**Figure 3.10**), using the features of the *Shiny* package; this package allows the build and the host standalone of interactive web apps straight from R. The app is available at:

https://bicciatolab.shinyapps.io/ardesia/

The web application consists of: (i) the home page, briefly describing the analyzed tools, datasets and analyses reported in the app; (ii) the "Dataset display" page, showing the immunological content within each single cell dataset; (iii) the "Deconvolution results", where results of all deconvolution analyses for each tool are shown; (iv) the "Bulk-single cell correlation", where the app reports statistics and plots of bulk-sc correlations; and (v) the "More" page, with the detail of the Materials and Methods used in the analyses and a dedicated FAQs (Frequently Asked Questions) section.

Figure 3.10. The home page of ARDESIA, describing objectives of the application and content of each page. The upper bar allows the navigation across the pages of datasets summary and analysis. In the “More” drop-down menu, pages containing both FAQs and materials & methods are available to the user.

4.1.5. Conclusions on performance evaluation of deconvolution tools

Analyses of profiles from purified cells for detection of the subpopulation from bulk profiles returned highly variable results among the used tools. In summary, we can highlight some general indications and drawbacks:

- CIBERSORT correctly classifies almost all populations from either physiological or pathological conditions with the exception of dendritic cells, even if this population originate from one single dataset. Furthermore, although designed and tested only on array data, it is also able to work with RNA-seq data, at least for the tested populations;
- EPIC, of note, is influenced by the distribution of expression data, so use of unlogged expression data is recommended; furthermore, B-cells populations are usually poorly detected and CD4 often fragmented in CD4 and CD8 cell subtypes;
- ssGSEA displays a general and unspecific enrichment in several and often uncorrelated cell types, even from both myeloid and lymphoid lineage simultaneously, for several purified cell types from different datasets;
- xCell performs correctly in several cell types from different conditions. Modifications performed on the distribution of the enrichment score improved the interpretation of populations in the mixtures respect to classical ES from ssGSEA. however, heterogeneity of the dataset is a constraint for its use, as showed for the CLL dataset.

As a final note, the p-value for each single population reported by enrichment algorithms is a practical instrument to outline populations in the mixture; on the contrary, assessment of abundance significance in fraction-based tools can be user-dependent. However, ES interpretation is more complex when several p-values related to either similar or different populations within the same samples are significant.

Deconvolution analyses of bulks from single-cell RNA-seq provided further indications in estimating both different cell types and different cell abundances:

- CIBERSORT fractions significantly correlate with almost all populations from scRNA-seq datasets: it also detects with high confidence two out of three populations from the real bulk dataset, the B-cells and T-cells. The mAD estimate indicates a high accuracy in detecting fraction from both *in silico* datasets, where CIBERSORT has the lower error;
- EPIC deconvolution generally correlates with sc fractions of *in silico* datasets; however, correlation scores are lower respect to CIBERSORT or Xcell tools and with some incorrect quantification of highly present populations, the T-cells of the Melanoma dataset. On the contrary, the gmAD estimate is the lower across tools;
- ssGSEA solves fractions for fewer populations than rest of tools; the low abundance population of macrophages from

Melanoma dataset is not detected in all samples at all; the mAD estimate indicates good accuracy in the Lung dataset, whereas the low mAD in the Melanoma dataset is affected by several not significant correlations.

- xCell reconstructs with high correlation fractions from both *in silico* sc datasets; moreover, there is high accuracy in detecting two out of three populations of the Breast dataset, as is the case of CIBERSORT. However, its gmAD is the highest, mostly for the incorrect estimate of myeloid and T-cells populations in the lung dataset.

Considering performances of tools in detecting both purified populations and fractions from single-cell data, in the second part of my thesis project, finally opted for the use of CIBERSORT to generate a custom framework able to detect heterogeneity in bulk transcriptional data other than immune populations. This purpose required multiple enforces, because CIBERSORT, as all the other deconvolution tools, has been specifically designed to detect distinct leucocyte populations by the use of a specific immunological gene signatures. Initially, I changed both the gene signature and the organism for which the tool has been created for, i.e., to detect immune populations in mouse, and I tested the tool responses afterwards. Specifically, we were interested to what extent deconvolution analysis can perform fine-grained tasks e.g., discriminate very similar populations with close transcriptional

profiles. Subsequently, I reapplied the pipeline just defined in mouse to assess molecular subtype heterogeneity in breast tumor bulks. This analysis addresses two main interrogations: if deconvolution tools can be adapted to a different context than immune populations and if the defined transcriptional heterogeneity can be associated to a specific clinical outcome in breast tumors.

4.2. Generation of a bioinformatics framework to identify cell subpopulations from bulk transcriptional data

4.2.1. Creation of a murine gene signature for immune heterogeneity

To evaluate to which extent the deconvolution framework could be modified, we changed the organism for which deconvolution has been designed for, switching the detection of immune subpopulations from human to mouse.

The mouse is still the leading model organism to study human diseases. In the context of the immune system, the Immunological Genome Project (ImmGen) represents the complete compendium of genome-wide data containing expression of protein-coding genes for all defined cell populations of the mouse immune system, comprising the expression profiles of more than 200 cell populations generated in rigorously standardized conditions [59]. These data could, therefore, be a perfect base to create a reference profile to monitor the presence of any subpopulation of immune cells in mouse. To build a new signature matrix for the deconvolution analysis, I created a large murine meta-dataset composed by a wide repertoire of murine immunological samples. Mouse leukocytes expression has been mainly collected from the ImmGen project and, for missing populations as e.g. alternative macrophages, from other publicly

available data. Since the datasets originate from different experiments, we paid particular attention to minimize the presence of batch effects.

Table 3.9. Detail of the number of samples per myeloid subtype per tissue of the meta-dataset.

Tissue	Dendritic cells	Granulocytes	Monocytes	MΦ	MΦ M1	MΦ M2
Blood	-	3	14	-	-	-
Bone Marrow	-	7	6	12	6	8
Brain	-	-	-	3	-	-
Iliac, axillary, inguinal LN	12	-	-	-	-	-
Kidney	3	-	-	-	-	-
Liver	2	-	-	-	-	-
Lung	7	-	-	8	-	-
Mesenteric LN	14	-	3	-	-	-
Peritoneal Cavity	-	6	-	18	-	-
Salmonella	-	-	-	8	-	-
Skin	2	-	-	-	-	-
Skin LN	3	-	-	6	-	-
Small Intestine	-	-	-	12	-	-
Spleen	22	-	-	3	-	-
Subcutaneous LN	12	-	-	-	-	-
Synovial Fluid	-	3	-	-	-	-
Thymus	6	-	-	-	-	-

Legend. LN: lymph nodes; Salmonella: Macrophages from peritoneal cavity after salmonella infection. MΦ=Macrophages

We collected expression data of mouse dendritic cells (DC), monocytes (MO), macrophages (MG) and granulocytes (GN) from various tissues of the ImmGen project (GSE15907 [33]). Despite a large number of available cell types, no classically (M1) or alternatively activate (M2) macrophages subtypes were available. Because of their emerging role in several cancers [10], we recovered their gene expression data from 3 further studies of resting and

activated macrophages [34–36]. Overall, the resulting meta-dataset was composed by 6 main leukocyte subtypes, DC, GN, MF, M1, M2, MO, from 17 different tissues, for a total of 209 samples (see **Table 3.9** for details); because not all the 6 cell types from all tissues were available, the final dataset consisted in 27 different leukocyte subtypes.

Initially, the expression profiles were normalized singularly for each dataset. We performed an exhaustive quality control for potential biases by using different types of analysis (e.g. pseudo images, expression boxplots, degradation plots) and found no specific defects in any single array. Then, we created a unique meta-dataset using the 16,804 common genes shared by the different Affymetrix array types. The following quality control analysis by PCA highlighted a clear batch effect with two different main clusters corresponding to the GSE/platform version, which was confirmed by unsupervised clustering using most variable genes (**Figure 3.11**). To reduce this batch effect, we tested different combinations of batch removal using the *sva* algorithm in R and the available meta-information (GSE, platform type, cell type, tissue, weeks of age). We evaluated the effects of the different combinations by unsupervised cluster analysis using 10% of the most modulated genes, e.g. with higher CV (Coefficient of Variation). While no correction perfectly clustered all samples of the same leukocyte subtype, some combinations, e.g. those considering the type of array, were excluded from subsequent analysis because fragmented M1 and M2 populations in multiple clusters (data not shown). The final batch correction used to create

the meta-dataset was set using the GSE as batch and subtypes as covariate (**Figure 3.12**).

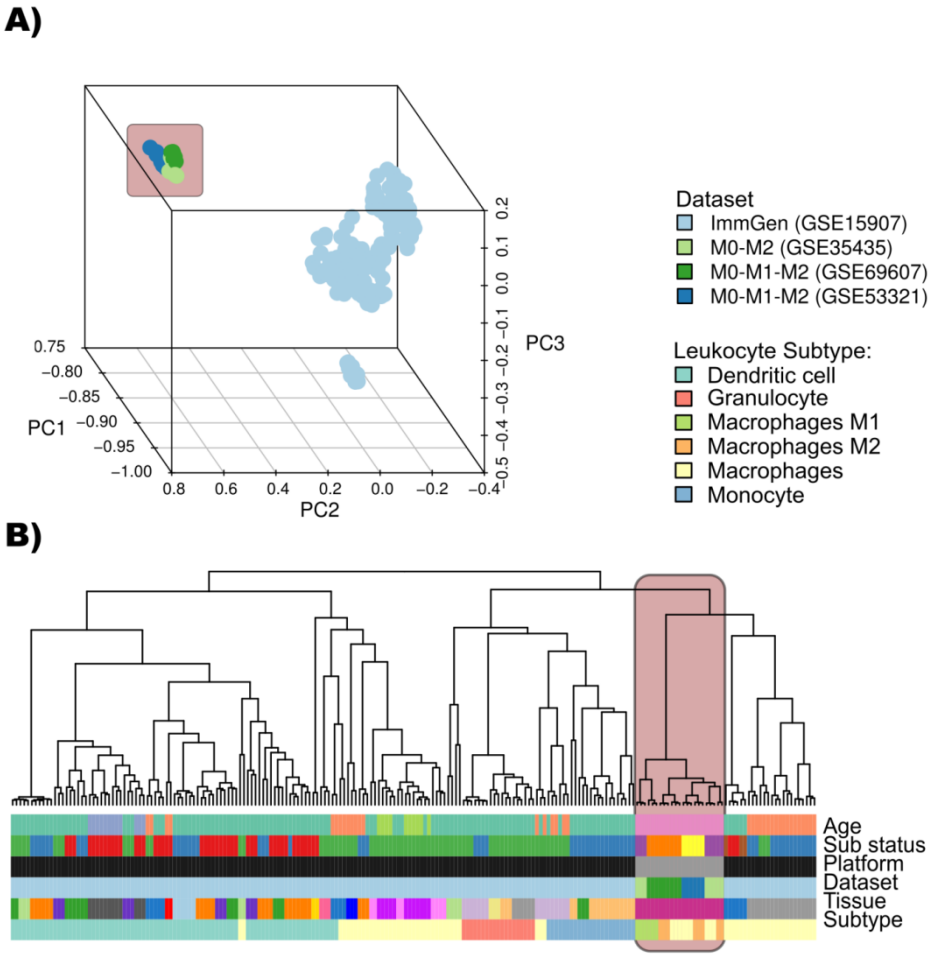


Figure 3.11. **A)** PCA analysis of the meta-dataset highlights a clear batch effect corresponding to the different datasets/platform used; samples from the datasets other than ImmGen (GSE15907) are highlighted by the light-red box. **B)** The Unsupervised clustering using 10% of most modulated genes confirmed this effect.

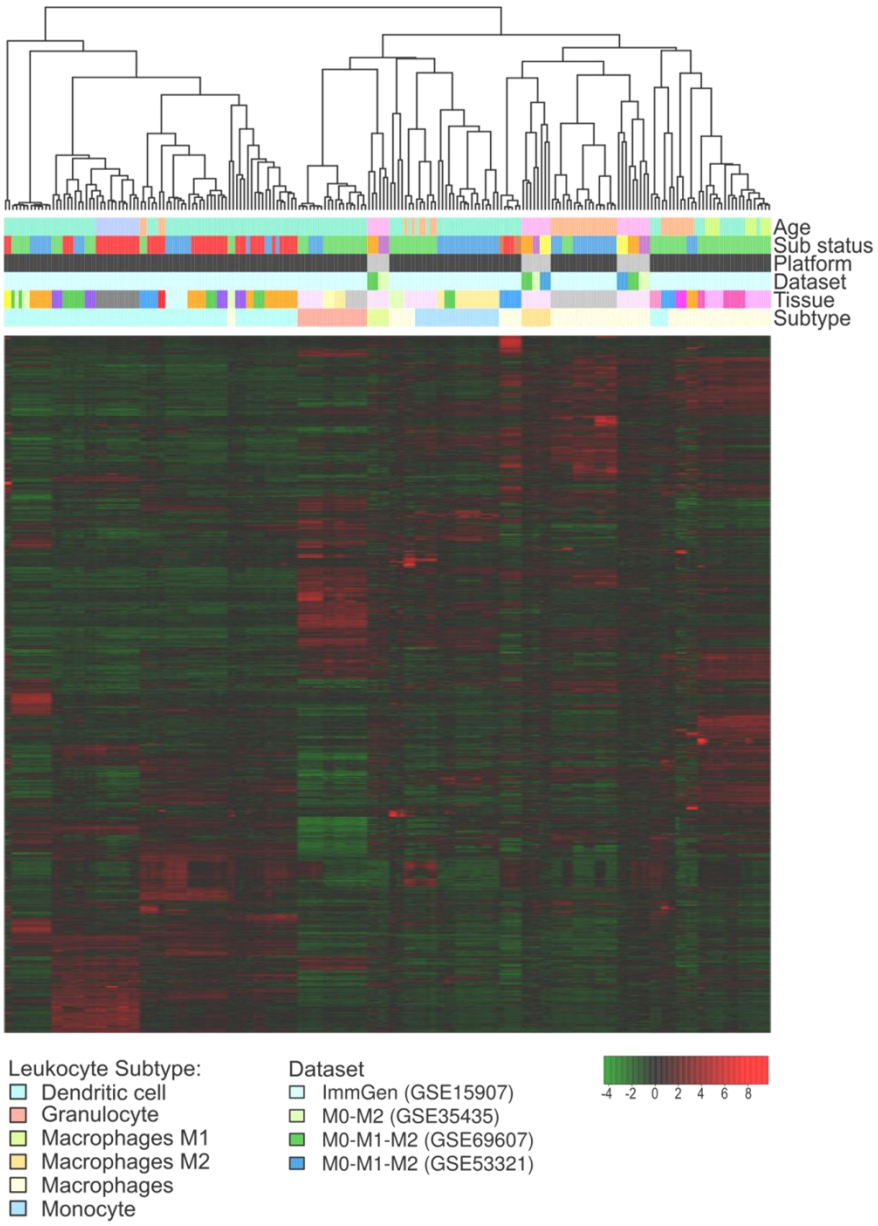


Figure 3.12. Unsupervised clustering using the whole transcriptional profile of the meta-dataset with the batch correction (GEO dataset set as batch and the subtype as a covariate). The Pearson's correlation was used as measure of distance and average as linkage method. Main annotations for each sample are summarized by color code. For brevity, we reported legends for dataset and leukocyte subtype annotations only. Color legend as in **Figure 3.11**.

Using the new murine meta-dataset, we defined a mouse gene signatures to distinguish among the above 27 combinations of cellular subtype and tissue. We performed all the pairwise comparisons across the different cell types: among the resulting significant genes, we progressively selected from 3 to 50 genes of each comparison. This approach resulted in the creation of 48 gene signatures, from where we finally selected the best gene signature according to the lower condition number (see Materials & Methods section for details). The final signature matrix was composed of 1,060 genes and, as expected, is enriched in several hematopoietic-related genes (**Figure 3.13**). Specifically, it includes several CD molecules, Receptor Tyrosine Kinases and Chemokines. The supervised clustering outlined clusters of up-regulated genes characteristics of the different leukocyte subtypes: e.g. DCs are mainly enriched in genes for T-cells activations (CDs molecules and HLA complex), whereas macrophages are enriched in cytokines (IL1A, IL6, IL10) and chemokines (CCL3/4/7/8).

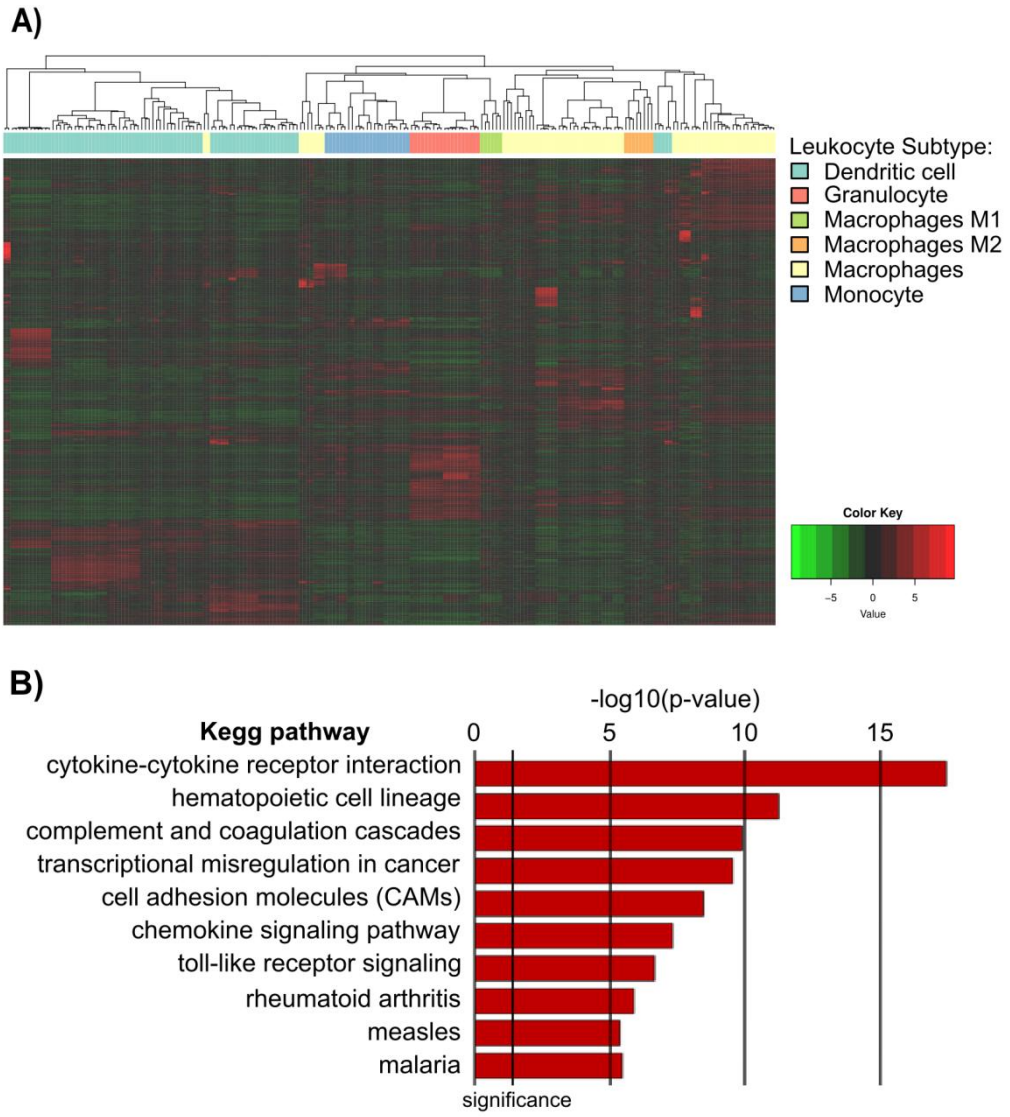


Figure 3.13. **A)** The expression of the 1060 hematopoietic genes from the murine signature matrix; the Pearson's correlation was used as a measure of distance and average as linkage method. **B)** Gene functional analysis on the murine signature matrix highlights a significant enrichment in genes pivotal for immune cells activity and signaling.

Then, the new murine reference matrix was tested using a transcriptional profile of several murine purified populations from independent public datasets. We collected samples from dendritic, macrophages and monocytes populations extracted from different mouse tissues. In general, the deconvolution on purified cells mainly classified each sample in the corresponding cell type annotation (**Figure 3.14**). Macrophages from interstitial tissue are the only exception: they are separated in similar proportions between macrophages, monocytes and dendritic cells. Interestingly, deconvolution with this gene signature distinguishes both the cell type and the tissue, or the most similar tissue, from which the cell has been extracted: e.g. dendritic cells from spleen are classified as a correct combination of cell and tissue with 51.3% on average and 78.5% on average when considering all DC subtypes.

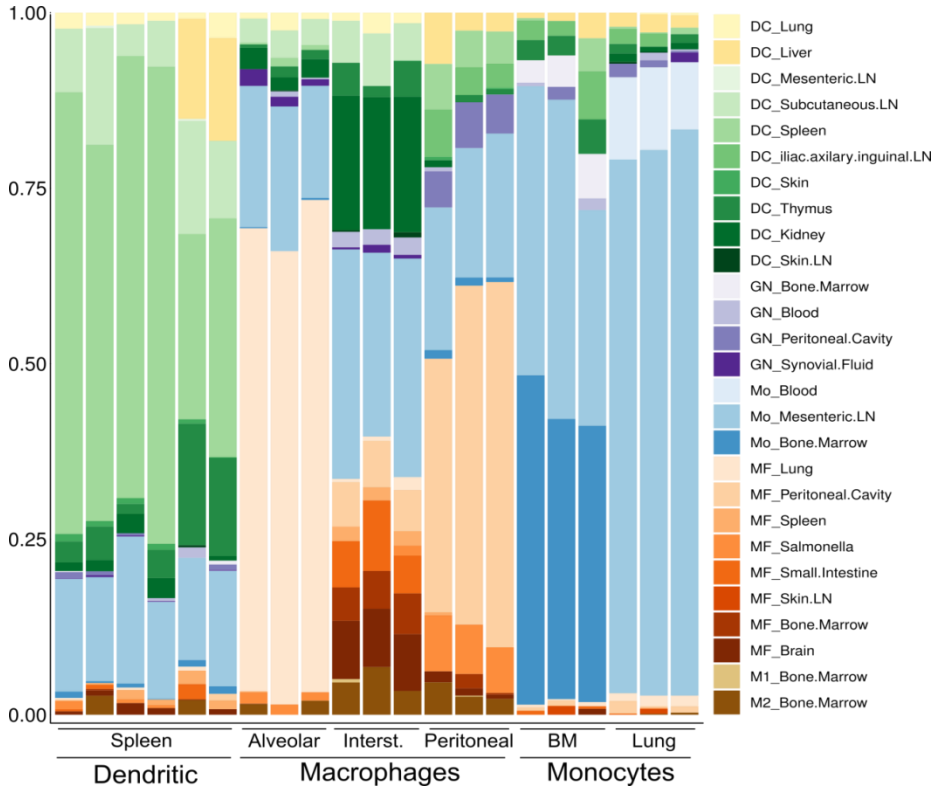


Figure 3.14. Test of the new mouse signature on different populations from multiple tissues. Expression profiles are a small collection of cell types from different public datasets. Interstiz.=interstizial tissue; LN=lymph node.

4.2.2. Define tumor heterogeneity in breast cancer

In tumor samples, the definition of both TME and bulk cells composition is critically important to outline prognosis on tumor evolution and to correctly plan its treatment. An additional and significant confounding variable is determined by the accumulation of different genetic alterations, which could generate the co-presence of multiple subtypes within the same lesion. Clinical evaluation of tumor heterogeneity is an emergent issue to improve clinical oncology; indeed, intratumoral heterogeneity is closely related to cancer progression, resistance to therapy, and recurrences [60]. For this reason, we built on CIBERSORT [17] to design a framework for the identification of cellular subpopulations of cancer samples from their bulk gene expression.

We started by collecting expression data of 57 breast cancer samples closely verified by immunohistochemistry (IHC) for expression of standard clinical markers in molecular subtype definition: estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor-2 (HER2), and Ki-67. Based on the different markers combination, samples were classified in 5 main molecular subtypes: Luminal A (LumA), Luminal B, HER2/ER+, HER2/ER- and TNBC (**Table 3.9**). By comparing each subtype against rest of samples, we generated a molecular gene signature composed by 230 genes, named BCsig (**Figure 3.15**); it includes common markers for BC subtyping, ERBB2 (erb-b2 receptor tyrosine kinase 2, alias HER2), ER, PR, and also other important players in

cell differentiation (GATA3, SOX10, SOX11), cell-cell signaling (APOE, MAPT, SSTR2, S100A9), metallopeptidases (MMP7, MMP1) or dysregulated in specific BC subtypes (FOXA1)[61].

Table 3.9. The number of samples and markers expression of each breast cancer subtype.

Subtype	#samples	Clinical markers			
		ER	PR	HER2	Ki-67
LumA	15	pos. or	pos.	neg. or low	neg. or low
LumB	14	pos.	pos.	neg. or low	high
HER2/ER+	8	pos.	/	pos.	/
HER2/ER-	6	neg.	/	pos.	/
TNBC	14	neg.	neg.	neg.	variable

neg.=negative; pos.=positive; low=up to 10%

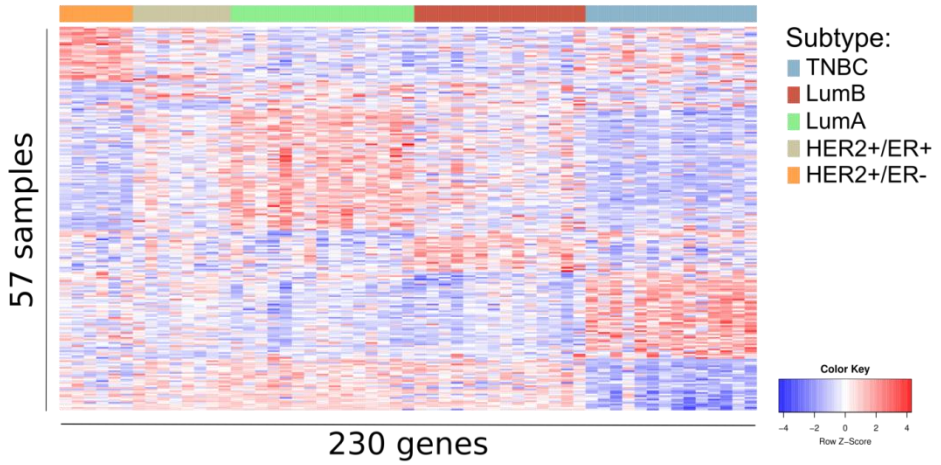


Figure 3.15 Heatmap with the 230 genes on the BC training set.

We used CIBERSORT with the *BCsig* to classify two homogenous and clinically-defined BC datasets of 283 samples profiled using

Illumina arrays (test dataset) and of 88 samples from the TCGA project profiled using Agilent arrays (validation dataset): of note, both datasets are composed only by TNBC primary tumors. By deconvolution analysis with the BCsig, the algorithm estimates the proportion of BC subtypes for each bulk. According to clinical data, in both datasets, there is a high prevalence of the triple-negative subtype also at molecular level (**Figure 3.16A**). Interestingly, when we looked at each single sample, we noticed a variable intratumoral heterogeneity of subtypes, with a modest or high presence of subtypes other than TNBC for about 40% of samples (**Figure 3.16B**). A subset of samples (n = 40, 13%) shows an almost complete absence of TNBC-like cells (TNBC fraction < 5%). Similar results were found for the samples from the TCGA dataset, where 15 samples (17%) are completely TNBC-negative.

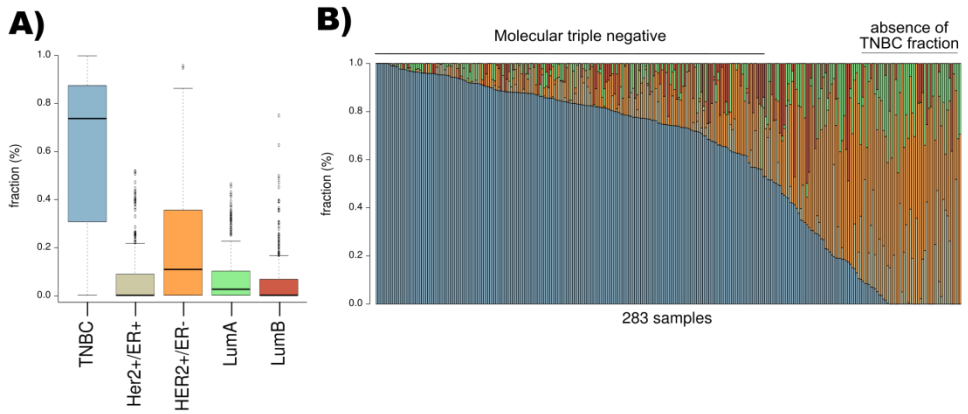


Figure 3.16. **A**) Subtype fraction distribution in the test dataset of 283 samples according to the deconvolution analysis with the BCsig; **B**) about two-third of samples are variably heterogeneous, with a modest or high presence of subtypes other than TNBC: about two-third of the dataset is concordantly TNBC at molecular level, whereas part of samples show a complete absence of the TNBC subtype. Colors as in panel A.

Finally, we tested the TNBC fraction for either clinical response to neoadjuvant chemotherapy or survival. In the test dataset, the analysis identified a poorer and significant prognosis in samples with lower fraction of TNBC cells (**Figure 3.17A**): specifically, patients with residual disease (RD) after treatment show a lower significant TNBC fraction (p=0.012). ROC curve confirms that the TNBC fraction can be efficiently used to identify patient samples which do not respond to therapy. A similar trend was observed also when we associated the TNBC fraction to survival (**Figure 3.17B**), both when we tested for differential distribution between alive or dead patients (boxplot, p=0.067) or for sensitivity and specificity in detecting patient outcome (ROC curve, p=0.082). The Kaplan-Meier curve still evidenced a better outcome in patient with high (>75%) TNBC fraction; even if the result was only almost statistically significant (p 0.0742), the two groups of TNBC high or TNBC low fraction showed different hazard of survival events, respectively 20.4% or 33.1%, and their curves clearly separate. These results suggest subtypes heterogeneity has greater influence over the clinical response and less on the survival of TNBC patients from this cohort. Of note, a significant association was found between the TNBC fraction (as TNBC or non-TNBC) with the neoadjuvant treatment type (p=0.029), and IHC markers KI67 ($p=4.43e^{-13}$) and HER2 (p=0.03).

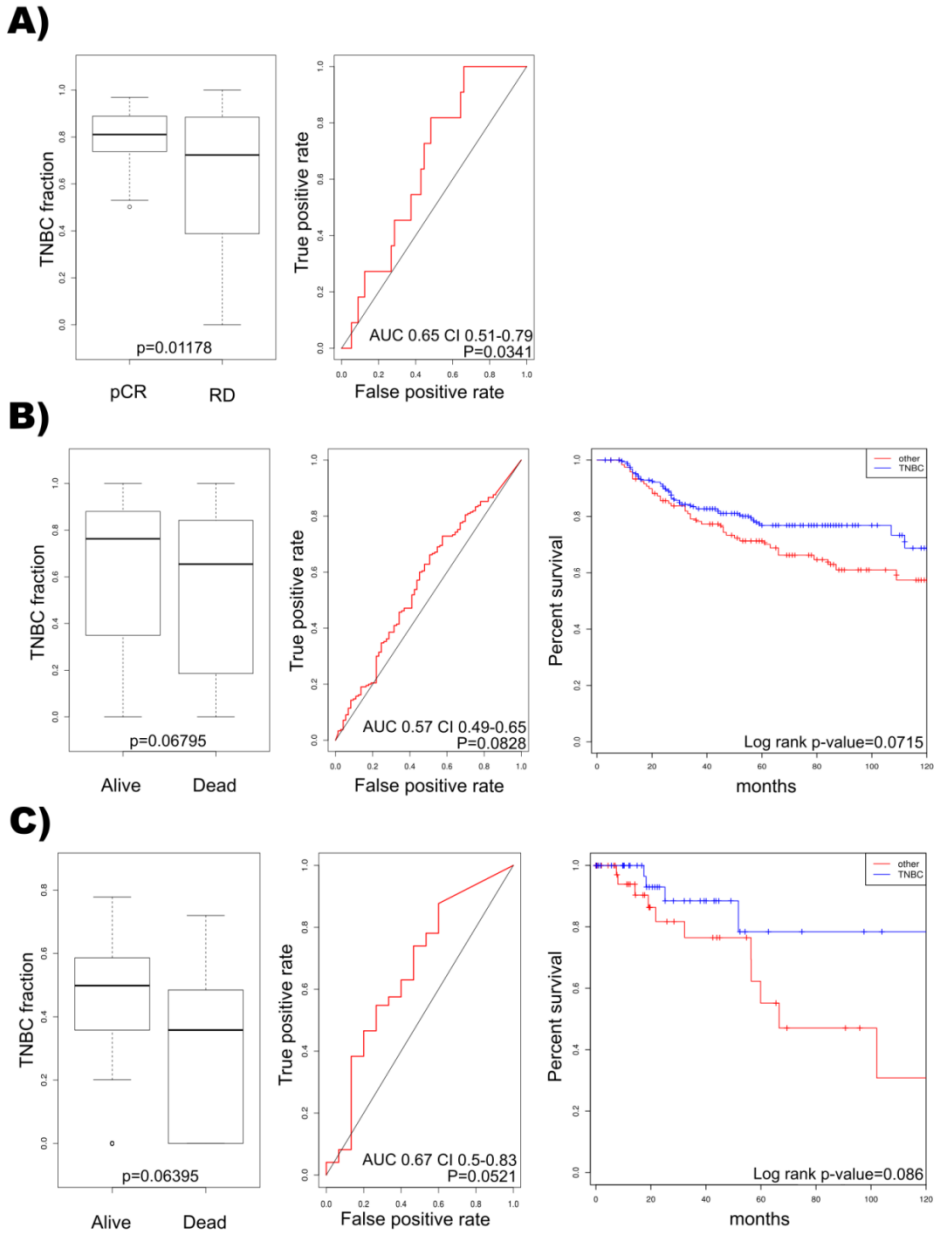


Figure 3.17. Evaluation of the predictive power of TNBC fraction for the response to neoadjuvant therapy and survival. **A**) Analysis of the association between the TNBC fraction and clinical response in the test dataset of 283 samples by boxplots of the TNBC fraction in complete response (pCR, n=11) and residual disease (RD, n=56) samples (left) or by ROC curve (right); **B**) Association with survival in the same dataset by boxplot of the TNBC fraction in alive (n=210) vs. dead (n=73)

patients (left), ROC curves (center) or Kaplan-Meier curves of samples with a detected fraction of at least 75% of TNBC cells vs. remaining samples (right); **C**) Analysis of the association between the TNBC fraction with survival in the TCGA dataset, tested by boxplot of the TNBC fraction in alive (n= 73) vs. dead (n=15) patients (left), ROC curves of the association with a poor prognosis (center), or Kaplan-Meier curves of samples with a detected fraction of at least 50% of TNBC cells vs a lower fraction (right).

To independently validate these findings, we verified the association between TNBC fraction and survival in the validation cohort of 88 triple-negative samples from the TCGA (**Figure 3.17C**). The fraction of molecular TNBC subtype was associated to survival, both by comparing this fraction in alive vs. dead patients (p=0.063) and by ROC curves analysis (p=0.052). The Kaplan-Meier plot confirmed a higher survival rate in patients with at least 50% of TNBC cells (p=0.086). Unfortunately, response to treatment was not available for this dataset.

In conclusions, deconvolution analysis uncovered intratumoral heterogeneity in clinically uniform cohorts of BC patients; thus, subtype fractions detected by deconvolution analysis can be a valid instrument to address triple-negative breast cancer patients to best existing therapy.

5. Discussion

In the first part of my PhD project, I evaluated the accuracy of tools for the deconvolution of gene expression data from bulk samples in different conditions. I tested all available tools and selected 4 of them, namely CIBERSORT, EPIC, ssGSEA and xCell, on the base of the presence of a proprietary gene signature and of their usability. Indeed, these 4 tools are easy-to-use also for non-expert users, since they have been implemented as a user-friendly web interface, as well as R or Java applications for more specific analysis. Tools included in my framework were previously published with validations on several datasets and conditions. However, deconvolution analyses are still lacking a perfect gold standard protocol for validation: to this aim, large datasets of solid tumors with paired bulk samples and independent measurements of different cellular proportions are required to optimally assess tools performances. Unfortunately, this type of data is not yet available. As such, I extended the analyses presented in the authors' paper using independent and selected studies, to highlight how deconvolution analysis performs on different cell types and data format.

To evaluate deconvolution response to variations in sample processing, I initially tested datasets composed of purified cells prepared with the same protocol but from different laboratories

(GSE28490 / GSE28491[28]) or for the same patient and extracted from different tissues (GSE50008[29]). Also, I performed further tests on tumor samples (GSE21029[30]), to evaluate tools performance on bulks from pathological conditions. As general considerations, results were characterized by high variability, depending both on the tool and on the analyzed dataset. For example, CIBERSORT showed high accuracy in almost all tested datasets and conditions, with the exception of few specific populations (mDC and pDC). Differently, EPIC results were characterized by poor accuracy for several populations (B-cells, CD8 and partially for monocytes), despite the low granularity of its signature; also, low performances in the identification of B Cells and CD8 were afterwards confirmed by sc analysis and also recently discussed elsewhere [62]. Furthermore, EPIC results are dependent on the kind of provided input (linear or logarithmic) (**Figure 3.2**): even though the tool has been designed to work with RNA-seq data, no limitation is set to the data distribution. Given the wide availability of expression data profiled by array, a query option to distinguish between array or RNA-seq input may be sufficient, as available in CIBERSORT and xCell pipeline. Conversely, ssGSEA displayed inaccuracy in many analyses, with a generic enrichment when using either purified or single-cell bulk populations, as detailed below. This outcome is at least partially caused by the distribution of the enrichment score, which is not linear and does not have a maximum limit value to be used as reference across different samples. Although some attempts at scaling the enrichment scores of each analysis to a 0-1 scale were made, no

clear improvement in the enrichment accuracy were achieved, nor in discriminating most significant populations from rest (data not shown). On the contrary, xCell detects with high consistency all cell types from samples of healthy subjects, despite the high parceling out of the gene signature. However, the main limitation of xCell remains the constraint of a heterogeneous dataset, meaning that the same sample has different enrichment scores depending on the dataset in which it is analyzed with, as occurred in the test of CLL samples.

Later, I used data from single-cell as the gold standard for the identification of populations heterogeneity in transcriptional bulks. Thus, my first step consisted in searching for studies with paired bulk-sc RNA-seq profiles and additional features for their use in deconvolution analysis, i.e. the presence of the tumor infiltrate. Interestingly and also surprisingly, this type of datasets is not currently available in public databases, with the only exception of the Breast dataset [32]. On the other hand, this lack of data emphasizes that nowadays single-cell technology still requires a lot of resources to build large datasets. To cope with this need, I generated two *in silico* bulk datasets using two of most large public sc experiments on solid tumor samples: lung cancer [4] and melanoma [21]. Before performing the analysis, I excluded samples with very low immune cell composition ($n<5$), thus partially reducing original samples size both in the Breast and Melanoma datasets. I performed some test considering the whole dataset (*data not shown*), but samples with

few immune cells produced biases in the relative fractions of immune cell types, as a single immune cell in a sample corresponds to 100% of that particular immune cell subtype. Thus, I opted for a more conservative solution, excluding these samples when comparing bulk deconvolution and sc fractions.

In bulk-sc correlation analysis, both CIBERSORT and xCell revealed a general high consistency, correctly quantifying very low populations, such as macrophages in the Melanoma dataset (2.7% of total cells). Two notes should be highlighted in xCell analysis: (i) correlations have been calculated on fractions generated with a custom pipeline and not directly on ES and (ii) despite authors clearly state that "*inferences are strictly enrichment scores, and cannot be interpreted as proportion*" [20], in our analysis the adjusted ES appears proportional with quantification, especially if we compare xCell score with the uncorrected ES obtained by ssGSEA. EPIC exhibited variable accuracy, with non-significant correlation for T-cells in the Melanoma dataset, despite the high abundance of this population; a new deconvolution analysis performed using the second gene signature available for this tool only partially improved T-cells estimation, thus suggesting that low correlation can be attributed, to some extent, to the algorithm itself. Of note, the bulk-sc correlation was previously calculated in the EPIC manuscript for the melanoma dataset reporting high consistency ($r=0.9$, $p<10^{-5}$), but all populations were analyzed jointly. Furthermore, we excluded two samples compared to their study due to their very low presence of leukocyte populations, for reasons stated above. Finally, both EPIC

and ssGSEA did not correctly detect any populations from the Breast dataset; ssGSEA especially showed low accuracy also for several populations from *in silico* datasets.

It is important to highlight the different behavior of all tools when performing deconvolution on real or *in silico* bulks. The analyses reported in this work show high and generally significant correlations when using generated bulks: this is somehow attended, because the inputted gene expression for deconvolution analysis and cell type fractions used for the correlation analysis both originate from the same high-throughput data, the scRNA-seq experiment. More troubling and less attended are the low correlations observed for the real bulks from the breast dataset; the limited number of samples, seven, is a reason of the poorer deconvolution performance, at least partially. Additionally, the Myeloid class includes a widely heterogeneous set of cells which, despite categorized in different populations, can be considered a continuum, with closely related transcriptional profiles that are more challenging to distinguish by deconvolution algorithmsThe analyses discussed in the previous chapters provided further insights into the deconvolution workflow. Comparing results derived from fraction-based and enrichment-based approaches was not straightforward, thus requiring the creation of a dedicated pipeline in the single-cell analysis. Moreover, it was also noticed that all tools often overestimate low abundant populations: some non-zero scores were attributed to non-existent cell populations, to the detriment of other cell types. This effect has been confirmed and more largely discussed in a recent paper [62].

In deconvolution analyses, it is complex to determine if low performances can be attributed to the algorithm or to the gene signature: a solution would be testing all pairwise combinations of signatures and tools, which is not straightforward due to different gene signatures format. Specifically, for each population, enrichment tools require a list of marker genes, whereas fraction-based ones require an expression profile to perform the linear regression analysis they are based upon. Moreover, the required format for gene signature expression profile differs between CIBERSORT and EPIC: the former exploits an expression matrix of 547 marker genes and their median expression in the 22 populations of the signature, while the latter leverages on an expression matrix containing the median expression and the variability (interquartile range) of the whole transcriptome for each of the 7 evaluated populations. Thus, in this work, I opted to test each tool with its own predefined gene signature.

Finally, from the user perspective, fraction-based results are more easily interpretable than enrichment results. However, evaluation of a fraction significance can be user-dependent when considering low-frequency populations. On the contrary, enrichment tools report a p-value for each enrichment giving a well-defined objective assessment of significance, even though occasionally several significant and unrelated populations are retrieved using either ssGSEA or xCell. This effect is more prominent for the ssGSEA but has been observed also for xCell and in other datasets not presented in this thesis. In

these cases, the definition of sample composition could be tricky for the user. In ssGSEA this effect may be due either to the inaccuracy of the algorithm, of signatures or both; instead, in xCell the reason is at least partially due to the use of uncorrected p-value, which can be overcome by applying an adjusted statistics, thus potentially improving the analysis through reduction of false positives.

In the second part of my work, I explore the possibility to apply the deconvolution pipeline to completely different scenarios. In the first phase, I crafted a dedicated gene signature for the deconvolution of hematopoietic system cells on a different organism, like mouse, and tested the tool response. The immune system plays a central role as mediator of many effective cancer immunotherapies in humans, as also recently reviewed [63]. In pathologies were systemic immunity is required for the effective cancer immunotherapy [64], detailed definition of the cellular composition of clinical samples could provide outstanding support in therapy selection. Mouse still remains a strong rationale to study human diseases; thus, definition of a mouse general framework for the identification of cell subpopulations from transcriptional data may provide strong potential, for cancer immunotherapy preclinical studies in particular. For this reason, I generated a murine gene signature able to discriminate a wide repertoire of immune cell types with potentially close transcriptional profiles, such as the same cellular subtype isolated from different tissues. Gene expression data from the ImmGen project has been the main reference for collecting publicly available dataset of

leukocyte cells; this database is a compendium of transcriptional profiles from immune populations generated by standardized protocols, thus reducing technical, conditions-based and also unknown possible biases. However, this database lacks immune populations from TME, meaning that classically (M1) or alternatively activate (M2) macrophages subtypes were recovered from other public datasets.

Usually, when creating large meta-datasets, a critical step is the selection of the method for expression data combination from different experiments; for this reason, all known and available variables that could affect expression data generation have been carefully considered and evaluated. Several efforts were spent in the batch correction procedure and the construction of the final meta-dataset; despite some analysis helped filtering out some combination of parameters, e.g. the array type was excluded after the unsupervised clustering, no “best” correction method was defined. The analyses indicated that CIBERSORT deconvolution with the murine signature can discriminate very similar populations with close transcriptional profiles. However, a limitation of the analysis still remains the heterogeneity of the datasets used to test the new signature, both in terms of real cellular purity and of protocol for cells isolation. Further tests using profiles of murine populations isolated in rigorous and uniform conditions are required to confirm specificity and sensitivity of the signature.

Subsequently, we applied the framework to outline subtypes heterogeneity in breast cancer (BC). Early transcriptomic studies classified this tumor into four molecular intrinsic subtypes (Luminal A, Luminal B, HER2 enriched, and Basal-like) and a Normal Breast-like group [65,66] by a 50-gene assay (known as the PAM50). The PAM50 classifier has provided independent predictive information of pathologic complete response (pCR) to neoadjuvant therapy across all subtypes [67]. Importantly, all subtypes can be found in immunohistochemically defined triple-negative breast cancer (TNBC), but the basal-like subtype plays a leading role, characterizing from 50% to 75% of TNBC samples [68]. I focused my study on the TNBC subtype since it is currently the most challenging to treat and with overall survival shorter than other subtypes [68]. Indeed, when restricting analyses to TNBC, none of the PAM50 subtypes at the time of diagnosis significantly correlated with pCR [69]. For this reason, even slight differences in survival can have an important clinical implication.

To create a gene signature for breast cancer subtyping, I took advantage of a cohort of 57 breast cancer samples profiled using Illumina array. Each sample was closely verified by immunohistochemistry by clinicians, proving this dataset as an optimal reference for the generation of a molecular subtype gene signature, named BCsig. To evaluate subtypes heterogeneity using the BCsig signature, I took advantage of two cohorts composed by primary, untreated TNBC: importantly, they are both uniforms for sample composition, since they contain only pre-treatment primary

tumors all clinically-defined negative for ER, PR and HER2 markers. The first dataset is composed by 283 samples profiled with the same platform of the training test, Illumina array, whereas the second dataset is composed by the TNBC samples from the TCGA-BRCA project: despite the lower number of samples (88 in total), TCGA samples have been profiled using Agilent array, allowing the test of the BCsig on a different types of expression arrays.

According to *in silico* deconvolution analysis with the BCsig, in both datasets almost 40% of samples showed a variable degree of heterogeneity. The association analysis between this TNBC fraction and either clinical response or survival defined a subgroup of TNBC patients characterized by poorer response and survival and by a heterogeneous composition of the tumor bulk. The outcome is more tightly related to treatment response ($p=0.03$ of ROC curve) than survival ($p=0.08$ and $p=0.05$ for the Illumina or TCGA dataset, respectively): however, clinical response data was not available for the TCGA dataset. It is now well established that, during their course, cancers generally become more heterogeneous: under therapeutic selective pressure, resistance to treatment can emerge as a result of the expansion of pre-existing subclonal populations or from the evolution of drug-tolerant cells [70]. Indeed, co-presence of multiple subtypes within the same lesion has been recently detected by single-cell RNA-seq in other tumors, for example in glioblastoma [71]. With these insights, deconvolution analysis using the BCsig might improve knowledge of TNBC and have potential therapeutic implications, by defining patients with higher heterogeneity and

poorer survival to be addressed to a more personalized treatment. However, the main limitation of this part of my study is the low number of samples used to validate the BCsig significance. For this reason, we are planning to perform deconvolution analyses on a large public dataset composed by more than 3,000 samples profiled by RNA-seq [72].

During the writing of this thesis, two relevant works regarding transcriptional deconvolution were published and a significant project was started. In July 2019, Sturm and colleagues [62] published the first benchmarking work on deconvolution methods, including the tools CIBERSORT, EPIC and xCell. They used transcriptional profile from single-cell data [73] to generate *in silico* bulk with known and increasing proportions of nine cellular subtypes. Their conclusions indicate in general high accuracy of the deconvolution tools but given well-defined and reliable gene signatures. Also, they defined substantial spillover between DC and B-cells in several methods, EPIC included, as also previously discussed here. The second paper presented an extension of the tool CIBERSORT, named CIBERSORTx [74]. This new version extends the application of previous algorithm to both RNA-seq and single-cell RNA-seq data, but basically the new method can infer cell-type-specific gene expression profiles, potentially improving to high extent the decoding of cellular heterogeneity in bulk tissues. Finally, an open challenge was recently launched to evaluate the ability of computational methods to deconvolve bulk expression data, reflecting mixture of cell types into individual immune components. All authors of

deconvolution methods have been invited to participate in this challenge where specifically generated *in vitro* and *in silico* admixtures will be analyzed. After a training phase, an intensive benchmarking of the tools will be performed on several datasets at different granularities. Together with the performance, several other figures will be investigated and compared, comprising collinearity as well as probe limits of detection.

The challenge is still in progress and results of both training and validation phases will be available at the end of the challenge, which is expected at the end of 2019 (https://www.synapse.org/#!Synapse:syn15589870/wiki/582446).

This project could be a cornerstone to determine the efficiency of existing methods and for the implementation of new algorithms for transcriptional deconvolution.

6. Conclusions

This thesis defined a framework to investigate heterogeneity in bulk transcriptional data of healthy and tumor samples through the use of deconvolution methods.

Lack of a gold standard to evaluate the performance of transcriptional deconvolution required a preliminary assessment of tool performances using independent and selected studies from bulk profiles of purified cells and of scRNA-seq experiments.

The second part of this work focused on the creation of molecular signatures for deconvolution in specific settings. We firstly generated a murine signature which discriminated cells with close transcriptional profiles, like the same cell type from different tissues. Then, we defined a gene signature to address intratumoral heterogeneity in breast cancer; its application highlighted the existence of a subgroup of TNBC patients whose heterogeneous composition of the bulk correlates with a poorer prognosis.

The first contribution of this work is the generation of the two molecular gene signatures, for either murine leukocytes or subtypes in breast cancer.

A second contribution is the characterization of the TNBC subtype, which is currently the most challenging to treat and with overall survival shorter than other subtypes. The analysis improved knowledge of TNBC with potential therapeutic implications by the identification of patients with higher heterogeneity.

Finally, a third contribution is the creation of a web interactive application to display the correlation between the deconvolution of bulk tissues obtained with four deconvolution tools, and the frequency of subpopulations assessed by single-cell profiling.

Perspectives and short-term objectives concern the evaluation of recent deconvolution tools to include in my pipeline and test of the breast cancer gene signature on other public datasets for the association on either clinical response or survival.

7. References

[1] Jandt U, Platas Barradas O, Pörtner R, Zeng AP. Mammalian cell culture synchronization under physiological conditions and population dynamic simulation. Appl Microbiol Biotechnol 2014. doi:10.1007/s00253-014-5553-6.

[2] Zhao Y, Simon R. Gene expression deconvolution in clinical samples. Genome Med 2010. doi:10.1186/gm214.

[3] Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. Nat Immunol 2014. doi:10.1038/ni.2787.

[4] Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 2018. doi:10.1038/s41591-018-0096-5.

[5] Hespel C, Moser M. Role of inflammatory dendritic cells in innate and adaptive immunity. Eur J Immunol 2012. doi:10.1002/eji.201242480.

[6] Noy R, Pollard JW. Tumor-Associated Macrophages: From Mechanisms to Therapy. Immunity 2014. doi:10.1016/j.immuni.2014.06.010.

[7] Cleator SJ, Powles TJ, Dexter T, Fulford L, Mackay A, Smith IE, et al. The effect of the stromal component of breast tumours on prediction of clinical outcome using gene expression microarray analysis. Breast Cancer Res 2006. doi:10.1186/bcr1506.

[8] Ali HR, Chlon L, Pharoah PDP, Markowetz F, Caldas C. Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study. PLoS Med 2016;13:1–24. doi:10.1371/journal.pmed.1002194.

[9] Xiong Y, Wang K, Zhou H, Peng L, You W, Fu Z. Profiles of immune infiltration in colorectal cancer and their clinical significant: A gene

expression-based study. Cancer Med 2018. doi:10.1002/cam4.1745.

[10] Conway EM, Pikor LA, Kung SHY, Hamilton MJ, Lam S, Lam WL, et al. Macrophages, inflammation, and lung cancer. Am J Respir Crit Care Med 2016. doi:10.1164/rccm.201508-1545CI.

[11] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013. doi:10.1038/nrg3542.

[12] Mohammadi S, Zuckerman N, Goldsmith A, Grama A. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proc IEEE 2017;105:340–66. doi:10.1109/JPROC.2016.2607121.

[13] Venet D, Pecasse F, Maenhaut C, Bersini H. Separation of samples into their constituents using gene expression data. Bioinformatics 2001. doi:10.1093/bioinformatics/17.suppl_1.S279.

[14] Campisi P, Egiazarian K. Blind image deconvolution: Theory and applications. 2017. doi:10.1201/9781420007299.

[15] Whalley K. Imaging without the blur. Nat Cell Biol 2009. doi:10.1038/ncb1949.

[16] Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife 2017;6:1–25. doi:10.7554/eLife.26476.

[17] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7. doi:10.1038/nmeth.3337.

[18] Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009. doi:10.1038/nature08460.

[19] Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep 2017;18:248–62. doi:10.1016/j.celrep.2016.12.019.

[20] Aran D, Hu Z, Butte AJ. xCell: Digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 2017. doi:10.1186/s13059-017-1349-1.

[21] Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (80-) 2016. doi:10.1126/science.aad0501.

[22] Feingold EA, Good PJ, Guyer MS, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. Science (80-) 2004. doi:10.1126/science.1105136.

[23] Stunnenberg HG, Abrignani S, Adams D, de Almeida M, Altucci L, Amin V, et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell 2016. doi:10.1016/j.cell.2016.11.007.

[24] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity 2018. doi:10.1016/j.immuni.2018.03.023.

[25] 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. doi:10.1073/pnas.0506580102.

[26] Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 2015. doi:10.1186/s13059-014-0560-6.

[27] Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (IRIS): Immune-specific genes identified from a compendium of microarray expression data. Genes Immun 2005. doi:10.1038/sj.gene.6364173.

[28] Allantaz F, Cheng DT, Bergauer T, Ravindran P, Rossier MF, Ebeling M, et al. Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression.

PLoS One 2012;7. doi:10.1371/journal.pone.0029979.

[29] Beliakova-Bethell N, Massanella M, White C, Lada SM, Du P, Vaida F, et al. The effect of cell subset isolation method on gene expression in leukocytes. Cytom Part A 2014;85:94–104. doi:10.1002/cyto.a.22352.

[30] McCoy JP, Liu P, Biancotto A, Gibellini F, White T, Stennett L, et al. The lymph node microenvironment promotes B-cell receptor signaling, NF- B activation, and tumor proliferation in chronic lymphocytic leukemia. Blood 2010;117:563–74. doi:10.1182/blood-2010-05-284984.

[31] Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 2014. doi:10.1371/journal.pone.0078644.

[32] Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun 2017. doi:10.1038/ncomms15081.

[33] Painter MW, Davis S, Hardy RR, Mathis D, Benoist C, Consortium IGP. Transcriptomes of the B and T lineages compared by multiplatform microarray profiling. J Immunol 2011.

[34] Martinez FO, Helming L, Milde R, Varin A, Melgert BN, Draijer C, et al. Genetic programs expressed in resting and IL-4 alternatively activated mouse and human macrophages: Similarities and differences. Blood 2013. doi:10.1182/blood-2012-06-436212.

[35] Jablonski KA, Amici SA, Webb LM, Ruiz-Rosado JDD, Popovich PG, Partida-Sanchez S, et al. Novel markers to delineate murine M1 and M2 macrophages. PLoS One 2015. doi:10.1371/journal.pone.0145342.

[36] Li L, Ng DSW, Mah WC, Almeida FF, Rahmat SA, Rao VK, et al. A unique role for p53 in the regulation of M2 macrophage polarization. Cell Death Differ 2015. doi:10.1038/cdd.2014.212.

[37] Rosas M, Davies LC, Giles PJ, Liao C Te, Kharfan B, Stone TC, et al. The transcription factor Gata6 links tissue macrophage phenotype and proliferative renewal. Science (80-) 2014. doi:10.1126/science.1251414.

[38] Edwards AD, Chaussabel D, Tomlinson S, Schulz O, Sher A, Reis e Sousa C. Relationships Among Murine CD11c high Dendritic Cell Subsets as Revealed by Baseline Gene Expression Patterns . J Immunol 2003. doi:10.4049/jimmunol.171.1.47.

[39] Sabatel C, Radermecker C, Fievez L, Paulissen G, Chakarov S, Fernandes C, et al. Exposure to Bacterial CpG DNA Protects from Airway Allergic Inflammation by Expanding Regulatory Lung Interstitial Macrophages. Immunity 2017. doi:10.1016/j.immuni.2017.02.016.

[40] Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol 2014. doi:10.1002/msb.134947.

[41] Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, et al. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics 2010. doi:10.1186/1471-2105-11-27.

[42] Gong T, Szustakowski JD. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics 2013;29:1083–5. doi:10.1093/bioinformatics/btt090.

[43] Erkkilä T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H. Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics 2010. doi:10.1093/bioinformatics/btq406.

[44] Frishberg A, Brodt A, Steuerman Y, Gat-Viks I. ImMquant: A user-friendly tool for inferring immune cell-type composition from gene-expression data. Bioinformatics 2016. doi:10.1093/bioinformatics/btw535.

[45] Bolen CR, Uduman M, Kleinstein SH. Cell subset prediction for blood genomic studies. BMC Bioinformatics 2011. doi:10.1186/1471-2105-12-258.

[46] Frishberg A, Steuerman Y, Gat-Viks I. CoD: Inferring immune-cell quantities related to disease states. Bioinformatics 2015.

doi:10.1093/bioinformatics/btv498.

[47] Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells. Cancer Res 2017. doi:10.1158/0008-5472.CAN-17-0307.

[48] Wang Y, Xia XQ, Jia Z, Sawyers A, Yao H, Wang-Rodriquez J, et al. In silico estimates of tissue components in surgical samples based on expression profiling data. Cancer Res 2010. doi:10.1158/0008-5472.CAN-10-0021.

[49] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013. doi:10.1038/ncomms3612.

[50] Finotello F, Mayer C, Plattner C, Laschober G, Rieder Di, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med 2019. doi:10.1186/s13073-019-0638-6.

[51] Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression [Genome Biol., 17 (2016), (218)]. Genome Biol 2016;17:1–20. doi:10.1186/s13059-016-1113-y.

[52] Gaujoux R, Seoighe C. CellMix: A comprehensive toolbox for gene expression deconvolution. Bioinformatics 2013;29:2211–2. doi:10.1093/bioinformatics/btt351.

[53] Tamborero D, Rubio-Perez C, Muiños F, Sabarinathan R, Piulats JM, Muntasell A, et al. A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations. Clin Cancer Res 2018. doi:10.1158/1078-0432.CCR-17-3509.

[54] Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002. doi:10.1093/nar/30.1.207.

[55] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003. doi:10.1093/nar/gkg091.

[56] Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res 2016. doi:10.1093/nar/gkv1045.

[57] Cao Y, Zhu J, Jia P, Zhao Z. scRNAseqDB: A database for RNA-seq based gene expression profiles in human single cells. Genes (Basel) 2017. doi:10.3390/genes8120368.

[58] Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res 2011. doi:10.1093/nar/gkq1019.

[59] Heng TSP, Painter MW, Elpek K, Lukacs-Kornek V, Mauermann N, Turley SJ, et al. The immunological genome project: Networks of gene expression in immune cells. Nat Immunol 2008. doi:10.1038/ni1008-1091.

[60] Stanta G, Bonin S. Overview on clinical relevance of intra-tumor heterogeneity. Front Med 2018. doi:10.3389/fmed.2018.00085.

[61] Nakshatri H, Badve S. FOXA1 in breast cancer. Expert Rev Mol Med 2009. doi:10.1017/S1462399409001008.

[62] Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics, 2019. doi:10.1093/bioinformatics/btz363.

[63] Tran E, Robbins PF, Rosenberg SA. Final common pathway' of human cancer immunotherapy: Targeting random somatic mutations. Nat Immunol 2017. doi:10.1038/ni.3682.

[64] Spitzer MH, Carmi Y, Reticker-Flynn NE, Kwek SS, Madhireddy D, Martins MM, et al. Systemic Immunity Is Required for Effective Cancer Immunotherapy. Cell 2017. doi:10.1016/j.cell.2016.12.022.

[65] Perou CM, Sørile T, Eisen MB, Van De Rijn M, Jeffrey SS, Ress CA, et al. Molecular portraits of human breast tumours. Nature 2000. doi:10.1038/35021093.

[66] Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 2010. doi:10.1186/bcr2635.

[67] Prat A, Parker JS, Fan C, Perou CM. PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. Breast Cancer Res Treat 2012. doi:10.1007/s10549-012-2143-0.

[68] Garrido-Castro AC, Lin NU, Polyak K. Insights into molecular classifications of triple-negative breast cancer: Improving patient selection for treatment. Cancer Discov 2019. doi:10.1158/2159-8290.CD-18-1177.

[69] Prat A, Lluch A, Albanell J, Barry WT, Fan C, Chacón JI, et al. Predicting response and survival in chemotherapy-treated triple-negative breast cancer. Br J Cancer 2014. doi:10.1038/bjc.2014.444.

[70] Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018. doi:10.1038/nrclinonc.2017.166.

[71] Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell 2019. doi:10.1016/j.cell.2019.06.024.

[72] Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, et al. Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative. JCO Precis Oncol 2018. doi:10.1200/po.17.00135.

[73] Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun 2017. doi:10.1038/s41467-017-02289-3.

[74] Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019. doi:10.1038/s41587-019-0114-2.

8. Appendix

Figure S1. Deconvolution analysis by CIBERSORT, EPIC, ssGSEA and xCell on GSE28490 dataset composed of 9 cell types of purified cells profiled by Affymetrix array.

Figure S2. Deconvolution analysis by CIBERSORT, EPIC, ssGSEA and xCell on GSE28491 dataset composed of 7 cell types of purified cells profiled by Affymetrix array.

Figure S3. Deconvolution analysis by CIBERSORT, EPIC and ssGSEA on GSE50008 dataset composed of 4 cell types purified by positive selection, negative selection or FACS and profiled by Illumina array.

Figure S4. Deconvolution analysis by CIBERSORT, EPIC, ssGSEA and xCell on GSE21029 dataset composed by samples from CLL patients profiled by Affymetrix array.

Figure S5. Correlation plots of bulk-single cell correlations in the Breast cancer sc dataset.

Figure S6. Correlation plots of bulk-single cell correlations in the Lung cancer sc dataset.

Figure S7. Correlation plots of bulk-single cell correlations in the Melanoma sc dataset.

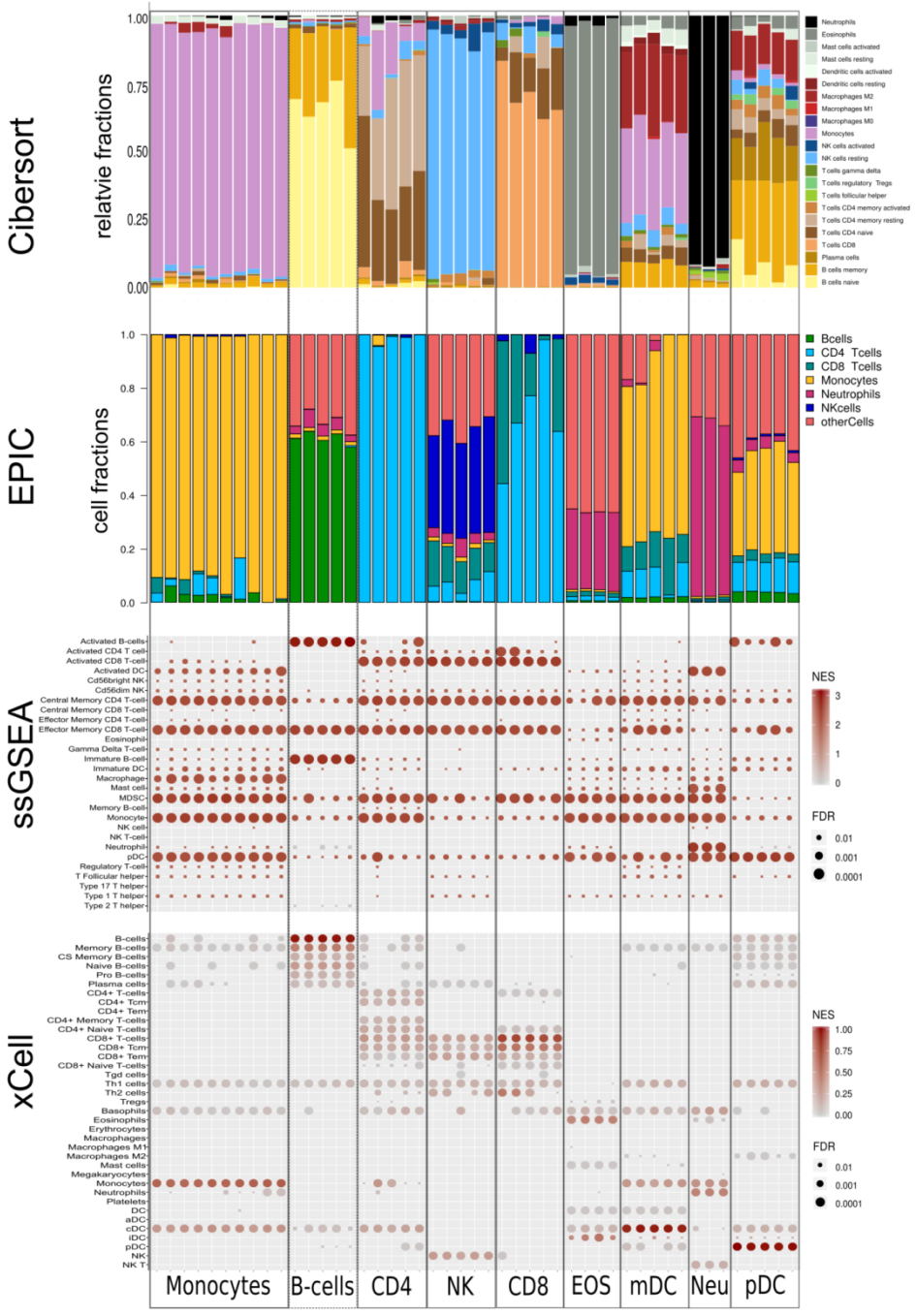


Figure S1. Deconvolution analysis on GSE28490. Neu=neutrophils, EOS=eosinophils.

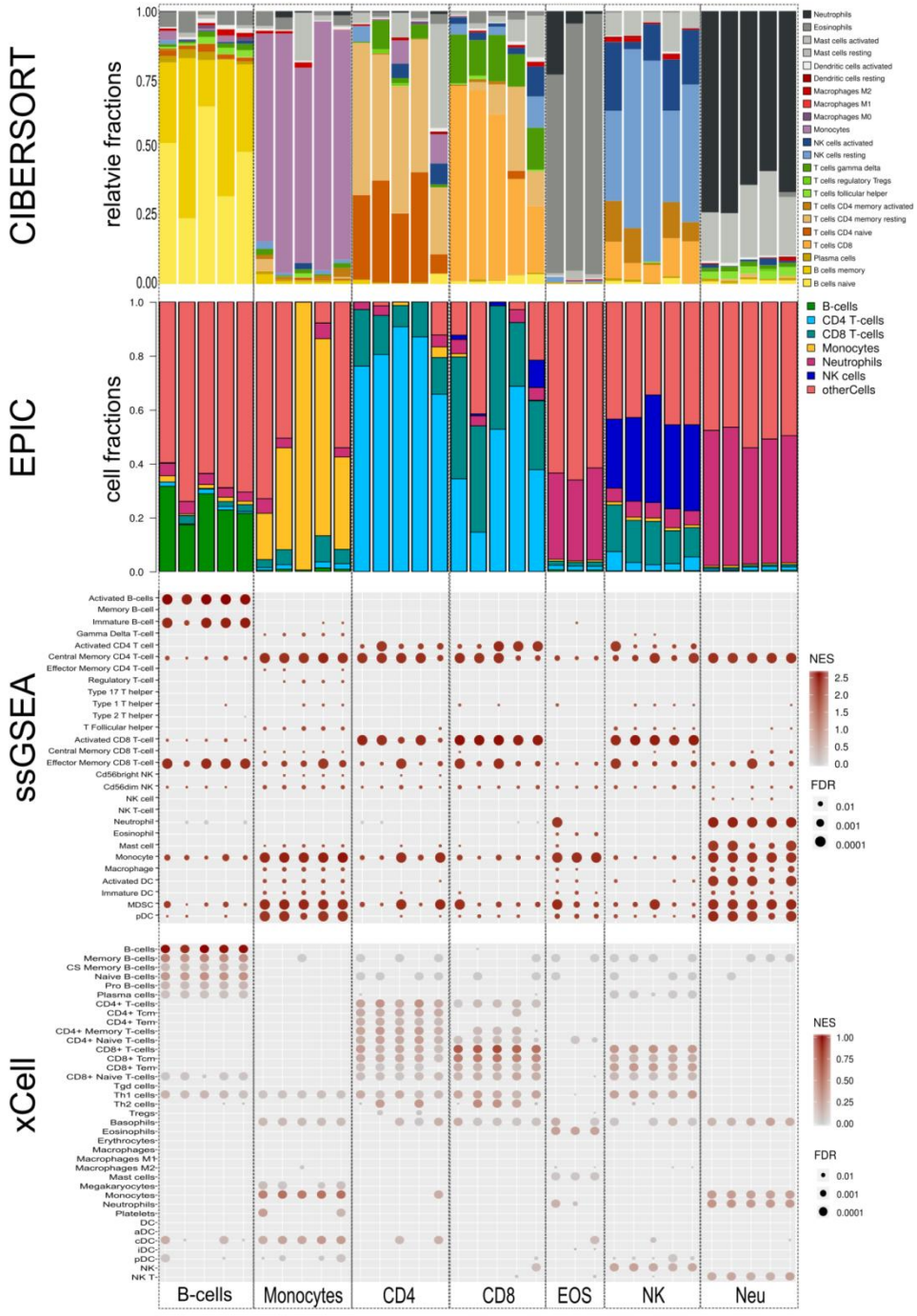


Figure S2. Deconvolution analysis on GSE28491.

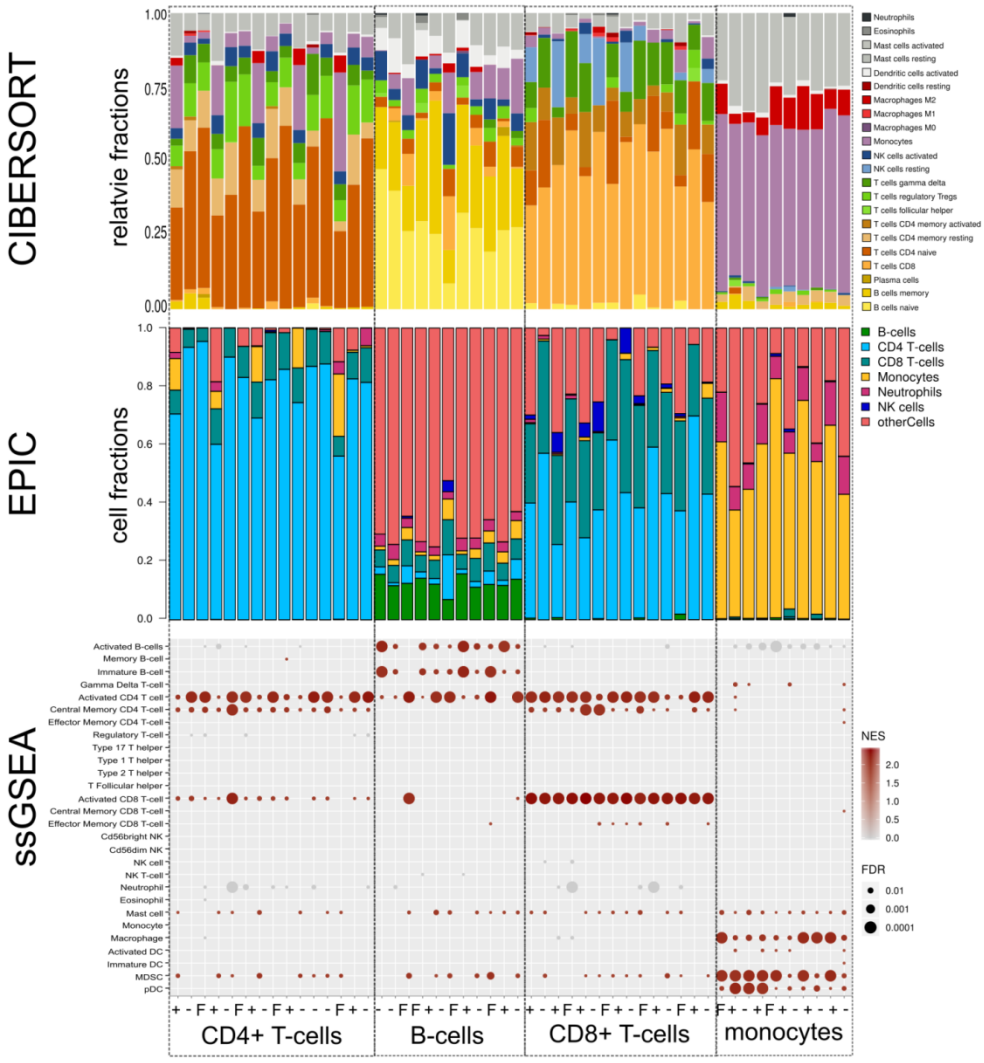


Figure S3. Deconvolution analysis on GSE50008. “+”: positive selection; “-”: negative selection; FACS: selection by FACS.

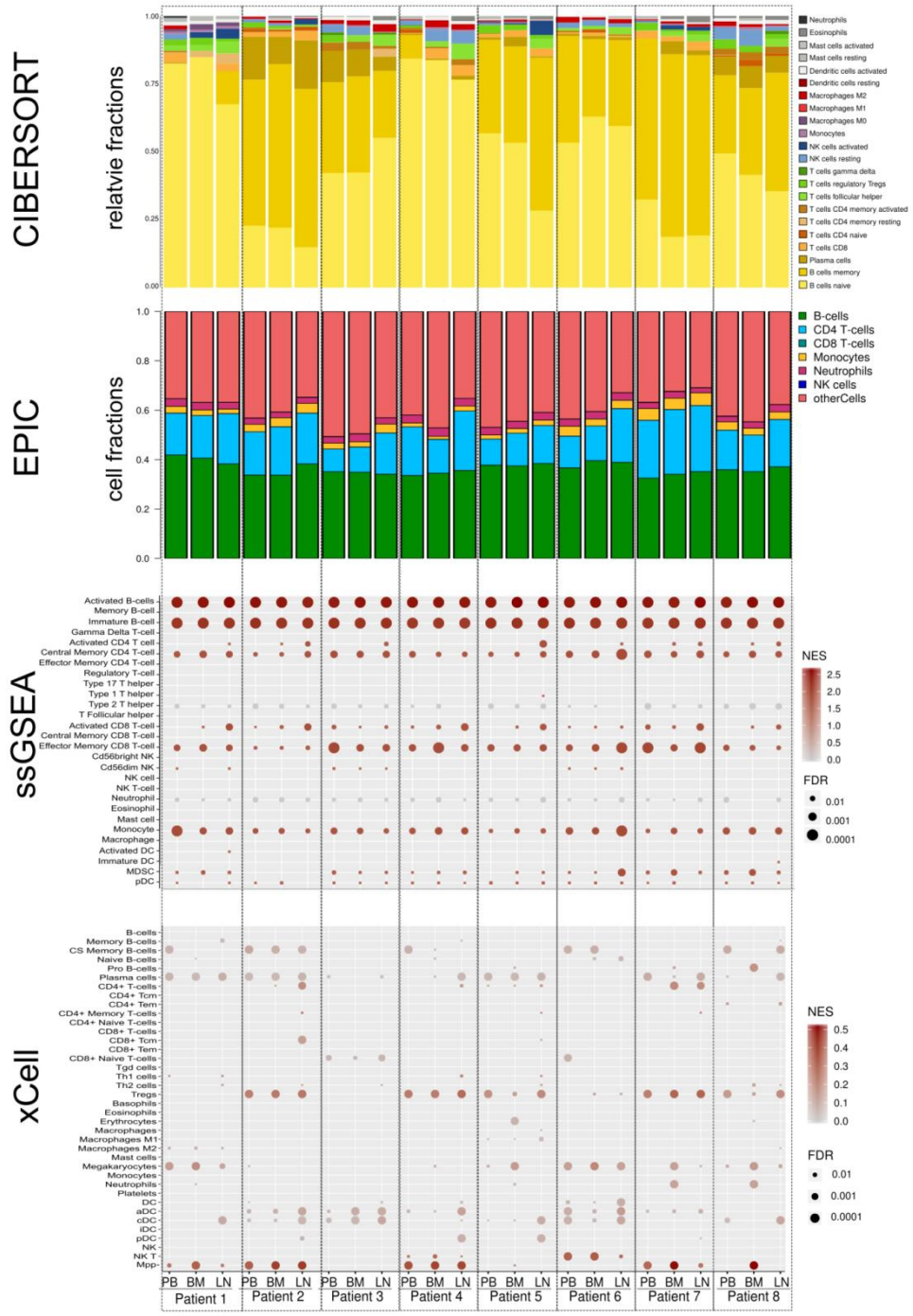


Figure S4. Deconvolution analysis on GSE21029.

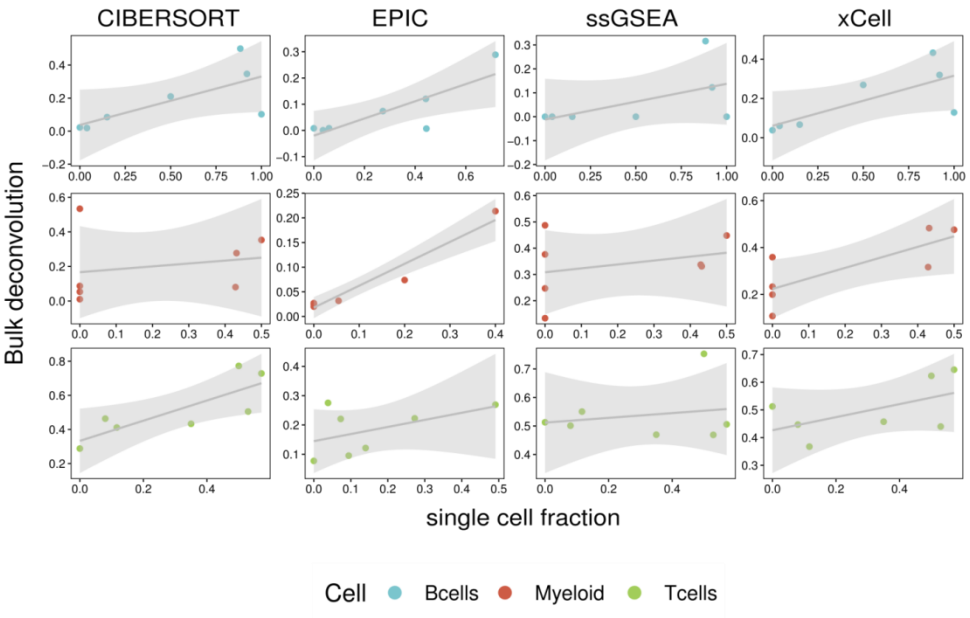


Figure S5. Bulk-single cell correlations in the Breast cancer dataset.

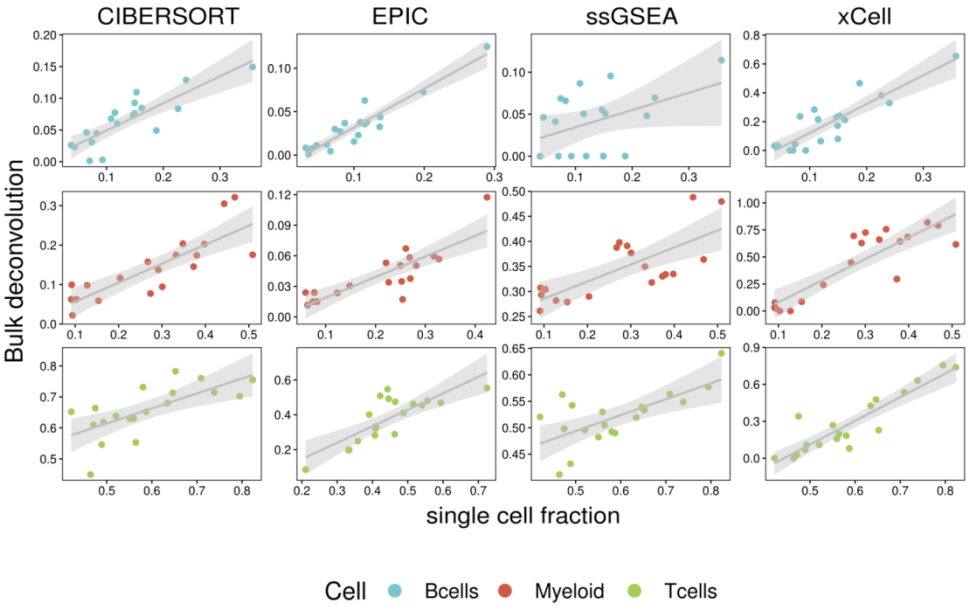


Figure S6. Bulk-single cell correlations in the Lung cancer dataset.

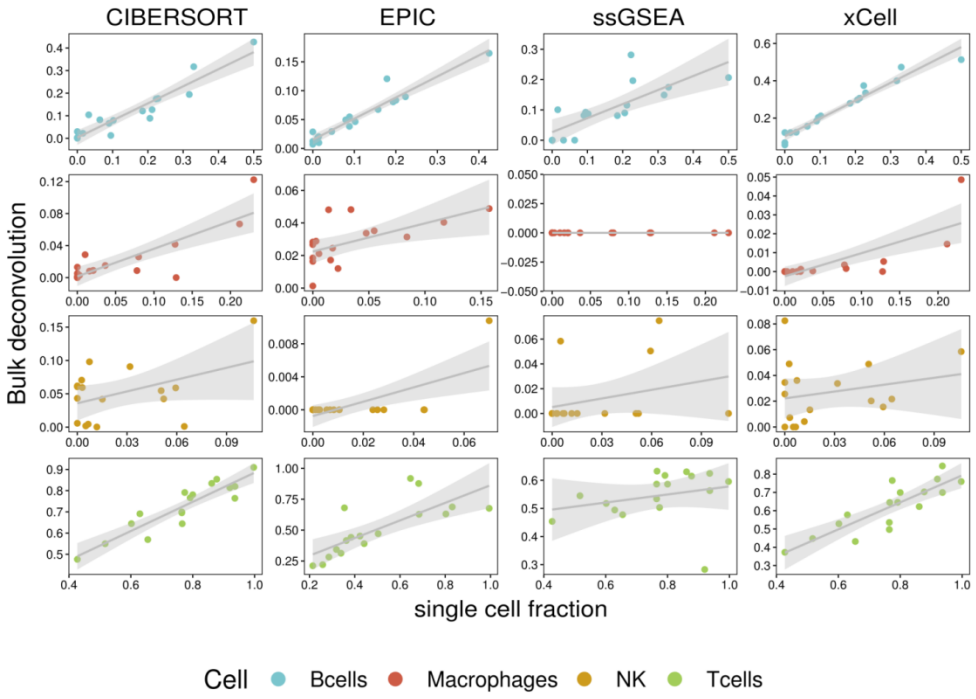


Figure S7. Bulk-single cell correlations in the Melanoma dataset.

9. Scientific products

Research Articles

- Rossi R., Falzarano M.S., Sabatelli P., Antoniel M., **Grilli A.**, Bicciato S., Fang M., Ferlini A., Gualandi F. *Urinary mesenchimal stem cells are a tool to study COLVI transcripts and proteins*. Submitted to Muscle and Nerve, September, 2019.
- Boraldi F, Lofaro FD, Romano O, **Grilli A**, Losi L, Moscarelli P, Bicciato S, Quaglino D. *Exome sequencing and bioinformatic approaches reveals rare sequence variants involved in cell signalling and elastic fibre homeostasis: new evidence in the development of ectopic calcification*. Cell Signal. 2019 Jul;59:131-140. doi: 10.1016/j.cellsig.2019.03.020. Epub 2019 Mar 26. Review. PubMed PMID: 30926389.
- Brummelman J, Mazza EMC, Alvisi G, Colombo FS, **Grilli A**, Mikulak J, Mavilio D, Alloisio M, Ferrari F, Lopci E, Novellis P, Veronesi G, Lugli E. *High-dimensional single cell analysis identifies stem-like cytotoxic CD8(+) T cells infiltrating human tumors*. J Exp Med. 2018 Oct 1;215(10):2520-2535. doi: 10.1084/jem.20180684. Epub 2018 Aug 28. PubMed PMID: 30154266.
- Mancarella C, Pasello M, Ventura S, **Grilli A**, Calzolari L, Toracchio L, Lollini PL, Donati DM, Picci P, Ferrari S, Scotlandi K. *Insulin-Like Growth Factor 2 mRNA-Binding Protein 3 is a Novel Post-Transcriptional Regulator of Ewing Sarcoma Malignancy.* Clin Cancer Res. 2018 Aug 1;24(15):3704-3716. doi: 10.1158/1078-0432.CCR-17-2602. Epub 2018 Apr 27. PubMed PMID: 29703820.
- **Grilli A**, Bengalli R, Longhin E, Capasso L, Proverbio MC, Forcato M, Bicciato S, Gualtieri M, Battaglia C, Camatini M. *Transcriptional profiling of human bronchial epithelial cell BEAS-2B exposed to diesel and biomass ultrafine particles*. BMC Genomics. 2018 Apr 27;19(1):302. doi: 10.1186/s12864-018-4679-9. PubMed PMID: 29703138.

- Risi E, **Grilli A**, Migliaccio I, Biagioni C, McCartney A, Guarducci C, Bonechi M, Benelli M, Vitale S, Biganzoli L, Bicciato S, Di Leo A, Malorni L. *A gene expression signature of Retinoblastoma loss-of-function predicts resistance to neoadjuvant chemotherapy in ER-positive/HER2-positive breast cancer patients.* Breast Cancer Res Treat. 2018 Jul;170(2):329-341. doi: 10.1007/s10549-018-4766-2. Epub 2018 Mar 22. PubMed PMID: 29564743.
- Cortesi F, Delfanti G, **Grilli A**, Calcinotto A, Gorini F, Pucci F, Lucianò R, Grioni M, Recchia A, Benigni F, Briganti A, Salonia A, De Palma M, Bicciato S, Doglioni C, Bellone M, Casorati G, Dellabona P. *Bimodal CD40/Fas-Dependent Crosstalk between iNKT Cells and Tumor-Associated Macrophages Impairs Prostate Cancer Progression.* Cell Rep. 2018 Mar 13;22(11):3006-3020. doi: 10.1016/j.celrep.2018.02.058. PubMed PMID: 29539427.
- Mancarella C, Casanova-Salas I, Calatrava A, García-Flores M, Garofalo C, **Grilli A**, Rubio-Briones J, Scotlandi K, López-Guerrero JA. *Insulin-like growth factor 1 receptor affects the survival of primary prostate cancer patients depending on TMPRSS2-ERG status.* BMC Cancer. 2017 May 25;17(1):367. doi: 10.1186/s12885-017-3356-8. PubMed PMID: 28545426.

Oral communications

- BBCC, 2019, Salerno (Accepted). *ARDESIA: a web app for the Automatic Report of DEconvolution tools by SIngle cell Annotation*. **Grilli A.**, Caroli J, Bicciato S., Battaglia C.
- 3° Workshop BIOMETRA, 2018, LITA, Milano. *A bioinformatic framework to identify cell subpopulations from bulk gene expression data of cancer samples.* **Grilli A**., Bicciato S., Battaglia C.

Posters

- ESHG, 2018, Milan. A *Bioinformatic framework to identify cell subpopulations from bulk gene expression data of cancer samples*. **Grilli A**., Battaglia C., Bicciato S.
- BBCC, 2017, Naples. *A Bioinformatic framework to identify cell subpopulations from bulk gene expression data of cancer samples*. **Grilli A**., Castellano S., Battaglia C., Bicciato S.
- EWOG, 2017, Rome. *Activation of specific transcriptional modules define maturation states in JMML BM and HSCs*. Forcato M., Pizzuto MS., **Grilli A**., Masetti R., Niemeyer CM., Locatelli F., te Kronnie G., Bicciato S., Basso G, Bresolin S.
- ESMO, 2017, Madrid. *Gene signatures as potential predictive markers of response to neoadjuvant chemotherapy in ER+/HER2+ breast cancer patients*. Risi E., **Grilli A**., Migliaccio I., Biagioni C., Guarducci C., Bonechi M., McCarteney A, Vitale S., Laura Biganzoli L, Bicciato S., Di Leo A., Malorni L.
- SABCS, 6-10 dicembre 2016, San Antonio. *A RB-1 loss-of-function gene-signature (RBsig) predicts resistance to neoadjuvant chemotherapy in HER2+/ER+ breast cancer patients*. Risi E., **Grilli A**., Migliaccio I., Biagioni C., Guarducci C., Bonechi M., Hart C.D., Biganzoli L., Bicciato S., Di Leo L., Malorni L.

Foundings

- Research Fellow, AIRC 5x1000 to Prof. Silvio Bicciato, with the project “Molecular basis for triple negative breast cancer metastasis: new tools for diagnosis and therapy”.
- Research Fellow, AIRC 5x1000 to Prof. Silvio Bicciato, with the project “Analisi bioinformatica di dati genomici in campioni ottenuti da pazienti con cancro della mammella triplo-negativo”.
- Research Fellow, Ricerca finalizzata 2016 to Dr. Mattia Forcato with the project “Rigenerative potential of fibro-adipogenic progenitors derived exosomes in the treatment of Duchenne Muscular Dystrophy with deacetylase inhibitors”.

10. Revision

This thesis was evaluated by 2 independent reviewers and 1 anonymous reviewer:

Chiara Romualdi, Ph.D., Professor
Professor in Moecular Biology and Statistics
Dipartimento di Biologia
Università degli studi di Padova
Via U. Bassi 58/b
35131 – Padova (Italy)

Paolo Serafini, Ph.D.
Assistant Professor of microbiology and Immunology
Department of Microbiology & Immunology
UM/Sylvester Comprehensive Cancer Center
Leonard M. Miller School of Medicine
University of Miami
1600 NW 10 Avenue, Miami, Fl

Research integrity declaration

I declare that my Phd project fulfills research integrity good practices:

- Research Environment ✓
- Training, Supervision and Mentoring ✓
- Research Procedures ✓
- Safeguards ✓
- Data Practices and Management ✓
- Collaborative Working ✓
- Publication and Dissemination ✓
- Reviewing, Evaluating and Editing ✓

Acknowledgments

I would like to thank my supervisors, Prof. Silvio Bicciato and Prof. Cristina Battaglia for their continuous support, help and motivations during this project. A sincere thanks to Silvio, since he is a continuous source of inspiration to raise up the level of my work. A grateful thanks to Cristina, for her continuous encouragements and suggestions on this project and also on what's next.

I want to express my sincere appreciation for all members of "AIRC 5x1000" and "Ricerca finalizzata 2016" projects that supported this work. I would like to thank the reviewers for their thoughtful comments and suggestions which helped to improve the quality of the final manuscript.

Thanks to all people from the Bicciato's lab, since for every kind of doubt or problem there's always someone willing to help; thanks to Mattia, he is an essential lifeline to solve any existing or unconceivable bioinformatics problem; thanks to Oriana for her valuable suggestions and help in presentations setting; thanks to Martina, for her support in writing this thesis and for introducing me in the fantastic world of vector images; a special thanks to Jimmy, for the unceasing, essential and invaluable help during the writing of this thesis and for his crucial contribution in the creation of ARDESIA, further to be a continuous source of surprises and laughs.

Also, thanks to Leonardo, for his constant smile and for his infinite curiosity for every image and color of this thesis, and for everything in the surrounding world; to Alessandro, for his great, infinite patience for constantly postponing games together while I was finishing this, to him endless, personal aim. Especially, I want to thank Elisa, she knows the start of this long journey and she is next to me since that moment; I would have not reached this exceptional objective without her help.