Loss-Function Learning for Digital Tissue Deconvolution

Dissertation zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Biologie und Vorklinische Medizin
der Universität Regensburg

vorgelegt von

Franziska Görtler

aus
Regensburg

im Jahr 2019

Loss-Function Learning for Digital Tissue Deconvolution

Dissertation zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Biologie und Vorklinische Medizin
der Universität Regensburg

vorgelegt von

Franziska Görtler

aus
Regensburg

im Jahr 2019

Das Promotionsgesuch wurde eingereicht am: 27.09.2019

Die Arbeit wurde angeleitet von Prof. Dr. Rainer Spang.

Unterschrift: Regensburg, den 27.09.2019

Publications

Parts of this thesis have been published in the proceedings of RECOMB 2018 [1]. This includes the abstract, the notations and the mathematical loss-function learning description in section 2.1 and 2.2 of chapter 2. Also parts of the results of the melanoma data set in chapter 3 were published as well as their discussion in chapter 5.

Danksagung

Mein Dank gilt meinem Doktorvater Rainer Spang sowie meinem Kollegen Michael Altenbuchinger für die Unterstützung bei der Durchführung der Doktorarbeit. Sowie meinen Kollegen fr die vielen fachlichen und manchmal vielleicht nicht ganz so fachlichen Gespräche am Lehrstuhl fr funktionelle Genomik.

Ebenso bedanken möchte ich mich bei Christian Schmiedl vom RCI Regensburg fr das zur Verfüngung stellen des CLL-Datensatzes.

Contents

Abstract

The gene expression profile of a tissue averages the expression profiles of all cells in this tissue. Digital tissue deconvolution (DTD) addresses the following inverse problem: Given the expression profile y of a tissue, what is the cellular composition c of that tissue? If X is a matrix whose columns are reference profiles of individual cell types, the composition c can be computed by minimizing $\mathcal{L}(y - Xc)$ for a given loss function $\mathcal{L}$. Current methods use predefined all-purpose loss functions. They successfully quantify the dominating cells of a tissue, while often falling short in detecting small cell populations.

In this here presented, newly developed approach training data are employed in order to learn the loss function $\mathcal{L}$ along with the composition c. This allows for adaption of the loss function to application-specific requirements, such as focusing on small cell populations or distinguishing phenotypically similar cell populations.

Loss-function learning is tested on two different single-cell RNA sequencing data sets. The first is generated from melanoma specimens and the second from peripheral blood samples of patients with Chronic Lymphocytic Leukemia (CLL). The CLL data were augmented by bulk sequencing data. It could be demonstrated that the here introduced method quantifies large cell fractions as accurately as existing methods and significantly improves the detection of small cell populations and the distinction of similar cell types. Furthermore, it is shown that the developed DTD models may be applied mutually to both sets of data. As a result the model on the melanoma data is also relevant for the CLL data set and vice versa.

Einleitung

In der Medizin wird das bösartige unkrontrollierte Vermehren und Wuchern von Zellen als Krebs bezeichnet. Bösartig heißt, dass es neben der Ausbildung des Primärtumors zur Streuung und somit Bildung von Metastasen kommt. Die Häufigkeit des Befalls der einzelnen Organe ist abhängig von Faktoren wie Alter, Geschlecht, Region und Lebenswandel. In Deutschland ist Krebs die zweithäufigste Todesursache nach Herz-Kreislauf-Erkrankungen. Wird rechtzeitig eine Therapie begonnen, oder tritt ein langsam verlaufender Krebs erst in hohem Lebensalter auf, so muss der Verlauf nicht tödlich sein. Die relativen 5-Jahres-Überlebensraten über alle Krebsarten in Deutschland betrugen 2017 65% bei Frauen und 59% bei Männern [2].

Besonders erbgutbeeinflussende Faktoren sind krebsserregend, da hier die Mutationen in alle nachfolgenden Tochterzellen weitergetragen werden. Während der Zellteilung ist die Zelle besonders anfällig für Mutationen, deshalb sind sich schnell teilende Zellen häufiger von Kreps betroffen. Die meisten Krebsarten (90-95% der Fälle) werden durch Umweltfaktoren ausgelöst [3]. Diese sind Umweltgifte und radioaktive, Röntgen- oder UV-Strahlung, die auch bei Untersuchungsmethoden wie CT-Scans [4] auftritt. Daneben gibt es biologische und therapeutische Einflüsse wie Onkoviren [5], Stammzelltherapie [6] sowie immunsuppressive Therapien nach Organtransplantation [7]. Ebenso haben die Lebensumstände und der Lebensstil einen großen Einfluss auf die Entstehung von Tumoren. Dabei handelt es sich beispielsweise um Übergewicht [8, 9], Tabak- sowie Alkoholkonsum.

Tumore bestehen nicht nur aus den entarteten Krebszellen sondern enthalten Blutgefäße zur Versorgung sowie Immunzellen. Die Zusammensetzung dieser Immunzellen ist abhänig von der Art des Tumors sowie dem Patienten. Zwischen Immun- und Tumorzellen gibt es komplexe Wechselwirkungen [10], diese haben Einfluss auf den Verlauf der Erkrankung [11] sowie die Heilungschancen [12]. Ebenso können die vorkommenden Immunzellen zur Immuntherapie der Tumore verwendet werden [13–15]. Krebszellen tarnen sich gegenüber den Immunzellen und werden von diesen somit nicht mehr erkannt. Schafft man es, diese Blockade zu lösen und das Immunsystem zu stimulieren, so ist es diesem wieder möglich, die Tumorzellen zu erkennen und zu vernichten [16, 17].

Es spielt eine Rolle, welche Immunzellen sich im und um den Tumor aufhalten, und in welcher Menge sie vorkommen. Übliche Methoden um diese Frage zu beantworten sind beispielsweise Immunhistochemie oder fluoreszensz aktivierte Zellsortierung (fluorescence-activated cell sorting = FACS). Bei der Immunhistochemie [18] werden Proteine oder andere Strukturen in Gewebe mit Hilfe von Antikörpern sichtbar gemacht. Tumorzellen können so identifiziert und klassifiziert werden, da in diesen bestimmte, nachweisbare Antigene exprimiert sind. So können Therapien bei morphologisch gleich erscheinenden Tumoren auf deren tatsächliche Tumoreigenschaften angepasst werden. Bei FACS werden die Zellen einer Probe analysiert, indem sie einzeln mit hoher Geschwindigkeit an einem Lichtstrahl oder einer elektrischen Spannung vorbeigeleitet werden. Dabei werden unterschiedliche Effekte erzeugt, abhängig von Form, Struktur und Zellfärbung, aus welchen die Zelleigenschaften abgeleitet werden.

Weitere Verfahren sind Einzelzell-RNA-Sequenzierung [19], Massenspektrometrie [20] und PT-PCR [19].

Neuere Methoden wie gene set enrichment analysis (GSEA) oder digital tissue deconvolution (DTD) sind computergestützt. GSEA [21, 22] ist eine Methode um Gen- oder Proteinklassen zu identifizieren, welche in einer großen Anzahl von Genen oder Proteinen über- oder unterrepräsentiert sind.

In dieser Arbeit stellen wir eine Methode zur DTD vor. Dabei werden anhand von Einzelzellmessungen diejenigen Gene bestimmt, welche bei der Dekonvolution des untersuchten Gewebes die optimalen Ergebnisse erzielen. Der große Vorteil ist, dass, so diese Gene einmal bestimmt sind, sie zur Dekonvolution von Bulk-Messungen verwendet werden können. Hierzu existieren viele verschiedene Algorithmen, einige davon werden in den Kapiteln 1.2.2 und 1.2.3 beschrieben. Die Verwendung von aus Einzelzellmessungen definierten Gensets zur Dekonvolution ist ein großer Vorteil, da Bulk-Messungen im Vergleich zu Einzelzellmessungen deutlich kostengünstiger sind. Bei einigen DTD Methoden werden Referenzprofile der zu untersuchenden Zelltypen verwendet, bei anderen nicht. Diese können ebenso aus den Einzelzellmessungen gewonnen werden. Die hier vorgestellte Methode zur Digital Tissue Deconvolution [1] gehört zu den ersteren Verfahren. Sie verwendet jedoch im Unterschied zu anderen Methoden zusätzlich zu Referenzprofilen und Einzelzellmessungen noch die Zellzusammensetzung bekannter Mischungen um die für die Dekonvolution aussagekräftigsten Biomarker zu bestimmen. Im Gegensatz zu anderen Methoden werden diese Gene je nach betrachteten Immunzelltypen algorithmisch bestimmt und nicht aufgrund von biologischem oder medizinischem Vorwissen. Damit ist diese Methode zur Bestimmung des Immunzellgehaltes von Proben einerseits sehr variabel andererseits sehr und anpassungsfähig, z.B. an die jeweiligen Zelltypen von Interesse. Der Nachteil dieser Methode ist, dass hierfür immer Daten zum Lernen notwendig sind, so z.B. von single-cell Sequenzierungen.

In der vorliegenden Arbeit werden im ersten Teil (Kapitel 1) die biologischen Grundlagen erklärt sowie etablierte und neue Methoden zur Bestimmung von zellulären Zusammensetzungen vorgestellt. Anschließend wird die Methode der Digital Tissue Deconvolution mathematisch beschrieben und numerische Simulationen dazu durchgeführt (Kapitel 2). Anhand zweier Datensets wird gezeigt, dass das beschriebene Verfahren zur Detektion der Immunzelltypen geeignet ist. Es wird zuerst ein Datenset aus Einzelzellmessungen von 19 Melanomen betrachtet (Kapitel 3). Beim zweiten Datenset handelt es sich um Einzelzellmessungen zu verschiedenen Zeitpunkten der Therapie bei vier Patienten mit chronischer lymphatischer Leukämie (Kapitel 4). Zudem wird für beide Datensets die vorgestellte Methode mit der aktuell führenden Methode in diesem Bereich, CIBERSORT [23], verglichen. In allen Vergleichen wurden bessere Resultate erzielt.

Introduction

In medicine, cancer is defined as a malign and rampant proliferation of cells. The term “malign” expresses that besides the development of a primary tumor there is also a dissemination of cells which leds to metastases. How often the individual organs are affected by this disease depends on factors like age, sex, residence and lifestyle. In Germany cancer is the second most common cause of death following cardiovascular diseases. If therapy is started early enough, or if the cancer is of a kind that progresses slowly and occurs in old age, cancer does not necessarily have to be deadly. The average five-year survival rates in 2017 across all cancer types in Germany were 65% for women and 59% for men [2].

Particular factor for inducing a carcinogenic progress are cell mutations which are passed on to all following daughter cells. During cell division the cell is especially vulnerable to mutations, so cells that multiply fast and often are affected more easily than other cells. Most cancer types (90-95% of all cases) are triggered by environmental factors [3], such as pollutants but also X-rays or UV-rays which are used for survey methods like CT-scans [4]. Furthermore there are also biological and therapeutic influences like oncoviruses [5], stem-cell therapy [6] as well as immunosuppressive therapies after organ transplantation [7]. Also environment and lifestyle factors contribute to tumor formation. These factors can be obesity [8, 9], tobacco and alcohol consumption.

Tumors not only consist of degenerated cancer cells but also contain blood vessels for supply of nourishing substances and immune cells. The particular composition of these immune cells depends on the tumor type and on the individual patient. There are complex interactions between immune and tumor cells [10], which influence the course of the disease [11] and the prospects of treatment [12]. The present immune cells can also be used for immunotherapy of the tumors. Cancer cells camouflage themselves against the immune cells and are thus no longer recognized by them. If it is possible to lift this mimicry and to stimulate the immune system, it is possible for the present immune dells to recognize and destroy the malignant cells [16, 17].

Here it matters which immune cells are current in and around the tumor, and in which quantity they are present. Common methods to answer these questions are for example immunohistochemistry or fluorescence-activated cell sorting (fluorescence-activated cell sorting = FACS). In immunohistochemistry [18] proteins or other structures in the tissue are visualized by means of antibodies. Tumor cells can thereby be identified and classified because they express certain detectable antigens. As a result therapies can be adapted to the actual tumor properties in tumors which have identical morphology. When using FACS the cells of a probe are analyzed by passing them one at a time through a light beam or an electrical voltage with high velocity. Different effects are produced, depending on shape, structure and cell dyeing. from the recorded specifics individual cell properties are derived.

Other methods are single-cell RNA sequencing [19], mass spectrometry [20] and PT-PCR [19].

Newer methods such as gene set enrichment analysis (GSEA) or digital tissue deconvolution (DTD) are fully computationally generated. GSEA [21, 22] is a method to identify gene or protein classes which are over- or underrepresented in a large number of genes or proteins.

In this thesis an advanced method for DTD is introduced. Based on single cell measurements genes which achieve the optimal results in the deconvolution of the examined tissue will be determined. A major benefit is that, once genes are determined, they can be used for deconvolution of bulk measurements. To this end there already exist many different algorithms, some of them are described in the chapters 1.2.2 and 1.2.3. The here presented approach uses data sets from expensive single cell measurements to train the method for application to the cost efficient bulk measurements. Some deconvolution methods use reference profiles of the cell types under study, others do not. These reference profiles can also be obtained from the single cell measurements. Our presented method for digital tissue deconvolution [1] is part of the methods mentioned first. However, in contrast to other methods it uses in addition to reference profiles and single cell measurements also mixtures with known cellular composition to determine the most relevant biomarkers for deconvolution. Unlike in other methods, the genes relevant for deconvolution are determined algorithmically, only depending on the immune cell types considered and not on the basis of prior biological or medical knowledge. Hence, our method for determining the immune cell content of samples, is quite variable and adaptable to cell types of particular interest. However, additional data with known cell composition is needed for determining the significant biomarkers.

In the present thesis the first part (chapter 1) is devoted to the technical biological terms as well as established and new methods for the determination of cellular compositions. Subsequently, the method of digital tissue deconvolution is described mathematically and numerical simulations for it are conducted (chapter 2). Based on two data sets, it is shown that the described method is suitable for the detection of various immune cell types. First, a data set with single cell measurements of 19 melanoma specimens is considered (chapter 3). A second data set consists of single cell measurements of chronic lymphocytic leukemia from four patients at different points in time (chapter 4). For both data sets the results of our newly established method is compared to the current state of the art method CIBERSORT [23]. In all cases superior results are produced.

Chapter 1

Biological and Algorithmic Basics

This chapter deals with the biological background of tumor infiltrating immune cells and the mathematical basics of digital tissue deconvolution. Section 1.1 gives an introduction to tumor infiltrating immune cells. Section 1.2 outlines the mathematical and algorithmic basics of Digital Tissue Deconvolution (DTD) and gives an overview of available deconvolution algorithms.

1.1 Tumor Infiltrating Immune Cells

The immune system is the biological defense system of higher life forms. It is a complex system constituted by sophisticated interplay of organs, cell types and molecules. Its function is to prevent tissue damage caused by pathogens, to eliminate alien substances, excrete microorganisms infiltrating the body, and to destroy defected body cells. There are two different mechanisms in the immune defense. On one hand there is the innate immunity which needs no training by pathogens. The reaction of the innate immune system occurs within minutes and is defined in the genetic information. On the other hand is the adaptive immunity. Here, the immune defense is acquired and hence specific for the pathogen. The adaptive immunity is characterized by the flexibility to adapt to new or altered pathogens. After initial contact For a complex immune response both parts of the immune system are necessary [24].

A major part of the immune system are the leukocytes, or white blood cells. In this work monocytes and macrophages which are part of the innate immunity are considered. These cells are scavenger cells as they absorb extraneous material and dispose it. After appropriate stimulation B cells produce specialized antibodies in order to defeat certain pathogens or other harmful substances. T cells as part of the adaptive immune system mediate between innate and adaptive immunity. Natural killer cells (NK cells) are part of the innate immune system as they do not have antigen-specific receptors. They detect and kill tumor and virus infected cells [25]. Endothelial cells participate in innate and adaptive immune response. They function as detectors of foreign pathogens and inflammatory processes and mobilize other immune-cell types like monocytes, macrophages and T cells [26].

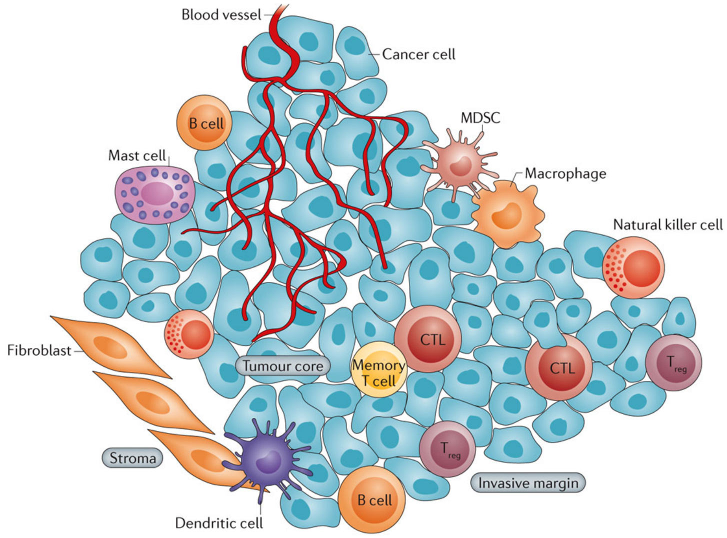


Figure 1.1: Here tumor cells (blue) are infiltrated by immune cells like B cells, mast cells, T cells (CTL, memory T cells, Tregs), natural killer (NK) cells and others. Blood vessels which provide nutrients for the tumor are shown in red. Picture from [12].

Immune cell types can be influenced by tumors. Cancer tissue needs immune cells to communicate with the surrounding immune system to keep it in check. In the beginning these immune cells try to limit tumor growth, but later they get inactivated or even help the tumor to grow. These cells originate from immunological cells. An example for such cells are the cancer associated fibroblasts (CAFs). CAFs are involved in the chronic inflammation of cancerous tissue which supports the tumor. Another sort of tumor supporting immune cells are the nurse like cells (NLC). These cells promote the survival of CLL lymphocytes by production of chemokines of antiapoptotic activity and they promote the expression of adhesion molecules [27].

The human body is composed of different tissues, which are characterized by different cellular compositions. In cancer cells the normal process of growth, cell division and apoptosis is altered. Thus, the cellular composition in tumor tissue differs from that in normal tissue. Tumor tissue additionally is infiltrated by immune cells and blood vessels. Immune cells and tumor cells interact within a complex network [10], dependent on the specific tumor type. Figure 1.1 visualizes a tumor tissue (blue), where tumor and immune cells interact with each other.

Tumor infiltrating immune cells or their composition affect disease progression [11] or treatment success [12]. Moreover, small subpopulations can be potential targets for immunotherapy. The immune cells interact with the tumor and some of these interactions even support the tumor. If they can be blocked by immunotherapies, immune cells can fight the cancerous cells and kill them. Thus, the success of treatment also depends on the presence, quantity, and molecular sub-type of the infiltrating immune cells [28]. There are several methods to estimate the cellular composition of tumor tissue. Fluorescence-activated cell sorting (FACS; e.g. [29]), cytometry by time-of-flight (CyTOF; e.g. [30]), and single-cell RNA sequencing [31] are common techniques.

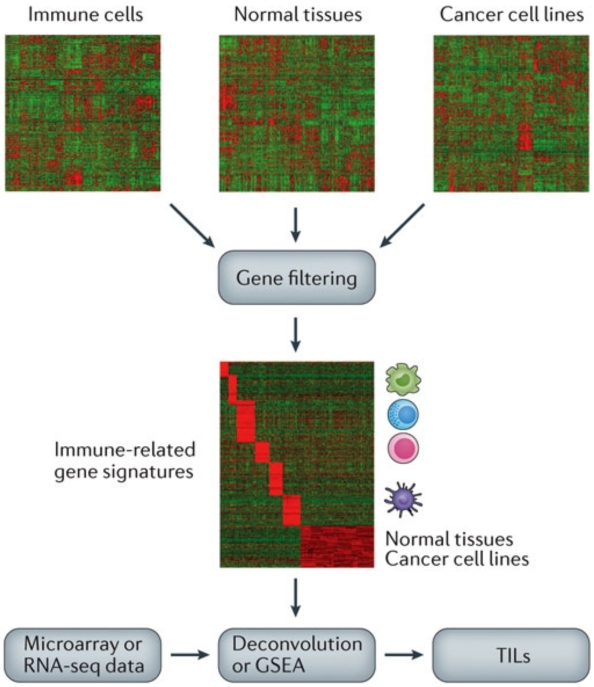


Figure 1.2: In gene set enrichment analysis (GSEA) one gets the gene signals related to the different immune cells from expression profiles of the different immune cell types of interest (TILs). Picture from [28].

1.2 Algorithmic basics

Several computational tools to predict the amount of immune cells in cancer tissue are already available. They can be categorized as supervised and unsupervised methods. Supervised methods use reference profiles of a set of preselected cell types. Unsupervised methods [32] can be applied without prior knowledge. Common input for DTD are gene expression data as well as methylation data, from sequencing [33, 34] and microarray technology [35, 36].

An alternative are gene-set enrichment methods [37, 38]. Here, the aim is to find statistically enriched genes which are involved in a pathway of interest or a certain cellular process [39]. Yet another method are single-sample approaches in which genes are ranked by the differential expression between two different biological conditions. Finally, one tries to estimate the bulk measurement by adjusting the content of the different cell types. Supervised DTD methods require cell-type-specific reference profiles. These algorithms solve an inverse problem as associated with gene set enrichment analysis (GSEA). Their aim is to provide the most accurate estimate for the cellular composition and not to give the best prediction of the bulk profile.

1.2.1 Mathematical Basics and Limitations

The bulk profile y is constructed by RNA-seq or methylation data values for all considered genes. In the columns of reference matrix X, the reference profiles for the cell types of interest are stored.

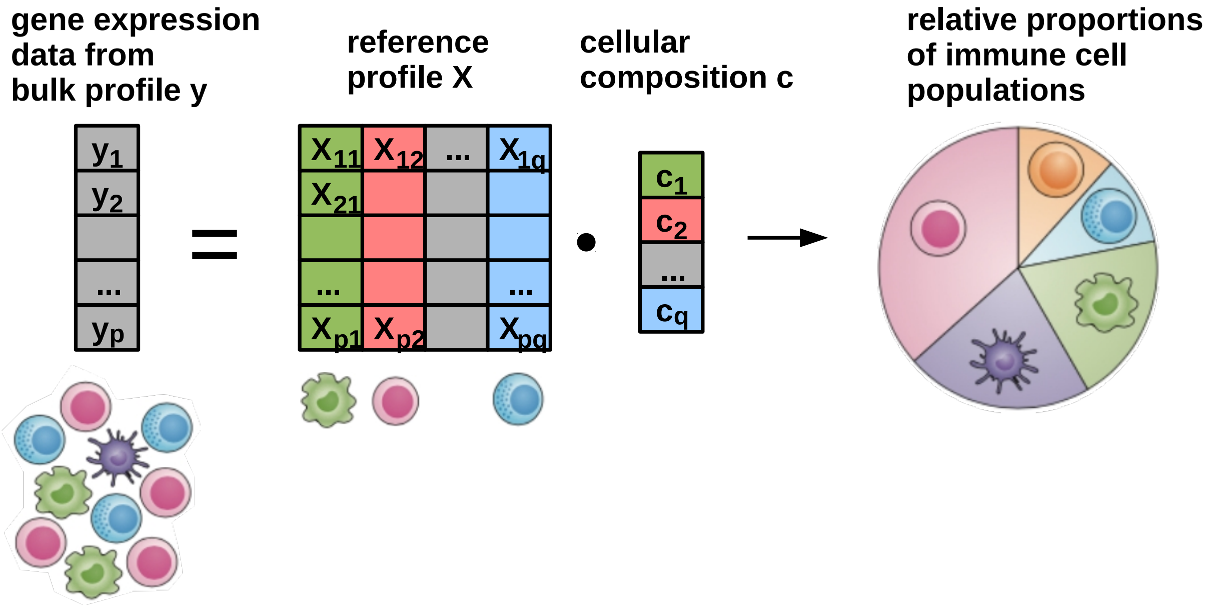


Figure 1.3: Deconvolution methods use the bulk profile y and the reference matrix X from the regarded cell types. With this information one can calculate the composition c of the immune cells in the bulk. Picture from [28].

The gene counts for the individual cell profiles are found in the rows of the matrix X. With given Y and X the cellular composition C of the specific cell type is estimated and one obtains the relative immune cell proportions. The bulk Y is a linear combination of the reference profiles in X. Figure 1.3 is an illustration of the problem. Mathematically, Figure 1.3 can be written as

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,q} \\ X_{2,1} & X_{2,2} & \dots & X_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,1} & X_{p,2} & \dots & X_{p,q} \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_q \end{pmatrix} = X \cdot c. \tag{1.1}$$

The naive solution would be

$$\underset{c}{\operatorname{argmin}} ||y - Xc||_2^2. \tag{1.2}$$

To get better deconvolution results, the naive solution in equation 1.2 is replaced by a given loss function $\mathcal{L}$. In order to calculate the cellular composition C of the bulk profiles Y, the loss function $\mathcal{L}(y - Xc)$ is minimized. The different competing DTD methods use different algorithms to find the best composition C of the profiles Y for a given predefined all-purpose loss function $\mathcal{L}$.

If the bulk profiles Y were exact mixtures of the reference profiles contained in X the existing deconvolution methods would work perfectly and for the true cellular distribution C the result of $Y - XC$ would be zero. However, the bulk profiles in Y are not exact mixtures, which causes several problems:

(1) **It is hard to quantify small cellular fractions.** In tumors the cell populations of the immune cells are mostly small. However the reaction of a tumor to immunotherapy may be determined by them. Therefore it is important to reflect the faint signals coming from the small cell populations with an appropriate weight in the DTD algorithm. This aspect allows for major approvements in the calculations an is the most sensitive adjustment tool.

(2) **Potential incompleteness of the reference profile collections.** Some cells in the analyzed tissue might not be covered by the reference profiles. This results in a not solvable global DTD problem. This issue is treated by increasing the frequency of the other cell types in order to compensate for the contributions of the not covered cells in the DTD-algorithm. In other words, if a reference profile is missing, the algorithm will overestimate another profile instead.

(3) **Similar expression profiles of two different cell types are hard to distinguish.** For not related immune cells the expression profiles differ greatly and they are easier to quantify. But for immunological sub-entities of a cell-type the differences between the corresponding reference profiles are more subtle. The distinction of two cell types becomes more difficult with the similarity of their cellular profiles.

To summarize, for different applications there are different approaches necessary. This can be done by adapting the loss function $\mathcal{L}$ to the specific problem. In the end, the aim is always to focus on a predefined gene set which is most helpful to deconvolute the cell types of interest. DTD results depend strongly on the gene set. For example distinguishing between immune cells based on a set of genes if those are not expressed is not possible. These genes are then dominating in the loss function $\mathcal{L}$. When using a helpful gene set even small cell populations with faint signals can be deconvoluted correctly. The problem is that it is not clear a priori, which genes are important for deconvolution and which ones are better to be ignored.

1.2.2 Historical Development of DTD and Early Works

IRIS One of the first attempts to digital tissue deconvolution was IRIS, immune response in silico [40]. A compendium of microarray expression data of six immune cell types, either in activated or differentiated states of major non-immune tissues, were used to identify immune-specific genes. A gene is defined as immune-cell specific, if its expression value in the immune-cell profile is higher than in any major organ tissue. IRIS groups genes within a cell type, or if they are specific for more than one cell type, in a lineage. For the statistical clustering an unweighted average method is used. In microarray experiments, the genes found by IRIS can be used as cell markers for analysis.

DeconRNASeq DeconRNASeq is a more recent deconvolution solution for mRNA-Seq data. It is provided by Gong et al. [41]. The algorithm uses a linear model of reference profiles. The cellular compositions are used as weights which are preserved to be non-negative. For calculating the proportions of specific cell types in a sample, a non-negative least-squares constraint problem is solved. For obtaining the global optimum in the solution quadratic programming is used. The method can also predict missing fractions in the bulk sample. The problem is given by

$$\min_{\mathrm{C}}(||CX - Y||^2), \qquad \text{such that} \begin{cases} \sum_i c_{ki} = 1 \\ c_{ki} \geq 0, \forall i \end{cases}, \tag{1.3}$$

where the matrix Y contains the bulk samples, X the normalized transcriptional measurements from pure tissues and C the proportions of the tissues over the samples.

1.2.3 State of the Art Deconvolution Methods

TIMER was developed to study the interactions of tumor-infiltrating immune cells with the surrounding cancer cells [42]. In Figure 1.4 the deconvolution workflow is shown. First, several preprocessing steps are carried out, e.g., the sample purity is calculated. The surrogate for tumor purity is the fraction of aneuploid cells (a). Then, batch effects need to be removed (b). Next, genes which are negatively correlated with tumor purity are selected. Those, which have expression levels strongly affected by the purity of the tumor are tested for immune signature enrichment (c). Next, the top 1% of the strongest expressed genes are removed, since they dominate the inference of results (d). Finally the deconvolution of a mixture Y is calculated by (e):

$$f = \underset{\forall r: f_r > 0}{\operatorname{argmin}} \sum_{g \in \{G_{Of}\}} \left(Y^g - \sum_{\text{all cell types } r} f_r X_r^g \right), \tag{1.4}$$

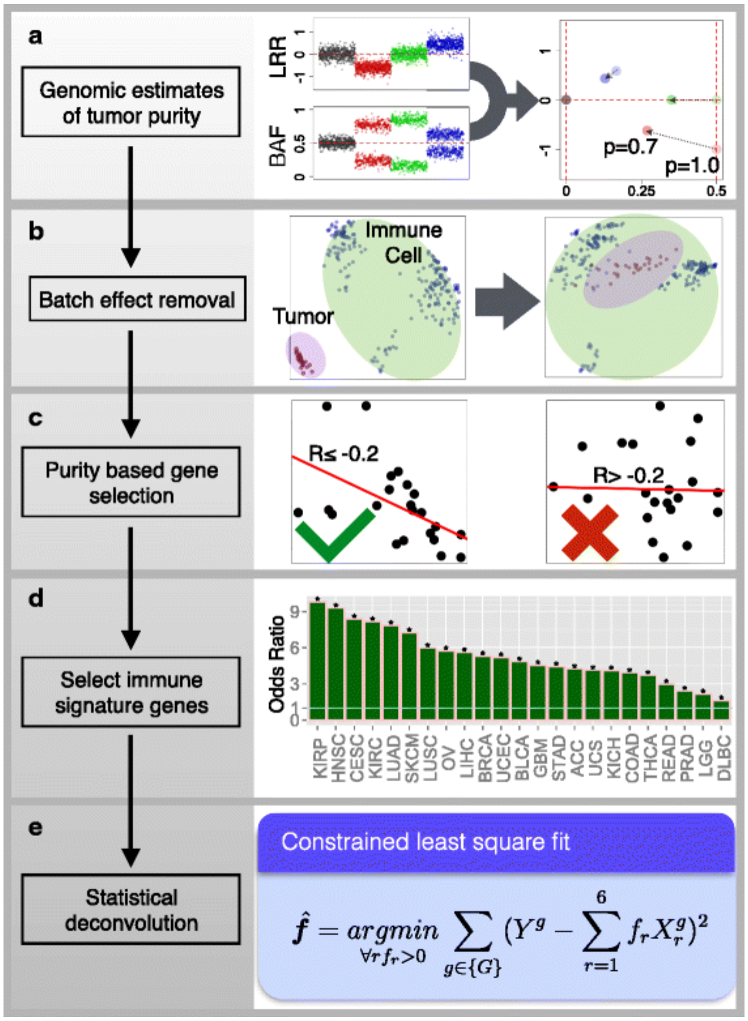
where Y^g is the gene expression of gene g, with $g \in G_{Of}$. Here G_{Of} collects the genes which remain after the filtering steps. The expression of gene g in cell type r is given by X_r^g and f_r is the amount of cell type r in the mixture Y.

CIBERSORT A second state-of-the-art method is CIBERSORT [43]. Here, a linear support vector regression (SVR) algorithm with adaptive feature selection is used to devonvolute bulk mixtures. Figure 1.5 gives an overview of the deconvolution process.

CIBERSORT needs a matrix of gene expression profiles X and a bulk gene-expression profile Y as an input. The cell fractions C are defined through

$$Y = XC. \tag{1.5}$$

CIBERSORT uses ν-support vector regression, which is an optimization method for binary classification problems. Here two classes are separated by a hyperplane with maximal margins. The



$$\hat{f} = \underset{\forall r f_r > 0}{argmin} \sum_{g \in \{G\}} (Y^g - \sum_{r=1}^{6} f_r X_r^g)^2$$

Figure 1.4: The five steps of TIMER for calculating the distribution of immune cells in a bulk sample. (a) calculation of the tumor purity and removing batch effects (b). Gene selection using tumor purity (c). Removing of the strongest expressed genes (d). Finally the cellular proportions are estimated by a constrained linear regression problem (e). Picture from [42]

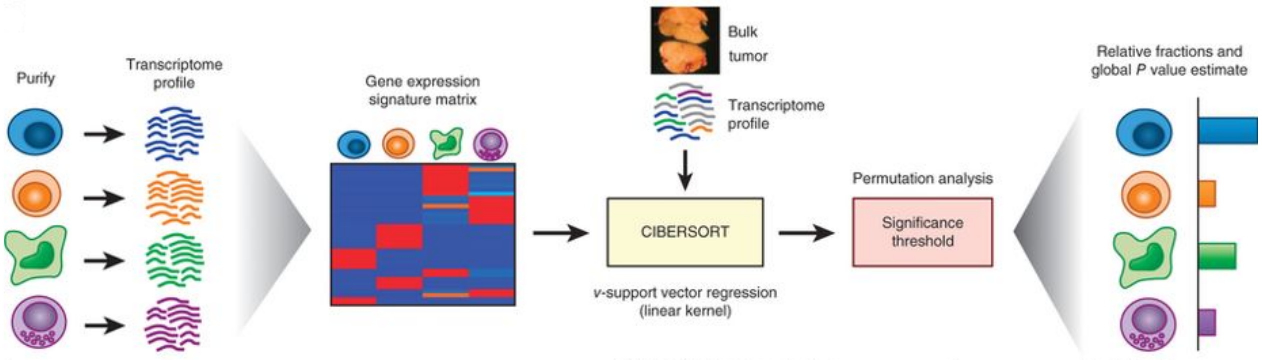

Purify
Transcriptome profile
Gene expression signature matrix
Bulk tumor
Transcriptome profile
CIBERSORT
ν-support vector regression (linear kernel)
Permutation analysis
Significance threshold
Relative fractions and global P value estimate

Figure 1.5: Schematic representation of CIBERSORT. Transcriptome profiles of purified cells are used to construct the gene expression signature matrix X. With this, the transcriptome profile of a tumor bulk Y is deconvoluted by a ν-support vector regression algorithm. After significance thresholding, the relative cell-type fractions C are returned. Picture from [43].

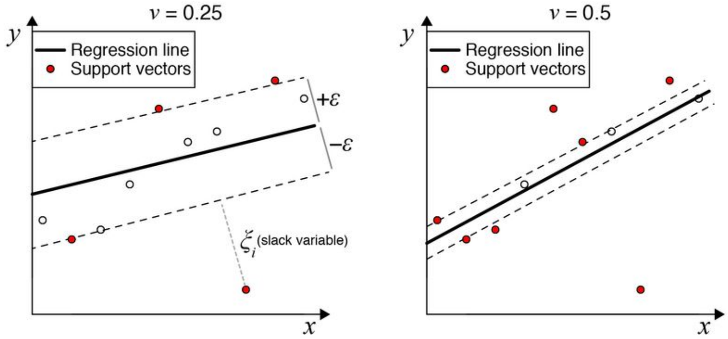


Figure 1.6: Visualization of ν-vector regression for two different choices of ν. The solid black line represents the regression line and the red points the support vectors. Picture from [43].

hyperplane boundaries are determined by a subset constituted by the input data. This subset supplements the vectors. Support vector regression fits a hyperplane with as many data points as possible within a distance ϵ. The points outside the ϵ environment of the regression line are the support vectors shown as red points in Figure 1.6. The algorithm minimizes a linear combination of two functions:

1. The loss function which measures the error associated with data fitting by a linear ϵ-insensitive loss function.

2. The penalty function that determines the complexity of the model. Here it is a L_2-norm penalty.

The resulting cell-type proportions C are normalized to a sum of one.

Chapter 2

Methods

In this chapter the mathematical background of loss-function learning for digital tissue deconvolution is presented. Section 2.1 introduces the used notation. Section 2.2 gives the mathematical background of loss-function learning. In section 2.3, it is proved that the corresponding optimization problem is non-convex. Finally, a line search algorithm with adaptive step size is presented in section 2.4 and it gets shown that this algorithm is able to find the most informative genes for deconvolution in a simulation study, as seen in section 2.5.

2.1 Notations

Let $X \in \mathbb{R}^{p \times q}$ be a matrix with cellular reference profiles $X_{\cdot,j}$ in its columns, where the dot stands for all row indices. X_{ij} is the reference expression value of gene i in cells of type j, p the number of genes, and q the number of cell types in X, respectively. Further a matrix $Y \in \mathbb{R}^{p \times n}$ with bulk profiles of n cell mixtures $Y_{\cdot,k}$ in its columns and a matrix $C \in \mathbb{R}^{q \times n}$ with the cellular compositions of the mixtures $C_{\cdot,k}$ as columns is introduced.

2.2 Loss-Function Learning

Following the established linear DTD algorithms, the mixture $Y_{\cdot,k}$ is approximated by a linear combination of reference profiles (the columns of X) with $C_{\cdot,k}$ as weights and the composition of the k-th mixture $C_{\cdot,k}$ is estimated by minimizing

$$\mathcal{L}_g(Y_{\cdot,k} - XC_{\cdot,k})\,, \tag{2.1}$$

where

$$\mathcal{L}_g = ||\text{diag}(g)(Y_{\cdot,k} - XC_{\cdot,k})||_2^2\,. \tag{2.2}$$

In contrast to standard DTD algorithms, which determine g by prior knowledge or separate statistical analysis, g is learned directly from data. To this end it is assumed that a training set of mixtures

$Y_{\cdot,k}$ from a specific application context with known cellular proportions $C_{\cdot,k}$ with sum one exists. The entries of g are the gene weights that define the loss function. The aim of the here presented algorithm is to learn g from the training data such that minimizing $\mathcal{L}_g(y - Xc)$ with respect to c yields accurate quantifications of cell populations for future samples with similar characteristics as those used for training.

The newly developed method has two nested objective functions: An outer function $L(g)$ and an inner function $\mathcal{L}_g$, which is here given by equation (2.2). L evaluates discrepancies between the estimated and the true cellular frequencies of cell types across samples by Pearson correlation:

$$L(g) = -\sum_{j=1}^{q} \mathrm{cor}(C_{j,\cdot}, \hat{C}_{j,\cdot}(g)) \quad \text{subject to } g_i \geq 0 \,\text{and } ||g||_2 = 1\,, \tag{2.3}$$

where the $\hat{C}_{j,\cdot}(g)$ are the estimates of $C_{j,\cdot}$ given g. To evaluate $L(g)$ it is necessary to calculate all $\hat{C}_{j,\cdot}(g)$, which requires optimising $\mathcal{L}_g$ with respect to all $C_{\cdot,k}$. Note that if $\hat{g}$ is a minimum of L, so is $\alpha\hat{g}$ for $\alpha > 0$. The constraint $||g||_2 = 1$ is thus needed to ensure unique solutions. Note that

$$\mathrm{cor}(C_{j,\cdot}, a_j\hat{C}_{j,\cdot}) = \mathrm{cor}(C_{j,\cdot}, \hat{C}_{j,\cdot})\,, \tag{2.4}$$

where a_j is an arbitrary positive constant. This symmetry is important, since bulk and reference profiles must be normalized to a common mean across genes or to a common library size. A normalized reference profile $X_{\cdot,j}$ of a cell type reflects the true RNA content $\tilde{X}_{\cdot,j}$ of these cells only up to an unknown factor: $X_{\cdot,j} = \alpha_j\tilde{X}_{\cdot,j}$. Large cells with a lot of RNA have smaller α_j than smaller cells. The same is true for the bulk profiles $Y_{\cdot,k}$, where the relation $Y_{\cdot,k} = \beta_k\tilde{Y}_{\cdot,k}$ holds. The deconvolution equation

$$\tilde{Y}_{\cdot,k} = \tilde{X}\tilde{C}_{\cdot,k} + \epsilon \tag{2.5}$$

yields estimates for $\tilde{C}_{jk}$ that reflect the number of cells of type j. However, $\tilde{Y}$ and $\tilde{X}$ are not observable in practice and consequently, $\tilde{C}$ is not accessible by DTD directly. So, one needs to work with X and Y instead.

Note that $C_{\cdot,k} = \tilde{C}_{\cdot,k}/\sum_{j=1}^{q}\tilde{C}_{jk}$. Consider now the hypothetical deconvolution formula with normalized Y but the unobservable true $\tilde{X}$

$$Y_{\cdot,k} = \tilde{X}C'_{\cdot,k} + \epsilon\,. \tag{2.6}$$

Here, it is assumed $C'_{\cdot,k} = c\,C_{\cdot,k}$ for all k, where c is a positive constant. In other words it is assumes that if the library size of $Y_{\cdot,k}$ is the same for all samples, roughly the same number of cells are needed to account for it. This assumption allows to replace $\tilde{Y}$ by Y.

The choice of the correlation in the definition of $L(g)$ also allows to replace $\tilde{X}$ by X. If Eq. (2.6) is written using X,

$$Y_{\cdot,k} = \sum_{j=1}^{q} \frac{1}{\alpha_j} X_{\cdot,j} C'_{jk} + \epsilon \tag{2.7}$$

is obtained Thus, the estimated cell frequencies are $\frac{1}{\alpha_j} C'_{j,\cdot} = \frac{c}{\alpha_j} C_{j,\cdot}$, and can be quite different from the training proportions $C_{j,\cdot}$ in absolute numbers. Nevertheless, they correlate with $C_{j,\cdot}$ and will thus generate small losses $L(g)$.

In summary, data normalization makes tissue deconvolution a non-standard deconvolution problem. The choice of correlation as loss function allows for estimation of cell frequencies independent of normalization factors.

The minimum of $\mathcal{L}_g$ can be calculated analytically, yielding

$$\hat{C}(g) = (X^T \Gamma X)^{-1} X^T \Gamma Y \tag{2.8}$$

with $\Gamma = \operatorname{diag}(g)$. Inserting this term into L leaves a single optimization problem in g. L is minimized by a gradient-descent algorithm. Let μ_j and σ_j be the mean and standard deviation of $C_{j,\cdot}$, respectively. For the gradient (for more detailed calculations see Appendix A)

$$\begin{aligned} \frac{\partial L(g)}{\partial g_i} &= -\sum_{j=1}^{q} \sum_{k=1}^{n} \frac{\partial \left(\operatorname{cor}(C_{j,\cdot}, \hat{C}_{j,\cdot}) \right)}{\partial \hat{C}_{jk}} \frac{\partial \hat{C}_{jk}(g)}{\partial g_i} \\ &= \sum_{j=1}^{q} \sum_{k=1}^{n} \frac{1}{\sigma_j \hat{\sigma}_j} \left(\frac{\operatorname{cov}(C_{j,\cdot}, \hat{C}_{j,\cdot})}{n \hat{\sigma}_j^2} \big(\hat{C}_{jk} - \hat{\mu}_j \big) - \frac{1}{n} (C_{jk} - \mu_j) \right) \frac{\partial \hat{C}_{jk}(g)}{\partial g_i} \,. \end{aligned} \tag{2.9}$$

is obtained. With equation 2.8 one gets for the partial distribution $\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}$

$$\begin{aligned} \frac{\partial \hat{C}_{jk}}{\partial g_i} &= \left(\frac{\partial}{\partial g_i} \left((X^T \Gamma X)^{-1} X^t \Gamma Y \right) \right)_{jk} \\ &= \left(-(X^T \Gamma X)^{-1} \left(X^T \delta(i) X \right) (X^T \Gamma X)^{-1} X^T \Gamma Y + (X^T \Gamma X)^{-1} X^T \delta(i) Y \right)_{jk} \,. \end{aligned}$$

Written as a matrix one gets

$$\frac{\partial \hat{C}(g)}{\partial g_i} = (X^T \Gamma X)^{-1} X^T \delta(i) \big(1 - X (X^T \Gamma X)^{-1} X^T \Gamma \big) Y \,, \tag{2.10}$$

where $\delta(i) \in \mathbb{R}^{p \times p}$ is defined as

$$\delta(i)_{jk} = \begin{cases} 1 & \text{if } i = j = k, \\ 0 & \text{else.} \end{cases} \tag{2.11}$$

The constraints $||g||_2 = 1$ and $g_i \geq 0$ were incorporated by normalizing g by its length and by restricting the search space to $g_i \geq 0$.

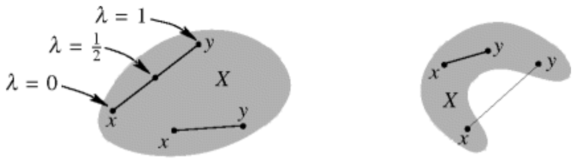


Figure 2.1: Left region is convex. For every two points in X, the connection line is also in X. The right region is not convex. One can find two points x and y where the line-segment joining these points x and y lies outside of the gray set X. Picture from [44].

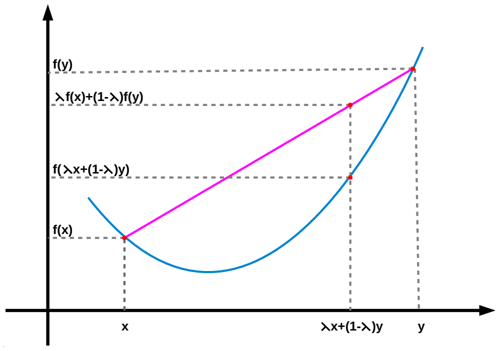


Figure 2.2: An example for a convex function. The connecting line for every $x, y \in X$ lies on or above the function f. Picture from [45].

2.3 Loss-Function Learning Problem is not Convex - Counterexample Shown for Hessian with Negative Eigenvalues

Here, it is shown that the loss-function Equation 2.3 is non-convex.

Definition 2.3.1 *$X \subset \mathbb{R}^n$ is a **convex set**, when for every pair of points within the region, every point on the straight line segment that joins the pair of points is also within the region, thus*

$$\forall x, y \in X : \forall \lambda \in [0,1] : x + \lambda(y - x) \in X. \tag{2.12}$$

*Let $X \subset \mathbb{R}^n$ be convex. A function $f : X \to \mathbb{R}$ is called **convex**, if*

$$\forall x, y \in X : \forall \lambda \in [0,1] : f(x + \lambda(y - x)) \leq f(x) + \lambda(f(y) - f(x)). \tag{2.13}$$

Figure 2.1 shows an example for a convex and a nonconvex region and Figure 2.2 shows a convex function.

Definition 2.3.2 *A matrix $A \in \mathbb{R}^{n \times n}$ is called **symmetric**, if*

$$A^T = A. \tag{2.14}$$

Definition 2.3.3 *A symmetric matrix A is called **positive semidefinite**, when the corresponding quadratic form q_A is positive semidefinite, that means*

$$q_A(x) \geq 0 \quad \forall x \in \mathbb{R}^n, x \neq 0. \tag{2.15}$$

The quadratic form is given by

$$q_A(x) = x^T Ax \quad \forall x \in \mathbb{R}^n, x \neq 0. \tag{2.16}$$

Definition 2.3.4 *Let $D \in \mathbb{R}$ be a non-empty, open subset. Let k be a non-negative integer. The function f is said to be of **differentiability class $\mathcal{C}^k$** if the derivatives f, f', ..., $f^{(k)}$ exist and are continuous.*

Theorem 2.3.5 ***Taylor's theorem with Peano remainder term*** *Let $f : [a,b] \to \mathbb{R} \in \mathcal{C}^{n+1}$, $n \in \mathbb{N}$, $x_0 \in (a,b)$. This implies the formula for the remainder term of Lagrange*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + R_n(x; x_0), \tag{2.17}$$

where $R_n(x; x_0)$ is given by

$$R_n(x; x_0) := \frac{f^{(n)}(\xi) - f^{(n)}(x_0)}{n!}(x - x_0)^n \text{ for a } \xi \in [0,1]. \tag{2.18}$$

Furthermore it holds for the remainder term in the n-th order that

$$R_n(x; x_0) = o((x - x_0)^n) \text{ for } x \to x_0. \tag{2.19}$$

proof: For proof see Königsberger, Analysis 2 [46]. □

The o in theorem 2.3.5 is little-o-notation by Landau. The notation $f = o(g)$ or $f \in o(g)$ means that f is growing slower than g.

Theorem 2.3.6 ***Criterion of convexity (see [46])***
Let $f : U \to \mathbb{R}$ be a $\mathcal{C}^2$-function on a convex and open set U. Then:

i) f is convex if and only if $f''(x) \geq 0 \quad \forall x \in U$.

ii) f is strictly convex if $f''(x) > 0 \quad \forall x \in U$.

proof: For proof see Königsberger, Analysis 2 [46]. □

As the calculation of U is not taking place in $\mathbb{R}$, but in $\mathbb{R}^p$, the theorem needs to be upgraded:

Theorem 2.3.7 *For $f \in \mathcal{C}^2$ and $X \subset \mathbb{R}^n$ open, convex and nonempty holds:*
f is convex on X if and only if $\forall x \in X : \quad \nabla^2 f(x)$ is positive semidefinite.

proof: $\Rightarrow$: Show first:
$f \in \mathcal{C}^1(\mathbb{R}^n), X \in \mathcal{R}^n$ is convex and not empty. Then:

$$f \text{ is convex in X } \Leftrightarrow \forall x, y \in X : f(y) \geq f(x) + (\nabla f(x))^T (y - x). \tag{2.20}$$

First the direction “⇒” of equation (2.20) is proven. Let f be convex on X, then

$$\begin{aligned} f(x+\lambda(y-x)) &\leq f(x)+\lambda(f(y)-f(x)) \quad \forall \lambda \in [0,1], \forall x, y \in X. \\ \Rightarrow\ \Psi(\lambda) &:= f(x)+\lambda(f(y)-f(x))-f(x+\lambda(y-x)) \geq 0. \end{aligned} \tag{2.21}$$

As $f \in \mathcal{C}^2(\mathbb{R}^n)$ it ensues that $\Psi(\lambda)$ is continuously differentiable as composition of continuously differentiable functions and $\Psi(0)=0$. So

$$\begin{aligned} \Psi'(\lambda) &= f(y)-f(x)-(\nabla f(x+\lambda(y-x)))^T(y-x) \\ \Psi'(\lambda) &= f(y)-f(x)-(\nabla f(x))^T(y-x). \end{aligned} \tag{2.22}$$

$\Psi(\lambda) \geq 0$ and $\Psi(0)=0$ implies $\Psi'(0) \geq 0$ because $\Psi \in \mathbb{C}^2(\mathbb{R}^n)$. It follows with equation (2.22) for $f(y)$ that

$$f(y) \geq f(x)+(\nabla f(x))^T(y-x). \tag{2.23}$$

Second, direction “⇐” of equation (2.20) is proven. Let y be given by $y = x+tc$ with $x \in X$, $c \in \mathbb{R}^n$ and $t > 0$ sufficiently small. Since X is convex and not empty it follows that for sufficiently small t also $y = x+tc \in X \quad \forall c \in \mathbb{R}^n$. With Equation (2.22) follows

$$0 \leq f(x+tc)-f(x)-t(\nabla f(x))^T c. \tag{2.24}$$

With the Taylor formulas in theorem 2.3.5 and $x_{\text{Taylor}} = x+tc$ and $x_{0,\text{Taylor}} = x$ follows for $f(y) = f(x+tc)$

$$\begin{aligned} f(x+tc) =& f(x)+(\nabla f(x))^T(x+tc-x)+\frac{1}{2!}(x+tc-x)^T\nabla^2(x+tc-x)^2 \\ &+ \underbrace{(x+tc-x)^T\frac{\nabla^2 f(\xi)-\nabla^2 f(x)}{2!}(x+tc-x)}_{o(t^2)}. \end{aligned} \tag{2.25}$$

With equation (2.25) it follows for equation (2.24)

$$\begin{aligned} f(x+tc)-f(x)-t(\nabla f(x))^T c =& \frac{1}{2}t^2 c^T\nabla^2 f(x)c+o(t^2) \geq 0 \qquad\qquad \mid : \frac{t^2}{2} \\ & c^T\nabla^2 f(x)c+\frac{2o(t^2)}{t^2} \geq 0 \end{aligned} \tag{2.26}$$

With $t \to 0+0$ one gets

$$\lim_{t\to 0+0}\left(c^T\nabla^2 f(x)c+\frac{2o(t^2)}{t^2}\right) \geq 0. \tag{2.27}$$

Therefore the Hessian is positive semidefinite.
Now the other direction of the theorem is shown.

$\Leftarrow$: Now the assumption is $\forall x \in X : \nabla^2 f(x)$ is positive semidefinite.
One has to show that f is convex (see definition).
As $f \in \mathcal{C}^2$ and X is open, convex and nonempty, there exists for every $x, y \in X$ a $\xi \in]0,1[$ such that

$$f(y) = f(x) + (\nabla f(x))^T (y - x) - \frac{1}{2}(y - x)^T \nabla^2 f(x + \xi(y - x))(y - x) \tag{2.28}$$

holds. As $\forall x \in X$ it holds that $\nabla^2 f(x)$ is positive semidefinite and so $x^T \nabla^2 f(x) x \geq 0 \; \forall \; x \in X$ (see definition 2.3.3). It follows

$$\frac{1}{2}(y - x)^T \nabla^2 f(x + \xi(y - x))(y - x) \geq 0. \tag{2.29}$$

With equation (2.28) this yields

$$f(y) \geq f(x) + (\nabla f(x))^T (y - x). \tag{2.30}$$

Let $\overline{x} := x + x + \lambda(y - x)$ with $\lambda \in]0,1[$. $\overline{x} \in X$ for $x, y \in X$ since X is convex. Using equation (2.30) with $(x, y) = (\overline{x}, x)$ and with $(x, y) = (\overline{x}, y)$ yields

$$\begin{array}{rrl@{\qquad\qquad}l} & f(x) \geq & f(\overline{x}) + (\nabla f(\overline{x}))^T (x - \overline{x}) & | \cdot (1 - \lambda) \\ + & f(y) \geq & f(\overline{x}) + (\nabla f(\overline{x}))^T (y - \overline{x}) & | \cdot \lambda \\ \hline\hline \multicolumn{2}{r}{(1 - \lambda) f(x) + \lambda f(y) \geq} & f(x + \lambda(y - x)) + 0 & \end{array} \tag{2.31}$$

which is the definition of convergence (see 2.3.1). □

To prove non-convexity, theorem 2.3.7 gets applied. It holds $g_i \geq 0 \;\; \forall \;\; i \in 1, \ldots, p$. Note that at least as many g_i need to be non-zero as cell types are included. Otherwise the corresponding system of equations remains under determined. Further the normalization $||g|| = 1$ is to ensure uniqueness of the solution. Here this constraint is neglected, since it does not change the subsequent arguments.

G is part of a $p-$dimensional sphere and G is nonempty and convex. Furthermore, G is not open since $g_i \geq 0 \;\; \forall \;\; i \in 1 \ldots p$. First the inner part $\mathring{G}$ of the region G is considered. Figure 2.3 illustrates the region G for $g \in \mathbb{R}^3$ for 3,2 and 1 cell type, respectively. For $q \geq 1$, $\mathring{G}$ is convex, nonempty and open. Furthermore, the covariance and the standard deviation are $\mathcal{C}^\infty$ and thus also $\hat{C}(g)$ is $\mathcal{C}^\infty$, where it gets assumed that the standard deviation remains non-zero in general.

The Hessian of the outer loss-function $L(g)$ is given by (for more detailed calculations see

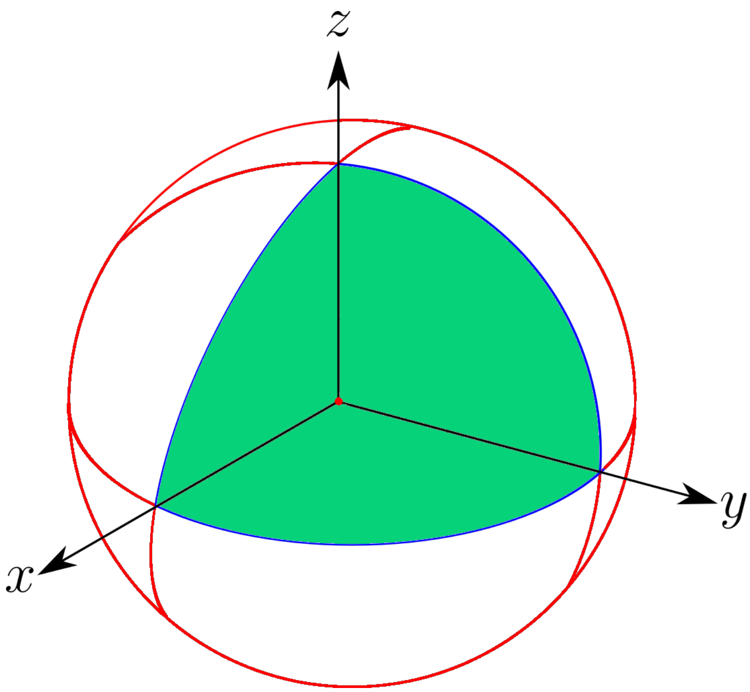


Figure 2.3: For three genes x, y and z and three cell types the allowed region G corresponds to the green area. For two cell types G corresponds to the green area and its border (blue), where the lines $x = y = 0$, $z = y = 0$, and $x = z = 0$ are excluded. Theorem 2.3.7 is only valid on G. Picture from [47].

Appendix A)

$$\begin{aligned}
H_{li}(g) =&\frac{\partial}{\partial g_l}\left(\frac{\partial L}{\partial g_i}\right) = \sum_{j=1}^{q}\sum_{k=1}^{n}\frac{\partial}{\partial g_l}\left(\frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,\cdot}))}{\partial \hat{C}_{jk}}\cdot\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right) = \\
=&\sum_{j=1}^{q}\sum_{k=1}^{n}\left\{\underbrace{\left[\frac{\partial}{\partial \hat{C}_{jk}}\left(\frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,\cdot}))}{\partial \hat{C}_{jk}}\right)\right]}_{\textcircled{A}}\frac{\partial \hat{C}_{jk}(g)}{\partial g_l}\cdot\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right. \\
&\left. + \frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,\cdot}))}{\partial \hat{C}_{jk}}\cdot\underbrace{\frac{\partial}{\partial g_l}\left(\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right)}_{\textcircled{B}}\right\}.
\end{aligned}\tag{2.32}$$

For Ⓐ and Ⓑ one gets

$$\begin{aligned}\textcircled{A} =&\frac{1}{n\sigma_j\hat{\sigma}_j^3}(\hat{\mu}_j - \hat{c}_{j,k})\left(\frac{\operatorname{cov}(C_{j,\cdot},\hat{C}_{j,\cdot})}{n\hat{\sigma}_j^2}(\hat{C}_{jk}-\hat{\mu}_j) - \frac{1}{n}(C_{jk}-\mu_j)\right)\\ &+\frac{1}{n^2\sigma_j\hat{\sigma}_j^3}\Bigg[(C_{jk}-\mu_j)(\hat{C}_{jk}-\hat{\mu}_j) - \frac{2}{\hat{\sigma}_j^2}\operatorname{cov}(C_{j,\cdot},\hat{C}_{j,\cdot})(\hat{C}_{jk}-\hat{\mu}_j)^2\\ &+(n-1)\operatorname{cov}(C_{j,\cdot},\hat{C}_{j,\cdot})\Bigg]\end{aligned} \tag{2.33}$$

and

$$\begin{aligned}\textcircled{B} =& \big((X^T\Gamma X)^{-1}X^T\left\{\delta(l)X(X^T\Gamma X)^{-1}X^T\delta(i)(-1+X(X^T\Gamma X)^{-1}X^T\Gamma)\right.\\ &\left.+\delta(i)X(X^T\Gamma X)^{-1}\left[X^T\delta(l)X(X^T\Gamma X)^{-1}X^T\Gamma - X^T\delta(l)\right]\right\}Y\big)_{j,k}\,.\end{aligned} \tag{2.34}$$

Non-convexity is shown by a counter example. Here, specific values for X, C, and Y are chosen, subject to the constraints

1. $X_{i,j} \geq 0\ \forall\ i \in \{1,\ldots,p\}$ and $\forall\ j \in \{1,\ldots,q\}$,
2. $C_{j,k} \geq 0\ \forall\ j \in \{1,\ldots,q\}$ and $\forall\ k \in \{1,\ldots,n\}$,
3. $Y_{i,k} \geq 0\ \forall\ i \in \{1,\ldots,q\}$ and $\forall\ k \in \{1,\ldots,n\}$.

For the reference matrix X

$$X = \begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 3\\ 0 & 0 & 4\end{pmatrix} \tag{2.35}$$

is chosen and for the distributions in C

$$C = \begin{pmatrix}2 & 1\\ 4 & 6\\ 3 & 2\end{pmatrix}. \tag{2.36}$$

The bulk profiles Y were chosen as

$$Y = \begin{pmatrix}1 & 5\\ 2 & 6\\ 3 & 7\\ 4 & 8\end{pmatrix}. \tag{2.37}$$

To prove non-convexity, it has to be shown that the Hessian matrix has negative Eigenvalues. For calculating the Hessian H the reference matrix X and the bulk profile Y are normalized to 100 counts in every column. The composition C is normalized to 1 for each column.

The estimated bulk profiles $\hat{C}$ are calculated with equation (2.8) by using $g = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Equation (2.32) yields the Hessian H,

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 14.959681 & 6.333319 \\ 0 & 0 & 6.333319 & -27.626320 \end{pmatrix}. \qquad (2.38)$$

The Eigenvalues are 15.88160, 0, 0 and -28.54824.

One of four of the Eigenvalues is negative. Thus, function equation (2.3) is non-convex for X, C, and Y as given by equations (2.35), (2.36), and (2.37). With theorem 2.3.7 the Hessian of the regarded loss-function learning problem is not convex in $\mathring{G}$ and therefore not on G.

2.4 Algorithm for Loss-function Learning

Here, the used algorithm for loss-function learning is presented. This algorithm uses a coordinate descent in condition with line search. The update step is

$$g_{\text{step s}} = g_{\text{starting point}} + \frac{\text{step size}}{s_{\max}} \cdot s \cdot \nabla \mathcal{L} \qquad \text{for } s \in \{0, 1, \ldots, s_{\max}\}, \qquad (2.39)$$

where ∇L is the gradient of the loss function. Step size is the step length in the direction ∇L and s_{max} is the maximum number of points that are evaluated along the gradient. Further negative entries in $g_{\text{step s}}$ are set to zero such that the constraint $g_i \geq 0$ holds. For every step $s \in \{0, 1, \ldots, s_{\max}\}$ one calculates the corresponding g_{s} and the value of the loss-function L. There are three possible cases where the position of the optimal s_{opt}, which corresponds to the minimum in the loss-function L, is localized. All cases are exemplified in Figure 2.4:

1. The loss-function L takes its minimum for $s_{\text{opt}} \in \{1, 2, \ldots, s_{\max} - 1\}$. In this case, it gets updated and the next iteration starts:

$$g_{\text{new start}} = g_{s_{\text{opt}}} \qquad (2.40)$$

2. L takes its minimum for $s_{\text{opt}} = s_{\max}$. Here, the maximum in the direction of the gradient is not yet reached. The search continues in the direction of the calculated gradient and gets not updated in the next iteration. The new starting point of the gene weight is

$$g_{\text{new start}} = g_{s_{\max}} \qquad (2.41)$$

and the step size in the next iteration is doubled.

3. L takes its minimum for $s_{\text{opt}} = 0$. So the starting point for g is very near to the maximum in the direction of gradient. To find a better solution, the same starting vector g is used and the step size is divided by 10. Here, an update of the gradient is not necessary.

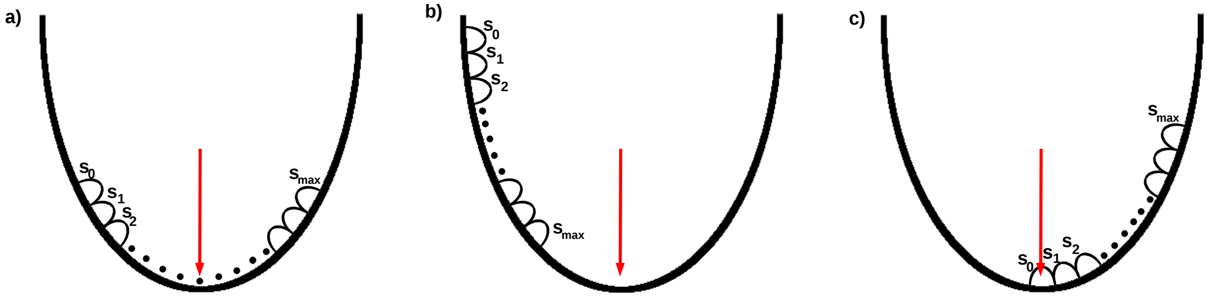


Figure 2.4: The three cases of minimum positions are shown here. The red arrow points to the minimum of the loss-function. In figure (a) the minimum is reached within the number of steps. In figure (b) the minimum is not reached after the maximal step number, therefore the step size is increased. In figure (c) the starting step s_0 is too near to the minimum of the loss function, which makes it necessary to reduce the step size.

As starting step size the number of steps is chosen in order to minimize the number of initial parameters. Then, in every iteration either the gradient or the step size is updated. When a minimum of the loss-function L is reached in $s_{\text{opt}} \in \{1, 2, \ldots, s_{\max} - 1\}$ a new gradient is calculated (Figure 2.4 a). The step size is updated when the minimum is not yet reached or if it is reached in the first step, receptively (Figure 2.4 b an c). The number of steps is given by the user. For initialization $g = (1, ..., 1)$ gets used.

2.5 Simulation Study on Artificial Data

In this section it gets shown that loss-function learning reliably identifies cell markers for DTD.

A reference matrix X_1 consisting of four reference profiles (columns) and six genes (rows) gets defined as

$$X_1 = \begin{pmatrix} 8 & 4 & 0 & 0 \\ 3 & 6 & 0 & 0 \\ 0 & 0 & 7 & 4 \\ 0 & 0 & 1 & 5 \\ 5 & 2 & 9 & 9 \\ 6 & 2 & 9 & 7 \end{pmatrix}, \tag{2.42}$$

and normalized column-wise to a value of 100.

The cellular compositions of the artificial bulk profiles were simulated by randomly drawing values between zero and one for each cell type and then normalized to the sum of 1. By multiplying the reference matrix X_1 with the cellular compositions in C, bulk profiles Y are obtained. 50 artificial bulk profiles for testing are simulated and normalized like the reference profiles. In order to proof the autonomy of the learned models from the training data 100 different cellular compositions C

are simulated to give 100 different artificial training mixtures. Every training mixture consists of 100 artificial bulk profiles.

In the presented loss-function learning scenario, column 3 and 4 of X get ignored and only the first two get used for loss-function learning.

The gene weight of $\sqrt{\frac{1}{6}}$ is assumed for every gene in the beginning. This yielded a mean correlation of 0.673 ± 0.039 for the training sets and of 0.673 in the test set for the two cell types. The estimated gene weights are given in Table 2.1.

std. model	0.637 ± 0.039
gene 1	0.748 ± 0.005
gene 2	0.252 ± 0.005
gene 3	0 ± 0
gene 4	0 ± 0
gene 5	0 ± 0
gene 6	0 ± 0
loss-fct. learned model	1 ± 0

Table 2.1: Mean value and standard deviation for 100 complete optimization runs applied to the ideal loss-function learning problem. Each was optimized for 100 different randomly chosen cellular compositions. Only for the two genes which carry information the algorithm calculated a gene weight greater than zero. The corresponding average correlation for the two considered cell types was close to 1 on simulated test data.

The algorithm obtained non-vanishing weights for gene 1 and 2, while the other gene weights were equal to zero. Thus, the algorithm correctly selected those genes which had a vanishing contribution in cell types 3 and 4. The results of the gene vectors after loss function learning were manipulated for all genes, to check their influence on the overall deconvolution result. For this manipulation a random value between zero and one for the gene of interest were taken and then the manipulated vector of gene weights g were renormalized again to one. With these new gene vectors the artificial bulk of the test set was deconvoluted. Results can be seen in table 2.2.

manipulation in	result of loss-fct. learned model
no manipulation	1 ± 0
gene 1	1 ± 0
gene 2	1 ± 0
gene 3	1 ± 0
gene 4	1 ± 0
gene 5	0.870 ± 0.098
gene 6	0.840 ± 0.106

Table 2.2: Mean value and standard deviation for 100 complete optimization runs applied to of the ideal loss-function learning problem before and after gene manipulation. The results are shown on the test set.

Manipulation in gene one and two had no influence on the deconvolution results. The loss-function learning problem in this case is purely analytically solvable. The gene weight corresponds to a multiplication of the first two rows of $Y = X\hat{C}$ with a constant value. This has no influence on the solution $\hat{C}$. Gene three and four did not contribute to the deconvolution and consequently the average correlation over the considered cell types did not change. Here the values in the reference matrix are zero and thus, according to equation (2.8), a non-zero value in g, i.e. in Γ, had no influence on the resulting cell-type distribution $\hat{C}(g)$. For the genes five and six a manipulation led to a decreased value for the overall correlation. For these genes the reference matrix has non-vanishing entries for all four cell types. Here, the deconvolution of cell type 1 and 2 is confounded by the amount of cells of type 3 and 4 and consequently the performance gets lower in accuracy.

Next, the reference matrix X was changed slightly and the first two cell types were considered for deconvolution, as previously. The new reference matrix becomes

$$X_2 = \begin{pmatrix} 8 & 4 & 5 & 0 \\ 3 & 6 & 0 & 0 \\ 0 & 0 & 7 & 4 \\ 0 & 0 & 1 & 5 \\ 5 & 2 & 9 & 9 \\ 6 & 2 & 9 & 7 \end{pmatrix}. \tag{2.43}$$

The remaining simulation study was designed as previously (100 training sets with 100 artificial bulk profiles each and 50 artificial bulk profiles as test set). The results are shown in Table 2.3:

std. model (training set)	0.655 ± 0.041
gene 1	0.751 ± 0.005
gene 2	0.248 ± 0.005
gene 3	0 ± 0
gene 4	0 ± 0
gene 5	0 ± 0.001
gene 6	0 ± 0
loss-fct. learned model (test set)	0.906 ± 0

Table 2.3: Mean value and standard deviation for 100 complete optimization runs applied to the ideal loss-function learning problem. Results for X_2. The results are shown for the test set.

Although the gene weights were similar to the first simulation study, the performance analytically from a correlation of 1 to 0.906. Note that the deconvolution problem is no longer perfectly solvable due to confounding contribution of cell types 3 and 4. The gene weights were manipulated as in the previous simulation study and yielded the results shown in Table 2.4.

manipulation in	result of loss-fct. learned model
no manipulation	0.906 ± 0
gene 1	0.906 ± 0.001
gene 2	0.906 ± 0
gene 3	0.906 ± 0
gene 4	0.906 ± 0
gene 5	0.781 ± 0.080
gene 6	0.784 ± 0.068

Table 2.4: Mean value and standard deviation for 100 complete optimization runs applied to the ideal loss-function learning problem before and after gene manipulation. The shown results correspond to reference matrix X_2 and are again evaluated on test data.

Manipulating the genes two to four had no influence on the results as discussed previously. Also manipulation of gene one had no influence. As the gene weights vectors were heavy on the first two genes and the others only showed vanishing entries, equation (2.8) led to a simplified equation system consisting of two equations for two variables. This reduced mathematical expression again is perfectly solvable. A change in the gene weight vector of gene one corresponds to a multiplication of the corresponding equation $gY = gXc$, with a constant factor and has therefore no influence on the results. The last two genes had entries in all cell types and thus a change in the gene weights led to lower correlations than in the first simulation study which was shown above.

As third example the reference matrix X_3 is considered,

$$X_3 = \begin{pmatrix} 0 & 0 & 8 & 4 \\ 0 & 0 & 3 & 6 \\ 7 & 4 & 5 & 7 \\ 1 & 5 & 3 & 4 \\ 9 & 9 & 5 & 2 \\ 9 & 7 & 6 & 2 \end{pmatrix} . \tag{2.44}$$

Now, there are zeros in the first two genes of the two investigated cell types. The realization of the experiment was the same as in X_1 and X_2. The results for the gene weights and their standard deviation are itemized in table 2.5.

std. model (training set)	0.954 ± 0.050
gene 1	0.013 ± 0.018
gene 2	0.013 ± 0.018
gene 3	0.025 ± 0.032
gene 4	0.031 ± 0.042
gene 5	0.897 ± 0.136
gene 6	0.020 ± 0.028
loss-fct. learned model (test set)	0.971 ± 0.002

Table 2.5: Mean value and standard deviation for 100 complete optimization runs applied to of the ideal loss-function learning problem. Results for X_3. The shown results correspond to the artificial bulk of the test set.

Again the resulting mathematical expression is not analytically solvable. The starting correlation for the standard set with equally distributed starting genes was quite high. The gene weight was mainly focused on gene five. The manipulation results are listed in table 2.6.

manipulation in	result of loss-fct. learned model
no manipulation	0.971 ± 0.002
gene 1	0.971 ± 0.002
gene 2	0.971 ± 0.002
gene 3	0.412 ± 0.192
gene 4	0.779 ± 0.060
gene 5	0.968 ± 0.012
gene 6	0.848 ± 0.070

Table 2.6: Mean value and standard deviation for 100 complete optimization runs applied to the ideal loss-function learning problem before and after gene manipulation. Calculations for the third reference matrix X_3. The results are shown for the artificial bulk of the test set.

Here, again, gene one and two had no influence on the results, as their entries in the matrix were zero. Gene three had a high influence since here the overall correlation dropped while manipulating the corresponding entry in the gene vectors. Even if gene five had a high value of the gene weight vector, the influence on the overall correlation result was small as the value dropped only slightly (0.968 vs 0.971) under manipulation. The other two genes also contributed to a decreased value in the overall correlation.

The last example considers a reference matrix X_4 with non zero entries in every gene for every cell type. The matrix was given by

$$X_4 = \begin{pmatrix} 8 & 4 & 5 & 2 \\ 3 & 6 & 1 & 6 \\ 5 & 7 & 7 & 4 \\ 3 & 4 & 1 & 5 \\ 5 & 2 & 9 & 9 \\ 6 & 2 & 9 & 7 \end{pmatrix} . \tag{2.45}$$

Here, obviously no gene is destined to be used by the algorithm for deconvolution. The numerical simulation was carried out as before. The results for deconvolution and gene weights are listed in table 2.7.

std. model (training set)	0.648 ± 0.045
gene 1	0.999 ± 0.004
gene 2	0.000 ± 0.000
gene 3	0.000 ± 0.001
gene 4	0.000 ± 0.000
gene 5	0.000 ± 0.002
gene 6	0.000 ± 0.001
loss-fct. learned model (test set)	0.936 ± 0.001

Table 2.7: Mean value and standard deviation for 100 complete optimization runs applied to of the ideal loss-function learning problem. Results for X_4. Results are shown for the artificial bulk of the test set.

The gene weight was concentrated on gene one. Manipulating the gene vectors led to the results listed in table 2.8.

manipulation in	result of loss-fct. learned model
no manipulation	0.936 ± 0.001
gene 1	0.934 ± 0.015
gene 2	0.758 ± 0.003
gene 3	0.842 ± 0.001
gene 4	0.665 ± 0.005
gene 5	0.068 ± 0.074
gene 6	0.204 ± 0.074

Table 2.8: Mean value and standard deviation for 100 complete optimization runs applied to the ideal loss-function learning problem before and after gene manipulation. Calculations for the fourth reference matrix X_4. The results are shown for the artificial bulk of the test set.

The results display that gene one has no significant influence on the deconvolution results, even if the corresponding value of the gene weight vector was very high. Gene two to four had more influence on the deconvolution result. Manipulating the last two genes had a significant influence on the deconvolution results. Here, a deconvolution was not possible any more.

Next, biological variability was studied (see section 1.2.1). The perfectly solvable example X_1 was chosen for simulation on 100 training sets with 100 artificial bulk sets each and a test set with 50 artificial bulks consistent with the previous procedures. For simulating the biological variability, an error, drawn from a normal distribution with mean zero, was added and an increasing standard deviation between 0.1% to 100% of the total number of counts in the bulk samples of the training sets. For all perturbations the g=1 model was compared to the model optimized by loss-function learning. The results of the mean correlation for training and test set are shown in figure 2.5. The standard model (blue) gives for every perturbation step high variance in the deconvolution results.

With increasing perturbation, the average correlation drops. For the loss-function learned model (green) only small variance in the results for small perturbations up to 1% are shown. Here the loss-function learning algorithm was able to calculate a model for deconvoluting the bulk profiles, despite the perturbations. For higher perturbations the algorithm showed less performance and the results within one perturbation step varied within a broader range. Also, for high perturbations the loss-function learned model led to better results than the standard model on average, which showed a poor performance in general. Figure 2.6 displays the gene weights of every set for all perturbations. With increasing perturbation, the range of the gene weight for the considered gene was increasing. The weights are for most sets concentrated on the first and second gene. Gene three and four stay at or near zero for all perturbations. Even for high perturbations the algorithm realizes that these two genes are not helpful for deconvolution. For higher perturbations gene five and six exhibit very high values for some data sets.

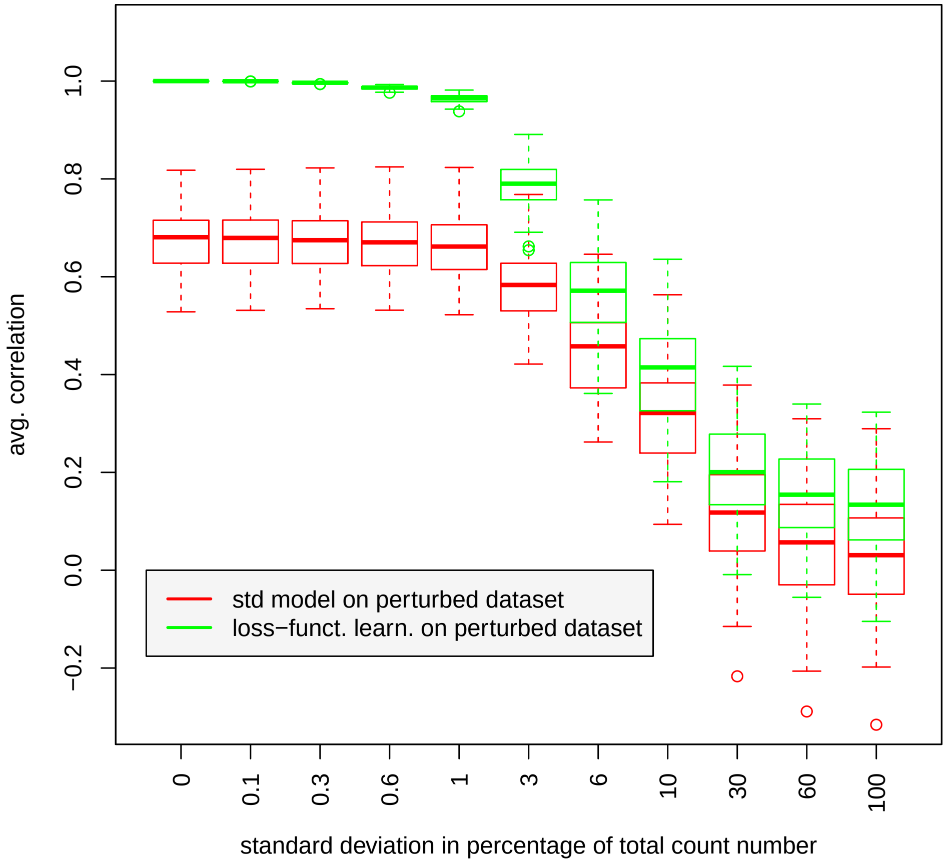


Figure 2.5: Average correlation of the standard model and the model from loss-function learning. The training sets were perturbed by an error simulated from a normal distribution with standard deviation ranging from 0% to 100% of the total count number of the bulk measurements. Small perturbations up to 1% have only a small influence on the results from loss-function learning (green). For higher perturbations the variance within one perturbation increases and the average correlation drops fast. The standard model (red) shows a similar distribution as the loss-function learned model. However the deconvolution results show higher variance in the results for small perturbations.

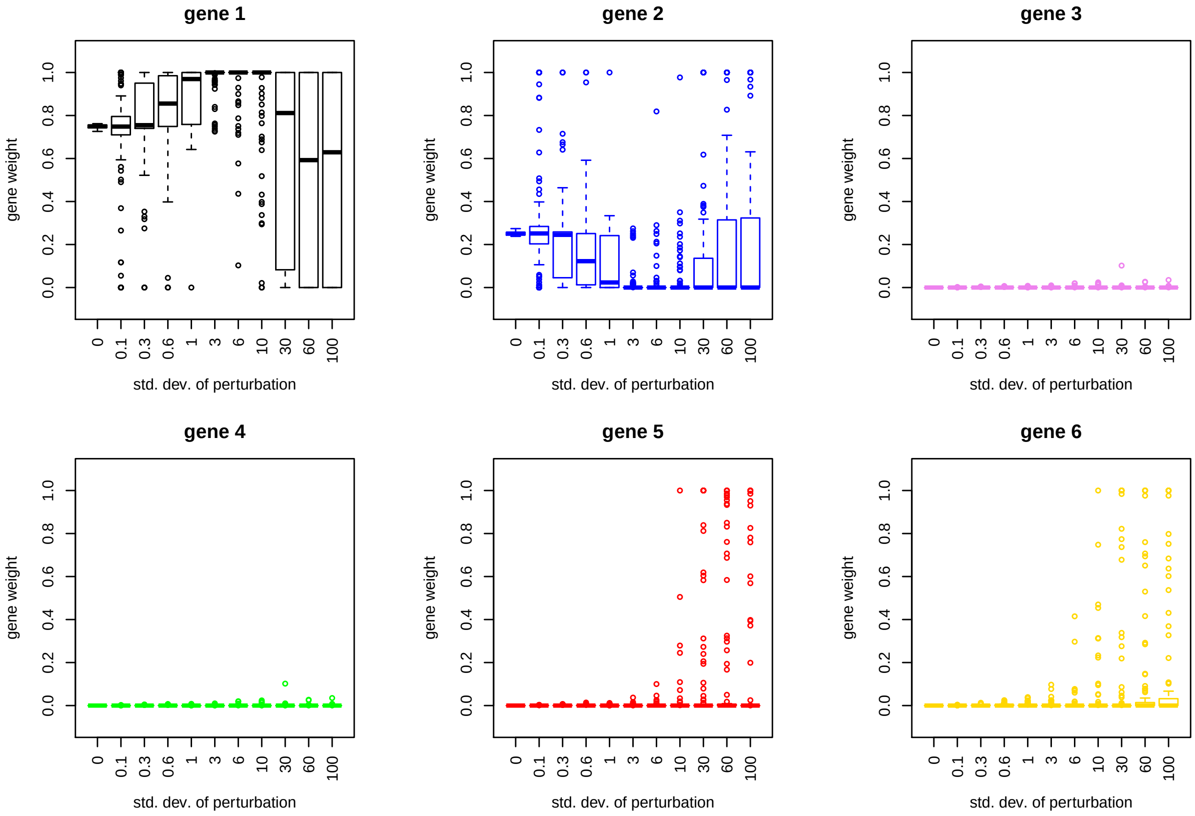


Figure 2.6: Gene weight distribution vs standard deviation for the six genes after loss-function learning. For every gene, gene weights versus perturbation is plotted for all samples. As for the unperturbed model gene one exhibits the highest value for most samples, followed by gene two. These two genes are the genes which held the deconvolution information in an unperturbed data set. Gene three and four continuously show very low gene weights. The weights in gene five and six respond with higher values to greater perturbation in some of the data sets.

Chapter 3

Results of DTD with Melanoma Data

In this chapter the developed loss-function learning algorithm for digital tissue deconvolution is applied to single-cell RNA sequencing data of melanomas. Section 3.1 gives a short overview over the data set. Next a biological description of melanoma tumors and the analyzed cell types follows in section 3.2. In section 3.4 - 3.6 it is shown that the presented loss-function learning algorithm improves deconvolution results in different settings. First, it is pointed out that incomplete reference data does not deteriorate the deconvolution (section 3.3). Next, it is proven that digital tissue deconvolution is able to quantify small cell populations (section 3.4) and that it can disentangle closely related cell types (section 3.5). Even for small training sets loss-function learning improves performance. The starting point of the deconvolution and the used training mixtures have only little influence on the results which is shown in section 3.6. Section 3.7 elaborates how high-performance computing was used for calculating a model with five times as many genes as before and demonstrates that the developed procedure is able to detect cellmarkers. Results for smaller and larger data sets are compared in section 3.8. Finally it is displayed that loss-function learning outperforms the state of the art method CIBERSORT (section 3.10). The discussion of the results follows in chapter 5.

3.1 Description of the Melanoma Dataset

For both training and validation, expression profiles of cellular mixtures of known composition needed to be available. Expression data of melanomas whose composition has been experimentally resolved using single-cell RNASeq profiling [48] got used for the calculations. The data included 4,645 single-cell profiles from 19 melanomas. The cells were annotated as T cells (2,068), B cells (515), macrophages (126), endothelial cells (65), cancer-associated fibroblasts (CAFs) (61), natural killer (NK) cells (52), and tumor/unclassified (1,758). The first 9 melanomas defined the validation cohort and the remaining 10 the training data. Figure 3.1 illustrated the data set.

First, data were transformed into transcripts per million. Then, for each cell cluster we sampled 20% of single-cell profiles in the training data were sampled, summed up, normalized to a common

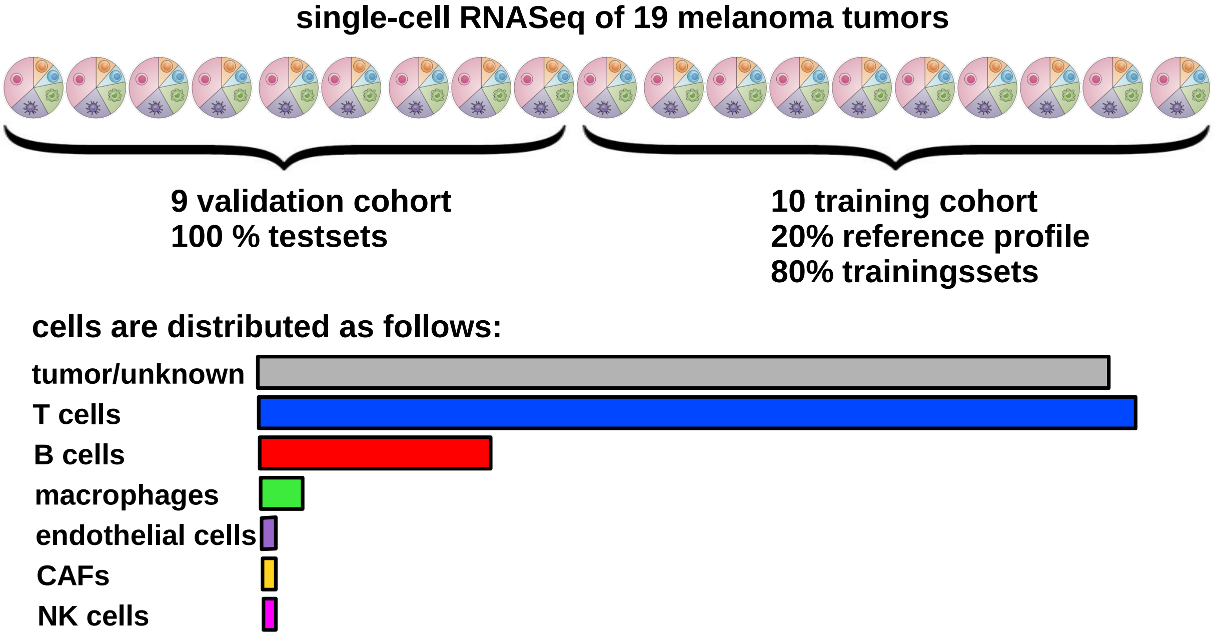


Figure 3.1: In upper part of the figure the fractions of patients in training and validation cohort are visualized. The lower part shows the distribution of the different cell types.

number of counts, and removed from the training data. This yielded reference profiles $X_{\cdot,j}$. The 1,000 genes with the highest variance across all reference profiles were used to train models.

The sum of all single-cell profiles of a melanoma yielded the bulk profiles. In addition, a large number of artificial bulk profiles were generated by randomly sampling single-cell profiles and summing them up. All bulk profiles were normalized to the same number of reads as those in $X_{\cdot,j}$.

3.2 Melanoma and Cell Type Characterisation

Melanomas are cancers which are derived from melanocytes, the pigment-containing cells. They normally occur in the skin, but rarely at other places like mouth, intestines or eye [49]. A large majority of melanomas are caused by ultraviolet light exposure in humans with fair skin [49, 50]. Others are derived from moles [49]. Increased risk for melanomas is indicated by the existence of an exuberant number of moles starting from childhood, family members with melanomas, poor immune function and some rare genetic defects [51]. Melanomas are detected through biopsy or analysis of suspicious skin lesions. When detected early, the prospects of successful treatment are high. The cancer is then totally removed by surgery. Yet, when the lymph nodes are affected, the healing prognosis is much worse. New therapies with immunotherapeuthical approaches for treatment of spread melanomas are currently developed and tested in clinical studies. Among other things, the therapeutical potential of melanoma specific T cells [52] and antigene presenting dendritic cells are tested [53, 54]. But in contrast to leukemia, which is also tested for several cellular immunotherapies, the melanoma is a firm and compact tumor. As such, it is not as easily reached by immune cells as leukemia cancers.

In healthy patients the main task of T cells is to identify and eliminate virus-infected cells. The assumption is, that T cells are also able to recognize and eliminate cancer cells. Due to mutations in the tumor, tumor-specific neo-antigens are arising [55, 56]. They help the tumor to protect itself against the immune system. T cells, which normally scan the cells for damage, do not recognize these mutated cells as invaders, therefore the malignant cells can proliferate. For therapeutic efficiency of immunotherapies these neoantigens are considered to be important [57], as it is suggested that neo-antigens are commonly recognized by intratumoral CD8+ T cells [57]. CD4+ T helper cells play a key role in the regulation of most antigen-specific immune responses. The response of CD4+ T helper cells to melanomas and other tumors helps to develop optimal anticancer vaccines and to create T cell related therapies [57].

B cells are part of the adaptive immune system. They produce antibodies and present antigens. B cell receptors on their cell membrane allow them to bind on specific antigens and initiate an antibody response to it [58]. The microenvironment of tumors is often infiltrated by B cells. Depending on the tumor, the immune response to the tumor can be positively or negatively regulated. It is found that particularly in melanomas without metastases, the content of tumor associated B cells is significantly higher as in melanomas with metastases. Furthermore the overall survival of patients is significantly correlated with a higher number of tumor associated B cells [59]. A comparison between melanoma-associated and peripheral blood-derived B cells showed that they are distinct in abundance, clonality and gene expression. The B cells in the tumoral content may act as antigen presenting cells. They help to initiate an anti-tumor immune response based on T cells [60]. B cells have also the ability for acquiring antigens via the B cell receptor. These antigens can be transferred to other antigen-presenting cells by direct cell contact, i.e. to macrophages which activate CD4+ T cells.

Macrophages are white blood cells which are part of the innate immune system. They help to activate the adaptive immune system by initiating defense mechanisms based on recruiting other immune cells. For example, they activate T cells by presenting them antigens [61]. In solid tumors, tumor-associated macrophages (TAM) are important components in the microenvironment of the tumor. They emerge from monocytes and exhibit in each differentiation state various immunosuppressive functions which maintain the microenvironment of the tumor. Drugs or stromal factors can simulate the tumor-associated macrophages to produce specific chemokines which recruit tumor-infiltrating lymphocytes. Therefore the macrophages constitute ideal targets for cancer immunotherapy [62]. TAMs are involved in all stages of tumor development. In an early stage they establish an inflammatory micorenvironment. Later they suppress the anticancer activity of the immune system. Another feature is that they enhance migration and invasion of cancer cells, thereby contributing to the metastatic process [63].

Endothelial cells play an important role in the cancerogeneous promotion. The endothelium constitutes the inner surface of blood and lymphatic vessels. As such a barrier between circulating blood or lymph in the lumen and the rest of the vessel wall is formed. The endothelium is involved in most disease states [64]. The tumor extravasation through the membrane of endothelial cells is a critical step in the metastatic progress. Tumor cell extravasation is promoted by the interaction of metastatic melanoma cells with the endothelial cells [65].

A further cell type of the immune system are the natural killer cells which belong to the lymphocytes. They classify and kill abnormal cells like tumor and virus cells. NK cells have no antigen specific receptors and belong to the innate immune system [66]. They are regarded for the development of novel therapies as they participate in the early immune response against melanoma. Also their interaction with dendritic cells and cytokine secretion assists in the development of an adequate response of the adaptive immune system. As the melanoma cells often escape the CD8+ T cell recognition due to the down-regulation of major histocompatibility complex class (MHC) I molecules, NK cells have the scope to detect and destroy melanoma cells which express low levels of these molecules which makes NK cells potential candidates for melanoma immunotherapy [67].

A cell type which supports melanomas are the cancer-associated fibroblats (CAFs). The tumor microenvironment displays a high number of these complex cells [68], which are unable to undergo apoptosis [69]. CAFs create an environment supportive to tumor growth and metastasis by producing cytokines and chemokines. Moreover they dispose pro-inflammatory and pro-angiogenic factors [69]. The strategy in targeting CAFs is to create a tumor-resistant microenvironment in order to suppress the growth of melanomas which carry different genetic mutations. However until now, the mechanism by which CAFs help melanomas to progress and how they contribute to drug resistance [70] is not precisely known.

3.3 Loss-Function Learning Improves DTD Accuracy in the Case of Incomplete Reference Data

From the training cohort 2,000 artificial cellular mixtures were generated. For each of these mixtures, 100 single-cell profiles are drawn randomly, their raw counts are summed up and normalized to a fixed number of total counts. Analogously, 1,000 artificial cellular validation mixtures were generated.

Then, the reference matrix X got restricted to three cell types (T cells, B cells and macrophages) after drawing 20% of every cell type including tumor/unknown to generate a very realistic distributions in the training sets. Hence endothelial cells, CAFs, NK cells and tumor/unclassified cells in the mixtures are not represented in X. The variance of all genes in X was calculated and then the 1,000 most variable ones were used for calculation of the loss function. For standard DTD with $g = (1, \dots, 1)$ correlation coefficients of 0.70 (T cells), 0.39 (B cells), and 0.52 (macrophages) between true and estimated cell population sizes for the validation mixtures were observed (scatter plots for validation set in Figure 3.2 (a)-(c)). These improved to 0.87 (T cells), 0.89 (B cells), and 0.84 (macrophages) for loss-function learning, after 1000 iterations of the gradient descent algorithm on the training data are ran (scatter plots for validation set in Figure 3.2 (d)-(f)).

In order to test whether additional information improves the learned model, the 1,000 most variable genes based on all six cell types including tumor cells were chosen. Minimized again only for T and B cells and macrophages, correlation coefficients of 0.87 (T cells), 0.89 (B cells) and 0.82 (macrophages) for the validation mixtures were obtained. Compared to the model above, which selected genes from only three of the considered cell types, the deconvolution results are nearly the same.

The loss-function learning algorithm converged for the training and test set (blue and green line in Figure 3.2 g) within the 1000 iterations run. The red line displays the mean correlation of the standard model.Further, it can be observed that the deconvolution improves fastest in the first few steps and then converges at about 300 steps. Note that no signs of overfitting are observed, as can be deduced from by the green line representing the test data.

The heatmap for the 50 most important genes (genes were ranked by $\hat{g}_i \times \text{var}(X_{i,\cdot})$) is shown in Figure 3.2 h. The map clusters genes characteristic of T cells, B cells, and macrophages, while no clusters for endothelial cells, CAFs and NK cells are observed. The latter cell types were part of the artificial bulk but not considered in the deconvolution.

Next, it got tested whether the calculated cellular composition using loss-function learning depends on the starting point g of the algorithm. For this purpose, the gradient descent algorithm is tested on the 100 most variable genes for 100 different, uniformly drawn starting points $g \in [0, 1]^p$. The maximal Euclidean distance between the resulting composition vectors c was 2%.

To test the limits of the approach, all but the macrophages are excluded, which account for less than 3% of all cells, from the reference data X. Here the 1,000 highest expressed genes for macrophages are chosen. It can be observed that standard DTD broke down, while loss-function learning yielded a model that predicted macrophage abundances that still correlated well ($r = 0.83$) with the true abundances (Figure 3.3). Figure 3.4 c shows the corresponding heatmap of the reference matrix X. It can be observed that the learned model focuses on genes that characterize macrophages.

3.4 Loss-Function Learning Improves the Quantification of Small Cell Populations

Data for the mixtures of T cells, B cells, macrophages, endothelial cells, CAFs, NK cells and tumor/unclassified cells were generated in the same way as before. All cells except the tumor cells were used in X. This time the abundance of B cells in the simulated mixtures at 0 to 5 cells, 5 to 15, 15 to 30, 30 to 50, and 50 to 75 out of 100 cells is controlled. Not surprisingly, small fractions of B cells were harder to quantify than large ones. Deconvolution results for the standard deconvolution model with $g = 1$ calculated for every B cell content are shown as red diamonds in Figure 3.4 a. Loss-function learning improved the accuracy for all amounts of B cells, but the improvements were greatest for small amounts (Figure 3.4 a). With only 0 to 5 cells in a mixture the accuracy improved from $r = 0.22$ to $r = 0.79$. Furthermore, it could be observed that loss-function learning on small

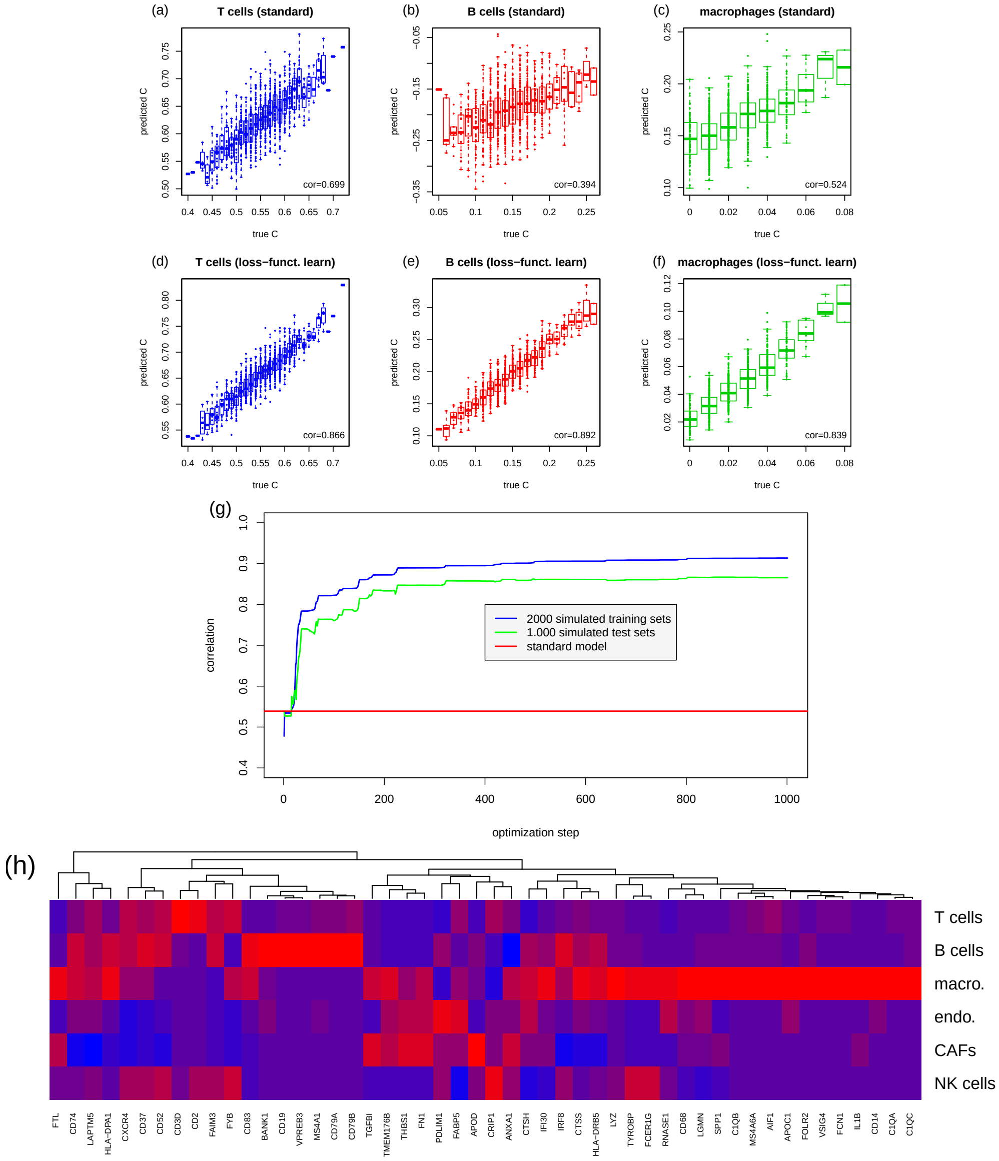


Figure 3.2: (a)-(c) Scatter plot for the validation sets for standard DTD with $g = (1, \ldots, 1)$. Loss-function learning improved the results for all cell types (d)-(f). (g) Average correlation versus optimization steps for loss-function learning. The blue line corresponds to the training data, the green line to the test data. For comparison the g=1 model, marked red, is included. Plot (h) shows the heatmap of the 50 most important genes (genes were ranked by $\hat{g}_i \times \mathrm{var}(X_{i,\cdot})$). The algorithm focuses on genes that separate T and B cells and macropages. Blue corresponds to low expression and red to high expression.

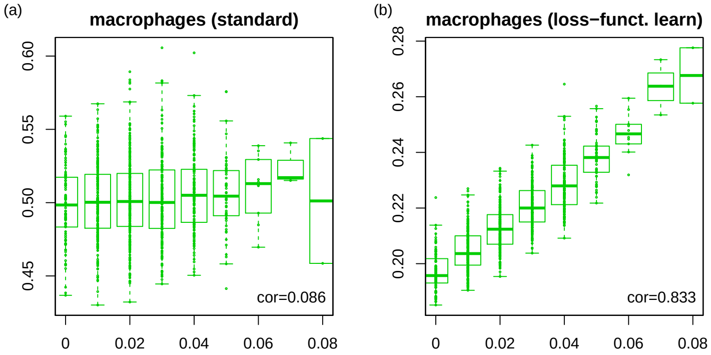


Figure 3.3: Deconvolution performance with only a single reference profile (macrophages). Predicted cell frequencies are plotted versus real frequencies. Results from the standard DTD model with $g = 1$ are shown in (a), for DTD with loss-function learning in (b).

B cell proportions yielded a model that was highly predictive of B cell contributions over the whole spectrum (Figure 3.4 a green stars). Furthermore the trained model got extrapolated to the sample mixtures containing a lower B-cell fraction of between 50% to 70%, (Figure 3.4 a orange crosses). Here the model lost performance compared to the loss-function learning model for every B cell step. With higher amounts of B cells in the mixtures it was not necessary for the model to specialize on small amounts of the cell type.

If the top-ranked genes of the model learned for the small B cell population are compared (Figure 3.4b) to that of the macrophage-focussed simulation (Figure 3.4 c), it can be observed that the former still comprises marker genes to distinguish all cell types, while the latter focuses on genes that characterize macrophages.

3.5 Loss-Function Learning Improves the Distinction of Closely Related Cell Types

The cell types that were annotated by Tirosh et al. [48] displayed very different expression profiles. If one is interested in T cell subtypes such as CD8+ T cells, CD4+ T-helper (Th) cells, and regulatory T cells (Tregs), reference profiles are more similar and DTD is more challenging. The fraction of annotated T cell profiles get subdivided as follows: all T cells with positive CD8 (sum of CD8A and CD8B) and zero CD4 count were labelled CD8+ T cells (1,130). Vice versa, T cells with zero CD8 and positive CD4 count were labelled CD4+ T cells (527). These were further split into Tregs if both their FOXP3 and CD25 (IL2RA) count was positive (64), and CD4+ Th cells otherwise (463). T cells that fulfilled neither the CD4+ nor the CD8+ criteria (411) contributed to the mixtures, but were not assessed by DTD. The reference matrix X, here consisting of T cells, B cells, macrophages, endothelial cells, CAFs and NK cells, was augmented by the selected T-cell types and thereby replaced the original T-cell labeling with the more specific profiles for CD8+ T cells, CD4+ Th and Tregs. Then 2,000 training and 1,000 test mixtures were simulated as outlined before.

For standard DTD with $g = 1$, correlation coefficients of 0.19 (CD4+ Th), 0.53 (CD8+), and

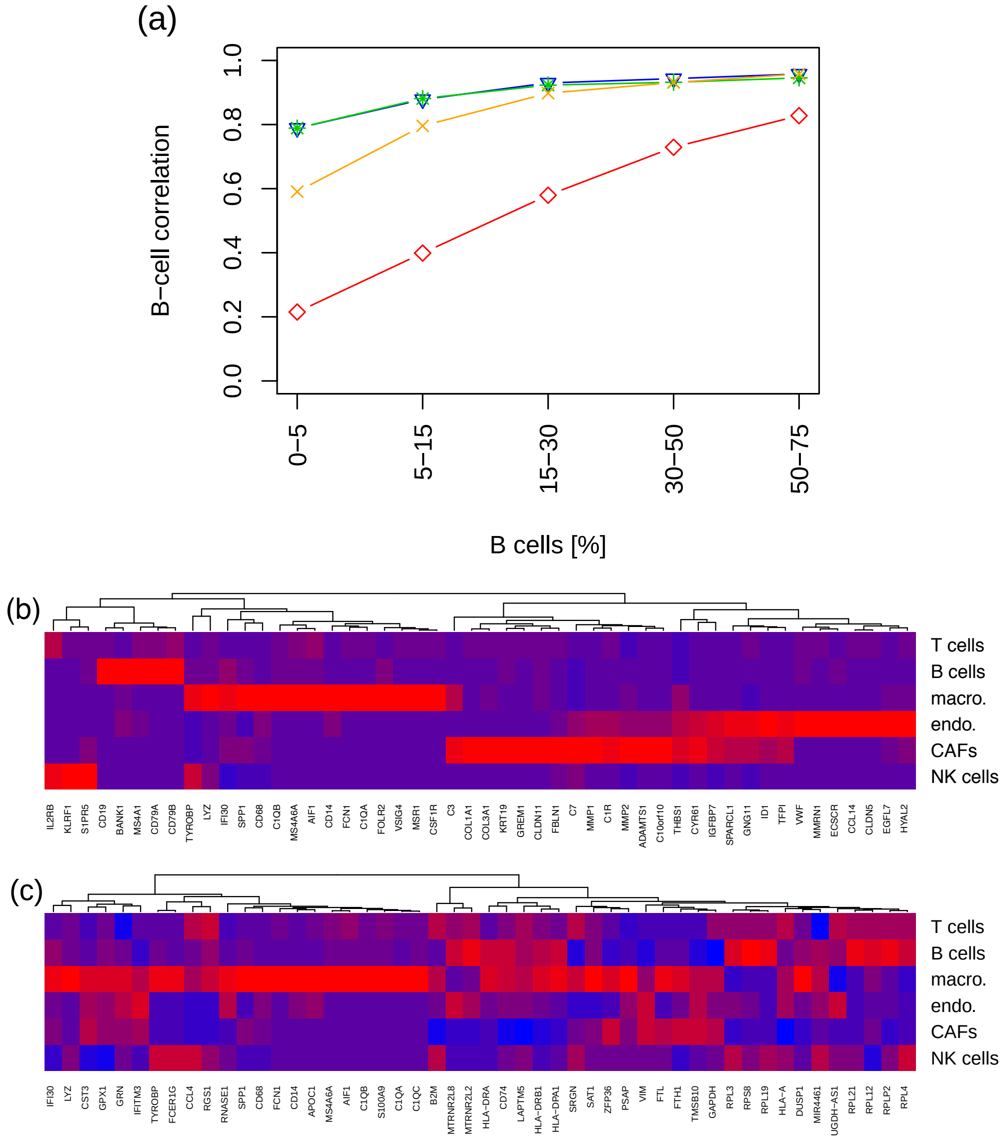


Figure 3.4: Plot (a) shows how the correlation between predicted and true cellular frequencies for B cells depends on the proportion of B cells. The blue triangles correspond to models from loss-function learning and red diamonds to the standard DTD model with $g = 1$. Furthermore, the green stars show how the model trained on mixtures with 0 to 5% B cells extrapolates to higher B cell proportions. The orange line in contrast was trained on mixtures with 50 to 75% B cells and extrapolates to lower B cell proportions. Plot (b) shows a heatmap of the 50 most important genes corresponding to the green star model (genes were ranked by $\hat{g}_i \times \text{var}(X_{i,\cdot})$). Plot (c) shows an analogous heatmap for loss-function learning on macrophages only. Blue corresponds to low expression and red to high expression.

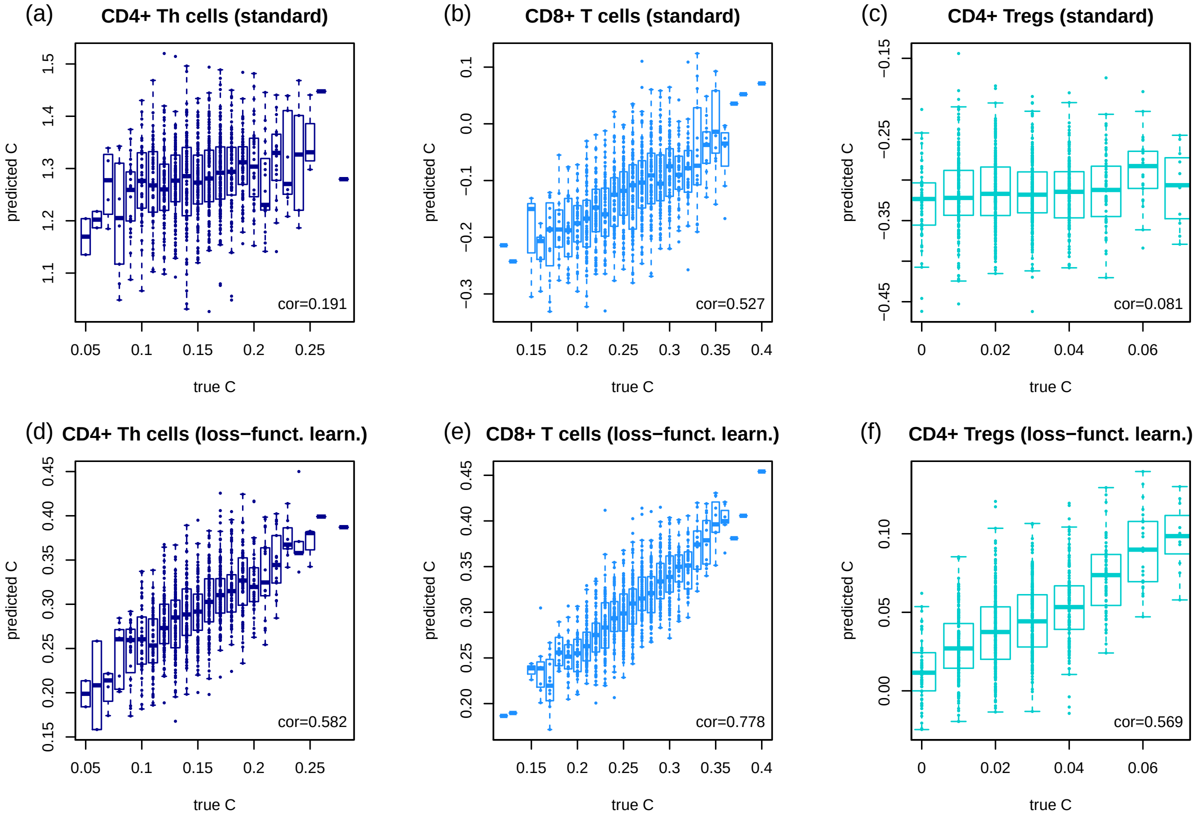


Figure 3.5: Deconvolution of T cell subentities. Results from the standard DTD model with $g = 1$ are shown in the upper row, plots (a-c), results from loss-function learning in the lower row, plots (d-e).

0.08 (Tregs) between true and estimated cell population sizes were observed. These improved to 0.58 (CD4+ Th), 0.78 (CD8+), and 0.57 (Tregs) for the here discussed method (Figure 3.5).

3.6 Loss-Function Learning is Beneficial Even for Small Training Sets, the Performance Improves as the Training Dataset Grows

The simulation in subsection 3.5 was repeated, but varied the size of the training data set. It was observed that loss-function learning improved accuracy for training data sets as small as 15 samples. Moreover, with more training data the performance improved and saturated only for training sets with more than 1,000 samples (Figure 3.6).

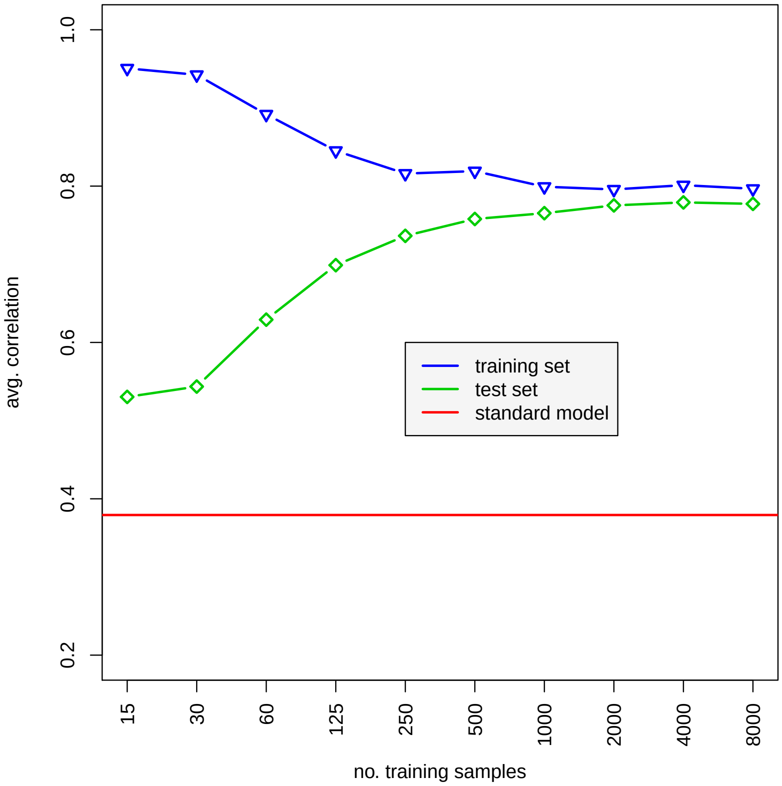


Figure 3.6: Performance with and without loss-function learning as a function training set size. The performance was assessed by calculating the average correlations between predicted and true cellular contributions over all cell types. The green and blue curve correspond to the performance of loss-function learning for validation and training mixtures, respectively. The performance of standard DTD with $g = 1$ is shown as a red line for the validation mixtures.

For 20 training sets the loss function and the average correlations for the different sizes of training sets were calculated. The test set contains always the same 1,000 bulk profiles (Figure 3.7). The training sets had a length of 8,000 simulated bulk samples. For 15 samples, the first 15 from the simulated bulks were selected, for 30 samples, the first 30 were selected and so on.

For small training sets the performance was sensitive to the individual simulation run, only as indicated by the large error bars in Figure 3.7. Here the calculated model adapted strongly on the given training set. This can be seen by the corresponding results for the test set, which also shows large error bars. However, for all evaluated simulation runs, the resulting model is better than the standard model (red line), even if only 15 training sets are used. With more training data the training performance decreased while the test performance increased. In both cases the error bars became tighter with more samples in the training set. Thus, for larger training sets convergence to a common value may be observed. Outliers with decrease performance are produced occasionally. It can be speculated that those may be contributed to the non-convexity of the Hessian (see section 2.3), where the model ends up in a local optimum which is not close to the global extremum.

Since our loss-function learning problem is not convex, the influence of the starting point g to the result of the deconvolution process was studied. For this purpose, the process was started at 20 different randomly calculated points ($g_{\text{start}=[0,1]^p}$). This was done again for training set lengths ranging from 15 up to 8,000 samples. The test set consisted of 1,000 simulated bulk mixtures and the loss-function learning was done for the 1,000 most variable genes, as previously. Figure 3.8 shows the results. For all randomly chosen starting points the deconvolution without loss-function learning led to compromised deconvolution results, as shown in red in Figure 3.8. After loss-function learning the results of training and test set improved for all starting points. As the yielded data of the learned models spread over a range of approximately 0.1 in correlation, the algorithm converges to different local optima. The correlation range in the test set turned out to be smaller than in the training set. In the later case all models led to similar correlation in the test set. To summarize, it was observed that the starting point exerts only a small effect on the deconvoulution results.

3.7 HPC-Empowered Loss-Function Learning Rediscovers Established Cell Markers and Complements Them by New Discriminatory Genes for Improved Performance

A final model, optimized on the 5,000 most variable genes gets introduced. For this purpose 25,000 training mixtures from the melanomas of the training data were generated. With standard desktop workstations the solution of this problem was computationally not feasible. A single computation of the gradient took 16 hours (2x Intel Xeon CPU [X5650; Nehalem Six Core, 2.67 GHz], 148 Gb RAM), and this needs to be computed several hundred times until convergence. Therefore a High-Performance-Computing (HPC) implementation of the code by parallelizing equations (2.3) and (2.10) with MPI, using the pbdMPI library ([71], [72]) as an interface was developed. Furthermore R was linked with the Intel Math Kernel Library for threaded and vectorized matrix operations. The algorithm was ran on 25 nodes of our QPACE 3 machine [73] with 8 MPI tasks per node and

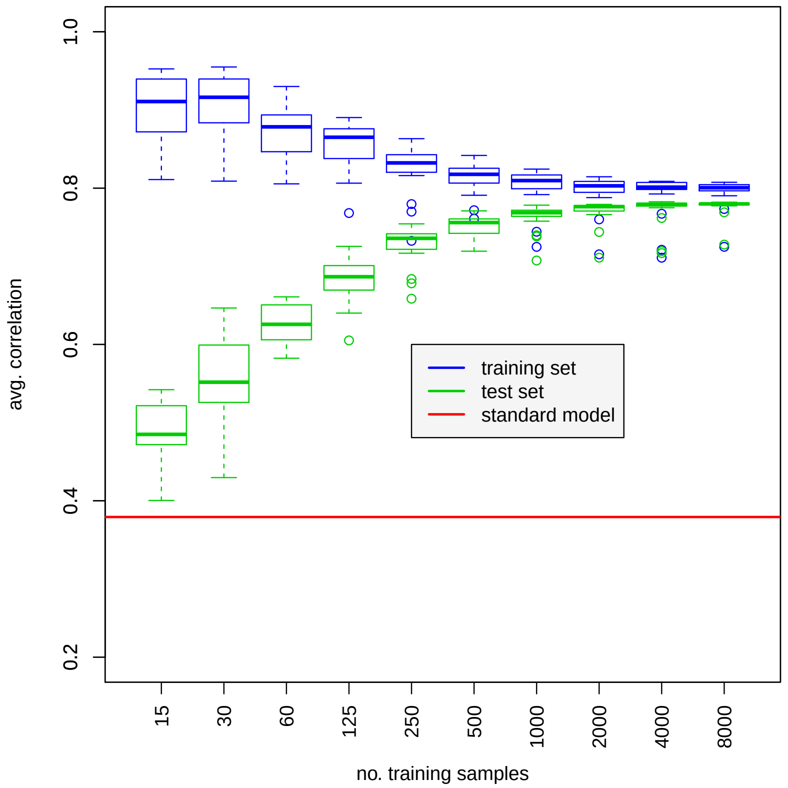


Figure 3.7: Performance with and without loss-function learning as a function of the size of the training set. Performance was assessed by calculating the average correlations between predicted and true cellular contributions over all cell types. Here, the calculations of Figure 3.6 were repeated 20 times, always with new training sets but with the same start vector $g = 1$. The performance of standard DTD with $g = 1$ is shown as a red line for the validation mixtures.

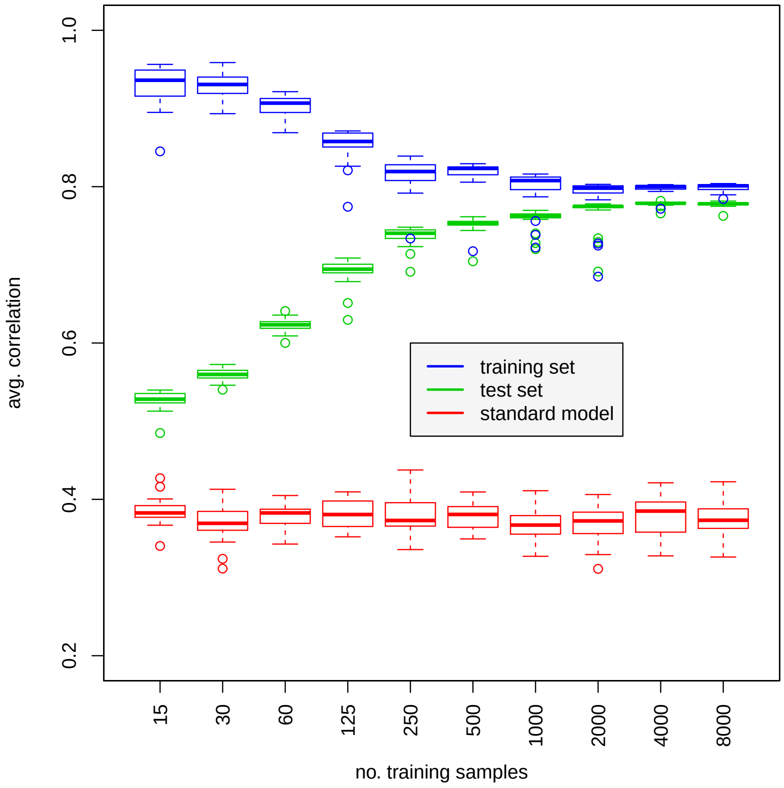


Figure 3.8: Loss-function learning for different starting points $g \in [0,1]^p$. The results for standard model (red), training sets (blue) and test sets (red) after loss-function learning for different length of the training set are shown. Always the same training set were used restricted to the respective number of training samples.

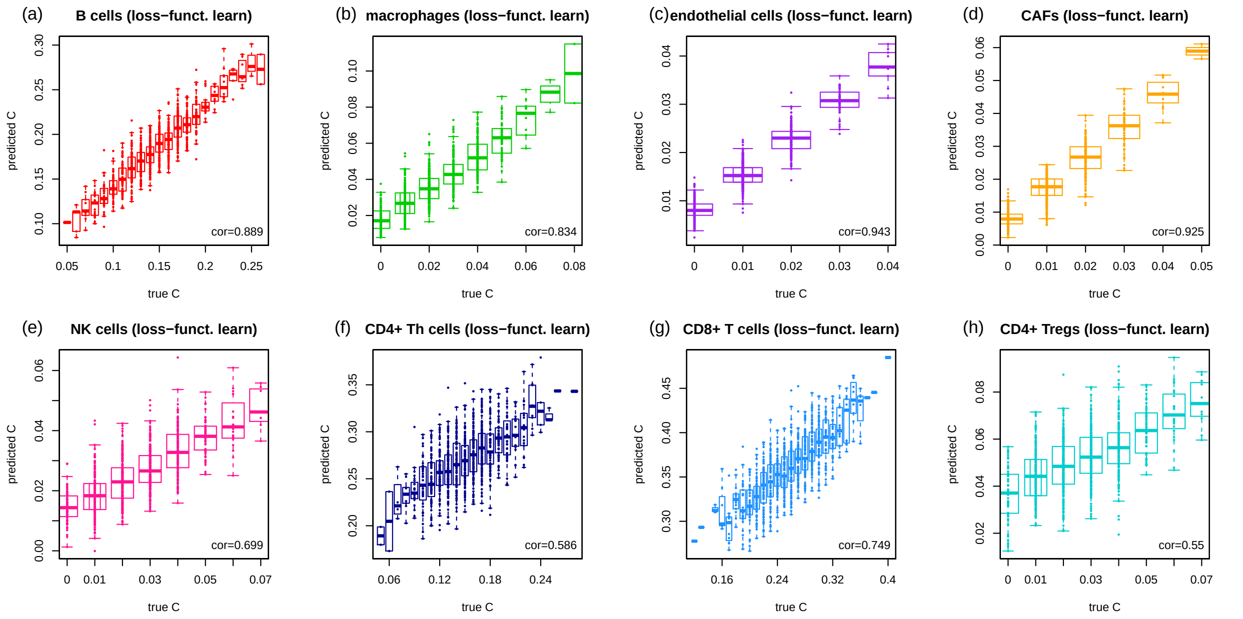


Figure 3.9: Scatter plots and correlation results for test sets for loss-function learning with the 5,000 most variable genes.

32 hardware threads per task, where each thread can use two AVX512 vector units. In 16 hours 5,086 iterations were finished, after which the loss defined in Equation 2.3 was stable to within 1%.

A test set consistent of 1,000 bulk profiles was simulated. The correlation results and scatter plots are shown in Figure 3.9. Despite for the two CD4+ T cell subtypes good deconvoluion results were achieved. The results for the CD+4 T cells were similar to the ones learned for the 1,000 most variable genes (see Figure 3.5). All cell types an their deconvolution results for both gene lenths are shown in Table 3.1. The comparison and discussion is in section 3.8. The high-performance model includes several genes, whose expression is characteristic for the cells distinguished in the present study. These include, among others, the CD8A gene, which encodes an integral membrane glycoprotein essential for the activation of cytotoxic T-lymphocytes [74] and the protection of a subset of NK cells against lysis, thus enabling them in contrast to CD8- NK cells to lyse multiple target cells [75]. As evident from Figure 3.11, NK cells are clearly set apart from all the other cell types studied by the expression of the killer cell lectin like receptor genes KLRB1, KLRC1, and KLRF1 [76]. B cells, on the other hand, are clearly characterized by the expression of (i) CD19, which assembles with the antigen receptor of B lymphocytes and influences B cell selection and differentiation [77], (ii) CD20 (MS4A1), which is coexpressed with CD19 and functions as a store-operated calcium channel [78], (iii) B Lymphocyte Kinase (BLK), a src-family protein tyrosine kinase that plays an important role in B cell receptor signaling and phosphorylates specifically (iv) CD79A at Tyr-188 and Tyr-199 as well as CD79B (not among the top 150 genes) at Tyr-196 and

Tyr-207, which are required for the surface expression and function of the B cell antigen receptor complex [79], and (v) BLNK, which bridges BLK activation with downstream signaling pathways [80]. The expression of FOXP3 is also highly cell specific. FOXP3 distinguishes regulatory T cells from other CD4+ cells and functions as a master regulator of their development and function [81]. Finally, CD4+ T-helper (Th) cells are distinguished indirectly from all the other aforementioned lymphocytes by the lack of expression of cell type-specific genes. In contrast to lymphocytes, macrophages, cancer-associated fibroblasts (CAFs), and endothelial cells, which line the interior surface of blood vessels and lymphatic vessels, are characterized each by a much larger number of genes. Exemplary genes include CD14, CD163, MSR1, STAB1, and CSF1R for macrophages. The monocyte differentiation antigen CD14, for instance, mediates the innate immune response to bacterial lipopolysaccharide (LPS) by activating the NF-κB pathway and cytokine secretion [82], while the colony stimulating factor 1 receptor (CSF1R) acts as a receptor for the hematopoietic growth factor CSF1, which controls the proliferation and function of macrophages [83]. CAFs, on the other hand, are distinguished by the expression of genes encoding extracellular matrix proteins such as fibulin-3 (EFEMP1), various collagens (COL1A1, COL3A1, COL6A1, COL6A3), versican (VCAN), a well known mediator of cell-to-cell and cell-to-matrix interactions [84] that plays critical roles in cancer biology [85], as well as the matrix metalloproteinases MMP1 and MMP2, two collagen degrading enzymes that allow cancer cells to migrate out of the primary tumor to form metastases [86]. Noteworthy is also GREM1, an antagonist of the bone morphogenetic protein pathway. Its expression and secretion by stromal cells in tumor tissues promotes the survival and proliferation of cancer cells [87]. Genes characteristic for endothelial cells include among others CDH5, a member of the cadherin superfamily essential for endothelial adherens junction assembly and maintenance [88], the endothelial cell-specific chemotaxis receptor (ECSCR) gene, which encodes a cell-surface single-transmembrane domain glycoprotein that plays a role in endothelial cell migration, apoptosis and proliferation [89], claudin-5 (CLDN5), which forms the backbone of tight junction strands between endothelial cells [90], and the von Willebrand factor (VWF), which mediates the adhesion of platelets to sites of vascular damage by binding to specific platelet membrane glycoproteins and to constituents of exposed connective tissue [91].

Of the top 150 genes shown in Figure 3.11, 28 genes were discussed. These genes have a total weight of 28% of all 5,000 gene weights (calculated as $\hat{g}_i \times \mathrm{var}(X_{i,\cdot})$). The developed algorithm complements this gene set with additional genes, including some that were not yet used to characterize cell types. An interesting example is CXorf36 (DIA1R), which has been described as being expressed at low levels in many tissues and deletion and/or mutations of which have been associated with autism spectrum disorders [92]. However, nothing is known about its function to date. Therefore, its observed overexpression in endothelial cells may provide an important clue for future study on its function.

3.8 Loss-Function Learning Results Depend on the Size of the Gene Space

The HPC calculated loss-function learned model for the 5,000 most variable genes were compared with the model for the 1,000 most variable genes that were obtained from calculations on a local desktop station. Table 3.1 gives an overview over the correlation results for training and test set in this scenario. The outcome for both loss-function learned models were averaged over all eight cell types. The computations for the overall correlation of the test set for 5,000 and 1,000 most variable genes turned out to be 0.772 and 0.776, respectively. The results for the different immune cell types fluctuated up to 4% (NK cells). On the training data the 5,000 gene model performed better than the 1,000 gene model. On the test data, however, this trend was not observed. There for every cell type the correlation results were better on the larger set than for the smaller gene set. On average, an improvement of 7.9% was recorded. Especially the smaller subpopulations, such as CD4+ Tregs, gave much higher correlation values. Interestingly, this trend was not observed on the test set. Thus, it may be assumed that the 5,000 gene model was overfitted in the test set.

cell type set	B	macro	endo	CAF	NK	CD4+ Th	CD8+	CD4+ Treg	mean
1,000, tr	0.880	0.902	0.908	0.904	0.671	0.670	0.846	0.539	0.790
5,000, tr	0.912	0.938	0.943	0.946	0.846	0.755	0.855	0.761	0.869
1,000, te	0.863	0.841	0.937	0.915	0.739	0.562	0.777	0.577	0.776
5,000, te	0.889	0.834	0.943	0.925	0.699	0.586	0.749	0.550	0.772

Table 3.1: Correlation results for training and test set using loss-function learning. Contrasted here are the results for the HPC model with 5,000 vs the local desktop calculations with the 1,000 most variable genes. The HPC model was calculated on 25,000 training and 5,000 test mixtures. For the local desktop model the standard 2,000 training and 1,000 test mixtures were used.

For several gene numbers the results of loss-function learning in training and test set, as well as the necessary computing time were compared. For gene numbers up to 1,000 genes the same simulated training and test set were used and restricted to the regarded number of genes. For comparison the results of training and test set for the HPC loss-function learning problem with 5,000 genes were plotted. The time component was neglected, as it was not possible to calculate this system on a normal desktop station. Figure 3.10 gives the results. The average correlation in the test set saturated with 1,000 genes (green). Due to overfitting effects the retrieved values for the training sets were still increasing (blue).

Concerning computation time, the calculation of the correlation was the most time consuming part for smaller number of genes. When the number of genes increased, the calculation of the cellular composition became more and more complex and resource consuming. Even if the overfitting effect for more than 1,000 genes was overcome by using more single cell measurements of more patients,

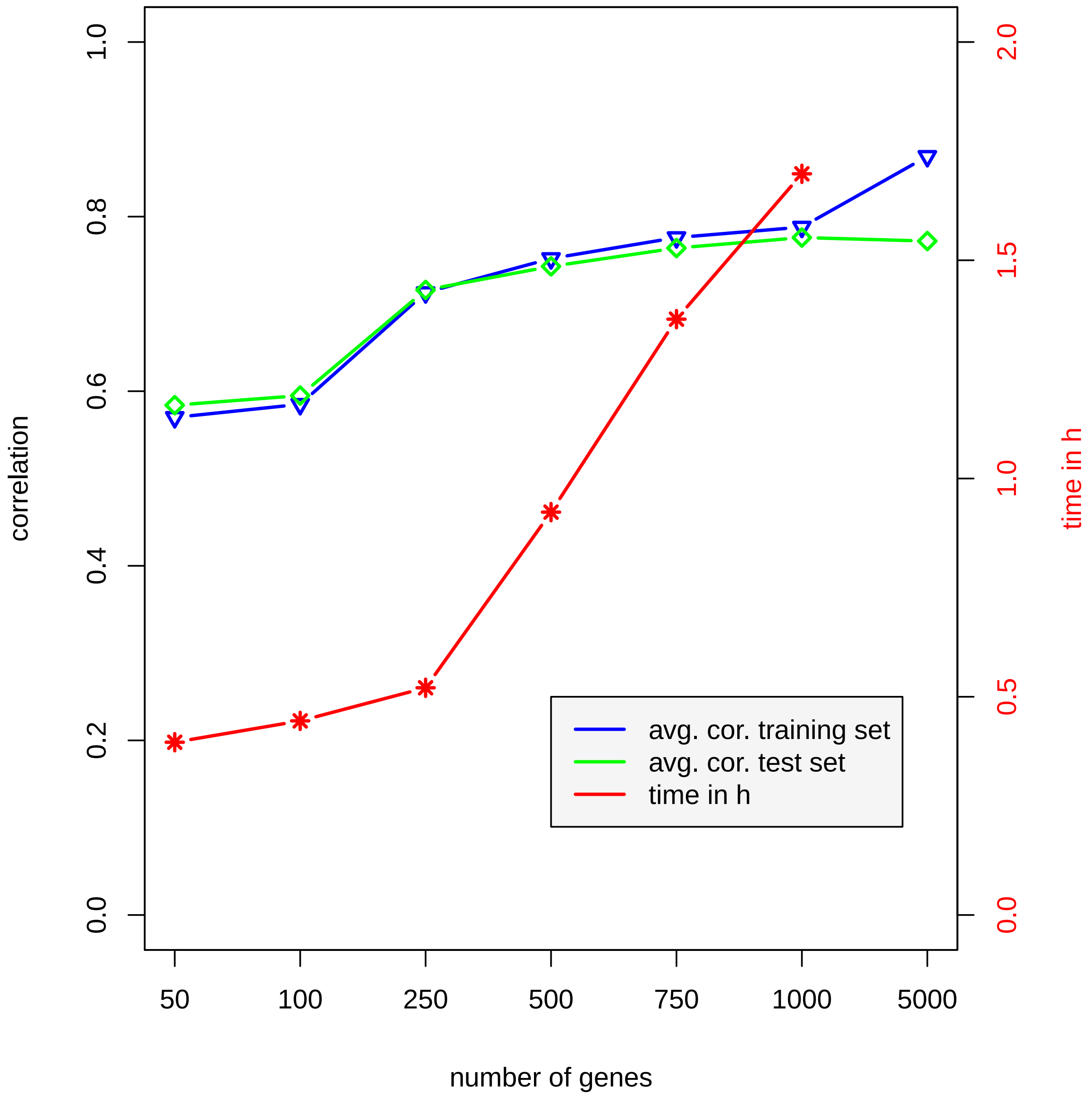


Figure 3.10: Average correlation/time plotted against the number of regarded genes in the loss-function learning problem for eight cell types. Blue: average correlation in the training set. Green: average correlation of the test set. Red: computation time. The correlation of the test set saturates for 1,000 genes and is compared to the training set. For the 5,000 genes 5,085 steps were calculated, for the other gene lengths 1,000 steps were computed.

the time component had to be kept in mind as it limits the maximal number of regarded genes in the calculations.

In summary, the deconvolution with more than 1,000 genes did not lead to better results than the HPC model with 5,000 genes. This was also the case for subtypes of CD4+ T cells, which showed the most compromised performance on test data. The performance gain on the training data did not persist on the test data.

3.9 Loss-Function Learning Shows Similar Performance as CIBERSORT for the Dominating Cell Populations and Improves Accuracy for Small Populations and in the Distinction of Closely Related Cell Types

Next the model trained in subsection 3.7 were compared to a competing method. For this, 1,000 test mixtures from the validation melanomas were generated. For comparison CIBERSORT [23] were chosen, because it was consistently among the best DTD algorithm in a broad comparison of five different algorithms on several benchmark data sets [23]. CIBERSORT were performed on the test mixtures, using two distinct approaches: first, the generated validation data were uploaded to CIBERSORT using their reference profiles. The specific cell types used by CIBERSORT are subsumed as follows:

- B cells: B cells naive and B cells memory.
- Macrophages: Monocytes, macrophages M0, macrophages M1, macrophages M2, dendritic cells resting and dendritic cells activated (as they belong to the mononuclear phagocyte system).
- Endothelial cells: Not available at CIBERSORT.
- CAFs: Not available at CIBERSORT.
- NK-cells: NK cells resting and NK cells activated.
- CD4+ T cells: T cells follicular helper, T cells CD4 naive, T cells CD4 memory resting and T cells CD4 memory activated.
- CD8+ T cells: T cells CD8.
- Tregs T cells: T cells regulatory (Tregs).

CIBERSORT further separates plasma cells, T cells gamma delta, mast cells resting, mast cells activated, eosinophils and neotrophils. For these cell types no corresponding cell types were labeled in the available melanoma data set.

The performance of CIBERSORT on the validation data is summarized in Figure 3.12 as CIBERSORT^a (yellow). It can be observed that the large population of B cells was estimated accurately, while smaller populations were inaccurate (NK cells, Tregs). Next, the melanoma our reference profiles were uploaded and used the CIBERSORT gene selection (CIBERSORT^b green) got used. It was found that highly abundant cell types (B cells and CD8+ T cells) were predicted with high accuracy. However, the distinction of similar cell types such as CD4+ T helper cells and Tregs was compromised, $r = 0.42$ and $r = 0.42$, respectively. Similarly, predictions for the small populations of CAFs were compromised. That might be explained by the fact that CIBERSORT does not take into account their distinction and thus appropriate marker genes might be missing. In a direct comparison to CIBERSORT the here presented method continuously showed similar or better performance.

Next, it was tested whether the developed method would have also worked for bulk profiles generated by a different technology than the reference profiles. The scRNASeq derived loss-function and the bulk profiles described above was used but the reference profiles in X were replaced by microarray data downloaded from the CIBERSORT web page. The microarray matrix X was rescaled such that the gene-wise means were identical to the scRNASeq data. Results are shown in Figure 3.12 in pink. Although accuracy was slightly reduced, the CIBERSORT results still get outperformed by the here presented method.

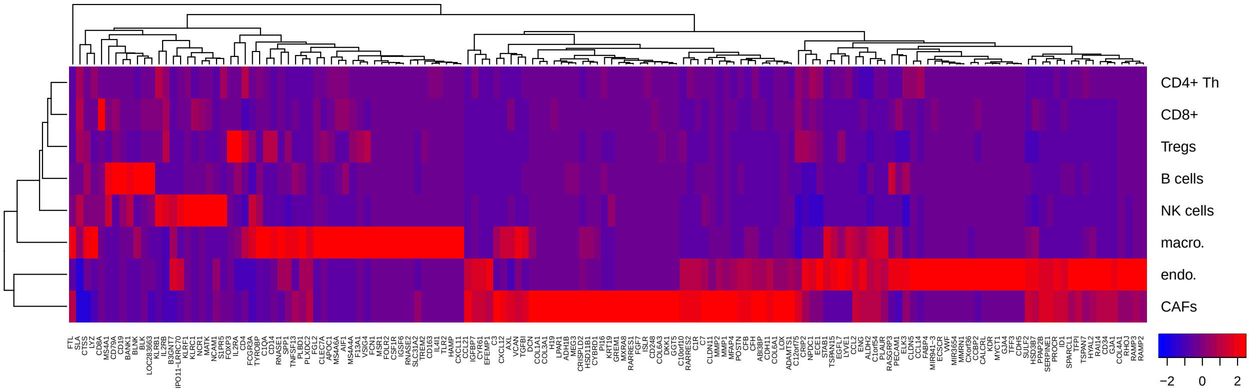


Figure 3.11: Heatmap of X for the features with the top 150 weights ($\hat{g}_i \times \mathrm{var}(X_{i,\cdot})$). Blue corresponds to low expression and red to high expression. The data were clustered by Euclidean distance.

The ranking of the predefined biomarkers in the CIBERSORT reference profile were compared with the highest weighted genes in the loss-function learning DTD models. In the HPC set of the 5,000 most variable genes 240 out of 547 genes of the CIBERSORT reference profile were present. This means the developed loss-function learning approach did not treat 307 genes which were included in the CIBERSORT standard program. Instead the loss-function learning model identified other available genes in the dataset as helpful for deconvolution. The gene weight was averaged

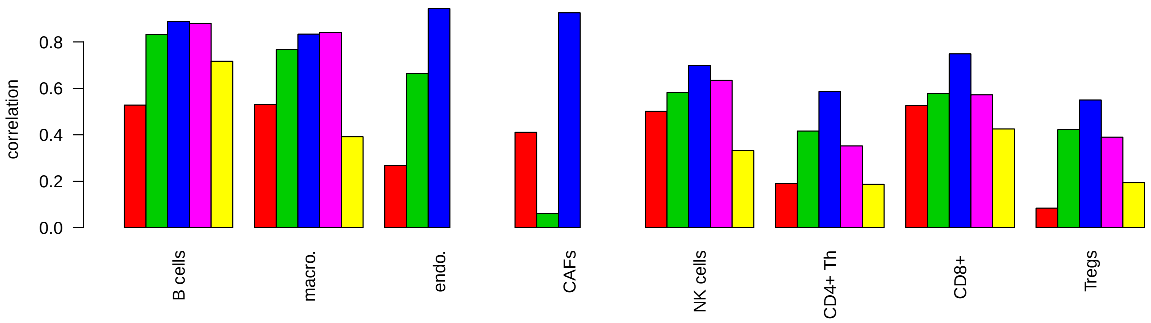


Figure 3.12: Performance comparison. The methods are from left to right: standard DTD with $g = 1$ on the 5,000 most variable genes (red), CIBERSORT^b (green), loss-function learning (blue), the study where bulk and reference profiles were generated with different technologies (pink), and CIBERSORT^a (yellow). Performance was calculated as correlation between predicted and true frequencies on 1,000 validation mixtures. Endothelial cells (endo.) and CAFs were not estimated by CIBERSORT^a and microarray reference profiles were not available. Thus no yellow and pink bars are shown.

over the full reference set, which yielded 0.0388 ± 0.109. In contrast, genes which were not included marked an average weight of 0.0159 ± 0.1250. This number also reflects the gene ranking resulting from the loss-function learning model. CIBERSORT genes were ranked 2.44 times higher than the other genes on average.

When considering the model from loss-function learning for the 1,000 most variable genes, only 109 genes of CIBERSORT were contained in the melanoma reference profiles. It was observed that CIBERSORT genes had an average gene weight of 2.911 ± 1.234 while the remaining genes had an average weight of 5.537 ± 5.905 .

3.10 Loss-Function Learning Improves the Decomposition of Bulk Melanoma Profiles

All mixtures discussed so far were artificial because only 100 single-cell profiles were chosen randomly. They might differ significantly from mixtures in real tissue. Therefore, 19 full bulk melanoma profiles were generated by summing up the respective single-cell profiles. These should reflect bulk melanomas [93]. The predicted results are contrasted with the true proportions in Figure 3.13. Only the predictions for Tregs were compromised with $r = 0.48$, while the predictions for all other cell types were reliable with correlations ranging from $r = 0.70$ (CD4+ Th) to $r = 0.99$ (CAFs) on the validation melanomas.

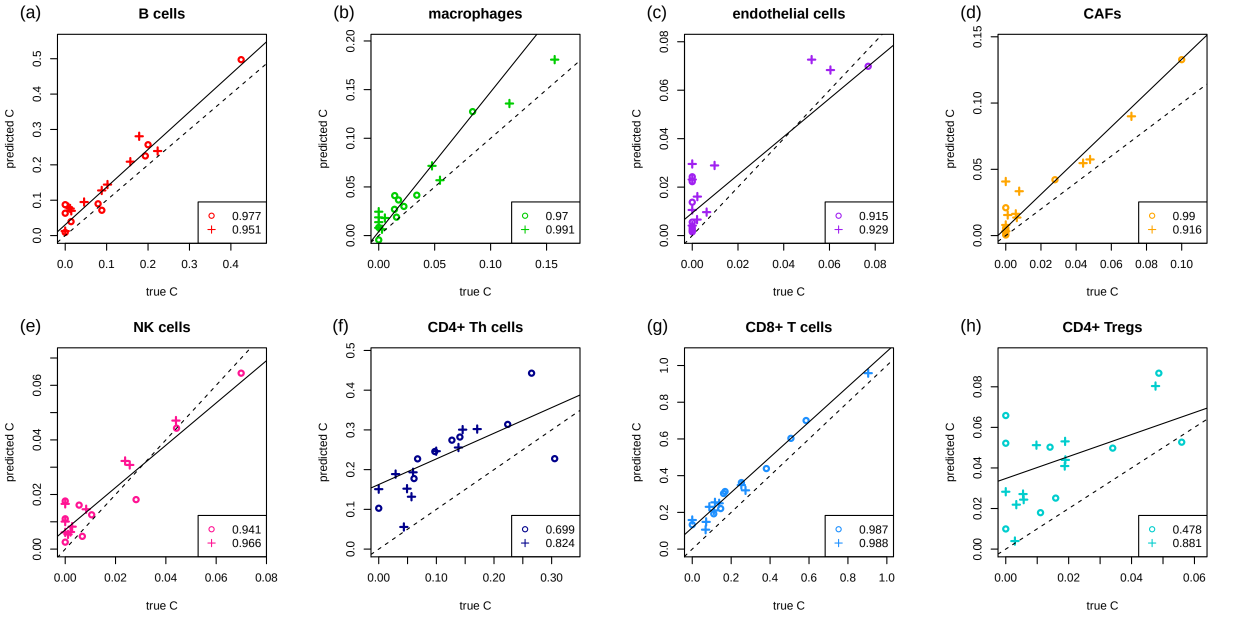


Figure 3.13: Deconvolution of melanoma tissues. The circles indicate melanomas from the validation data and plusses from the training data. Figure (a) to (h) correspond to B cells, macrophages, endothelial cells, CAFs, NK cells, CD4+ Th cells, CD8+ T cells, and CD4+ Tregs, respectively. The solid black lines show the corresponding linear regression fits on the validation data, the dashed lines the identity.

Chapter 4

Deconvolution of Blood Specimens from Patients with Chronic Lymphozytic Leukemia

This chapter and the following are built on a similar structure as chapter 3. First, the CLL-data set is introduced (section 4.1). Then, the characterization of single-cell measurements by t-SNE (section 4.2) are reviewed and the biological background of the relevant cell types is given (section 4.2). In section 4.4 follows the application of loss-function learning to the CLL data set. It is also demonstrated that the deconvolution model leads to reasonable biomarkers (section 4.5). Section 4.7 points out that the model from loss-function learning can be generalized to bulk gene-expression profiles of CLL blood specimens. The comparison with the state of the art deconvolution tool CIBERSORT follows in section 4.8. In subsection 4.9, the DTD model from the melanoma data set gets applied to the CLL data and vice versa. In chapter 5 follows the discussion of the results.

4.1 Description of the CLL Dataset

A data set of 43069 single-cell RNA sequencing profiles from CLL blood specimens was provided by Christian Schmidl [1][94]. The datas were pheripheral blood samples of routine examinations retrieved from patients. CLL is characterized by an over expression of CD19 and CD5. The data were measured using the 10x technology, which yields potentially lower read counts. Cells were collected at therapy start and then once or multiple times after 30, 120, 150 or 280 days following treatment with Ibrutinib. Ibrutinib is a BTK protein kinase inhibitor. It disturbs the survival signal of B and CLL cells, thereby inducing apoptosis in this cells. Ibrutinib also helps transporting B and CLL cells from the bone marrow and lymph nodes into the blood stream, which contributes to break up nests of malign cells in the tissue. All afflicted patients were treated with other drugs

[1] Mail: christian.schmidl@ukr.de

before starting on Ibrutinib as a single agent therapy.

The sampling frequency and it's corresponding time points are shown in Fig. 4.1 as well as in the first column of Table 4.1. There were several cells, defined by only a few hundred genes which showed non-zero count. As a consequence, cells with less than 200 detected genes were discarded from further analysis. The single cells were distributed over the cell types as follows: Tumor cells (27285), CD4+ T cells (1811), CD8+ T cells (8420), monocytes (2233), natural killer (NK) cells (1024), nurse like (NL) cells (388) and cells of unknown type (1908). The data were normalized to transcripts per million (TPM). Table 4.1 illustrates the distribution of single cell measurements over different patients and time point of measurements.

cell type patient	CLL	CD4+	CD8+	mono.	NK	NL	unknown	sum
PT 1, d 0	2674	12	55	20	2	6	10	2779
PT 1, d 120	921	42	27	26	3	0	57	1076
PT 5, d 0	6361	66	254	33	52	6	62	6834
PT 5, d 30	3170	539	2292	338	325	62	419	7145
PT 5, d 150	477	710	3301	789	283	39	468	6067
PT 6, d 0	1103	69	994	788	52	227	329	3562
PT 6, d 30	4967	73	510	22	24	4	114	5714
PT 6, d 120	1877	21	161	18	4	1	30	2112
PT 6, d 280	3048	91	470	73	13	9	66	3770
PT 8, d 0	1876	26	115	122	75	33	15	2262
PT 8, d 30	465	12	82	2	37	0	12	610
PT 8, d 120	346	150	159	2	154	1	326	1138
total	27285	1811	8420	2233	1024	388	1908	43069

Table 4.1: Distribution of single-cell CLL data from different patients and time point of measurements. PT abbreviates patient and is followed by its assigned number. d quantifies the number of days after treatment. Further short notations are: CLL for tumor cells, CD4+ for CD4+ T cells, CD8+ for CD8+ T cells, mono for monocytes, NK for natural killer cells and NL for nurse like cells.

There are more single cell measurements from patient five and six than from patient one and eight. Note that, as a consequence, there is a dominant contribution from patients five and six and their specific characteristics in the calculations. In order to compensate for this overcontribution, one patient with high and one with a low number of measured cells was paired in the training and validation sets. Thus, the training mixtures and reference profiles were dominated by patients six, the validation set by patient five. Depending on the method of measurement the gene count per cell might retrieve several zero count entries.

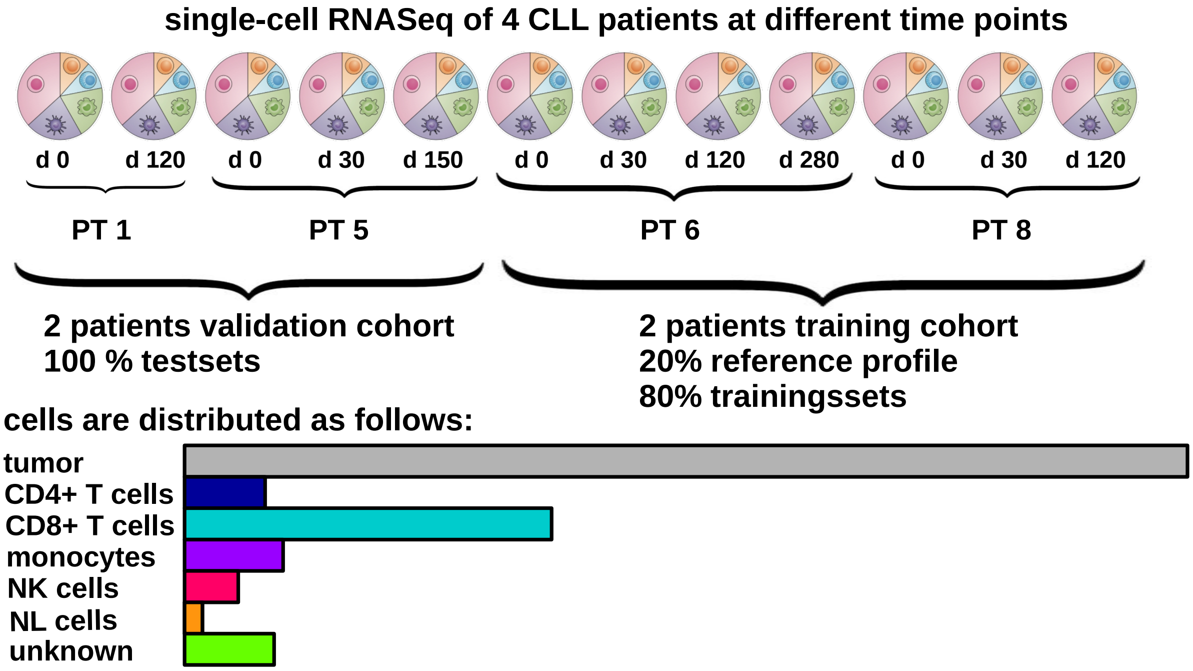


Figure 4.1: In the upper part of the graph distribution of the patients in the training and validation cohort as well as the different measurement points for each patient are visualized. In the lower part of the illustration, the distribution of the single cells over the different cell types is shown. Note, the presence of a high amount of tumor cells compared with the number of immune cells.

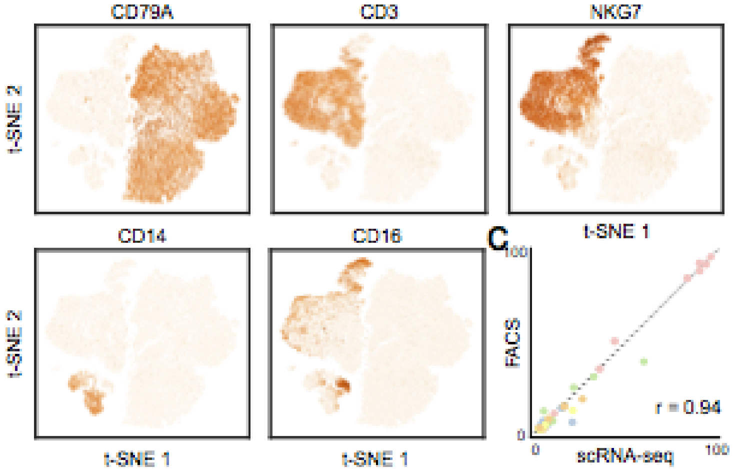


Figure 4.2: The first five pictures show the t-SNE clustering of the single cell RNASeq measurements with pigmented marker genes (CD79A, CD3, NKG7, CD14 and CD16). In the last picture the results of the RNASeq clustering are plotted against the results of cell type determination by FACS.

4.2 Classification of Single-Cell RNASeq Data in Cell Types by t-SNE

T-SNE is a method to reduce high dimensional data to lower dimensional representations [95]. Here, the 2D representation was used to separate the data into clusters. In the regarded case the high dimensional space is given by the number of genes. In figure 4.2 the results of the t-SNE clustering are shown. t-SNE 1 and t-SNE 2 are the coordinates of the two dimensional space. The single cell measurements were pigmented by several marker genes to determine the corresponding cell type of the observed clusters.

The cluster of CLL cells was characterized by a high expression of CD79A, which is a popular marker for B cells. The cluster of CD4+ and CD8+ T cells were characterized by a high expression of the marker gene CD3. Further, the monocyte cluster was identified through the expression of CD14 and the NK cluster by expression of the NKG7 gene. For a description of CLL and the immune cell types of this data set see section 4.3.

4.3 Chronic Lymphocytic Leukemia (CLL) and Characterization of the Cell Types by Single Cell RNASeq Measurements

An accumulation of monoclonal CD5+ mature B cells in lymphoid tissues, peripheral blood and bone marrow [96] leads to one of the most common B cell malignancies in older adults: chronic lymphocytic leukemia (CLL) [97], which is a low-grade, leukemic B-cell-non-hodgkin-lymphoma. The CLL arises by clonal augmentation of mature and small celled B lmphocytes. These accumulated

CD5+ B cells are resistant to apoptosis [98]. It is assumed that genetic variations are the activator of CLL. CLL is not curable by therapy with antibodies or by chemotherapy.

The malignant CLL cells and the surrounding tissue have an extensive interaction [99]. This interaction is decisive for their survival and marks the resistance to therapy and the generation of a milieu which suppresses the immune system [100]. When CLL proliferates in the tissue, the interactions of the tumor cells are taking place in particular with nurse like (NL) cells [101].

In general, nurse cells help other cells, provide food and stability to the cells in their surrounding environment. As specialized macrophages, they assist in the development of new red blood cells in the bone marrow, helping them to mature. In CLL, the NL cells differentiate from CD14+ cells [102] into large, round and adherent cells [98]. These NL cells protect the malignant CLL cells also from apoptosis [101, 103–105]. The expression of CD68 [102] and CD163 [106] is characteristic for NL cells. Their gene expression pattern resembles that of tumor associated macropages [106]. Like these macrophages in solid tumors, it is expected that NL cells mediate the resistancy to chemotherapy [107] and have, like in other cancers, an influence on overall and progression free survival of patients [108, 109].

Monocytes are the largest type of leukocytes. They can differentiate into macrophages and myeloid lineage dendritic cells. In healthy people the monocytes get instructed by B- and T- cells to eat malignant cells. In CLL, immunosuppressive genes, for instance PTGR2, RAP1GAP or CDC42EP3 [77], are alternated. Furthermore genes which are associated with phagocytosis and inflammation are deregulated in monoyctes [77]. Additionally the proliferation of T cells is blocked by the contact with these altered monocytes in the CLL patients [77].

T cells are white blood cells and part of the immune defense. Together with B cells, the T cells constitute the acquired immune response. They migrate through the organism and control the membrane receptors of other cells in order to find morbid transformations. There are several different subgroups of T cells, here CD4+ and CD8+ T cells are considered. In general, the CD4+ T cells are helper cells. When they recognize an ill cell, they use cytocines to call other immune cells for help, where as the CD8+ T cells kill the affected cell directly.

In CLL the immune functions of the T cells are downregulated. The T cells exhibit an irregular distribution of subtypes, showing higher expressions in their immune checkpoints and a higher amount of proliferated cells than T cells in normal tissue. Due to disease activity and disease treatments the T cell profiles in the CLL patients differ substantially [110].

Finally there are natural killer (NK) cells. They belong to the lymphocytes and can identify abnormal cells, like tumor cells or virus infected cells, and kill them [66]. The NK cells are part of the native immune response. They recognize bacterial and fungus cell walls and annihilate invading those cells. In CLL, the NK cells are exposed to a high amount of tumor cells. As a consequence their phenotype and function is altered. The patients express higher numbers of NK cells, but these are less mature. They show signs of being worn out and the cell degranulation process gets dysfunctional [111].

4.4 Loss-Function Learning Applied to the CLL Dataset

The second two patients (patient six and eight) of the CLL data set were used for creating a training set and reference profiles. Data from the other patients (one and five) were used for creating a test set only. Training and test set as well as the reference profiles for the different cell types were created as described in section 3.1. For the reference profile 20% of the single cells were drawn randomly from every cell type, summed up and normalized. As for the melanoma data set, 1,000 normalized bulk profiles in the test set and 2,000 in the training set were created. For calculation of the loss-function learned model only the five non-malignant known cell types (CD4+ and CD8+ T cells, monocytes, NK and NL cells) were used in the reference profile. Thus, about 70% of the cells in the mixtures were not covered by reference profiles. For calculations the 1,000 genes with the highest variance were used. The loss-function learned model converged in the training set to a mean correlation over the five considered cell types of 0.841 after 1,000 optimization steps. For the test set the obtained correlation was 0.738. For the standard model a correlation of 0.347 (training set) vs 0.227 (test set), respectively, was obtained. In Figure 4.3 the convergence of training and test set for the loss-function learning problem is shown. Here the average correlation is plotted against the calculation steps. The curve of the test set (green) remains slightly below the training set (blue), which could be expected since the loss-function learning model adapts to the training set. Results for standard and loss-function learning model are shown in Figure 4.4. In the first row the deconvolution results for the standard model of the test set are shown (picture (a) - (e)). Results for the training set after loss-function learning are shown in the second row (picture (f) - (j)). Results of the validation set can be found in the last row (picture (k) - (o)). For all cell types better results for the learned model were achieved than for the standard model. Particularly for CD4+ T cells, NK and NL cells the improvements in the training set were remarkable. Here, the standard model was not able to detect them accurately, quite in contrast to the loss-function learned model. The result for the CD4+ T cells in the validation set was surprising. In the training set high improvements for this cell type compared to the standard model could be observed, however in the test set the performance was much lower. The correlation score matched the success of the training set roughly only by half. Due to the strict segmentation of patient data in training and test set, it is likely that the CD4+ T cell subtypes were unevenly distributed over the two sets. Thus, the reference matrix, which was constructed from the same patient data as the training set, did not equally represent the CD4+ T cells in the test set. Different subgroups of the CD4+ T cells like conventional CD4+ T cells (Tconv), regulatory T cells (Treg) or activated conventional CD4+ T cells (act Tconv) may have been present in the different patients. In the calculations however, the training and validation set was dominated by one patient only. Therefore it can be speculated that the underlying CD4+ T cell distribution between both patients were not cohorent. To test this hypothesis patients were not separated in training and test set. All single cells were combined together in one pool. Out of it 20% of every cell type were drawn for creating the reference profiles. The rest was separated in two equal parts. One was used for creating the test bulk profiles, the other for the training ones. With this procedure it was ensured that the biological variability of the CD4+ T cells was captured each in the reference profile, training and test set.

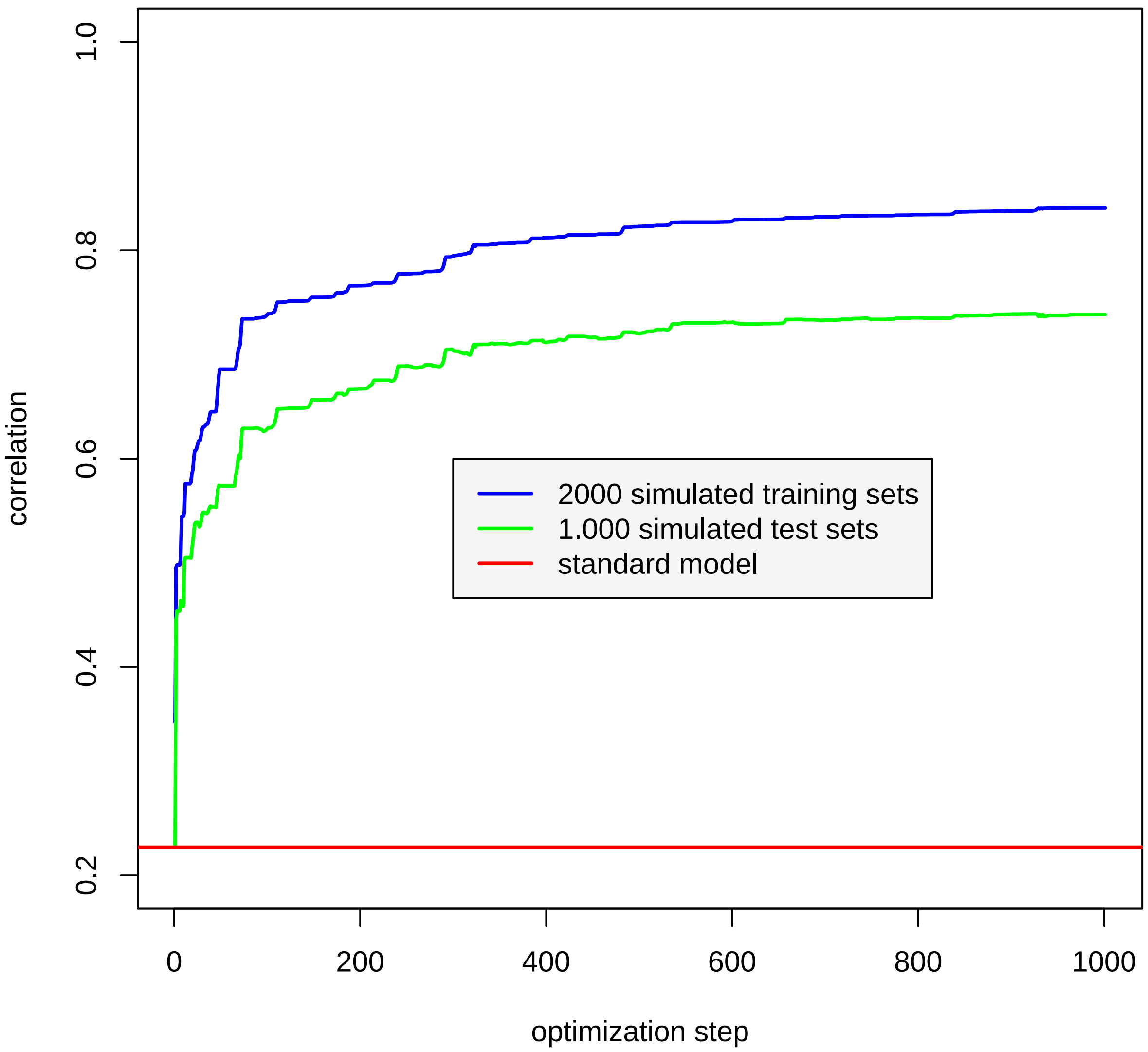


Figure 4.3: Visualization of minimization for loss-function learning, where patient data were strictly separated into training and validation data. Loss-function learning was calculated for all five immune cell types. The mean correlation over all cell types is plotted against the respective optimisation steps. The red line is the mean correlation of the standard model, calculated for the validation set. In blue the correlation for the training set is shown. The mean correlation of the test set is shown in green. In both cases the correlation saturated for high numbers of optimization steps.

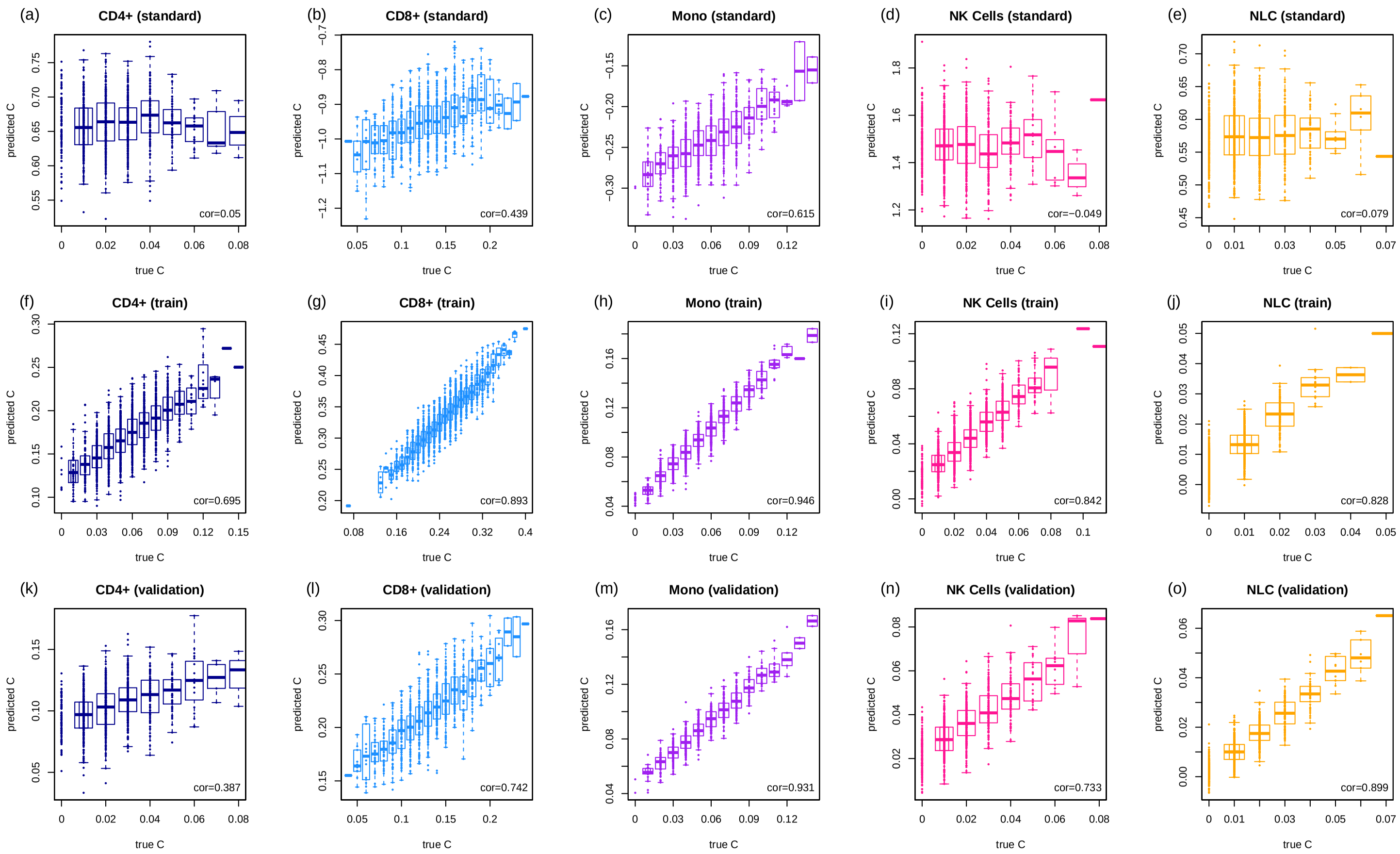


Figure 4.4: Individual loss-function learning calculation results for training and validation set for the five immune cell subtypes of the CLL data set. (a) to (e) show the correlation between real and predicted cellular distribution for the standard model ($g = (1, \ldots, 1)$) for all different cell types. (f) to (j) give the corresponding results for the loss-function in the cases when strong discrepancies between the performance on the training and validation data were observed. In the last row ((k) to (o)) loss-function learning results for the validation set can be seen. For CD4+ T cells after loss-function learning of training and validation set high discrepancies were found.

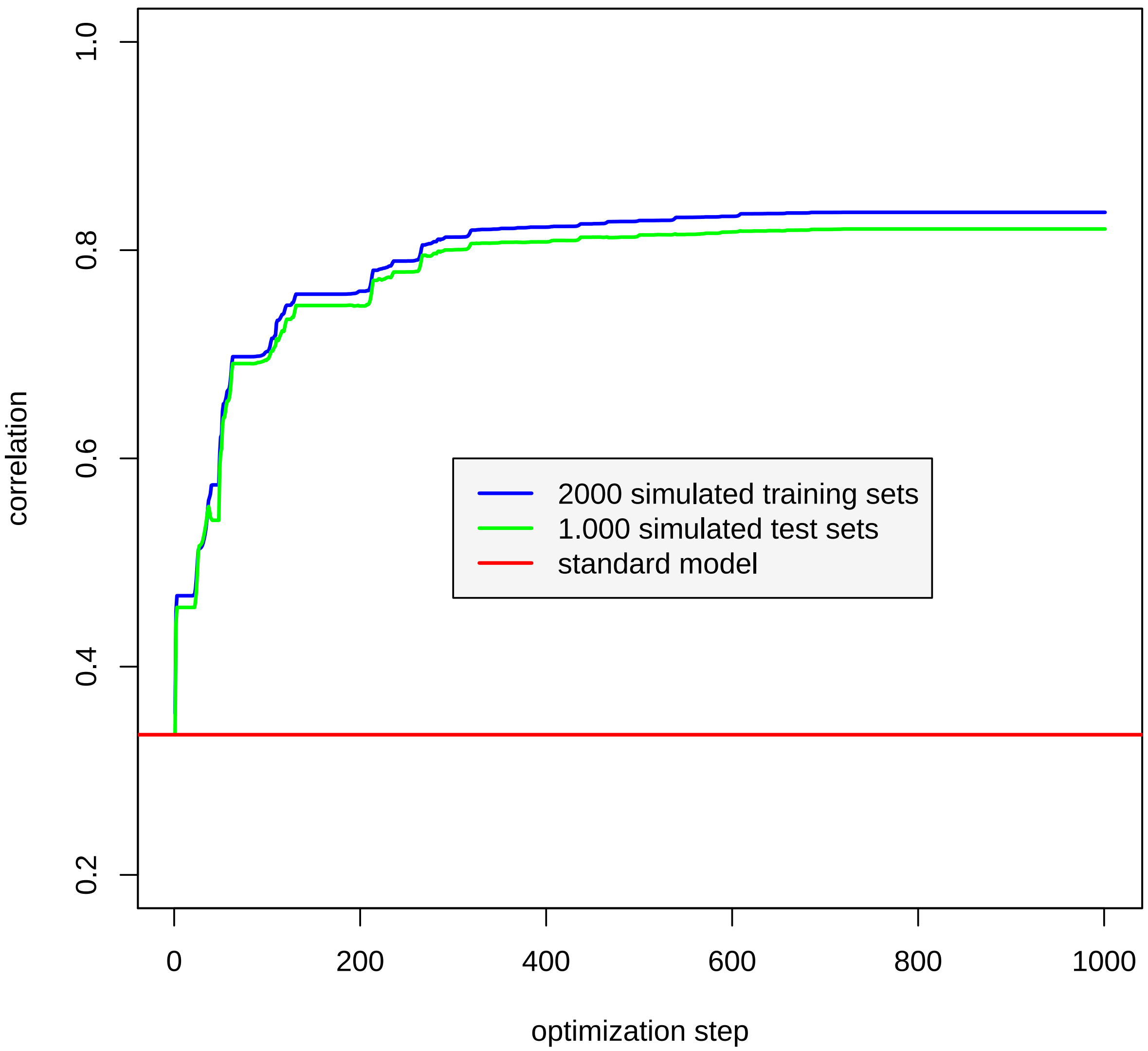


Figure 4.5: Visualization of minimization for loss-function learning, including all five immune cell types, where the single cells were separated randomly in reference profiles, training and validation set. The mean correlation of all cell types in the reference matrix is plotted against the respective optimization steps. The red line is the mean correlation of the standard model, calculated for the validation set. In blue the correlation for the training set is shown. In the beginning, high improvements are observed, while the loss function saturates for high numbers of optimization steps. The mean correlation on the test set is shown in green. Here, the differences between the results of training and test set are much smaller than previously in Figure 4.3.

Figure 4.6 shows the results for the standard model on the test set in the first row, as well as the loss-function learned model on the training and test set in the second and last row. The results for CD4+ T cells in the training set was performing slightly lower than the CD4+ result, calculated with the strictly separated training and validation group (see Figure 4.4). Now the higher variability of the single CD4+ T cells, utilized for creating the training sets and the reference profile, prohibited a higher adaption of the loss-function learning model to all of the CD4+ T cell subtypes of all four patients. As opposed to the strictly separated case, a specialization of the loss function to one or two subgroups of the CD4+ T cells was not possible any more. In the test set it was the other way round. Here much better results could be achieved due to the better coverage of the CD4+ T cell subtypes by the reference profile and the training set. The CD8+ T cells and the NK cells also showed improvement in the test set for the randomly mixed test and training set compared to the strictly separated case. They turned out to produce similar correlation results for the training set in both cases. However, their values for the test set proved to be significantly better. On the other hand results for monocytes didn't change qualitatively, which supports the hypothesis of subtype mingling. For NK cells no similar conclusions could be drawn due to their overall low number. In their case the test results turned out to be even better than in the training set, compared to the strictly separated case.

In summary, the predictions for the test set were better than previously in the strictly separated case (0.822 vs. 0.738). This can also be verified from the convergence graphs in Figure 4.5. Here the optimization results for the test set (green) are closer to the training set (blue) when compared to the graphs for the strictly separated test and training set (see Figure 4.3).

To further examine the hypothesis of diverse, patient-specific CD4+ T cell subtypes information from other CD4+ T cell entities is required. These data, however, are currently not available from the t-SNE labeling and so the task is left for future investigations.

The observation that one gets better results when the training and validation cohort are not strictly separated by patient is intuitively clear. However, the strength of this effect was surprisingly high, suggesting that also other sources of inconsistencies need to be taken into account, like for instance, batch effects.

4.5 Loss-Function Learning is Able to Detect Known Biomarkers

As in section 3.7, again the 50 most important genes of the loss-function learning model calculated from all available single-cell RNASeq measurements were analyzed. The genes were ranked from the CLL model by the following score: First, the variance for each gene in the reference matrix was calculated. Then, the variance was multiplied with the gene weight g_i to form the trained model:

$$\text{score}_i = g_i \times \text{var}(X_i). \tag{4.1}$$

In Figure 4.7 the results for the 50 most important genes are shown. The left side of the heatmap a dendogram shows that the cell types cluster into two groups. One consisting of NL cells and monocytes and the other one of CD4+ and CD8+ T cells together with the NK cells. The calculated

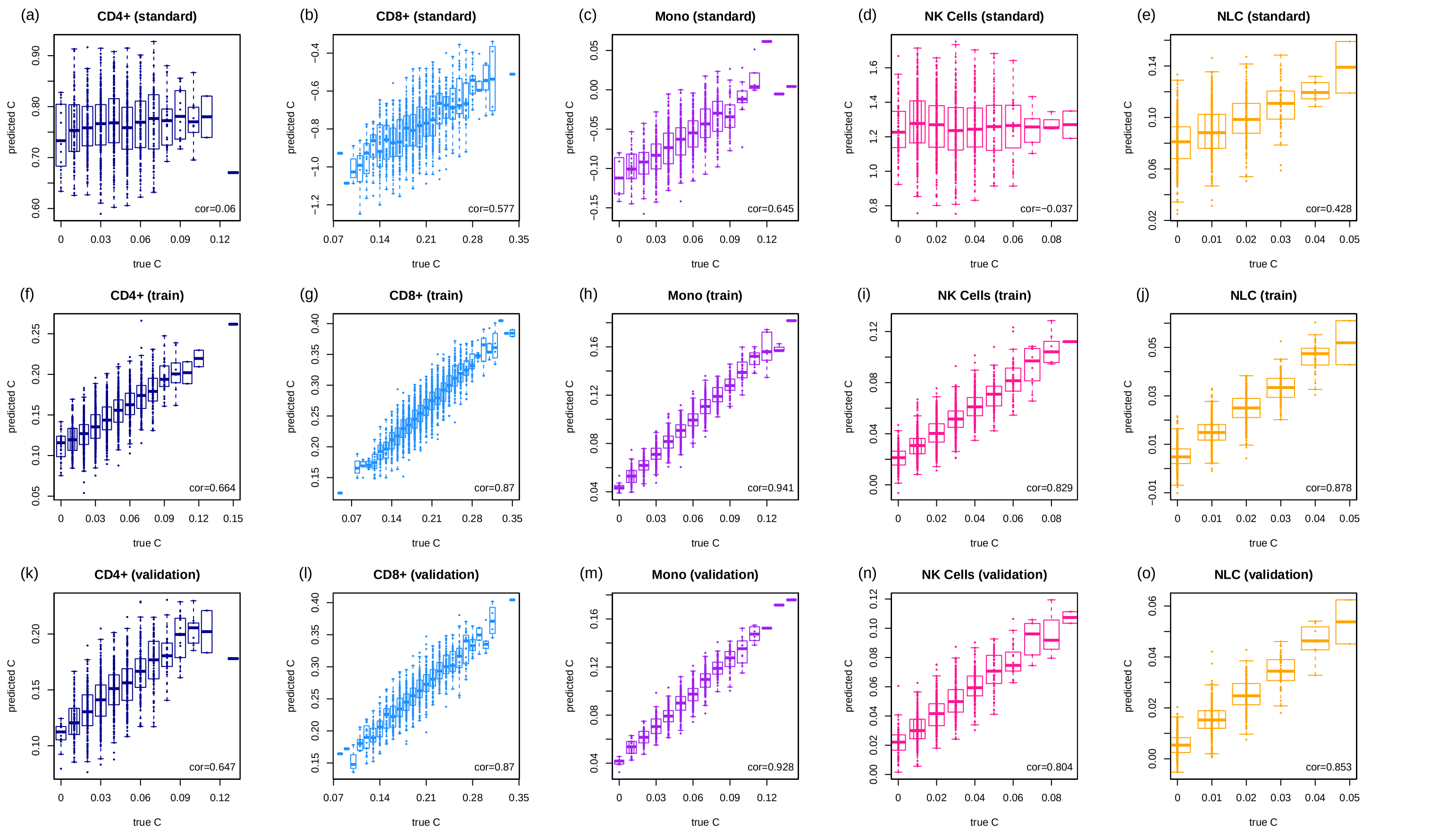


Figure 4.6: Training and validation set were created from all four patients. They were not strictly separated into test and training set, although the single-cell profiles were kept separated accordingly. The results for the CD4+ T cells after loss-function learning for training and test set support the hypothesis of diverse, patient-specific, CD4+ subpopulations. (a) to (e) show the correlation between real and predicted cellular distribution for the standard model ($g = (1, \ldots, 1)$) and for the test set for different cell types. (f) to (j) show the outcome for the loss-function learned model on the training set. In the last row ((k) to (o)) the results of the test set using loss-function learning are shown. In all graphs the predicted cellular composition is plotted against the true one.

model includes several genes, whose expression is characteristic for the cells distinguished in the present study. Below some of the most important genes for every cell type are discussed.

The NL cells, which contribute to the microenvironment of CLL, express, among others, CDKN1C. CDKN1C is upregulated in CLL lypmhocytes cultured with NL cells [112] and it is suggested to act as a tumor suppressor gene. CD68 is a biomarker for NL cells in CLL and their tissue-associated counterpart, the macrophages [113]. It is a marker of survival chance in cancer patients [114, 115]. NK cells are clearly set apart from all the other cell types by the expression of the killer cell lectin-like receptor genes KLRB1 and KLRF1 [76, 116] and by an overexpression of the CST3 gene, which occurs in both monocytes and nurse like cells, in the CLL samples. The overexpression of CST3 in CLL, compared to normal B cells, may affect the regulation in the immunoreactivity process and in protein degranulation [117]. Another exemplary gene in NL cells and monocytes is the allograft inflammatory factor 1 (AIF1), which is highly expressed in activated macrophages around inflamed tissue. AIF1 appears in the macrophages related cell types monocytes and NL cells. The B cell receptor (BCR) plays an important role in the interaction of the microenvironment of germinal centers with B cells. The germinal centers account for proliferation. It is believed that the BCR contributes to pathogenesis and clinical evoluion in CLL [118]. BCR activates, among other genes, FOS which is overexpressed in monocytes and NL cells [119, 120] and the C10orf54 gene [121]. FOS proteins regulate proliferation, differentiation and transformation of cells. The expression of FOS has also been associated with apoptosis in some cases [122]. The C10orf54 gene is an immunoregulatory receptor which inhibits the T cell response [123].

Moncytes are set apart from the NL cells for example by expression of MS4A6A and LGALS2. In CD16+ monocytes MS4A6A, which is necessary for signal transduction, is higher expressed than in other monocyte subtypes [124]. In CLL patients higher numbers of CD16+ monocytes were detected compared to healthy patients [125]. A second, very important, gene marker for monocytes is LGALS2. T cells, NK cells, macrophages, neutrophil granulocytes and other immune cells are affected by this gene. It influences immune surveillance, molecular trafficking, apoptosis, metastasis, inflammation and lots of other critical functions in cancer biology. As it supports cancer survival, this molecule constitutes a critical part of the tumor microenvironment. It kills T cells and infers with functions of the NK cells to suppress the immune system and support metastasis [126, 127].

On the other hand T cells are set apart form the NK cells by the T cell marker CD3D [128] which is expressed in CD4+ and CD8+ T cells. GO-term analysis of CLL patients showed that this gene indicates biological functions like cell growth and proliferation, response to inflammation and immunological disease [129]. The same holds for the CD247 gene marker, which is also expressed in the NK cells. Another T cell marker is TRAC, which is a T cell receptor [130]. There are also genes characteristic for both NK and T cells, for example CCL5. CCL5 is a proinflammatory chemokine which is involved in activated T cells in the glucose uptake, covering the high demands of energy in T cells. It also regulates trafficking of, among other things, T cells and NK cells [131]. A further biomarker for NK- and T cells is GNLY which is cytolytic against tumors and microbes [132]. It also activates the expression in many cytocines, i.e. in CLL5 [133]. IL32 is expressed in NK and activated T cells, especially in T cells which undergo apoptosis [134]. The serine protease GZMB and CTSW are biomarkers for NK and CD8+ T cells [135, 136]. The natural killer cell granule

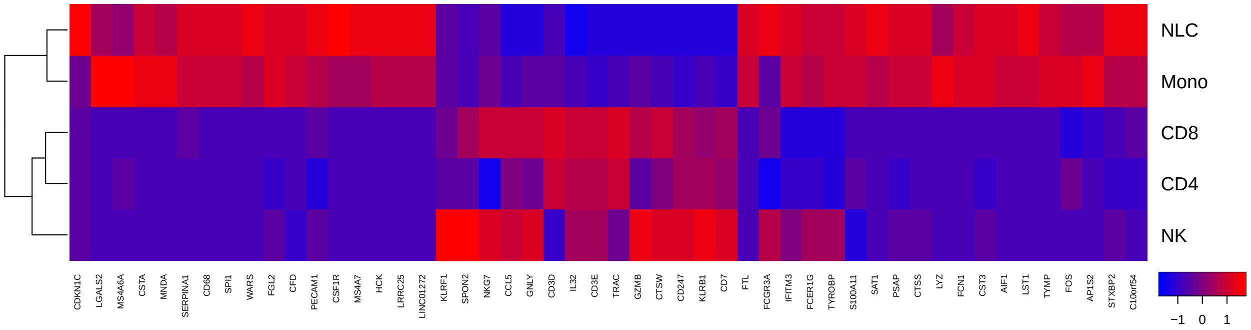


Figure 4.7: Heatmap of the reference profile X for the features with the top 50 weights ($\hat{g}_i \times \text{var}(X_{i,\cdot})$). Blue corresponds to low expression and red to high expression. The data were clustered by Euclidean distance. Reference profiles and training and test set were generated out from all four CLL patients.

protein 7 (NKG7) is also a marker gene for NK and CD8+ T cells [137].

The here developed algorithm completes the discussed gene set with an additional set of genes, which proved through calculations, to be among the 50 most relevant genes for cancerogenesis. Some of those supplemental genes may be further examined and may lead to biological insights. They potentially give hints on new biomarkers for the investigated cell types.

4.6 Expressed Genes in the Experimental Data are also Found to be Highly Expressed in the Loss-Function-Learning Model

In Figure 4.8 the highly expressed genes for the five labeled immune system cell types (CD4+ T cells, CD8+ T cells, CD14+ myeloid cells, CD56+ NK cells and nurse like cells) as well as the tumor cells (CD19+ CLL cells) are shown. The diagram lists the single cell datas for all patients at different points in elapsed time. It can be seen that the CD4+ T cells show no individual characteristic marker genes which would be characteristic for them alone. Their biomarkers are shared with the other cell types. The CD4 marker is also found in myeloid cells, CCR7 in CLL cells, CD3G, CD3D and IL32 in CD8+ T cells. Hence it is difficult to label the CD4+ T cells correctly. As a result the deconvolution results by loss-function learning get affected.

Loss-function learning was done for the five listed cell types and the tumor cells. The simulated bulk and reference profiles were totally separated, patient one and five were used for creation of training set and reference profile, patient six and eight were used for the test set. The results before and after loss-function learning for training and test set for simulated bulk sets can be seen in Figure 4.9. Here the deconvolution for CLL tumor cells, monocytes and NL cells gave high quality results.

The ones for CD8+ T cells and NK cells gave a slightly lesser quality. Again, only the results for the CD4+ T cells were compromised.

The heatmap in Figure 4.10 shows the 150 most variable genes. The highly expressed genes in Figure 4.8 also give major contribution to the high ranked genes after loss-function learning. For the CD4+ T cells CD3G, CD3D and IL32 were found. Regarding the genes of the CD8+ T cells all expressed genes of the experiment were reproduced by our data. The genes CD14, CST3 and CD68 were high expressed for myeloid cells and monocytes, respectively. For the NK cells we found FCGR3A, TRDC, GZMB, NKG7 and IL32 in both sets. The NLC shared the genes CST3, FCGR3A and CD68. Only for the CLL cancer cells no genes highly expressed in the experimental data in the high ranked genes of the loss-function learning model were found. Good deconvolution results for the CLL cells were achieved. The CLL cells give the main part in the mixtures. Therefore, as seen in section 3.4, a small number of biomarkers is sufficient for deconvolution.

[2] Christian Schmidl, mail: christian.schmidl@ukr.de

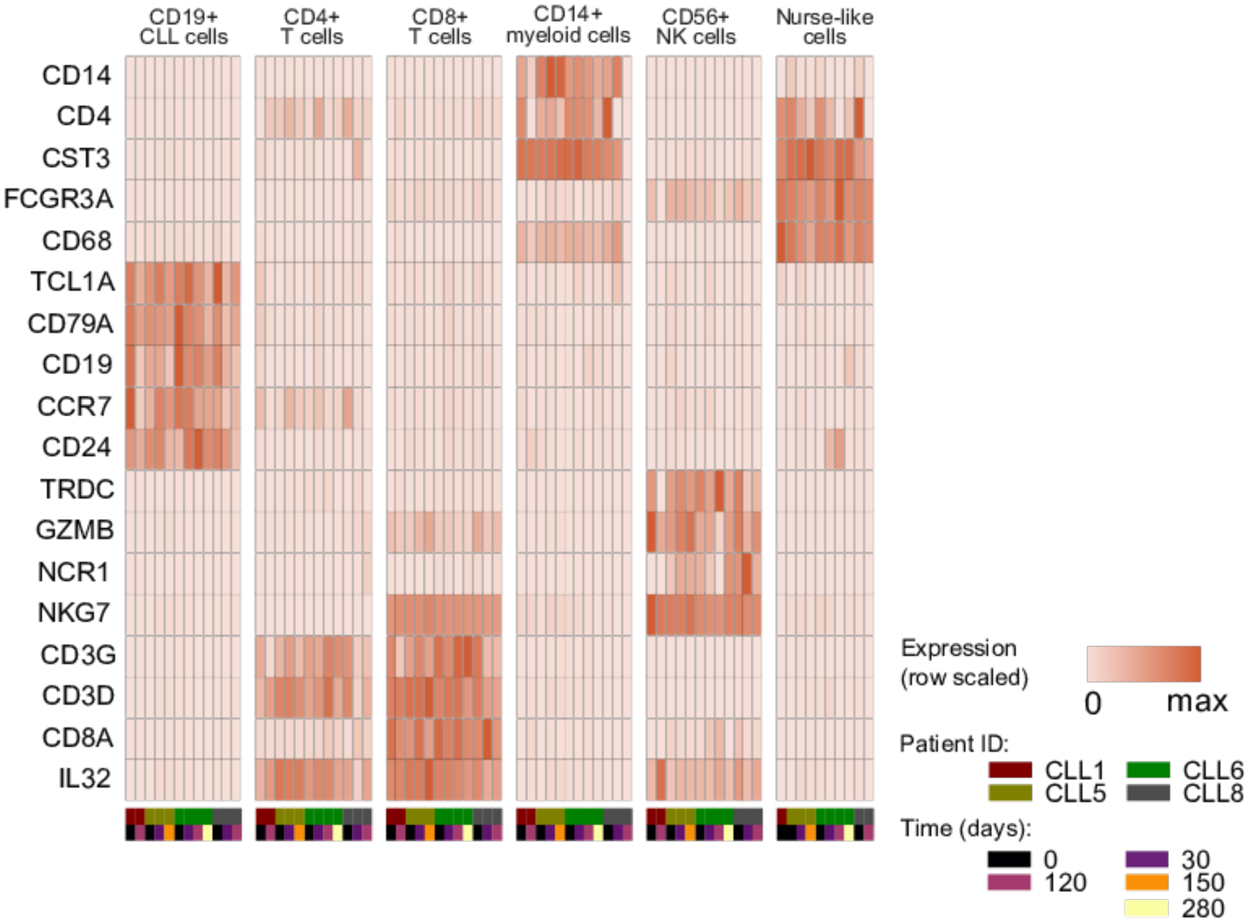


Figure 4.8: Highly expressed genes in the experimental single cell data listed for every cell type. The expression is shown separated by patient and time. Graphic provided by experimenter Christian Schmidl[2].

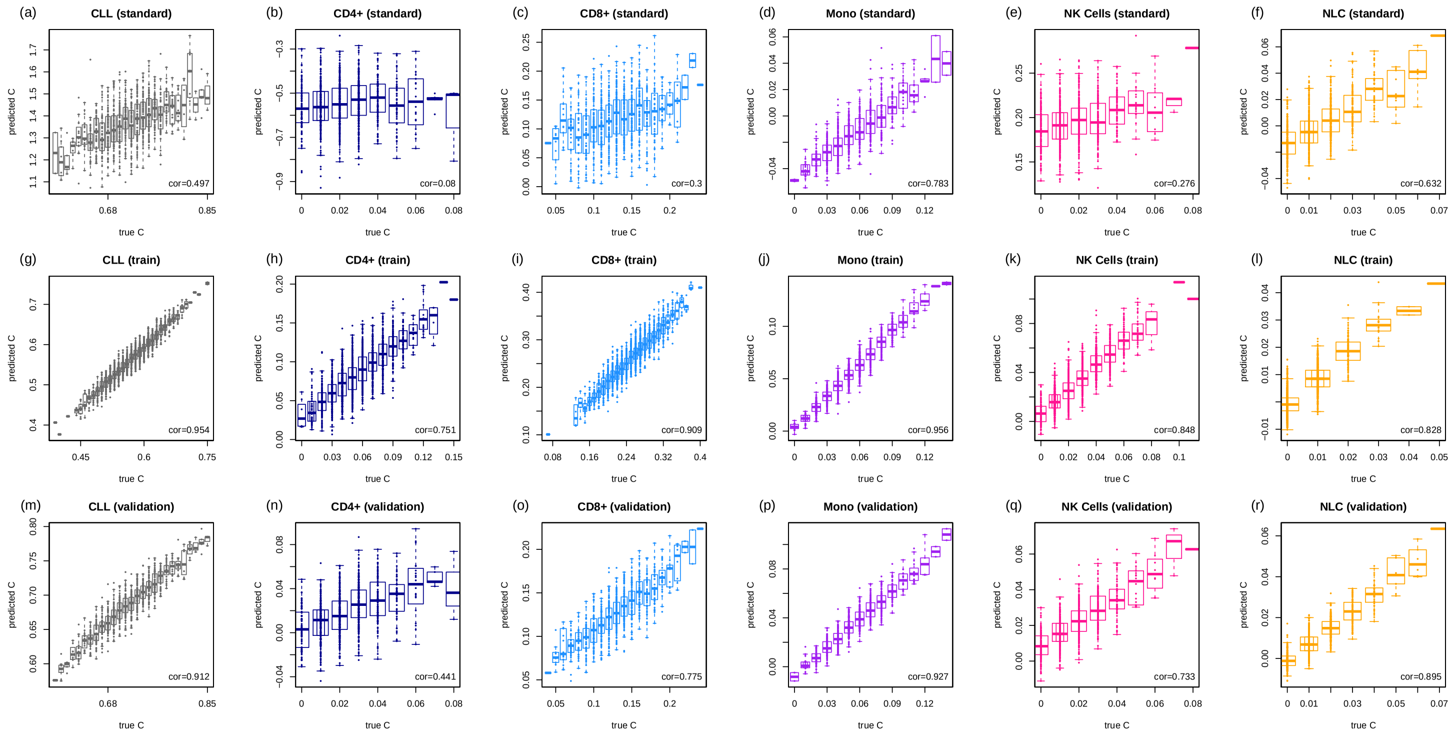


Figure 4.9: Loss-function learning results for training and validation set for the five evaluated immune cell subtypes and the CLL cells of the CLL data set. The training set and reference profiles are simulated from patients one and five data sets, validation set of patient six and eight. (a) to (e) show the correlation between real and predicted cellular distribution for the standard model ($g = (1, \ldots, 1)$) for the different cell types. (f) to (j) show the corresponding results. For the loss-function strong discrepancies between the performance on the training and validation data were observed. In the last row ((k) to (o)), loss-function learning results for the validation set are shown. For CD4+ T cells, after loss-function learning of training and validation set high discrepancies were found.

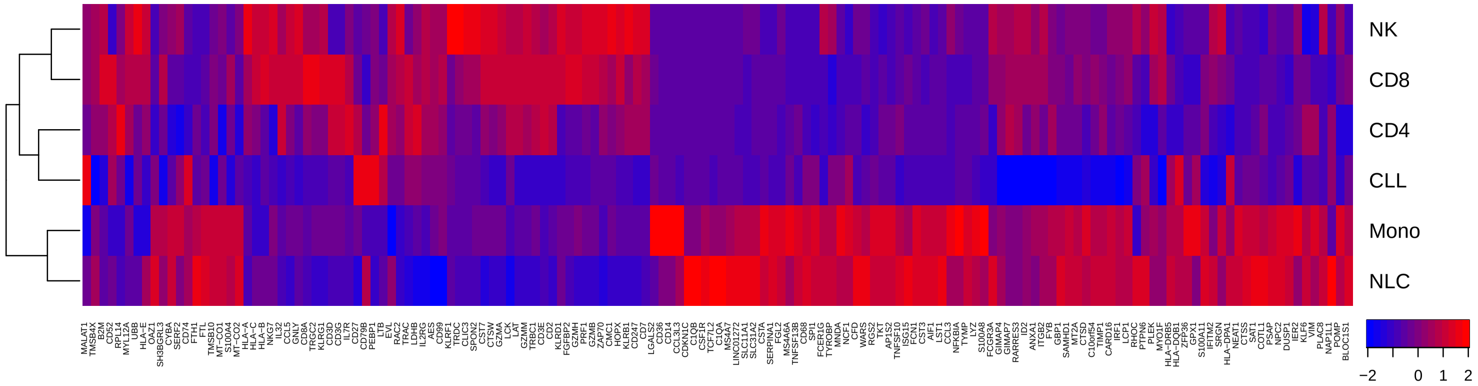


Figure 4.10: Heatmap of X, featuring the top 150 weights ($\hat{g}_i \times \text{var}(X_{i,\cdot})$). Blue corresponds to low expression and red to high gene expression. The data were clustered by Euclidean distance. The depicted heatmap corresponds to the simulation of Figure 4.9.

4.7 Application of the Loss-Function Learning Model on Bulk Sequencing Data

Eight of the twelve CLL specimens were measured by bulk RNA sequencing. These bulk data were augmented by additional four CLL bulk measurements [94]. For all bulk data, cellular compositions were estimated by FACS sorting. Here, it was tested whether the cellular compositions of these data can be predicted accurately using the models estimated from loss-function learning. Figure 4.8 gives an overview of the bulk data, which were measured at different time points, and the corresponding single cell data.

Firstly it was validated that FACS sorting provides accurate estimates of the cellular compositions. For this purpose, the eight samples with corresponding single-cell measurements were used. The composition of the cell types determined over both ways, by FACS and by single cell sequencing, were compared. The focus was put on the cell types which were considered in both, single-cell RNA sequencing and FACS, namely CD4+ and CD8+ T cells, monocytes and NK cells. The cellular proportions for both technologies were normalized to one. This ensures that results are comparable. Then, for each of the five concordant cell types cellular proportions from FACS versus the corresponding values from single-cell sequencing were plotted. The corresponding results together with respective correlations are shown in Figure 4.12. For all cell types which were present in both data sets, very high correlations, more than 94%, were achieved. The graphics visualize the slight deviation of the fitted data (solid line) from the ideal ratio indicated by the line through the origin (dashed line). Thus, the total percentages of cells in both sets were very similar and no shifts between the two measurements were observed. The last picture Figure 4.12f summarizes the results as a histogram. The mean correlation over all cell types was 97.7%. Thus, i that the distribution measured by FACS may be taken as cellular composition of the bulk profiles. Next, the model from loss-function learning estimated from single cell RNASeq measurements was applied to the bulk samples. All single cells of the four single cell samples were used without corresponding bulk sample (PT1 d120, PT6 d280, PT8 d30 and PT8 d120) as reference profiles and as training. For the creation of the reference profiles 20% of all cell types were sampled. Out of the remaining cells 100 cells were drawn by chance for every training set and a simulated bulk profile was calculated. In total 2,000 bulk profiles were simulated. The reference profiles and the mixtures in the training set were normalized to a fixed count number. As test set all twelve bulk profiles were used. For the cellular composition of the bulk FACS results were applied. For the calculation of the loss-function learning model the five cell types which were labeled in FACs and single cell measurements (CLL, CD4+ T cells, CD8+ T cells, monocytes and NK cells) were used. The results for the deconvolution of the bulk profiles are shown in Figure 4.13. It can be observed that the results from DTD correlate well with the estimates taken from FACs (from 0.729 for NK cells to 0.983 for CLL cells).

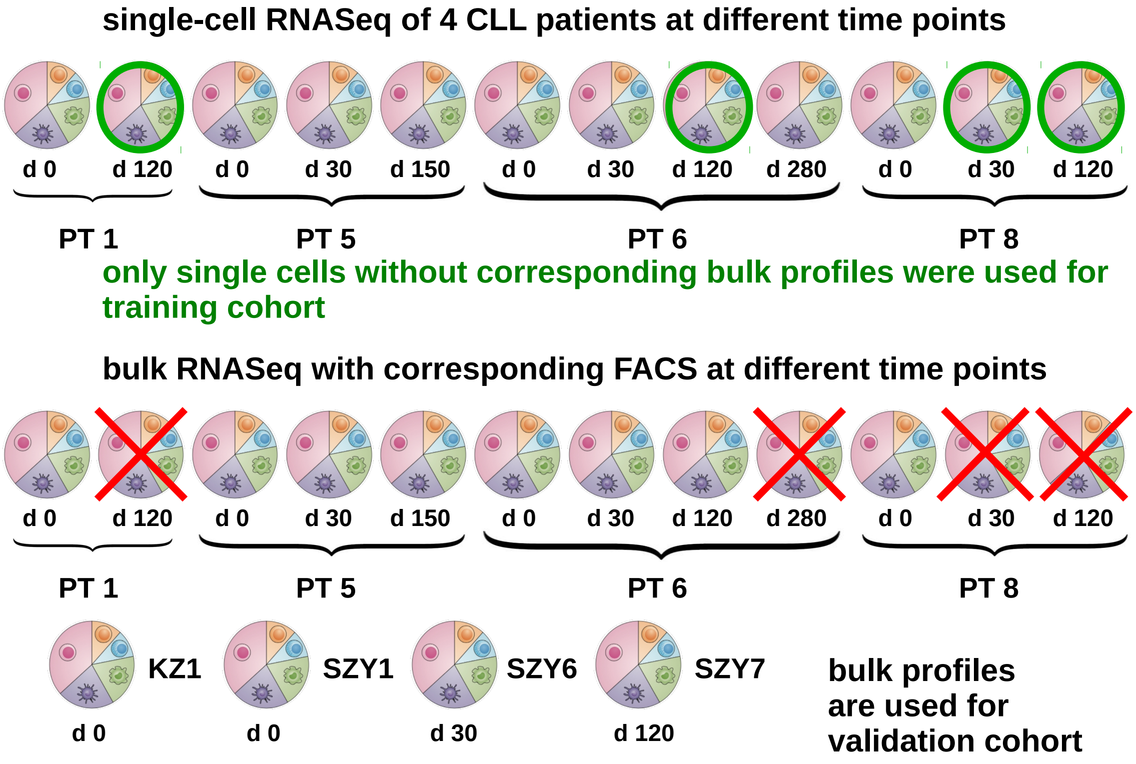

single-cell RNASeq of 4 CLL patients at different time points
d 0
d 120
d 0
d 30
d 150
d 0
d 30
d 120
d 280
d 0
d 30
d 120
PT 1
PT 5
PT 6
PT 8
only single cells without corresponding bulk profiles were used for training cohort
bulk RNASeq with corresponding FACS at different time points
d 0
d 120
d 0
d 30
d 150
d 0
d 30
d 120
d 280
d 0
d 30
d 120
PT 1
PT 5
PT 6
PT 8
KZ1
SZY1
SZY6
SZY7
d 0
d 0
d 30
d 120
bulk profiles are used for validation cohort

Figure 4.11: In the upper part of the graphic, the distribution of the patients at the different time points of measurement is visualized. In the lower one the corresponding FACs profiles are shown. Eight of the FACS measurements showed consistency with the single cell measurements. Four of the single cell measurements did not show accordance.

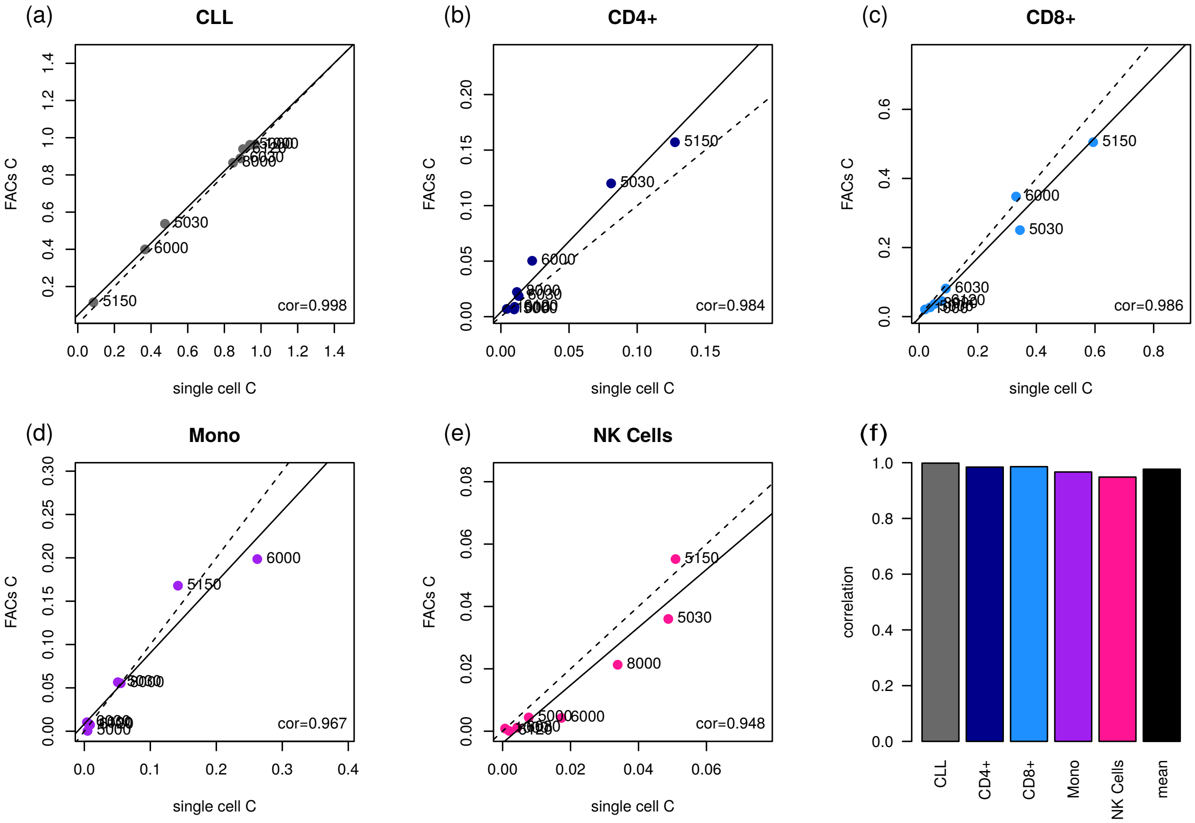


Figure 4.12: Scatter plot for comparison of cell-type content in bulk and single-cell measurements. The x-axis gives the cell type contents from single-cell sequencing measurements, the y-axis the composition of the bulk determined by FACs. The continuous line marks the best fit, the dashed line indicates an equal ratio with slope 1, running through the origin. The point labels are composed by the patients number in the first digit and the measuring time point in the following three digits. E. g., number 5030 means patient number five measured 30 days after treatment.

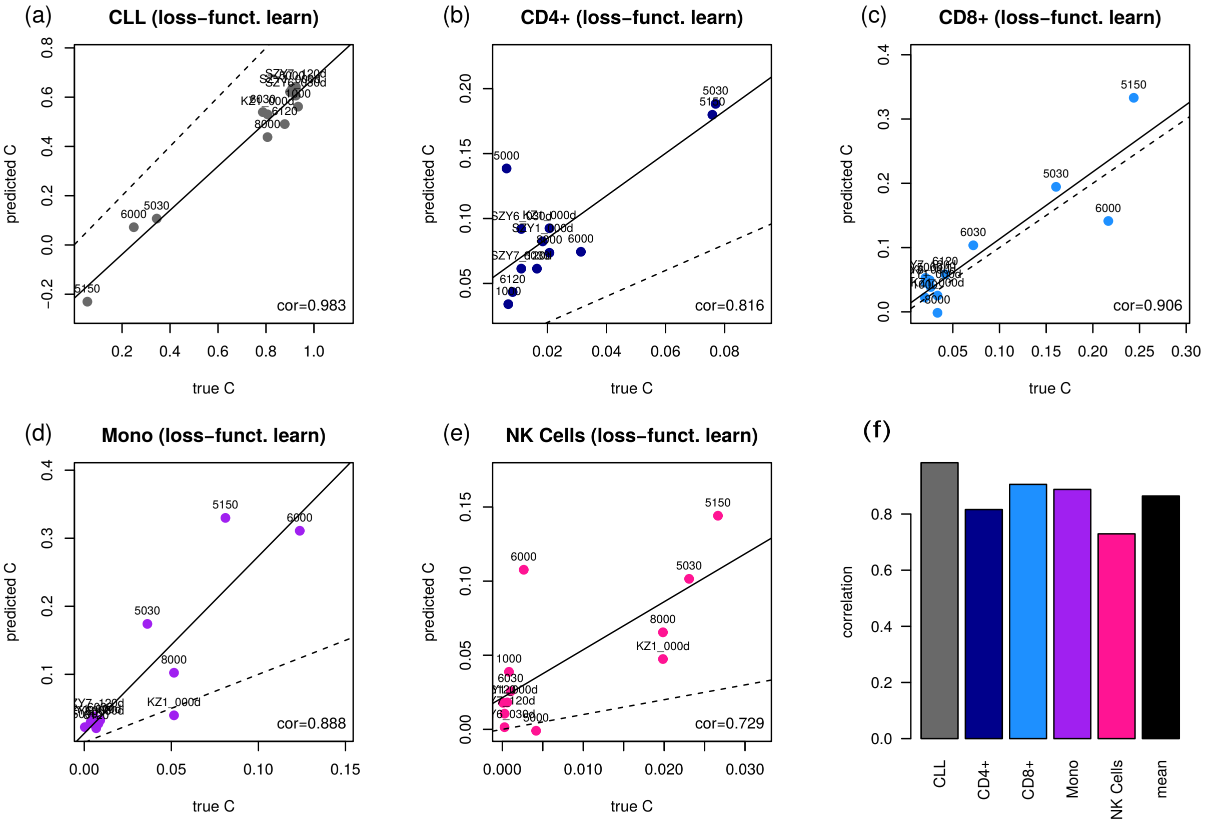


Figure 4.13: Results of bulk deconvolution for a model learned from training sets consisting of single cells without corresponding bulk profiles. The CLL-tumor cells were treated as a cell type. The measurement points were labeled like in figure 4.12. For eight patients both, FACs and single cell measurements were available. The labeling of the other four patients corresponds to the notation given by the experimenter and can be seen in Figure 4.11. a) to e) are the results of the loss-function learned model, f) summarizes the results in a barplot.

4.8 Comparison with CIBERSORT

Like in chapter 3.9, here the results of the CLL data set get compared with CIBERSORT calculations. Cell types of CIBEROSRT got assigned to the immune cell types of the CLL dataset as follows:

- CLL: B cells naive and memory.
- CD4+ T cells: T cells CD4 naive, memory resting and memory activated, T cells regulatory (Tregs), T cells follicular helper.
- CD8+ T cells: T cells CD8.
- Monocytes: Monocytes, macrophages M0, M1 and M2, dendritic cells resting and activated.
- NK cells: NK cells resting and activated.

For comparison of both deconvolution methods initially the genes of the CIBERSORT reference profile were used for comparison and later the 1,000 most variable genes as done previously. The reference profile of CIBERSORT consist of 547 genes. Out of these, 489 were covered by the generated single cell profiles. Those were the genes utilized for loss-function learning. As in section 4.7, only the single cells without corresponding bulk profiles were used for loss-function learning. 20% of the single cells of every cell type were taken for creating reference profiles, the remaining fraction was utilized to simulate a training set of 2,000 mixtures. The results are shown in Figure 4.14.

The standard model with it's reference profiles and without loss-function learning is shown in red. There it can be observed that this model performs remarkably well, except for CD8+ cells. When CIBERSORT reference profiles were applied, the B cell results were compared with the tumor cell content from CLL, since CLL is a malignancy of mutated B cells. For the standard model (red) an overall correlation of 0.743 was achieved. The CIBEROSRT algorithm (green) in conjunction with with the reference profiles generated by the developed algorithm shows compromised results, compared to the standard model. Loss-function learning improved the results of the standard model (blue). CIBERSORT with its own reference profiles gives better results than the standard model (yellow). However, it did not outperform the model from loss-function learning. One has to keep in mind that only a few bulk samples were available for the development of the here presented new approach. Thus, the correlation is expected to have a large uncertainty.

Due to the small validation set of only twelve bulk samples, the results for the test set from chapter 4.4, where the single cells were separated by chance in reference profiles, training and validation set, got incorporated. Now, however, no strict separation of patients in the sampling of the test and training set took place. The predefined gene set of CIBERSORT was used for the calculations and all the five immune cell types existing in the single cell data set got considered. The results can be seen in Figure 4.15. NL cells can not be deconvoluted by CIBERSORT with their own reference profiles since they are not part of the CIBERSORT reference profiles. Thus,

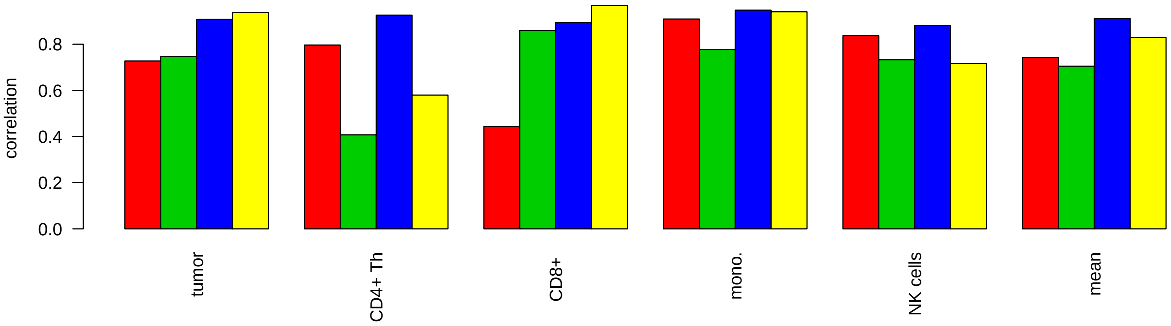


Figure 4.14: Deconvolution results for the bulk profiles evaluated with different devonvolution methods. Comparison of loss-function learning model with CIBERSORT, calculated for the genes of CIBERSORT. Red: Reference profiles from CLL data for the CIBERSORT genes. Standard model. Green: Reference profiles CLL, deconvoluted in CIBERSORT. Blue: Reference profiles CLL, loss-function learned. Yellow: CIBERSORT with CIBERSORT reference profiles.

the figure shows no bar for the NL cells. In this case they are also not involved in the calculation of the mean correlation. Again the loss-function learning model led to the best results (blue), now followed by CIBERSORT combined with the loss-function learning reference profile (green), followed by the naive model (red). Here again the CD4+ T cells gave the least improved results and were not predictable in the simulated mixtures with the standard model. CIBERSORT with their own reference profiles showed the worst performance (yellow).

In the case of simulated bulk mixtures the results for all deconvolution methods performed not as good as for the bulk profiles in Figure 4.14. Due to the reason that there are only twelve bulk measurements, each single measurement gives a much higher contribution to the correlation than a single one out of the 1,000 simulated bulk profiles. Consequently, estimates for the correlation are expected to be much more stable than in the previous study on the bulk profiles.

Deconvolution results of the predefined gene set chosen by known biomarkers of CIBERSORT were compared with our gene set chosen computationally by the highest variance in the reference profile. First the deconvolution results for the simulated bulk set using the 1.000 most variable genes were investigated. Only in the CIBEROSRT comparison, where the reference profiles from CIBEROSRT were used, we used the smaller gene set predefined by CIBERSORT. The results can be seen in Figure 4.16.

Next deconvolution results of both gene sets, the 1.000 most variable genes optimized by loss-function learning problem and the predefined gene set of CIBERSORT got compared. The results are shown in Figure 4.17.

When solving the deconvolution problem with loss-function learning better results were obtained for the genes chosen by variability (dark blue) than for the predefined gene selection of CIBERSORT

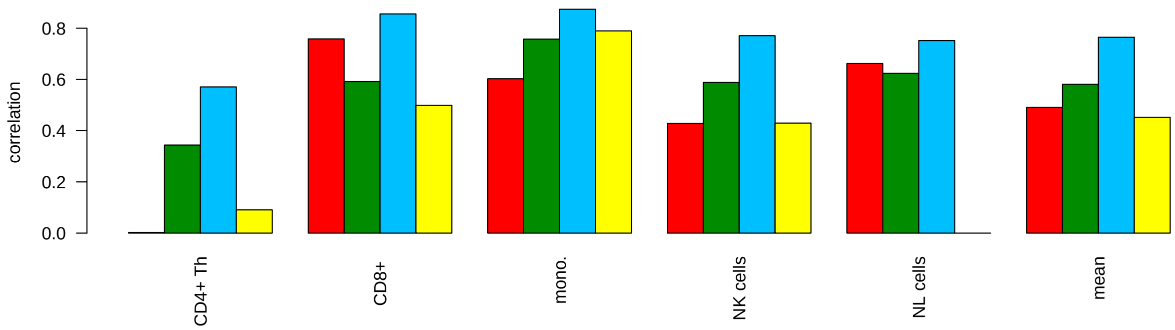


Figure 4.15: Deconvolution results for generated bulk profiles of the single cell measurements were genes from the CIBERSORT reference profiles were used. Comparison of loss-function learning model with CIBERSORT. Red: Standard model. Green: CIBERSORT combined with the reference profiles generated by the developed algorithm. Blue: Loss-function learned model. Yellow: CIBERSORT with CIBERSORT reference profiles.

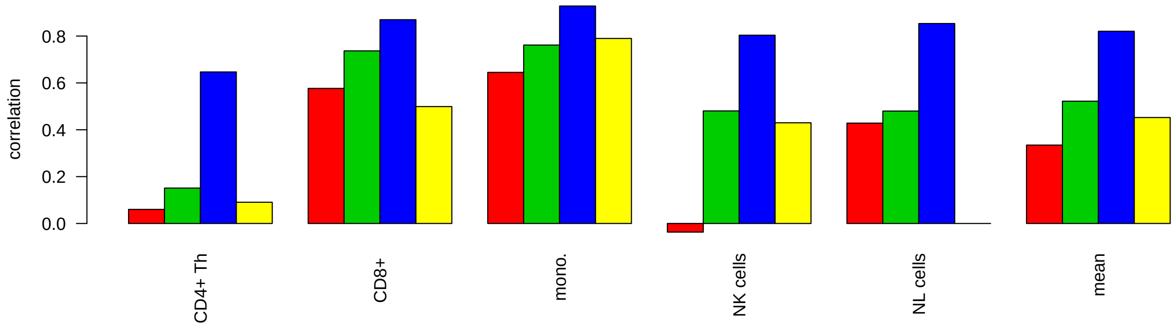


Figure 4.16: Comparison of loss-function learning model with CIBERSORT. Deconvolution results for generated bulk profiles from the single cell measurements. The 1,000 most variable genes were used in the reference profile. For CIBERSORT deconvolution with the CIBERSORT reference profile the smaller CIBERSORT data set was used. Red: Standard model. Green: CIBERSORT combined with the reference profiles generated by the developed algorithm. Blue: Loss-function learned model. Yellow: CIBERSORT with CIBERSORT reference profile.

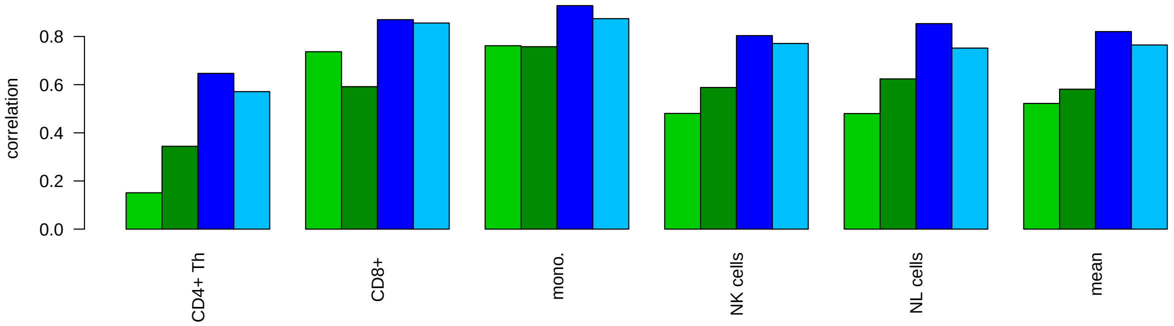


Figure 4.17: Deconvolution for generated bulk profiles of the single cell measurements. Comparison of loss-function learning model with CIBERSORT calculated for the 1.000 most variable genes and the CIBERSORT gene choice. From left to right: Green: CIBERSORT with our reference profile for the 1.000 most variable genes. Dark green: CIBERSORT with our reference profile for the CIBERSORT genes. Blue: Loss-function learned model for 1.000 most variable genes. Light blue: Loss-function learned model for CIBERSORT genes.

(light blue). This ranking was also observed for each cell type individually. For the CIBERSORT algorithm the opposite relation was found: the results were better for the gene selection given by CIBERSORT (dark green) than for the choice taken by the loss-function learning algorithm(light green). However, in both cases the loss-function learning method outperformed the linear regression-based deconvolution algorithm of CIBERSORT. In the here newly developed approach the method of gene selection is adaptive to the cell types of interest therefore better performance could be achieved. Please note, that NL cells could not be deconvoluted by CIBEROSRT since they were not considered initially for the gene selection.

4.9 Deconvolution of Bulk Profiles with Deconvolution Models Learned of Foreign Data

There are still few data sets available, that are comprised of a high amount of labeled single cell RNAseq measurements of one cancer type. Thus, it was of special interest to find out whether loss-function learning models calculated with one data set are transferable to a second data set of a different cancer type. As two data sets of single cell RNAseq measurements were available, the melanoma one discussed in chapter 3 consisting of 19 melanoma tumors and the CLL set which is discussed now, the loss-function learning model got calculated with one data set and the other one served as test set for deconvolution.

In order to prove transferability for both data sets reference profiles and a set of 2,000 simulated bulk profiles were created. For the reference profiles 20% of every cell type were taken and for the bulk profiles out of the remaining stock 100 single cells got drawn by chance. Reference and bulk profiles were normalized to a fixed count number as previously. The melanoma data set consisted of seven cell types (B cells, macrophages, endothelial cells, CAFs, NK cells, CD4+ and CD8+ T cells) where as the CLL data comprised six different types (CLL cells, CD4+ and CD8+ T cells, monocytes, NK and NL cells). For both data sets deconvolution models got evaluated using loss-function learning. Using those models the other respective data set for the four cell types which were contained in both sets (CD4+ and CD8+ T cells, NK cells, as well as monocytes/macrophages) got deconvoluted. Macrophages (melanoma data set) were matched with monocytes (CLL data set) as most monocytes differentiate to macrophages.

Two different deconvolution approaches were taken. First, the reference profile matrix got limited to the cell types contained in both data sets. Second, all available cell types were deconvoluted and subsequently only the cell types contained in both data sets got compared. For both approaches, the standard model $g = (1, \ldots, 1)$ and the loss-function learning model got tested. Figure 4.18 corresponds to the first scenario, where the references profiles got restricted to the available cell types. Figure 4.19 gives results for where the models were trained on the full references profiles and where they were constrained subsequently for comparison.

In both scenarios and for both cancer types the deconvolution with loss-function learning led to better results than the standard model. The standard models led to similar results for both approaches (preselected celltypes vs. all available cell types). In the case of preselected cell types the results improved to 0.607 for the melanoma bulk profiles and to 0.667 for the CLL bulk profiles with loss-function learning. For the deconvolution using all available cell types, values of 0.597 for melanoma bulk and of 0.693 for CLL bulk got obtained. Similar outcome can be observed for both strategies. The two T cell subtypes, CD4+ Th and CD8+, were hard to deconvolute and the deconvolution was compromised by the use of external reference profiles.

The HPC model got applied for deconvoluting the CLL bulk profiles. Before evaluation the reference matrix got restricted to the cell types in both data sets (CD4+ and CD8+ T cells, macrophages and NK cells). The macrophages result for the cellular composition was compared to the corresponding outcome for the monocytes. A summary is shown in Figure 4.20. From the plots a) - d) one can take the results for the standard model with $g = (1, \ldots, 1)$, in e) - h) the results for deconvolution with the HPC loss-function learning model is shown. For the monocytes/macrophages and NK cells solid agreement is testified. Also the CD4+ T cells improved a lot compared to the standard model. For CD8+ T cells however a compromised performance was obtained.

Furthermore the CLL bulk profiles for all eight cell types in the reference profile got evaluated. After deconvolution the cell types of both data sets got compared. For the results of the CD4+ T cells the results of the CD4+ Tregs and CD4+ T-helper cells were summed up and compared with the CD4+ references of the FACS profiles, as the CD4+ T cells get separated in the HPC model. Figure 4.21 shows the deconvolution results for the standard model (scatter plot a) - d)) and HPC loss-function learning model (scatter plot e) - h)). The deconvolution results for the HPC model now turned out to be significantly better than for the standard model, as seen in Figure 4.20, despite the lack of a sufficient large number of CD8+ T cells in the data set. On average the performance

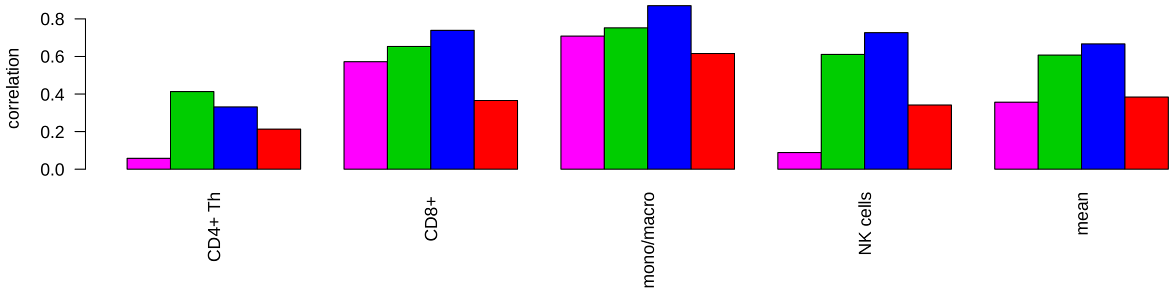


Figure 4.18: Results with deconvolution only for CD4+, CD8+, monocytes/macrophages and NK cells, which were contained in both data sets. The mean correlation of the values for the four considered cell types is labeled as mean in the histogram. From left to right: Magenta: Standard model with CLL reference profile evaluated on simulated Tirosh bulk data. Green: Loss-function learned model on CLL data, evaluated on Tirosh data set. Blue: Loss-function learned model on Tirosh data, evaluated on CLL data set. Red: Standard model with Tirosh reference profile evaluated on CLL data set.

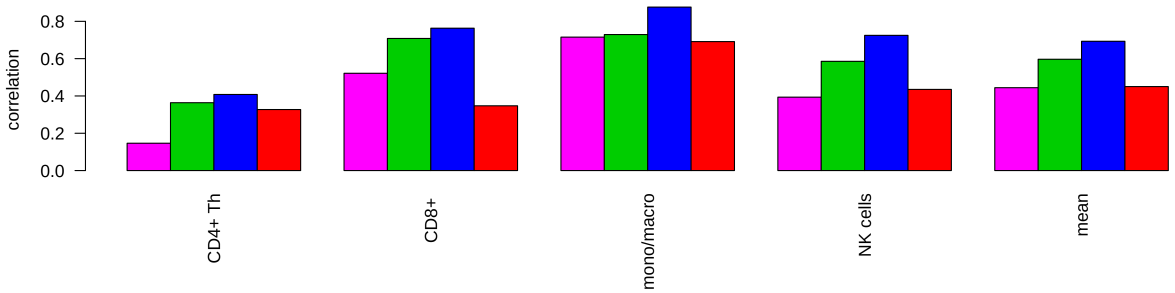


Figure 4.19: Cell-type reduction to the cell types contained in both data sets after deconvolution. The mean correlation of the values for the four considered cell types is labeled as mean in the histogram. From left to right: Magenta: Standard model with CLL reference profile evaluated on simulated Tirosh bulk data. Green: Loss-function learned model on CLL data, evaluated on Tirosh data set. Blue: Loss-function learned model on Tirosh data, evaluated on CLL data set. Red: Standard model with Tirosh reference profile evaluated on CLL data set.

got increased to 65.6% for the overall correlation compared to 14,3% without loss function learning. Better deconvolution results were also achieved for the two T cell subtypes, i.e. 89% and 91%, compared to 48% and 37%, obtained by the model were the reference matrix was restricted in the beginning. The NK cells lost slightly in performance compared to the standard model. The overall performance in the actual loss-function learning model mounted to 84.6% compared to 62.9% attained by the loss-function learning model with restricted reference matrix in the beginning.

In the second deconvolution case the results after deconvolution got restrained to the shared cell types of both data sets. By this procedure better results for the two T cell subtypes could be realized. CLL is caused by degenerated B cells, which belong, like the T cells, to the lymphocytes. When restricting the reference matrix before devconvolution, some of the CLL cancer cells seem to get mislabeled as T cells as their reference profiles are too similar. From Figure 4.20 e) and f) it can be concluded that, especially bulk profiles with a lower number of CD4+ and CD8+ T cells, get overestimated. In these bulks the fraction of cancer CLL cells, and therefore degenerated B cells, is very high. It seems that there a lot of CLL cells got assigned to the T cell portion. In the case of the full reference profile these mislabeled CLL cells may be classified as B cells and thus are not attributed to the T cells. This artifact can be seen in Figure 4.21 e) and f), where for smaller T cell contents much lower values get achieved compared to the restricted matrix case depicted in Figure 4.20.

Therefore, it seems reasonable to apply the loss-function learning models of one data set to a different cancer type, when including all reference profiles of the cell types which were used in model calculation. Additionally it seems necessary to employ reference profiles or profiles for a very similar cell type (B cells for CLL cells e.g.) for all cell types which compromise a high fraction of the bulk mixtures. This holds especially true when dealing with different cell types belonging to the same subgroup, e.g. the here presented T cells.

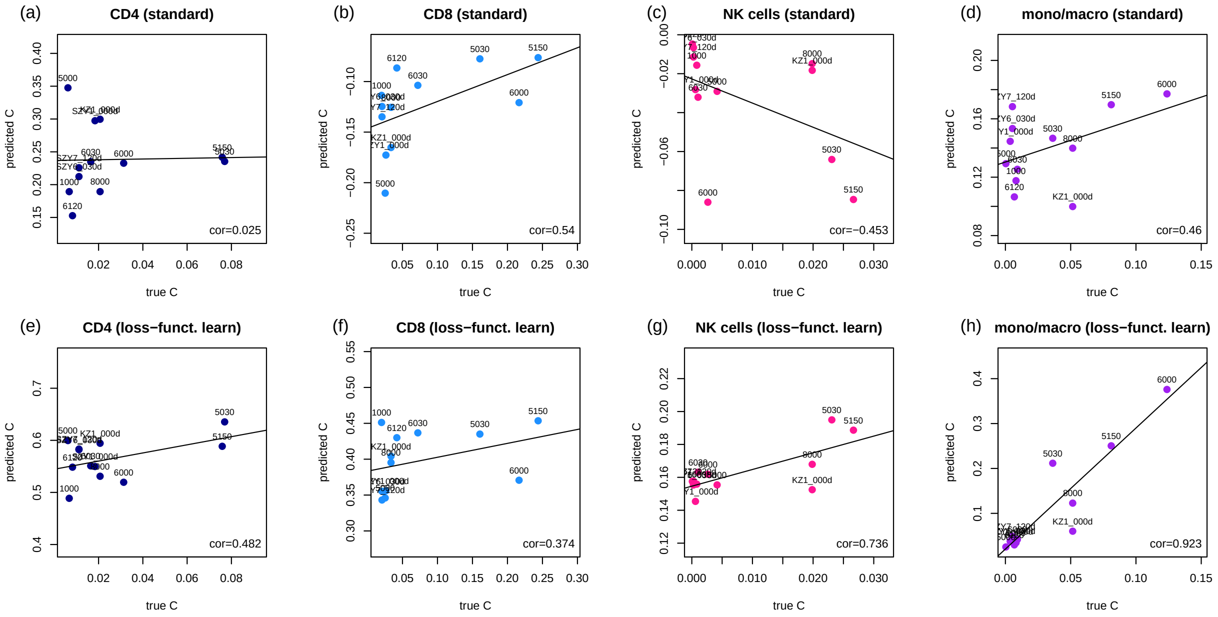


Figure 4.20: CLL RNAseq bulk measurements deconvoluted with HPC model. The reference matrix was restricted to cell types shared of both data sets before deconvolution. The standard model is shown in pictures a) - d), the loss-function learning model in pictures e) - h).

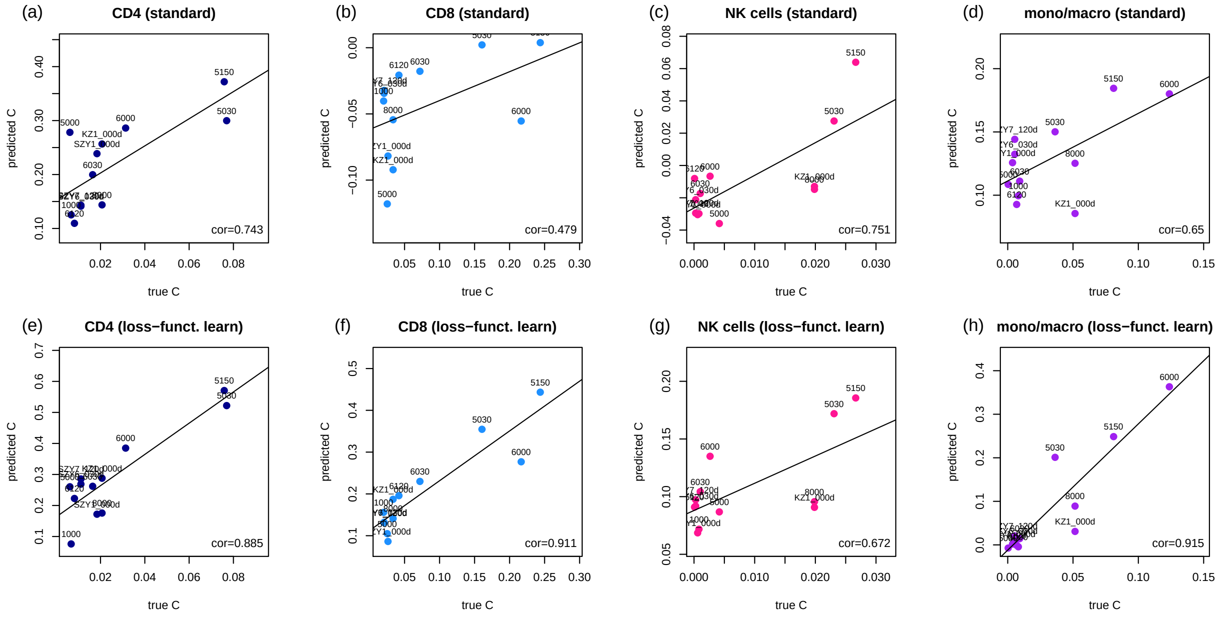


Figure 4.21: CLL RNAseq bulk measurements deconvoluted with HPC model. All eight cell types were deconvoluted in the reference matrix and the results for the cell types available in both data sets were compared. The standard model is shown in pictures a) - d), the loss-function learning model in pictures e) - h).

Chapter 5

Discussion

Application of training data for loss-function learning for digital tissue deconvolution is suggested to adapt the deconvolution algorithm to the requirements of specific application domains. The concept is similar to an embedded feature-selection approach in regression or classification problems. In both contexts feature selection is directly linked to a prediction algorithm and not treated as an independent preprocessing step.

The main limitation of the here presented method is the availability of training data. Other methods do not use, and cannot use, training data. In fact, the strength of loss-function learning results primarily from the additional information in training data with known cellular compositions. Such data is not always available, but with current improvements in FACS and single-cell sequencing technology, it is becoming increasingly available.

A specific instance of loss-function learning using squared residuals for $\mathcal{L}_g$ got introduced and evaluated by real patient data sets. The concept is not limited to this specific type of inner loss function and can also be used in combination with other loss functions such as those from penalized least-squares regression [138], l_1 regression, or support vector regression [23]. However, the least-squares loss function allows for stating the outer optimization problem in a closed analytical form, reducing computational burden.

The outer loss function L evaluates the fit of estimated and true cellular proportions in the training samples. The correlation of estimated versus true quantities across samples got chosen instead of non-absolute measure of deviation such as $||c-\hat{c}||_2^2$, which does not fulfill symmetry (2.4). Moreover, it was not required that the estimated proportions $\hat{C}_{\cdot,k}$ for tissue k to sum up to one. Consequently, the estimated cellular composition for a given cell type is comparable between tissues, but the estimated cellular composition across cell types is not. When testing the here presented method it was not looked at absolute deviations of true versus estimated cell proportions, but only at their correlation. It was intended to take account of how many cells of a specific type (e.g., T cells) are in a tissue (Figure 3.13), nor whether they constituted 10% or 20% of the cells in this tissue. However, if two tissues were considered and it got estimated that there were more cells of that type in the first tissue compared to the second, this relation was also found in the true cell populations.

In summary, loss-function learning got introduced a new machine-learning approach to the digital tissue deconvolution problem. It allows for adaption to application-specific requirements, such as focusing on small cell populations or delineation of similar cell types. In simulations and in an application to melanoma tissues the use of training data allowed to quantify large cell fractions as accurately as existing methods, and significantly improved the detection of small cell populations as well as the distinction of similar cell types.

Chapter 6

Summary and Outlook

A novel method for digital tissue deconvolution got introduced. In contrast to other deconvolution methods no predefined gene set for determining the immune cell content in bulk measurements was used. An algorithm that adapts the gene set for deconvolution to the specific biological problem is proposed. For this purpose, predefined training mixtures generated from single-cell RNA sequencing measurements got employed. Since the newly developed algorithm learns the best loss function for deconvolution, this new method is called ”loss-function learning”.

It got proved that the used loss-function learning function is not convex. Further, it was shown that the deconvolution results for different starting points do only weakly depend on the initialization of the algorithm.

It got verified that the developed algorithm can accurately estimate immune cell compositions in two different single cell RNASeq data sets, one corresponding to melanoma metastases the other to chronic lymphozytic leukemia (CLL). Further it was demonstrated that the estimation of small cell proportions and on distinguishing similar cell types is improved compared to the current technique. Moreover, it was proved that missing reference profiles in the deconvolution can be compensated.

For both data sets the new algorithm outperformed the state of the art algorithm CIBERSORT. Particularly, because the new method is unbiased with respect to the selection of gene. Thus, it potentially facilitates the detection of new cell markers.

For the CLL data set a corresponding bulk was available that was used for final validation of the method. Here the tumor content reliably could be predicted, although the performance was slightly compromised for NK cells, which made out only a small proportion of the immune cell content.

The newly developed method is limited mostly by the necessity of training data. However, it was feasible to transfer the loss-function learning models from one dataset to another, apon which the results were slightly compromised but substantially better than achieved by a standard deconvolution approach.

Outlook

Loss-function learning is a machine learning method and can be prone to overfitting. For this purpose, it is planned to use penalty terms to regulate optimization problems. Possible strategies are l_1 and l_2 penalties, corresponding to LASSO and ridge regression. This work is partly done and will further be done by Marian Schn [1]. Further, for instance user friendly software will be provided and computation speed may be improved by using Python or C/C++ for instance. Further advances may be achieved by the application of other functions for loss-function learning, such as a least squares distance between predicted and true cellular compositions and non-negativity constraints on the estimated compositions.

[1] Mail: Marian.Schoen@klinik.uni-r.de

Appendix A

Appendix: Auxiliary Calculationis used for Calculating Gradient and Hessian

The following definitions for mean, variance, covariance and correlation are used:

$$\begin{aligned}
\overline{a} &= \frac{1}{n} = \sum_{i=1}^{n} a_i, && \text{(A.1)}\\
\operatorname{var}(a) &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})^2, && \text{(A.2)}\\
\operatorname{cov}(a,b) &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b}), \operatorname{cor}(a,b) &&= \frac{\operatorname{cov}(a,b)}{\sigma_a \sigma_b}, \qquad \text{(A.3)}
\end{aligned}$$

where σ_a is the standard deviation of vector a.

One gets

$$\begin{aligned}
&\frac{\partial}{\partial \hat{C}_{jk}} \hat{\sigma}_j = \\
&\frac{1}{2}\left(\frac{1}{n}\sum_{p=1}^{n}(\hat{C}_{jp} - \hat{\overline{C}}_{j,.})^2\right)^{-\frac{1}{2}} \cdot \frac{1}{n}\left[2\left(\hat{C}_{jk} - \frac{1}{n}\sum_{q=1}^{n}\hat{C}_{jq}\right)\left(1 - \frac{1}{n}\right) + 2\sum_{\substack{p=1\\ p\neq k}}^{n}\left(\hat{C}_{jp} - \frac{1}{n}\sum_{q=1}^{n}\hat{C}_{jq}\right)\left(-\frac{1}{n}\right)\right] \\
=&\frac{1}{n\hat{\sigma}_j}\left(\hat{C}_{jk} - \hat{\mu}_j\right). \qquad \text{(A.4)}
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \hat{C}_{jk}} \mathrm{cov}(C_{j,\cdot}, \hat{C}_{j,\cdot}) =& \frac{\partial}{\partial \hat{C}_{jk}} \frac{1}{n} \sum_{p=1}^{n} \Big[(C_{jk} - \overline{C}_{j,\cdot}) \hat{C}_{jp} - (C_{jp} - \overline{C}_{j,\cdot}) \hat{\overline{C}}_{j,\cdot} \Big] \\ =& \frac{1}{n}(C_{j,k} - \hat{C}_{j,\cdot}) \underbrace{- \frac{1}{n} \sum_{p=1}^{n} C_{j,p} \cdot \frac{1}{n} + \frac{1}{n} \sum_{p=1}^{n} \overline{C}_{j,\cdot} \cdot \frac{1}{n}}_{-\frac{1}{n}\overline{C}_{j,\cdot} + \frac{1}{n}\overline{C}_{j,\cdot}} \\ =& \frac{1}{n}(C_{jk} - \mu_j) \qquad \text{(A.5)}\end{aligned}$$

The gradient of the loss function $L(g)$ (equation (2.3)) is calculated by

$$\begin{aligned}\frac{\partial L(g)}{\partial g_i} &= -\sum_{j=1}^{q} \sum_{k=1}^{n} \frac{\partial \left(\mathrm{cor}(C_{j,\cdot}, \hat{C}_{j,\cdot}) \right)}{\partial \hat{C}_{jk}} \frac{\partial \hat{C}_{jk}(g)}{\partial g_i} \\ &= -\sum_{j=1}^{q} \sum_{k=1}^{n} \frac{1}{\sigma_j \hat{\sigma}_j} \left(\frac{1}{n}(C_{jk} - \mu_j) - \frac{\mathrm{cov}(C_j, \hat{C}_j)}{n \hat{\sigma}_j^2} (\hat{C}_{jk} - \hat{\mu}_j) \right) \frac{\partial \hat{C}_{jk}(g)}{\partial g_i} \\ &= \sum_{j=1}^{q} \sum_{k=1}^{n} \frac{1}{\sigma_j \hat{\sigma}_j} \left(\frac{\mathrm{cov}(C_{j,\cdot}, \hat{C}_{j,\cdot})}{n \hat{\sigma}_j^2} \big(\hat{C}_{jk} - \hat{\mu}_j \big) - \frac{1}{n}(C_{jk} - \mu_j) \right) \frac{\partial \hat{C}_{jk}(g)}{\partial g_i} \, . \qquad \text{(A.6)}\end{aligned}$$

where

$$\begin{aligned}\frac{\partial \hat{C}_{jk}}{\partial g_i} &= \left(\frac{\partial}{\partial g_i} \left((X^T \Gamma X)^{-1} X^t \Gamma Y \right) \right)_{jk} \\ &= \left(-(X^T \Gamma X)^{-1} \left(\frac{\partial}{\partial g_i} X^T \Gamma X \right) (X^T \Gamma X)^{-1} X^T \Gamma Y + (X^T \Gamma X)^{-1} X^T \left(\frac{\partial}{\partial g_i} \Gamma \right) Y \right)_{jk} \\ &= \left(-(X^T \Gamma X)^{-1} \left(X^T \delta(i) X \right) (X^T \Gamma X)^{-1} X^T \Gamma Y + (X^T \Gamma X)^{-1} X^T \delta(i) Y \right)_{jk} \, .\end{aligned}$$

The Hessian of the outer loss-function $L(g)$ is given by

$$
\begin{aligned}
H_{li}(g) =& \frac{\partial}{\partial g_l}\left(\frac{\partial L}{\partial g_i}\right) = \sum_{j=1}^{q}\sum_{k=1}^{n}\frac{\partial}{\partial g_l}\left(\frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,.}))}{\partial \hat{C}_{jk}}\cdot\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right) \overset{\text{product rule}}{=} \\
=& \sum_{j=1}^{q}\sum_{k=1}^{n}\left\{\left[\frac{\partial}{\partial g_l}\left(\frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,.}))}{\partial \hat{C}_{jk}}\right)\right]\cdot\frac{\partial \hat{C}_{jk}(g)}{\partial g_i} + \frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,.}))}{\partial \hat{C}_{jk}}\cdot\frac{\partial}{\partial g_l}\left(\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right)\right\} \\
=& \sum_{j=1}^{q}\sum_{k=1}^{n}\left\{\underbrace{\left[\frac{\partial}{\partial \hat{C}_{jk}}\left(\frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,.}))}{\partial \hat{C}_{jk}}\right)\right]}_{\textcircled{A}}\frac{\partial \hat{C}_{jk}(g)}{\partial g_l}\cdot\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right. \\
&\left. + \frac{\partial(-\mathrm{cor}(C_{j,\cdot},\hat{C}_{j,.}))}{\partial \hat{C}_{jk}}\cdot\underbrace{\frac{\partial}{\partial g_l}\left(\frac{\partial \hat{C}_{jk}(g)}{\partial g_i}\right)}_{\textcircled{B}}\right\}, \qquad \text{(A.7)}
\end{aligned}
$$

with

$$
\begin{aligned}
\textcircled{A} =& \frac{1}{n\sigma_j\hat{\sigma}_j^3}(\hat{\mu}_j - \hat{c}_{j,k})\left(\frac{\mathrm{cov}(C_{j,\cdot},\hat{C}_{j,\cdot})}{n\hat{\sigma}_j^2}\big(\hat{C}_{jk}-\hat{\mu}_j\big) - \frac{1}{n}(C_{jk}-\mu_j)\right) \\
&+ \frac{1}{n^2\sigma_j\hat{\sigma}_j^3}\left[(C_{jk}-\mu_j)(\hat{C}_{jk}-\hat{\mu}_j) - \frac{2}{\hat{\sigma}_j^2}\mathrm{cov}(C_{j,\cdot},\hat{C}_{j,\cdot})(\hat{C}_{jk}-\hat{\mu}_j)^2\right. \\
&\left. +(n-1)\mathrm{cov}(C_{j,\cdot},\hat{C}_{j,\cdot})\right] \qquad \text{(A.8)}
\end{aligned}
$$

and

$$
\begin{aligned}
\textcircled{B} =& \left(\frac{\partial}{\partial g_l}\left((X^T\Gamma X)^{-1}X^T\delta(i)\big(1 - X(X^T\Gamma X)^{-1}X^T\Gamma\big)Y\right)\right)_{j,k} \\
=& \Big((X^T\Gamma X)^{-1}X^T\left\{\delta(l)X(X^T\Gamma X)^{-1}X^T\delta(i)(-1 + X(X^T\Gamma X)^{-1}X^T\Gamma)\right. \\
&\left. +\delta(i)X(X^T\Gamma X)^{-1}\left[X^T\delta(l)X(X^T\Gamma X)^{-1}X^T\Gamma - X^T\delta(l)\right]\right\}Y\Big)_{j,k}. \qquad \text{(A.9)}
\end{aligned}
$$

Bibliography

[1] Franziska Görtler, Stefan Solbrig, Tilo Wettig, Peter J. Oefner, Rainer Spang, and Michael Altenbuchinger. Research in Computational Molecular Biology: 22nd Annual International Conference, RECOMB 2018, Paris, France, April 21-24, 2018, Proceedings (Lecture Notes in Computer Science). Springer, 2018.

[2] Krebs in Deutschland fr 2013/2014, 2017.

[3] Preetha Anand, Ajaikumar B. Kunnumakara, Chitra Sundaram, Kuzhuvelil B. Harikumar, Sheeja T. Tharakan, Oiki S. Lai, Bokyung Sung, and Bharat B. Aggarwal. Cancer is a preventable disease that requires major lifestyle changes. Pharmaceutical Research, 25(9):2097–2116, jul 2008.

[4] Amy Berrington de González. Projected cancer risks from computed tomographic scans performed in the united states in 2007. Archives of Internal Medicine, 169(22):2071, dec 2009.

[5] Krebs - eine Nebenwirkung der Evolution?, 2007.

[6] Cancer warning over stem cell therapies, 2007.

[7] Claire M. Vajdic and Marina T. van Leeuwen. Cancer incidence and risk factors after solid organ transplantation. International Journal of Cancer, 125(8):1747–1754, oct 2009.

[8] Gillian K Reeves, Kirstin Pirie, Valerie Beral, Jane Green, Elizabeth Spencer, and Diana Bull. Cancer incidence and mortality in relation to body mass index in the million women study: cohort study. BMJ, 335(7630):1134, nov 2007.

[9] Eugenia E Calle. Obesity and cancer. BMJ, 335(7630):1107–1108, nov 2007.

[10] Robert D. Schreiber, Lloyd J. Old, and Mark J. Smyth. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. Science, 331(6024):1565–1570, 2011.

[11] Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science, 313(5795):1960–1964, 2006.

[12] Wolf Herman Fridman, Franck Pages, Catherine Sautes-Fridman, and Jérôme Galon. The immune contexture in human tumours: impact on clinical outcome. Nature Reviews Cancer, 12(4):298, 2012.

[13] Jian-Qing Gao, Naoki Okada, Tadanori Mayumi, and Shinsaku Nakagawa. Immune cell recruitment and cell-based system for cancer therapy. Pharmaceutical Research, 25(4):752–768, sep 2007.

[14] Manfred Schuster, Andreas Nechansky, and Ralf Kircheis. Cancer immunotherapy. Biotechnology Journal, 1(2):138–147, feb 2006.

[15] Thomas Hinz, Christian J. Buchholz, Ton van der Stappen, Klaus Cichutek, and Ulrich Kalinke. Manufacturing and quality control of cell-based tumor vaccines: A scientific and a regulatory perspective. Journal of Immunotherapy, 29(5):472–476, sep 2006.

[16] Gavin P. Dunn, Lloyd J. Old, and Robert D. Schreiber. The immunobiology of cancer immunosurveillance and immunoediting. Immunity, 21(2):137–148, aug 2004.

[17] José A Guevara-Patiño, Mary Jo Turk, Jedd D Wolchok, and Alan N Houghton. Immunity to cancer through immune recognition of altered self: Studies with melanoma. In Advances in Cancer Research, pages 157–177. Elsevier, 2003.

[18] Gudrun Lang. Histotechnik: Praxislehrbuch fr die Biomedizinische Analytik (German Edition). Springer, 2012.

[19] W. W. Soon, M. Hariharan, and M. P. Snyder. High-throughput sequencing for biology and medicine. Molecular Systems Biology, 9(1):640–640, apr 2014.

[20] Eric J. Lanni, Stanislav S. Rubakhin, and Jonathan V. Sweedler. Mass spectrometry imaging and profiling of single cells. Journal of Proteomics, 75(16):5036–5051, aug 2012.

[21] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, 2005.

[22] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 34(3):267–273, June 2003.

[23] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12(5):453–457, 2015.

[24] Jos A. M. Borghans, Andr J. Noest, and Rob J. De Boer. How Specific Should Immunological Memory Be? The Journal of Immunology, 163(2):569575, 1999.

[25] Maren Claus, Johann Greil, and Carsten Watzl. Comprehensive analysis of NK cell function in whole blood samples. Journal of Immunological Methods, 341(1-2):154–164, feb 2009.

[26] Jietang Mai, Anthony Virtue, Jerry Shen, Hong Wang, and Xiao-Feng Yang. An evolving new paradigm: endothelial cells – conditional innate immune cells. Journal of Hematology & Oncology, 6(1):61, 2013.

[27] Agata A. Filip, Bogumiła Ciseł, Dorota Koczkodaj, Ewa Wasik-Szczepanek, Tomasz Piersiak, and Anna Dmoszynska. Circulating microenvironment of CLL: Are nurse-like cells related to tumor-associated macrophages? Blood Cells, Molecules, and Diseases, 50(4):263–270, apr 2013.

[28] Hubert Hackl, Pornpimol Charoentong, Francesca Finotello, and Zlatko Trajanoski. Computational genomics tools for dissecting tumour-immune cell interactions. Nature Reviews Genetics, 17(8):441–458, 2016.

[29] Sherrif F Ibrahim and Ger van den Engh. Flow cytometry and cell sorting. In Cell Separation, pages 19–39. Springer, 2007.

[30] Sean C Bendall, Erin F Simonds, Peng Qiu, D Amir El-ad, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science, 332(6030):687–696, 2011.

[31] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nature Methods, 11(1):41–46, 2014.

[32] E. Andres Houseman, Molly L. Kile, David C. Christiani, Tan A. Ince, Karl T. Kelsey, and Carmen J. Marsit. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics, 17(1), jun 2016.

[33] Yongjun Chu and David R. Corey. RNA sequencing: Platform selection, experimental design, and data interpretation. Nucleic Acid Therapeutics, 22(4):271–274, aug 2012.

[34] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1):57–63, jan 2009.

[35] E. TAUB FLOYD, JAMES M. DeLEO, and E. BRAD THOMPSON. Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. DNA, 2(4):309–327, dec 1983.

[36] Alexander R. Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLOS ONE, 4(7):1–16, 07 2009.

[37] Santhilal Subhash and Chandrasekhar Kanduri. GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. BMC Bioinformatics, 17(1), sep 2016.

[38] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1):44–57, jan 2009.

[39] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, sep 2005.

[40] Abbas A R, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, Fong S, M van Lookeren Campagne, Godowski P, Williams P M, Chan A C, and Clark H F. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes And Immunity, 6:319, mar 2005.

[41] Ting Gong and Joseph D. Szustakowski. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics, 29(8):1083–1085, 2013.

[42] Li Bo, Severson Eric, Pignon Jean-Christophe, Zhao Haoquan, Li Taiwen, Novak Jesse, Jiang Peng, Shen Hui, Aster Jon C., Rodig Scott, Signoretti Sabina, Liu Jun S., and Liu X. Shirley. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biology, 17(1):174, 2016.

[43] Newman Aaron M, Liu Chih Long, Green Michael R, Gentles Andrew J, Feng Weiguo, Xu Yue, Hoang Chuong D, Diehn Maximilian, and Alizadeh Ash A. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12:453, mar 2015.

[44] Oliver Stein. Konvexe Optimierungsprobleme, pages 37–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018.

[45] N. Bourbaki. Konvexe und konkave Funktionen, 2018.

[46] Konrad Königsberger. Analysis 2. Springer Berlin Heidelberg, 2004.

[47] Johannes Schneider. Datei:Kugelkoordinaten 2.svg, 2015.

[48] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352(6282):189–196, 2016.

[49] International Agency for Research on Cancer. World Cancer Report 2014 (International Agency for Research on Cancer). World Health Organization, 2014.

[50] Holly E. Kanavy and Meg R. Gerstenblith. Ultraviolet radiation and melanoma. Seminars in Cutaneous Medicine and Surgery, 30(4):222–228, dec 2011.

[51] Sad C. Azoury and Julie R. Lange. Epidemiology, risk factors, prevention, and early detection of melanoma. Surgical Clinics of North America, 94(5):945–962, oct 2014.

[52] Scanning for melanoma , 2010.

[53] R. J. Friedman, D. S. Rigel, and A. W. Kopf. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. CA: A Cancer Journal for Clinicians, 35(3):130–151, may 1985.

[54] J. M. Mascaro. The dermatologist's position concerning nevi: A vision ranging from ”the ugly duckling” to ”little red riding hood”. Archives of Dermatology, 134(11):1484–1485, nov 1998.

[55] Alexandrov Ludmil B., Nik-Zainal Serena, Wedge David C., Aparicio Samuel A. J. R., Behjati Sam, Biankin Andrew V., Bignell Graham R., Bolli Niccol, Borg Ake, Brresen-Dale Anne-Lise, Boyault Sandrine, Burkhardt Birgit, Butler Adam P., Caldas Carlos, Davies Helen R., Desmedt Christine, Eils Roland, Eyfjrd Jrunn Erla, Foekens John A., Greaves Mel, Hosoda Fumie, Hutter Barbara, Ilicic Tomislav, Imbeaud Sandrine, Imielinski Marcin, Jger Natalie, Jones David T. W., Jones David, Knappskog Stian, Kool Marcel, Lakhani Sunil R., Lpez-Otn Carlos, Martin Sancha, Munshi Nikhil C., Nakamura Hiromi, Northcott Paul A., Pajic Marina, Papaemmanuil Elli, Paradiso Angelo, Pearson John V., Puente Xose S., Raine Keiran, Ramakrishna Manasa, Richardson Andrea L., Richter Julia, Rosenstiel Philip, Schlesner Matthias, Schumacher Ton N., Span Paul N., Teague Jon W., Totoki Yasushi, Tutt Andrew N. J., Valds-Mas Rafael, van Buuren Marit M., van t Veer Laura, Vincent-Salomon Anne, Waddell Nicola, Yates Lucy R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi Jessica, Andrew Futreal P., McDermott Ultan, Lichter Peter, Meyerson Matthew, Grimmond Sean M., Siebert Reiner, Campo Elas, Shibata Tatsuhiro, Pfister Stefan M., Campbell Peter J., and Stratton Michael R. Signatures of mutational processes in human cancer. Nature, 500:415, aug 2013.

[56] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer Genome Landscapes. Science, 339(6127):15461558, 2013.

[57] Nienke van Rooij, Marit M. van Buuren, Daisy Philips, Arno Velds, Mireille Toebes, Bianca Heemskerk, Laura J.A. van Dijk, Sam Behjati, Henk Hilkmann, Dris el Atmioui, Marja Nieuwland, Michael R. Stratton, Ron M. Kerkhoven, Can Kemir, John B. Haanen, Pia Kvistborg, and Ton N. Schumacher. Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. Journal of Clinical Oncology, 31(32):e439e442, 2013. PMID: 24043743.

[58] Kenneth Murphy. Janeway's Immunobiology (Immunobiology: The Immune System (Janeway)). Garland Science, 2011.

[59] Kanika Garg, Margarita Maurer, Johannes Griss, Marie-Charlotte Brggen, Ingrid H. Wolf, Christine Wagner, Niels Willi, Kirsten D. Mertz, and Stephan N. Wagner. Tumor-associated B cells in cutaneous primary melanoma and improved clinical outcome. Human Pathology, 54:157164, 2016.

[60] Daniel Abate-Daga, Theresa A. Alexander, Evgeny Arons, Mitchell Ho, Paul F. Robbins, Steven A. Rosenberg, and Richard A. Morgan. Melanoma-associated B cells are distinct from peripheral blood-derived B cells, and may serve as a source of tumor-targeting antibodies. Journal for ImmunoTherapy of Cancer, 1(1):P180, Nov 2013.

[61] Kenneth M. Murphy, Paul Travers, and Mark Walport. Janeway Immunologie (German Edition). Spektrum Akademischer Verlag, 2009.

[62] Fujimura Taku, Kambayashi Yumi, Fujisawa Yasuhiro, Hidaka Takanori, and Aiba Setsuya. Tumor-Associated Macrophages: Therapeutic Targets for Skin Cancer. Frontiers in oncology, 8:33, jan 2018.

[63] Pieniazek Malgorzata, Matkowski Rafal, and Donizy Piotr. Macrophages in skin melanoma-the key element in melanomagenesis. Oncology letters, 15(4):53995404, apr 2018.

[64] William C. Aird. 3 - Endothelium. In Craig S. Kitchens, Craig M. Kessler, and Barbara A. Konkle, editors, Consultative Hemostasis and Thrombosis (Third Edition), page 3341. W.B. Saunders, Philadelphia, third edition edition, 2013.

[65] Yukihiko Kato, L. U. Zhang, and Roberto Pili. Endothelial cells promote metastatic melanoma cell invasion via alpha(v)beta(3) integrin. Cancer Research, 65(9 Supplement):889889, 2005.

[66] Maren Claus, Johann Greil, and Carsten Watzl. Comprehensive analysis of NK cell function in whole blood samples. Journal of Immunological Methods, 341(1):154–164, 2009.

[67] Tarazona Raquel, Duran Esther, and Solana Rafael. Natural Killer Cell Recognition of Melanoma: New Clues for a More Effective Immunotherapy. Frontiers in immunology, 6:649649, jan 2016.

[68] Cirri Paolo and Chiarugi Paola. Cancer associated fibroblasts: the dark side of the coin. American journal of cancer research, 1(4):482497, mar 2011.

[69] De Veirman Kim, Rao Luigia, De Bruyne Elke, Menu Eline, Van Valckenborgh Els, Van Riet Ivan, Frassanito Maria Antonia, Di Marzo Lucia, Vacca Angelo, and Vanderkerken Karin. Cancer associated fibroblasts and tumor growth: focus on multiple myeloma. Cancers, 6(3):13631381, jun 2014.

[70] Zhou Linli, Yang Kun, Andl Thomas, Wickett R Randall, and Zhang Yuhang. Perspective of Targeting Cancer-Associated Fibroblasts in Melanoma. Journal of Cancer, 6(8):717726, jun 2015.

[71] Wei-Chen Chen, George Ostrouchov, Drew Schmidt, Pragneshkumar Patel, and Hao Yu. pbdMPI: Programming with Big Data – Interface to MPI, 2012. R Package, URL https://cran.r-project.org/package=pbdMPI.

[72] Wei-Chen Chen, George Ostrouchov, Drew Schmidt, Pragneshkumar Patel, and Hao Yu. A Quick Guide for the pbdMPI Package, 2012. R Vignette, URL https://cran.r-project.org/package=pbdMPI.

[73] Peter Georg, Daniel Richtmann, and Tilo Wettig. DD-αAMG on QPACE 3. arXiv.org, 1710.07041, 2017.

[74] André Veillette, Michael A. Bookman, Eva M. Horak, and Joseph B. Bolen. The CD4 and CD8 T cell surface antigens are associated with the internal membrane tyrosine-protein kinase p56lck. Cell, 55(2):301–308, 1988.

[75] Elena G. Addison, Janet North, Ismail Bakhsh, Chloe Marden, Sumaira Haq, Samia Al-Sarraj, Reza Malayeri, R. Gitendra Wickremasinghe, Jeffrey K. Davies, and Mark W. Lowdell. Ligation of CD8 on human natural killer cells prevents activation-induced apoptosis and enhances cytolytic activity. Immunology, 116(3):354–361, 2005.

[76] Alessandro Moretta, Cristina Bottino, Massimo Vitale, Daniela Pende, Claudia Cantoni, Maria Cristina Mingari, Roberto Biassoni, and Lorenzo Moretta. Activating Receptors and Coreceptors Involved in Human Natural Killer Cell-Mediated Cytolysis. Annual Review of Immunology, 19(1):197–223, 2001. PMID: 11244035.

[77] R. C. Rickert, K. Rajewsky, and J. Roes. Impairment of T-cell-dependent B-cell responses and B-l cell development in CD19-deficient mice. Nature, 376(6538):352–355, jul 1995. 10.1038/376352a0.

[78] Haidong Li, Linda M. Ayer, Jonathan Lytton, and Julie P. Deans. Store-operated Cation Entry Mediated by CD20 in Membrane Rafts. Journal of Biological Chemistry, 278(43):42427–42434, 2003.

[79] Robert C. Hsueh and Richard H. Scheuermann. Tyrosine kinase activation in the decision between growth, differentiation, and death responses initiated from the B cell antigen receptor. In Advances in Immunology, volume Supplement C (75), pages 283–316. Academic Press, 2000.

[80] Jürgen Wienands, Jutta Schweikert, Bernd Wollscheid, Hassan Jumaa, Peter J Nielsen, and Michael Reth. SLP-65: a new signaling component in B lymphocytes which requires expression of the antigen receptor for phosphorylation. Journal of Experimental Medicine, 188(4):791–795, 1998.

[81] Shohei Hori, Takashi Nomura, and Shimon Sakaguchi. Control of Regulatory T Cell Development by the Transcription Factor Foxp3. Science, 299(5609):1057–1061, 2003.

[82] Alain Haziot, Enza Ferrero, Frank Köntgen, Naoki Hijiya, Shunsuke Yamamoto, Jack Silver, Colin L Stewart, and Sanna M Goyert. Resistance to Endotoxin Shock and Reduced Dissemination of Gram-Negative Bacteria in CD14-Deficient Mice. Immunity, 4(4):407–414, 1996.

[83] Charles J. Sherr, Carl W. Rettenmier, Rosalba Sacca, Martine F Roussel, A. Thomas Look, and E. Richard Stanley. The c-fms proto-oncogene product is related to the receptor for the mononuclear phagocyte growth factor, CSF 1. Cell, 41(3):665–676, 1985.

[84] Y. J. Wu, D. P. La Pierre, J. Wu, A. J. Yee, and B. B. Yang. The interaction of versican with its binding partners. Cell Research, 15(7):483–494, jul 2005.

[85] William Du, Weining Yang, and Albert J Yee. Roles of versican in cancer biology - Tumorigenesis, progression and metastasis. Histology and Histopathology, 28(6):701–713, 03 2013.

[86] Anshita Gupta, Chanchal Deep Kaur, Manmohan Jangdey, and Swarnlata Saraf. Matrix metalloproteinase enzymes and their naturally derived inhibitors: Novel targets in photocarcinoma therapy. Ageing Research Reviews, 13(Supplement C):65–74, 2014.

[87] Julie B Sneddon, Hanson H Zhen, Kelli Montgomery, Matt van de Rijn, Aaron D Tward, Robert West, Hayes Gladstone, Howard Y Chang, Greg S Morganroth, Anthony E Oro, et al. Bone morphogenetic protein antagonist gremlin 1 is widely expressed by cancer-associated stromal cells and can promote tumor cell proliferation. Proceedings of the National Academy of Sciences USA, 103(40):14842–14847, 2006.

[88] S. Gory-Faure, M.H. Prandini, H. Pointu, V. Roullot, I. Pignot-Paintrand, M. Vernet, and P. Huber. Role of vascular endothelial-cadherin in vascular morphogenesis. Development, 126(10):2093–2102, 1999.

[89] Chunwei Shi, Lu Jia, Wu Wen, Ma Fanxin, Georges Joseph, Huang Hanju, Balducci James, Chang Yongchang, and Huang Yao. Endothelial Cell-Specific Molecule 2 (ECSM2) Localizes to

Cell-Cell Junctions and Modulates bFGF-Directed Cell Migration via the ERK-FAK Pathway. PloS One, 6(6):1–15, 06 2011.

[90] Reiner F Haseloff, Sophie Dithmer, Lars Winkler, Hartwig Wolburg, and Ingolf E Blasig. Transmembrane proteins of the tight junctions at the blood–brain barrier: structural and functional aspects. In Seminars in cell & developmental biology, volume 38, pages 16–25. Elsevier, 2015.

[91] J. Evan Sadler. Biochemistry and Genetics of von Willebrand Factor. Annual Review of Biochemistry, 67(1):395–424, 1998. PMID: 9759493.

[92] Azhari Aziz, Sean P Harrop, and Naomi E Bishop. DIA1R is an X-linked gene related to Deleted In Autism-1. PLoS One, 6(1):e14534, 2011.

[93] Georgi K Marinov, Brian A Williams, Ken McCue, Gary P Schroth, Jason Gertz, Richard M Myers, and Barbara J Wold. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Research, 24(3):496–510, 2014.

[94] André F. Rendeiro, Thomas Krausgruber, Nikolaus Fortelny, Fangwen Zhao, Thomas Penz, Matthias Farlik, Linda C. Schuster, Amelie Nemc, Szabolcs Tasnády, Marienn Réti, Zoltán Mátrai, Donat Alpar, Csaba Bödör, Christian Schmidl, and Christoph Bock. Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib drug response in chronic lymphocytic leukemia. bioRxiv, 2019.

[95] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. Journal of Machine Learning Research, 9:2579–2605, 2008.

[96] Nicholas Chiorazzi, Kanti R. Rai, and Manlio Ferrarini. Chronic Lymphocytic Leukemia. New England Journal of Medicine, 352(8):804815, 2005. PMID: 15728813.

[97] Federico Caligaris-Cappio, Donata Gottardi, A Alfarano, Alessandra Stacchini, M G Gregoretti, P Ghia, M T Bertero, Anna Novarino, and L Bergui. The nature of the B lymphocyte in B-chronic lymphocytic leukemia. 19:601–13, 02 1993.

[98] Frdric Boissard, Jean-Jacques Fourni, Camille Laurent, Mary Poupot, and Loc Ysebaert. Nurse like cells: chronic lymphocytic leukemia associated macrophages. Leukemia & Lymphoma, 56(5):15701572, 2015. PMID: 25586606.

[99] L. Lagneaux, A. Delforge, D. Bron, C. De Bruyn, and P. Stryckmans. Chronic Lymphocytic Leukemic B Cells But Not Normal B Cells Are Rescued From Apoptosis by Contact With Normal Bone Marrow Stromal Cells. Blood, 91(7):2387–2396, 1998.

[100] Jan A. Burger, Maite P. Quiroga, Elena Hartmann, Andrea Bürkle, William G. Wierda, Michael J. Keating, and Andreas Rosenwald. High-level expression of the T-cell chemokines CCL3 and CCL4 by chronic lymphocytic leukemia B cells in nurselike cell cocultures and after BCR stimulation. Blood, 113(13):3050–3058, 2009.

[101] J.A. Burger, N Tsukada, Meike Burger, N.J. Zvaifler, M Dell'Aquila, and Thomas Kipps. Blood-derived nurse-like cells protect chronic lymphocytic leukemia B cells from spontaneous apoptosis through stromal cell-derived factor-1. 96:2655–63, 11 2000.

[102] N. Tsukada. Distinctive features of "nurselike" cells that differentiate in the context of chronic lymphocytic leukemia. Blood, 99(3):1030–1037, feb 2002.

[103] Nishio Mitsufumi, Endo Tomoyuki, Tsukada Nobuhiro, Ohata Junko, Kitada Shinichi, Reed John C, Zvaifler Nathan J, and Kipps Thomas J. Nurselike cells express BAFF and APRIL, which can promote survival of chronic lymphocytic leukemia cells via a paracrine pathway distinct from that of SDF-1. Blood, 106(3):10121020, mar 2005.

[104] J. Hoellenriegel, S. A. Meadows, M. Sivina, W. G. Wierda, H. Kantarjian, M. J. Keating, N. Giese, S. O'Brien, A. Yu, L. L. Miller, B. J. Lannutti, and J. A. Burger. The phosphoinositide 3'-kinase delta inhibitor, CAL-101, inhibits b-cell receptor signaling and chemokine networks in chronic lymphocytic leukemia. Blood, 118(13):3603–3612, jul 2011.

[105] S. Deaglio. CD38 and CD100 lead a network of surface receptors relaying positive signals for b-CLL growth and survival. Blood, 105(8):3042–3050, apr 2005.

[106] Loic Ysebaert and Jean-Jacques Fournié. Genomic and phenotypic characterization of nurse-like cells that promote drug resistance in chronic lymphocytic leukemia. Leukemia & Lymphoma, 52(7):1404–1406, jun 2011.

[107] F Boissard, Jean Jacques Fournie, Quillet-Mary Anne, Loic Ysebaert, and Mary Poupot. Nurse-like cells mediate ibrutinib resistance in chronic lymphocytic leukemia patients. 5:e355, 10 2015.

[108] Catharina Medrek, Fredrik Pontén, Karin Jirström, and Karin Leandersson. The presence of tumor associated macrophages in tumor stroma as a prognostic marker for breast cancer patients. BMC Cancer, 12(1), jul 2012.

[109] Francesco Marchesi, Mariangela Cirillo, Antonella Bianchi, Michela Gately, Odoardo M. Olimpieri, Elisabetta Cerchiara, Daniela Renzi, Alessandra Micera, Bjorn O. Balzamino, Stefano Bonini, Andrea Onetti Muda, and Giuseppe Avvisati. High density of CD68+/CD163+ tumour-associated macrophages (m2-TAM) at diagnosis is significantly correlated to unfavorable prognostic factors and to poor clinical outcomes in patients with diffuse large b-cell lymphoma. Hematological Oncology, 33(2):110–112, apr 2014.

[110] Marzia Palma, Giusy Gentilcore, Kia Heimersson, Fariba Mozaffari, Barbro Näsman-Glaser, Emma Young, Richard Rosenquist, Lotta Hansson, Anders Österborg, and Håkan Mellstedt. T cells in chronic lymphocytic leukemia display dysregulated expression of immune checkpoints and activation markers. Haematologica, 102(3):562–572, 2017.

[111] Alexander W. MacFarlane, Mowafaq Jillab, Mitchell R Smith, R. Katherine Alpaugh, Marion E. Cole, Samuel Litwin, Michael M. Millenson, Tahseen I. Al-Saleem, Adam D. Cohen, and Kerry S. Campbell. Natural Killer Cell Dysfunction in Chronic Lymphocytic Leukemia Is Associated with Loss of the Mature KIR3DL1+ Subset. Blood, 124(21):3318–3318, 2014.

[112] Filip Agata A, Ciseł Bogumiła, and Wsik-Szczepanek Ewa. Guilty bystanders: nurse-like cells as a model of microenvironmental support for leukemic lymphocytes. Clinical and Experimental Medicine, 15:73–83, nov 2013.

[113] F Boissard, C Laurent, A G Ramsay, Quillet-Mary Anne, Jean Jacques Fournie, Mary Poupot, and Loic Ysebaert. Nurse-like cells impact on disease progression in chronic lymphocytic leukemia. 6:e381, 01 2016.

[114] Chistiakov Dimitry A, Killingsworth Murry C, Myasoedova Veronika A, Orekhov Alexander N, and Bobryshev Yuri V. CD68/macrosialin: not just a histochemical marker. Laboratory Investigation, 97:4, nov 2016.

[115] Zhao Xia, Zhang Wei, Wang Li, and Zhao Wei-Li. Genetic methylation and lymphoid malignancies: biomarkers of tumor progression and targeted therapy. Biomarker Research, 1:24–24, aug 2013.

[116] Roda-Navarro Pedro, Arce Ignacio, Renedo Mónica, Montgomery Kate, Kucherlapati Raju, and Fernández-Ruiz Elena. Human KLRF1, a novel member of the killer cell lectin-like receptor gene family: molecular characterization, genomic structure, physical mapping to the NK gene complex and expression analysis. European Journal of Immunology, 30(2):568–576, 1.

[117] Li Jie, Qin Yi, and Zhang Haiyan. Identification of key miRNA-gene pairs in chronic lymphocytic leukemia through integrated analysis of mRNA and miRNA microarray. Oncology Letters, 15(1):361367, oct 2017.

[118] Yair Herishanu, Patricia Prez-Galn, Delong Liu, Anglique Biancotto, Stefania Pittaluga, Berengere Vire, Federica Gibellini, Ndegwa Njuguna, Elinor Lee, Lawrence Stennett, Nalini Raghavachari, Poching Liu, J. Philip McCoy, Mark Raffeld, Maryalice Stetler-Stevenson, Constance Yuan, Richard Sherry, Diane C. Arthur, Irina Maric, Therese White, Gerald E. Marti, Peter Munson, Wyndham H. Wilson, and Adrian Wiestner. The lymph node microenvironment promotes B-cell receptor signaling, NF-B activation, and tumor proliferation in chronic lymphocytic leukemia. Blood, 117(2):563574, 2011.

[119] Dimitar G. Efremov, Stefania Gobessi, and Pablo G. Longo. Signaling pathways activated by antigen-receptor engagement in chronic lymphocytic leukemia B-cells. Autoimmunity Reviews, 7(2):102108, 2007. B Cell Targeted Therapies.

[120] Valerie Pede, Rombout Ans, Vermeire Jolien, Naessens Evelien, Mestdagh Pieter, Robberecht Nore, Vanderstraeten Hanne, Van Roy Nadine, Vandesompele Jo, Speleman Frank, Philipp Jan, and Verhasselt Bruno. CLL Cells Respond to B-Cell Receptor Stimulation with a MicroRNA/mRNA Signature Associated with MYC Activation and Cell Cycle Progression. PLOS ONE, 8(4):112, 04 2013.

[121] Pede Valerie, Rombout Ans, Vermeire Jolien, Naessens Evelien, Mestdagh Pieter, Robberecht Nore, Vanderstraeten Hanne, Van Roy Nadine, Vandesompele Jo, Speleman Frank, Philipp Jan, and Verhasselt Bruno. CLL Cells Respond to B-Cell Receptor Stimulation with a MicroRNA/mRNA Signature Associated with MYC Activation and Cell Cycle Progression. PLoS ONE, 8(4):e60275, feb 2013.

[122] FOS Gene(Protein Coding), 2019.

[123] VSIR Gene(Protein Coding), 2019.

[124] Frankenberger Marion, Hofer Thomas P.J., Marei Ayman, Dayyani Farshid, Schewe Stefan, Strasser Christine, Aldraihim Asaad, Stanzel Franz, Lang Roland, Hoffmann Reinhard, Costa Olivia Prazeres da, Buch Thorsten, and Ziegler-Heitbrock Loems. Transcript profiling of CD16-positive monocytes reveals a unique molecular fingerprint. European Journal of Immunology, 42(4):957–974, 1.

[125] Maffei Rossana, Bulgarelli Jenny, Fiorcari Stefania, Bertoncelli Linda, Martinelli Silvia, Guarnotta Carla, Castelli Ilaria, Deaglio Silvia, Debbia Giulia, De Biasi Sara, Bonacorsi Goretta, Zucchini Patrizia, Narni Franco, Tripodo Claudio, Luppi Mario, Cossarizza Andrea, and Marasca Roberto. The monocytic population in chronic lymphocytic leukemia shows altered composition and deregulation of genes involved in phagocytosis and inflammation. Haematologica, 98(7):1115–1123, jan 2013.

[126] Peter P. Ruvolo. Galectin 3 as a guardian of the tumor microenvironment. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1863(3):427–437, 2016. Tumor Microenvironment Regulation of Cancer Cell Survival, Metastasis, Inflammation and Immune Surveillance.

[127] Martin H Deininger, Richard Meyermann, and Hermann J Schluesener. The allograft inflammatory factor-1 family of proteins. FEBS Letters, 514(2-3):115121.

[128] Alarcón B, Ley S C, Sánchez-Madrid F, Blumberg R S, Ju S T, Fresno M, and Terhorst C. The CD3-gamma and CD3-delta subunits of the T cell antigen receptor can be expressed within distinct functional TCR/CD3 complexes. The EMBO Journal, 10(4):903–912, apr 1991.

[129] Zhang Jie, Xiang Yang, Ding Liya, Keen-Circle Kristin, Borlawsky Tara B, Ozer Hatice Gulcin, Jin Ruoming, Payne Philip, and Huang Kun. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. BMC Bioinformatics, 11(Suppl 9):S5S5, oct 2010.

[130] TRAC T cell receptor alpha constant [Homo sapiens (human)] , 2018.

[131] Olivia Chan, J. Daniel Burke, Darrin F. Gao, and Eleanor N. Fish. The chemokine CCL5 regulates glucose uptake and AMP kinase signaling in activated t cells to facilitate chemotaxis. Journal of Biological Chemistry, 287(35):29406–29416, jul 2012.

[132] Steffen Stenger, Dennis A. Hanson, Rachel Teitelbaum, Puneet Dewan, Kayvan R. Niazi, Christopher J. Froelich, Tomas Ganz, Sybille Thoma-Uszynski, Agustı´n Melin, Christian Bogdan, Steven A. Porcelli, Barry R. Bloom, Alan M. Krensky, and Robert L. Modlin. An Antimicrobial Activity of Cytolytic T Cells Mediated by Granulysin. Science, 282(5386):121125, 1998.

[133] Anmei Deng, Sunxiao Chen, Qing Li, Shu-chen Lyu, Carol Clayberger, and Alan M. Krensky. Granulysin, a Cytolytic Molecule, Is Also a Chemoattractant and Proinflammatory Activator. The Journal of Immunology, 174(9):52435248, 2005.

[134] Chiho Goda, Taisuke Kanaji, Sachiko Kanaji, Go Tanaka, Kazuhiko Arima, Shigeaki Ohno, and Kenji Izuhara. Involvement of IL-32 in activation-induced cell death in t cells. International Immunology, 18(2):233–240, jan 2006.

[135] Carol A. Dahl, Fritz H. Bach, Wing Chan, Kay Huebner, Giandomenico Russo, Carlo M. Croce, Thomas Herfurth, and J. Scott Cairns. Isolation of a cDNA clone encoding a novel form of granzyme B from human NK cells and mapping to chromosome 14. Human Genetics, 84(5):465470, Apr 1990.

[136] CTSW cathepsin W [Homo sapiens (human)] , 2018.

[137] L. G. Hidalgo, G. Einecke, K. Allanach, and P. F. Halloran. The Transcriptome of Human Cytotoxic T Cells: Similarities and Disparities Among Allostimulated CD4+ CTL, CD8+ CTL and NK cells. American Journal of Transplantation, 8(3):627636.

[138] Zeev Altboum, Yael Steuerman, Eyal David, Zohar Barnett-Itzhaki, Liran Valadarsky, Hadas Keren-Shaul, Tal Meningher, Ella Mendelson, Michal Mandelboim, Irit Gat-Viks, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Molecular Systems Biology, 10(2):720, 2014.