A Two-Scale Map of Global Gene Expression for Characterising *in Vitro* Engineered Cells

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Michael Lenz

Berichter: Universitätsprofessor Dr. rer. nat. Andreas Schuppert
Universitätsprofessor Dr. rer. nat. Martin Zenke
Tag der mündichten Prüfung: 05.05.2015

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Abstract

In recent years the possibility to engineer cells *in vitro* has encountered significant progress. This engineering of cellular states, which is tightly coupled to the field of stem cell research, is considered to be a very useful technology for the generation of specialised cells for drug development, disease modelling, or regenerative medicine. One important part of this process is the quality control, i.e. the detailed characterisation of the end products, to ensure that the transformed cells are similar to their *in vivo* counterparts.
Many markers and functional assays exist that can be used for quality control of these cells. However, most of them focus on specific, relatively narrow properties of the cells, neglecting a global overall comparison to the desired cell type.
Here, we present a genome-wide gene expression microarray based approach to cell characterisation, providing complementary information to the commonly used single gene or morphological markers. We use a dimension reduction approach to localise newly generated microarray data in the high dimensional expression space. Using a combination of unsupervised and supervised dimension reduction methods, we establish a two-scale map of global gene expression with phenotypic interpretation of the coordinates.
This two-scale map is used to characterise several different samples. It is first validated on a dataset of 24 different tissues and cell lines as well as on two datasets of artificially mixed tissues. Using these datasets, it is shown that the developed method outperforms three existing methods for RNA based global cell characterisation and that it provides increased information compared to the purely unsupervised or purely supervised dimension reduction methods.
Application of the two-scale map to characterise *in vitro* transformed cells prooves to be useful in providing complementary information to the typical marker based or morphological criteria. In this respect, we could identify two examples of *in vitro* transformed cells where the transformation process is incomplete on a global expression level. Furthermore, we can show that *in vitro* differentiation of pluripotent stem cells results in immature cells that are similar to embryonic of fetal tissues of the respective type.
Using microarray data from artificial mixtures of different tissues, we can observe clear non-linear effects in the data that fit well to the current understanding of the relationship between the RNA content of cells and the measurement signal of microarrays. Such non-linear effects are currently not captured by the proposed linear dimension reduction approach and give important hints for further improvements of the method.
In addition to quality control of in vitro transformed cells, the two-scale decomposition approach developed in this thesis may also be useful for a number of other applications, such as the analysis of drug response profiles or disease progression.

Zusammenfassung

Die gezielte Modifizierung von Zellen in der Kulturschale hat in den letzten Jahren signifikante Fortschritte gemacht. Diese eng mit der Stammzellfoschung zusammenhängende Technologie hat großes Potenzial zur Herstellung von spezialisierten Zellen für die Medikamentenentwicklung, die Modellierung von Krankheiten, sowie die regenerative Medizin. Ein wichtiger Schritt im Herstellungsprozess der Zellen ist die Qualitätskontrolle, das heißt die detaillierte Charakterisierung der zellulären Endprodukte, die sicherstellen soll, dass die generierten Zellen den entsprechenden körpereigenen Zellen möglichst stark ähneln.
Diese Qualitätskontrolle der Zellen kann mit vielen verschiedene Markern und funktionalen Untersuchungsmethoden bewerkstelligt werden. Die meisten dieser Methoden analysieren jedoch nur einen relativ beschränkten Bereich der zellulären Eigenschaften und vernachlässigen dabei den globalen Vergleich zum angestrebten Zelltyp.
In der vorliegenden Arbeit präsentieren wir einen Ansatz der auf der Analyse der genomweiten Genexpression mittels Microarrays basiert und somit komplementäre Informationen zu den üblicherweise genutzten Einzelgen oder Morphologie-basierten Markern liefert. Dazu nutzen wir einen Dimensionsreduktionsansatz um neu generierte Microarray Daten im eigentlich hochdimensionalen Genexpressionsraum zu lokalisieren. Wir kombinieren dabei unsupervidierte und supervidierte Dimensionsreduktionsmethoden und etablieren damit eine zwei-Skalen Landkarte der globalen Genexpression mit der Möglichkeit der direkten phenotypischen Interpretation der Koordinaten.
Durch die Anwendung dieser zwei-Skalen Landkarte auf *in vitro* transformierte Zellen können wir zeigen, wie nützlich dieser Ansatz zur Generierung komplementärer Informationen zu den typischen Markern oder morphologischen Informationen ist. In diesem Zusammenhang zeigen wir zwei Beispiele *in vitro* transformierter Zellen, bei denen der Transformationsprozess auf globaler Genexpressionsbasis nicht vollständig abgeschlossen ist. Desweiteren können wir zeigen, dass durch die *in vitro* Differenzierung pluripotenter Stammzellen meist nicht vollständig ausgereiften Zellen entstehen, die eine hohe Ähnlichkeit zu embryonalen oder fetalen Zellen des entsprechenden Typs aufweisen.
Basierend auf Microarray Daten von künstlichen Mischungen verschiedener Gewebe, beobachten wir deutliche nicht-lineare Effekte in den Daten, die gut zum momentanen Verständnis der Beziehung von RNS Menge und Messsignal passen. Diese nicht-linearen Effekte werden aktuell von dem hier vorgestellten linearen Dimensionsreduktionsmodell nicht berücksichtigt, geben aber wertvolle Hinweise für mögliche zukünftige Verbesserungen des Modells.
Zusätzlich zur Qualitätskontrolle *in vitro* transformierter Zellen, könnte der hier entwickelte zwei-Skalen Ansatz auch für einige andere Anwendungen, wie die Analyse der Medikamentenwirkung auf Zellen oder den Vergleich verschiedener Stadien der Krankheitsentwicklung, nützlich sein.

Acknowledgements

First of all I would like to thank my supervisor Prof. Andreas Schuppert for the proposition of this interesting topic, his very valuable advises and the fruitful discussions about the mysteries of biological data analysis. Andreas, I enjoyed (and still enjoy) the time in your group very much and I thank you for your important comments on my thesis and the high flexibility and independence you gave me for the conduction of my research. The plenty of anecdotes you told us during lunch breaks were always a very welcome escape from work and I could learn about some common pitfalls in management and daily life from them.
I also want to thank my second advisor Prof. Martin Zenke for his continuous support of my work, the helpful comments on my thesis and the very valuable discussions about the results from the microarray data analyses. Martin, I could learn a lot about stem cell biology from you and it was and is a pleasure to work together with you in the StemCell-Factory project.
I thank all my collaborators from the various research projects for the very interesting and fruitful joint research that we conducted. I want to name especially Wolfgang Wanger, Bernhard Schuldt, Franz-Josef Müller, and Ben MacArthur with whom I had several attractive collaborations. Many thanks go also to Daniela Malan, Philip Sasse, Carolin Haubenreich, Oliver Brüstle, Paul Wanek, Hatim Hemeda, and all the other partners from the StemCellFactory project for the joint work in the quality control and the overall very inspiring teamwork in building up this fancy machine. Thanks to Joana Frobel, Arne Schenk, and Roman Goetzke for the joint analysis of the methylation data and to Qiong Lin for the helpful conversations about bioinformatics and data analysis.
I want to thank especially my colleagues Steffen Schaper, Nina Kusch, Maryam Montazeri, Ali Hadizadeh, Max Eck, Jayesh Bhat as well as my former colleagues Sonja Vivas, Bernhard Schuldt, and Arne Schenk for the interesting and inspiring time with them. I enjoyed the private as well as work-related discussions and felt always very comfortable in our group. The same holds true for the plenty of colleagues from AICES with whom I had some nice retreats in Monschau and Vaals, funny conversations during lunch break, and overall a very good time. Special thanks go to my room mates in the Rogowski building (Callum, Thang, Marcus, Atanas, Bernhard, Sonia, Raheel, Arianna) for the great working environment during my first two years.
I am also very thankful to Hülya Ulu-Esser, Nicole Faber, Nadine Bachem, and all the other members of the service team for their administrative support.
Many thanks go to Steffen, Arne and my brother Christian for proof-reading of this thesis. Your comments and corrections were highly appreciated and helped me a lot in finalising

this thesis.
Last but not least I want to thank my family for their continuous support during my time as a doctoral candidate, my studies, and basically throughout my whole life. Special thanks go to my wife Katrin who pushed me forward when I was in a crisis and who helped me to relax from my work when I needed a time out. Thank you for your patience and your belief in me and my work!
Thanks to all of you and to all other people that supported me and that are not explicitly named here!

Contents

List of Figures

List of Tables

1. Introduction

Embryonic stem cells (ESCs) have been of interest to researchers for many years due to their ability to differentiate into any cell type of the human body [1]. They can therefore be used to generate new cells of a specific type for regenerative medicine, disease modelling, or drug development. However, the ethical issues associated with these cells prevented intensive research and further use of these cells in many countries.

The discovery of Takahashi, Yamanaka, and colleagues that mouse [2] as well as human [3] fibroblasts can be reprogrammed into induced pluripotent stem cells (iPSC) by specific transcription factors, has made pluripotent stem cell research significantly less critical from an ethical perspective. This led to a sharp increase of research activities in the field, with many groups working on the improvement of *in vitro* reprogramming and differentiation protocols.

However, despite several improvements made in the last years, it is still observed that especially *in vitro* differentiated cells do not fully resemble the corresponding cells from primary tissues [4, 5]. Therefore, it is very important to characterise the transformed cells appropriately and to compare them to their *in vivo* counterparts.

Characterisation of cells is classically done by histological analysis, i.e. manual morphological examination or antibody based staining of specific proteins. Furthermore, several functional tests are used to check the typical functionalities of a specific cell type. However, these methods are often either labour intensive, difficult to quantitate, inaccurate, or not well standardised [6, 7]. Therefore, there is a pressing need to develop well standardised and quantitative ways to characterise these cells.

In the last years, the way we tackle biological questions started to change. With the advent of high throughput technologies and systematic storage of omics data, important biological or medical findings are not only drawn from experiments but also from *in silico* studies [8, 9]. Furthermore, significant methodological progress has been made, allowing an integration of the available data to exploit their full potential. Such an integration of data from various studies can be used to characterise cells based on a comparison of the new data with well annotated retrospective data using multivariate statistical methods.

One way to tackle this task is the use of gene expression microarray data, which are abundantly available and give relevant information about the cell type. Existing approaches use these data to build a multi-class classifier for the diverse human tissue types [10], or calculate an enrichment score for each tissue [11], aiming to assign the new data to one specific class, i.e. one specific tissue or cell type. However, this classification into a discrete set of tissue types is not well suited to characterise cells that deviate from the known *in*

vivo tissues or that are mixtures of different cell types. Therefore, it may be better to use dimension reduction techniques, spanning a low dimensional space with physiologically relevant dimensions to allow a more continuous characterisation of cells in terms of a location in a well interpretable expression 'landscape'.
So far, studies with this aim focused on unsupervised dimension reduction techniques [11, 12], ending up with a very low dimensional approximation that allows only a very rough characterisation of cells. We extend this low dimensional space with a residual tissue specific space that is created in a supervised manner. In combination, these two spaces are referred to as a 'two scale map of global gene expression' that can be used to characterise *in vitro* differentiated cells in comparison to well annotated cell and tissue types.
This introduction gives first some further background information on stem cell research, its potentials, as well as the need to rigorously classify *in vitro* transformed cells. It then describes possibilities of using large sets of publicly available gene expression data for cell characterisation and motivates the application of dimension reduction methods to deal with the high dimensionality of these data. The chapter ends with a summary of the main contributions of this thesis and an outline of its structure and content.

1.1. Reprogramming and differentiation - cellular plasticity opens new perspectives

Humans are multicellular organisms with trillions of individual cells [13]. These cells can be classified into hundreds of different cell types [14, 15] with specific characteristics, such as a specific morphology, functionality, or expression of certain genes. All of these cells originate from a single cell, the zygote, by replication and differentiation. In this developmental process, the zygote first self-renews several times before the first differentiation step occurs, leading to formation of the early blastocyst [16]. The outer part of the early blastocyst consists of so called trophectoderm cells that become part of the placenta. The inner part, the so called inner cell mass, consists of pluripotent cells that have the potential to generate any cell type of the human body [16].
During the next steps of embryonic development, three germ layer (endoderm, mesoderm, and ectoderm) evolve. These layers were identified as distinct anatomical regions in the embryo, giving rise to different organs at later stages [17]. This early anatomical division into different parts led to a hierarchical concept of cell differentiation. According to this concept, cells take several decision steps during their differentiation. Each decision step reduces the cellular plasticity, narrowing the set of possible cell types that can evolve from that cell [18, 19]. This model of differentiation is supported by further experimental evidence, showing that cells that are transplanted from one location to another in the embryo follow their old differentiation lineage despite of the changed microenvironment [17].
These and other observations [20] indicate that cell type and lineage choices are to some extent stable with respect to disturbing factors. However, many results from the last decade

showed that the differentiated cellular states are less restricted, i.e. more plastic, than previously thought. In 2002, Theise and Krause [17] postulated that "any cell containing the entire genome, without transpositions, multiplications or deletions, has the potential to display features of any cell type of the organism from which it was derived". It may be that only a small part of these transformations are of physiological relevance [17], but the possibility to transform certain differentiated cells *in vitro* into any other cell type has been experimentally validated and is currently shaping research in the fields of stem cells, regenerative medicine, and disease modelling.

The increasing interest in these fields started with the seminal work of Takahashi and Yamanaka [2] who reprogrammed mouse fibroblasts into iPSCs by induction of four transcription factors. In 2007, a similar procedure was successfully applied for reprogramming of human fibroblasts [3, 21, 22, 23]. Since then, many different types of cells have been reprogrammed to iPSCs with a diverse set of protocols. Starting with the use of retroviral or lentiviral approaches with genomic integration, more clinically relevant, integration-free methods, have been developed [24, 25] and various small molecules have been suggested to increase the usually very low reprogramming efficiency [26]. Furthermore, it was observed that some cell types, e.g adult neural stem cells, are easier to reprogram than others, requiring the use of only two [27] or even a single [28] instead of the usual four transcription factors in the case of neural stem cells.

IPSCs are highly similar to ESCs [3], being able to self-renew indefinitely and to differentiate into all three embryonic germ layers [1]. Thus, they serve as a potential source for all cell types of the human body, including those with low accessibility, e.g. cardiomyocytes or neural cells. In contrast to ESCs, which are derived from the early blastocyst [1], generation of iPSCs avoids the use of human embryos, resulting in less ethical controversies [29] and reducing political discussions about usage regulations and their impact on scientific progress [30, 31]. In addition to this ethical advantage, a key promise of iPSCs is their use for the creation of disease-specific or personalised cells of any type, i.e. with a disease specific or personalised genome. This may improve *in vitro* disease models and reduce rejection reactions in cell therapies, having a great impact for drug development, and regenerative medicine [32, 33, 34, 35].

For modelling of diseases that are caused by single genetic mutations, the ability to create disease specific cells, i.e. with a characteristic genetic disorder, is very important. For example, many neurodegenerative diseases show a pathological processing of proteins caused by gene deletions or mutations [36, 37]. One example is the spinal muscular atrophy (SMA) disease that is caused by a deletion or mutation of the survival of motor neuron 1 (SMN1) gene resulting in a loss of alpha motor neurons in the spinal cord [38]. SMA has previously been studied in different animal models or by the analysis of fibroblasts lacking SMN1. These models have the disadvantages of species specific or cell type specific differences. However, by reprogramming of the affected fibroblasts to iPSCs and subsequent neuronal differentiation, one can circumvent both disadvantages as performed in a study by Ebert and colleagues [38, 39]. A similar procedure may be beneficial in the analysis of cardiac

dysrhythmia, e.g. in the investigation of the long-QT syndrome [40].
Despite of the promises of iPSC-derived cells for disease modelling and drug development, the usage of these cells in industrial applications is still comparably low. The reasons for this are the high costs and time consuming manual processes, as well as a lack of standardisation and sometimes insufficient quality of the products, among others [41]. The issues

Figure 1.1.: A picture of the StemCellFactory (www.stemcellfactory.de) which was developed and build up within the StemCellFactory project. The six meters long and more than two meters wide machine was constructed at RWTH Aachen and is now located and maintained at Life & Brain GmbH in Bonn. It was designed for a fully automatic reprogramming of fibroblasts into iPSCs and a subsequent differentiation of these into neural cells and cardiomyocytes. One essential step in the StemCellFactory is the quality control of the end products in order to assure their usability as an *in vitro* disease model.

of automation and standardisation are currently tackled by an interdisciplinary research project building a prototype that performs reprogramming of fibroblasts or mesenchymal stem cells (MSCs) to iPSCs and subsequent neural or cardiac differentiation [36, 42]. The StemCellFactory (Figure 1.1) is a novel automatic production unit for reprogramming, cultivation and differentiation of iPSCs, transferring the complex manual laboratory processes to an automatic production process.
However, the issue of the quality of the cellular end products can not be solved entirely by automation of the current manual protocols, but requires the improvement of biological protocols [4]. There are several concerns about the quality of the cells. For instance, mutations may occur during reprogramming as well as during prolonged culturing of pluripotent stem cells [43, 44, 45]. Another important issue are the relatively strong differences between *in vitro* differentiated cells and their *in vivo* counterparts in their gene expression status [5]. These differences open the question whether the *in vitro* generated cells are in fact of another type than their *in vivo* counterparts.

These concerns are also very important for regenerative medicine, where iPSC-derived cell products present a promising resource for patient specific cells. Despite significant progress in regenerative medicine in the last years [46], important questions remain regarding the safety of cell transplants [47], especially due to their observed tumourigenic potential [48]. Perhaps the most dramatic observation in this regard was the formation of a tumour in a boy that was treated with neural stem cells [49]. This boy with ataxia telangiectasia, an inherited neurodegenerative disease, developed a glioneuronal neoplasm roughly four years after treatment and it could be shown that the tumour cells were of non-host origin [49]. Therefore, it is important to develop criteria to control the quality of reprogrammed cells. These quality criteria should be well standardised, objective and as reliable as possible. One possibility to characterise cells is through the use of gene expression microarray data. These data are comparably well standardised and are frequently deposited in public databases, so that other researchers have access to them. Therefore, it is possible for other researchers to evaluate the results of a study based on these data. One prominent example where this has been done led to a retraction of a high profile article [50] and even to some associated legal issues. In this article, the authors claimed to generate pluripotent stem cells from adult human testis [50]. However, based on the gene expression data of these cells, another group of researchers could show that the generated cells are more similar to fibroblasts than to pluripotent stem cells [51], eventually leading to the retraction of the former article.
The present thesis contributes to this research field by integrating large amounts of publicly available microarray data to characterise *in vitro* reprogrammed and differentiated cells according to their large-scale gene expression patterns. This system-wide analysis complements typical marker-based, genomic, functional or morphological characterisations of cells.

1.2. Exploiting large public microarray datasets for a global characterisation of cells

In the last years, the way we tackle biological questions started to change. With the advent of high throughput sequencing technologies [52] and the increasing amount of datasets that are available from public databases [53, 54, 55], biology is essentially entering the field of big data and important biological or medical findings are not only drawn from experiments but also from *in silico* studies [8, 9]. These changes in biomedical research pose not only challenges to bioinformatics for storage and effective handling of the rapidly increasing amount of data [56], but also require new methods for information extraction and modelling, in order to make sense out of the data and interpret the results appropriately. For the latter challenge, it is necessary to integrate heterogenous data and to utilise the already available knowledge for interpretation of new experiments, bringing them into a broader context [57].
Data integration approaches for microarray gene expression data have been implemented

at different levels. The classical analyses started with calculation of differential gene expression between two phenotypes in order to detect genes that are associated with the phenotypical difference [58, 59]. For the interpretation of the results, knowledge from gene annotation databases is used, providing information about the genomic location of a single gene, as well as the structure, functionality, and cellular location of the associated protein [60, 61].
These analyses were then extended to the level of gene sets. Methods like gene set enrichment analysis (GSEA) [62, 63, 64] use information from gene ontology [65] or pathway databases [66] to build gene sets and identify potentially important pathways, processes or functionalities based on the differential expression of the genes in a predefined set. This analysis accounts for the fact that most cellular functionalities are accomplished by the joint action of several proteins making it necessary to study not only a single gene in isolation, but to take a broader view on the expression of a set of genes.
Looking at an even broader scale, it becomes clear that the molecular pathways themselves also do not act in isolation, but are embedded into a large network [67]. Furthermore, phenotypic changes do not only lead to a gene expression change in an individual pathway, but rather to a global genome-wide change in the expression profile [11]. Consequently, methods of data integration should also consider this genome-wide scale in order to correctly interpret system-wide effects that may lead to counterintuitive or misleading results when interpreted at fine-grained scales only [11, 68]. On such a genome-wide scale, phenotypes can be compared more directly, e.g. by linkage of gene expression changes in the experiment of interest with gene expression patterns that are associated with specific tissues, clinical parameters, or changes in the cellular environment [69, 70, 71].
For the focus of the present thesis, the linkage of global gene expression during *in vitro* differentiation experiments with tissue specific expression patterns is of high interest [72]. The required comparison of expression patterns from different sources, i.e. primary tissue samples and cultivated cells, faces several challenges due to environmental differences, biological heterogeneity in clinical samples [73], lab dependent effects, as well as technical noise. Furthermore, we aim to characterise single samples, i.e. we do not make use of replicate measurements to reduce the level of noise. Therefore, it is of utmost importance to utilise the knowledge from retrospective data to distinguish meaningful biological information from pure measurement errors.

1.3. Locating new samples in reduced dimensional gene expression spaces

The goal of this thesis is to characterise new samples, especially from *in vitro* differentiated cells, by a global gene expression based comparison to well annotated retrospective data. For this purpose, we use large amounts of publicly available microarray data. Microarray experiments measure the expression of roughly 20,000 genes at the same time in a

high-throughput manner. From a mathematical perspective, this means that one has to deal with 20,000 variables from which it is initially not known to which extend they are related to each other. Therefore, any large scale analysis of microarray data must deal with the 'curse of dimensionality' [74] and the associated counterintuitive challenges of high dimensions [75, 76]. The term 'curse of dimensionality' describes some very general features of high dimensions that are associated, for instance, with the exponential increase of data points that are necessary to sample spaces of increasing dimensionality with a sufficient density. This often leads to a prevalence of empty space in high dimensions due to restricted sample sizes. This prevalence of empty space as well as the other features of high dimensional spaces impose some problems that affect many data mining and machine learning techniques, e.g. due to the measure concentration phenomenon [75, 77]. Therefore, the use of any of these methods must be considered carefully.

The specific challenge in this thesis is to compare the location of new microarray data to that of retrospective data in the roughly 20,000 dimensional gene expression space. Due to the high dimensionality of the expression space, it is not possible to directly visualise or to imagine the location of a sample in the expression space for a direct manual interpretation of the location. Therefore, several data mining and machine learning tools exist that can assist in the interpretation of the data.

One possible tool that can be used in this respect is a k-nearest neighbour classifier. This classifier evaluates the distance of the new sample to all retrospective data and classifies the new sample based on the annotation of the k nearest neighbours from the retrospective dataset. A special case of the k-nearest neighbour classifier is the 1-nearest neighbour classifier, which classifies the new sample according to the annotation of the retrospective sample with smallest distance.

One problem of the k-nearest neighbour classifier is that it is not well suited for the analysis of high dimensional data. This is due to the measure concentration phenomenon. The measure concentration phenomenon describes the fact that with increasing dimensionality the distance of one sample to the nearest and farthest other sample become more and more similar. That means, that the distance of one sample to all other samples is almost constant in very high dimensions, making the concept of neighbourhood meaningless.

Another problem that was described for the k-nearest neighbour approach in high dimensions is the 'hubness' phenomenon [78]. The hubness phenomenon basically states that in high dimensions some points eventually become hubs, i.e. these points are very often among the k-nearest neighbours of all other points, although they are not necessarily similar to them from a biological point of view.

Instead of using a k-nearest neighbour approach for classification, several other classification algorithms can be applied, e.g. decision trees, neural networks, or support vector machines. All of these classifiers have different strengths and weaknesses and the best choice depends always on the particular application [79]. In the case of high dimensional data, it is important to consider the different pitfalls that are associated with the curse of dimensionality. The most critical problem for classification in high dimensions is the

very sparse sampling of the space by the training data, leading to a prevalence of empty space [80]. This is especially critical for highly flexible non-linear classifiers, e.g. neural networks, that can easily be overfitted to the training data. Less flexible classifiers that have more strongly constrained classification boundaries, e.g. linear classification methods, are usually less effected by the overfitting problem. However, these more constrained methods may be unable to find a proper boundary for the training data due to inappropriate constraints. Therefore, the best method choice is always a trade-off between flexibility and the risk of overfitting [79]. Any available information about appropriate constraints for the classification boundary can help to improve the methods.

The methods described so far focus on a classification of cells into a discrete set of classes, i.e. they are typically used as multi-class classifiers. However, this approach has a major drawback, since it does not incorporate the possibility that a new sample is only partially similar to an existing cell type. This can be critical in cases where tissues are mixtures of different cell types or when *in vitro* transformed cells have not fully reached the desired cell type.

One way to overcome this problem is the use of classifiers that provide probability values for each class, allowing relatively smooth transitions from one cell type to another [81]. Another approach is the use of dimension reduction techniques that create a space of lower dimensionality with direct phenotypical interpretation of the coordinates. In this simplified representation, microarray samples are not restricted to a discrete set of different classes, i.e. cell types, but can vary continuously on the specified sub-manifold. When such a reduced dimensional space has been created, new samples can be located in that space, and the location in the space can be directly used for characterisation of the cells. This is feasible due to the relatively low dimensionality of the space and the direct phenotypical interpretability of the coordinates.

Reduction of dimensionality is commonly achieved via unsupervised methods, such as principal components analysis (PCA) or non-negative matrix factorisation (NMF). However, these unsupervised methods do not directly provide an interpretation of the coordinates in terms of phenotypes. This problem can in principle be solved by supervised dimension reduction approaches that use the available annotation to determine relevant directions, e.g. based on regression methods. However, for these methods it is necessary to know the structure of the low dimensional space in advance and to have appropriate annotations of the data. Therefore, both kinds of methods have their own advantages and disadvantages and the specific choice of the method must be carefully done in an application-specific manner.

The aim of the present thesis is to create a low dimensional representation of the gene expression space that allows a relatively straightforward characterisation of new data in terms of their location in the expression space. For the derivation of this low dimensional space, we compare the strengths and weaknesses of supervised and unsupervised methods for this particular application. Based on this evaluation, we develop a new workflow that combines both approaches in order to utilise their complementary strengths.

1.4. Main contributions of the thesis

A Two-scale map of global gene expression for the characterisation of *in vitro* engineered cells

The main contribution of the present thesis is the development of a two-scale gene expression map that can be used for an absolute characterisation of cells based on gene expression microarray data. This new method uses a reduced dimensional approximation of the expression space with physiologically relevant coordinates to improve the interpretability of the location of these cells in the expression space.

We compare the new method to three existing methods [10, 11, 81] for mRNA-based cell characterisation that have been published recently and show that it outperforms all of these methods in some way. In comparison to the 'Concordia: Phenotypic concept enrichment' method of Schmid et al. [11], we can show that our method has a significantly higher accuracy of identifying the correct cell or tissue type. Furthermore, the Concordia method is only described for a single microarray platform, whereas we show that our method can be used across different microarray platforms. A further advantage of our method lies in the above mentioned ability of dimension reduction based methods to characterise cells that match only partially to a pure tissue, being able to quantify and interpret the differences in terms of a direction with a phenotypical meaning. This ability is especially important for the characterisation of *in vitro* differentiated cells that differ from any *in vivo* tissue. In contrast, the method of Schmid et al. provides only a similarity score for each tissue type, having the associated disadvantages of distance based approaches.

The 'Unknown RNA sample annotation' (URSA) method of Lee et al. [10] is applicable across platforms and can even be used to characterise RNA sequencing data. Furthermore, it delivers more specific results than the Concordia tool. However, the URSA tool is also a multi-class classifier that has problems to characterise mixture samples or *in vitro* differentiated cells that differ from the tissues in the training dataset. Thus, it is not possible, for instance, to draw appropriate conclusions about the direction of differentiation in time series experiments or to evaluate the grade of differentiation.

Finally the 'CellNet' method of Cahan et al. [81] is well suited for the characterisation of *in vitro* differentiated cells. However, it requires relatively large amounts of training data per tissue type and is therefore currently restricted to a relatively small amount of tissues. Furthermore, it is retrained for every microarray platform, making it necessary to have such abundant datasets available for each of the platforms. In contrast, our method can cope with a relatively small amount of training data per tissue and can be transferred to other microarray platforms. Therefore, it incorporates many more different tissues and can also distinguish between subgroups, e.g. between different regions of the brain.

A workflow for the combination of unsupervised and supervised dimension reduction techniques

The second contribution of the thesis is the development of a workflow (Fig. 1.2) for the combination of supervised and unsupervised dimension reduction techniques that combines the benefits of both individual approaches and that may be of advantage as a general workflow for other applications as well.

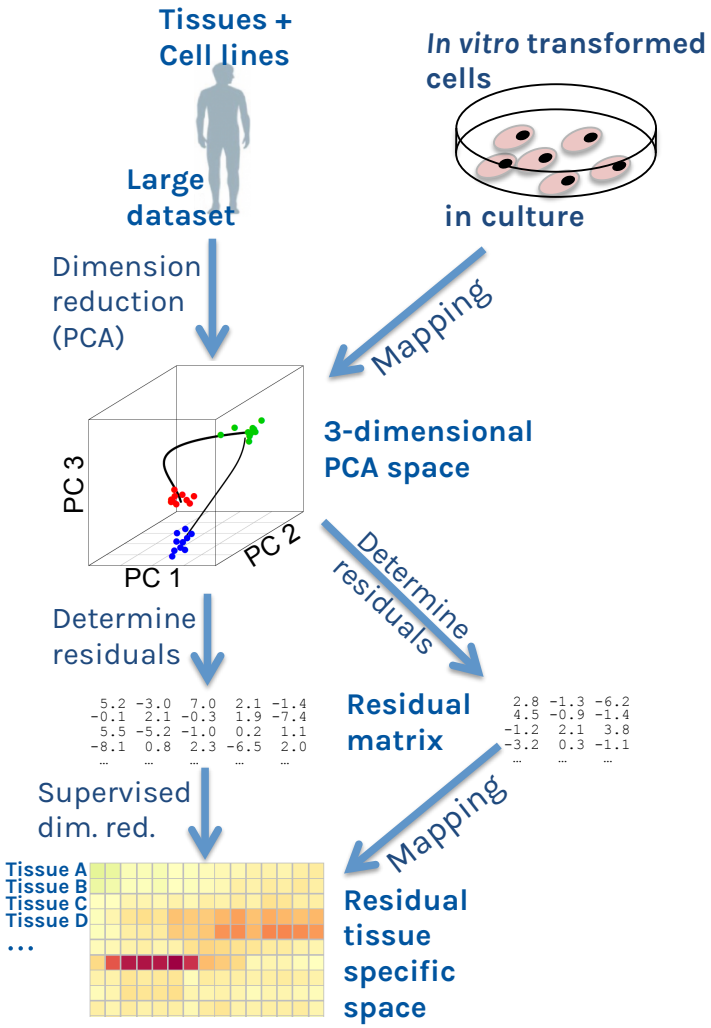


Figure 1.2.: Workflow of the two-scale map generation and mapping procedure. A large retrospective microarray dataset incorporating many different tissues and cell lines is first explored by an unsupervised dimension reduction technique, here principal components analysis (PCA). The first 3 principal components (PCs) build the 3-dimensional PCA space. Afterwards, the residual matrix of this dataset is calculated and used for the supervised dimension reduction using a fold-change based criterion on each annotated tissue or cell type. New data, e.g. from *in vitro* transformed cells, can be mapped to these two spaces by first mapping the original data to the PCA space, calculation of the residual matrix, and subsequent mapping to the residual tissue specific space.

So far, dimension reduction techniques applied to gene expression microarray data were either purely unsupervised [12], or purely supervised as is the case for the PhysioSpace method which we developed [72]. The unsupervised PCA based method of Lukk et al. [12] is not able to find all biologically relevant directions and does not provide biologically meaningful interpretation for most principal components (except of the first three).
Our own purely supervised method [72] is better in this respect, but suffers a bit from the shared processes, e.g. proliferation, that lead to high correlations between several different tissues.
The combined approach developed in this thesis applies unsupervised and supervised dimension reduction in a subsequent manner. The unsupervised PCA based dimension reduction is applied first, covering mainly shared processes. The supervised methodology is subsequently applied to the residuals of the PCA, capturing mainly tissue specific information. Thus, the challenges of the purely supervised method to cope with shared processes between different tissues were significantly reduced by the PCA based decomposition, making it possible to dissect general processes that are shared by several tissues and more tissue specific expression patterns.
Such a two-scale approach may be beneficial for other applications as well. One potential application is the analysis of disease progression that involves changes in mixture fractions of different cells as well as changes in the gene expression in single cell types. This combination of two effects could be observed for disease progression in leukaemia [82].

A complementary quality control criterium for *in vitro* transformed cells

The third contribution is more focused on the biological aspects of the thesis. We describe and discuss several existing methods for the quality control of *in vitro* transformed cells and show that our global gene expression based method can give complementary information to the commonly used methods. Specific molecular markers or functional assays alone provide often only relatively narrow information about a specific aspect of the cellular state, whereas our global comparison provides a broader view. Therefore, it is important to combine the complementary broad scale and the more detailed information in order to decide about the quality of the cells. We show examples where the success of the transformation process is seemingly validated by individual markers in the respective study, whereas our method hints towards an incomplete transformation of the cells.
Furthermore, due to the general challenge of defining what we mean by a cell type, and what we mean by good quality, we point out that it is important to interpret the various quality criteria in an application-specific way. In order to do this, it is very important to understand how observed differences to the desired cell type can be interpreted and whether they affect the intended application of the generated cells. The two-scale landscape can provide an interpretation of observed differences due to the use of reduced dimensional spaces with directional information that are linked to phenotypes. In contrast, other global gene expression based methods focus often on a pure comparison of similarity to specific

cell types or tissues without proper directional information of observed differences.

1.5. Content and structure of the thesis

Chapter 2 of the present thesis focuses on the classification of cells into different types. It starts with a general introduction of commonly used measurement techniques for the characterisation of cells. It then focuses on the characterisation of pluripotent stem cells, discussing and comparing the properties of different methods and pointing out some inconsistencies that open the question on how cell types can be defined in general. This question is then discussed from a more theoretical point of view and the concept of cell types as attractors of dynamical systems is described.
One way to use this concept in a practical setting is to localise cells in the gene expression space. Therefore, the next section of chapter 2 introduces gene expression microarrays, which can be used to measure the gene expression of cells on a transcriptome-wide basis. The specific focus of this section is on the quantitative relationship between the RNA content and the observed measurement signal, combined with some information about measurement noise. These information will be important for later parts of the thesis, especially for the comparison of linear and non-linear mapping methods in chapter 4.
The last part of chapter 2 introduces three existing methods for the RNA based characterisation of cells. These methods will be used in chapter 4 for a comparison to our newly developed method.
In chapter 3 we focus on the dimension reduction of the high dimensional gene expression space. To this end, existing approaches for unsupervised and supervised dimension reduction and especially their application to large gene expression datasets are described. The strengths and weaknesses of these methods are discussed, using an illustrative statistical model to clarify the described effects.
After that, the development of the two-scale gene expression map, combining unsupervised and supervised dimension reduction techniques is described. In order to evaluate the information content in the low dimensional PCA space and the residual complement, we use an information ratio criterion [83] that measures the amount of relevant information that is not captured by the first three principal components.
The mapping of new data onto the two-scale map is described in chapter 4. First, two different mapping procedures, a linear and a non-linear mapping, as well as the approach used for cross-platform analyses are described. These procedures are then used to map various tissues as well as artificial mixtures of different tissues onto the two-scale map for validation purposes. These data are used to compare the linear and non-linear mapping as well as to compare the newly developed two-scale map to existing approaches as well as to a direct supervised dimension reduction approach.
After the validation of the new method, it is applied to the analysis of various human *in vitro* transformed cells, namely iPSC derived mesenchymal stromal cells, cardiomyocytes, neurons, neural stem cells and astrocytes, as well as neural progenitor derived astrocytes

and neural stem cells that were directly converted from astrocytes.
In chapter 5 we discuss the general idea of using a multi-scale decomposition approach for the analysis of gene expression microarray data, i.e. to extend the presented two-scale approach to further scales. Furthermore, we discuss the complementarity of the quality information that can be received by the two-scale landscape in comparison to commonly used approaches and propose to combine these information in an application-specific manner to decide about the quality of the generated cells. Finally, we comment on the relevance of the presented workflow for other applications in the biomedical area.
Chapter 6 concludes the thesis with a summary of the main insights.
Finally, we present more detailed information on the analysed datasets and utilised software packages as well as some methodological details in the supplement along with additional figures and tables supporting the presented results.

2. Classifying cells into different types using high-dimensional gene expression data

The usual goal of *in vitro* reprogramming or differentiation experiments is the generation of cells representing a specific cell type, e.g. pluripotent stem cells, neurons or cardiomyocytes. An essential requirement in this context is the assessment of the adopted cell type, i.e. the proper characterisation of the cells.
In standard biological literature, cells are often assigned to a specific cell type based on functional assays, marker expression, morphological information or the cellular history [84]. These possibilities of cell type determination are briefly introduced and discussed in the first part of this chapter, focussing on the example of characterising pluripotent stem cells. Assessment of pluripotency is a rather well suited classification problem, due to its clear functional definition, i.e. the potential to differentiate into derivatives of all three embryonic germ layers, that can be directly tested by appropriate functional assays. These functional assays are, however, usually costly, time consuming or critical from an ethical point of view. Therefore, they are often replaced by surrogate markers that are less expensive and less critical.
Despite the rather well defined functional definition of pluripotency, researchers often distinguish between different types of pluripotent stem cells, e.g. depending on the origin of the cells. While all of these cells are pluripotent, it is not clear whether they differ in some other properties, e.g. their differentiation efficiency or their tumourigenic potential, making one cell type more appropriate for specific applications than another.
Thus, despite the joint characteristic of being pluripotent, the cells are not necessarily equivalent with regard to all their properties. This problem of testing cells for their phenotypical equivalence is even more complicated for *in vitro* differentiated cells compared to their *in vivo* counterparts. It motivates the topic of the second part of this chapter, discussing the question on how a cell type can be defined in a consistent and rigorous way [85, 86]. This rather theoretical question is not easy to answer. In contrast to e.g. chemical elements, which can be classified according to their number of protons, cells are living organisms and there does not seem to exist any essential property that can be used to classify them in a similar way as the chemical elements [86]. Therefore, the cell type as attractor concept was introduced [67, 87], which is based on modelling the cell as a dynamical system.

Classifying cells based on this concept requires high-throughput measurements such as gene expression microarrays in order to locate the cells in the high dimensional expression space. Therefore, we describe the microarray technology in the third part of this chapter. The focus of this description is on the quantitative relationship between mRNA content and measurement signal, as well as on the measurement noise. These quantitative properties are important to consider when models for mRNA-microarray based characterisation of cells are developed and evaluated.
Three existing models for mRNA based characterisation of cells are introduced in the last part of this chapter. These models use different methods to reach this aim and have also quite different characteristics. Two of them are not well suited for the analysis of cells that fit only partially to an existing cell type, while the third one is limited to a relatively small amount of different tissues since it requires relatively large amounts of data from each cell type for the training of the classifiers.

2.1. Assays for the characterisation of *in vitro* transformed cells

The discovery of different cell types evolved stepwise during the history of biological research. More and more subtle differences were detected, leading to the definition of various subtypes, such as the various types of neural cells, ending up with some hundreds of different cell types in the human body that are known to date [15].
Most of these cell types were defined based on specific properties of the cells in the human body, i.e. *in vivo*. For example, cardiomyocytes are myocytes that are located in the heart and motor-, sensory-, and interneurons are distinguished based on their connection to muscles, sensory organs, or other neurons, respectively. When put into culture these cells were usually still considered to be of the same type as before. While this strategy may be useful in case of a pure culturing of cells without the intention of changing the cell type, the cellular origin cannot be used to classify *in vitro* transformed cells anymore. These cells must therefore be characterised based on molecular, functional, or morphological characteristics rather than their *in vivo* origin.

2.1.1. Overview of commonly applied assay types

There is a large variety of different measurement techniques available to characterise cells. Three main classes of techniques are briefly introduced here, having different strengths and weaknesses and complementing each other in terms of spatial resolution, molecular information content, and direct assessment of phenotypic information.

Imaging and immunohistochemistry

The cellular morphology and expression of specific markers that can be stained using antibodies (immunohistochemical analysis) are frequently used in histological analyses of

tissue biopsies. Due to their spatial resolution, these methods are very useful for the analysis of tissues that are mixtures of different cell types. These imaging based methods can also be used for the analysis of cultured cells and *in vitro* transformed cells. They are especially useful for a fast and cheap initial analysis of the cells and are well suited for automation purposes since they only need liquid handling and microscopy facilities.
One drawback of imaging based methods is the subjectivity due to the usually applied manual evaluation without proper quantification. This challenge is, however, currently tackled by image processing approaches trying to create more objective criteria for image based cell classification [88].
More fundamental disadvantages of these methods concern the information content that is restricted to morphological information, being hard to directly associate with specific functionalities or cellular behaviours, or to the expression of individual genes or proteins, neglecting information about the other molecules that are necessary for a systems-wide characterisation of the cells.

High-throughput omics measurements

The latter disadvantage is resolved by the use of high-throughput measurement techniques, which determine the amount of e.g. RNA or proteins on a systems wide level, i.e. they measure the amounts of ten thousands different genes, proteins, or other molecules simultaneously. Furthermore, these measurements are quantitative and usually comparably well standardised. The quality of different measurement techniques varies, but it is often considerably higher than initially anticipated [89] and initial artefacts and quality issues are resolved by specific preprocessing and normalisation methods.
One disadvantage of high-throughput measurements are the relatively high costs, compared to imaging or staining techniques. However the costs are rapidly decreasing due to technical improvements, especially in sequencing methodologies [90].
A further disadvantage is the missing spatial resolution, as well as the measurement of a pool of thousands of cells, which may consist of an unknown mixture of different cell types. The latter issue is currently tackled by single cell sequencing technologies [91], which will probably provide new interesting insights in future studies. Alternatively, deconvolution approaches can be used to computationally dissect the measured signals into parts corresponding to specific cell types [92, 93].

Functional assays

Despite the high diversity of functional assays, making it hard to generalise the advantages and disadvantages of these measurement techniques, some common features can be stated. Functional assays can provide phenotypical information that, in some cases, directly test the defining properties of specific cell types, e.g. the pluripotent differentiation potential of pluripotent stem cells. This is a very valuable information that can be used as gold standard for the assessment of the corresponding cell type.

Common disadvantages of functional assays compared to the omics and imaging techniques are the relatively narrow information content, which is focused on a single functionality, and often a missing spatial or single cell resolution. Furthermore, some functional assays are very cost and time intensive as exemplified by the pluripotency assays introduced in the next section.

2.1.2. Pluripotency assays

Many different assays have been developed to test cells for their pluripotent differentiation potential, including imaging, molecular and high-throughput based as well as functional assays. Due to the functional definition of pluripotency, i.e. the ability of a cell to differentiate into derivatives of all three embryonic germ layers, functional assays are a natural choice for this task. However, due to ethical limitations, high costs, and long durations of some functional assays, surrogate markers have been developed that have beneficial properties in these regards (Figure 2.1). The most prominent assays are introduced and discussed in the following.

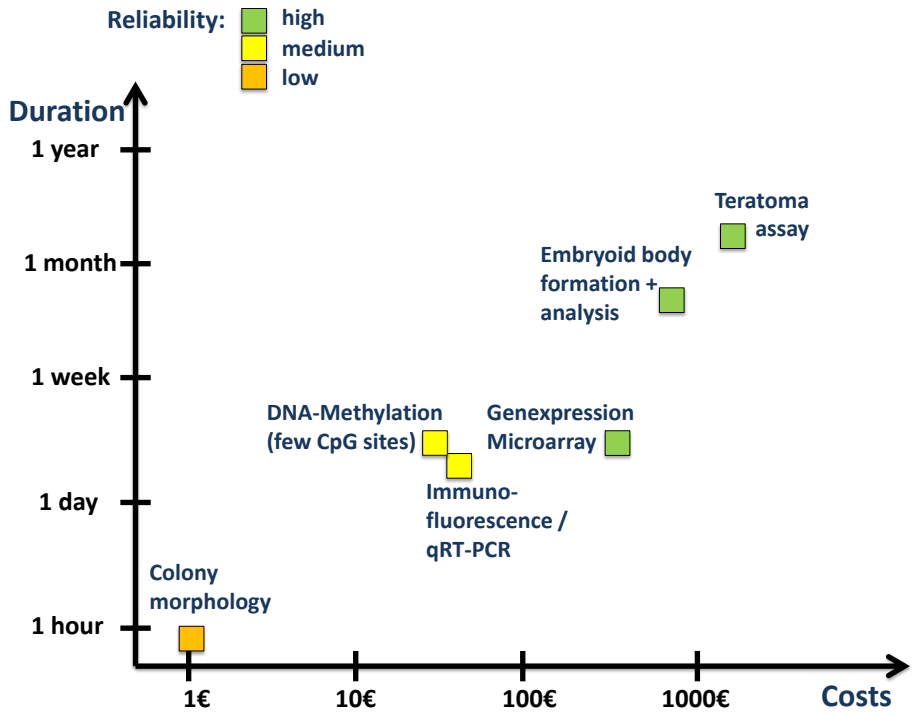


Figure 2.1.: The choice of specific iPSC characterisation assays is a trade-off between the cost and duration on one hand and the reliability on the other hand. Several assays are evaluated here based on approximate values for their costs and durations [94, 95, 96], as well as a rough rating of their reliability [70, 96, 97]. Exact values of costs and durations depend on the service provider and the concrete realisation of the assays. The scaling and the tick marks on the ordinate are therefore rather indicative than mathematically meaningful.

Tetraploid complementation and chimera assays

The tetraploid complementation assay as well as the chimera assay are regarded as gold standards for pluripotency assessment in mice. In these tests, pluripotent stem cells are injected into the blastocyst of either a tetraploid embryo, for the tetraploid complementation assay, or a genetically different embryo with normal chromosome number, for the chimera assay [98, 99]. In the first case, the tetraploid cells do not contribute to formation of the mouse fetus, but form the extra-embryonic tissue. Therefore, the whole developing mouse, if it develops, originates from the injected pluripotent stem cells. This is only possible if the injected cells were truly pluripotent, and is therefore a very stringent pluripotency assay.
The chimera assay tests whether the developing mouse consists of a mixture of the two genetically different cells, i.e. the cells from the initial embryo and the injected pluripontent stem cells. Thus, the injected cells do not have to form the whole mouse, as in the tetraploid complementation assay, but must still contribute to the formation of tissues from all three embryonic germ layers to fulfil the criteria of this pluripotency assay.
These two stringent pluripotency assays can not be used to test the potency of human cells, due to ethical reasons. Therefore, another functional assay, the teratoma assay, is currently considered as gold standard for pluripotency assessment in human cells.

Teratoma assay

In the teratoma assay, pluripotent stem cells are injected into immunodeficient mice, so called SCID (severe combined immunodeficiency) mice [96, 6]. After several weeks of *in vivo* growth (4 to 15 weeks, depending on the protocol in use [100]) it is checked whether a benign tumour, consisting of tissues from all three embryonic germ layers, has formed, validating the pluripotency of the injected cells. Thus, the teratoma assay tests, whether the injected cells can form tissues of all three germ layers in an *in vivo* environment, but it does not check whether the cells can contribute to the development of a full living animal. Criticism about the teratoma assay includes animal welfare as well as ethical concerns [6]. Furthermore, the teratoma assay is relatively cost and time intensive (Figure 2.1) and lacks a proper standardisation [6, 100], although first attempts to standardise the assay have recently been published [101].
Taking these criticisms and the available alternatives (see below) into account, it is not recommended to use the teratoma assay as a standard for a pure testing of pluripotency anymore. However, one may use the teratoma assay for other purposes, especially for studying the tumourigenic potential of pluripotent cells [6].

Embryoid body formation

An alternative functional assay for pluripotency evaluation is the embryoid body (EB) assay, which is based on an *in vitro* stochastic differentiation into cells of all three embryonic germ layers [102]. This assay has many advantages compared to the teratoma

assay, since it does not use animals, it is cheaper, usually faster and easier to standardise [102]. Furthermore, EB formation can be used to extract dynamical information of early differentiation events, which are not accessible from teratomas since the latter are only evaluated at a single endpoint [6].
The main disadvantage of the EB assay is the *in vitro* culture environment, which may be inappropriate for tests of the tumourigenic potential of the studied cells. There may be many different stimuli or interactions that promote cells to develop a malignant phenotype and the relevant conditions, whatever they are, may be better represented in an *in vivo* mouse model than *in vitro*. However, it is also not clear how good mouse models represent the relevant conditions in a human and how reliably results can be transferred between both species [103].

Molecular markers

Molecular markers can be used as surrogates to the functional assays, and do not directly test the definition, i.e. the differentiation capacity, of pluripotent stem cells. Therefore, they need to be validated rigorously with relatively large amounts of cells.
Many different molecular markers have been proposed to characterise pluripotent stem cells, including the surface markers TRA-1-60 and TRA-1-81 as well as SSEA3 and SSEA4 [104]. Furthermore, several nuclear markers are commonly used to test cells for their pluripotency, including OCT4 (the corresponding gene is called POU5F1), NANOG, SOX2, and several others [104]. Many of these markers are, however, not entirely specific to ESCs and iPSCs, but are also expressed by some cancer stem cells [104]. Apart from that, systematic evaluations of several pluripotency markers during early differentiating events indicate differences in the expression dynamics [97].
In an other study, fully reprogrammed cells (SSEA4, TRA1-60, and NANOG positive) were distinguished from partially reprogrammed cells (SSEA4 positive, TRA1-60 negative, low expression of NANOG) based on typical pluripotency markers, but established lineages of both cell types were able to form teratomas [105]. Furthermore, in the same study, TRA1-60 positive and SSEA4 negative cells occurred during reprogramming. Unfortunately, these cells could not be established as a cell line in order to check for teratoma formation.
Interestingly, the results of this study are sometimes used to question the gold standard of the teratoma assay [102, 106]. This is a reasonable debate, since it seems to be clear that the established partially reprogrammed cells from [105] differ from well known ESCs in the expression of important pluripotency markers. On the other hand, the established partially reprogrammed cell line seemed to be able to form tissues derived from all three embryonic germ layers. Thus, the argumentation on one or the other side depends on how we define pluripotent stem cells. They can be either defined based on the functional property, having a pluripotent differentiation capacity, or based on the similarity to the pluripotent cells in the inner cell mass of the blastocyst.
Summarising the aforementioned results on molecular pluripotency markers, it can be

stated that most of these markers can nicely distinguish between pluripotent stem cells and somatic cells (Figure 2.2, right graphic), while there are some conflicting results with respect to partially reprogrammed cells and some cancer cells. Therefore, it is still under discussion which marker, or which combination of markers is most accurate for the characterisation of pluripotent stem cells.

Gene expression microarrays - PluriTest

The so called 'PluriTest' is a bioinformatics assay that uses two whole genome based gene expression scores, termed pluripotency score and novelty score, to distinguish pluripotent stem cells from other cells [70, 107]. PluriTest was trained on a large set of well characterised gene expression microarray data from the Illumina Human HT12 v3 platform involving 264 pluripotent stem cells (223 ESCs and 41 iPSCs) and 204 somatic cells or tissues [70, 108]. It uses non-negative matrix factorisation (NMF) [109, 110] for the identification of pluripotency modules that are the basis of the pluripotency score. The novelty score is determined as the distance of a sample to a NMF-based convex hull approximation of the typical pluripotent stem cell location in the high-dimensional gene expression space. Thus, a high pluripotency score can be regarded as being indicative for the pluripotent differentiation potential of the analysed cells, while a low novelty score rather indicates high transcriptomic similarity to the ESCs and iPSCs from the training data.
Therefore, the novelty score is especially useful to detect pluripotent stem cells with abnormal or novel (i.e. not present in the training data) transcription patterns. It could be shown that it is capable of distinguishing germ cell tumors and parthenogenic pluripotent stem cells from ESCs and iPSCs [70]. This can usually not be accomplished with the single gene molecular markers that were introduced before (Figure 2.2).
Initially, PluriTest was only developed for one specific microarray platform, the Illumina Human HT12 v3 platform, for which it is available as an online tool (`http://www.pluritest.org/`). However, it could be shown that PluriTest can also be used with other arrays and to some extend also for mouse cells by ortholog probe matching, when an appropriate reference dataset is available [70, 111, 112]. The reference dataset is needed to define new cutoff values for the pluripotency and novelty score, since the absolute values of these scores change for different platforms.
Interestingly, while it can be expected that the performance for other platforms is slightly worse, the PluriTest assay seems still to be capable of distinguishing embryonic carcinoma (EC) cells from iPSCs and ESCs when it is projected to the Affymetrix Human U133 Plus 2.0 microarray platform (Figure 2.2, left graphic), maintaining its beneficial properties compared to standard single gene pluripotency markers (Figure 2.2, right graphic). Furthermore, karyotypic abnormal ESCs from the same platform show an increased novelty score compared to karyotypic normal ESCs.
In summary, PluriTest is a reliable alternative to functional pluripotency assays that is well standardised, comparably cheap and fast (Figure 2.1), and avoids any kind of animal

experiments.

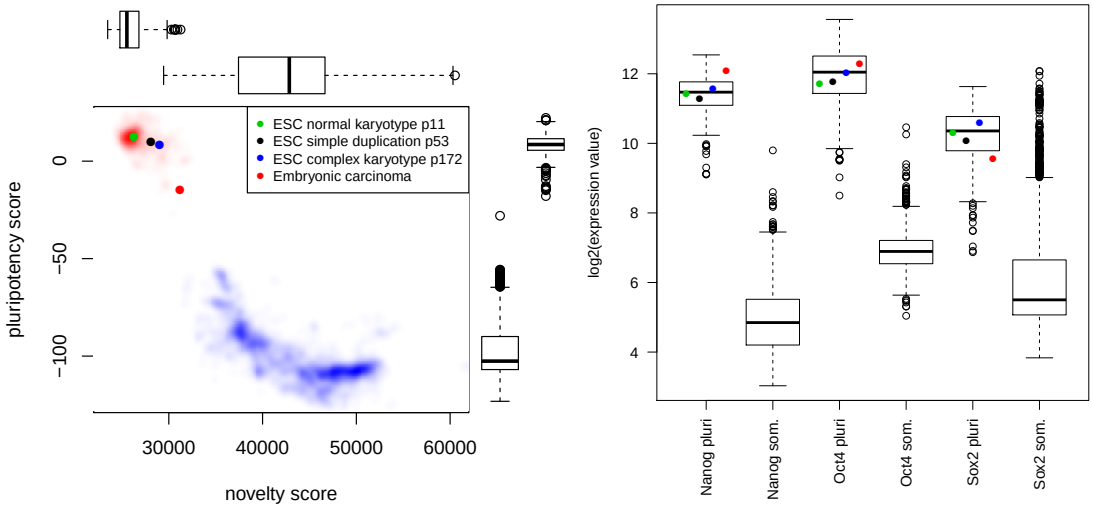


Figure 2.2.: The PluriTest assay (projection to the Affymetrix Human U133 Plus 2.0 array) can distinguish between ESC and embryonic carcinoma (EC) cells and shows an increased novelty score for karyotypic abnormal ESCs (left), whereas typical pluripotency markers show no clear trend (right). Compared are ESCs with normal karyotype at passage 11 (green filled dots), ESCs with a simple karyotypic duplication at passage 53 (black filled dots), ESCs with complex karyotypic changes at passage 172 (blue filled dots), and EC cells (red filled dots) retrieved from the Gene Expression Omnibus database (accession number GSE7234). The red and blue background colours represent a reference dataset of 368 pluripotent (red) and 1607 somatic (blue) samples smoothed with a kernel density estimator. Box plots on the top and right side of the left figure show the corresponding marginal distributions of the novelty score and pluripotency score. The same colour codes and reference data are used for the right graphic, showing the log2-expression value of three pluripotency markers (Nanog, Oct4, and Sox2). The relative large overlap of pluripotent and somatic samples for Sox2 is mainly due to its high expression in neural cells.

DNA-Methylation

Another way to assess the pluripotency of cells is through the analysis of epigenetic DNA modifications. Especially the DNA methylation status of pluripotent stem cells has been compared to that of several other cell types in various studies [113, 114, 115]. These studies, however, did not propose a certain marker for evaluation of pluripotency. This has recently be done by us [116], by a detection of only three CpG sites that are well suited to classify cells into pluripotent and non-pluripotent. This methylation based evaluation does not provide the possibility to dissect abnormal pluripotent cells, e.g. EC cells, from normal pluripotent stem cells, as can be done using PluriTest. However, it represents a cheap and fast alternative for pluripotency evaluation that is well standardised in contrast to most

immunofluorescence or qRT-PCR measurements of single marker genes or proteins.

Morphology

Pluripotent stem cells are usually round and small and tend to grow in densely packed colonies with smooth borders. These morphological characteristics distinguish them from many other cells, including fibroblasts, which are often used as a source for reprogramming experiments. Therefore, a cheap and fast possibility to get a first impression of the reprogramming success is phase contrast microscopy (Figure 2.1). However, so far, the morphological characterisation is rather subjective and neither quantitative nor standardised, and there are only few first attempts to develop quantitative morphological markers [88]. Furthermore, some partially reprogrammed cells were qualitatively described to have morphological characteristics that are similar to fully reprogrammed cells. Therefore, it remains to be evaluated by systematic comparisons how reliably pluripotent stem cells can be characterised based on morphological criteria. An important first step in this direction would certainly be a storage of microscopy images in public databases similar to the established databases for omics measurements and an evaluation of the preprocessing and standardisation needs to make images from different microscopes comparable.

2.1.3. Conflicting evidence - comparing EC, ES, and iPS cells

For the characterisation of *in vitro* transformed cells, it is important to check whether the generated cells represent the corresponding *in vivo* cell type in a sufficient manner. In order to achieve this goal, it would be very helpful to have (*in vitro*) testable definitions of cell types that are fully based on the current molecular or phenotypical state of the cells rather than their history of derivation. This is the case for the definition of pluripotent stem cells and several assays for pluripotency evaluation were introduced in the previous section that test the defining property of pluripotent stem cells in a direct or indirect manner. Nevertheless, some differences between some pluripotent stem cells were observed, e.g. in the expression of certain markers. Furthermore, some pluripotent stem cells are commonly distinguished based on their origin. ESCs are derived from the preimplantation or periimplantation embryo [1], EC cells can be isolated from teratocarcinomas [43], and iPSCs can be reprogrammed from various somatic cells by ectopic expression of usually four transcription factors [3].
Despite the different origins of these cells, they all express typical pluripotency markers and it could even be shown for all three of them that they are able to form chimeras or whole mice by the tetraploid complementation assay [98, 117, 118, 119]. Thus, they fulfil stringent pluripotency criteria and can thus be considered as pluripotent stem cells.
Taking this result alone, one may argue that EC, ES, and iPS cells should be considered as equivalent cell types. However, EC cells are derived from malignant and life threatening germ cell tumors, while ESCs are derived from the embryo and are thus tightly connected to the formation of new life without a malignant phenotype. Furthermore, EC cells usu-

ally have chromosomal abnormalities and tend to form malignant teratocarcinomas, while ESCs usually have a normal karyotype and tend to form benign teratomas when injected into immunodeficient mice [43, 6, 117, 120, 121]. Yet, there are also reports of ESCs or ESC derived progenitor cells that form malignant teratocarcinomas [122, 123].
Thus, a pure distinction based on the original location *in vivo* is problematic. This seems to be particularly relevant when cells are maintained *in vitro* for a long time, since prolonged culturing of ESCs can introduce genomic modifications that are typical for EC cells, leading to teratocarcinoma formation in immunodeficient mice [43]. However, genomic modifications can occur frequently and it is not clear which modifications are responsible for a malignant transformation, and whether the transformation can be explained by genomic changes alone [123, 124].
The tumourigenic potential is also a major concern for iPSCs, since they can have additional genomic abnormalities due to mutations in the somatic cells that are used as source for the reprogramming or due to genomic integrations during the reprogramming process [124]. Furthermore, single mutated cells, e.g. with a TP53 mutation, may have a selective advantage during the reprogramming process [125].
In addition, there are many discussions about the similarity or dissimilarity of ESCs and iPSCs [126]. Newly derived iPSCs with low passage numbers seem to have some epigenetic memory of the cells that were used as a source for the reprogramming, while this memory diminishes during prolonged culturing [111, 127]. Furthermore, it is speculated that there are different steps of reprogramming which are only partially passed by some iPSCs [126], leading to some variation between different iPS cell lines. However, not only iPSCs, but also ESCs show considerable variation in their gene expression, epigenetic status as well as their differentiation propensity [128]. Therefore, one might even ask the question whether different ES cell lines can be regarded as equivalent.
In summary, conflicting evidence of EC, ES, and iPS cell similarity together with considerable variations between different EC, ES, or iPS cell lines pose the question on how a cell type can defined in a rigorous way, such that all cells of a specific type have well defined common properties. This question will be the topic of the next section, showing that there are many complications that challenge current approaches for defining cell types in an absolute way.

2.2. The challenge of defining cell types in a rigorous way

Pluripotent stem cells are relatively well defined via their pluripotent differentiation potential and the ability for prolonged undifferentiated proliferation. Nevertheless, the differentiation efficiency towards various tissues as well as several other properties differ between different pluripotent stem cells, as indicated in the previous section. Similar problems apply also for various somatic cell types. Furthermore, many cell types lack a clear testable definition as it is present for the pluripotent stem cells. Finally, assessing the cell type of cultured cells is challenged through observations that cultured cells and especially *in vitro*

transformed cells have some significantly different properties than their *in vivo* counterparts [12, 5]. Therefore, it is recognised [86] that the question on how to define a cell type should be tackled in a more systematic way.

2.2.1. The essentialism approach

This question has been recently addressed by Slater [86] who prefers the term "cellular kind" instead of "cell type". He indicates that there seems to be no essential property that can be used to define cell types. An essential property in this context is an "intrinsic property" that is "possessed by all and only the members of a kind", and that explains "why members of the kind have a series of superficial properties more or less in common" [86]. A typical example of such an essential property, in a different context, is the number of protons defining a chemical element. This is an intrinsic property fulfilling the above criteria, which naturally classifies atoms into a discrete set of different types, while it still allows differences between two instances of the same kind, e.g. in the number of neutrons of ^{12}C and ^{13}C atoms. Nevertheless, atoms of one kind have several properties in common and this can be explained through an intrinsic property, the number of protons, that is very stable with respect to environmental changes of the atoms.
In the case of cell types, there seems to be no such essential property. While in earlier times, it was thought that cells selectively loose specific genes during differentiation, suggesting that the DNA content may be such an essential property [86, 129], we now know that this does not happen. In fact, for most cell types of an individual, except e.g. mature red blood cells that do not have a nucleus, the DNA content seems to be the same. Furthermore, the method of somatic cell nuclear transfer, i.e. the fusion of an enucleated egg cell with a nucleus from a somatic cell, indicates that no nuclear properties can be regarded as essential properties, since the fused cells eventually become pluripotent, rather than becoming the cell type of the somatic cell from which the nucleus was taken [130, 131]. Instead, the distinction into different cell types may be due to differences in the gene expression, i.e. in the mRNA and protein content of the cell.
However, using the mRNA or protein content as essential property is problematic as well. This is due to the fact that cells live, and are therefore not at thermodynamic equilibrium. The mRNA and proteins are constantly degraded and newly expressed, making the cellular state dependent on a constant supply from the environment. Furthermore, dynamical changes such as the cell cycle take place, leading to a dynamic alteration of the mRNA and protein content, making these properties inappropriate for the essentialism approach (see [86] for more details).
So, rather than searching for an essential property to define cell types, it may be more appropriate to account for the dynamic changes and the non-equilibrium state in which cells live. This is possible through an interpretation of a cell as a dynamical system, with cell types being (possibly complex) attractor states [67, 87, 132, 133], which is explained in the following.

2.2.2. Cell types as attractor states - a dynamical systems view

In the dynamical systems view, the molecular state of a cell is described by a high dimensional state vector, corresponding e.g. to the mRNA content of the cell, where a single component of the vector corresponds to a single gene. The space spanned by all possible configurations of this vector is called the state space and a cell is represented by a point in that state space. Thus, in the case of gene expression data, the state space is approximately 20,000 dimensional, where each dimension specifies the expression status of one gene.

The possible states that a cell can adopt are restricted due to interactions between different molecular components of a cell as represented for instance by gene regulatory networks, imposing certain constraints on the set of possible states. These regulatory networks also define possible dynamic changes of the molecular state. These dynamic changes can, in principle, be described by a dynamical model, e.g. a set of differential equations. According to these equations, the cell moves in a coordinated way through the state space, eventually approaching an attractor state, where it resides. Such an attractor state can be either a fixed point or, accounting for the cell cycle and other dynamic changes, a dynamic attractor such as a limit cycle or strange attractor [133, 134].

Many dynamical systems have multiple attractors that are approached depending on the initial position of the cell in the state space. The set of all initial positions from which a cell will approach a certain attractor is called the 'basin of attraction'. Intuitively, this can be visualised as an energy-like landscape (sometimes called attractor landscape) with stationary attractors as local minima [133, 135]. However, this visualisation may be a bit misleading since the cell lives at a non-equilibrium state and minima in an energy-landscape usually correspond to equilibrium states.

In the cell type as attractor concept, each attractor in the high dimensional state space corresponds to a cell type, meaning that a cell that lies in the basin of attraction of one specific attractor is considered as being of the type that corresponds to this attractor. This concept leads naturally to a distinction into a certain number of cell types, where all cells of that type have certain common properties as long as they are not too much disturbed, i.e. pushed away from the attractor. These properties may change over time, in case of a dynamic attractor, but will recur after a certain time.

Defining cell types as attractors of a dynamical system is an interesting concept from a theoretical point of view and there are also some confirmations of the theory from simulations as well as wet lab experiments. Indeed, experimental results show that cells that are perturbed by different stimuli can eventually approach the same molecular state, indicating that this molecular state may be considered as an attractor state [87]. Another example is the reprogramming into induced pluripotent stem cells, which can be performed with different sets of induced transcription factors [136], leading (more or less) to the same final state. This final state can therefore also be considered as an attractor state. Furthermore, simulations of large random transcriptional networks with boolean approaches

indicate that the number of cell types in humans corresponds well to the number of observed attractor states in the simulation [67], assuming constant environmental conditions and model parameters. This makes the cell type as attractor concept plausible.
Thus, there is some experimental evidence for the cell type as attractor concept and the theoretical implications of this concept are far reaching, including changes in the understanding of cancer initiation and progression [134, 137]. However, the concrete implementation of this concept in terms of a system of differential equations is currently not possible due to the restricted knowledge of the complicated biological mechanisms. Thus, it is either possible to restrict the dynamic analysis to a reasonable subsystem [138, 139], or to concentrate on specific aspects of the dynamical systems concept.
Therefore, the focus of current research is more on the qualitative description of the attractor landscape. This is done in such a way that the intuitive understanding of the landscape matches to the observed experimental phenomena. One example is the usually very low efficacy of the reprogramming to iPSC, which suggests some stochasticity in the reprogramming process [140]. This rare event of reprogramming may be explained by high energy barriers in the attractor landscape [141]. Interestingly, it has been reported in a recent study that this high barrier can be significantly reduced, leading almost to 100% efficacy of the reprogramming process [142].
Another qualitative behaviour of cells that needs to be explained by a valid theory is the pluripotency or multipotency of stem cells as opposed to unipotent cells. That means that the qualitative difference, in terms of the attractor landscape, of the different types of cells having different potentials to differentiate spontaneously into other cell types has to be explained. One possible answer would be that stem cells do not lie in a basin of the attractor landscape, but rather on a saddle point, i.e. a non-stable fix point. However, this can not explain the potential to maintain pluripotent stem cells for long durations in culture. Therefore, it is more likely that there is a small local basin on this saddle point, similar to a mountain lake on the top of a mountain [141].
Another experimental observation that challenges the qualitative understanding of the attractor landscapes is the following interesting behaviour of cell populations that have a certain heterogeneity with respect to some markers. It was reported that in cultures of mouse haematopoietic progenitor cells, there are some spontaneous outlier cells with an extremely high or an extremely low amount of a certain marker protein [143]. Interestingly, when these outlier cells were sorted and cultured separately, the complete evolving population reconstituted the original population of the system before sorting. This would indicate that the cells lie in an attractor and are stochastically pushed to some border from which they return eventually. However, what was a bit surprising is the long time, i.e. more than one week, these cells needed to relax to the original population [143]. Therefore, this observation may indicate that there are local sub-attractors within a larger basin of attraction that lead to a slower recurrence of the disturbed cells[141].
These qualitative refinements of the attractor landscape will probably continue in future research, incorporating e.g. descriptions about the effects of changes in the environment or

mutations in the DNA [134]. For the aim of this thesis, mainly the concept of considering a cell as a point in the high-dimensional gene expression space and a cell type as a specific region in this space will be utilised, trying to characterise cells based on their location in this space. For this purpose, we need some measurements of whole genome gene expression, e.g. gene expression microarrays, which will be introduced in the following.

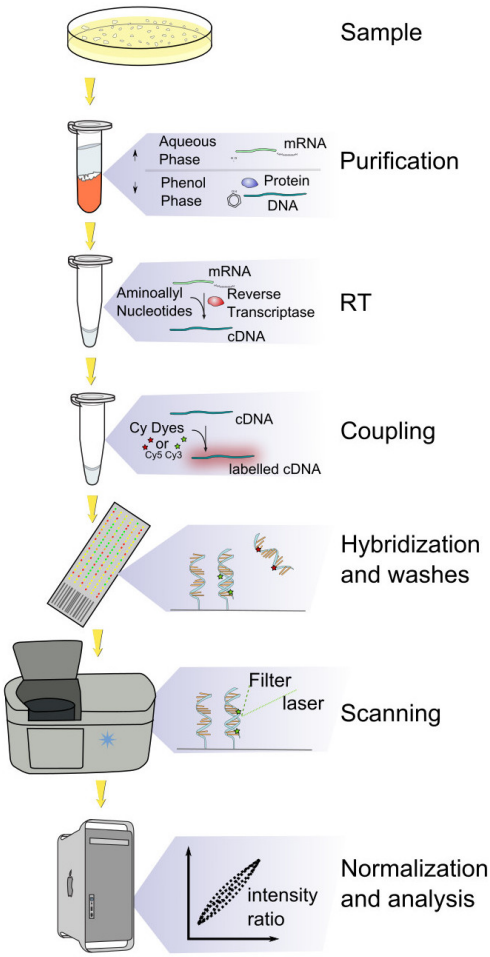
2.3. Gene expression microarray data

Gene expression microarrays can measure the expression of thousands of genes simultaneously. Roughly 20 years ago [144], this initiated a revolution in molecular biology, allowing to study not only a handful of genes but to look at systems-wide gene expression. Since then, the number of publicly available microarray data experienced a rapid increase [53] and the knowledge about the reliability of microarray data [89], common causes and distributions of measurement errors [145], differences between microarray platforms [146], as well as appropriate preprocessing strategies [147] has improved. This knowledge is important for the development of advanced analysis tools that fit well to the properties of gene expression data. Therefore, a summary of the most relevant characteristics of the microarray technology is given in the following.

2.3.1. General principles and platform differences

Generally, two types of microarrays are distinguished. Two-colour microarrays measure the RNA content of two samples simultaneously, focusing on the differential expression between these samples, while one-colour microarrays measure the expression of a single sample, providing absolute intensity values [148]. This difference has some impact on the preprocessing and the statistical design of the experiment. Especially, the integration of heterogeneous data from different studies and various cell types is easier to manage with one-colour microarrays, since two-colour microarrays may be hybridised with different references across studies. This makes it easier to work with one-colour microarrays in meta-analyses. Furthermore, it could be shown that the performance of one-colour microarrays for assessment of differential expression is similar to that of two-colour arrays [149, 150]. Therefore, we focus on the analysis of the more commonly used single-colour microarrays in this thesis.

The main steps of a microarray experiment, from RNA extraction to data preprocessing and normalisation are summarised in Figure 2.3 [151]. These steps can be roughly divided into three phases. The first phase concerns the sample preparation. It consists of RNA extraction, purification, reverse transcription, amplification, and labelling. The second phase focuses on the measurement itself, consisting of the hybridisation, washing, and scanning of the array. The third phase finally deals with data preprocessing and normalisation. In all of these steps, differences in the protocol, measurement or processing parameters as



http://en.wikipedia.org/wiki/DNA_microarray_experiment

accessed at May 20th, 2014

Figure 2.3.: Overview of core steps in a microarray experiment. RNA is extracted from roughly 10^5 to 10^6 cells, purified, reverse transcribed and labeled with a fluorescent dye. Additionally, there is often an amplification step, which is not depicted in the figure. The labeled cDNA is then put onto the array, where it hybridises to complementary oligonucleotide strands and is thus separated into the different genes that are located at different spots on the array. Loosely binding cDNA is washed off from the array to minimise cross-hybridisation of non-specific cDNA. The amount of RNA that is hybridised to each spot is then indirectly measured via fluorescence imaging and subsequent image processing of the excited fluorescent dye. Subsequent preprocessing and normalisation steps are finally used to improve the comparability between arrays and to remove measurement artefacts. For two-colour microarrays, a second RNA sample from reference cells is simultaneously prepared, labeled with a different dye and hybridised together with the first sample.

well as in manual processing can introduce technical noise to the data. This noise can be tremendously reduced by a high standardisation of chip design, hybridisation, and data processing protocols [152]. Such highly standardised arrays from commercial companies, e.g. Affymetrix, Agilent, or Illumina, have only very low between-array variability [89] and can therefore be reliably used for biological inference.

However, while arrays from one specific microarray platform can be directly compared, arrays from different platforms are less comparable, sometimes hindering the direct comparison of data from two studies of interest. This is due to substantial differences in absolute expression values between platforms that result from differences in the array design, as well as differences in hybridisation protocols and scanners, that overall clearly exceed common batch effects [152]. Parts of the platform dependent differences can be explained by differences in the lengths and exact sequences of the probes that are attached to the array. Chips from Affymetrix usually have attached oligonucleotides of 25 bases length, whereas Illumina probes are 50 bases long and Agilent probes have a length of 60

bases [146]. Furthermore, arrays from the same company sometimes differ in their probe-sequences, which may match to different positions within a gene, e.g. close to the 3' end or equally distributed over the entire gene [153]. Different nucleotide sequences lead to differences in the binding affinity of probe and target and therefore to different intensity values, complicating direct comparisons between different arrays.
Overall, there are numerous possible influences on the intensity measurement of a microarray spot and it is important to study these effects in detail to facilitate a better understanding and modelling of the data.

2.3.2. From RNA content to measurement signal

The goal of gene expression microarray studies is to measure the mRNA content of cells for each gene. However, as described in the previous section, the mRNA content is not measured directly. Instead, a fluorescence intensity is measured for each spot on the array that is assumed to depend, apart from measurement noise, monotonically on the mRNA content corresponding to the respective gene. The exact relationship of the measured signal and the mRNA content is still under investigation [154], but some crucial mechanisms have already been modelled and some approximations can be made that are very useful for practical analyses [145, 154, 155, 156].

Notation

Let $\underline{y}$ be the vector of measured intensity signals from the microarray and let $\underline{x}$ be the vector of mRNA concentrations in the target, where each component represents a protein-coding gene. It is then of interest to determine the function F that relates both vectors, possibly incorporating an error term $\underline{\epsilon}$

$$\underline{y} = F(\underline{x}, \underline{\epsilon}). \tag{2.1}$$

Here, it is assumed for simplicity of the following descriptions that the dimensions of $\underline{x}$ and $\underline{y}$ are equal with a one to one correspondence of the components in $\underline{x}$ and $\underline{y}$. Thus, it is assumed that probesets were already summarised and that the summarisation only reduces noise or probe-specific variability, i.e. it only affects $\underline{\epsilon}$ and the probe-specific parameters of F but not the form of F per se. This is certainly not entirely true, but the detailed description of the various summarisation methods in combination with platform specific differences in probeset compositions goes beyond the scope of this thesis.
For a further simplification of the notation, we focus in the following on a single component y_i of $\underline{y}$, implying that the form of F is the same for every component, but the parameters can be probeset-specific. For further notational simplicity, we skip the index i and distinguish a single component y from the vector $\underline{y}$ by the missing underline. Also, probeset-specific parameters are not explicitly indexed with i. The vector $\underline{x}$ is simplified to a two-component vector $(x_S, x_N)^T$, where x_S is the amount of specific target, i.e. of the gene that specifically

binds to the investigated probeset, and x_N summarises the non-specific targets into a single component (it is straightforward to extend the description to the full vector $\underline{x}$).

Langmuir model based on physicochemical parameters

We start with neglecting the error terms and describing the relationship of RNA content and measurement signal in terms of physicochemical parameters, inspired from [157]. First, the measurement signal y can be described by an affine relationship to the fraction of bound probe θ:

$$y = a + b\theta. \tag{2.2}$$

where a is the background luminescence and b is the slope depending on the scanner settings (e.g. the exposure time or laser power) and dye properties. The fraction of bound probe θ can be determined using chemical equilibrium calculations [157]. With some relatively mild assumptions and some reformulations, this leads to the following expression [157]:

$$\theta = \frac{\omega_N X_N + \omega_S K_S(x_S - p\theta_S)}{1 + X_N + K_S(x_S - p\theta_S)} \tag{2.3}$$

Here, $\omega_S, \omega_N \in [0, 1]$ are factors representing the fraction of specific and non-specific target, respectively, that stays at the probe after the washing steps, and $X_N = K_N x_N$ is the non-specific binding strengths before washing. K_S and K_N are equilibrium constants for specific and non-specific targets, p is the total amount of probe, and θ_S is the fraction of bound probe that is bound by a specific target. The term $p\theta_S$ accounts for the depletion of the specific target due to binding to the probe [157] and can be calculated by solving the following equation for θ_S:

$$\theta_S = \frac{K_S(x_S - p\theta_S)}{1 + X_N + K_S(x_S - p\theta_S)}. \tag{2.4}$$

Plugging equation (2.3) in equation (2.2) and defining the following new constants:

$$K = \frac{K_S}{1 + X_N} \tag{2.5}$$

$$A = a + b\omega_N \frac{X_N}{1 + X_N} \tag{2.6}$$

$$B = b(\omega_S - \omega_N \frac{X_N}{1 + X_N}) \tag{2.7}$$

leads to the simplified equation [157]:

$$y = A + B\frac{K(x_S - p\theta_S)}{1 + K(x_S - p\theta_S)}, \tag{2.8}$$

where we treat X_N as a constant, focussing on the relationship between x_S and y. For small values of Kx_S, θ_S can be approximated by a linear term cx_S for some appropriate constant c [157]. With this approximation, the previous equation becomes a classical

Langmuir equation

$$y = A + B\frac{\hat{K}x_S}{1+\hat{K}x_S}, \tag{2.9}$$

where $\hat{K} = K(1-pc)$. The typical shape of such a Langmuir equation can be observed in microarray spike-in datasets, where the actual amount of RNA for specific genes is experimentally controlled (Figure 2.4 left). It starts almost linearly with slope $\hat{B} = B\hat{K}$ for very small values of x_S and saturates to its supremum $A + B$ for very large values of x_S.

Gene expression data are typically log2-transformed before analysis. The reasons for this are twofold. First, for many biological questions the relative change in expression, i.e. the fold change, is supposed to be of more relevance than the absolute change in expression and a log transformation allows to use additive models for assessing relative changes. Second, the technical error in microarray data is mainly multiplicative, i.e. it increases with increasing RNA concentrations, and it is heavily skewed to the right. Therefore, a log-transformation results approximately in additive noise with a less skewed distribution, although the distribution is still not normal and there is still some intensity dependence in the noise (see paragraph "Noise models" below).

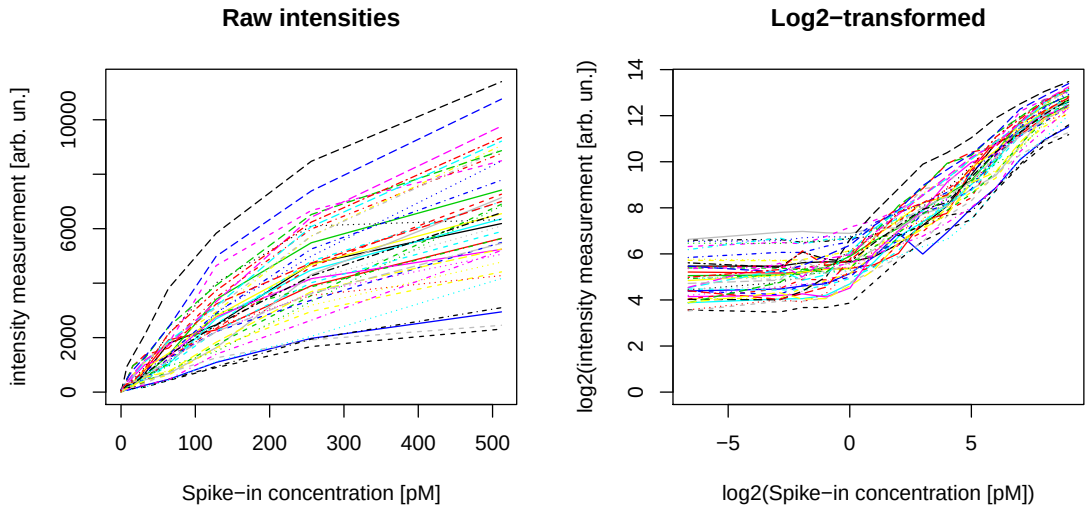


Figure 2.4.: The RNA content of 42 different genes from spike-in experiments [158] is plotted against the corresponding signal intensity of the Affymetrix Human Genome U133A microarray. Each gene is represented by a different colour or line type (data points are linearly interpolated). The raw intensities (left), i.e. without log-transformation, have a typical shape corresponding to the Langmuir equation with varying parameters for the 42 different genes. Log2-transformation of the concentrations as well as the intensity values (right) result in a sigmoidal function with three phases: An initial flat phase, a linear phase with slope close to 1, and a saturation phase which is only slightly visible in the available spike-in range. Note that the spike-in concentration of 0 was set to 0.01 for the right graphic to allow log-transformation.

When we apply a log2-transformation to both x_S and y, i.e. we define variables $lx = \log_2(x_S)$ and $ly = \log_2(y)$, equation (2.9) becomes

$$ly = \log_2(A + \hat{B}\frac{2^{lx}}{1 + \hat{K}2^{lx}}), \tag{2.10}$$

which, for appropriate parameters A, $\hat{B}$, and $\hat{K}$, is a sigmoidal function that can be divided into 3 parts: an initial flat part, a middle part with slope of almost 1, and a late saturation part. The parameter A determines the initial value, the fraction $\frac{A}{\hat{B}}$ mainly specifies the length of the initial flat part, and the value of $\hat{K}$ determines the beginning of the saturation part. This characteristic shape can be observed in the spike-in experiment from [158] (Figure 2.4 right), although the spike-in concentrations are too small in this experiment to appropriately visualise the saturation phase. It is noticeable that the curves are fairly similar for the 42 different probes, indicating that the probe-specific parameters do not vary too much. This seems to be true also for other platforms [159]. These results will be important for the mapping across microarray platforms, which will be discussed in section 4.1.2.

Noise models

The development of appropriate noise models for microarray data is important for the choice of appropriate statistical models for data analysis. However, despite of many investigations and significant progress, the question about the best model is still not fully solved. Initial models incorporated an intensity independent, additive error term with normal distribution [160], or a multiplicative noise term, accounting for the intensity dependence of the error variance [161]. These models could be refined by a linear combination of an additive and a multiplicative component [162], providing a reasonably good first order approximation to the functional dependence of the error on the intensity value. However, more recent investigations suggest that these models do not suffice to fully describe the observed error terms, suggesting a very general intensity-dependent noise model without a specific functional form [163].
The complicated nature of the noise in a microarray study can be partially understood by the physicochemical model described above. If we assume, for simplicity, that every parameter in equation (2.9) can vary by an additive term and additionally incorporate biological noise, we get the following equation:

$$y = (A + \epsilon_A) + (B + \epsilon_B)\frac{(\hat{K} + \epsilon_{\hat{K}})(x_S + \epsilon_{\text{biol}})}{1 + (\hat{K} + \epsilon_{\hat{K}})(x_S + \epsilon_{\text{biol}})}. \tag{2.11}$$

Here, the additive noise component ϵ_A accounts for differences in the background luminescence among other intensity independent effects (equation (2.6)), while the almost multiplicative noise component ϵ_B mainly incorporates differences in scanner settings, dye properties, or washing efficiency. Furthermore, there are two noise terms that can not be

considered as additive or multiplicative, namely the hybridisation associated error $\epsilon_{\hat{K}}$ as well as biological noise ϵ_{biol}, which have more complicated relationships to the intensity value due to the saturation effect. Apart from this, Poisson-distributed error terms have been reported that may occur due to the probabilistic nature of the hybridisation and read out processes as well as the inverse transcription and amplification processes [145]. Such Poisson noise can also not be represented by additive or multiplicative terms. Finally, biological noise is hardly understood and may itself also depend on the mRNA content, meaning that the additive representation of the biological noise in equation (2.11) may be inappropriate. Thus, a first order approximation to the intensity dependence of microarray noise may be reasonable in a certain intensity range. However the saturation effect for high intensity values, affecting mainly prehybridisation and hybridisation noise, as well as stochastic Poisson distributed effects require second or higher order terms.
These very rough theoretical investigations are further complicated by the different data preprocessing and normalisation methods that are commonly applied. These methods try to remove array specific effects and therefore have an additional influence on the noise distribution and intensity dependence. Many articles discussing the intensity dependence of microarray noise for Affymetrix data, use the 'Affymetrix Microarray Suite' (MAS) or 'Probe Logarithmic Intensity Error' (Plier) [164] algorithms for data preprocessing. These algorithms make use of the mismatch (mm) probes on Affymetrix chips and seem to have higher intensity dependence of the error term after log-transformation than the 'Robust Multi-array Average' (RMA) [165] algorithm, which uses only positive match (pm) probes (Figure 2.5). Due to this result, in combination with the fact that newer microarray chips from Affymetrix, e.g. the Human Gene 1.0 st-v1 chip, do not have any mismatch probes, the RMA algorithm will be used for data preprocessing of Affymetrix chips in the present thesis.
Apart from the intensity dependence of the error term, its distribution is also of high interest. It is generally accepted that the noise distribution of non-transformed data is heavily skewed to the right. Therefore, a log-transformation is usually performed to reduce the skewness, almost resulting in a normal distribution of the error term. Nevertheless, there is still some evidence that the distribution is slightly skewed and heavy-tailed [166], which may be, at least partially, attributed to the remaining intensity dependence of the noise term after log-transformation [163].

Correlated noise

Many of the above mentioned error sources, e.g. changes in background intensity, scanner settings or dye properties, as well as hybridisation, washing and amplification processes, are usually not independent between probes, but affect many, if not all probes, in a similar way. Furthermore, biological variability is known to result in correlated noise as well. Only the poisson like stochasticity in the amplification and hybridisation processes can be assumed to be uncorrelated between different genes.

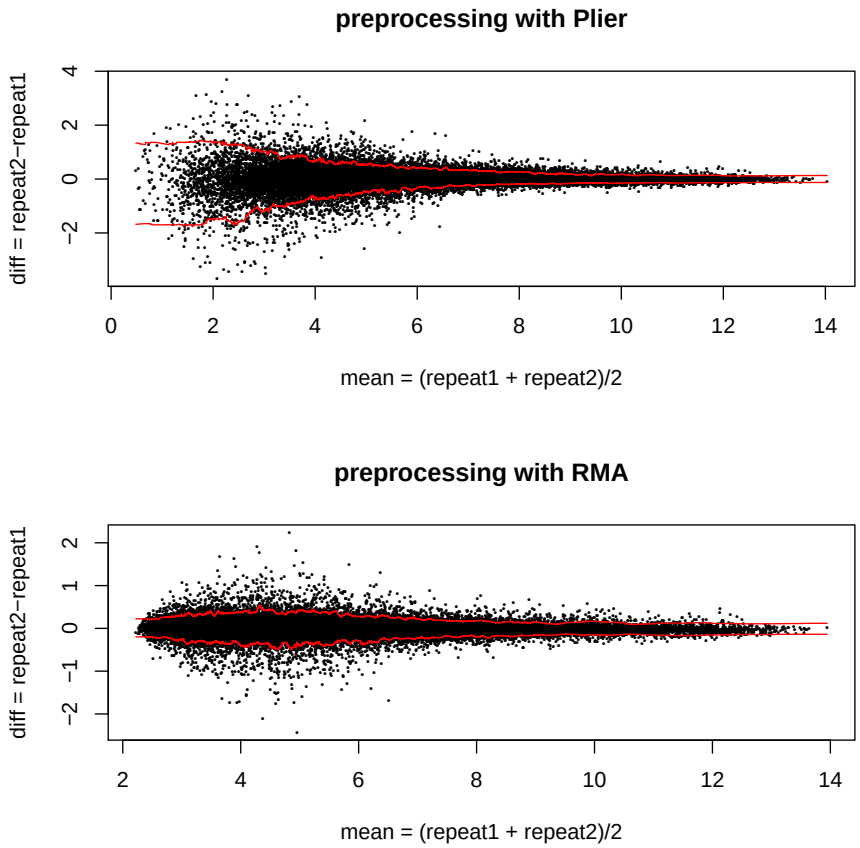


Figure 2.5.: The functional dependence of microarray noise on the signal intensity changes with different preprocessing algorithms. Data preprocessed by the 'Probe Logarithmic Intensity Error' (Plier) algorithm have a relatively high intensity dependence after log-transformation (top), while this effect is considerably reduced for data that were preprocessed by the 'Robust Multi-array Average' (RMA) algorithm (bottom). Shown are the mean vs. the difference of two technical replicates from the T47D human breast carcinoma cell line (dataset GSE19921 from the Gene expression omnibus database [167]). Red lines indicate the 5% and 95% quantiles calculated on a moving window of 500 data points width.

Normalisation methods remove parts of the systematic errors that affect all genes in the same way. However, some between-array differences, like a variability in washing, have different effects on different probes [156], complicating a removal by normalisation methods. Furthermore, there is cross hybridisation as well as some spatial bias found in microarray data, also leading to local rather than global correlations among probes [168, 169]. Overall, it can be expected that the noise in microarray data is substantially correlated between different genes and that normalisation methods can only partially remove these correlated noise terms. Such correlation can severely affect the outcome of data analysis and must be carefully considered. In the framework of this thesis, this is especially important to consider when performing principal components analyses or similar unsupervised dimension

reduction methods that exploit the correlation structure in the data.

2.4. Existing approaches for whole transcriptome based cell characterisation

The idea to characterise cells based on their global gene expression status is relatively new due to the restricted availability of appropriate data in earlier times. However, the acceptance of such methods is increasing due to great success stories, e.g. for the characterisation of pluripotent stem cells [70, 71, 107, 128]. However, these methods were often restricted to single cell or tissue types. In the recent years, three different methods were developed that can be used to determine the cell or tissue type of a specific sample, depicting a multi-class classifier [10, 11, 81]. These methods are described in the following and their relevant properties are compared from a theoretical point of view.

2.4.1. Concordia: Phenotypic concept enrichment

The first of these three methods is the Concorida tool, which was published in 2012 [11]. This tool uses a dataset of 3030 microarray samples from the Affymetrix Human U133 Plus 2.0 platform as a reference for the characterisation of new data. These samples were classified into 1489 anatomy and disease concepts according to the Unified Medical Language System (UMLS) [170], where each sample can have multiple concept associations and each concept was represented by at least three samples.
The characterisation of new data is based on an enrichment score for each UMLS concept, which is calculated as follows. First, the Spearman correlation between the new data and all of the reference data is calculated. Afterwards, for each concept and each sample of the new data, an enrichment score is calculated based on a Kolmogorov-Smirnov like test statistic, comparing the Spearman correlations of the reference samples that are associated with the concept to those that are not associated with the concept [11]. The enrichment score is exactly 1, if the correlations to all samples of the respective UMLS concept are consistently higher than those to all other samples. As a result of the method, the enrichment score is reported for each UMLS concept, depicting a measure of similarity to each concept.
In the original publication, the Concordia tool is only used for a single microarray platform, namely the Affymetrix Human U133 Plus 2.0 platform, on which the reference samples were hybridised. A transformation to other platforms is certainly possible with an appropriate reference dataset on the respective platform. However, the use of a single reference dataset for cross platform analyses may reduce the accuracy of the method and needs to be tested.
One disadvantage of the method can be explained with an idealised example. Suppose, for simplicity, that each sample of the reference dataset is only associated with a single UMLS concept. Suppose further that there is almost no variation between the reference samples belonging to a single UMLS concept, i.e. there is no biological or technical variation between tissues of the same type. Then, if we analyse a new sample that fits to one

specific concept, e.g. a heart biopsy, the enrichment score for the 'heart' concept would be exactly 1. Now, if we analyse another sample, which does not exactly match to any of the concepts, but which is partially similar to one concept, e.g. an *in vitro* differentiated cardiomyocyte, the enrichment score for the 'heart' concept would be also 1. Essentially, for any sample that we analyse, there is (almost) necessarily always a concept that is most similar to this sample, and the enrichment score for this concept is exactly 1, irrespective of the correlation values themselves, i.e. irrespective on how similar the sample is to the samples corresponding to this concept. Therefore, if the new cells do not exactly match to any of the tissues or cell types included in the UMLS concepts from the reference dataset, this will not be recognised by the enrichment score. Similar effects are true for the concept with second highest similarity. The value of this concept can always be expected to be relatively high and it depends only on the sample sizes of the samples that belong to the first and second concept, e.g. it becomes a little bit lower if there are more samples corresponding to the best fitting concept.
Thus, the method does not give real information on how similar the new sample is to a specific concept, but only about the ranking of concepts. This problem is a bit reduced in the real case, where we have some biological and technical variation in the reference dataset. However, especially for cases with a relatively high signal-to-noise ratio the problem persists and it generally prevents the proper characterisation of samples that do not fit exactly to any of the UMLS concepts.

2.4.2. Unknown RNA Sample Annotation (URSA)

The 'Unknown RNA Sample Annotation' (URSA) tool [10] uses more advanced classification techniques to determine the cell or tissue type of new samples. The method was trained on a large dataset of more than 14,000 samples from the Affymetrix Human U133 Plus 2.0 microarray platform. These samples were manually classified into 244 different tissue or cell type terms according to the BRENDA Tissue Ontology (BTO) [171].
URSA is a multi-class classifier, which is able to classify unknown samples accurately into one of the tissue ontologies. This classification is achieved based on a support vector machine (SVM) classification and a subsequent Bayesian correction utilising the hierarchical structure in the BTO.
To be more precise, a SVM classifier is trained for each BTO term, with positive labelling of all samples that correspond to this BTO term or one of its descendants. Samples that are only annotated to an ancestor of the BTO term are excluded and all other samples are labeled negative [10]. Thus, the SVM classifier utilises already the tree like structure of the BTO.
The results of all SVM classifications are then used in a post-processing step to calculate posterior probabilities for each BTO term based on a hierarchical Bayesian correction method [172] that takes the hierarchical structure of BTO into consideration. This post-processing step increases the precision and recall of the classifier [10].

Although the URSA tool was trained on a single microarray platform, it could be shown that it is applicable for other platforms and even for RNA sequencing data [10]. For cross platform analyses, the genes were matched according to their Entrez gene identifiers and the new samples were quantile normalised to adjust their distribution to that of the training samples. Furthermore, a permutation test is used for post-processing, setting each non-significant ontology to zero, in order to suppress unspecific associations with tissue ontology terms [10].
Overall, the URSA tool is a highly accurate and widely applicable tool for RNA based sample annotation. However, one disadvantage with respect to characterisation of *in vitro* differentiated cells is the binary classification for each ontology term, preventing the identification of partial similarities. Furthermore, the Bayesian correction and the permutation test used for post-processing may suppress weak signals that can for instance occur due to a mixture effect in the sample. Therefore, we recommend the tool for a characterisation of pure tissues, but not for characterisation of samples that may deviate from pure tissues, e.g. due to mixture effects or incomplete differentiation. Thus, it is also not well suited for the goal of the present thesis.

2.4.3. CellNet: A Network based classification of cells

CellNet is a very new method published by Cahan et al. in August 2014 [81]. It is specialised for the characterisation of *in vitro* transformed cells, providing information about the global similarity to several different tissues and additionally about possible relevant transcription factors that may be used to improve the transformation to the desired cell type.
CellNet was initially developed on the Affymetrix mouse 430.2 and the Affymetrix human HG U133 Plus 2.0 microarray platforms. It was additionally trained for the Affymetrix mouse and human Gene 1.0 arrays and for the Illumina Human 8v2 array [81]. However, for every platform it is retrained based on a platform specific dataset. It uses data from roughly 20 different tissues with at least 60 (for human) or 100 (for mouse) samples [81].
The training of CellNet starts with the data-based reconstruction of a gene regulatory network (GRN) using the 'context likelihood of relatedness' (CLR) algorithm [173]. The CLR algorithm determines relationships between transcriptional regulators and target genes based on a correlation measure with a context specific significance cutoff. The method is used to construct a global GRN with all data as well as additional GRNs for all samples of "common developmental origin" [81]. These GRNs are then combined to a single GRN per cell or tissue type which is further processed by applying a cutoff for removing lower performing relationships that is determined by a comparison to a Gold Standard. The Gold Standard is determined using appropriate data from transcription factor binding measurements, gene expression profiling of mouse ESCs (mESC) modified for inducible expression of 94 transcription factors, and chromatin immunoprecipitation microarray (ChIP-chip) and sequencing (ChIP-seq) data of 54 transcription factors in mESC.

After that, the GRNs are divided into subnetworks of high density interconnection. Each of the subnetworks is then tested for its tissue specificity by applying a gene set enrichment analysis that compares the expression of one specific tissue to all others on the gene set comprising all genes of the subnetwork. All subnetworks that were significant for a specific tissue are then combined to a single GRN per tissue or cell type [81]. Finally, all genes of this GRN were used to train a Random Forest classifier [174] for each cell or tissue type, which can then be used for classification of new samples.
Besides this characterisation of cells using the Random Forest classifier, CellNet incorporates also a transcription factor focused evaluation of possible ways to improve the cellular transformation [81]. This aspect is neglected in the present thesis since we focus on the characterisation of cells.
One major drawback of the CellNet method is the high data requirement for the construction of the GRNs, preventing the inclusion of more different cell or tissue types due to limitations of available data. Therefore, it is currently limited to roughly 16 to 20 tissues, depending on the organism, i.e. mouse or human, as well as the specific microarray platform. Furthermore, the Random Forest classifier is again a binary classifier, with the disadvantage as discussed for the URSA classifier, although in practice the probability associated with the classification results gives some indication on how well the respective tissue is represented in the newly analysed data. This is different in the URSA tool, where most probabilities are either close to zero or close to one.

3. A two-scale map of global gene expression - combining unsupervised and supervised dimension reduction techniques

Due to the high dimensionality of gene expression microarray data, it is hard to directly interpret the data in terms of their location in the expression space. Therefore, dimension reduction techniques, such as principal components analysis (PCA) [175, 176], can be used to reduce the expression space to the lower intrinsic dimensionality, by exploiting correlations between variables. However, finding the biologically relevant dimensions is a challenge of its own and typical unsupervised methods can suffer from noise and depend heavily on the choice of samples. Furthermore, the interpretation of dimensions determined in an unsupervised way is sometimes ambiguous. Supervised dimension reduction methods can help in this respect, but they have their own disadvantages, associated with incomplete annotation of important processes, or linear dependencies between the detected directions. Therefore, we propose a combination of unsupervised and supervised dimension reduction, in order to join the strengths of both methods.

3.1. An illustrative statistical model for gene expression data

We will use an illustrative statistical model of gene expression to explain some advantages and disadvantages of unsupervised and supervised dimension reduction techniques in the following. This statistical model describes the expression measurements $\underline{y_j}$ of sample $j, j = 1, \ldots, n$ as a linear combination of two qualitatively different types of biological signal plus two qualitatively different types of noise.

Among the two types of biological signal, one is always associated with a binary weight, e.g. describing whether the sample is a heart tissue or not, while the other one is associated with a continuous weight, e.g. describing the proliferation rate of the sample. Thus, for d_b orthogonal biological dimensions $\underline{u_i}, i = 1, \ldots, d_b$ of the first type, the corresponding term of the model can be described as $\sum_{i=1}^{d_b} a_{ij}\underline{u_i}$, where the coefficient a_{ij} is a binary variable stating whether sample j is associated with e.g. the tissue, cell line, or disease i or not. The term associated with the second type of biological signal can be described as $\sum_{l=1}^{d_c} c_{lj}\underline{w_l}$, where d_c is the number of orthogonal biological signals of this type, c_{lj} is a continuous variable describing the phenotype, e.g. the proliferation status, of sample j,

and $\underline{w_l}$ describes the direction of the effect in the expression space.
The noise terms can be distinguished in one term describing correlated noise and one describing uncorrelated noise. They are described as d_n orthogonal noise terms $\underline{v_k}, k = 1, \ldots, d_n$, representing noise terms that are correlated between genes, and an independent identically distributed (i.i.d.) error term $\underline{\epsilon_j}$, i.e. without any correlation between genes.
Putting all of these terms together, we arrive at the following linear illustrative model of gene expression:

$$\underline{y_j} = \sum_{i=1}^{d_b} a_{ij}\underline{u_i} + \sum_{l=1}^{d_c} c_{lj}\underline{w_l} + \sum_{k=1}^{d_n} b_{kj}\underline{v_k} + \underline{\epsilon_j}., \tag{3.1}$$

where all components u_i, w_l, and v_k are assumed to be orthogonal to each other. Here, the noise coefficient b_{kj} is a random variable with mean 0 and variance 1, the coefficient c_{lj} has also variance 1, and the coefficient a_{ij} is a binary variable describing the tissue, cell line, or disease affiliation. That is, a_{ij} is 1 for the n_i samples that are affiliated with the specific phenotype, e.g. for all hematopoietic cells, and 0 for the $n - n_i$ samples that are not affiliated with the phenotype. Each component of the m dimensional vector $\underline{\epsilon_j}$ of the noise term is assumed to be normally distributed with mean 0 and standard deviation σ. Here, m is the length of the vectors in equation (3.1), i.e. the number of genes measured by the microarray.
This model is certainly very simplistic and does not represent the truth. However, it can be used to explain some conceptual differences as well as associated advantages and disadvantages of supervised and unsupervised dimension reduction techniques. The reasons for the choice of these different terms in the model are the following. First of all, it is necessary to distinguish between biological variation and measurement noise in order to separate the effects of interest from disturbing effects. Second, the distinction between the two types of noise is necessary due to the clearly different statistical behaviour of correlated and uncorrelated noise terms. In many cases of statistical analyses of microarray data, the correlated noise is neglected since it is more complicated to model. This can have significant confounding effects, e.g. for the analysis of gene sets [177]. Third, the separate modelling of the two biological signal types is mainly due to the availability of annotation. In most studies that are deposited in public databases only binary phenotypes, e.g. the cell or tissue type, are annotated. Therefore, this term can be nicely determined by supervised methods. In contrast, continuous variables, e.g. the proliferation rate, are typically not annotated and can therefore only be determined by unsupervised methods.

3.2. Exploring gene expression spaces with unsupervised dimension reduction techniques

Unsupervised dimension reduction methods have the general advantage of being independent of available annotations. Therefore, they can be regarded as unbiased with respect to specific aims of the analysis and do also not suffer directly from any annotation errors. In

contrast, a common disadvantage of unsupervised methods is the more complicated and sometimes ambiguous interpretation of the respective dimensions in terms of a biological meaning, although methods like non-negative matrix factorisation (NMF) [109] have been designed that usually facilitate a more intuitive interpretation of the detected components. Apart from these annotation and interpretation related advantages and disadvantages, there are also more quantitative facts to consider as will be described in the following. For this analysis, we focus on one of the most simple and most frequently used unsupervised dimension reduction techniques, namely PCA [175, 176]. We first describe the general concept of PCA and discuss its optimality for dimension reduction in certain cases. Afterwards, we apply PCA to large gene expression datasets and discuss some general problems that arise in this analysis.

3.2.1. Principal components analysis - a (sometimes) optimal linear dimension reduction technique

Concept of principal components analysis

PCA [175, 176] is a technique in multivariate statistics and data mining that is useful when several variables under investigation are correlated with each other. Due to these correlations, there is some redundant information in the variables which may be problematic for further analysis, e.g. due to singular matrices in regression analysis, and which may be exploited to reduce the dimensionality of the variable space.
The concept of PCA is to successively detect orthogonal directions with highest variance in the data. That means that it first detects the direction with highest variance in the complete variable space, which is called the first principal component. In a second step, the analysis is restricted to the orthogonal space to this first principal component and a second principal component is detected with the highest variance in this space. This procedure is then continued iteratively, resulting in a new coordinate system with rotated directions (Fig. 3.1).
The directions detected by this procedure correspond to the eigenvectors of the covariance matrix of the data. Thus, one procedure to calculate a PCA is an eigendecomposition of the covariance matrix. However, due to numerical stability reasons a singular value decomposition (SVD) on the centred data, i.e. with mean 0 for each gene, is usually preferred.
SVD factorises the centred data matrix $X \in \mathbb{R}^{m \times n}$ into three matrices

$$X = U \Sigma V^T, \tag{3.2}$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are matrices whose columns are orthogonal unit vectors, called the left and right singular vectors. The matrix $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal with positive numbers on the diagonal which are called the singular values. The right singular vectors, i.e. the columns of V equal the eigenvectors of the sample covariance matrix of the data, which are used as new coordinate axes. The multiplication of the two other matrices,

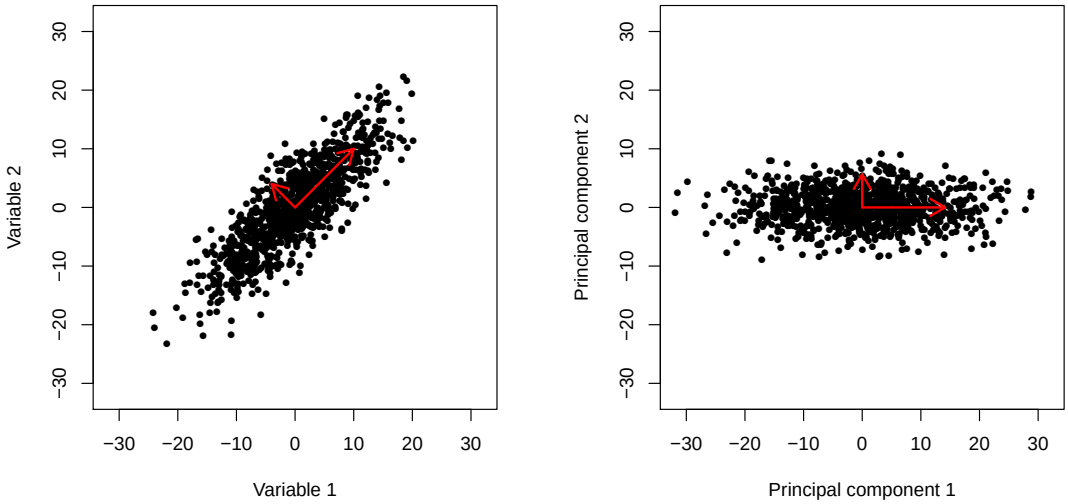


Figure 3.1.: Illustration of principal components analysis (PCA). PCA detects successively the orthogonal directions of highest variance. Here, the two variables (Variable 1 and Variable 2, left graphic) are positively correlated. The first principal component (PC) has the direction of the longer red arrow. The second PC is orthogonal to the first PC, showing in the direction of the shorter red arrow. The right graphic depicts the data in the PCA-rotated coordinate system.

i.e. $U\Sigma$ are consequently the coordinate values of the data points in this new coordinate system.

In many practical cases, PCA is used for dimension reduction by keeping only the first p dimensions of the PCA-rotated coordinate system. These first p dimensions contain large amounts of the variance in the data due to the variance maximisation criterium used to determine these principal components. One essential question in this procedure is how to choose p, i.e. the number of principal components that are kept after dimension reduction. This can be done, e.g. based on the amount of variation that is explained by the respective components or based on the relevant information that is captured by the components.

Another question that arises is whether or in which circumstances the PCA provides a good or even an optimal dimension reduction and how this optimality is measured. This question will be investigated in the following.

PCA and its optimality for normally distributed data

It is often noted that PCA assumes normally distributed data or is at least optimal for normally distributed data. However, in many applications such a normal distribution can not be assumed. Therefore, it is worth to describe, in more detail, the properties of the PCA in these contexts.

First of all, we would like to note that PCA does not assume any normal distribution. That means that it is valid also for non-normally distributed data, in contrast to many parametric statistical tests, e.g. the t-test. Therefore, the main question to address is the

optimality of PCA and the dependence of this optimality on the distribution of the data. When we talk about optimality, it is necessary to define some optimality criteria. These criteria certainly depend on the application of interest. Therefore, we will distinguish three different cases in the following for which we discuss the optimality of the PCA.
In the first case, we assume that we have an input-output system with p independent input variables $\underline{x} = (x_1, \ldots, x_p)^T$ that control the system and m output variables $\underline{y} = (y_1, \ldots, y_m)^T$ that depend linearly on $\underline{x}$:

$$\underline{y} = A\underline{x} \tag{3.3}$$

The goal is to determine the values of the p input variables as well as the linear mapping A given n observations of the m output variables. This problem is well known as the source separation problem and it amounts to finding p independent components.
PCA can be used to solve this problem under certain assumptions. PCA imposes orthogonality constraints on the detected directions and removes all correlations. This is sufficient to find independent components when the data follow a multivariate normal distribution, i.e. $\underline{x} \sim N(\underline{\mu}, \Sigma)$, since the normal distribution is fully characterised by the mean vector and covariance matrix. Thus, a correlation of zero implies independence of the components. However, this is not true for other distributions, for which PCA is not the optimal choice for source separation. In this case, it is important to use a more general concept of independence as is done for example in independent component analysis (ICA) [178].
Apart from that, it is important that the linearity assumption is fulfilled, i.e. that the outputs depend linearly on the inputs. Otherwise, non-linear methods have to be used.
The second case we want to consider involves some noise. Assume again that we have p input variables and m output variables that are now corrupted by noise:

$$\underline{y} = A\underline{x} + \underline{\epsilon}, \tag{3.4}$$

where $\underline{\epsilon}$ is independent of $A\underline{x}$. The goal now is not to detect precisely the values of the p input variables, but to find a p-dimensional subspace that contains the signal. Thus, we try to find a reduced dimensional representation that captures the true signal as good as possible. In the first case described above, the entire signal lies in the p-dimensional PCA-space in case of a linear model, even if the independent components can not be directly estimated by the PCA, i.e. even when the data are not normally distributed. This is due to the reason that we do not have any noise. However, with the presence of noise this may change.
In fact, PCA minimises the squared error between the data and the reduced dimensional representation. Hence, given n observations of the output variables, it determines the matrices $\hat{A} \in \mathbb{R}^{m \times p}$ and $\hat{X} \in \mathbb{R}^{p \times n}$ such that $||Y - \hat{A}\hat{X}||_2$ is minimised.
This minimisation of the squared error is equivalent to maximising the likelihood for normal distributed noise. However, when the noise is not normally distributed, the maximum

likelihood estimate does not equal the classical PCA solution and some extensions, e.g. to the exponential family [179], are required.
Thus, in this second case, only the noise term, but not the data themselves need to be normally distributed in order to receive optimal solutions in terms of a maximisation of the likelihood.
The third case we consider here is actually the same as the second case, i.e. we aim to reduce the dimensionality, capturing as much of the relevant signal as possible. However, we use an information theoretic optimality criterium instead of a likelihood maximisation. This optimality criterium is related to mutual information $I(X,Y)$, which is defined for two random variables X and Y with values in Ω_X and Ω_Y as

$$I(X,Y) = \sum_{x\in\Omega_X, y\in\Omega_Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{3.5}$$

In fact, the optimality criterion that we consider here is the minimisation of the relevant information loss by the dimensionality reduction. Here, the relevant information loss is defined as the difference in mutual information of the data and the relevant signal before and after dimension reduction [180]. In terms of the model in equation (3.4) the relevant information loss $L_{AX}(Y,\hat{Y}_p)$ between the original data Y and the projected data on the reduced dimensional subspace $\hat{Y}_p$ can be expressed as [180]

$$L_{AX}(Y,\hat{Y}_p) = \sup\left(I(AX,Y) - I(AX,\hat{Y}_p)\right), \tag{3.6}$$

where the supremum is taken over all partitions of the sample space.
Geiger and Kubin [180] show that this information loss is zero when PCA is applied for dimension reduction and the noise $\underline{\epsilon}$ is independently identically normally distributed with zero mean. This result does not require any specific distribution of the data themselves, but only of the noise. This statement can even be extended to non-normally distributed noise, as long as the noise is independently identically distributed (i.i.d.) and is more similar to a normal distribution than the signal. Here, the similarity of two distributions is measured by the Kullback-Leibler divergence [180, 181].
This result is very interesting for our approach since we try to reduce the dimensionality while keeping as much relevant information as possible. Furthermore, the normality of the noise component is almost fulfilled for logarithmically transformed gene expression microarray data, while the data themselves are far from being normally distributed. Therefore, PCA should be well suited with respect to these requirements.
However, there are two problems that cannot be neglected. First, the noise can not be completely regarded as being independent, i.e. there is correlated noise present in microarray data. Second, there are some additional problems that come along with finite sample sizes in high dimensions, meaning that the results of PCA depend on the specific choice of samples. These problems will be further investigated in the next sections.

3.2.2. Principal components analysis on large microarray datasets

Lukk et al. [12] applied PCA for dimension reduction purposes on a large dataset of 5372 samples from the Affymetrix Human U133 A array, representing 369 different tissues, cell lines, or disease states. They detected, that the first principal component (PC) of their analysis separates hematopoietic cells from all other cells, and termed it the hematopoietic axis (Fig. 3.2 top left). The second PC was associated with malignancy or proliferation, distinguishing clearly between fast proliferating cell lines and slower proliferating tissues. Interestingly, this separation on the second component suggests a clustering of all cell lines into a single cluster, irrespective of and separated from their tissue of origin. This was also confirmed by a hierarchical cluster analysis applied to the mean expression vectors for each tissue or cell line [12]. The third PC was associated with neural tissues, and can therefore be termed the neural axis (Fig. 3.2 top right). Thus, the first three principal components could be clearly associated with a biological meaning. However, this was not the case for the fourth component, which correlates with an array quality metric, suggesting that it represents measurement noise [12]. Therefore, Lukk et al. suggest that the gene expression space, besides smaller scale differences, has a very low intrinsic dimensionality. Schmid et al. [11] performed PCA on a different dataset of 3030 samples and achieved slightly different results. They could confirm a separation of hematopoietic and neural tissues from all others on their first two principal components, but did not find a separate clustering of cell lines on the first two components. Unfortunately, they do not show any results for further components.

In order to evaluate the results of Lukk et al. [12] on an independent dataset, we assembled our own independent dataset of 7100 samples from the Affymetrix Human U133 Plus 2.0 array and performed PCA on it. The results of this analysis are shown in Fig. 3.2 (bottom), confirming a separation of neural cells, cell lines, and hematopoietic cells from all other cells on the first three PCs. However, the order of the separation of these cell types is different as in the Lukk [12] data set (Fig. 3.2) and the biological interpretation of PC 2 is a bit ambiguous. Furthermore, the fourth PC is not associated with an array quality metric, but has a clear biological meaning, separating liver cells (grey dots at the bottom) from all other cells. This dataset dependent and sometimes uncertain biological interpretability is a common challenge for unsupervised dimension reduction approaches. Further principal components of both datasets are depicted in the supplemental Fig. B.1 and B.2.

This visual comparison of the first principal components of both datasets can also be quantified by statistical analyses. A linear regression model trying to explain each of the first three PCs of our own dataset using the first three PCs of the Lukk [12] dataset results in R-squared values of 0.74, 0.66, and 0.70. These are relatively high values, considering the fact that the comparison of the PCs is also affected by platform specific differences. In contrast, trying to explain the fourth PC of the own dataset by the first 5 PCs of the Lukk dataset results in a R-squared value of 0.07, and it increases to a still comparably low value of 0.50 when the first 10 PCs of the Lukk dataset are used as regressors. Thus,

this analysis supports the visual observation that the first three PCs of both datasets span similar spaces, while there are more clear differences in further principal components (this is true also for other than the fourth PC, data not shown).

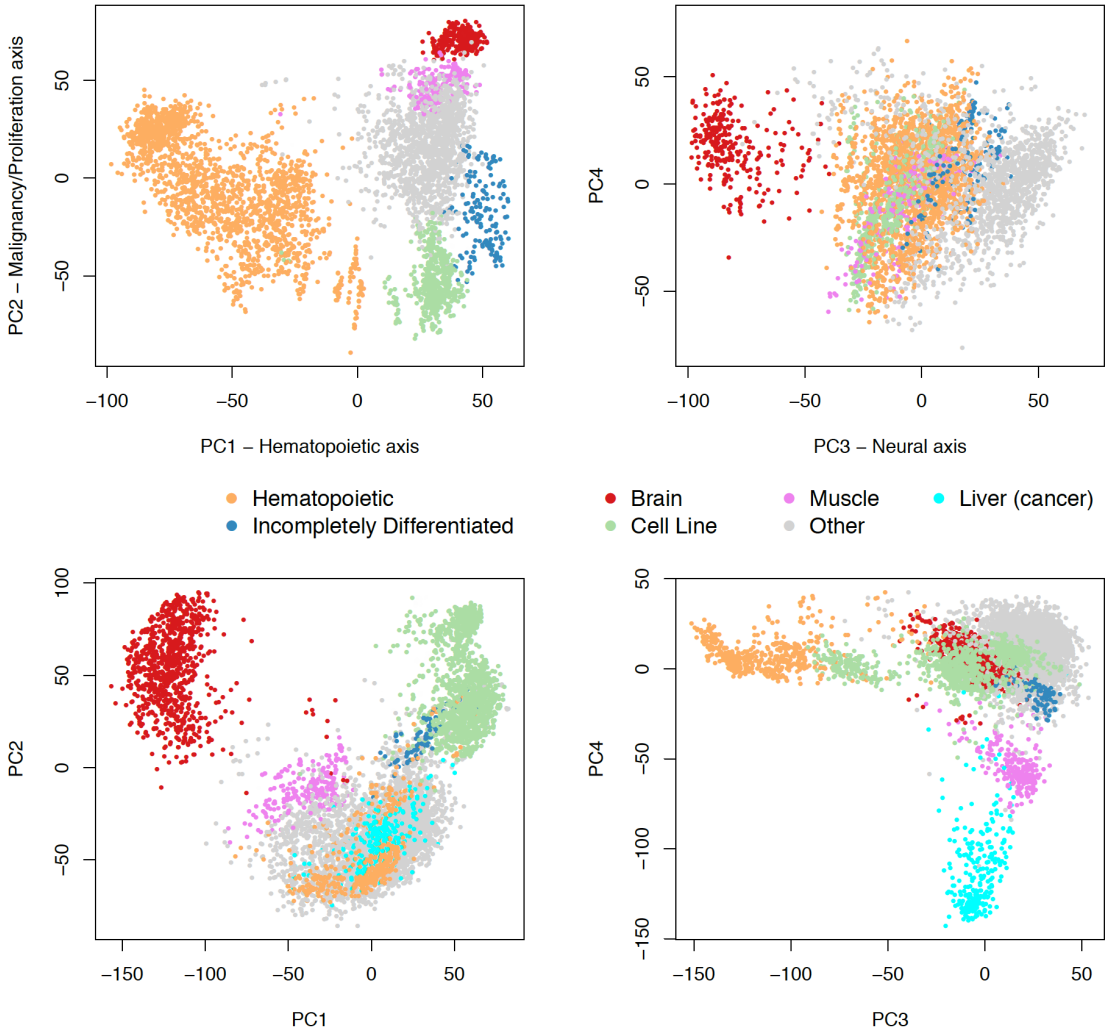


Figure 3.2.: PCA applied to the Lukk dataset [12] (top) and the self created dataset (bottom) reveals differences in the tissue associations of the respective PCs. For the Lukk (own) dataset, PC1 (PC3) is associated with hematopoietic cells (orange), PC2 (partially PC1) is associated with proliferation, and PC3 (partially PC2) is associated with neural cells (red). PC4 in the Lukk dataset is associated with an array quality metric [12], while PC4 in the own dataset mainly distinguishes between liver and liver cancer cells (cyan) and all other cells. Liver cells are only separately coloured for our own dataset, not for the Lukk et al. [12] dataset, where they are among the grey dots. For both datasets, PCs 1 to 3 consistently distinguish mainly between neural cells, hematopoietic cells, cell lines, and all other cells.

3.2.3. Effects of sample size and measurement noise on principal components analysis

There are two interesting observations from the previous section, which we want to investigate in more detail in terms of the illustrative statistical model described above. First, the observation by Lukk et al. [12] that the fourth PC of their dataset is correlated with an array quality metric. This may suggest that the intrinsic dimensionality of the gene expression space is very low dimensional, at least if only those dimensions are counted that exceed the level of noise. Second, the observation that principal components differ between the datasets of Lukk et al. [12] , Schmid et al. [11], and our own dataset. This effect can be expected to be caused by differences in the sample sizes of the relevant cell types or tissues.

In order to get a rough understanding of the effects seen in the PCA, it is necessary to determine the eigenvalues of the sample covariance matrix. In the classical statistical setting of a relatively small number of variables m, and a large number of samples n, approximated by the limit of $n \to \infty$ for fixed m, the eigenvalues λ_r of the sample covariance matrix converge to the eigenvalues of the population covariance matrix [182]. In the illustrative statistical model described in section 3.1, we do not have a normal distribution in all terms. Therefore, it is not straightforward to explicitly determine the population eigenvalues. However, we can state the variability in the direction of the different components that are included in the illustrative statistical model, which are given by

$$\tilde{\lambda}_r = \begin{cases} \frac{n_i}{n}(1-\frac{n_i}{n})||u_i||^2 + \sigma^2, & r = I_{d_b}(i) \\ ||w_l||^2 + \sigma^2, & r = I_{d_c}(l) \\ ||v_k||^2 + \sigma^2, & r = I_{d_n}(k) \\ \sigma^2, & r > d_b + d_c + d_n \end{cases} , \tag{3.7}$$

Here, $I_{d_b}(i)$ ($I_{d_c}(l)$,$I_{d_n}(k)$) is the i-th (l-th, k-th) element of the index set corresponding to biological components of type one (biological components of type two, correlated noise components). The components u_i, w_l, and v_k were assumed to be orthogonal to each other. Although, the $\tilde{\lambda}_r$ do not equal the eigenvalues, a higher value of $\tilde{\lambda}_r$ means that this component of the statistical model will be more strongly represented in the first principal components. Therefore, the $\tilde{\lambda}_r$ are directly associated with the eigenvalues.

The asymptotic convergence of sample eigenvalues to population eigenvalues can only be reasonably used for the case $n \gg m$. For microarray data, in contrast, m is usually at least in the same order of magnitude as n, often even considerably larger. Therefore, it is more realistic to consider the asymptotic case of n and m jointly approaching infinity, i.e. $n \to \infty$, $m \to \infty$ and $\frac{m}{n} \to \gamma \in (0, \infty)$, resulting in fundamentally different statistics.

In this "large n, large m" case, the sample eigenvalues do not converge to the population eigenvalues. Instead, the eigenvalues of the noise term, having population eigenvalues

$\lambda_l = \sigma^2, \forall l$, converge to a limiting distribution F, with density [183, 184]:

$$f(x) = F'(x) = \begin{cases} \frac{1}{2\pi\gamma x\sigma^2}\sqrt{(b-x)(x-a)}\mathbf{1}_{\{a\leq x\leq b\}}, & \gamma \leq 1 \\ \frac{1}{2\pi\gamma x\sigma^2}\sqrt{(b-x)x}\mathbf{1}_{\{0<x\leq b\}}, & \gamma > 1 \end{cases}, \tag{3.8}$$

where $a = \sigma^2(1-\sqrt{\gamma})^2$, $b = \sigma^2(1+\sqrt{\gamma})^2$, and $\mathbf{1}$ is the indicator function. For $\gamma > 1$ there is an additional point mass $1 - \frac{1}{\gamma}$ at $x = 0$. Equation (3.8) specifies the limiting distribution of all sample eigenvalues. Of special interest is the distribution of the largest eigenvalue of the noise term, since it is interesting to compare it to the eigenvalues corresponding to biological components. In [185, 186] it is shown that the largest eigenvalue converges almost surely to $b = \sigma^2(1+\sqrt{\gamma})^2$, i.e. the upper boundary of the support of f, under mild conditions. Thus, for the Lukk dataset with $\gamma \approx 4$, the largest eigenvalue corresponding to the independent noise term is $9\sigma^2$ instead of σ^2. This limited sample size effect is even considerably stronger for smaller datasets, due to an increased value of γ.

Another sample size associated effect, affecting the eigenvalues of the covariance matrix, is the decrease of the eigenvalues corresponding to biological signals of type one with more extreme values of $\frac{n_i}{n}$, e.g. for the case of fixed $n_i \leq \frac{n}{2}$ and increasing n. To give an example, consider a typical cell type of the Lukk dataset with roughly $n_i = 15$ samples (5372 samples divided into 369 distinct tissues or cell types). The $\tilde{\lambda}_r$ value corresponding to this tissue specific component can be calculated according to equation (3.7) to $0.0028||u_i||^2 + \sigma^2$. The same component on a subset of the data with all 15 samples of this tissue and another 15 samples of other tissues would result in a value of $0.25||u_i||^2 + \sigma^2$ and would thus be almost 100 times higher (assuming that σ^2 is negligible). At the same time, the sample variance corresponding to correlated noise is almost not affected by these sample size differences, when we assume that it is independent of the sample choice and phenotypes. This may explain the occurrence of a noise related component already in the fourth PC of the Lukk dataset. Two examples of the relevance of these sample size effects are shown in Fig. 3.3. One example shows the first two PCs on a subset of our our own dataset, including all 28 muscle samples, 28 heart samples and 10 further samples, 2 of each of the 5 main classes distinguished in Fig. 3.2, i.e. brain, cell line, hematopoietic, incompletely differentiated, and others. Due to the changes in the sample size, the first PC now separates muscle tissues from all other samples, and the second PC distinguishes between heart and muscle tissues on one hand, and the other 10 samples on the other hand (Fig. 3.3, left). This result is of course only a toy example, and the outcome may have been expected by intuition. However, the mathematical principle behind is valid for the complete dataset and statements about the intrinsic dimensionality based on first occurrences of noise associated components must be taken with care.

The second example is based on a PCA on the brain tissues from the Lukk dataset (Fig. 3.3, right), showing that the first PC separates cerebellum and other tissues, whereas the second PC separates frontal cortex tissues from more interior brain tissues, i.e. hypothalamus and caudate nucleus. Thus, for the reduced dataset, these two biologically interpretable

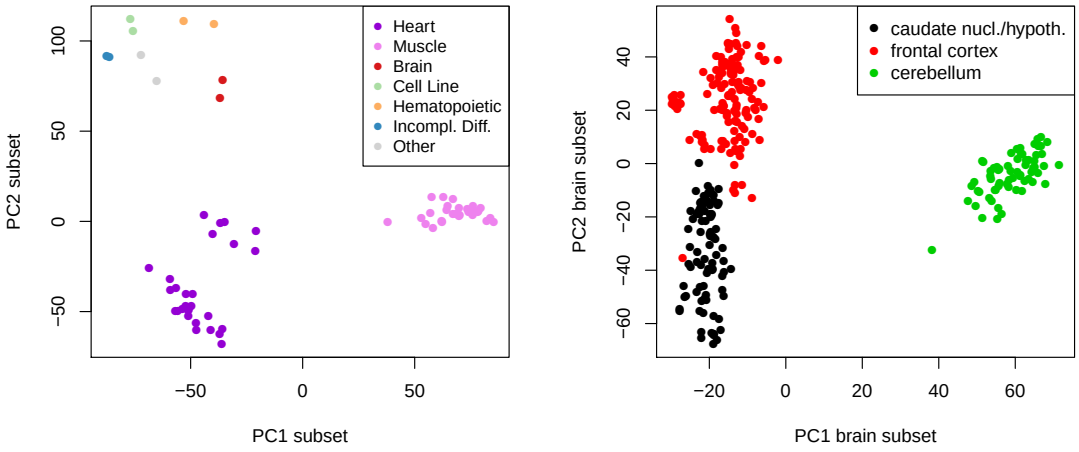


Figure 3.3.: PCA analyses are severely affected by the number of samples for individual tissues/cell types (left) and can detect further tissue specific patterns when applied to subsets of the data (right). On the left, the first and second PCs of a subset of the own dataset are shown, distinguishing mainly between the 28 heart samples (violet), 28 muscle cells (magenta), and the 10 other samples. On the right, PCA is applied to the brain tissues of the Lukk dataset, distinguishing nicely between different brain regions.

directions in the expression space have higher variances than any noise terms, as opposed to the results from the global analysis. Similar sample size effects are probably responsible for the differences in the PCs of the Lukk et al. [12], Schmid et al. [11], and our own dataset. For example, the Schmid et al. [11] dataset contains considerably less cell line samples than the other two datasets, the Lukk et al. [12] dataset contains a very large amount of hematopoietic samples, and our own dataset incorporates comparably many healthy or diseased liver tissues, explaining the direction of the fourth PC.

Beside the sample size dependence, the proper identification of the biological dimensions depends of course also on the effect size, i.e. the length of the vectors $\underline{u_i}$ or $\underline{w_l}$, in comparison to the strengths of the correlated noise terms as well as the variance σ of the uncorrelated noise.

Another problem of unsupervised methods is that the interpretation of the detected directions can be obscured since they may represent a linear combination of biologically relevant directions. This problem is actually associated with the non-optimality of PCA in detecting independent components when the data are not normally distributed (as discussed in section 3.2.1).

In summary, the PCA based dimension reduction has especially problems to detect the phenotypes that are represented by relatively few samples, i.e. the vectors $\underline{u_i}$ with a small corresponding n_i.

3.3. Supervised dimension reduction of gene expression spaces

In supervised dimension reduction approaches, some information about the data are used to find appropriate directions for the reduced coordinate system. Taking the illustrative statistical model from above as an example, one possibility would be to determine the coefficients a_{ij} or c_{lj} based on the information that are *a priori* available, instead of learning them from the data. Using these coefficients, supervised dimension reduction approaches can then determine the reduced coordinate system, i.e. the vectors $\underline{u_i}$ and $\underline{w_l}$.

3.3.1. Linear regression analysis for supervised dimension reduction

A classical way of doing this is linear regression analysis, which requires the specification of a statistical model to describe the data. In analogy to the illustrative statistical model used above (equation (3.1)), a multivariate linear regression model could be specified as

$$\underline{y_j} = \sum_{i=1}^{d_b} a_{ij}\underline{u_i} + \sum_{l=1}^{d_c} c_{lj}\underline{w_l} + \underline{\epsilon_j}, \tag{3.9}$$

without the explicit incorporation of the correlated noise terms. The choice of the specific method for calculation of the unknown variables $\underline{u_i}$ and $\underline{w_l}$ depends on the distribution model for the error term $\underline{\epsilon_j}$. In most cases, the error term is assumed to be normally distributed with mean zero and covariance matrix Σ, i.e. $\underline{\epsilon_j} \sim N(0, \Sigma)$.
In a very general setting without any restrictions to Σ, the correlated noise can be captured by $\underline{\epsilon_j}$. However, for microarray experiments such a general unrestricted setting for Σ can not be used, since the high dimensionality requires too high sample sizes to allow an estimation of all parameters. Therefore, many approaches neglect the correlations by choosing Σ to be diagonal, i.e. with all off-diagonal elements set to zero. More advanced methods account for the correlated noise terms using a multi-step approach, combining the regression model with a PCA analysis on the residuals to identify so called 'surrogate variables' that may have confounding effects [187]. Other approaches restrict the flexibility of Σ even more, by assuming a certain similarity of the diagonal elements in an empirical Bayes setting [188].
These different methods can have very large effects on the statistical significance assessment, increasing the power of detecting significant differences or reducing bias that can result from unmodelled correlated noise. However, we are not directly interested in the statistical significance, but aim to get estimates of the vectors $\underline{u_i}$ and $\underline{w_l}$. These estimates are independent of the covariance matrix and are therefore not affected by the specific structure of the covariance matrix. Therefore, we assume for simplicity a normal distribution with diagonal covariance matrix Σ and no further constraints. This results essentially in a linear model for each gene, i.e. it can be evaluated gene wise. The unknown variables $\underline{u_i}$ and $\underline{w_l}$ can then be determined by an ordinary least squares fitting of the model to the

data, which is the best linear unbiased estimate according to the Gauss-Markov theorem [189, 190].
The ability to determine appropriate directions in a supervised manner depends strongly on the availability and the reliability of *a priori* phenotypical information. In the following, we use the cell type annotation in the Lukk et al. [12] dataset for supervised dimension reduction. Thus, we use a regression model with binary information only, which can be described by

$$\underline{y_j} = \sum_{i=1}^{d_b} a_{ij}\underline{u_i} + \underline{\epsilon_j}, \tag{3.10}$$

In this special case the a_{ij} have distinct support, meaning that a single sample is always associated with a single cell type. Therefore, the regression estimates $\underline{u_i}$ equal the mean values of all samples corresponding to a specific cell type. When we include an intercept for each gene, $\underline{u_i}$ equal the difference between the mean values of all samples corresponding to a specific cell type and the overall mean of all samples.
This results in one direction or signature per tissue type or cell line, i.e. in the case of the Lukk dataset this approach would generate 369 signatures corresponding to the respectively annotated groups of samples. Geometrically, each of these signatures is the vector pointing from the overall mean to the mean of the respective tissue or cell type in the log-transformed expression space.
Due to the use of log-transformed data, the difference between the two mean values equals the logarithmic fold change. Other approaches, e.g. a (modified) t-test based comparison [72] or support vector machines, can also be used, but do not have such a straightforward interpretation in terms of a vector between two well defined points.
The approach described so far does not have any restriction with respect to orthogonality of the tissue specific vectors. Therefore, it is not surprising that there are relatively strong correlations between the signatures (Fig. 3.4). These correlations are to a large extent due to general joint processes like proliferation, or similarities between cells or tissues with similar functions or cellular compositions, e.g. between hematopoietic cells or different brain regions, as can be seen in the PCA analysis in section 3.2.2.

3.3.2. Advantages and disadvantages of supervised dimension reduction

The advantages of a supervised dimension reduction in comparison to an unsupervised dimension reduction are the direct interpretability of the directions due to the direct link to a phenotypical annotation as well as the ability to detect directions that are represented by very few samples. Such directions are often not detected by unsupervised methods due to their sample size dependence as described above.
The main disadvantage of supervised dimension reduction techniques is that the specific structure in the data must be known beforehand, i.e. it is necessary to specify a precise statistical model, and all independent variables in this model must be known *a priori*. Furthermore, the directions that are found by the supervised method are not necessarily

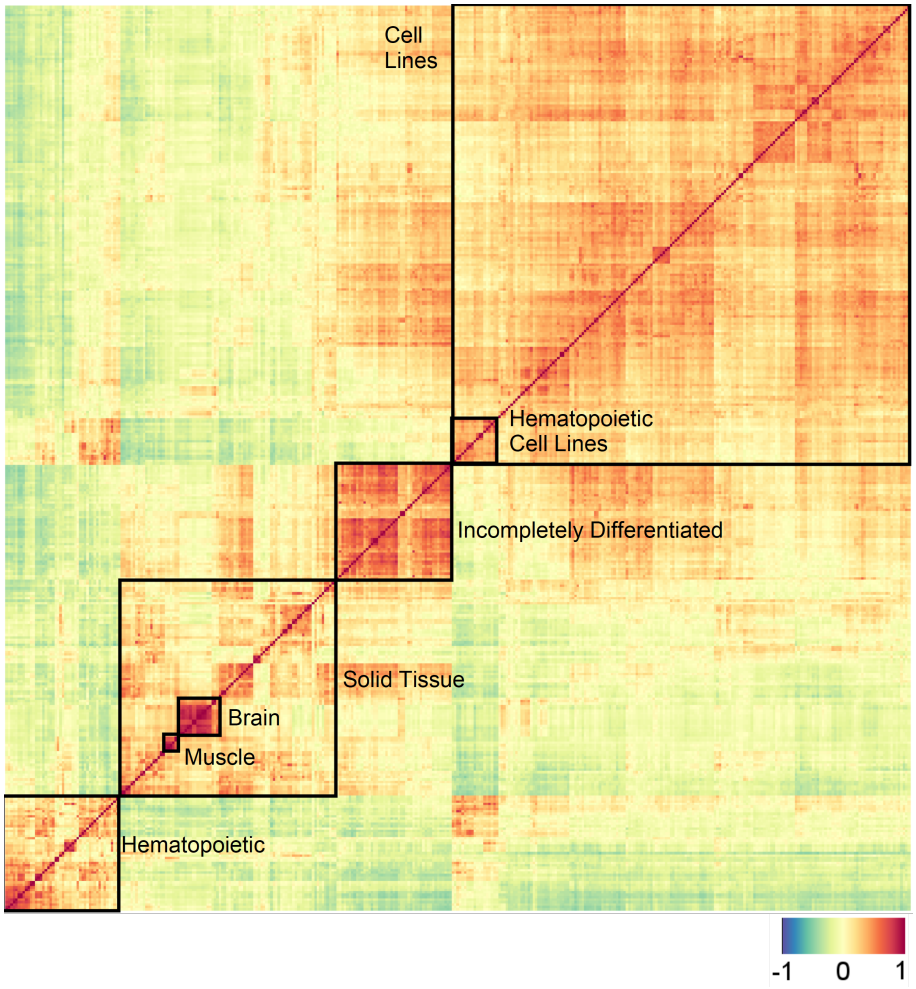


Figure 3.4.: Correlation matrix of the 369 tissue specific signatures derived from the Lukk [12] dataset, showing strong correlations between different signatures. Several clusters can be identified (boxes and text) that are similar to the main clusters identified by PCA in Fig. 3.2. The ordering is based on hierarchical clustering of the columns and rows of the correlation matrix (euclidean distance, complete linkage).

orthogonal. In fact, these directions can have very high correlations as can be seen in the example of cell or tissue type based dimension reduction of the Lukk [12] dataset (Fig. 3.4). This complicates a proper projection onto the space spanned by these directions, especially when singular matrices occur.

The requirement to know all independent variables of the model can be critical when using retrospective datasets without a standardised annotation.

Furthermore, especially continuous variables that are represented by the coefficients c_l in the illustrative statistical model (equation (3.1)) are often not available at all. A prominent example is the proliferation rate of the cells, which has a strong effect on the overall gene expression [68], as can be seen by the PCA analysis (Fig. 3.2). The information on the proliferation rate of cells is usually not available. Therefore, it is not possible to detect such directions in a supervised setting.

The non-orthogonality of the directions is critical for several tissues or cell types. There are several groups of tissues that have relatively high correlations, such as the brain tissues, hematopoietic cell types, incompletely differentiated cells, muscle tissues, solid tissues, or cell lines in general (Fig. 3.4). Together with the finding from the PCA analysis in section 3.2.2, this indicates that there is some tree-like structure in the expression space. According

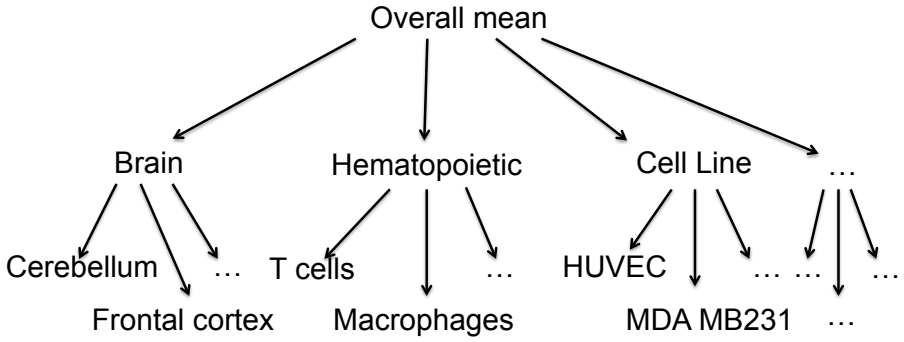


Figure 3.5.: Illustration of the tree-like structure in the gene expression space. There are some directions that are shared by several cell types, e.g. by brain tissues from different regions or by different cell lines. However, at a finer scale these tissues or cell types are also different from each other. The joint directions of these more detailed tissue types lead to a relatively strong correlation in their gene expression.

to this structure, larger groups of different tissues can be first separated from each other and split into smaller groups afterwards (Fig. 3.5). These joint directions of several cell or tissue types lead to relatively strong correlations among them. Therefore, the supervised dimension reduction with vectors pointing directly to the annotated cell or tissue types leads to relatively large correlations between directions.

Supervised methods that impose a tree structure, e.g. decision trees, use directly the given variables, i.e. genes, to find an appropriate tree for classification. Therefore, they do not search for new directions, i.e. combinations of genes, to build up the tree. Thus, as long as the detailed tree structure is not known, it is not directly possible to account for it by supervised methods.

Now, one might argue that there is also a tree-like structure in tissue ontologies, which is also exploited by existing methods, e.g. by the URSA method [10] (section 2.4.2). However, this does not necessarily fit to the tree-like structure in the data. For example, cell lines are usually close to their tissue of origin in the BRENDA tissue ontology, e.g. a breast cancer cell line can be found under breast - breast cancer cell - breast cancer cell line. However, the data analysis based tree structure suggests that cell lines are relatively far apart from their tissue of origin and constitute a separate branch in the tree.

One alternative possibility would be to use orthogonalisation strategies in order to reduce the correlation structure in the tissue specific vectors. While this may be of advantage, it reduces the interpretability of the directions and it still does not impose a tree-like structure as depicted in Fig. 3.5. Instead, it keeps one vector as it is, e.g. the vector pointing

directly to cerebellum tissues, and forces all other vectors to be orthogonal to this vector. This procedure is then repeated, leading to a subsequent orthogonalisation of all vectors. Therefore, the vector pointing to cerebellum includes also the general brain expression pattern, while the vectors associated with all other brain regions do not contain this general expression pattern. This leads to a successively decreased information content and does not allow a proper comparison of the different brain regions due their unequal treatment. In summary, unsupervised methods are well suited to detect joint processes, whereas supervised methods are better suited to detect tissue or cell type specific expression patterns. This complementarity in the strengths of the individual methods is used in the following to combine them by a subsequent application of the two approaches.

3.4. Combining unsupervised and supervised dimension reduction techniques

3.4.1. Construction of the two-scale map

As indicated in the previous sections, the unsupervised and supervised methods have complementing strengths and weaknesses with respect to finding biologically relevant directions in the expression space. Therefore, we will now combine the unsupervised PCA with the supervised regression-based dimension reduction approach.

To this end, we use the first three PCs of the Lukk [12] dataset to define a principal components space (PCA space) and complement it by a supervised tissue specific space. The PCA space is restricted to the first three PCs since the fourth PC is already associated with an array quality metric as discussed before. Furthermore, the first three PCs of the Lukk dataset [12] and our own dataset span similar spaces (see section 3.2.2), being indicative of some robustness of the three dimensional space with respect to differences in data composition.

The supervised tissue specific space is then determined by the fold change based procedure described above, which is now applied to the PCA residual space, i.e. to the residual expression which is not captured by the three dimensional PCA. The advantage of such a decomposition into two different spaces is that much of the correlation is captured in the orthogonal PCA space. Therefore, correlations in the residual space are significantly reduced (Fig. 3.6, left).

Lukk et al. [12] hypothesise that the expression space has a very low intrinsic dimensionality, apart from more detailed differences. In this respect, we could partly confirm their results based on the analysis of an alternative large dataset. However, we hypothesise that there is significant and relevant information in the residual space that should not be neglected. Hence, we argue that it is worth to extend the analyses of [12].

To argue for the hypothesis of significant and relevant information content in the residual space, we would first like to hint to the differences between the PCA analyses of the Lukk dataset [12] and our own dataset, especially on the fourth component. Furthermore, we

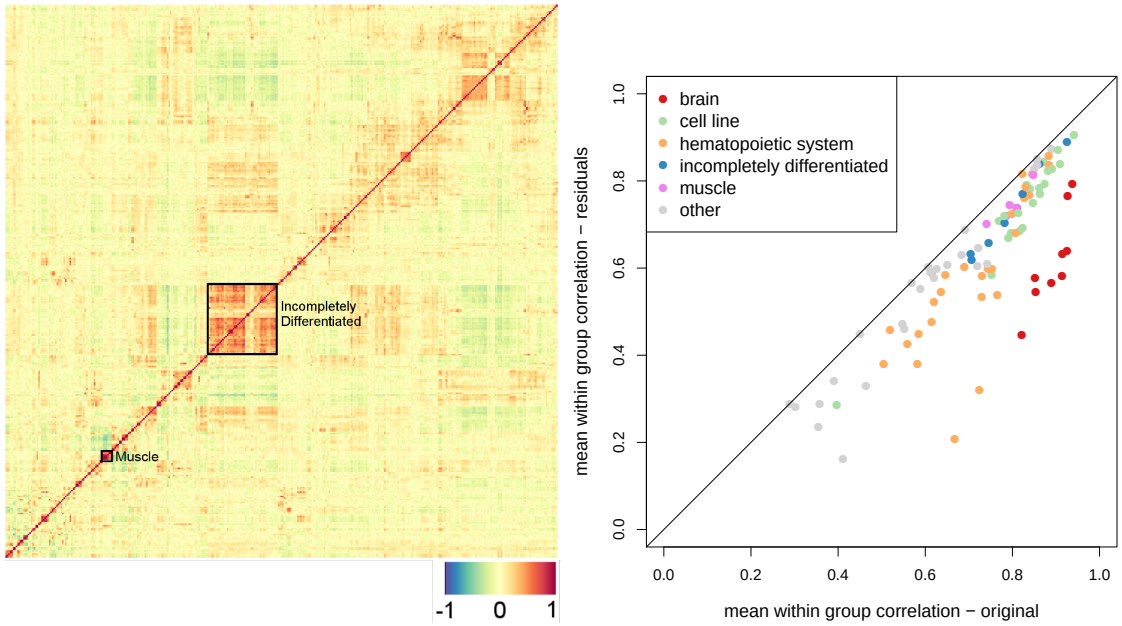


Figure 3.6.: Residual tissue specific signatures after PCA based decomposition show strongly reduced correlations between each other (left). Only very few clusters remain in comparison to Fig. 3.6 (the ordering of signatures is kept constant). In contrast, correlations between samples within a specific tissue or cell line remain high (right). Shown are the within group correlations of the 96 groups that contain at least 10 samples before (abscissa) and after (ordinate) PCA based decomposition.

could show that principle components analysis on subsets of the data is able to find more detailed structures that fit well to the annotation of cell types. Apart from that, we could explain the possible reasons for detecting measurement noise already in the fourth principle component, and showed that different choices of data can reveal other directions that exceed the noise level (Fig. 3.3, left).

Another validation of our hypothesis is depicted in Fig. 3.6 (right). It shows the correlation of different samples within individual tissue or cell types. According to the low dimensional hypothesis, one would expect that these correlations within a specific group of samples would be around zero in the residual space, since almost the entire tissue specific signal would be captured by the PCA space. However, this is not the case. Indeed, the within group correlation is only slightly reduced in the residual space (Fig. 3.6, right) and is significantly larger than zero for all of the 96 cell types with at least 10 samples that were used for this analysis. Furthermore, a manual investigation of the residual space vectors suggests that typical tissue or cell type specific genes, e.g. POU5F1, NANOG, or LIN28A for the ESC signature, are among the genes with highest residual expression for the respective tissue. Therefore, the residual space still contains important tissue or cell line specific information.
Similar results can be obtained for our own dataset (Fig. 3.7 and Fig. B.3). Therefore, it

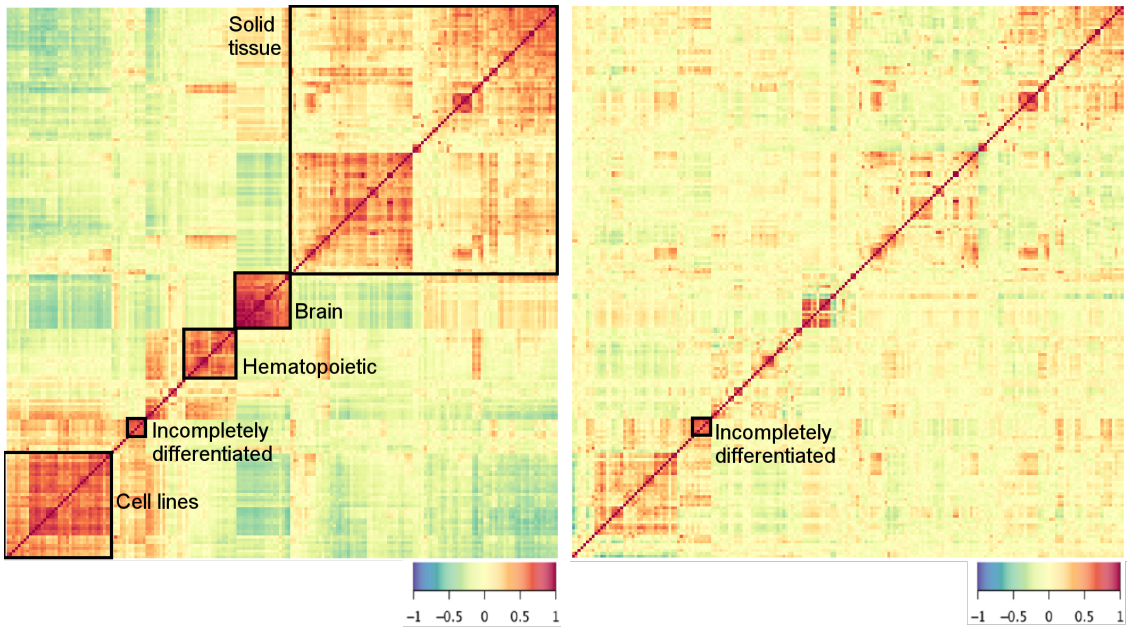


Figure 3.7.: The tissue or cell line specific signatures derived from our own dataset show similarly strong correlations as those derived from the Lukk dataset [12] (left). Correlations diminish strongly after PCA based decomposition (right).

does make sense to extend the three dimensional PCA space by a tissue specific residual space that is determined by supervised methods as follows.

Two-scale map generation using a reference dataset

Input: Reference dataset X (columns correspond to genes), factor f specifying the group membership (e.g. according to tissue/cell type)

1. Determine the gene-wise mean μ of the entire reference dataset: $\mu = \frac{1}{n}\sum_{j=1}^{n} X_j$.
2. Perform a PCA, via a singular value decomposition (svd) on the centered reference dataset: $U \cdot D \cdot V^T = \mathrm{svd}(X - \mu)$
3. Choose the first principal components that have a clear biological meaning (visual evaluation according to the available annotation, here: first 3 PCs) as PCA space: $\mathrm{PS} = (V_{\cdot i})_{i=1,2,3}^{T}$
4. Determine the residual expression: $R = X - \mu - (U_{\cdot i})_{i=1,2,3} \cdot (D_{ji})_{i,j=1,2,3} \cdot \mathrm{PS}$
5. For each factor level k of f:
   a) Determine the gene-wise mean $\tilde{\mathrm{RS}}_{k\cdot}$ of all samples corresponding to this group: $\tilde{\mathrm{RS}}_{k\cdot} = \frac{1}{n_k}\sum_{j\in\mathrm{level}(k)} R_j$.
   b) Determine the residual tissue specific space RS by a normalisation of the rows of $\tilde{\mathrm{RS}}$ to length one: $\mathrm{RS}_{k\cdot} = \tilde{\mathrm{RS}}_{k\cdot}/||\tilde{\mathrm{RS}}_{k\cdot}||$

We refer to this combination of the PCA space and the residual tissue specific space as a two-scale map of gene expression, since the two spaces capture different scales, fitting well to the concept of a tree-like structure in the expression space. The PCA space captures the joint processes, e.g. proliferation, or joint expression patterns of the brain and hematopoietic tissues/cell types. Thus, it operates rather on the middle layer in the tree-like representation (Fig. 3.5). The residual tissue specific space extends this rough separation then to the more fine-scaled differences in the lower layer of the tree structure. In the next chapter we will apply the method for the characterisation of several different microarray samples and compare the two-scale approach to the purely supervised dimension reduction. We will concentrate on the Lukk dataset [12] for this purpose as its three leading PCs offer a simpler biological interpretation, but will sometimes switch to our own dataset, when the cell types included in this dataset fit better to the respective application. Before doing so, we want to evaluate the relevant information content in the PCA space and provide further evidence that there is a lot of relevant information not captured by the PCA.

3.4.2. Estimating the amount of relevant information in the PCA and residual space via the information ratio

We have described the concept of information loss and the properties of PCA for minimisation of it (see section 3.2.1). In this section we will use a slightly different information criterion to determine the amount of relevant information that is captured by the three dimensional PCA space compared to the information in the other dimensions. The so-called 'information ratio' criterion was introduced by Schneckener et al. [83]. It determines the ratio of information with respect to a specific phenotypic difference that is contained in the residual space, i.e. in the space that is not captured by the first three or four PCs. Here, the information with respect to a phenotypic difference is measured by the negative logarithmic p-value of a t-test between the samples of two phenotypes. The ratio of information in the residual space is determined for each gene and combined as a weighted sum as described in more detail in the following.

Calculation of the information ratio (IR)

Input: Dataset X (columns correspond to genes), binary factor f specifying the group membership

1. Determine the gene-wise mean μ of the entire reference dataset: $\mu = \frac{1}{n}\sum_{j=1}^{n} X_j$.
2. Perform a PCA, via a singular value decomposition (svd) on the centered reference dataset: $U \cdot D \cdot V^T = \mathrm{svd}(X - \mu)$
3. Choose the first k principle components as projection space and determine the

projected data: $\mathrm{P} = \mu + (U_{\cdot i})_{i=1,\ldots,k} \cdot (D_{ji})_{i,j=1,\ldots,k} \cdot (V_{\cdot i})^T_{i=1,\ldots,k}$

4. Determine the residual expression: $R = X - P$
5. Calculate for the datasets X, P, and R for each gene the $\log_{10}$ p-values between the two groups specified in f using a t-test. These log-p-values are named lpX, lpP, and lpR, respectively.
6. Determine a weight $w_i = e^{-\lambda \cdot \mathrm{lpX}_i}$, for each gene i where λ is the slope of the logarithmic density of the lpX values, which typically follow an exponential distribution [83].
7. Calculate the information ration IR as

$$\mathrm{IR} = \frac{1}{\sum_i w_i} \sum_i w_i \frac{\mathrm{lpR}_i}{\mathrm{lpR}_i + \mathrm{lpP}_i}$$

One important finding reported by Schneckener et al. [83] is that, in their datasets, almost all phenotypic information with stable differential gene expression is contained in the first four PCs. This finding suggests that almost all relevant information content can be captured by four dimensions. This result would support the finding of Lukk et al. [12], stating that the overall dimensionality of the gene expression space is very low. Note that we use only the first three PCs, since the fourth PC of the Lukk dataset [12] represents noise. Using the first four PCs leads to the same conclusions (data not shown).
In stark contrast to this result, an analysis of the number of controlling nodes in the gene expression network that are required to fully control the network results in much higher numbers. In fact, it was estimated that roughly 80% of all nodes must be controlled [191], which results in a very high number for a network with roughly 20,000 genes. However, as noted in [192], this analysis does not take into account the co-regulation and feedback loops that restrict the admissible space to a rather low dimensionality. Furthermore, we can currently not expect to be able to control any node in a gene regulatory network. Therefore, we will mainly measure stable states of the gene regulatory network that correspond to attractors of the regulatory network.
In this context, it is very interesting to investigate in more detail whether we can really find information beyond the low dimensional PCA space in large gene expression datasets, using the information ratio criterion. As indicated above, this was not the case for the datasets investigated in [83]. However, their datasets were usually focused on a more narrow set of cell types or tissues, e.g. on breast cancers or lung cancers. In contrast, we analyse a very heterogeneous dataset with hundreds of different cell or tissue types. Therefore, it will be interesting to see how much relevant information is contained in the first three PCs for the Lukk et al. [12] dataset.
To this end, we determine the information ratio for comparisons of various different tissues or cell types. Each of these cell or tissue types has at least 10 samples in the dataset. We

choose the PCA dimensionality as $k = 3$, i.e. we determine the amount of information in the three dimensional PCA space compared to its complement. Some representative results of this investigation (six comparisons) are depicted in Fig. 3.8, where we compare the log10 p-values in the PCA space and its complement to those in the original data. Here, the PCA space is determined based on the full dataset and not only on the set of tissues/cell types that are compared.

These examples show, that in cases where both groups are in the same large scale group, e.g. where both are cell lines (pc3 prostate cancer vs. ssMCF7 breast cancer), both are solid tissues (heart vs. kidney), both are hematopoietic cells (B cells vs. T cells), or both are brain tissues (cerebellum vs. caudate nucleus), the main information is contained in the residual space. In contrast, when the compared groups are in different large scale groups, e.g. cerebellum vs. B cell, or ssMCF7 breast cancer vs. breast cancer, there is more information in the PCA space. This is due to the large scale differences between these groups that are captured by the PCA space. Nevertheless, there is still some significant information also in the residual space, as can be seen by the significant p-values and by the IR values that are still above 0.3. These results fit very well to our previous results, showing that the PCA space distinguishes mainly between the 6 large scale groups (hematopoietic, brain, incompletely differentiated, cell line, muscle, and solid tissue) and finer differences lie mainly in the residual space.
Furthermore, these results confirm that there is a significant amount of tissue specific information in the residual space and that at least the linear intrinsic dimensionality of the gene expression space is clearly higher than 3.

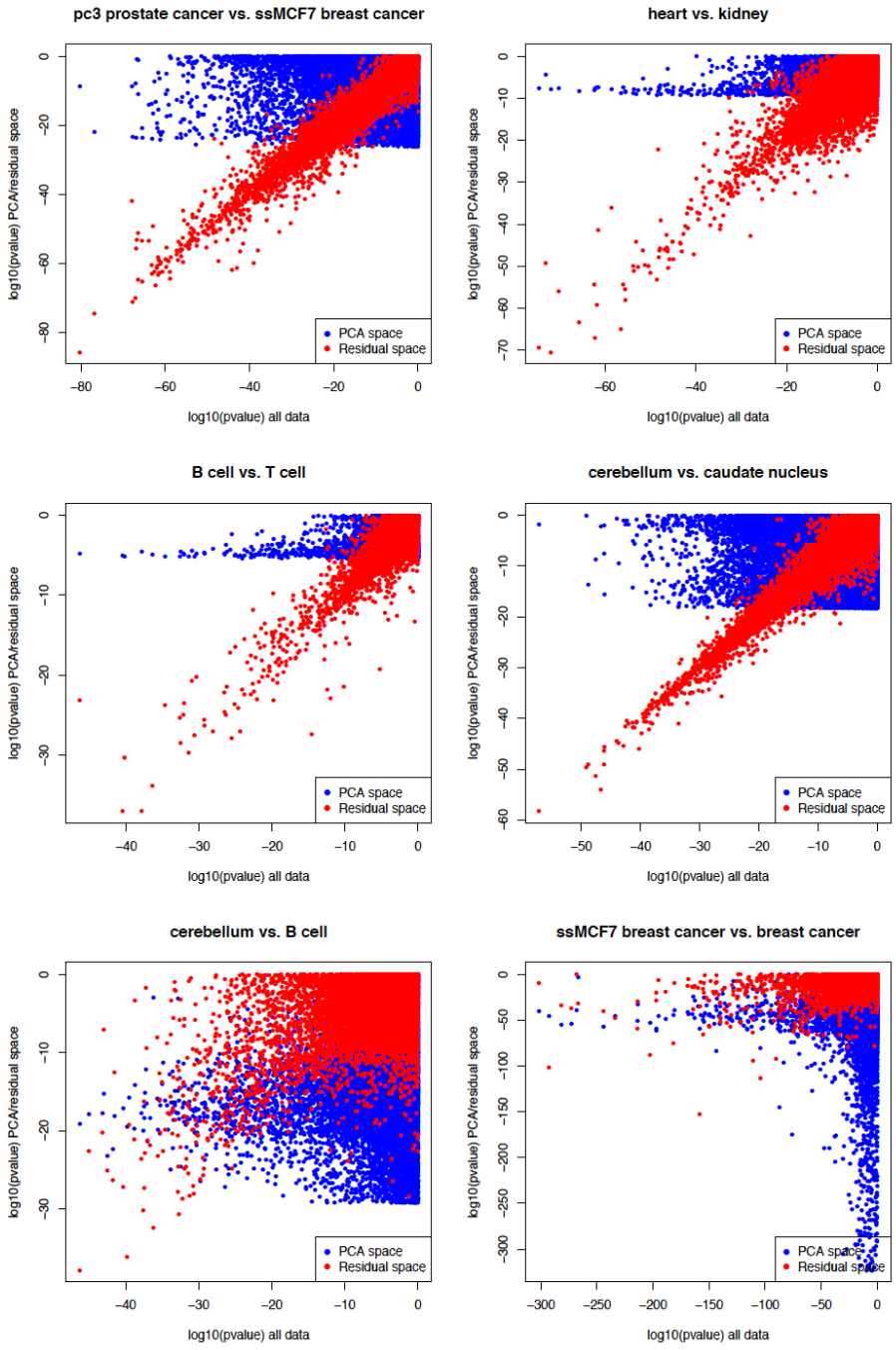


Figure 3.8.: Information partition between the PCA space and its complement for 6 different comparisons. Blue (red) points depict the log10 p-value in the PCA space (the residual space) in comparison to the log10 p-value in the original dataset. The corresponding IR values are 0.7436 (pc3 prostate cancer vs. ssMCF7 breast cancer), 0.8887 (heart vs. kidney), 0.8364 (B cell vs. T cell), 0.7188 (cerebellum vs. caudate nucleus), 0.3041 (cerebellum vs. B cell), and 0.3350 (ssMCF7 breast cancer vs. breast cancer).

4. Using the two-scale map to characterise new data

4.1. Mapping new data onto the two scale map

4.1.1. Linear mapping

Having constructed the PCA and residual tissue specific spaces, it is now possible to map new data onto these spaces in order to characterise them in terms of their location relative to well known tissues or cell lines. Such a mapping can be done in a linear way by the following two subsequent steps. First, the new data are mapped to the PCA space. This is done by subtraction of the overall mean vector from the Lukk [12] dataset and subsequent scalar multiplication with the three principal components from the PCA space. Second, the residual expression of the new data is determined and scalar multiplied with the normalised (to length one) tissue specific vectors of the residual tissue specific space.

Linear mapping of new data onto the two-scale map

Input: New data Y (columns correspond to genes), PCA space PS, residual tissue specific space RS

1. Subtract the gene-wise mean μ of the platform specific reference dataset from the data: $Y_{\text{res}} = Y - \mu$
2. Project the data onto the PCA space: $\text{PS}_Y = \text{PS} \cdot Y'_{\text{res}}$
3. Determine the residual expression: $R = Y_{\text{res}} - \text{PS}'_Y \cdot \text{PS}$
4. Project the residuals onto the tissue specific space: $\text{RS}_Y = \text{RS} \cdot R'$

4.1.2. Mapping to different microarray platforms

This mapping is only directly applicable to new samples from the same microarray platform. However, there are many different microarray platforms and new samples are usually not hybridised to the relatively old Affymetrix Human U133A array, which was used to generate the Lukk [12] dataset. Therefore, it is necessary to transform the two-scale landscape to other microarray platforms. This can be done by a matching of probe identifiers via their corresponding genes.

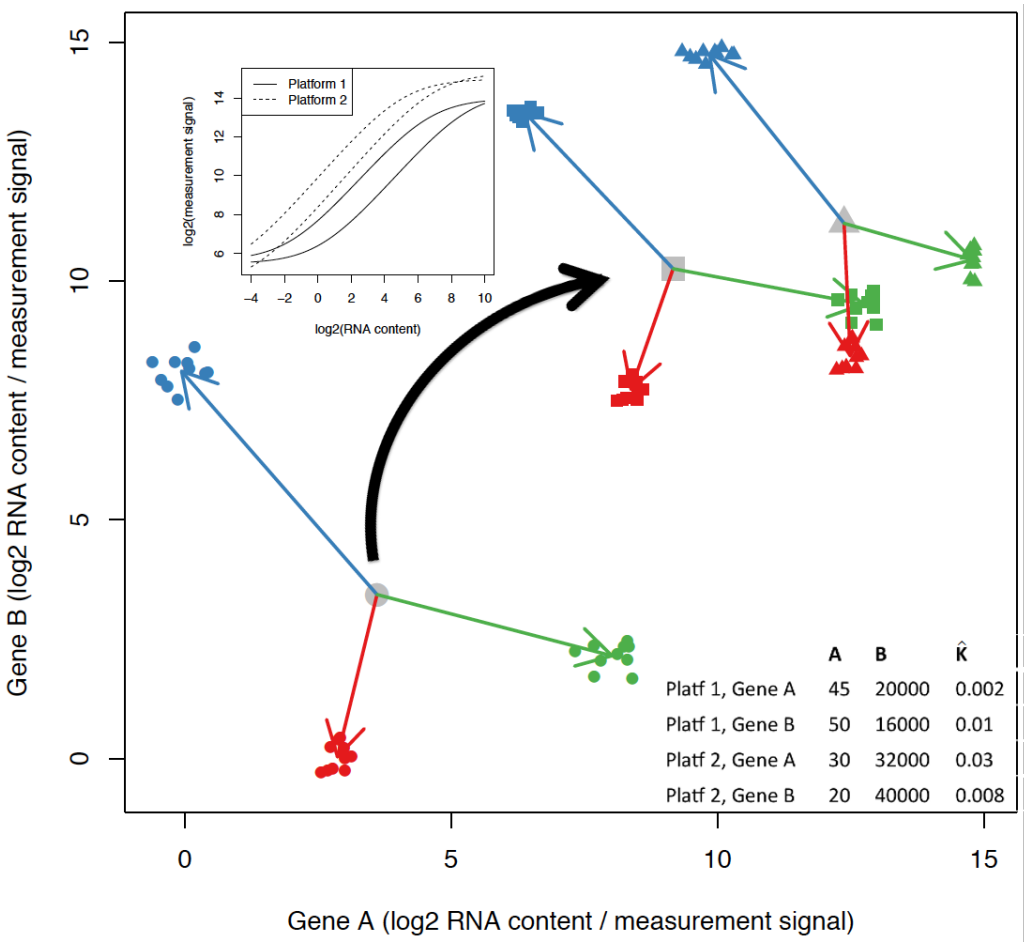


Figure 4.1.: Illustration of the differences in the expression space for measurement signals from two different platforms (squares and triangles) compared to the initial RNA content (circles). Plotted are simulated RNA contents and measurement signals (both in log2-scale) of two genes (gene A on the abscissa, gene B on the ordinate) from three different cell types (blue, green, and red). The RNA to signal intensity transformations (inset) are calculated based on the Langmuir equation (2.9) with the parameters as specified in the figure. A substantial shift in the expression values can be immediately observed, while the shortening of the arrows (due to the flattening of the hybridisation curve at both ends), as well as the slight change in their directions, is less pronounced. Grey dots represent the mean values of the RNA content (circle) and the signal intensities for the two platforms (square and triangle). The black arrow illustrates the transformation from RNA content to signal intensities via the curves in the accompanying inset.

However, absolute values are usually not directly comparable between different platforms, since differences in the parameters of the curve describing the relationship of RNA content and measurement signal (section 2.3.2) often lead to clear shifts in the expression space (Fig. 4.1). Therefore, it is important to correct for this shift. In the present thesis, this is done by exchanging the mean expression value of the Lukk [12] dataset in the above mapping procedure by the mean value of a reference dataset of the new microarray

platform. Due to the large number of different tissues and cell lines in each of the used reference datasets, the mean values should not contain large amounts of tissue specific expression. Therefore, exchanging the mean value of the Lukk dataset by the mean value of the new reference dataset should remove the platform dependent shift relatively well. A comparison to the use of the mean from the Lukk dataset showed that using the platform specific mean gives better results (data not shown). Besides the shift in the expression space, the lengths of vectors pointing to a specific tissue will also change do to differences in the RNA to measurement signal curves (Fig. 4.1) as well as differences in the number of (matching) probes on the array. However, such differences in the lengths of vectors are not critical since the values are compared to a reference dataset from the same platform anyways. What is more important is the similarity of directions on both arrays, i.e. that for a fixed underlying RNA content, the tissue specific vectors on both array types have similar directions (Fig. 4.1).
This is not critical for the almost linear phase of the log-log curve describing the relationship of RNA content and measurement signal. More critical are both ends of the curves, where non-linearities occur. However, the non-linearities have a similar shape for all investigated microarray platforms, resulting in relatively small differences in directions.
Methods for batch effect removal, e.g. using empirical Bayes methods [193], can also be used for cross-platform normalisation [194]. Such methods usually account for gene wise additive and multiplicative batch effects and thus concentrate on removal of noise. This is especially well suited to adjust measurement signals within one specific group of samples, i.e. from one specific phenotype. That is, they can be nicely used to integrate data across platforms for determination of differentially expressed genes between two phenotypes. However, such approaches rely on a relatively well balanced number of samples for each phenotype and microarray platform. In contrast, the reference datasets which we consider here may not at all contain a sample of a specific tissue which is present in the Lukk dataset. Therefore, there is an increased risk of removing valuable biological information along with the platform differences. Therefore, these methods seem not to be well suited to remove the non-linear effect seen in the RNA to signal curves.

A first explicit evaluation of the mapping across platforms is shown in Fig. 4.2, projecting the first three principal components of the Lukk [12] dataset to either the Affymetrix Human Gene 1.0 st-v1 platform (Fig. 4.2, top) or the previously described own dataset from the Affymetrix Human U133 Plus 2.0 platform (Fig. 4.2, bottom). Both projections show very nice results with a very good separation of the hematopoietic cells, brain tissues, and cell lines from all other samples.

4.1.3. A non-linear rank based mapping method with gene set selection

As an alternative to the linear mapping method, it is also possible to use non-linear mapping methods, e.g. in order to increase the influence of specific genes on the mapping or to account for non-linearities in the data. Here, we describe one non-linear mapping method

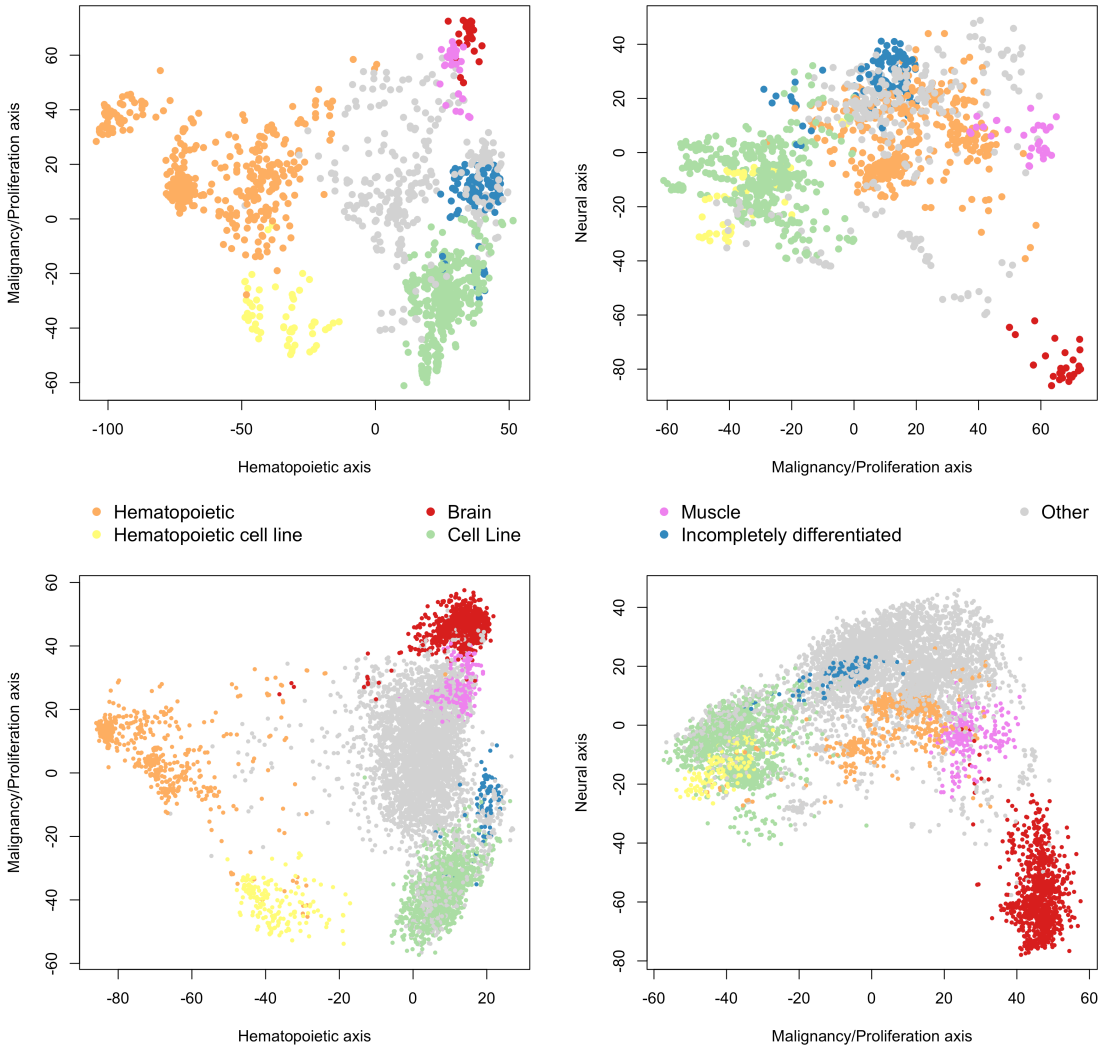


Figure 4.2.: Mapping of the first three principal component of the Lukk [12] dataset (Affymetrix Human U133A array) to two different Affymetrix platforms, namely the Affymetrix Human Gene 1.0 st-v1 (top), and Affymetrix Human U133 Plus 2.0 (bottom) platforms. Both mappings reveal a clear separation of hematopoietic cells, brain tissues, and cell lines from other samples. This confirms the possibility to robustly project the three dimensional PCA space from [12] to other Affymetrix arrays.

with gene set selection that we use as an alternative to the linear mapping.

Non-linear rank based methods with gene selection have been proposed by several groups [69, 72, 195] for determination of signature associations. These methods are similar to gene set enrichment analysis (GSEA) [63, 64, 196], but perform gene selection on the microarray data, rather than using predefined gene sets. The method used in this thesis is similar to that from Lenz et al. [72], but with a different selection of gene sets. It is implemented according to the following workflow:

Non-linear mapping of new data onto the residual tissue specific space
Input: Residuals R from the new data (columns correspond to genes), residual tissue specific space RS
For each row i of the residual tissue specific space:

1. Determine two gene sets G_T and G_B consisting of the top and bottom 1% genes of $\mathrm{RS}_{i\cdot}$, respectively
2. Calculate a Wilcoxon rank sum test between the genes in G_T and G_B for each row $R_{i\cdot}$ in the residuals of the new data
3. Determine the signed $\log_{10}$ p-value of the Wilcoxon test as mapping quantity

The performance of the linear and non-linear mappings will be compared in detail in the following section. However, some principal advantages of the linear mapping over the non-linear mapping can already be stated at this point. The linear mapping does not depend on any cutoff parameters or other choices of methodological details, preventing discussions about their best choices. Furthermore, the linearity simplifies the relative comparison of different samples, since the order of operations does not matter. That means, it does not matter whether the difference building to the other sample or the mapping is performed first due to the linearity of both operations. This property facilitates an easy comparison to multiple reference samples without a recalculation of the mapping.
Due to these practical advantages, we will focus on the linear mapping unless the non-linear mapping turns out to have significant advantages with respect to the accuracy of the results. Therefore, we will compare both kinds of mapping in the following evaluation part.

4.2. Evaluation

4.2.1. Mapping various tissues to the two-scale map

For a first evaluation of the two-scale map approach, 24 samples from the Gene expression omnibus (GEO) dataset GSE18674 are used, consisting of 22 different tissues and 2 cell lines that were hybridised to the Affymetrix Human U133 Plus 2.0 array. This array has many probes in common with the Affymetrix Human U133A array. Therefore, platform specific effects should be relatively small.

Mapping to the PCA space

Mapping of the 24 samples to the three dimensional PCA space leads already to a clear separation of several tissues. Especially brain tissues (fetal brain, cortex, cerebellum, and spinal cord) and hematopoietic tissues (bone marrow, spleen, and thymus) as well

as cell lines (HELA and SHSY5Y) are clearly separated (Fig. 4.3). Furthermore, the neuroblastoma cell line SHSY5Y reveals a certain similarity to neural tissues on the third principal component, showing that the tissue of origin has still a certain influence. A surprising observation is the tendency of the lung sample towards a hematopoietic similarity (Fig. 4.3). This may be explained by a commonly observed relatively strong immune activity in the lung [197].

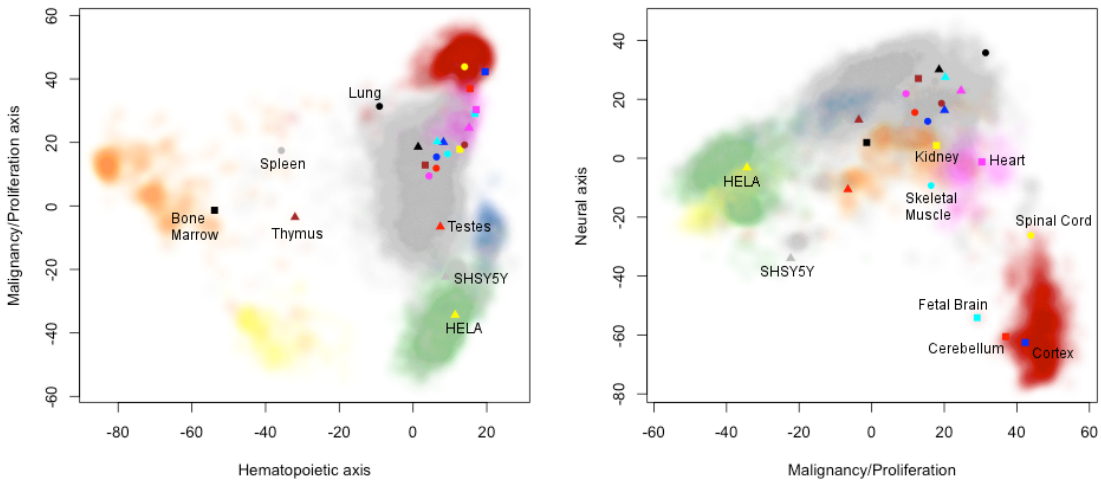


Figure 4.3.: Mapping of 24 tissues or cell lines from GSE18674 to the PCA space. Hematopoietic tissues (bone marrow, spleen, and thymus) are separated on the first principal component (PC) and cell lines (HELA, and SHSY5Y) are separated on the second PC (left). Brain tissues (cortex, cerebellum, fetal brain, and partially spinal cord) and the neuroblastoma cell line SHSY5Y have distinct expression on the third PC (right). Background colours correspond to the samples from Fig. 4.2, i.e. orange: hematopoietic, yellow: hematopoietic cell line, red: brain, green: cell line, magenta: muscle, blue: incompletely differentiated, grey: other

Mapping to the residual tissue specific space

Among the 22 tissues, several can be expected to have very similar expression values due to the similar functions of these tissues, e.g. different regions of the brain or tissues from the colon and the small intestine. It is therefore of special interest how good these tissues can be separated in the residual tissue specific space. The result of the linear mapping of the data to the residual tissue specific space is depicted in Fig. 4.4. This mapping shows a very specific classification into the different tissues. Notably, even different brain regions, i.e. cerebellum (second column), cortex (fourth column), and spinal cord (15th column) show clear differences in corresponding signatures of the residual tissue specific space. The cerebellum sample is specifically associated with the cerebellum signature, the cortex sample is associated with the brain signature, and the spinal cord sample is most

strongly associated with the hypothalamus signature.

Sample	**Two-scale linear**	**One-scale linear**	**Two-scale non-linear**
Bone marrow	Bone	Bone	Chronic myeloid leuk.
Cerebellum	Cerebellum	Cerebellum	Cerebellum
Colon	Colon mucosa disease	Colon mucosa disease	Colon mucosa disease
Cortex	Brain	Brain	Brain bipolar disorder
Fetal brain	Brain	Brain	Brain
Heart	Heart disease	Heart disease	Heart disease
Kidney	Kidney epithelium	Kidney epithelium	Kidney epithelium
Liver	Cirrhosis	Cirrhosis	Cirrhosis
Lung	Lung	Lung	Lung
Pancreas	Cirrhosis	Cirrhosis	Kidney
Prostate	Prostate gland	Prostate gland	Prostate gland
Salivary gland	Thyroid gland	Thyroid gland	Oropharynx
Skeletal muscle	Skeletal muscle dis.	Skeletal muscle dis.	Skeletal muscle
Small intestine	Small intestine	Small intestine	Small intestine
Spinal cord	Hypothalamus	Hypothalamus	Hypothalamus
Spleen	Lymph node	Lymph node	Lymph node
Stomach	Esophageal adenoc.	Lung	Bladder mucosa
Testes	Testis	Testis	Testis
Thymus	Thymocyte	Thymocyte	Thymocyte
Thyroid	Thyroid adenoc.	Thyroid adenoc.	Thyroid gland
Trachea	Bronchial epithelia	Esophagus Barretts	Bronchial epithelia
Uterus	Myometrium	Myometrium	Myometrium
HELA	HeLa cervical adenoc.	HeLa cervical adenoc.	HeLa cervical adenoc.
SHSY5Y	SK-N-SH neuroblast.	SK-N-SH neuroblast.	SK-N-SH neuroblast.

Table 4.1.: Tissue signatures with highest values for the 24 different tissues/cell lines from GSE18674. For several tissues (bone marrow, colon, fetal brain, pancreas, salivary gland, spinal cord, spleen, stomach, and trachea) there is no directly matching signature in the tissue specific space. Note that myometrium is the middle layer of the uterine wall and thus represents a correct identification of the uterus sample. Furthermore, the SHSY5Y cell line was derived from the SK-N-SH cell line to which it has the highest association.

The fetal brain sample (5th column) has, besides signals in the brain and cortex signatures, also some association with neuroblastomas, which can be attributed to the more immature, incompletely differentiated status of fetal brain compared to adult brain.
The SHSY5Y neuroblastoma cell line shows also a specific signal in neuroblastoma signatures, being most strongly associated with the SK-N-SH neuroblastoma cell line, from which it was originally derived [198].
The other tissues and cell lines can also be correctly identified in many cases (Table 4.1), but for several tissues there is no matching tissue signature in the Lukk [12] dataset, preventing the possibility of correct identification. This issue can potentially be resolved by the use of different datasets or by adding further data to the Lukk dataset. For example,

using the own dataset instead of the Lukk dataset leads to a correct identification of all 22 tissues (Fig. B.4). However, it has to be admitted that the data from GSE18674 are already part of the own dataset and were therefore used in the generation of this alternative two-scale map.

Comparison to the one-scale approach

Another interesting point is the comparison of the two-scale and the one-scale approach, i.e. with and without PCA based decomposition. Use of the one-scale approach on the Lukk dataset results in a less detailed and specific distinction of different brain regions (Fig. 4.5, note the different selection and order of the visualised tissue signatures, i.e. rows), although the top scoring signatures are in most cases the same (Table 4.1). This less clear distinction reflects the considerably higher correlation in the tissue specific space without PCA based decomposition observed above (Fig. 3.4). Therefore, we will focus on the two-scale approach in the remaining part of this chapter.

Linear versus non-linear mapping

The non-linear mapping has similar top scoring signatures as the linear mapping (Table 4.1), with some slight differences mainly for those samples that do not have a directly matching signature in the two-scale map, e.g. pancreas or stomach. However, a detailed visual inspection of the heat map depicting the non-linear mapping to the residual tissue specific space suggests an even clearer distinction of the top scoring signatures from the less associated ones (Fig. 4.6). This could be interpreted as an advantage of the non-linear mapping.

A more quantitative investigation of the differences between the linear and the non-linear mapping reveals interesting insights into their relation. Two different effects can be observed in this quantitative investigation that explain the visual advantages of the non-linear mapping. First, the score obtained by the non-linear mapping can be nicely represented by a signed quadratic dependence on the linear mapping in a certain range (Fig. 4.7, Fig. B.5, appendix A). This can be attributed to the almost quadratic relationship between the test statistic and the log10 p-value for normally distributed test statistics (appendix A) and would suggest that there is an almost linear relationship between the Wilcoxon test statistic and the linear mapping in a certain range. The quadratic relationship leads to a relative suppression of scores with a lower test statistic compared to those with a higher test statistic, resulting in a seemingly more specific distinction of the respective signatures. The second effect, which can be seen in the quantitative investigation, is a saturation effect. The Wilcoxon test has a certain minimal possible p-value due to the rank based test statistics. This results in a maximum absolute value of roughly 37 for the non-linear mapping in the present example (horizontal line in Fig. 4.7 and Fig. B.5; appendix A). This saturation may be interpreted in such a way that the non-linear mapping provides information about the similarity of directions, neglecting information about the length of the vector showing

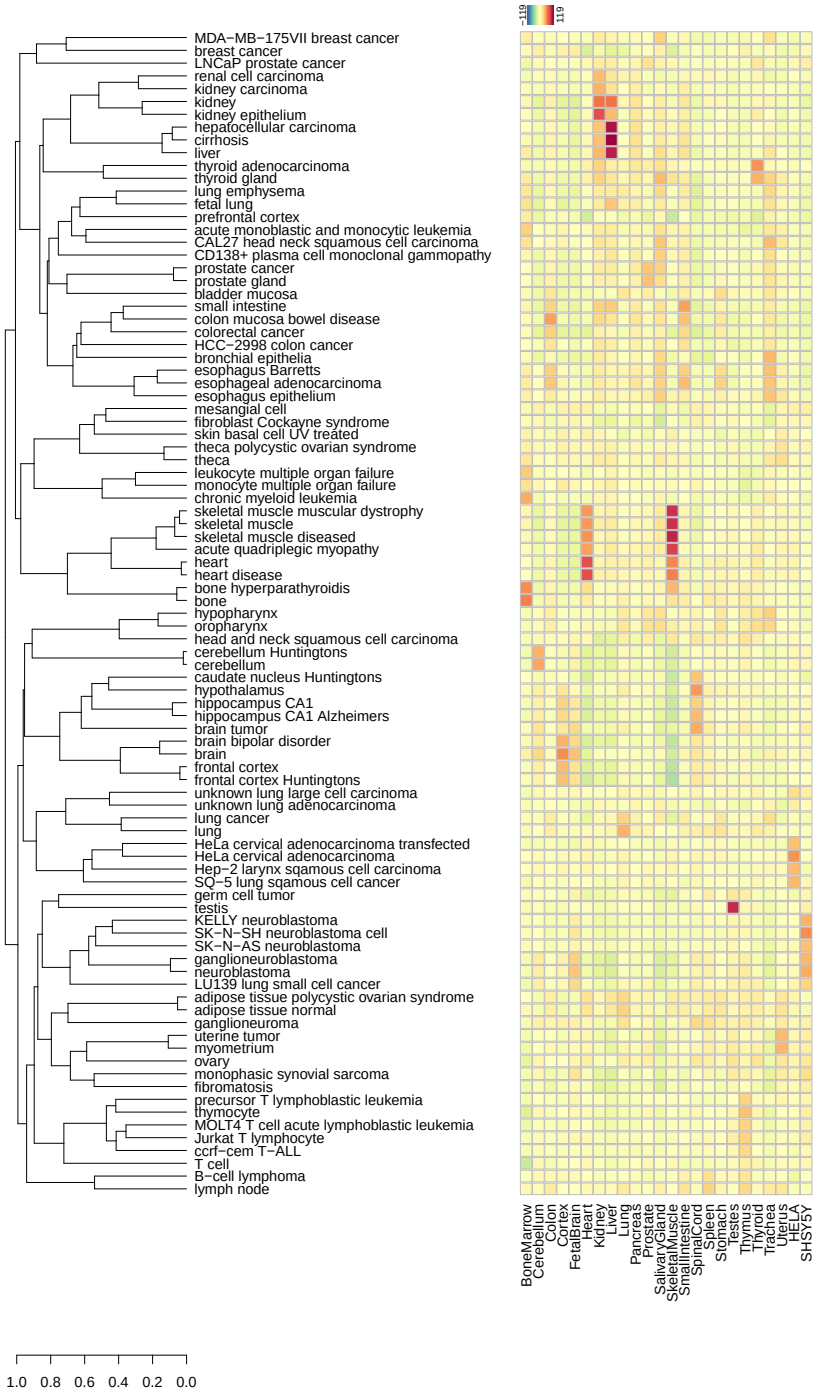


Figure 4.4.: Linear mapping of the 24 tissues or cell lines from GSE18674 (columns) to the residual tissue specific space of the Lukk [12] dataset (rows). Visualised is a selection of the 369 tissue specific signatures, including the most significant ones (see section A). Colours range from blue and green (negative association with the signature) over yellow (no association) to orange and red (positive association) as indicated by the colour bar at the top.

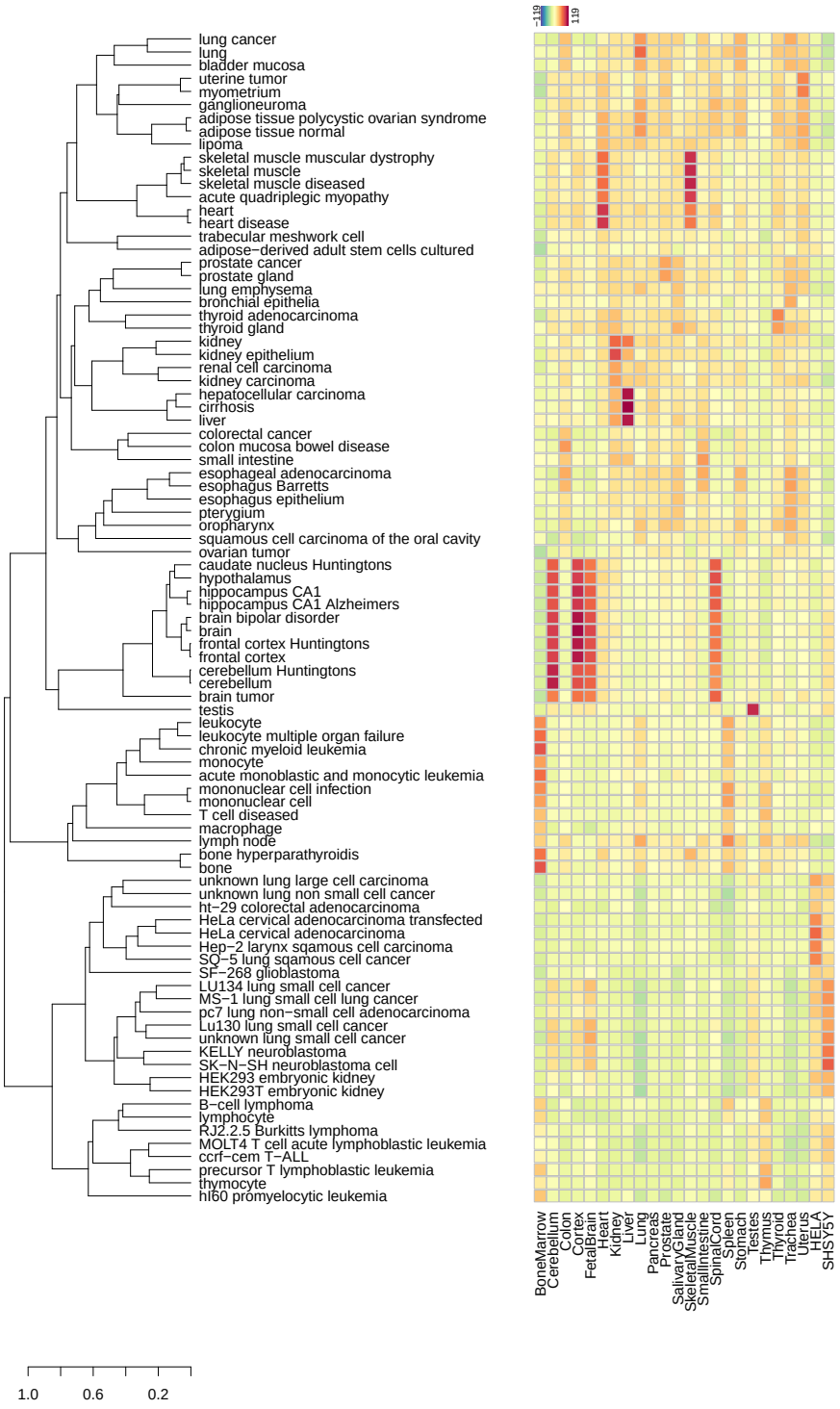


Figure 4.5.: Linear mapping of the 24 tissues or cell lines from GSE18674 (columns) to the tissue specific space without PCA based decomposition (rows). The heatmap clearly reveals the higher correlations between different signatures compared to Fig. 4.4, especially between those corresponding to different brain regions.

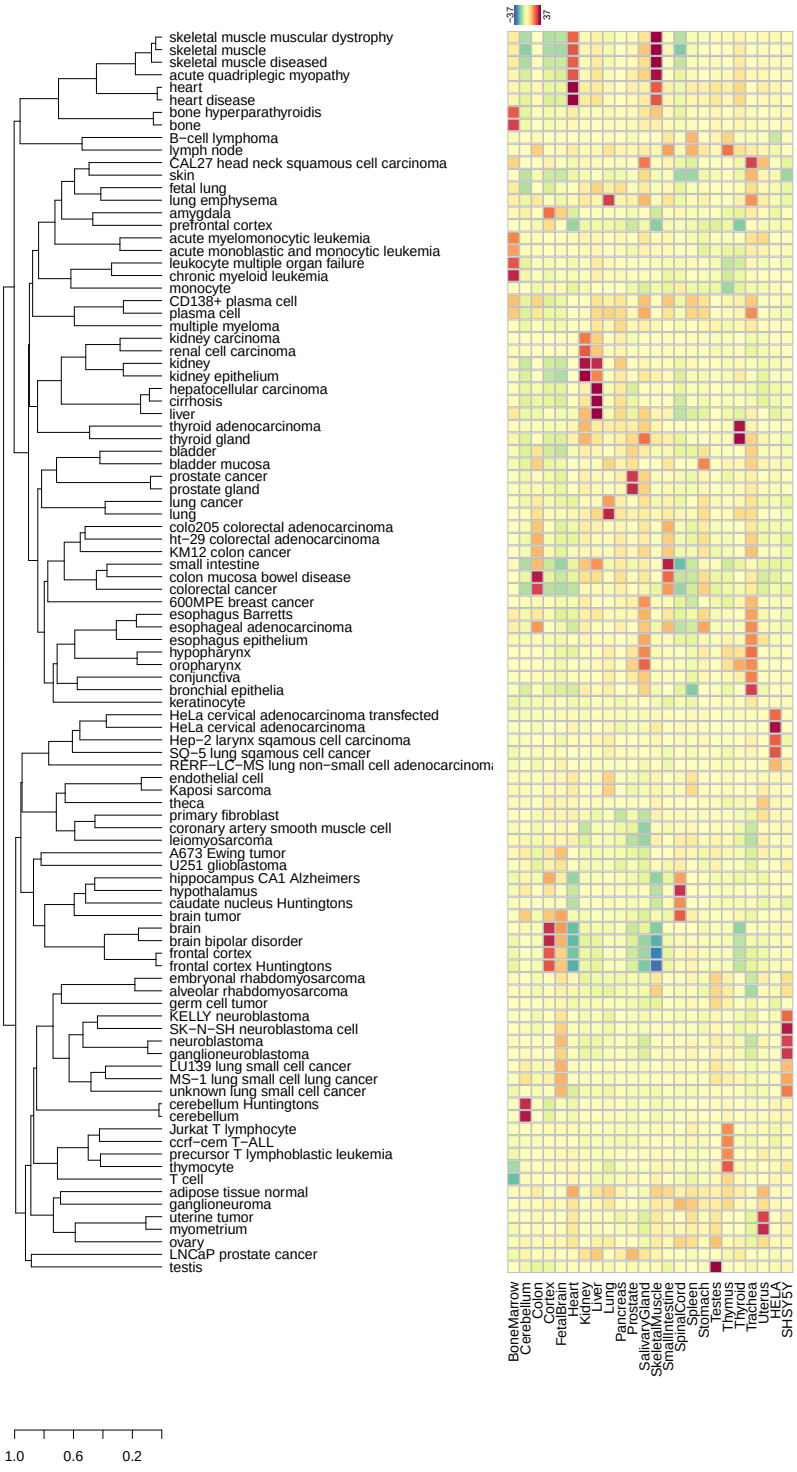


Figure 4.6.: Non-linear mapping of the 24 tissues or cell lines from GSE18674 (columns) to the residual tissue specific space (rows). The results are more distinct as for the linear mapping, with a higher difference between best matching signatures and more weakly associated ones.

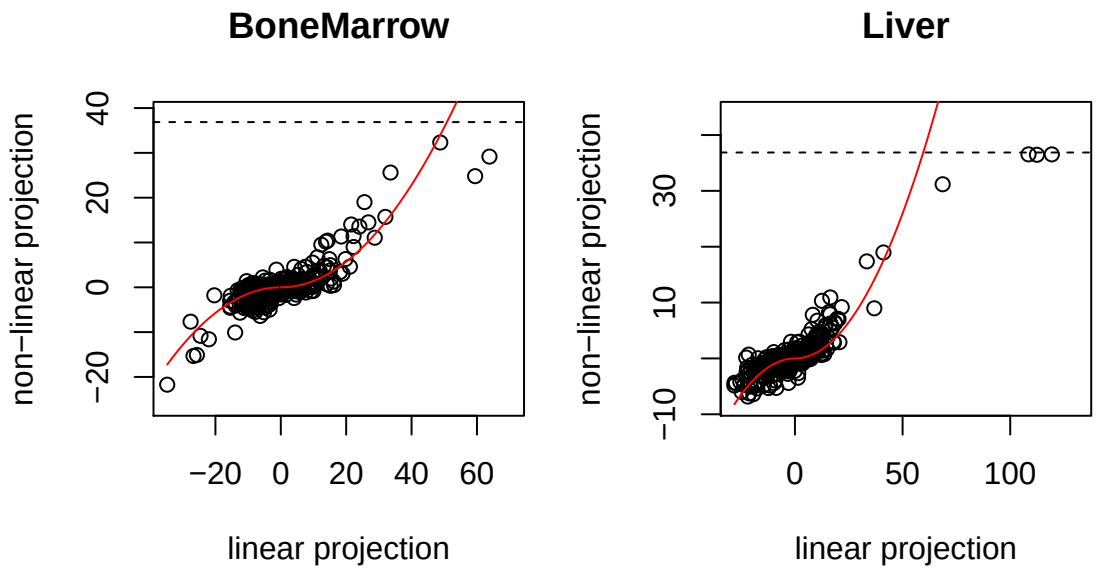


Figure 4.7.: Comparison of the linear and non-linear mapping methods exemplified at a bone marrow and a liver sample from dataset GSE18674. The non-linear mapping can in a certain range be approximated by a sign-preserving quadratic transformation of the linear mapping. However, at higher values a saturation effect can be observed due to the p-value limitations of the Wilcoxon rank sum test for fixed sample sizes (here number of probes).

in the respective direction (appendix A). Thus, the maximal value corresponds to an exact match of the directions, limiting the similarity score to a certain value.
The effect of this limited values on the heat map representation in Fig. 4.6, is that the maximal values in each column are more similar to each other since they are all limited by the value of 37. In contrast, for the linear mapping the maximal values of the samples (columns) differ more strongly, resulting in a less pronounced colouring for some of the samples. The linear mapping can be adjusted to determine similarity of directions by a division through the length of the residual data vector. This introduces also a limitation in the values and leads to more similar maximal values between the columns in the heat map (Fig. B.6).
In summary, for the 24 tissues and cell lines from dataset GSE18674, the linear and non-linear mappings result in similar rankings of the tissue signatures and visual advantages of the non-linear mapping can be removed by non-linear transformations of the linear mapping scores. However, this is not generalisable to all data, as can be shown by the examples in the next section.

4.2.2. Liver and breast cancer mixture data

So far the considered samples had a very clear association with single signatures due to the purity of the tissues. However, cell preparations may be mixtures of different cell

or tissue types, or may have only a relatively weak association with a specific signature. Mixture samples are a good means to check the behaviour of the two-scale map in these circumstances, since the true composition of the sample is still known, while the association with a specific tissue may be weak, e.g. due to a relatively low mixture fraction of this tissue. In the following, two different mixture datasets are discussed. One mixture of liver and breast cancer tissues which is investigated in the present section and one of heart and brain tissues, which will be investigated in the next section.
Dataset GSE33116 contains several samples of mixed liver and breast cancer tissues with mixture fractions of 0%, 5%, 10%, 15%, 20%, 25%, 50%, 75%, 80%, 85%, 90%, 95%, and 100% liver and the reverse amounts of breast cancer tissue. There are several replications of these mixture fractions in the dataset, but we consider only a single sample per mixture fraction (see appendix A for more details). The mixed RNA of this dataset has been hybridised to the Affymetrix Human U133A array, i.e. the same array which was used for the Lukk [12] dataset. Therefore, the results are not influenced by any platform dependent effects.

Mapping to the PCA space

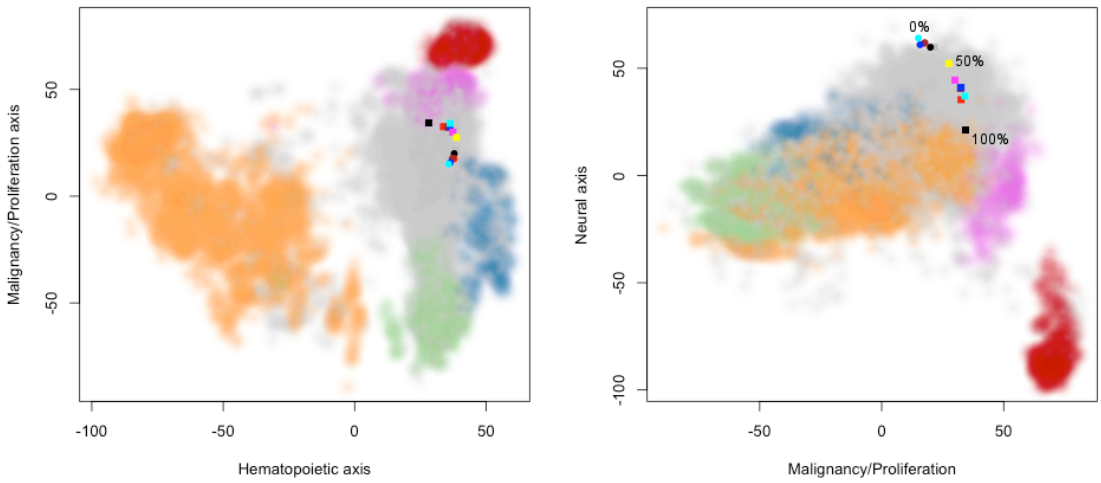


Figure 4.8.: The PCA map of the liver and breast cancer mixtures shows slight changes with changing mixture fractions. Most significant changes can be observed on the third principal component, where breast cancer samples (0% liver) lie at relatively extreme values, while liver samples (100%) lie in the middle range.

The mapping of these data to the PCA space shows a monotone curve on the neural axis (Fig. 4.8), where the breast cancer sample is on the opposite site of the brain tissues and the liver sample lies in the middle of the non-neural tissues. The extreme value of the breast cancer sample on the neural axis may indicate that the neural axis does not only have an association with brain tissues, but also partly with breast cancer tissues, though

with opposite direction. This is a general disadvantage of unsupervised methods, such as PCA, that there might be some minor associations to certain tissues or cell types that are hard to detect. That means, that the principal components may be associated not only with a single tissue or general process, but with multiple tissue at the same time, where the association to each individual tissue can be relatively weak. This problem is not very pronounced in the first three principal components in the present case, but is significantly more critical in further PCs.

Mapping to the residual tissue specific space

More interesting is the mapping to the residual tissue specific space, since this space contains directions that are directly associated with breast cancer and liver. The linear mapping, reveals a clear association of the samples with high liver fraction to the liver, cirrhosis, and hepatocellular carcinoma signatures, with a monotonously decreasing association for decreasing liver fractions (Fig. 4.9). The association to the breast cancer signature is much less clear, but a slightly increasing score with increasing breast cancer fraction can be observed as well. There may be several reasons for this comparably weak signal, including the generally high heterogeneity of cancer tissues or a relatively high similarity of the breast cancer samples to many other samples in the Lukk [12] dataset, resulting in a relatively short residual vector to the breast cancer samples. The latter explanation can be approximately quantified by a determination of the distance of the mean liver and mean breast cancer samples from the three dimensional PCA space. This distance is five times higher for the mean liver sample than for the mean breast cancer sample. Therefore, the linear projection onto the liver signature is substantially higher than for the breast cancer signature.

Comparison to the one-scale approach

This difference in signal strengths is reduced in the tissue specific space without PCA based decomposition, where we take the distance to the centre instead of the three dimensional space. Compared to the previous case with PCA based decomposition, the distance of the mean liver sample stays almost equal, while the distance of the mean breast cancer sample doubles. This reflects the observation from above that some breast cancer specific expression may already be contained in the neural axis of the PCA space, capturing some of the breast cancer specific expression. Consequently, the breast cancer signal is more pronounced in the tissue specific space without PCA based decomposition (Fig. B.8), depicting an advantage of the one-scale approach. However, there are also considerably higher associations to other cancer signatures in the one-scale approach. Furthermore, the breast cancer signature is even slightly associated with the 100% liver sample in the one-scale approach. This association is due to a positive correlation between liver and breast cancer signatures, which also partially explains the already relatively strong breast cancer signal for low breast cancer mixture fractions.

In contrast to that, an opposite effect can be observed in the residual tissue specific space of the two-scale map (Fig. 4.9), where a slightly negative breast cancer score is present for the 100% liver sample. Thus, the relative change in the breast cancer score from 100% liver to 0% liver is similar for both approaches, but the absolute values differ more strongly.

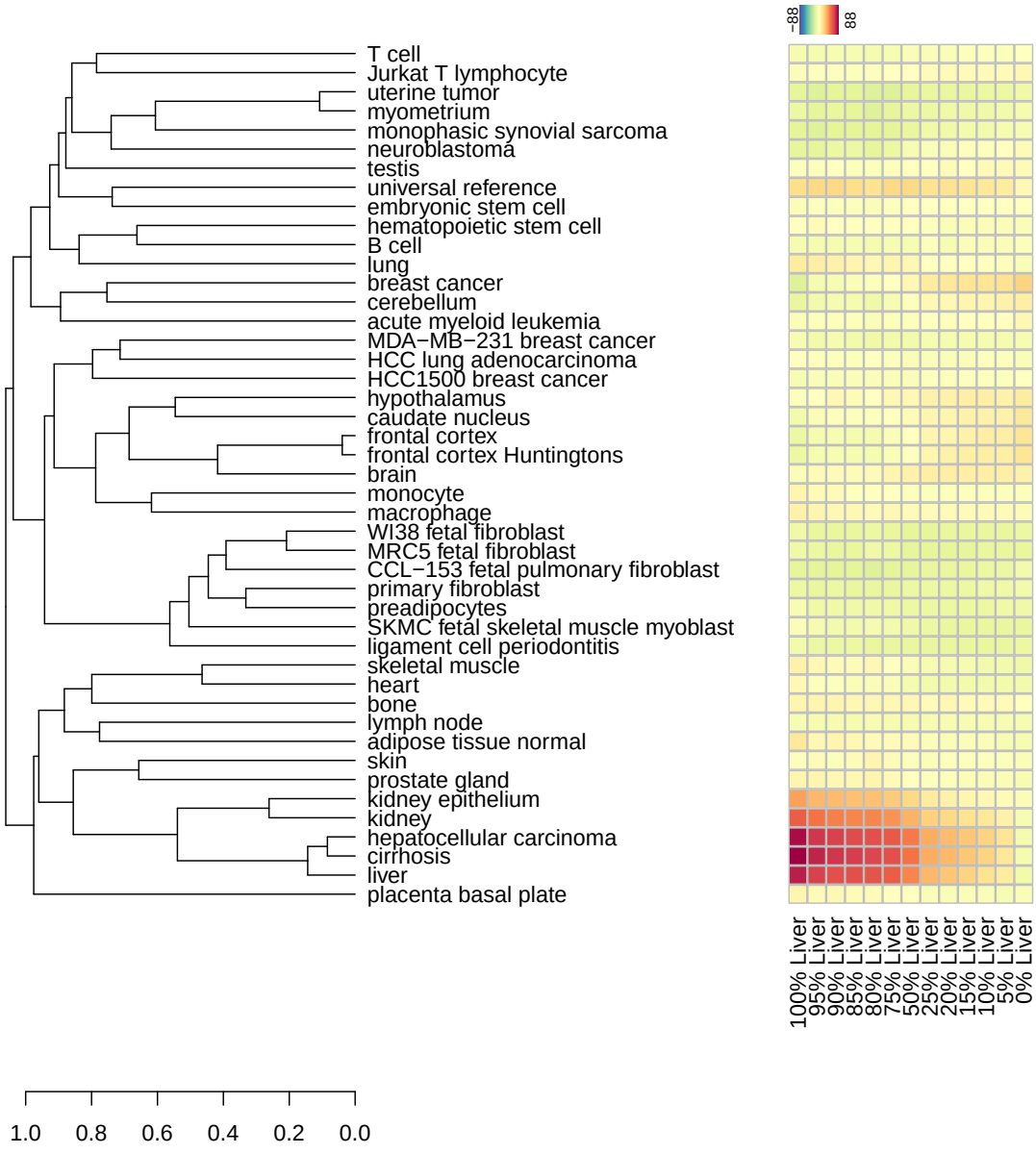


Figure 4.9.: Linear mapping of the mixtures of liver and breast cancer tissues (GSE33116) to the residual tissue specific space. The monotonous decrease of the liver association is paralleled by a slightly increasing association with the breast cancer signature.

Linear versus non-linear mapping

Interestingly, for this mixture example, the non-linear mapping shows a much stronger association with the breast cancer signature than the linear mapping (Fig. 4.10).

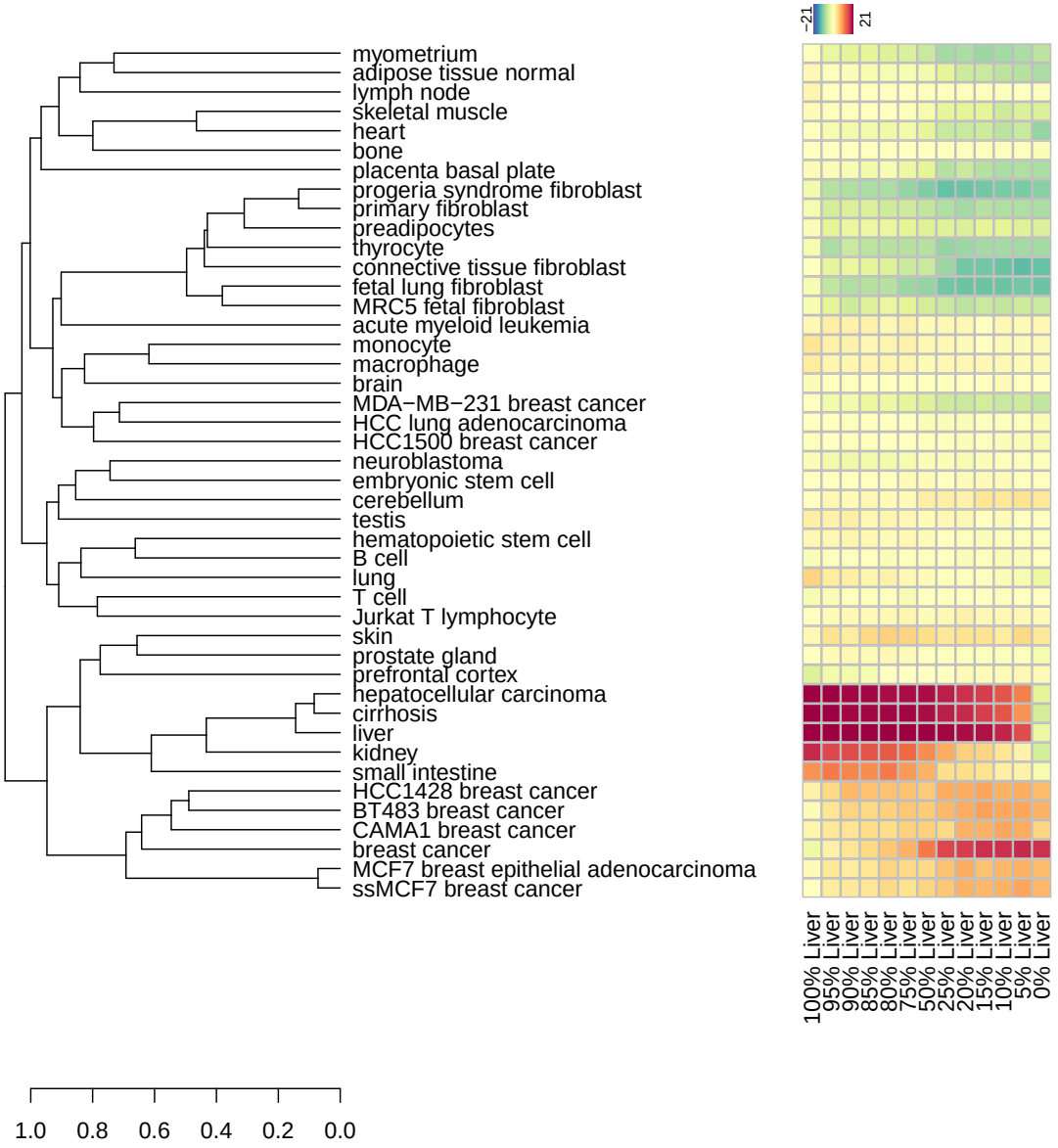


Figure 4.10.: The non-linear mapping of the liver and breast cancer mixtures to the residual tissue specific space shows a considerably stronger signal for the breast cancer signature as the linear mapping (Fig. 4.9).

Furthermore, a very strong association with the liver signature can be already detected for the sample with only 5% liver RNA. Thus, according to this example, the non-linear mapping is much better suited to detect very low fractions of specific cell types or tissues. A similar result can be seen for the non-linear mapping applied to the one-scale tissue

specific space, i.e. without PCA based decomposition (Fig. B.9). However, as for the linear mapping, the increased correlations to other tissues leads to a positive signal in other cancer signatures. Furthermore, an interesting observation is that these associations to other primary cancer tissues are stronger than the association to breast cancer cell lines in the one-scale approach, whereas it is the other way around in the two-scale approach. That means, the second highest association of the 100% breast cancer sample is with the lung cancer primary tissue in the one-scale approach. In contrast, in the two-scale approach, the second highest association is with a breast cancer cell line. The reason for this is the cell line specific expression which is captured in the second principal component. Therefore, the two-scale map can better detect similarities between cancer cell lines and their tissue of origin by removing cell line specific effects.
In contrast to the good performance for detection of tissues with small mixture fractions, the non-linear method is not well suited for the estimation of mixture fractions, especially for relatively large fractions. This can be exemplified at the example of the liver score, which does hardly change between the 20% and 100% liver mixture fractions. This is also reflected in Fig. 4.11, indicating that the liver and breast cancer scores have only slight non-linear dependence on the mixture fractions for the linear mapping, while there is a significantly more pronounced non-linearity for the non-linear mapping.

The main reason for these non-linearities is the logarithmic transformation of the measurement signals, leading to sometimes very strong non-linear relationships between the mixture fraction and the logarithmic measurement signals on which the calculations are done (Fig. 4.12). This effect is additionally enhanced by the saturation effect in the relationship between RNA content and measurement signal (section 2.3.2). It is generally stronger for genes with a high fold change between the two mixed tissues than for genes with a low fold change (Fig. 4.12). Therefore, it is not very surprising that the non-linearity is considerably stronger for the non-linear mapping, which concentrates on the genes with highest tissue specific expression. Due to the different strengths of the non-linear effect, depending on the fold change of the respective gene between the two mixed tissues, there is some non-linear structure in the residuals after mapping of the mixed samples to both spaces (Fig. 4.13). That means, that the mixture samples depend non-linearly on the respective tissue scores. This effect is not present for the pure tissue samples with 100% liver or 100% breast cancer tissue (Fig. 4.13). Due to this non-linear effect, there are the observed large differences between the linear and non-linear mapping approaches.

4.2.3. Heart and brain mixture data

Similar to the liver and breast cancer mixture data from the previous section, this section considers Affymetrix mixture data with RNA of heart and brain samples mixed at different mixture fractions and subsequently hybridised to a microarray [199]. In contrast to the previous section, the samples were hybridised to the Affymetrix Human Gene 1.0 st-v1 array. Thus, platform specific differences may obscure the results and it will be interesting

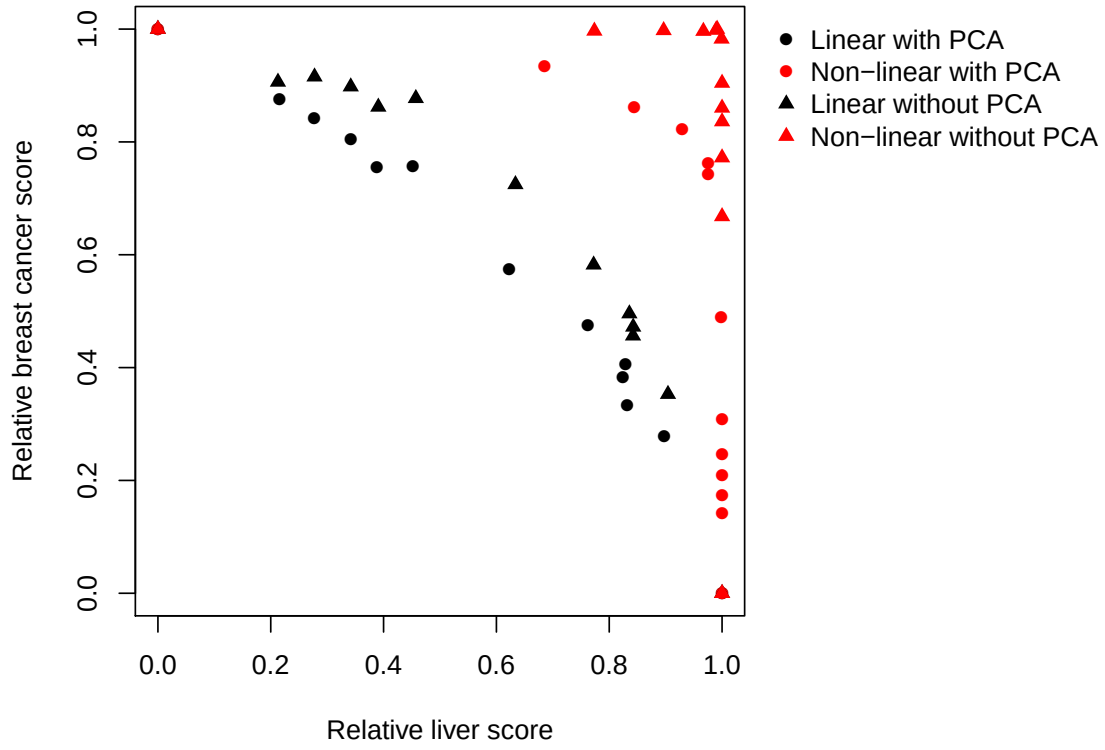


Figure 4.11.: Comparison of different mapping approaches for the liver and breast cancer mixture data. The linear mapping with PCA based decomposition is closest to a linear relationship between the liver (abscissa) and breast cancer (ordinate) signatures. The values for both signatures were normalised to have values between 0 and 1. The maxima of both scores differ strongly for the linear mapping with PCA (Fig. 4.9), while they are more similar for the other mappings (Fig. 4.10, B.8, and B.9)

to see whether the mixtures can be identified correctly.
The investigated dataset contains samples with a mixture fraction of 0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100% brain tissue and the reverse amounts of heart tissue. Each mixture fraction is available in triplicates, except of the 50% mixture, for which nine replicate measurements are available.

Mapping to the PCA space

The mapping to the three dimensional PCA space is of high interest for this dataset, since the third principal component is mainly associated with brain tissues, which are part of the mixture data. The result of the mapping is depicted in Fig. 4.14, showing a smooth transition from the magenta area corresponding to heart tissues to the red area corresponding to brain tissues, with very little variation between replicate measurements. Furthermore, a slight non-linearity in the distance between different mixture fractions can be observed on the neural axis, due to the unproportional high change from the 0%

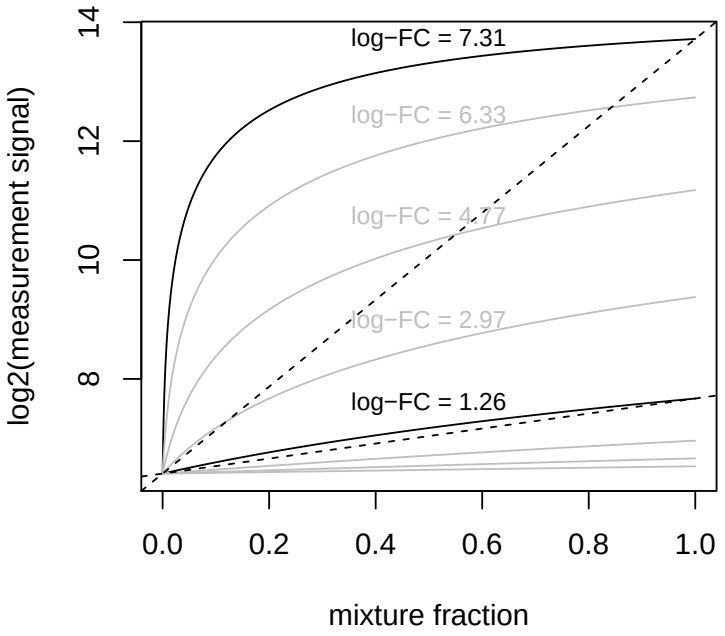


Figure 4.12.: Explanation of the non-linearity found in the liver and breast cancer mixture data (Fig. 4.13). Due to the non-linear RNA to measurement signal functions, combined with the applied logarithmic transformation (section 2.3.2), a linear increase in the RNA of a specific gene leads to a non-linear increase in the measurement signal. The non-linearity becomes considerably stronger for genes with a higher fold change between the two mixed tissue types. Solid lines represent an example of log2-measurement signals for different simulated genes over the mixture fraction. Dotted lines represent the approximation made by the linear mapping approach. Differences between solid and dotted lines thus depict the residual value, which increases with increasing fold change.

to the 5% brain tissue sample. This non-linearity can be explained by the logarithmic measurement signal as discussed in the previous section. The slight non-linearity on the malignancy/proliferation axis may have a similar reason.

Mapping to the residual tissue specific space

The linear mapping to the residual tissue specific space also results in a smooth increase and decrease of the heart and brain signature scores, respectively (Fig. 4.15). The variability between replicates is again very small. Similar as for the mapping of the 24 tissues, the two-scale approach succeeds to distinguish relatively well between different brain regions, while the one-scale approach suffers from the strong correlations between these regions (Fig. B.10).

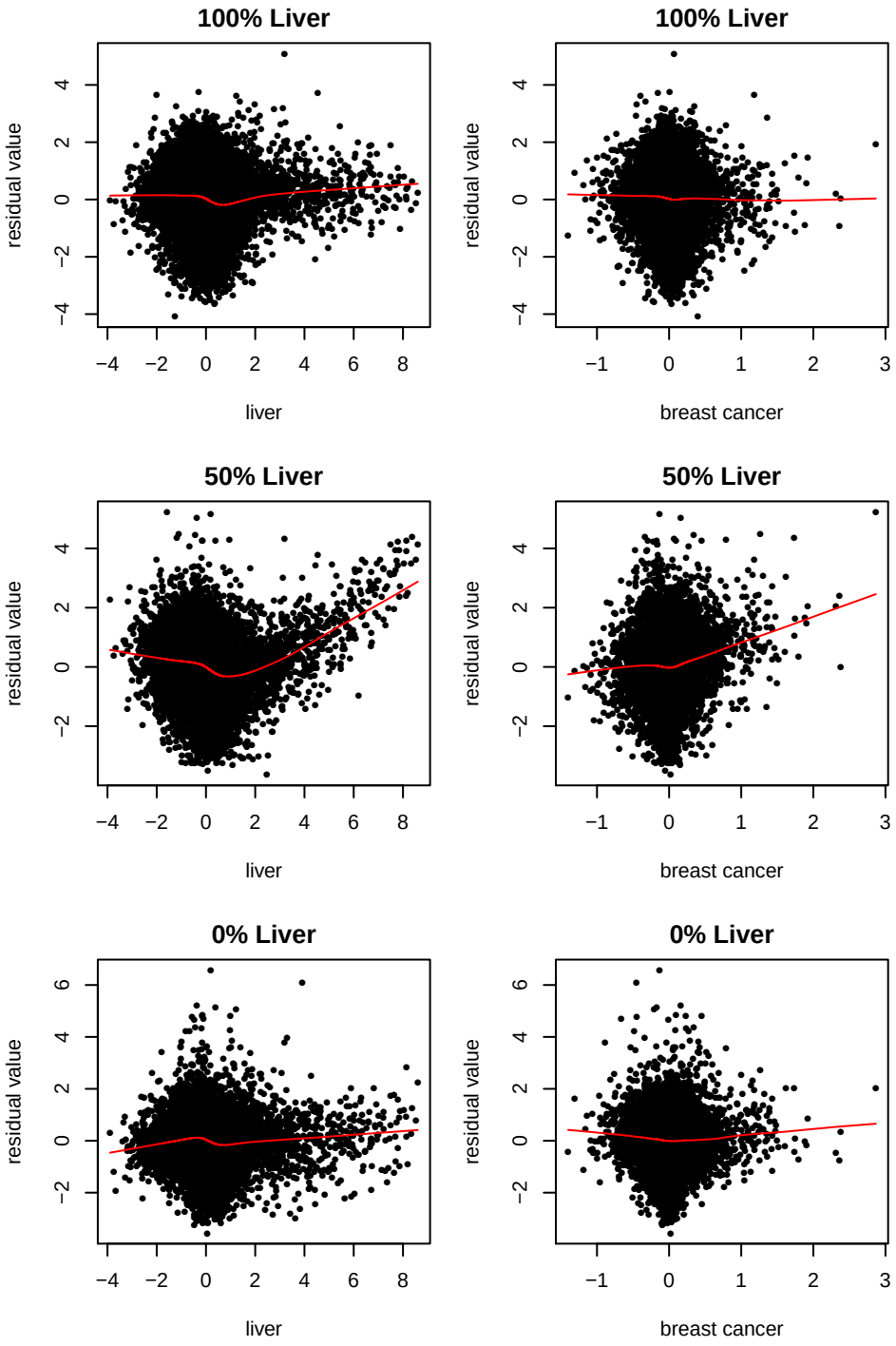


Figure 4.13.: Non-linear dependence of liver and breast cancer mixture data on corresponding signatures. Depicted is the residual value after projection to both the PCA and residual tissue specific spaces versus the liver or breast cancer signature for different mixture fractions. A clear non-linearity can be seen in the 50% mixture sample, while such an effect is not visible in the pure, i.e. 100% liver and 100% breast cancer samples.

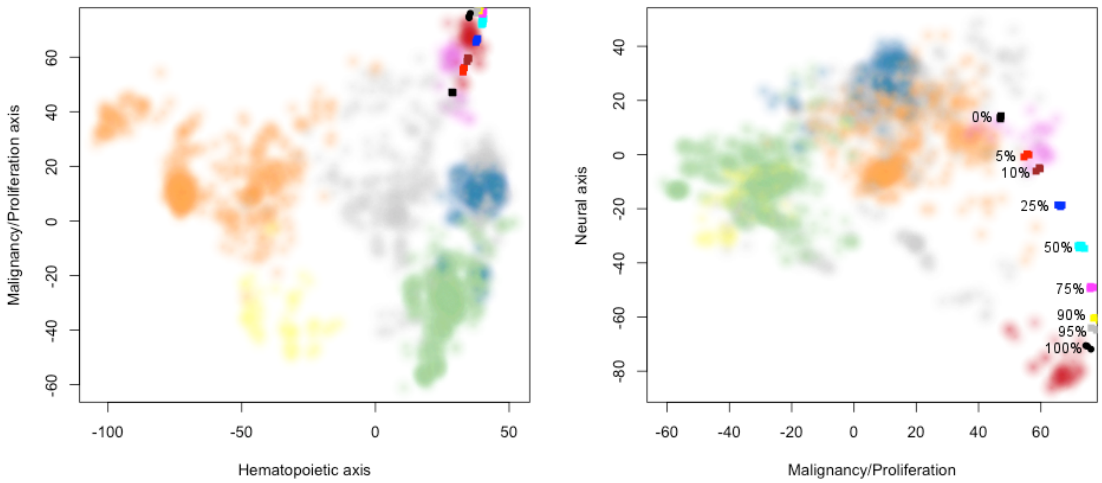


Figure 4.14.: Mixtures of brain and heart tissues from the Affymetrix Human Gene 1.0 st array [199] mapped to the PCA space. The 33 samples with different percentages of heart and brain tissue (0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100%) describe a nice curve with a monotonously changing projection on the neural axis.

Furthermore, it can be observed that the heart and brain signatures are negatively correlated, leading to slightly negative values for the pure samples in the opposite signature, i.e. a negative heart score for the 100% brain sample and a negative brain score for the 100% heart sample.

Linear versus non-linear mapping

This negative correlation is important for the non-linear mapping, since it leads to an unexpected behaviour in the brain signature (Fig. 4.16, Fig. B.11). In contrast to the non-linear effects seen in the previous chapter, the brain score shows a relatively sharp increase for low mixture fractions, stays then almost constant with increasing mixture fraction and increases again sharply for high mixture fractions (Fig. 4.16, red circles). This effect can be explained by the negative correlation of the brain and heart scores in the residual tissue specific space (Fig. 4.17) in combination with the unproportional high expression of the most prominent heart specific genes due to the non-linear relationship between the mixture fraction and the logarithmic measurement signal as described in the previous section (Fig. 4.12 and Fig. 4.13). The unproportional high expression of these genes prevents an increase in the brain signature score for intermediate fractions of heart tissue.

Apart from this unexpected phenomenon, the results depicted in Fig. 4.16 resemble those of the similar Fig. 4.11 from the previous section.

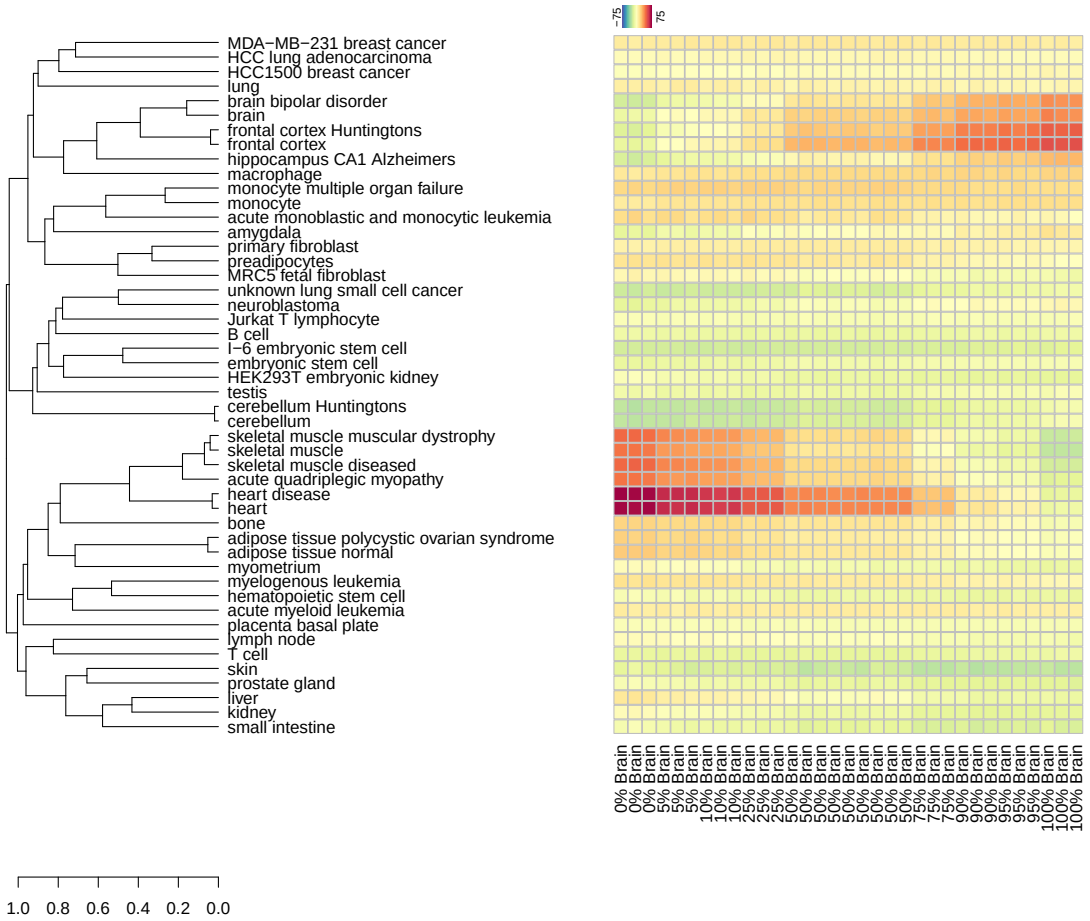


Figure 4.15.: Linear mapping of the mixtures of brain and heart tissues from Affymetrix [199] to the residual tissue specific space. A monotonous increasing association with brain signatures and a decreasing association with heart signatures can be observed.

4.3. Comparison to existing approaches

In this section, we compare the two-scale map to two existing approaches for an RNA microarray based cell characterisation, namely the "Unknown RNA Sample Annotation (URSA)" [10] and the "Concordia: Phenotypic concept enrichment" [11] tools. The comparison to the "CellNet" tool [81] will be shortly described afterwards, within the analysis of the *in vitro* differentiated cardiomyocytes. The reason for this is that the tool is currently restricted to a relatively small number of cell or tissue types, permitting a rigorous evaluation with the datasets used above.

4.3.1. Unknown RNA Sample Annotation (URSA)

The "Unknown RNA Sample Annotation (URSA)" tool [10] characterises new data based on a tissue ontology. This characterisation is done based on a support vector machine

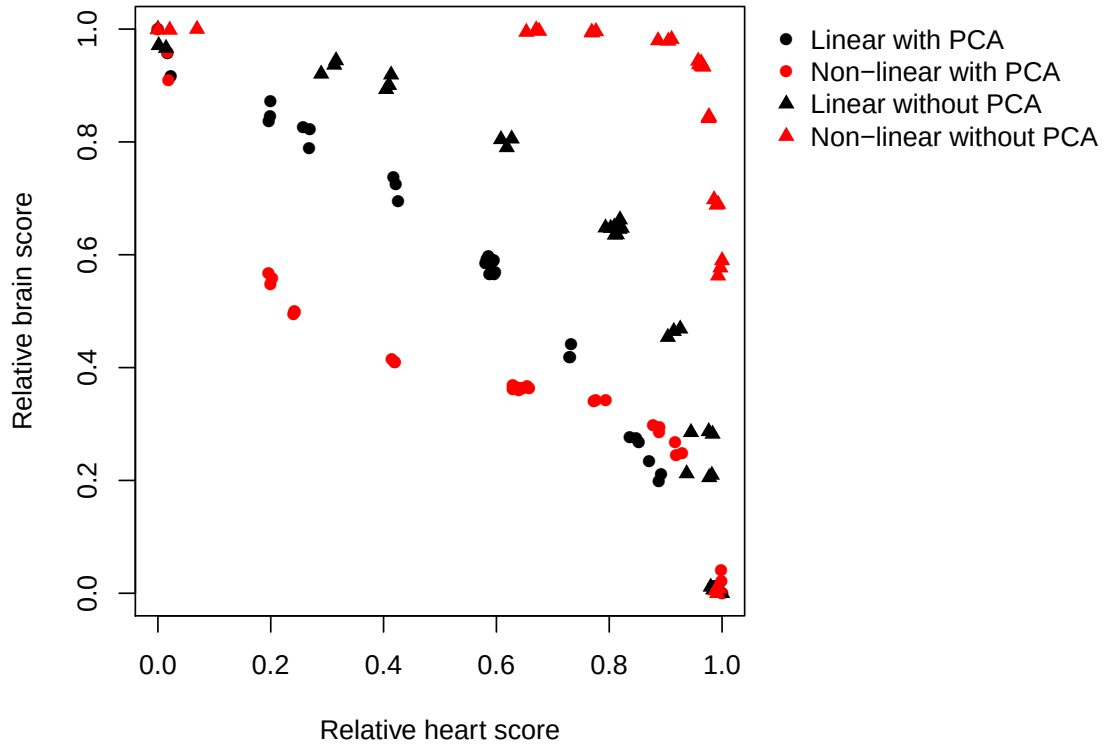


Figure 4.16.: Comparison of different mapping approaches for the brain and heart mixture data. Plotted is the heart score versus the brain score, each normalised to values between 0 and 1, for four different mapping procedures, i.e. the linear and non-linear mappings with and without PCA based decomposition. The linear mapping is closer to a linear relationship between the heart and brain score than the non-linear mapping. The non-linear mapping with PCA based decomposition shows clearly the signed-quadratic relationship discussed in the previous chapter.

classifier combined with a bayesian approach utilising the tree like structure of the ontology. The classifier was trained on a very large dataset from the Affymetrix human U133 Plus 2.0 array, but is applicable to data from any platform and even to RNA-seq data [10]. The URSA tool is accessible via an internet platform (http://ursa.princeton.edu/) and requires only a ranked list of genes representing the order of expression values from a single sample. The output of the URSA analysis is a list of tissue ontology terms that are predicted to be associated with the new data, along with probability values for the association. Thus, it is essentially a binary classification for each ontology term along with a measure of confidence. This is a structural difference to the two-scale landscape approach, which allows an interpretation in terms of a distance between two samples in a certain direction. This interpretation is especially helpful for mixture data or for the characterisation of samples that do not exactly fit to a certain tissue. An advantage of the URSA tool is that it does not require a reference dataset for characterisation of samples from different platforms.

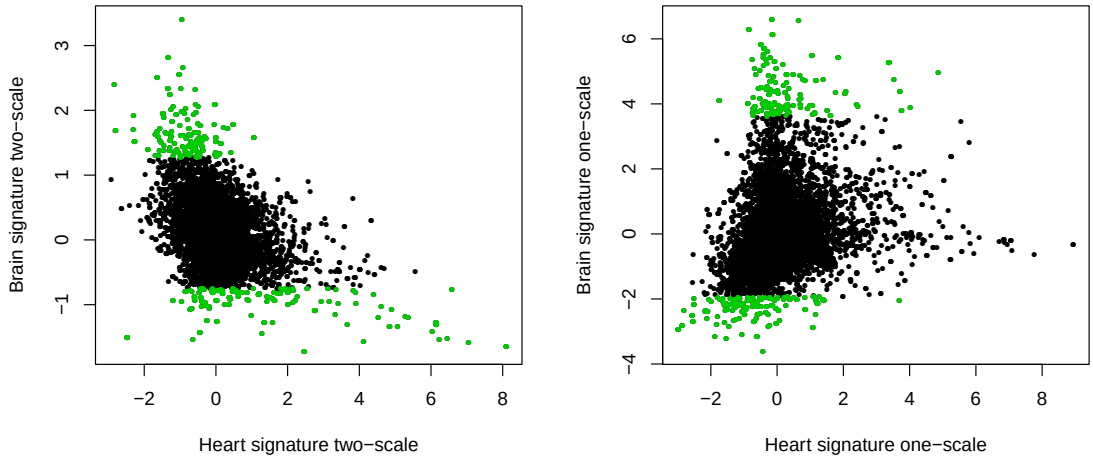


Figure 4.17.: The heart and brain signatures in the residual tissue specific space are negatively correlated and several prominent heart specific genes are amongst the most negative genes in the brain score (left). This negative correlation is not present in the tissue specific space without PCA based decomposition (right). The green colour marks all genes that are used for the non-linear mapping to the brain signature, i.e. the 1% of genes with highest and lowest expression in the signature.

In the following, the results of the URSA tool applied to the three datasets analysed so far are discussed.

Expression data from various tissues

We start with the 24 different tissues and cell lines from dataset GSE18674, which do not contain any mixtures or unknown cell types. These samples are hybridised on the Affymetrix Human U133 Plus 2.0 microarray. This platform was also used for the development of URSA. For these tissue samples, the URSA tool delivers very specific results and matches quite well to the true tissue type (Fig. 4.18). However, some tissues could not be identified, i.e. they have very low probability values for all ontology terms that are included in the URSA tool. These are the lung, salivary gland, and thymus tissues as well as the SHSY5Y cell line. Furthermore, it has to be noted that some relatively similar tissues could not be properly distinguished. For example, the colon and small intestine tissues show almost the same result, being associated with alimentary canal, colon, colorectum, gastrointestinal tract, intestine, large intestine, and viscus. The small intestine is additionally associated with the ontologies digestive gland, endocrine gland, and pancreas, but not with small intestine. The stomach sample is also similarly classified as the small intestine sample, without an association to viscus and endocrine gland. Thus, it is also falsely classified as colon and large intestine.

Figure 4.18.: Unknown RNA sample annotation (URSA) tool applied to the 24 different tissues and cell lines from dataset GSE18674. The results are very specific and most tissues are correctly identified, similar to the two-scale map. However, several tissues like lung, salivary gland, thymus, and the SHSY5Y neuroblastoma cell line are not associated with any ontology at all. Furthermore, very similar tissues like different brain regions are partly not properly distinguished. Colours correspond to probability values ranging from zero (light yellow) to one (red).

Another example are the adult cortex and fetal brain samples, having the same matching ontologies, except of an additional similarity of fetal brain to brain cancer. Finally, spinal cord and cerebellum have exactly the same associations. These small scale differences between different brain regions could be better resolved by the two-scale landscape approach.

Liver and breast cancer mixture data

The liver and breast cancer mixture dataset (GSE33116) is more challenging for the URSA tool for two reasons. First the samples were hybridised on a different microarray platform, namely the Affymetrix Human U133A array. Second, and more important, the dataset contains samples that are no pure tissues, but mixtures of two different tissues. Therefore, it will be interesting to see how the binary classification for each ontology can deal with such data. The results of this analysis are depicted in Fig. 4.19.

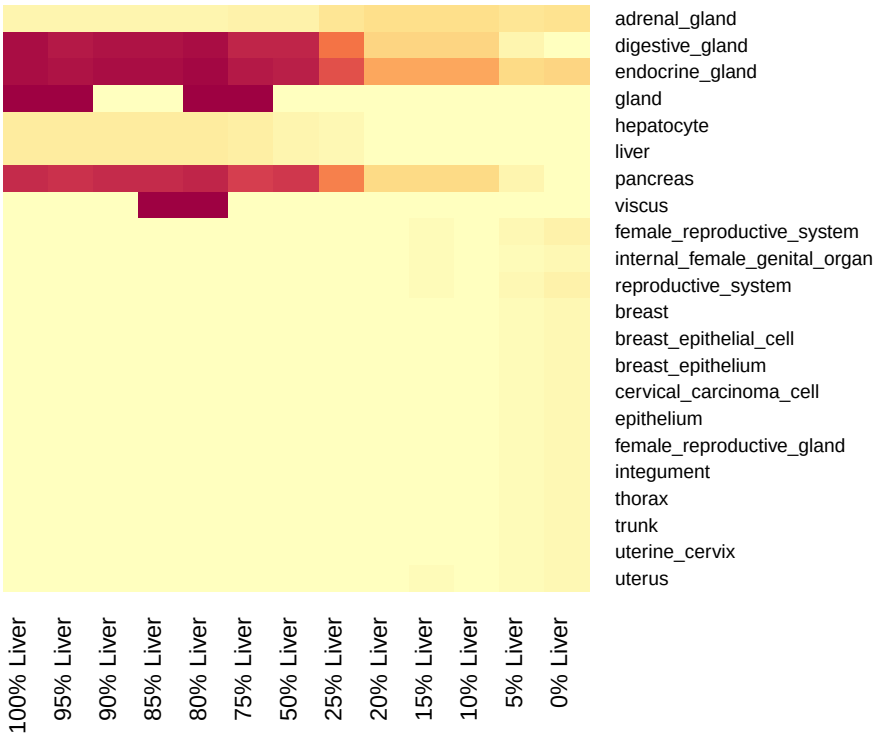


Figure 4.19.: Results of the URSA tool applied to the liver and breast cancer mixture data from GSE33116. Neither the pure nor the mixed samples are correctly classified.

Surprisingly, the URSA tool is not able to correctly identify even the pure samples. The pure liver sample is falsely characterised as pancreas, whereas the pure breast cancer sample does not have a clear association to any specific tissue. The most probable reported associations are endocrine gland (24% probability) and adrenal gland (18% probability).

Furthermore, the samples with 50% to 100% liver fraction do not show any consistent differences which may be attributed to different mixing fractions, i.e. they are predicted to be essentially the same cells. For lower liver fractions the ontology associations stay similar but with decreasing probability values.

Heart and brain mixture data

The second example of tissue mixtures are the heart and brain mixture data. For these data, the URSA tool should have less problems to correctly classify the pure samples since the expression patterns of heart and brain tissues are very distinct from all other tissues. The relevant tissue ontologies of this analysis are depicted in Fig. 4.20.

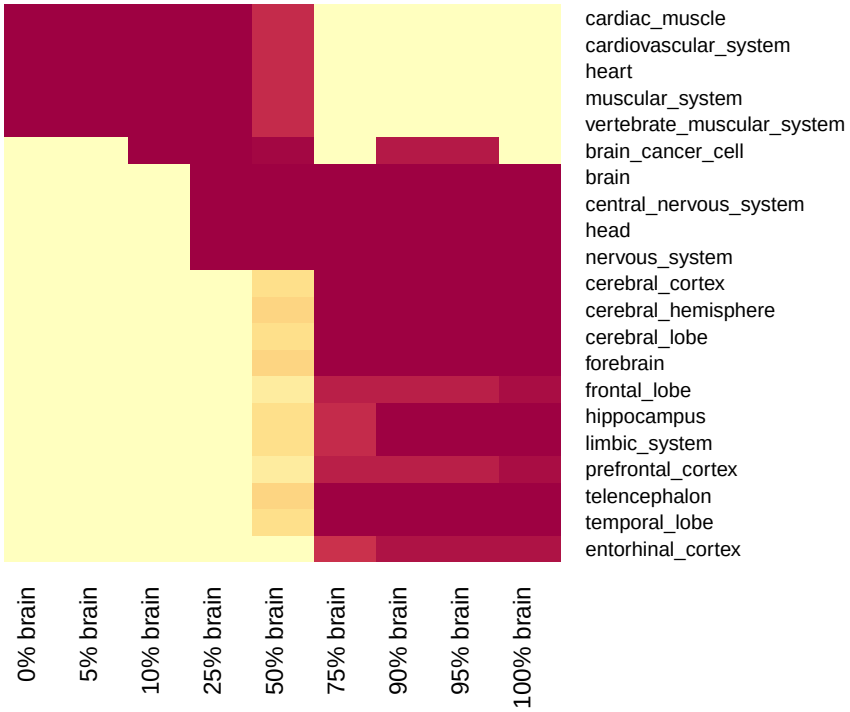


Figure 4.20.: Results of the URSA tool applied to the Affymetrix heart and brain mixture data. Pure samples are nicely distinguished and correctly classified. However, no smooth transition can be found for mixed samples. Instead, classifications change in a binary way. Interestingly, the 25% brain fraction suffices for a 100% association with heart as well as brain, whereas the opposite 75% fraction leads to a classification as pure brain tissue without any association to heart.

Indeed, the URSA tool nicely detects the right tissues, i.e. heart and brain, for the pure samples. However, it is not really able to identify mixtures of cells, as it either classifies the samples as pure heart or brain tissue, for the 5%, 10%, 75%, 90%, and 95% brain mixture fractions, or as 100% of both, brain and heart, as is the case for the mixtures with 25% and 50% brain RNA. Thus, there is no smooth transition due to the binary classification

scheme.
Overall, the URSA tool suppresses unspecific tissues much better than the two-scale landscape, which suffers a bit from the correlations which are still present between different tissue signatures. However, due to this suppression, the URSA tool is not able to detect smaller amounts of mixtures, i.e. it can not nicely detect smooth transitions. This might be critical for the analysis of *in vitro* differentiated cells and especially differentiation time courses, which do not exactly match to one specific tissue. In summary, URSA is a valuable and easy to use tool for expression based cell characterisation, but seems not to be well designed for the analysis of cellular transformations, what is the focus of the present thesis.

4.3.2. Concordia: Phenotypic Concept Enrichment

The Concordia tool characterises new microarray samples based on their Spearman correlation to a large reference database of annotated samples, combined with an enrichment score calculated with a Kolmogorov-Smirnov test statistic [11]. The tool was developed on the Affymetrix Human U133 Plus 2.0 microarray platform and no platform transformation for cross-platform analyses is described. Therefore, the tool is restricted to a single microarray platform.

For an evaluation of the tool we use the 24 tissues and cell lines from GSE18674 that are hybridised on the required microarray platform. Initially, we wanted to use the online platform (http://concordia.csail.mit.edu) described in [11]. However, the website was not accessible when we tried to use it (early September 2014). Therefore, we reimplemented the tool using the own dataset from the Affymetrix Human U133 Plus 2.0 array with manual annotation of 191 different cell or tissue types. Of these 191 groups, 150 have 3 or more associated samples and the analyses were restricted to these samples as suggested in [11]. The results of this analysis (Fig. 4.21) are significantly worse than the respective results of the two-scale map with the same reference dataset (Fig. B.4). Although the correct concept is in most cases among the top scoring ones, it is very hard to actually identify the tissue type since there are several enrichment scores with relatively high values which sometimes do not really fit together. For instance, urethra has the second highest score for the heart sample. The highest is pericardium and heart itself is only at position three. Interestingly, the performance measure used in [11] reports results that seem to be quite good at a first view (e.g. an average "accuracy" of 92.8%). However, this performance measure evaluates the classification performance of an individual concept over all samples and averages then over all concepts. That means, it compares the enrichment scores of all samples for an individual concept and reports the area under the receiver operating characteristic (ROC) curve. For the own dataset used in the present thesis, this results in a mean AUC of 97.5%, which is even higher than the 92.8% reported in [11]. The difference is due to the use of different datasets, and the dataset used here is larger, resulting in better AUC values [11].

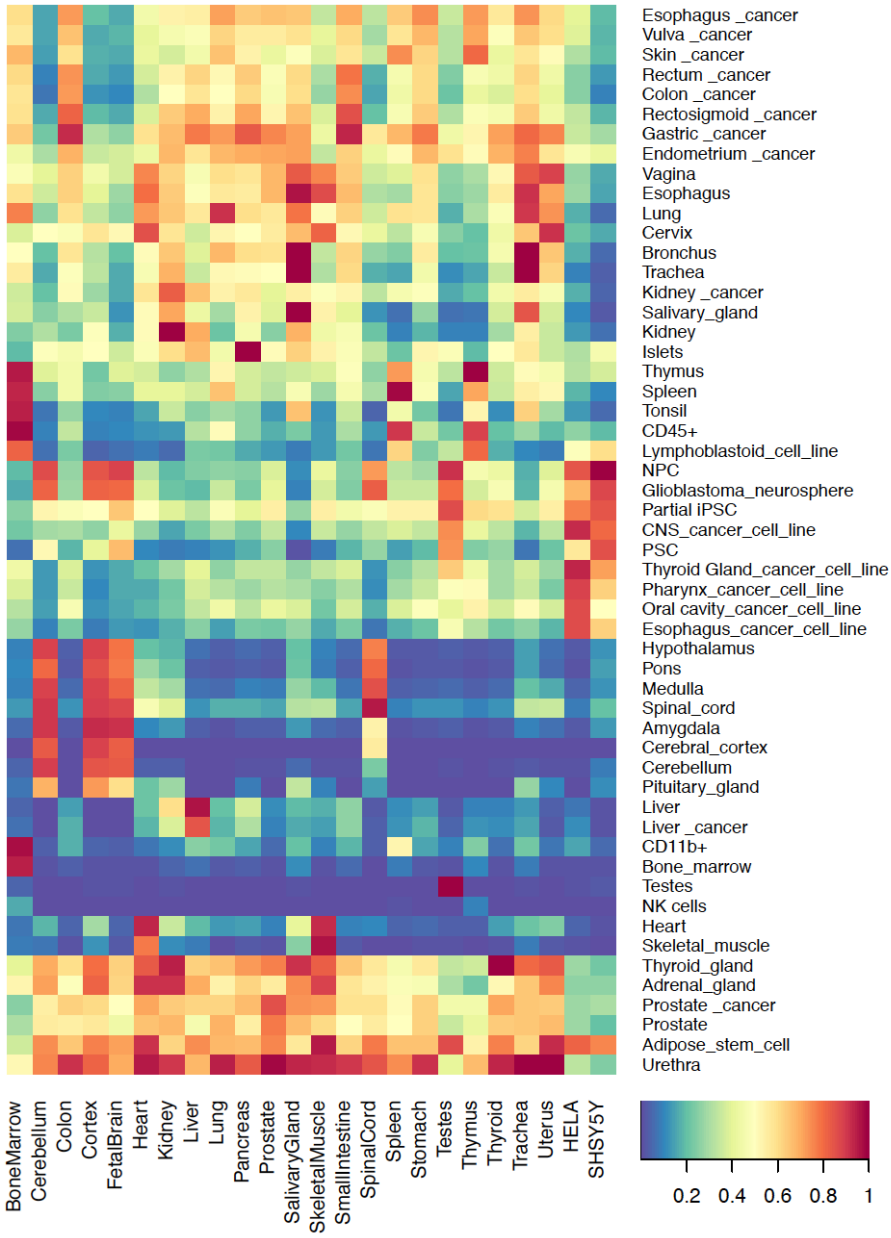


Figure 4.21.: Results of the Concordia tool applied to the 22 tissues and 2 cell lines of dataset GSE18674. Plotted is the enrichment score calculated as described in [11] using the own dataset described in the present thesis as reference. The results are significantly worse than those of the two-scale landscape.

In contrast, the aim for characterisation of cells is to compare different concepts for a single sample and to identify the sample using the concepts with highest enrichment scores. For an evaluation of the performance matching to this aim, one can compare the actual tissue type of a sample to the concept with highest enrichment score for this sample. This results in a correct identification of only 20.4% of all samples in the own dataset using the Concordia tool. For both kinds of evaluation we used a leave-one-out cross validation scheme as described in [11].

Due to the fact that some concepts are very similar, e.g. liver and fetal liver, it may happen that some exactly matching concepts have only the second or third highest score. Therefore, we also checked the fraction of samples where the exact matching concept is among the top 2 (30.9%), top 3 (39.7%), top 5 (51.4%) or top 10 (73.0%) scores. These values are still relatively low. So, according to this more strict criterion, the performance of the Concordia tool is significantly worse than indicated by the AUC criterion. This fits well to the observations made for the 24 tissues or cell lines from dataset GSE18674 which are hard to identify (Fig. 4.21).
For a comparison to the two scale map, we did the same leave-one out cross validation for the two-scale map, using the score of the residual tissue specific space. This results in a mean AUC of 98.3%, which is slightly higher than the 97.5% of the Concordia tool. More interesting is the alternative evaluation, comparing the top scoring signatures to the actual tissue type. This results in a correct identification of 71.4% of the samples, which is significantly higher than the 20.4% of the Concordia tool. As for the Concordia tool, this value increases if we check for the top 2 (82.1%), 3 (86.3%), 5 (89.5%) or 10 (92.3%) scores. All of these results are significantly better than those of the Concordia tool.
In summary, the Concordia tool performs significantly worse than the two-scale map and seems to be not well suited for characterisation of *in-vitro* reprogrammed or differentiated cells.

4.4. Application to *in vitro* differentiated cells

4.4.1. Mesenchymal stromal cells

Mesenchymal stromal cells (MSCs) isolated from bone marrow of three different donors were reprogrammed to iPSCs [111] and subsequently redifferentiated back to mesenchymal stromal cells (iPS-MSC) [200] (Fig. 4.22). The differentiation was accomplished by a simple exchange of the culture medium to the initial MSC-culture medium. That means, that iPS-MSCs and MSCs were cultured under the same conditions and are therefore well suited for an evaluation of the effects of reprogramming and differentiation processes. The redifferen-

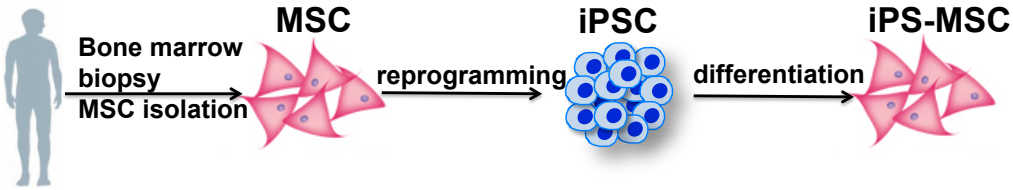


Figure 4.22.: Overview of the mesenchymal stromal cell (MSC) experiment. MSCs isolated from bone marrow of different donors were reprogrammed to induced pluripotent stem cells (iPSC) and redifferentiated to MSC (iPS-MSC). Gene expression microarray data were generated from the initial MSCs, the iPSCs, as well as during the course of differentiation at day 7, 14, 21, 28, and 35.

tiated iPS-MSCs fulfil the "minimal criteria for defining multipotent mesenchymal stromal

cells" [200, 201], but show some differences in gene expression and methylation compared to the original MSCs [200]. The expression changes mainly indicate a reduced immune modulatory function of iPS-MSCs. This could be validated by wet-lab experiments measuring the T cell activation of MSCs [200]. Interestingly, the methylation differences point towards an epigenetic rejuvenation of iPSCs and iPS-MSCs. That means, two methylation based markers for age prediction [202, 203] predict that iPSCs are age-reprogrammed to age zero and redifferentiation towards iPS-MSCs does not lead to an ageing of cells, i.e. they remain at age zero.

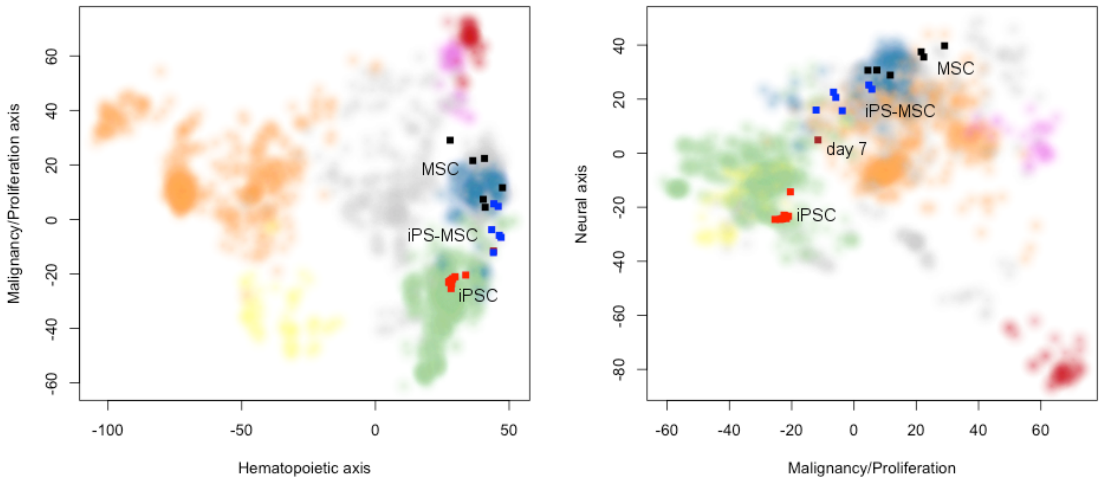


Figure 4.23.: Projection of bone marrow or iPSC derived mesenchymal stromal cells (MSCs) onto the PCA map. According to the second principal component, MSCs derived from iPSC (iPS-MSC, blue dots) seem to proliferate faster than bone marrow derived MSC (black dots). In general, the iPS-MSCs lie on a path between iPSCs and bone marrow derived MSCs.

Analysis of the expression data with the two-scale landscape also confirms the proper differentiation towards MSCs (Fig. 4.23 and Fig. 4.24). The main difference between MSCs and iPS-MSCs in the two scale map is the higher proliferation of the latter as indicated by a difference on the malignancy/proliferation axis, i.e. PC2 of the PCA space (Fig. 4.23). This higher proliferation could also be confirmed by wet lab experiments [200]. Similarly to the epigenetic rejuvenation of the cells, this is indicative of an immature status of the redifferentiated iPS-MSCs.

On the residual tissue specific space, there are only very tiny differences between MSCs and iPS-MSCs. For both cell types, several incompletely differentiated signatures, e.g. associated with MSCs, fibroblasts, or preadipocytes, have a high score. This fits well to results from the literature, describing the remarkable similarity of MSCs and fibroblasts [204].

In strong contrast to the MSCs, iPSCs are clearly classified as being similar to ESCs.

Together with the PluriTest results [200], this confirms the proper reprogramming of iPSCs and is a further validation of the biological relevance of the two-scale map.
Apart from these two rather pure cell types, i.e. iPSC and MSC, it is interesting to characterise the cells during the course of differentiation. A first finding from the two-scale map as well as from other analyses [200] is that the differentiation is already completed at day 14, showing almost no differences in gene expression and methylation at later time points (except of slight changes in the proliferation). Thus, the most interesting intermediate time point between iPSCs and iPS-MSCs is day 7. At this time point, there is an intermediate signal for both, incompletely differentiated and ESC scores, indicating a relatively smooth transition from iPSC to MSC without changing to a completely different cell type in between.

4.4.2. Cardiomyocytes

Analysis with the two-scale landscape

For the analysis of *in vitro* differentiated cardiomyocytes from human ES cells and iPS cells we reanalyse data from three studies, which we downloaded from ArrayExpress (E-MEXP-2654 [5]), or Gene Expression Omnibus (GEO) databases (GSE28191, and GSE35108 [205]). These studies contain one differentiation time series from day 0 to day 11 (GSE28191), differentiated cardiomyocytes at week 3 and 7 (E-MEXP-2654), as well as 4 weeks differentiated cardiomyocytes from iPSC of healthy people and dilated cardiomyopathy (DCM) patients (GSE35108). The DCM samples are "from a three-generation family of DCM patients carrying a point mutation (R173W) in exon 12 of the TNNT2 gene" [205].
Analysis of these samples with the two-scale gene expression map reveals a continuous change during differentiation on the proliferation axis (Fig. 4.25), as well as on the embryonic stem cell and heart dimensions of the residual tissue specific space (Fig. 4.26). Thus, in comparison to heart tissues (green points and magenta background on Fig. 4.25), the *in vitro* differentiated cardiomyocytes proliferate faster and are not fully differentiated to heart cells. The latter point is further indicated by a similarity of *in vitro* differentiated cardiomyocytes to incompletely differentiated cells, e.g. mesenchymal precursors, embryonic myoblasts, or preadipocytes, on the PCA space (Fig. 4.25, blue background), as well as the residual tissue specific space (Fig. 4.26). In contrast, fetal and adult heart tissues do not have such a similarity, but are slightly associated with adult adipose tissue (Fig. 4.26). This may hint to an immature status of the differentiating cells.
Having a more detailed look at the ESC and heart scores of the residual tissue specific space, one can observe a linear relationship between both scores during the course of differentiation (Fig. 4.27). That means, that the decrease of the ESC score goes hand in hand with an increase of the heart score. A deviation from the straight line can only be observed in the very early differentiation (day 2 and day 5), where almost no increase in the heart score can be observed, and for the DCM cardiomyocytes, which have already lost

Figure 4.24.: Linear projection of bone marrow or iPSC derived MSCs onto the residual map suggests that the differentiation is already completed at day 14. In general, MSCs and iPS-MSCs have very similar tissue associations with a dominant association to incompletely differentiated cells. Differentiating cells at day 7 exhibit similarities to the ESC signature as well as to signatures associated with incompletely differentiated cells, suggesting a location on the path between iPSC and MSC.

much of their pluripotency, but have still a comparably low association to heart tissues. One may ask which kind of transformation the cells undergo and whether the *in vitro* differentiated cells are mixtures of heart cells and pluripotent stem cells. To this end, we investigated the expression of several typical pluripotency markers, i.e. NANOG, POU5F1,

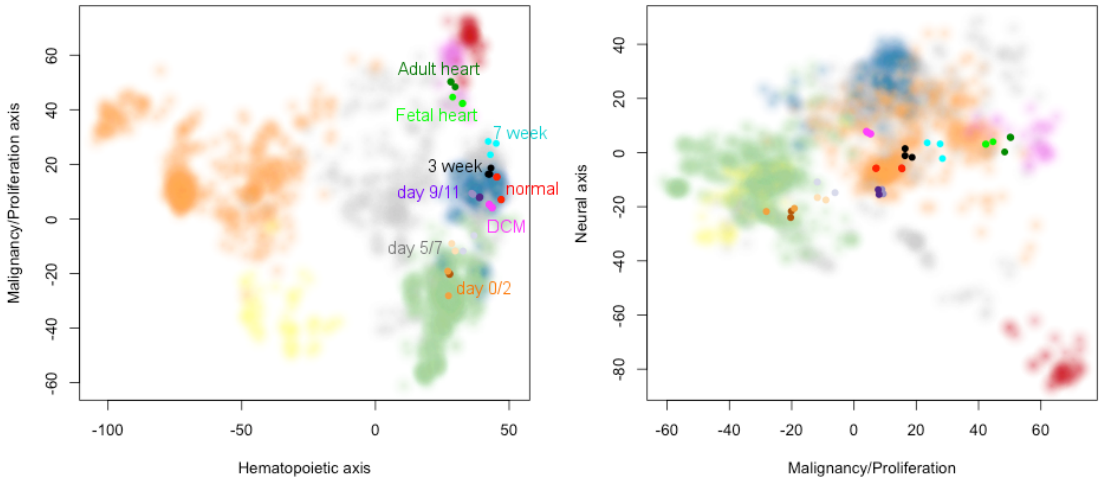


Figure 4.25.: Projection of heart tissues and differentiated cardiomyocytes onto the PCA map reveals a decreasing proliferation during differentiation, rendering the differentiating cells more and more similar to heart tissues. Colour code: brown: diff. time course day 0, orange: day 2, light orange: day 5, light purple: day 7, purple: day 9, dark purple: day 11, black: 3 week differentiation, cyan: 7 week differentiation, red: 4 week differentiation from normal iPSC, magenta: 4 week differentiation from DCM iPSC, light green: Fetal heart, dark green: adult heart.

DPPA4, DNMT3B, SOX2, L1TD1, ZSCAN10, LIN28A, and ZFP42. All of these markers show a sharp early decrease of the expression during the course of differentiation, while the ESC score decreases only marginally in this time (Fig. 4.28). In contrast, in case of a mixture effect, one would expect a logarithmic curve with a relatively high expression for early differentiated cells and a sharper drop down at the very end of the differentiation (Fig. 4.12). Therefore, the joint similarity to ESCs and heart tissues is not due to a mixture effect.

Analysis using CellNet

So far, we did not compare the two-scale landscape method to the CellNet tool [81], which was explicitly designed for the characterisation of *in vitro* transformed cells. As stated before, the tool requires relatively large amounts of data for the training and is therefore currently restricted to relatively few tissues or cell types. Unfortunately, neither neural tissues, nor mesenchymal stromal cells are part of this set of tissues for the Affymetrix human gene 1.0 st v1 microarray platform. Therefore, we restrict our analyses to the *in vitro* differentiated cardiomyocytes.

Using the CellNet online tool at `http://cellnet.hms.harvard.edu`, this leads to very similar results as obtained using the two-scale landscape (Fig. 4.29), albeit with a much higher training effort of the tool. Therefore, according to this limited analysis, both tools

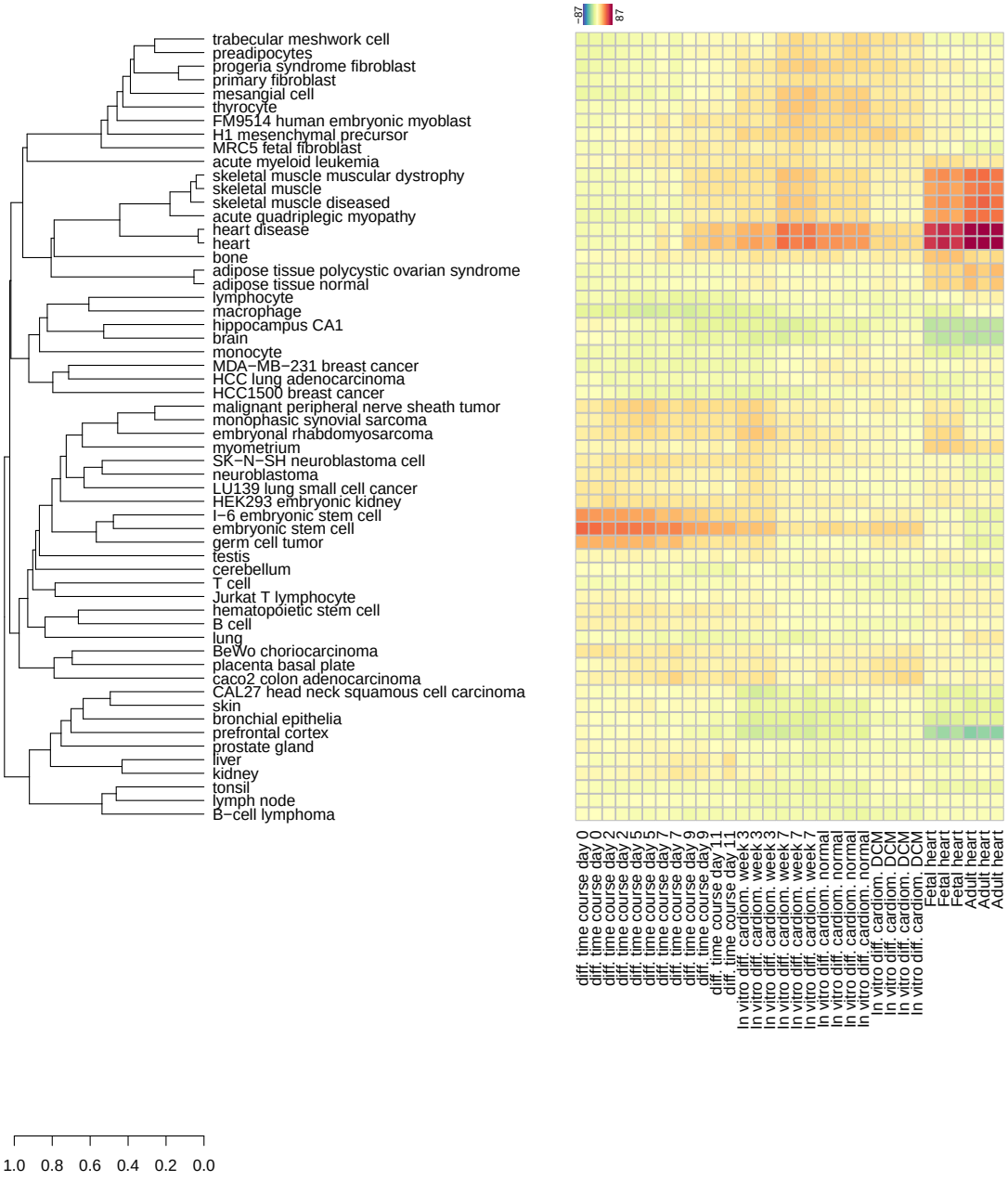


Figure 4.26.: Linear projection of heart tissues and differentiated cardiomyocytes onto the residual map reveals an increasing similarity to heart signatures and a decreasing similarity to ESC signatures with prolonged differentiation. Cardiomyocytes derived from normal iPSCs seem to be more similar to heart cells than those from dilated cardiomyopathy (DCM) patients. *In vitro* differentiated cardiomyocytes have some similarities to incompletely differentiated cell types, in contrast to fetal and adult heart samples.

can be considered as roughly equally good in those cases where the appropriate cell type is available in the CellNet tool. Certainly, a more detailed comparison of both tools is necessary to reveal strengths and weaknesses of one or the other tool. However, for most

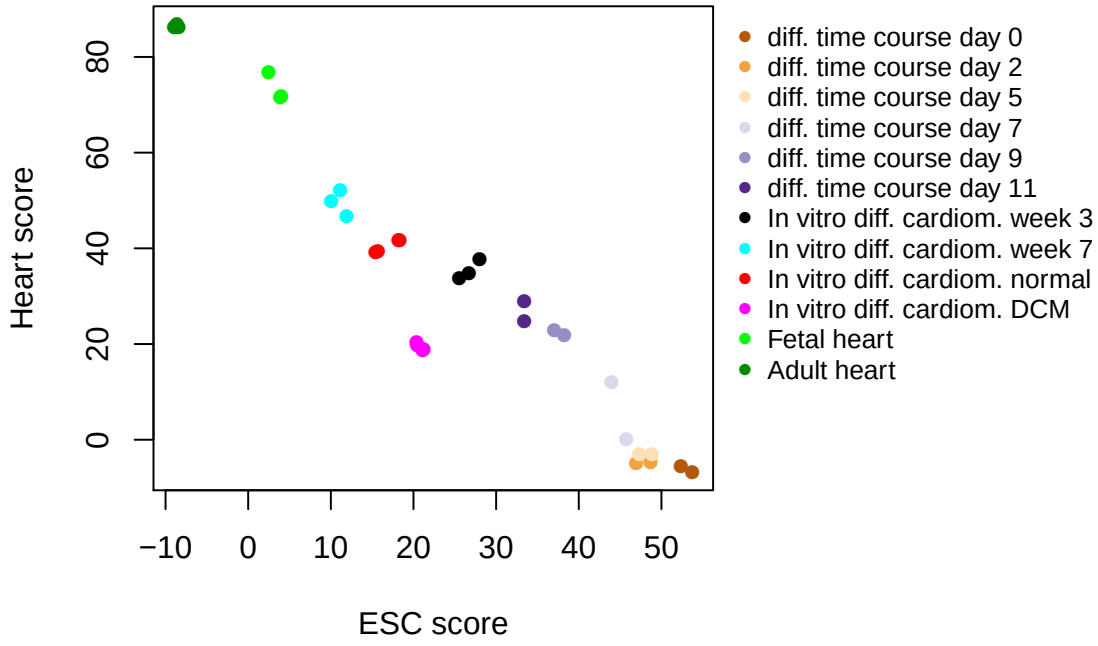


Figure 4.27.: Detailed comparison of the heart and ESC scores during cardiomyocyte differentiation. An almost linear relationship can be observed with decreasing ESC score and increasing heart score during differentiation.

of the data analysed here, the CellNet tool does not have appropriate reference data. This is also the case for the following examples of neurons, astrocytes, and neural stem cells.

4.4.3. Neurons, neural stem cells, and astrocytes

Brain tissue during human development

Many different cell types can be found in the human brain [15], including different types of neurons, neural progenitors or neural stem cells, as well as glial cells, e.g. astrocytes. The tissue composition is different in different regions of the brain and it changes strongly during early brain development [206]. For a detailed investigation of the expression patterns of brain tissue, Kang et al. [206] generated and analysed more than 1300 microarray samples from embryonic, fetal, and postnatal to adult brain, incorporating data from different brain regions at each developmental status.

The analysis of this large data resource showed that the strongest differences in expression occur during prenatal brain development [206]. Therefore, it is interesting to map a time course of brain development to the two-scale map in order to reveal the location of embryonic and fetal brain tissues for a comparison to *in vitro* differentiated cells.

We focused our analysis on samples from the cerebral cortex (or cerebral wall for early time points) and mapped samples from 15 different age periods [206] to the two-scale land-

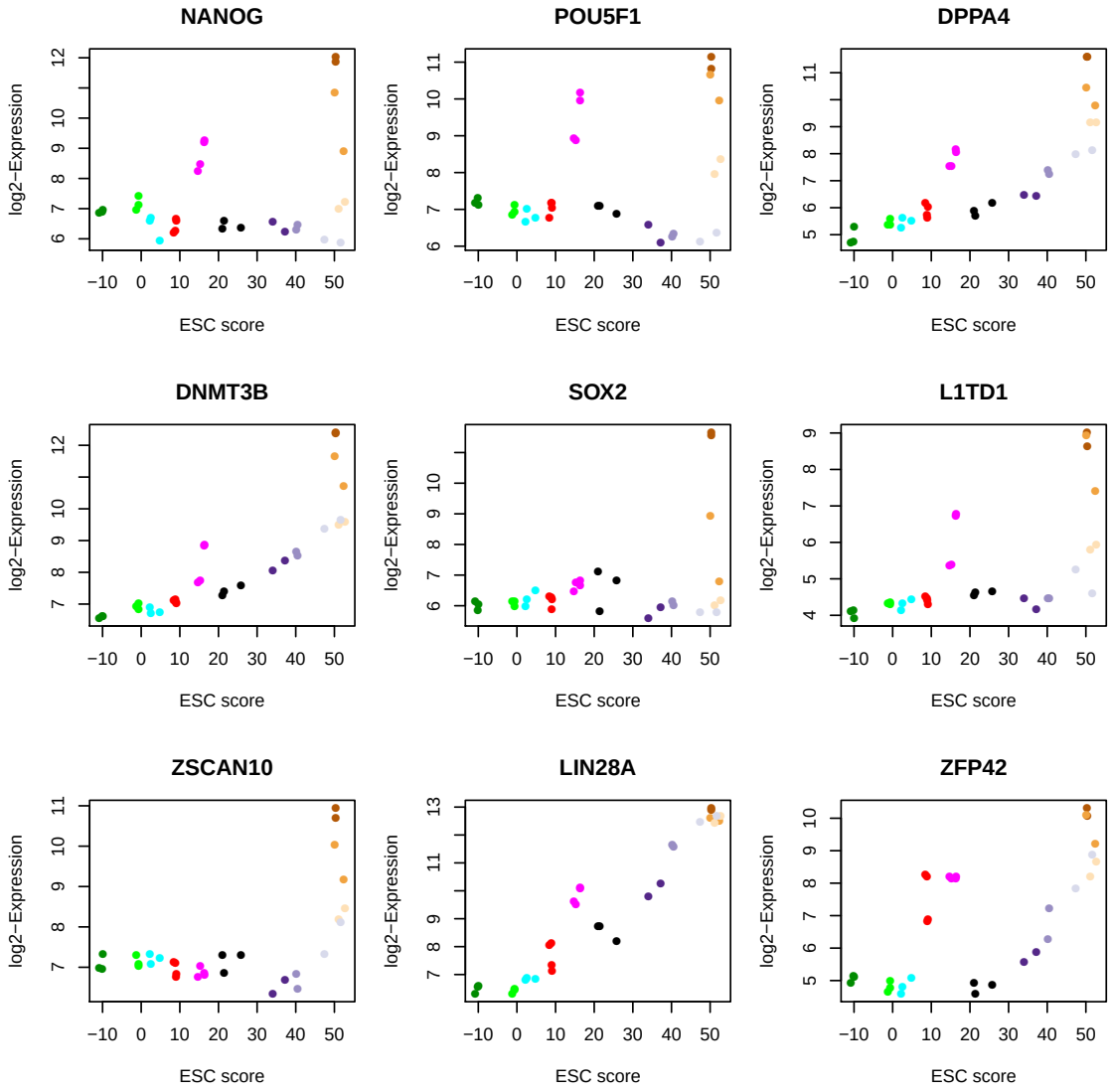


Figure 4.28.: Expression of typical pluripotency markers in the course of differentiation towards cardiomyocytes. All pluripotency markers show a sharp decrease at early stages of differentiation, where the overall similarity to ESCs has only slightly changed. In contrast, when the *in vitro* differentiated cells would be mixtures of pluripotent stem cells and heart cells, one would expect a logarithmic shape of the curve, similar to Fig. 4.12.

scape. In order to do this, we mapped the data that were generated on the Affymetrix Human Exon 1.0 st-v1 array to the Affymetrix Human Gene 1.0 st-v1 array and quantile normalised them together, since we do not have an appropriate reference dataset for the Exon array (see appendix A). This mapping may lead to small differences on the PCA space compared to the reference dataset used as background for orientation (Fig. 4.30, Fig. B.13).

The mapping of these samples to the PCA space and residual tissue specific space developed based on the Lukk dataset are depicted in Fig. B.13 and Fig. B.14. On the residual

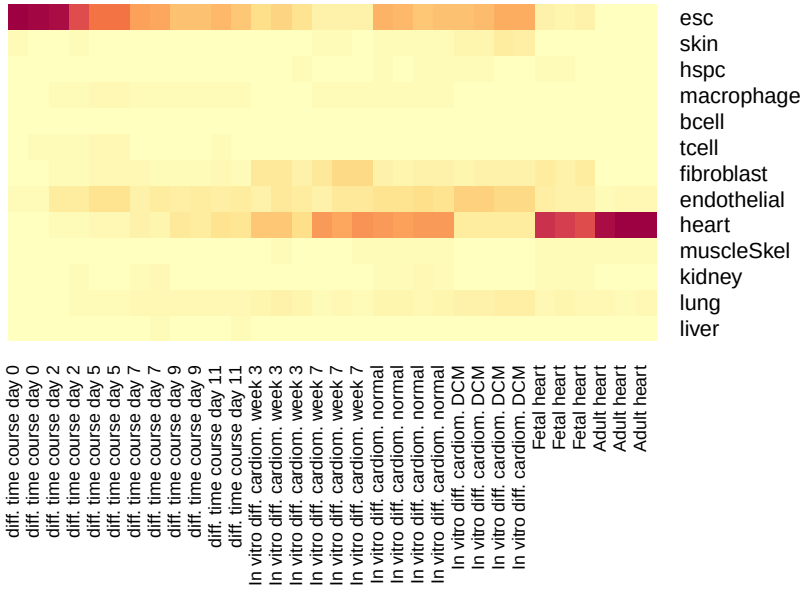


Figure 4.29.: Analysis of the *in vitro* differentiated cardiomyocytes using CellNet [81]. Shown are all 13 tissue/cell line scores that are currently included in the CellNet tool for the Affymetrix Human Gene 1.0 st v1 microarray platform. The results are very similar to those of the two-scale map, showing overall an increasing similarity to heart and and a decreasing similarity to ESC during prolonged differentiation. The values range from 0 (light yellow) to a maximum of 0.84 (dark red).

tissue specific space (Fig. B.14) a similarity of prenatal brain samples to various brain cancer and lung cancer tissues or cell lines can be observed. This fits well to the described similarity of some brain cancer cells to neural progenitors or neural stem cells [207, 208]. Unfortunately, the Lukk dataset does not incorporate samples from fetal brain or neural progenitors. Therefore, it is not possible to detect similarities to these tissue or cell types. This is different in the own dataset consisting of 7100 samples from the Affymetrix Human U133 Plus 2.0 array, which we manually classified into 191 different cell or tissue types. Therefore, we used this dataset to create an alternative two-scale map.
This two-scale map also consists of the first three PCs (Fig. 3.2 bottom) as PCA map and a residual tissue specific space that is constructed in the same way as for the Lukk [12] dataset, incorporating 191 dimensions that correspond to the annotated 191 cell or tissue types. The map was already shortly mentioned in section 4.2.1, where we showed that it is capable of correctly identifying all 22 tissues from dataset GSE18674 (Fig. B.4), although we note that these data are already part of the reference dataset itself.

Due to the neural progenitor and fetal brain samples in the own dataset we use this alternative two-scale map for the analysis of the brain tissues and *in vitro* differentiated neural cell types. Mapping of the 15 samples of the Kang [206] dataset to this two-scale map reveals an increasing similarity to adult brain tissues on the PCA space over time (Fig. 4.30). This trend occurs throughout embryonic and fetal development and is almost

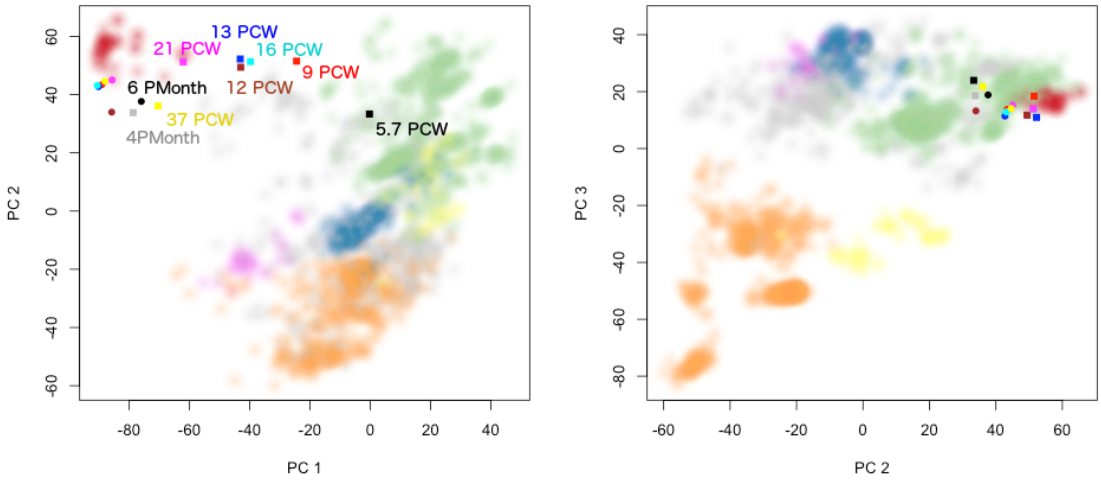


Figure 4.30.: Samples from human cortical development mapped onto the PCA space from the own dataset. Plotted are cortical brain tissues from embryonic and fetal development (5.7, 9, 12, 13, 16, 21, and 37 post-conceptional weeks) as well as from youth to adulthood (4 months, 6 months, 1 year, 8 years, 15 years, 36 years, 40 years, and 70 years). During brain development, the expression patterns get continuously closer to those of adult brain tissue (red background). Note that the data were generated on the Affymetrix Human Exon 1.0 st-v1 array, while the reference dataset used for the background colours was hybridised to the Human Gene 1.0 st-v1 array. This may explain the slight differences between the adult brain tissues and the red background on PC2.

completed at birth with only relatively slight differences between newborns and adults on the PCA space.

On the residual tissue specific space, a similarity of early embryonic cerebral wall tissue (5.7 post-conceptional week (PCW)) to pluripotent stem cells, neural progenitor cells, and fetal brain tissue can be observed (Fig. 4.31). The similarity to pluripotent stem cells decreases very rapidly during development, the similarity to neural progenitor cells increases first for the sample at PCW 9 and decreases then, while the similarity to fetal brain tissue increases first and stays high until birth. During late fetal development, the similarity to adult cerebral cortex and hippocampus increases and stays high throughout the human life span.

As already observed in previous sections, the two-scale landscape can distinguish between specific brain regions. Thus, the cerebral cortex samples show high similarities to cerebral cortex and hippocampus after birth, while samples from the cerebellum show a higher similarity to cerebellum from 4 months to 36 years after birth (Fig. B.15). The fact that this similarity decreases again in older aged people is a bit surprising and goes beyond the scope of the present thesis.

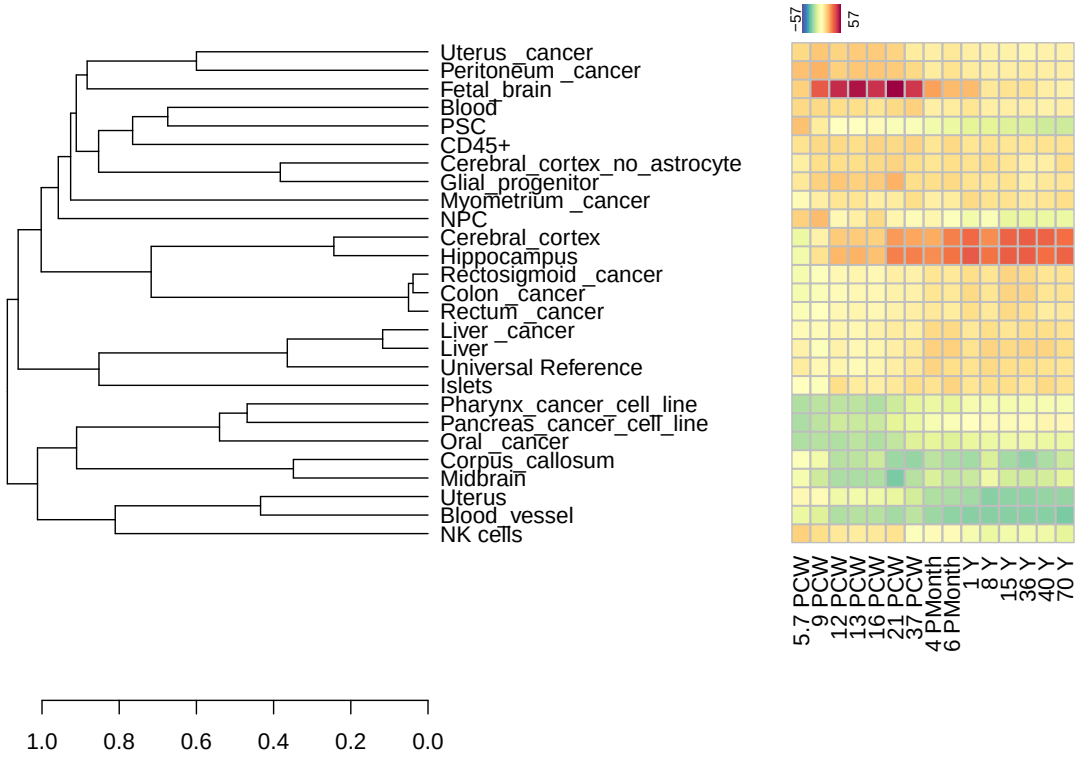


Figure 4.31.: 15 cortical brain tissue samples from post-conceptional week 5.7 to adulthood (70 years) mapped to the residual tissue specific space from the own dataset. The sample from PCW 5.7 has some similarity to fetal brain, as well as to neural progenitor cells (NPC), and pluripotent stem cells (PSC). The latter two decrease rapidly during development, while the similarity to fetal brain increases first and stays high until birth. It then decreases again with the parallel increase of some adult brain scores, i.e. the cerebral cortex and hippocampus scores.

In vitro differentiated neurons

In vitro differentiation of pluripotent stem cells into neurons is an interesting topic for molecular studies of neural development and especially for studying differences in neurons derived from iPSCs with disease associated genetic defects compared to those derived from healthy iPSCs. A key requisite for these kinds of studies is a proper differentiation into neurons that resemble neurons from *in vivo* brain tissue. Here, we analyse data from two studies of *in vitro* neural differentiation.

In the first study (dataset GSE51214) iPSCs are differentiated towards dopaminergic neurons and either sorted for CORIN positive cells at day 12 or not [209]. CORIN is a floor plate marker, i.e. it is expressed in the region of the developing plate where dopaminergic progenitor cells are usually located [209]. Therefore, the CORIN sorted cell preparations represent a more pure culture, with a higher percentage of midbrain dopaminergic progen-

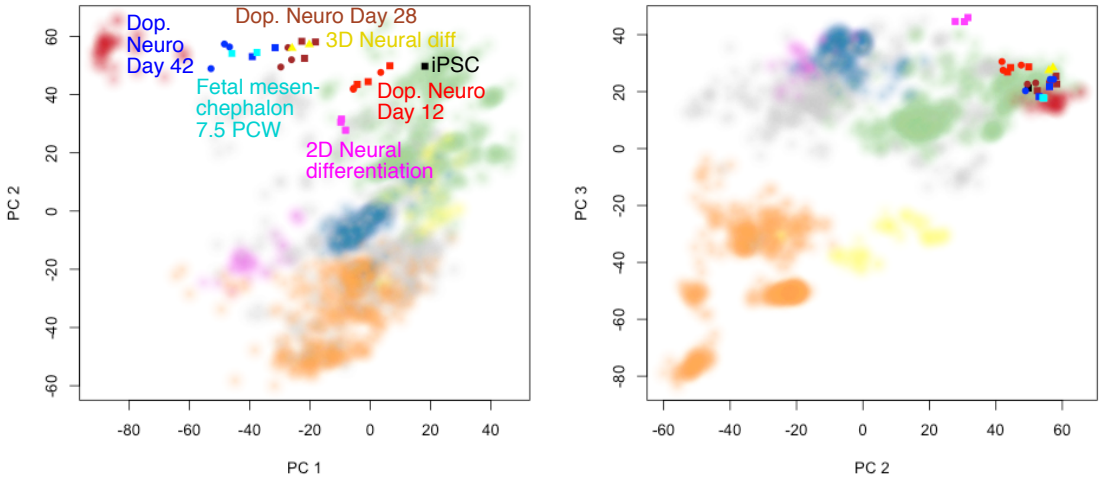


Figure 4.32.: *In vitro* differentiated neurons from datasets GSE51214, and GSE43235 mapped onto the PCA space. The differentiation time course from iPSC towards dopaminergic neurons (GSE51214, iPSC, Dop. Neuro Day 12, Dop. Neuro Day 28, and Dop. Neuro Day 42) shows increasing similarity to adult brain tissue. The samples from the latest differentiation time point (day 42) lie close to the fetal brain from post-conceptional week 7.5. CORIN sorted cell preparations (circles) at day 28 and day 42 of differentiation lie closer to brain tissues than unsorted samples (squares). The neural differentiations in 2D vs. 3D culture conditions from dataset GSE43235 show clear differences.

itor cells and reduced expression of several markers that are typical for other brain regions [209]. Furthermore, at day 28 of differentiation, the percentage of proliferating cells was higher in the unsorted samples than in the CORIN sorted samples [209]. Analysis of these samples with the two scale map reveals a monotonously increasing similarity to brain tissue over time on the PCA space (Fig. 4.32). Notably, CORIN sorted samples at day 28 and 42 have a higher similarity to brain tissue than unsorted samples. This potentially reflects the higher number of proliferating cells in the unsorted samples at day 28. On the residual tissue specific space (Fig. 4.33) the differentiation time course gets increasing similarity to fetal brain, whereas the similarity to pluripotent stem cells decreases. At day 42 of differentiation, the cells are very similar to the fetal mesenchyphalon from post-conceptional week 7.5 that are also included in the study.

In addition to the difference between CORIN sorted and unsorted cells on the PCA space, there are also slight differences on the residual tissue specific space. Especially for day 42 of differentiation, there is a notable difference in the hippocampus score, which is higher for sorted than for unsorted cells.

The second study of *in vitro* differentiated neurons (dataset GSE43235, no corresponding article published) investigates differences between *in vitro* differentiations in a classical 2D environment versus a 3D environment. They differentiate the cells for 28 days in both environments. Interestingly, these two differentiations result in very different cells. While the

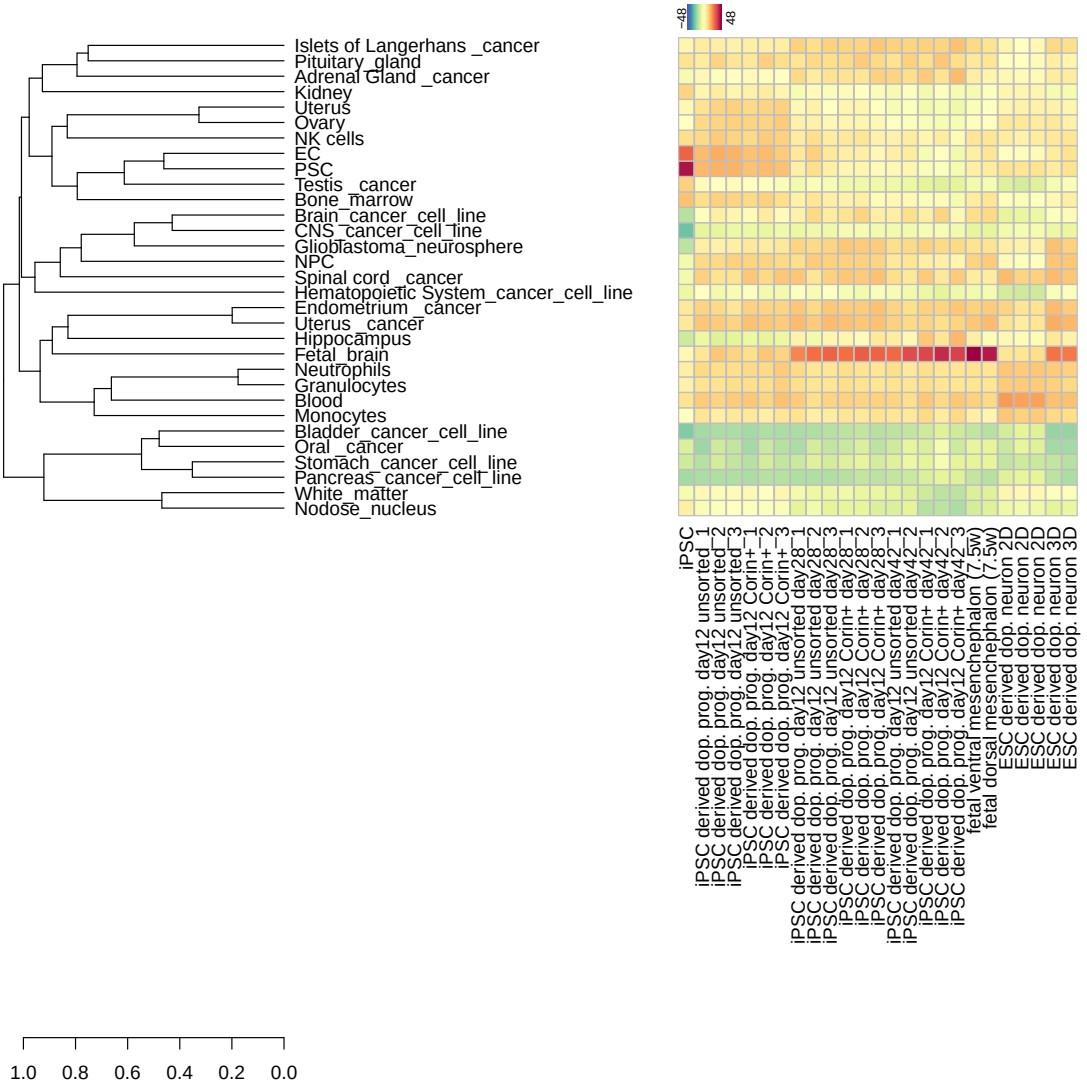


Figure 4.33.: Mapping of *in vitro* differentiated neurons to the residual tissue specific space. The *in vitro* differentiation time course from GSE51214 shows increasing similarity to fetal brain tissue, ending at similar expression patterns as the fetal mesenchyphalon from post-conceptional week 7.5 (column 1-20). Differentiation of pluripotent stem cells towards neurons in a 3D culture produces cells that are more similar to neurons than a similar differentiation in a 2D culture system (last 5 columns). Both differentiations were accomplished for 28 days (GSE43235).

3D differentiation produces cells that are similar to the differentiated cells of the previous study at day 28, the 2D culture system produces cells that have only very slight similarities to neural cells (Fig. 4.32 and Fig. 4.33).

Astrocytes and neural progenitors

Two other brain related cell types that receive much interest in studies of *in vitro* transformed cells are neural stem or progenitor cells and astrocytes, a special form of neuroglia. In addition to the *in vitro* differentiated neurons discussed above, we therefore analysed three further studies. One study investigates the *in vitro* differentiation of fetal brain (8 PCW) derived neural progenitor cells (NPCs) towards progenitor derived astrocytes (GSE43794) [210]. The second study (GSE36145) is concerned with the direct reprogramming of astrocytes to neural stem cells (NSCs). The third study (GSE43382) describes the reprogramming of fibroblasts to iPSCs and the subsequent differentiation of these towards NSCs, neurons, and astrocytes [211]. Furthermore, primary astrocytes are included for comparison purposes.

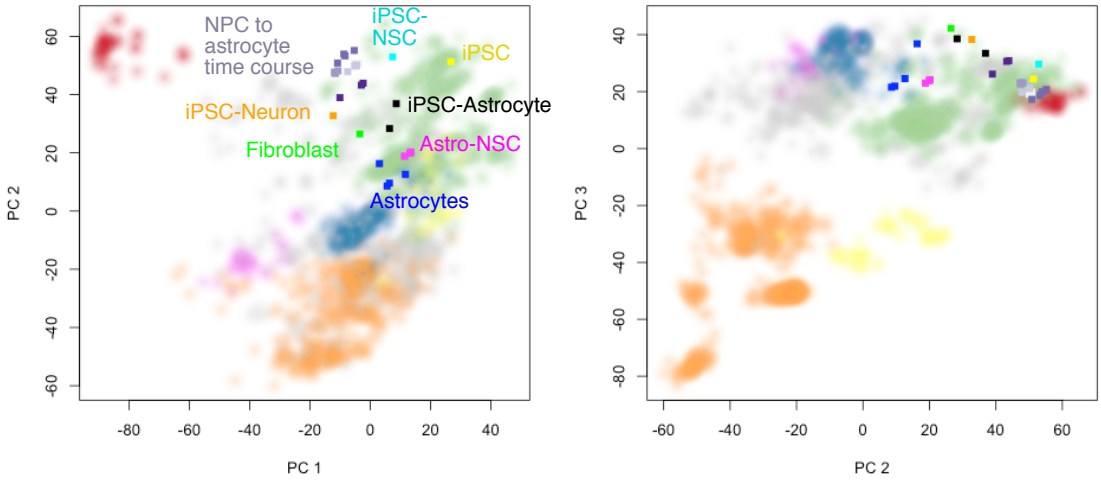


Figure 4.34.: *In vitro* differentiated astrocytes and neural progenitor/stem cells from datasets GSE43794, GSE36145, and GSE43382 mapped onto the PCA space. The NPC to astrocyte differentiation (GSE43794) induces almost no changes on the PCA space and even after 30 days (dark violet points) the NPC derived astrocytes are still relatively far away from primary astrocytes (blue points). The iPSC derived astrocytes from GSE43382 (black points) are more similar to primary astrocytes, although there is still some difference. Similarly, astrocyte derived NSCs (magenta points) are still very similar to the primary astrocytes from which they were derived (blue points) and relatively far apart from fetal brain derived neural progenitor cells (light grey points). iPSC derived NSCs (cyan point) are more similar to the NPCs.

From the mapping to the PCA space (Fig. 4.34), it was observed that the NPC to astrocyte differentiation time course (GSE43794) does not show many changes in the global expression status. Even so the cells at the latest time point (day 30 of differentiation) tend slightly in the direction were the primary astrocytes lie, they are still relatively far apart from them. The iPSC-derived astrocytes are more similar to the primary astrocytes,

although there are still clear differences. Similarly, the astrocyte-derived NSCs are still very similar to the primary astrocytes themselves and far apart from the fetal brain derived NPCs. The iPSC-derived NSCs are again more similar to the target cell type.

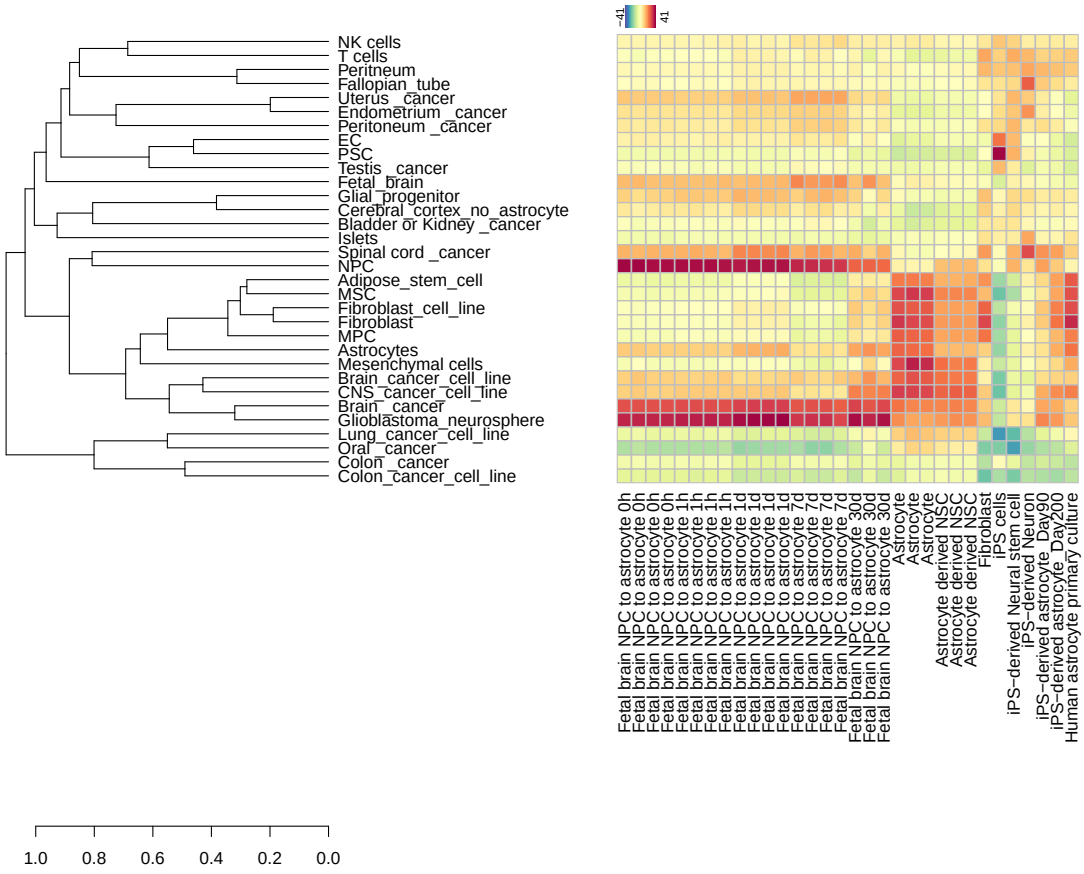


Figure 4.35.: *In vitro* differentiated astrocytes and neural progenitor/stem cells mapped onto the residual tissue specific space. The differentiation time course from fetal brain derived neural progenitor cells (NPCs) does not show any significant changes on the residual tissue specific space (columns 1 to 19). Astrocyte derived NSCs are more similar to neural progenitor cells than the astrocytes from which they were derived. However, a relatively strong similarity to astrocytes remains (columns 20 to 25). Reprogramming of fibroblasts to iPSCs and subsequent differentiation to NSCs, neurons, or astrocytes (GSE43382, last 7 columns) shows clear differences between the individual cell types. The respective cell types are nicely recognised by the residual tissue specific space, indicating a proper reprogramming and differentiation, although some unexpected signatures, e.g. fallopian tube for the neural cells, have relatively high scores.

Similar results can be seen on the residual tissue specific space. Here, it is confirmed that the NPC to astrocyte differentiation as well as the direct reprogramming of astrocytes to NSCs lead to suboptimal results, whereas the differentiation from iPSCs shows markedly

better results. This observation that trans-differentiated cells are less similar to the target cell type compared to iPSC derived cells was also made using CellNet [81].
These examples show nicely that despite of an activation of specific molecular markers as well as characteristic changes in the morphology of *in vitro* transformed cells [210], the global gene expression may be still clearly distinct from the target cell type. This is a nice example of complementary information that can be gained using global expression analyses, such as the mapping onto the two-scale landscape presented here.

5. Discussion and directions for future work

The first part of this chapter is concerned with the two-scale map of gene expression. It discusses the general idea of a two-scale decomposition and comments on a possible extension to further scales, resulting in a multi-scale decomposition of the gene expression space.
The second part considers the quality control of *in vitro* transformed cells. It describes the different aspects of a good quality that have different relevances for different fields of application.
In the third part of this chapter, the relevance of the two-scale decomposition approach for other research fields is described.

5.1. From global to local effects - Using multi-scale decompositions in gene expression analysis

In this thesis, we have developed a two-scale map of gene expression by a decomposition into a global scale, capturing common processes of different tissues or cell types and a finer scale, distinguishing between different tissues or cell types. We could show on the example of breast cancer samples that such a decomposition is very helpful in order to preserve an influence of confounding effects, such as a higher proliferation rate or general culturing associated differences, on tissue specific expression patterns. Similar effects have also been described for breast cancer outcome, which was supposed to be associated with an ESC signature [212]. Subsequent analyses could then show that almost all signatures are significantly associated with breast cancer outcome, since there is an underlying global effect, namely proliferation, which changes the expression of large amounts of genes [68]. Further details on the confounding influence of variation in large scales are discussed in [11].
Several methods have been developed so far that are dealing with the problem of confounding, non-phenotype associated variability. Leek et al. [187] propose a multi-step procedure, called surrogate variable analysis, which learns the main directions of non-phenotype associated variability using PCA on the residual vector after excluding phenotype associated effects. In a recent publication, we have proposed a method that uses a spherical transformation to reduce variability from rather large-scale non-phenotype associated effects [72]. This method was additionally very useful to generate proper null distributions in sample-

permutation based approaches for assessment of significant directional associations.
Both of these methods are very useful to account for confounding non-phenotype associated variability in the data. However, confounding effects, like culturing induced differences, are not always found as variability in the data under study or cannot be detected by the above described methods due to improper randomisation, i.e. a perfect correlation of the confounding variable with the variable of interest. This can be the case, for instance, when we compare cell line expression data to the expression of various tissues. In these cases, it is necessary to know the confounding effects a priori in order to be able to account for them. This can be done by a detection of these effects based on large scale retrospective data, as was done by Lukk et al [12]. This knowledge can then be used to subtract these effects from the data before analysing cells for their tissue relationship, as shown with the two-scale map in this thesis.
So far, we have restricted our analyses to a two-scale decomposition. In general, it would also be possible to extend the approach to even finer scales, i.e. to make a multi-scale decomposition. However, there are two reasons why this is not straightforward. First, in order to do this, it is necessary to have appropriate data with the respective annotation of the finer scale effects in order to relate these directions to biological knowledge. Second, it is not clear to which extent finer scales exist, i.e. how large the intrinsic dimensionality of gene expression data actually is. Third, in the case that finer scales exist, it is questionable whether they can be appropriately distinguished from noise. In fact, it has been observed that some biological phenotypes are very hard to detect in gene expression data, irrespective of the mathematical method applied [213], and that lists of differentially expressed genes are often very unstable for these phenotypes [83].
One principal possibility to extend the two-scale map by a further scale is to determine the residuals of the second mapping and to analyse these using gene set enrichment approaches or to focus on single genes that show a large deviation. This requires the use of the linear mapping approach, in order to have properly defined residual values. Furthermore, it requires to use some regularised least squares fitting due to the existing correlation in the residual space, which leads in practice to an over-fitting of the data (appendix A).
Thus, it is in principle possible to have a third scale and it may make some sense. However, there are two reasons that prevent a robust application of this approach. The first reason is that the noise level in the data will be too high in many cases. This is due to relatively large lab dependent effects, biological heterogeneity, or measurement noise, which in the past even led to a common belief that different microarray studies can usually not be analysed together.
A second reason is that the linear model applied is inappropriate in several cases, as we could show using the mixture data. For these data we calculated the residuals and saw a clear non-linear effect. This non-linear effect dominates a residual gene ontology analysis in some cases (Fig. 4.13), detecting gene ontology terms that are related to one of the mixed tissue types (data not shown). Thus, there will be no important additional biological information revealed by this analysis.

However, as we have shown, a residual analysis can be of high importance to detect non-linear effects, as we have done, and to improve the model accordingly in the future. This will be especially important when analysing mixture effects.

5.2. Quality control of *in vitro* reprogrammed and differentiated cells

There are many different methods to determine the quality of *in vitro* transformed cells, including functional, molecular and morphological criteria. However, as already discussed in chapter 2, these markers can have conflicting outcomes, leaving some questions about the actual quality of the cells open.

We introduced a method for the characterisation and quality control of *in vitro* transformed cells based on large scale gene expression data, comparing the overall gene expression state of the transformed cells to that of many different tissues and cell types. This method provides complementary information to the commonly used single gene or functional assays as we could show, e.g. at the example of converting neural progenitor cells to astrocytes (section 4.4.3). While commonly used markers indicate a successful transformation of the cells [210], our method indicates that the global gene expression status did not change very strongly and that even after 30 days of differentiation the cells were still very similar to neural progenitor cells.

This complementary information is very important for judgements about the quality of the generated cells and it should be part of the standard quality control of *in vitro* transformed cells. Despite of this importance of the overall expression of the cells, we want to point out that other quality criteria are important as well and that information gained from different criteria should be considered as complementary. For example, information about mutations at specific disease-relevant genes can usually not be detected using such global gene expression criteria, unless the mutation leads to a transformation of the cells, e.g. to a switch from one attractor region to another. However, such mutational information can be very important especially for safety concerns in regenerative medicine.

Another example of complementary information are DNA methylation profiles. These profiles could for instance provide information about the cellular age or senescence [200], which could be very relevant for decisions about the applicability of the cells for modelling of age-related diseases. Interestingly, in reprogramming and differentiation experiments it can be observed that epigenetic signs of senescence and ageing are reversed upon reprogramming to iPSCs [202, 203, 214] and that re-differentiated cells stay rejuvenated and acquire epigenetic signs of cellular senescence continuously over time [200].

Apart from these complementary information, it is also important to consider some potentially confounding effects that may influence the interpretation of the results. Such effects could be the different environmental conditions of cultured cells compared to primary tissues, as well as the mix of different cell types in primary tissues [81]. This has an influence

on the overall gene expression and should be kept in mind when interpreting the results. Therefore, it is important to decide about the quality of the generated cells in an application-dependent manner, using different quality criteria in combination. This will be exemplified in the following based on two main classes of applications in stem cell research. These are the use of *in vitro* transformed cells for regenerative medicine or for disease modelling.
For regenerative medicine, it is of interest whether the *in vitro* generated cells behave similarly to their *in vivo* counterparts in the moment they are injected into a human, i.e. when they are back in the *in vivo* environment. Cultured human keratinocytes, for instance, have a different molecular state than their *in vivo* counterparts but are still successfully used for burn therapy to regenerate the skin [215]. Thus, it is not of primary interest whether the cells change their molecular state or their behaviour when they are cultured *in vitro*, as long as this change is reversed when they are back in the *in vivo* environment. In contrast, for drug screenings and disease models, the *in vitro* behaviour of the cells should resemble the *in vivo* behaviour, in order to get a disease model that is useful for an accurate prediction of the drug effect. However, the large amounts of failures in clinical studies that could not be predicted in preclinical research [216] as well as explicit studies comparing 2D and 3D culture conditions [217, 218, 219, 220] indicate that the environment strongly effects the cellular behaviour and drug response.
Changes in the genome may occur during reprogramming, especially when vector DNA integrates into chromosomal DNA, and were observed to take place during long term culturing of pluripotent stem cells [45]. Such changes are not reversed in the *in vivo* environment and are therefore a major safety concern in regenerative medicine, potentially promoting a malignant transformation of the cells. While genomic changes can also have an impact on disease modelling, concerns about a possible future malignant transformation do not have any relevance for this application. Thus, as long as a mutation is not associated to the studied disease, it has usually a rather low impact for disease modelling and may be neglected in many cases.
Finally, the developmental immaturity that is often observed for *in vitro* differentiated cells is mainly a concern for modelling of age-related diseases, while the reprogramming-associated rejuvenation of cells may be regarded as a positive effect for regeneration of old-aged and senescent organs or tissues.
Further new technologies will contribute to this field of characterising cells in the near future. One example are single cell gene expression measurements [221] which became feasible in the recent years. Especially the possibility of single cell RNA sequencing is very promising for a whole-transcriptome based characterisation of individual cells, instead of measuring the expression of thousands of cells simultaneously [221]. These analyses will probably lead to new insights.
However, so far they are still at their infancy and there exist many problems with the dynamic changes in expression during the cell cycle, as well as stochastic effects [222, 223]. Furthermore, even if these problems can be solved in the future, a generic problem may be that the expression of these cells cannot be measured without isolating the cells before.

Thus, they are put into a different environment, potentially affecting their gene expression. It remains to be seen how strongly these cells are affected by this procedure. However, there may be a general problem of changing the objects that we want to measure through the measurement procedure itself, which is somewhat similar to the uncertainty relationship in physics [224]. This problematic has been described by Theise and Krause already in 2002 [17] and may introduce another level of complexity when trying to strictly define what we mean by a cell type and how the quality of transformed cells can be assessed.

5.3. Relevance for other biomedical research fields

In this thesis we focused on the application of the two-scale map to *in vitro* engineered cells. However, the map is certainly not restricted to this application field. A proper characterisation of cells is very important for development of cancer therapies and for optimising individual therapies in a personalised way. Furthermore, as already indicated above, many cancer researchers are interested in the differentiation status of cancer cells, and try to find associations with ESC-like patterns. In this respect, it is of specific interest whether the similarity to ESCs is specific or whether it is rather a side effect due to correlations with other phenotypic changes. These two possibilities can be nicely distinguished using the two-scale map, or even by the direct one-scale approach, as was shown in Lenz et al. [72]. For some types of cancer we could not validate previous findings of ESC similarities [72], e.g. for prostate cancer [225], as well as for the previously described case of breast cancer [212]. However, for another kind of cancer, namely Burkitt lymphomas, we could detect an association to ESCs [226]. Such an effect could also be seen in the PluriTest method, but it could at the same time be shown that the cells are still far away from being pluripotent [226].

In the case of comparing cancers from different stages, where we have usually many samples, it is important to have some significance assessment, which could be achieved using sample label permutation [72]. For this, it is helpful to use a spherical transformation for data preprocessing as indicated in section 5.1 [72]. This improves the generation of a null distribution especially in the case where we have a highly elliptical data distribution [72].

Apart from a direct use of the two-scale map in other research fields, it is also possible to use the general idea of a multi-scale decomposition for other applications. One example would be the analysis of stress response data. The response to drugs is to some extend cell type-specific, i.e. it depends on the cells to which the drug is applied. Therefore, it may be helpful to make a two-scale decomposition, distinguishing between generic cell type specific stress responses and drug specific effects. This is especially important since cell lines are clearly different from real tumor tissues. Thus, it would be highly beneficial to know which part of the response is drug specific and which depends on the cells to which the drug is applied. However, these effects have to be studied in more detail in future work, in order to see whether a two-scale decomposition may be beneficial for this type of application.

Another example is the usage of such an approach in disease progression, e.g. for chronic

myeloid leukaemia, where the mixture fractions of specific cell types change over time and where these mixture effects dominate changes in individual cells [82].

6. Summary and conclusion

We created a two-scale map of global gene expression that consists of a PCA based three-dimensional space and a residual tissue specific space. The PCA space was already described by Lukk et al. [12] and distinguishes mainly between hematopoietic tissues, cell lines, and neural tissues. Accordingly, the first PC is associated with hematopoietic cells, the second is mainly associated with the proliferation status of the cells, distinguishing between cell lines and primary tissues, and the third PC separates brain tissues from all others.

Based on an information theoretic view and utilising the information ratio criterion [83], we could show that the three dimensional PCA fails to capture a large amount of tissue specific information. Therefore, we extended this three-dimensional space by a residual tissue-specific space generated in a supervised manner using the classification of the samples into 369 different tissues, cell lines, or disease states [12]. Using a relatively simple and easy to interpret fold-change based method, we defined the 369 coordinates of the residual tissue specific space as the vectors pointing from the PCA space orthogonally to the mean of the respective tissue or cell type. This procedure results in 369 directions that are not necessarily orthogonal to each other. However, we could show that the correlation between the 369 vectors could be significantly reduced through the PCA based decomposition into two different spaces.

Using the same method on a dataset we collected independently leads to a similar decomposition. We could show that the first three PCs of this alternative dataset span a similar space as the first three PCs of the Lukk [12] dataset. This hints towards a certain stability of these dimensions. However, we could also show that a PCA-based decomposition is in general highly dependent on the composition of samples in the dataset. This effect was particularly pronounced when comparing further PCs of the two datasets as well as by specific choice of certain subsets of the data.

Apart from this investigation, the second dataset was very useful for the analysis of *in vitro* transformed neural cells, since it incorporates samples from neural progenitors as well as fetal brain, that were highly relevant for this applications but that are not present in the Lukk [12] dataset.

We described and explained two different methods of mapping new data onto the developed two-scale map and were able to correctly identify the tissue or cell type of several samples with these methods. Furthermore, we could show that the two-scale map can deal with mixture data, being able to identify a smooth transition from one tissue type to the other in artificial mixture data. While our method outperforms existing methods

(Concordia [11], URSA [10]) with respect to identification of mixtures, we point out that the non-linearities in the RNA content to measurement signal and especially the applied logarithmic transformation lead to a non-linear relationship between the mixture fraction and the preprocessed microarray data. This non-linearity is not captured by our linear model, as shown by the remaining non-linear structure in the residuals, and lead to clear differences between the two proposed mapping procedures. The non-linear mapping is better suited to detect low mixture fractions in the data, while the linear mapping is better suited for an estimation of mixture fractions. However, both mappings do not fully capture the described non-linearity. Therefore, there is space for improvements in future research in this direction.

We compared the two-scale landscape method to three existing methods of mRNA-based cell characterisation and showed that our method has some superior properties. The existing approaches are either less accurate (Concordia tool [11]), not well suited for characterisation of cells that do not fit exactly to a specific tissue or cell type (Concordia [11] and URSA [10] tools), or incorporate limited amounts of different cell types due to a high training data demand (CellNet tool [81]).

Existing dimension reduction approaches for large scale gene expression based cell characterisation focused either on unsupervised methods, i.e. principal components analysis [12], or on purely supervised methods [72]. We combined here supervised and unsupervised dimension reduction methods due to their complementary strengths and weaknesses.

With the two-scale map approach, we were able to extend the purely unsupervised approach by a more fine-scaled residual space, allowing a very detailed investigation of tissue or cell type similarity. We could also show that the method outperforms the purely supervised "one-scale" approach due to an increased information content through the decomposition into joint processes and tissue specific patterns. Furthermore, correlations between different tissues could be markedly reduced, making the interpretation of the results easier.

With the example of the breast cancer and liver mixture data, we could show that the PCA space covers much of the differences between cell lines and primary tissues, such that on the residual space, there is a relatively increased similarity of cell lines and their respective tissue of origin. This finding may be very important for drug development processes, where results from cell line assays are translated to the clinics.

We then applied the two-scale map to several different datasets of *in vitro* transformed cells. Using these analyses, we could show that *in vitro* differentiated cells are in most cases immature representatives of the desired cell types, having a high similarity to embryonic or fetal tissues of the respective type. Furthermore, we showed two examples where the transformation process was not successful at a global gene expression level, namely the differentiation of NPCs to astrocytes, as well as the direct reprogramming of astrocytes to NSCs. These examples show that the information gained from our global analysis gives additional complementary insights to the more commonly used single marker based or morphological information, leading sometimes to contradicting conclusions about the success of the transformation process.

Such contradictions lead to the question on how a good quality of the transformed cells can be defined. We discussed this question in general as well as more specifically for the case of pluripotent stem cells, coming to the conclusion that there is no single uniform property that can be used to assess the quality. Instead, the quality criteria should be used in an application-specific manner and complementary information should be combined to arrive at a more complete characterisation of the generated cells.
Finally, we propose that a two-scale decomposition of expression data may also be useful for other applications. Therefore, it will be very interesting to test similar approaches for drug response data or for the analysis of expression changes during disease progression. Due to the generality and wide applicability of our method, it is not restricted to the characterisation of engineered cells, but can also be used for characterising naturally occurring cells with unclear properties and mixture fractions, e.g. cancer cells.
Thus, we describe a general framework for the characterisation of cells that has a wide applicability and that can easily be extended to focus on specific cell types of interest. We expect that the proposed method will be extended and refined in the future to tackle the pressing needs of linking wet lab experiments to clinics and of distinguishing expression changes in single cells from population effects that are caused by changing mixture fractions.

Bibliography

[1] Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, et al. Embryonic stem cell lines derived from human blastocysts. *Science*, 282:1145–1147, 1998.

[2] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–676, 2006.

[3] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. Induction of pluripotent stem cells from human fibroblasts by defined factors. *Cell*, 131:861–872, 2007.

[4] Yamanaka S. Induced pluripotent stem cells: past, present, and future. *Cell Stem Cell*, 10:678–684, 2012.

[5] Synnergren J, Améen C, Jansson A, Sartipy P. Global transcriptional profiling reveals similarities and differences between human stem cell-derived cardiomyocyte clusters and heart tissue. *Physiol. Genomics*, 44:245–258, 2012.

[6] Buta C, David R, Dressel R, Emgård M, Fuchs C, et al. Reconsidering pluripotentcy tests: Do we still need teratoma assays? *Stem Cell Research*, 11:552–562, 2013.

[7] Ho AD, Wagner W, Franke W. Heterogeneity of mesenchymal stromal cell preparations. *Cytotherapy*, 10:320–330, 2008.

[8] Lee MJ, Ye AS, Gardino AK, Heijink AM, Sorger PK, et al. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signalling networks. *Cell*, 149:780–794, 2012.

[9] Marx V. Biology: The big challenges of big data. *Nature*, 498:255–260, 2013.

[10] Lee YS, Krishnan A, Zhu Q, Troyanskaya OG. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29:3036–3044, 2013.

[11] Schmid PR, Palmer NP, Kohane IS, Berger B. Making sense out of massive data by going beyond differential expression. *Proc Nat Acad Sci*, 109:5594–5599, 2012.

[12] Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, et al. A global map of human gene expression. *Nature Biotechnology*, 28:322–324, 2010.

[13] Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, et al. An estimation of the number of cells in the human body. *Ann Hum Biol.*, 40:463–471, 2013.

[14] Galvão V, Miranda JGV, Andrade RFS, Andrade Jr. JS, Gallos LK, Makse HA. Modularity map of the network of human cell differentiation. *Proc Natl Acad Sci*, 107:5750–5755, 2010.

[15] Vickaryous MK, Hall BK. Human cell type diversity, evolution, development and classification with special reference to cells derived from the neural crest. *Biol Rev*, 81:425–455, 2006.

[16] Department of Health and Human Services. Regenerative Medicine, August 2006.

[17] Theise ND, Krause DS. Toward a new paradigm of cell plasticity. *Leukemia*, 16:542–548, 2002.

[18] Fisher AG. Cellular identity and lineage choice. *Nature Reviews Immunology*, 2:977–982, 2002.

[19] Hall PA, Watt FM. Stem cells: the generation and maintenance of cellular diversity. *Development*, 106:619–633, 1989.

[20] Holmberg J, Perlmann T. Maintaining differentiated cellular identity. *Nature Reviews Genetics*, 13:429–439, 2012.

[21] Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol*, 26:101–106, 2008.

[22] Park IH, Zhao R, West JA, Yabuuchi A, Huo H, et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature*, 451:141–146, 2008.

[23] Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318:1917–1920, 2007.

[24] González F, Boué S, Belmonte JCI. Methods for making induced pluripotent stem cells: reprogramming à la carte. *Nature Reviews Genetics*, 12:231–242, 2011.

[25] Ho R, Chronis C, Plath K. Mechanistic insights into reprogramming to induced pluripotency. *J Cell Physiol*, 226:868–878, 2011.

[26] Federation AJ, Bradner JE, Meissner A. The use of small molecules in somatic-cell reprogramming. *Trends in Cell Biology*, pii: S0962-8924(13)00168-2:doi: 10.1016/j.tcb.2013.09.011, 2013.

[27] Kim JB, Zaehres H, Wu G, Gentile L, Ko K, et al. Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature*, 454:646–650, 2008.

[28] Kim JB, Greber B, Araúzo-Bravo MJ, Meyer J, Park KI, et al. Direct reprogramming of human neural stem cells by OCT4. *Nature*, 461:649–653, 2009.

[29] Yamanaka S. Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell Stem Cell*, 1:39–49, 2007.

[30] Schuldt BM, Guhr A, Lenz M, Kobold S, MacArthur BD, et al. Power-laws and the use of pluripotent stem cell lines. *PLoS ONE*, 8:e52068, 2013.

[31] Scott CT, McCormick JB, Derouen MC, Owen-Smith J. Federal policy and the use of pluripotent stem cells. *Nat Methods*, 7:866–867, 2010.

[32] Grskovic M, Javaherian A, Strulovici B, Daley GQ. Induced pluripotent stem cells - opportunities for disease modelling and drug discovery. *Nature Reviews Drug Discovery*, 10:915–929, 2011.

[33] Inoue H, Yamanaka S. The use of induced pluripotent stem cells in drug development. *Clinical Pharmacology & Therapeutics*, 89:655–661, 2011.

[34] Keller G. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes & Dev.*, 19:1129–1155, 2005.

[35] Wu SM, Hochedlinger K. Harnessing the potential of induced pluripotent stem cells for regenerative medincine. *Nature Cell Biology*, 13:497–505, 2011.

[36] Haupt S, Wanek P, Marx U, Zenke M, Brüstle O. Industrialisierung der iPSC Herstellung. *Laborwelt*, 4:26–27, 2012.

[37] Kunkanjanawan T, Noisa P, Parnpai R. Modeling neurological disorders by human induced pluripotent stem cells. *Journal of Biomedicine and Biotechnology*, 2011:Article ID 350131, 2011.

[38] Ebert AD, Svendsen CN. Stem Cell Model of Spinal Muscular Atrophy. *Arch. Neurol.*, 67:665–669, 2010.

[39] Ebert AD, Yu J, Rose FF, Mattis VB, Lorson CL, et al. Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature*, 457:277–280, 2009.

[40] Malan D, Friedrichs S, Fleischmann BK, Sasse P. Cardiomyocytes obtained from induced pluripotent stem cells with long-QT syndrome 3 recapitulate typical disease-specific features in vitro. *Circ. Res.*, 109:841–847, 2011.

[41] Engle SJ, Puppala D. Integrating human pluripotent stem cells into drug development. *Cell Stem Cell*, 12:669–677, 2013.

[42] Marx U, Schenk F, Behrens J, Meyr U, Wanek P, et al. Automatic Production of Induced Pluripotent Stem Cells. *First CIRP Conference on Biomanufacturing. Procedia CIRP*, 5:2–6, 2013.

[43] Andrews PW, Matin MM, Bahrami AR, Damjanov I, Gokhale P, Draper JS. Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin. *Biochem Soc Trans*, 33:1526–1530, 2005.

[44] Blasco MA, Serrano M, Fernandez-Capetillo O. Genomic instability in iPS: time for a break. *The EMBO Journal*, 30:991–993, 2011.

[45] Laurent LC, Ulitsky I, Slavin I, Tran H, Schork A, et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell*, 8:1–13, 2011.

[46] Liu Z, Zhou J, Wang H, Zhao M, Wang C. Current status of induced pluripotent stem cells in cardiac tissue regeneration and engineering. *Regenerative Medicine Research*, 2013:1–6, 2013.

[47] Oricchio E, Papapetrou EP, Lafaille F, Ganat YM, Kriks S, et al. A cell engineering strategy to enhance the safety of stem cell therapies. *Cell Reports*, 8:1677–1685, 2014.

[48] Doi D, Morizane A, Kikuchi T, Onoe H, Hayashi T, et al. Prolonged maturation culture favors a reduction in the tumorigenicity and the dopaminergic function of human ESC-derived neural cells in a primate model of Parkinson's disease. *Stem Cells*, 30:935–945, 2012.

[49] Amariglio N, Hirshberg A, Scheithauer BW, Cohen Y, Loewenthal R, et al. Donor-derived brain timor following neural stem cell transplantation in an ataxia telangiectasia patient. *PLoS Med*, 6:e1000029, 2009.

[50] Conrad S, Renninger M, Hennenlotter J, Wiesner T, Just L, et al. Generation of pluripotent stem cells from adult human testis. *Nature*, 456:344–349, 2008.

[51] Ko K, Araúzo-Bravo MJ, Tapia N, Kim J, Lin Q, et al. Human adult germ line stem cells in question. *Nature*, 465:E1; discussion E3, 2010.

[52] Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. *Mol Syst Biol*, 9:640, 2013.

[53] Baker M. Gene data to hit milestone. *Nature*, 487:282–283, 2012.

[54] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. NCBI GEO: archive for functional genomics data set – 10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–10, 2011.

[55] Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. ArrayExpress update - from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37:D868–D872, 2008.

[56] Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet*, 14:333–346, 2013.

[57] Schneider MV, Jimenez RC. Teaching the fundamentals of biological data integration using classroom games. *PLoS Comput Biol*, 8:e1002789, 2012.

[58] Smyth GK. Limma: linear models for microarray data. In Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editor, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[59] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98:5116–5121, 2001.

[60] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 13:163, 1997.

[61] The UniProt Consortium. Activities at the Universal Protein Resources (UniProt). *Nucleic Acids Res.*, 42:D191–D198, 2014.

[62] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8:e1002375, 2012.

[63] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes. *Nat Genet*, 34:267–273, 2003.

[64] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci.*, 102:15545–15550, 2005.

[65] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.

[66] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[67] Kauffman SA. Metabolic stability and epigenesist in randomly connected nets. *J Theoret Biol*, 22:437–467, 1969.

[68] Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, 7:e1002240, 2011.

[69] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313:1929–1935, 2006.

[70] Müller F-J, Schuldt BM, Williams R, Mason D, Altun G, et al. A bioinformatic assay for pluripotency in human cells. *Nature Methods*, 8:315–317, 2011.

[71] Williams R, Schuldt B, Müller F-J. A guide to stem cell identification: Progress and challenges in system-wide predictive testing with complex biomarkers. *BioEssays*, 33:880–890, 2011.

[72] Lenz M, Schuldt BM, Müller F-J, Schuppert A. PhysioSpace: Relating gene expression experiments from heterogeneous sources using shared physiological processes. *PLoS ONE*, 8:e77627, 2013.

[73] Rodriguez-Gonzalez FG, Mustafa DAM, Mostert B, Sieuwerts AM. The challenge of gene expression profiling in heterogeneous clinical samples. *Methods*, 59:47–58, 2013.

[74] Bellman RE. *Adaptive control processes: A guided tour*. Princeton University Press, Princeton, 1961.

[75] Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer*, 8:37–49, 2008.

[76] Wang Y, Miller DJ, Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. *British Journal of Cancer*, 98:1023–1028, 2008.

[77] Millman VD. A new proof of the theorem of A. Dvoretzky on sections of convex bodies. *Funct Anal Appl*, 5:28–37, 1971.

[78] Radovanović M, Nanopoulos A, Ivanović M. Hubs in space: popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

[79] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference and prediction.* Springer Series in Statistics, 2nd edition, 2009.

[80] Skillicorn DB. *Understanding high-dimensional spaces.* Springer Briefs in Computer Science, New York, 2012.

[81] Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ. CellNet: network biology applied to stem cell engineering. *Cell*, 158:903–915, 2014.

[82] Schuppert A, Koschmieder S, Montazeri M, Bennemann K, Copland M, et al. Combining population dynamics and entropy modelling reveals critical states in CML disease progression. submitted.

[83] Schneckener S, Arden NS, Schuppert A. Quantifying stability in gene list ranking across microarray derived clinical biomarkers. *BMC Medical Genomics*, 4(73), 2011.

[84] Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biology*, 6:R21, 2005.

[85] Schwartz SM. The definition of cell type. *Circulation Research*, 84:1234–1235, 1999.

[86] Slater MH. Cell types as natural kinds. *Biological Theory*, 7:170–179, 2013.

[87] Huang S, Eichler G, Bar-Yam Y, Ingber DE. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.*, 94:128701, 2005.

[88] Wakao S, Kitada M, Kuroda Y, Ogura F, Murakami T, et al. Morphologic and gene expression criteria for identifying human induced pluripotent stem cells. *PLoS ONE*, 7:e48677, 2012.

[89] MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006.

[90] Mardis ER. A decade's perspective on DNA sequencing technology. *Nature*, 470:198–203, 2011.

[91] Nawy T. Single-cell sequencing. *Nature Methods*, 11:doi:10.1038/nmeth.2771, 2014.

[92] Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29:2211–2212, 2013.

[93] Zuckerman NS, Noam Y, Goldsmith AJ, Lee PP. A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS Comput Biol*, 9:e1003189, 2013.

[94] Affymetrix company. Affymetrix services price list, accessed at December 11, 2013. `http://www.microarrays.ca/services/Affymetrix_Prices.pdf`.

[95] CCHMC/UC Pluripotent Stem Cell Facility. Human iPSC derivation costs: fibroblasts, accessed at December 11, 2013. `https://research.cchmc.org/stemcell/sites/bmidrupalpstemcell.chmcres.cchmc.org/files/admin/contributor/docs/Fibro_iPSC_fees.pdf`.

[96] Zhang WY, de Almeida PE, Wu JC. Teratoma formation: A tool for monitoring pluripotency in stem cell research. In *StemBook [Internet]*. Cambridge (MA): Harvard Stem Cell Institute, 2012. Available from: `http://www.ncbi.nlm.nih.gov/books/NBK133287/`.

[97] Ramirez J-M, Gerbal-Chaloin S, Milhavet O, Qiang B, Becker F, et al. Benchmarking Human Pluripotent Stem Cell Markers During Differentiation Into the Three Germ Layers Unveils a Striking Heterogeneity: All Markers Are Not Equal. *Stem cells*, 29:1469–1474, 2011.

[98] Kang L, Wang J, Zhang Y, Kou Z, Gao S. iPS cell can support full-term development of tetraploid blastocyst-complemented embryos. *Cell Stem Cell*, 5:135–138, 2009.

[99] Tam PPL, Rossant J. Mouse embryonic chimeras: tools for studying mammalian development. *Development*, 130:6155–6163, 2003.

[100] Müller F-J, Goldmann J, Loser P, Loring JF. A call to standardize teratoma assays used to define human pluripotent cell lines. *Cell Stem Cell*, 6:412–414, 2010.

[101] Gropp M, Shilo V, Vainer G, Gov M, Gil Y, et al. Standardization of the teratoma assay for analysis of pluripotency of human ES cells and biosafety of their differentiated progeny. *PLoS ONE*, 7:e45532, 2012.

[102] Sheridan SD, Surampudi V, Rao RR. Analysis of embryoid bodies derived from human induced pluripotent stem cells as a means to assess pluripotency. *Stem Cells International*, 2012:Article ID 738910, 2012.

[103] Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci USA*, 110:3507–3512, 2013.

[104] Zhao W, Ji X, Zhang F, Li L, Ma L. Embryonic stem cell markers. *Molecules*, 17:6196–6236, 2012.

[105] Chan EM, Ratanasirintrawoot S, Park I-H, Manos PD, Loh Y-H, et al. Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nature Biotechnology*, 27:1033–1037, 2009.

[106] Dolgin E. Putting stem cells to the test. *Nature Medicine*, 16:1354–1357, 2010.

[107] Goldmann JE, Schuldt BM, Lenz M, Müller FJ. PluriTest molecular diagnostic assay for pluripotency in human stem cells. In Loring JF, Peterson SE, editor, *Human stem cell manual, second edition*. Elsevier Inc, 2012.

[108] Müller F-J, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, 455:401–406, 2008.

[109] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[110] Lee DD, Seung SH. Algorithms for nonnegative matrix factorization. *Neural Inform. Process. Syst.*, 13:556–562, 2001.

[111] Shao K, Koch C, Gupta MK, Lin Q, Lenz M, et al. Induced pluripotent mesenchymal stromal cell clones retain donor-derived differences in DNA methylation profiles. *Molecular Therapy*, 21:240–250, 2012.

[112] MacArthur BD, Sevilla A, Lenz M, Mueller F-J, Schuldt BM, et al. Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. *Nat Cell Biol*, 14:1139–1147, 2012.

[113] Bibikova M, Chudin E, Wu B, Zhou L, Garcia EW, et al. Human embryonic stem cells have a unique epigenetic signature. *Genome Res*, 16:1075–1083, 2006.

[114] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454:766–770, 2008.

[115] Nazor KL, Altun G, Lynch C, Tran H, Harness JV, et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell*, 10:620–634, 2012.

[116] Lenz M, Goetzke R, Schenk A, Schubert C, Veeck J, et al. Epigenetic biomarker to support classification into pluripotent and non-pluripotent cells. *Scientific Reports*, in revision.

[117] Bulić-Jakuš F, Ulamec M, Vlahović M, Sinčić N, Katušić A, et al. Of mice and men: teratomas and teratocarcinomas. *Coll Antropol*, 30:921–924, 2006.

[118] Polejaeva I, Mitalipov S. Stem cell potency and the ability to contribute to chimeric organisms. *Reproduction*, 145:R81–R88, 2013.

[119] Solter D. From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. *Nature Reviews Genetics*, 7:319–327, 2006.

[120] Blelloch RH, Hochedlinger K, Yamada Y, Brennan C, Kim M, et al. Nuclear cloning of embryonal carcinoma cells. *Proc Natl Acad Sci USA*, 101:13985–13990, 2004.

[121] Damjanov I, Andrews PW. The terminology of teratocarcinomas and teratomas. *Nature Biotechnology*, 25:1212, 2007.

[122] Germain ND, Hartman NW, Cai C, Becker S, Maegele JR, Grabel LB. Teratocarcinoma formation in embryonic stem cell-derived neural progenitor hippocampal transplants. *Cell Transplant*, 21:1603–1611, 2012.

[123] Hovatta O, Jaconi M, Töhönen V, Béna F, Gimelli S, et al. A teratocarcinoma-like human embryonic stem cell (hESC) line and four hESC lines reveal potentially oncogenic genomic changes. *Cell Transplant*, 21:1603–1611, 2012.

[124] Ben-David U, Benvenisty N. The tumorigenicity of human embryonic and induced pluripotent stem cells. *Nature Reviews Cancer*, 11:268–277, 2011.

[125] Yi L, Lu C, Hu W, Sun Y, Levine AJ. Multiple roles of p53-related pathways in somatic cell reprogramming and stem cell differentiation. *Cancer Res*, 72:doi:10.1158/0008–5472.CAN–12–1451, 2012.

[126] Bilic J, Belmonte JCI. Concise review: induced pluripotent stem cells versus embryonic stem cells: close enough or yet too far apart. *Stem Cells*, 30:33–41, 2012.

[127] Kim K, Zhao R, Doi A, Ng K, Unternaehrer J, et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. *Nat Biotechnol.*, 29:1117–1119, 2011.

[128] Bock C, Kiskinis E, Verstappen G, Gu H, Boulting G, et al. Referene maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, 144:439–452, 2011.

[129] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell.* Garland Science, New York, 2008.

[130] Campbell KH, McWhir J, Ritchie WA, Wilmut I. Sheep cloned by nuclear transfer from a cultured cell line. *Nature*, 380:64–66, 1996.

[131] Gurdon JB. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *J Embryol Exp Morphol*, 10:622–640, 1962.

[132] Kauffman S. A proposal for using the ensemble approach to understand genetic regulatory networks. *J Theoret Biol*, 230:581–590, 2004.

[133] MacArthur BD, Ma'ayan A, Lemischka IR. Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol*, 10:672–681, 2009.

[134] Huang S, Ingber DE. A non-genetic basis for cancer progression and metastasis: self-organising attractors in cell regulatory networks. *Breast Disease*, 26:27–54, 2007.

[135] Waddington CH. *The Strategy of the Genes.* London: Allen & Unwin, 1957.

[136] Ma T, Xie M, Laurent T, Ding S. Progress in the reprogramming of somatic cells. *Circ Res*, 112:562–574, 2013.

[137] Davies PCW, Demetrius L, Tuszynski JA. Cancer as a dynamical phase transition. *Theor Biol Med Model*, 8:30, 2011.

[138] MacArthur BD, Please CP, Oreffo ROC. Stochasticity and the molecular mechanisms of induced pluripotency. *PLoS ONE*, 3:e3086, 2008.

[139] Li C, Wang J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLoS Comput Biol*, 9:e1003165, 2013.

[140] Hanna J, Saha K, Pando B, van Zon J, Lengner CJ, et al. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462:595–601, 2009.

[141] Huang S. Reprogramming cell fates: reconciling rarity with robustness. *BioEssays*, 31:546–560, 2009.

[142] Rais Y, Zviran A, Geula S, Gafni O, Chomsky E, et al. Deterministic direct reprogramming of somatic cells to pluripotency. *Nature*, 502:65–70, 2013.

[143] Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453:544–547, 2008.

[144] Schena M, Shalon D, Davies RW, Brown PO. Quantitative motoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[145] Tu Y, Stolovitzky G, Klein U. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Science*, 99:14031–14036, 2002.

[146] Hardiman G. Microarray platforms - comparison and contrasts. *Pharmacogenomics*, 5:487–502, 2004.

[147] Wu Z. A review of statistical methods for preprocessing oligonucleotide microarrays. *Stat Methods Med Res*, 18:533–541, 2009.

[148] Schulze A, Downward J. Navigating gene expression using microarrays - a technology review. *Nature Cell Biology*, 3:E190–E195, 2001.

[149] Oberthuer A, Juraeva D, Li L, Kahlert Y, Westermann F, et al. Comparison of performance of one-color and two-colour gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. *Pharmacogenomics J*, 10:258–266, 2010.

[150] Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu T-M, et al. Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. *Nature Biotechnology*, 24:1140–1150, 2006.

[151] Holloway AJ, van Laar RK, Tothill RW, Bowtell DDL. Options available - from start to finish - for obtaining data from DNA microarrays II. *Nature Genetics*, 32:481–489, 2002.

[152] Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, 2:351–356, 2005.

[153] Pradervand S, Paillusson A, Thomas J, Weber J, Wirapati P, et al. Affymetrix Whole-Transcript Human Gene 1.0 ST array is highly concordant with standard 3' expression arrays. *BioTechniques*, 44:759–762, 2008.

[154] Harrison A, Binder H, Buhot A, Burden CJ, Carlon E, et al. Physico-chemical foundations underpinning microarray and next-generation sequencing experiments. *Nucleic Acids Research*, 2013:1–18, 2013.

[155] Binder H, Preibisch S. GeneChip microarrays - signal intensities, RNA concentration and probe sequences. *Journal of Physics: Condensed Matter*, 18:537–566, 2006.

[156] Binder H, Krohn K, Burden CJ. Washing scaling of GeneChip microarray expression. *BMC Bioinformatics*, 11:291, 2010.

[157] Burden CJ, Binder H. Physico-chemical modelling of target depletion during hybridization on oligonucleotide microarrays. *Phys. Biol.*, 7:016004, 2010.

[158] Affymetrix company. Latin square data for expression algorithm assessment, accessed at July 2, 2014. `http://www.affymetrix.com/support/technical/sample_data/datasets.affx`.

[159] Gharaibeh RZ, Fodor AA, Gibas CJ. Accurate estimates of microarray target concentration from a simple sequence-independent Langmuir model. *PLoS ONE*, 5:e14464, 2010.

[160] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Nat Acad Sci*, 98:31–36, 2001.

[161] Sáski R, Calvo E, Corbeil J. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics*, 18:1633–1640, 2002.

[162] Rocke DM, Durbin B. A model for measurement error for gene expression arrays. *J Comp Biol*, 8:557–569, 2001.

[163] Zeisel A, Amir A, Köstler WJ, Domany E. Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes. *BMC Bioinformatics*, 11:400, 2010.

[164] Affymetrix company. *Guide to probe logarithmic intensity error (plier) estimation.* Affymetrix, Inc, 2005. `http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf`.

[165] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[166] Hardin J, Wilson J. A note on oligonucleotide expression values not being normally distributed. *Biostatistics*, 10:446–450, 2009.

[167] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30:207–210, 2002.

[168] Langdon WB, Upton GJ, da Silva Camargo R, Harrison AP. A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Trans Comput Biol Bioinform*, 7:647–653, 2010.

[169] Serhal P, Lemieux S. Correction of spatial bias in oligonucleotide array data. *Adv Bioinformatics*, 2013:167915, 2013.

[170] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32:D267–D270, 2004.

[171] Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res*, 39:D507–D513, 2011.

[172] Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22:830–836, 2006.

[173] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. Large-scale mapping and validation of Escheria coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5:e8, 2007.

[174] Breiman L. Random Forests. *Mach Learn*, 45:5–32, 2001.

[175] Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[176] Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[177] Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11:574, 2010.

[178] Comon P. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[179] Schapire RE Collins M, Dasgupta S. A generalization of principal components analysis to the exponential family. In Ghahramani Z Dietterich TG, Becker S, editor, *Advances in Neural Information Processing Systems 14*, pages 617–624, 2001.

[180] Geiger BC, Kubin G. Signal enhancement as minimisation of relevant information loss. *Proc. ITG Conf. on Systems, Communication and Coding*, 2013.

[181] Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[182] Anderson TW. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34:122–148, 1963.

[183] Bai ZD. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, 9:611–677, 1999.

[184] Marčenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik*, 1:457–483, 1967.

[185] Bai ZD, Silverstein JW, Yin YQ. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26:166–168, 1988.

[186] Geman S. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8:252–261, 1980.

[187] Leek TJ, Storey JD. Capturing heterogeneity in gene expression studies by 'surrogate variable analysis'. *PLoS Genetics*, 3:e161, 2007.

[188] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article 3, 2004.

[189] Gauss CF. Theoria Combinationis Observationum Erroribus Minimis Obnoxiae. *Parts 1, 2 and suppl. Werke*, 4:1–108, 1821, 1823, 1826.

[190] Graybill FA. *Theory and applications of the linear model.* Duxbury, North Scituate, MA (Waldsworth and Brooks/Cole, Pacific Grove, CA), 1976.

[191] Liu Y, Slotine J, Barabási A. Controllability of complex networks. *Nature*, 473:167–173, 2011.

[192] Müller FJ, Schuppert A. Few inputs can reprogram biological networks. *Nature*, 478:E4, 2011. doi:10.1038/nature10543.

[193] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118–127, 2006.

[194] Turnbull AK, Kitchen RR, Larionov AA, Renshaw L, Dixon JM, Sims AH. Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. *BMC Medical Genomics*, 5:35, 2012.

[195] Gower AC, Spira A, Lenburg ME. Discovering biological connections between experimental conditions based on common patterns of differential gene expression. *BMC Bioinformatics*, 12:381, 2011.

[196] Jiang Z, Gentleman R. Extensions to gene set enrichment analysis. *Bioinformatics*, 23:306–313, 2007.

[197] Hasenberg M, Stegemann-Koniszewski S, Gunzer M. Cellular immune reactions in the lung. *Immunol Rev*, 251:189–214, 2013.

[198] Biedler JL, Roffler-Tarlov S, Schachner M, Freedman LS. Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones. *Cancer Res*, 38:3751–3757, 1978.

[199] Affymetrix company. Gene 1.0 ST Array Data Set, accessed at August 6, 2014. `http://www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx`.

[200] Frobel J, Hemeda H, Lenz M, Abagnale G, Joussen S, et al. Epigenetic rejuvenation of mesenchymal stromal cells derived from induced pluripotent stem cells. *Stem Cell Reports*, 3:1–9, 2014.

[201] Dominici M, Blanc KL, Mueller I, Slaper-Cortenbach I, Marini FC, et al. Minimal criteria for defining multipotent mesenchymal stromal cells. *Cytotherapy*, 8:315–317, 2006.

[202] Horvath S. DNA methylation age of human tissues and cell types. *Genome Biology*, 14:R115, 2013.

[203] Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology*, 15:R24, 2014.

[204] Haniffa MA, Collin MP, Buckley CD, Dazzi F. Mesenchymal stem cells: the fibroblasts' new clothes? *Haematologica*, 94:258–263, 2009.

[205] Sun N, Yazawa M, Liu J, Han L, Sanchez-Freire V, et al. Patient-specific induced pluripotent stem cells as a model for familial dilated cardiomyopathy. *Sci Transl Med*, 4:130ra47, 2012.

[206] Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. Spatio-temporal transcriptome of the human brain. *Nature*, 478:483–489, 2011.

[207] Dirks PB. Brain tumor stem cells: The cancer stem cell hypothesis writ large. *Molecular Oncology*, 4:420–430, 2010.

[208] Vescovi AL, Galli R, Reynolds BA. Brain tumor stem cells. *Nat Rev Cancer*, 6:425–436, 2006.

[209] Doi D, Samata B, Katsukawa M, Kikuchi T, Morizane A, et al. Isolation of human induced pluripotent stem cell-derived dopamiergic progenitors by cell sorting for successful transplantation. *Stem Cell Reports*, 2:337–350, 2014.

[210] Ferenczy MW, Johnson KR, Marshall LJ, Monaco MC, Major EO. Differentiation of human fetal multipotential neural progenitor cells to astrocytes reveals susceptibility factors for JC virus. *J Virol*, 87:6221–6231, 2013.

[211] Kondo T, Asai M, Tsukita K, Kutoku Y, Ohsawa Y, et al. Modeling Alzheimer's disease with iPSCs reveals stress phenotypes associated with intracellular Aβ and differential drug responsiveness. *Cell Stem Cell*, 12:487–496, 2013.

[212] Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genetics*, 40:499–507, 2008.

[213] Shi L, Campbell G, Jones WD, Campagne F, Wen Z, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28:827–838, 2010.

[214] Koch CM, Reck K, Shao K, Lin Q, Joussen S, et al. Pluripotent stem cells escape from senescence-associated DNA methylation changes. *Genome Research*, 23:248–259, 2013.

[215] Fuchs E. The impact of cell culture on stem cell research. *Cell Stem Cell*, 10:640–641, 2012.

[216] Arrowsmith J. Trial watch: Phase II failures: 2008-2010. *Nat Rev Drug Discovery*, 10:328–329, 2011.

[217] Bissell MJ, Radisky DC, Rizki A. The organizing principle: microenvironmental influences in the normal and malignant breast. *Differentiation*, 70:537–546, 2002.

[218] Myers TA, Nickerson CA, Kaushal D, Ott CM, Höner zu Bentrup K, et al. Closing the phenotypic gap between transformed neuronal cell lines in culture and untransformed neurons. *J Neurosci Methods*, 174:31–41, 2008.

[219] Pampaloni F, Reynaud EG, Stelzer EHK. The third dimension bridges the gap between cell culture and live tissue. *Nature Reviews Molecular Cell Biology*, 8:839–845, 2007.

[220] Zschenker O, Streichert T, Hehlgans S, Cordes N. Genome-wide gene expression analysis in cancer cells reveals 3D growth to affect ECM and processes associated with cell adhesion but not DNA repair. *PLoS ONE*, 7:e34279, 2012.

[221] Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*, 21:1160–1167, 2011.

[222] Germain RN. Open questions: A rose is a rose is a rose - or not? *BMC Biology*, 12:2, 2014.

[223] Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res*, 24:496–510, 2014.

[224] Heisenberg W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik (in German)*, 43:172–198, 1927.

[225] Markert EK, Mizuno H, Vazquez A, Levine AJ. Molecular classification of prostate cancer using curated expression signatures. *Proc Nat Acad Sci*, 108:21276–21281, 2011.

[226] Wagener R, Lenz M, Schuldt B, Schuppert A, Siebert R, Müller FJ. Genome-wide analysis of pluripotency-associated signatures in lymphomas. submitted.

[227] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. `http://www.R-project.org/`.

[228] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[229] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4:1184–1191, 2009.

[230] Bellera CA, Julien M, Hanley JA. Normal approximations to the distributions of the Wilcoxon Statistics: Accurate to what N? Graphical insights. *Journal of Statistics Education*, 18:1–17, 2010.

[231] Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2. `http://CRAN.R-project.org/package=lars`.

A. Supplemental Materials and Methods

Datasets and software

Software

Almost all analyses were conducted using the statistical software R [227] version 3.0.2 with the Bioconductor software for computational biology and bioinformatics [228] version 2.22.0.
Data preprocessing was performed using the apt-probeset-summarize program of Affymetrix Power tools (`http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx`) with the robust multi array (RMA) [165] normalisation method. Mappings of genes and probe identifiers between different platforms were performed using the biomaRt package [229] version 2.18.0 in R.

Datasets

Reference datasets for the two-scale map

Two different versions of the two-scale map were derived that are based on two different datasets, namely the Lukk dataset [12], and a self assembled dataset from the Affymetrix Human U133 Plus 2.0 microarray platform. The Lukk [12] dataset was downloaded as preprocessed dataset from the Array Express database (`www.ebi.ac.uk/arrayexpress`, accession number E-MTAB-62). The annotation of tissues, cell lines and disease states are also available from the Array Express database. Table C.1 provides information about the sample sizes for each of the 369 different groups.
The own dataset was created based on data from the Gene expression omnibus database (`http://www.ncbi.nlm.nih.gov/geo/`), including the studies listed in table C.3. The raw data from these studies were downloaded and jointly preprocessed with the RMA normalisation method. Annotation of different tissues and cell lines was done manually based on the information provided in the gene expression omnibus database. An overview of the sample numbers per tissue or cell type is provided in table C.2.
Several analysed datasets were generated on the Affymetrix Human Gene 1.0 st-v1 microarray platform. Therefore, we also created a reference dataset from this platform, which was used for creation of the background in plots of the PCA space. This dataset consists of the samples from the following GEO series and Array Express dataset that were jointly preprocessed with RMA: GSE12592, GSE12865, GSE13949, GSE14581, GSE14733, GSE14908, GSE15163, GSE15791, GSE17504, GSE17757, GSE18072, GSE18171, GSE18296, GSE18376,

GSE18377, GSE18592, GSE18662, GSE18794, GSE18816, GSE18893, GSE19357, GSE19539, GSE19719, GSE19736, GSE19846, GSE19976, GSE20546, GSE20549, GSE20581, GSE20631, GSE20671, GSE20679, GSE20963, GSE21037, GSE21129, GSE21244, GSE21262, GSE21296, GSE21315, GSE21348, GSE21655, GSE21800, GSE22895, GSE23340, GSE23413, GSE23884, GSE24434, GSE24533, GSE24621, GSE24997, GSE25557, GSE25673, GSE26250, GSE26336, GSE26549, GSE26887, GSE26946, GSE27362, GSE27447, GSE27667, GSE28191, GSE28498, GSE28542, GSE29397, GSE29880, GSE30004, GSE30038, GSE30448, GSE30650, GSE30915, GSE32527, GSE35108, E-MEXP-2654.

Datasets analysed using the two-scale map

The first dataset that was analysed using the two-scale map (GSE18674) is already contained in the own dataset from the Affymetrix Human U133 Plus 2.0 platform and was extracted from this dataset for analysis.

The liver and breast cancer mixture data were downloaded as preprocessed data from Gene expression omnibus (GSE33116). For the analysis, we used the first mixture data, i.e samples GSM820180 to GSM820192.

The heart and brain mixture data were downloaded as raw data from the Affymetrix website [199] and preprocessed using RMA.

The MSC data were downloaded as raw data from GEO. They consist of the six untreated samples from GSE46019, i.e. GSM909604 to GSM909609, and all samples from GSE38806 and GSE54766. All data were downloaded as raw data and jointly preprocessed with RMA.

The cardiomyocyte and heart data were already part of the reference dataset from the Affymetrix Human Gene 1.0 st-v1 platform and were therefore directly extracted from this reference dataset. For the analysis, we used all data from GSE28191, all except of the undifferentiated data from E-MEXP-2654, and the *in vitro* differentiated cardiomyocytes from GSE35108.

The brain time course was downloaded from GEO (GSE25219) as preprocessed data and the probes from this Affymetrix Exon array were matched to the Affymetrix Human Gene 1.0 st-v1 probes via the biomaRt R package. The 15 samples from the cerebral cortex that were analysed were chosen randomly as follows: GSM708007, GSM708196, GSM708279, GSM708298, GSM708432, GSM708450, GSM708482, GSM708640, GSM708710, GSM708786, GSM708896, GSM708926, GSM708955, GSM708987, GSM709315. The 15 cerebellum samples have accession numbers GSM708006, GSM708137, GSM708201, GSM708297, GSM708431, GSM708448, GSM708480, GSM708639, GSM708709, GSM708785, GSM708894, GSM708948, GSM708953, GSM708985, and GSM709313.

The dataset of *in vitro* differentiated neurons consists of all samples from GSE51214 and GSE43235 that were downloaded as raw data and jointly preprocessed. The astrocyte and neural progenitor cells (all data from GSE43794, GSE36145, and GSE43382) were preprocessed in the same way.

All data were quantile normalised with the reference dataset of the respective platform before being mapped to the two-scale landscape.

Methodological details

PluriTest analysis

The PluriTest bioinformatics assay was originally developed on the Illumina Human HT12 v3 microarray platform. For a transformation to the Affymetrix Human U133 Plus 2.0 platform, probes were mapped via the biomaRt package. Pluripotency and novelty scores were calculated based on the NMF-components described in [70]. Due to the platform transformation, the calculation of the novelty score, is a bit different than in [70]. The novelty score is basically a distance measure from the NMF space spanned by the pluripotent stem cells in the training dataset. This distance is calculated by the eight-norm in [70]. Here, we use the squared two-norm instead, resulting in a less strong focus on single outlier genes.

Relationship between linear and non-linear mapping

We presented and compared two different mapping methods in the present thesis. The linear mapping method is based on an ordinary linear projection to each coordinate of the reduced dimensional spaces. The non-linear mapping is only based on a subset of the genes, namely the 1% of most strongly negatively and positively expressed genes in the respective signature. The mapping is then accomplished by a comparison of these two sets of genes via the Wilcoxon rank sum test on the new data.
This non-linear mapping procedure is therefore only based on the ranking of genes, neglecting any information about the effect size, i.e. the length of the vector that is mapped to the space. This means, that two vectors that have the same direction, but differ in their lengths, will have exactly the same non-linear mapping result. Therefore, the non-linear mapping can be considered as comparing only directions, whereas the linear mapping also considers the length of the vectors. This is one difference between the two mappings that can be easily removed by a normalisation of the vectors to length one, resulting also in a comparison of directions for the linear mapping.
Beside this difference between the two mappings, we describe an almost quadratic relationship between the two mappings for some data. This can be explained in the following way. For sufficiently large and balanced sample sizes, the Wilcoxon rank sum statistic can be very good approximated by a normal distribution [230]. Furthermore, the value of a normal distributed random variable has an almost quadratic relationship to the corresponding logarithmic p-value (Fig. A.1). Therefore, there is an almost quadratic relationship between the Wilcoxon rank sum statistic and the corresponding logarithmic p-value.
However, the Wilcoxon rank sum statistic is limited by an upper value that is dependent on the sample size [230]. In the present example this limit is at a log10 p-value of ap-

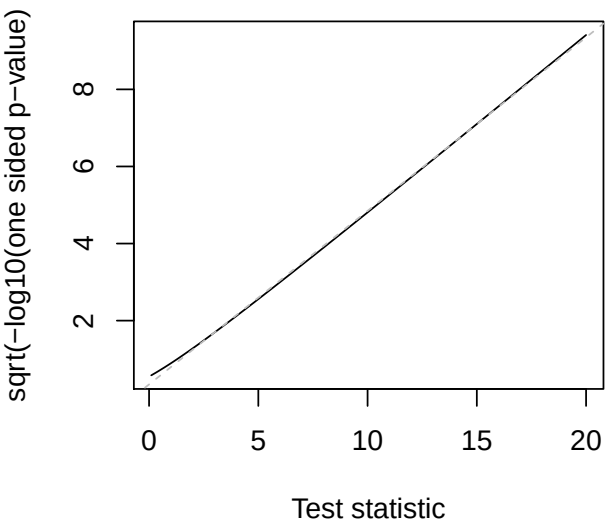


Figure A.1.: The value of a standard normal distributed test statistic versus the square root of the logarithmised p-value shows an almost linear relationship. Therefore, the log p-value is quadratically dependent on the test statistic.

proximately 37. Therefore, the described quadratic relationship is only valid in a certain range, which is limited by this upper bound. These two effects together can explain some of the differences between the linear and non-linear mappings. However, when there are additional non-linear effects in the data, as described for the mixture data, the relationship between both mapping procedures is more complex.

According to the above described quadratic relationship, we have fitted a signed quadratic curve to the data in Fig. 4.7 and Fig. B.5. This fit was accomplished through a fit of a model of the square root of the non-linear mapping scores depending linearly on the linear mapping scores. Due to the described saturation effect, the fit is only performed on those signatures with a linear projection value of at most 50% of the maximal absolute linear projection value. This avoids the disturbing influence of saturated signatures on the quadratic fit.

Residual analysis

For the example of mixed brain cancer and liver tissues, we analyse the residuals after the linear mapping to both reduced dimensional expression spaces, i.e. to the two-scale map. Such a determination of residuals is straight forward for the PCA space due to the orthogonality of the coordinates. However, in the residual space there is still some correlation present, e.g. between the liver and cirrhosis signatures, or between heart and heart disease. Therefore, the true dimensionality of the residual space is lower than 369. In fact, the tissue specific signatures in the residual space are not linearly independent, i.e.

the rank of the matrix of signatures is only 368 and not 369 (according to the rankMatrix function in R). However, this rank is most probably only that high due to the noise in the data that artificially increases the rank of the matrix. This has implications for the mapping of the data to the residual space.
The mapping presented in the thesis is a linear mapping to each signature of the residual space. However, this cannot be used to calculate the residuals, due to the correlations mentioned above. To do this, it is necessary to project the data onto the whole space at once, instead of projecting it onto each coordinate individually. One possibility to do this is a least squares fitting. However, as indicated above, the true intrinsic dimensionality of the residual space might be lower than the number of signatures in the residual space. Therefore, some dimensions may correspond to pure noise. Thus, in order to avoid a fitting of noise, it is necessary to include some regularisation in the regression model.
In this thesis this is done via the lars regression [231], which constraints the sum of the absolute regression coefficients. A successive increase of the constraining parameter results in a stepwise increase of the model size, representing an alternative to the classical forward selection of regression models [231].
A final choice that has to be made in this procedure is the step at which we stop with increasing the model, i.e. the choice of the model size. This is usually done based on a likelihood ratio test, the Akaike information criterion, or similar concepts. For the lars regression, an other approach, building a trade-off between the squared error and the degrees of freedom in the model is proposed [231]. However, these concepts usually require an independence of the repeated measurements in the data. In our case, the repeated measurements correspond to the different genes, which are certainly not independent (due to correlated noise terms). Therefore, the classical approaches result in very large models. Due to the fact that this residual analysis is only a side topic in the present thesis, we use a rather practical approach to solve this problem, without going into more statistical details of the best model choice. Thus, we reason that it is unlikely that the analysed data (especially for the known mixture of two tissues) incorporate more than four different tissue specific signatures and restrict the model size to four components. Other choices, i.e. slightly larger or smaller models, do not change the statements made in the thesis about the residual analysis (data not shown).

Visualisation

All plots were generated with the statistics software R. The background colours in the PluriTest and PCA space representations were basically done by a kernel density estimation, where the transparency in the colours corresponds to the density of the respective class at this location. Different classes are represented by different colours. The graphics of the residual space contain always only a subset of the 369 or 191 signatures. This subset includes always the lowest as well as the five highest scores for each sample. Due to this choice that depends on the analysed samples, the chosen signatures are different from

graphic to graphic. Sometimes, some additional fixed signatures were also included.

Implementation of the Concordia tool

The website of the Concordia tool was not accessible at the time we tried to use it. Therefore, we reimplemented the tool using the own dataset with the 191 different groups as phenotypic concepts. For the implementation, we strictly followed the methodological description in [11]. The evaluation of the performance, that even outperforms that of the original paper, indicates that the implementation was performed correctly.
The evaluation of the tool was performed with a leave-one-out cross validation scheme. The same scheme was also used for the comparative evaluation of the two-scale map. Only the PCA space (which was not used for calculation of performance measurements) was not recalculated at every step of the cross-validation scheme, meaning that it was determined based on all data. However, due to the large number of samples, the effect on the results should be negligible.

B. Supplemental figures

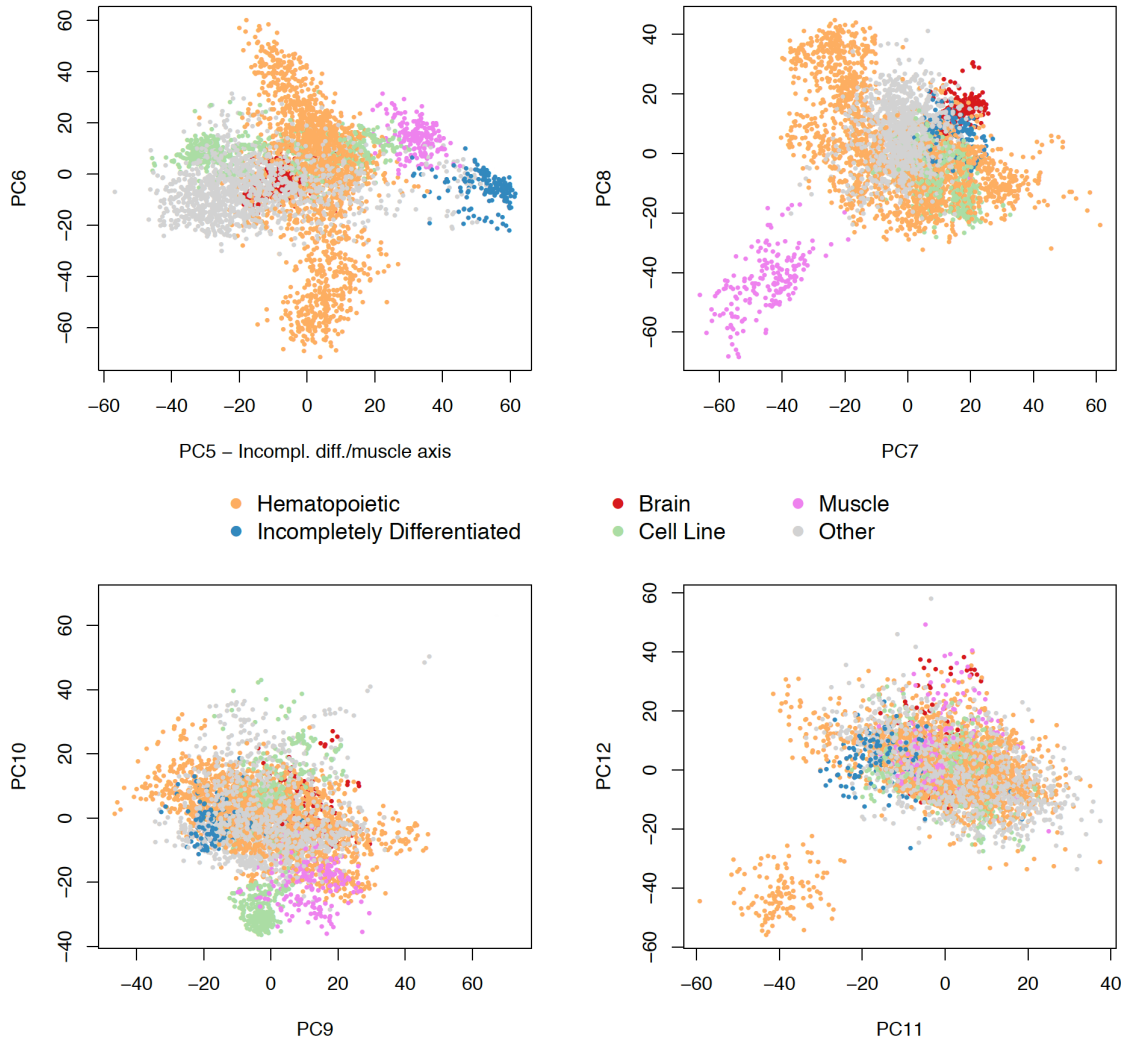


Figure B.1.: Principal components 5 to 12 of the Lukk [12] dataset. PC 5 separates incompletely differentiated and, to a less extent, muscle tissues from the rest. PCs 7 and 8 separate muscle tissues from all others. The separated cluster in PCs 11 and 12 corresponds mainly to CD138 positive plasma cell myelomas. PCs 6, 9, and 10 have unclear or no biological associations.

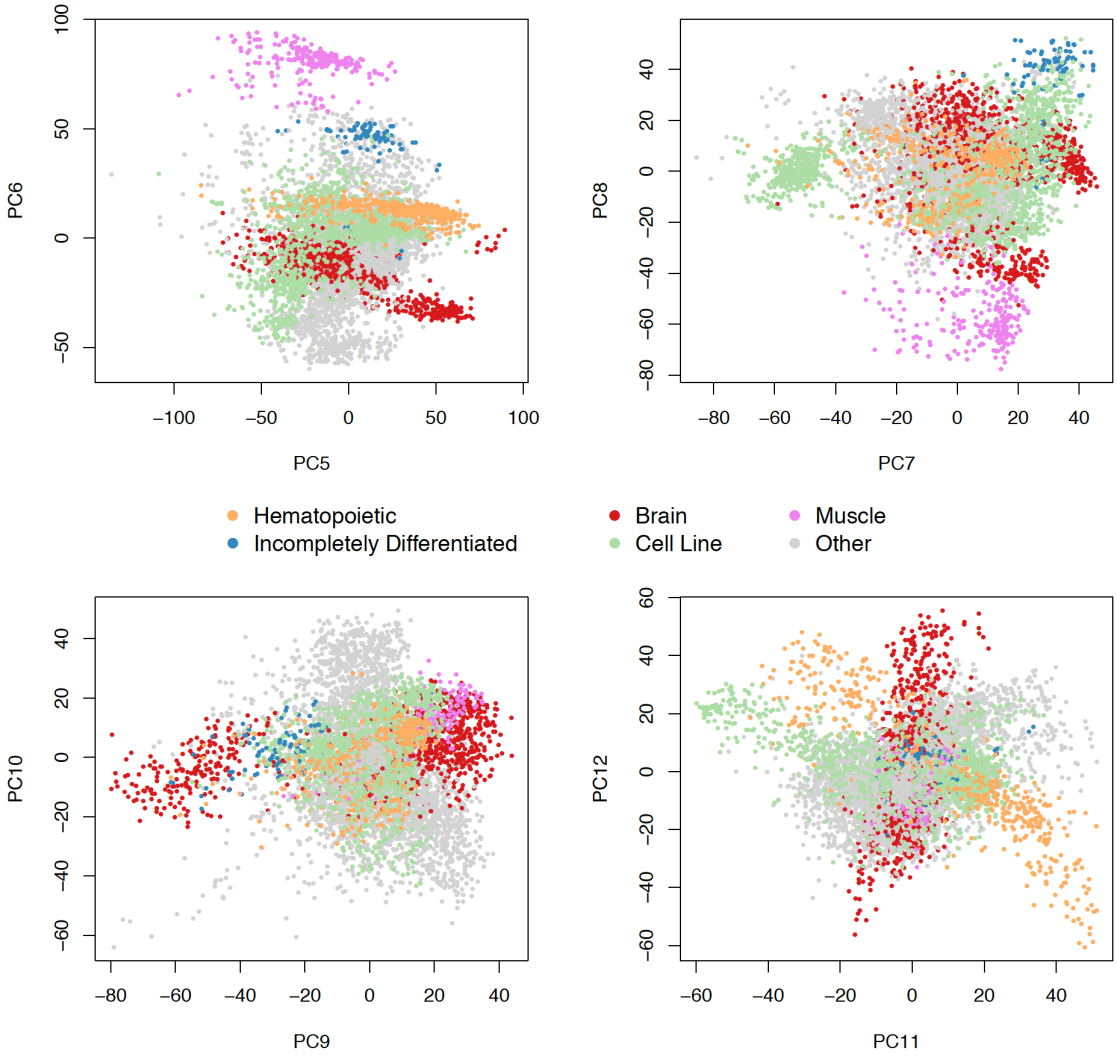


Figure B.2.: Principal components 5 to 12 of the own dataset. PC 6, and to some extent PC8, are specific for muscle tissues. PC7 separates pluripotent stem cells (green cluster on the left side) from the rest. All other depicted principal components do not show any clear specificity for a certain tissue or cell type.

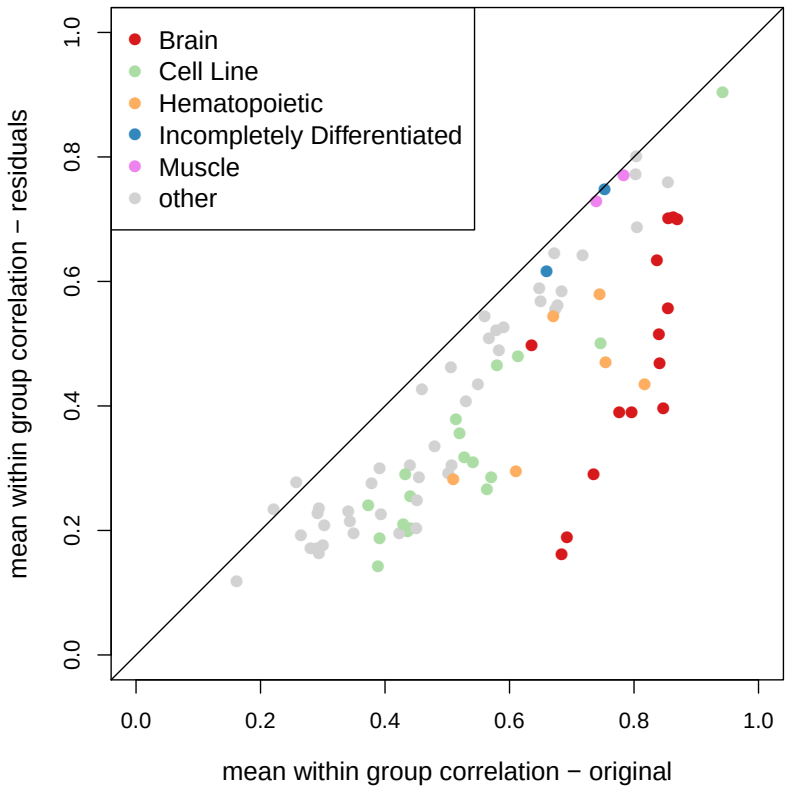


Figure B.3.: Within group correlations in the own dataset before and after PCA based decomposition. The within-group correlations in the residual space are all clearly positive and for several tissues they are similarly high as in the original data. This indicates that there is a significant amount of tissue specific information in the residual space.

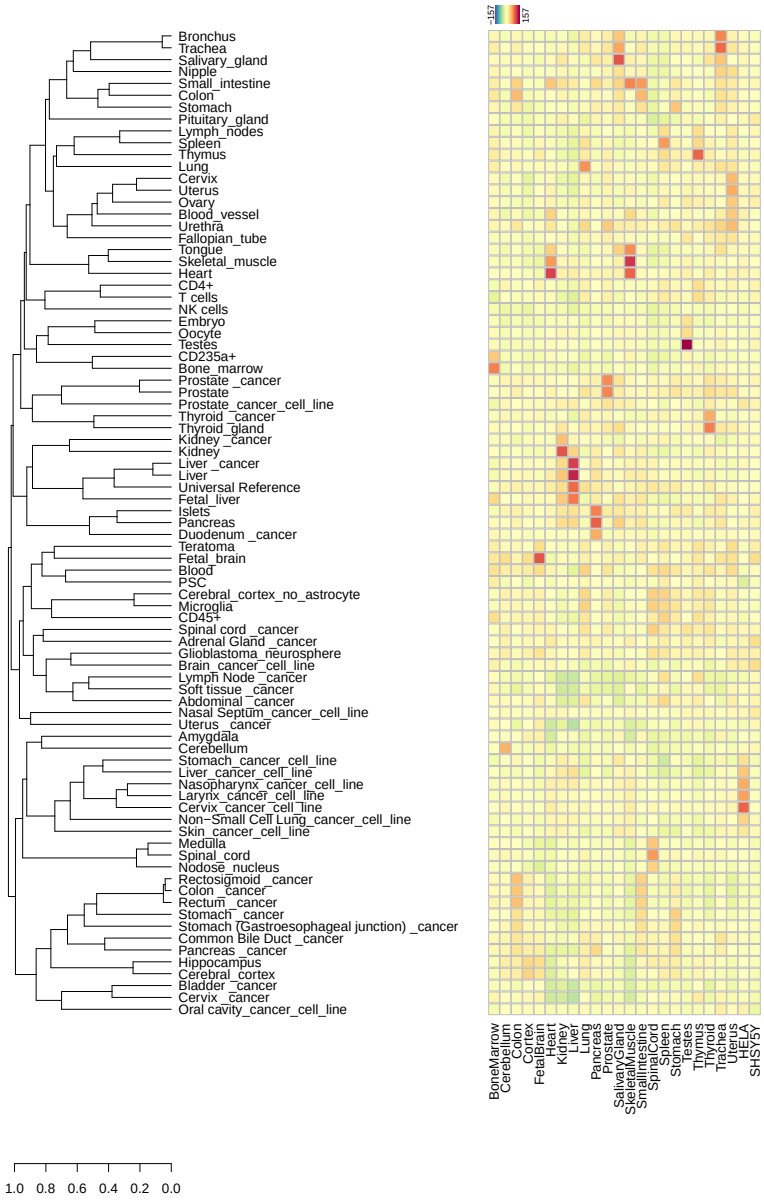


Figure B.4.: Mapping of 24 tissues or cell lines from GSE18674 to the residual space of the own dataset. All 22 tissues are correctly identified, i.e. the signature with the highest association matches with the true tissue type. It has to be noted that the data are part of the own dataset itself.

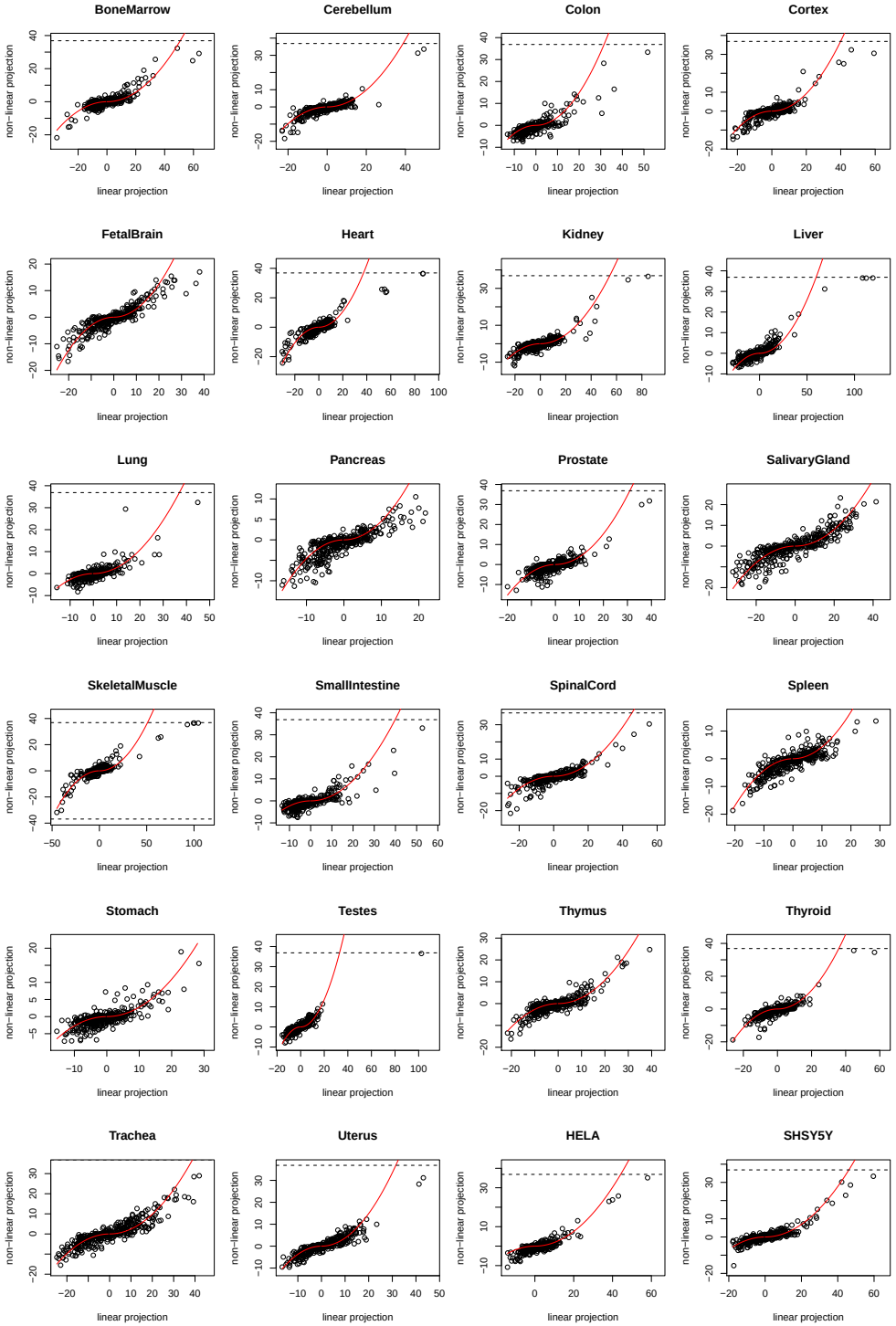


Figure B.5.: Comparison of the linear and non-linear mapping methods for all 24 sample from dataset GSE18674. The non-linear mapping can in a certain range be approximated by a sign-preserving quadratic transformation of the linear mapping. However, at higher values a saturation effect can be observed due to the p-value limitations of the Wilcoxon rank sum test for fixed sample sizes (here number of probes), leading to a maximal (minimal) value of the 37 (-37) for the non-linear mapping (dotted lines).

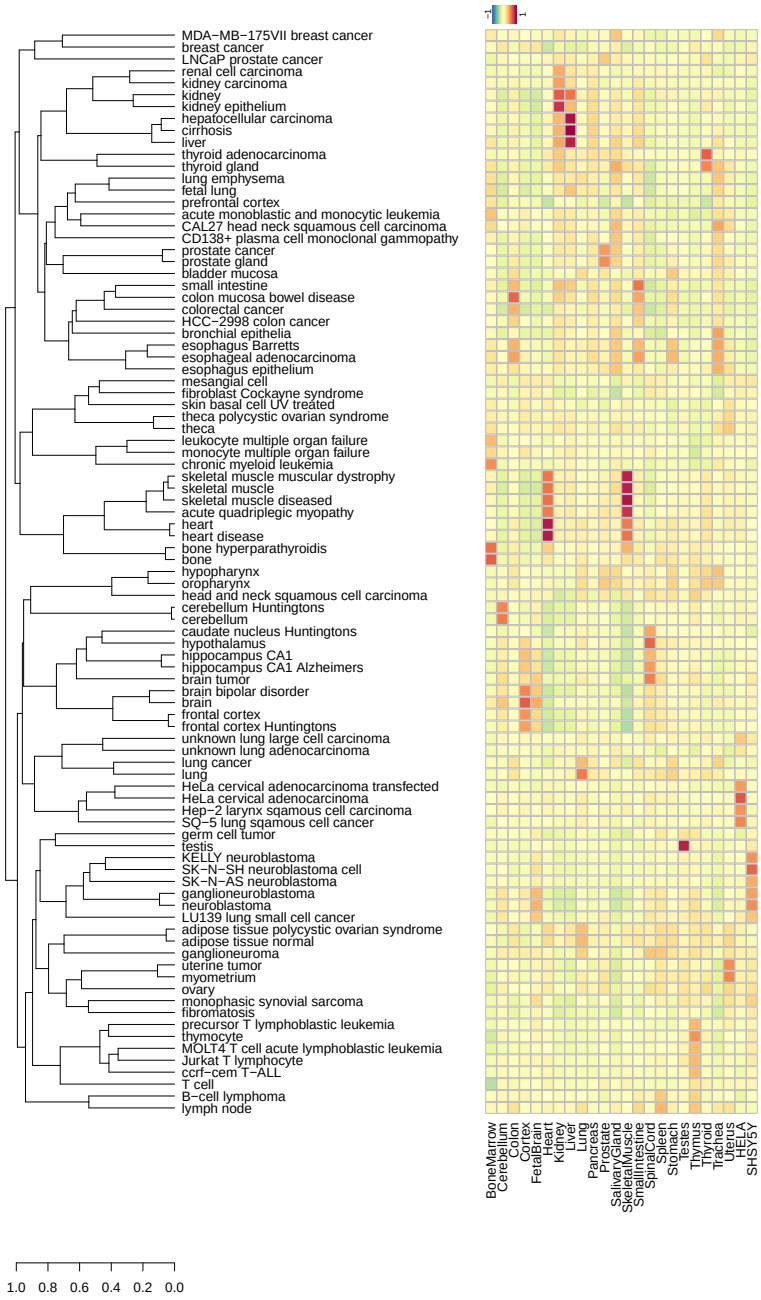


Figure B.6.: Linear mapping focusing on the directions of the 24 tissues/cell lines from GSE18674 in the residual space. The residual vectors of the 24 samples were normalised to length one in order to focus on the similarity of directions, neglecting the length of the vectors. This introduces an upper bound of 1 to the depicted values, leading to less differences between the different columns.

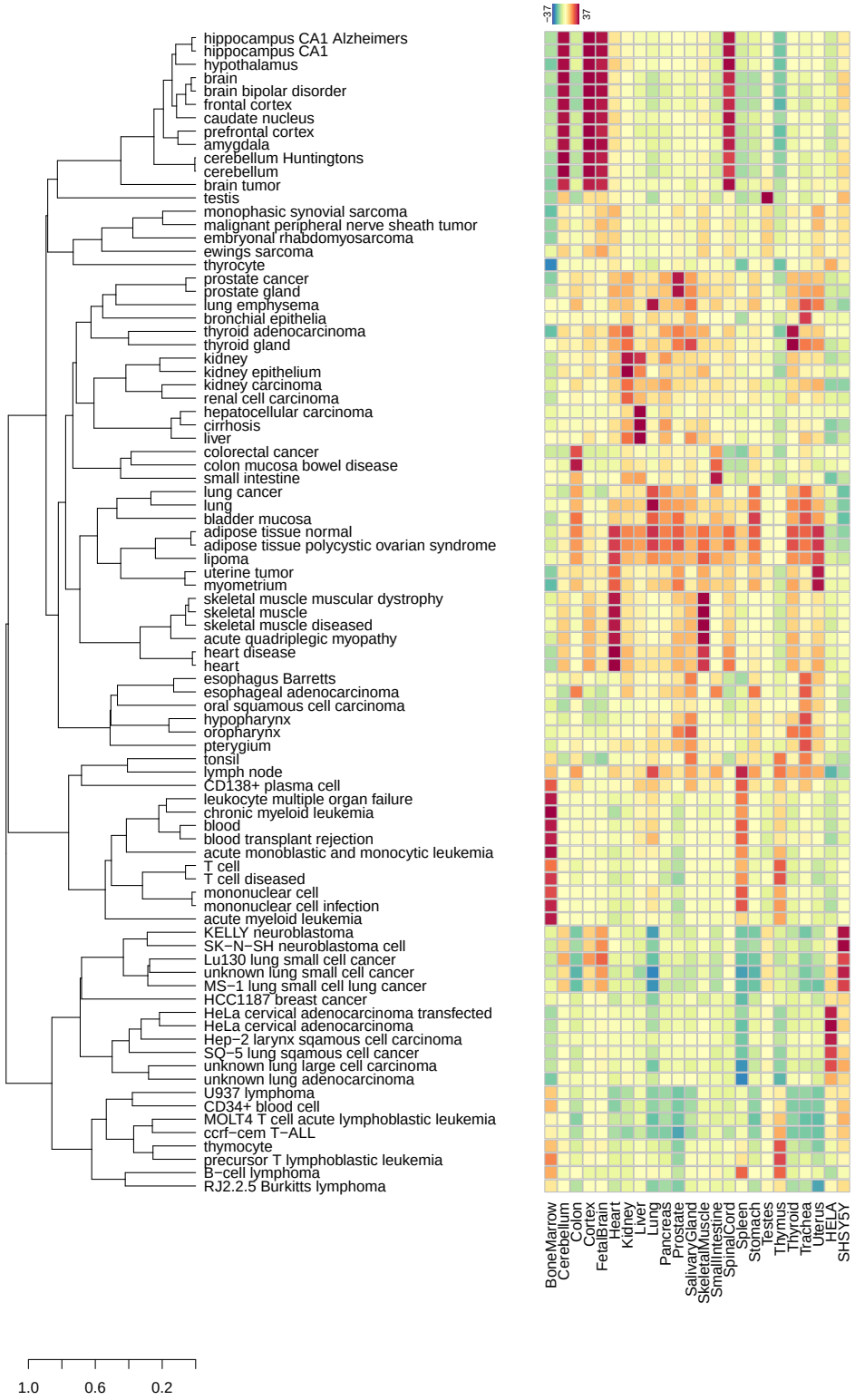


Figure B.7.: Non-linear mapping of the 24 tissues or cell lines from GSE18674 (columns) to the tissue specific space without PCA based decomposition (rows). The higher correlations between different signatures compared to the case with PCA based decomposition (Fig. 4.4) can be especially well seen for those signatures corresponding to specific brain regions.

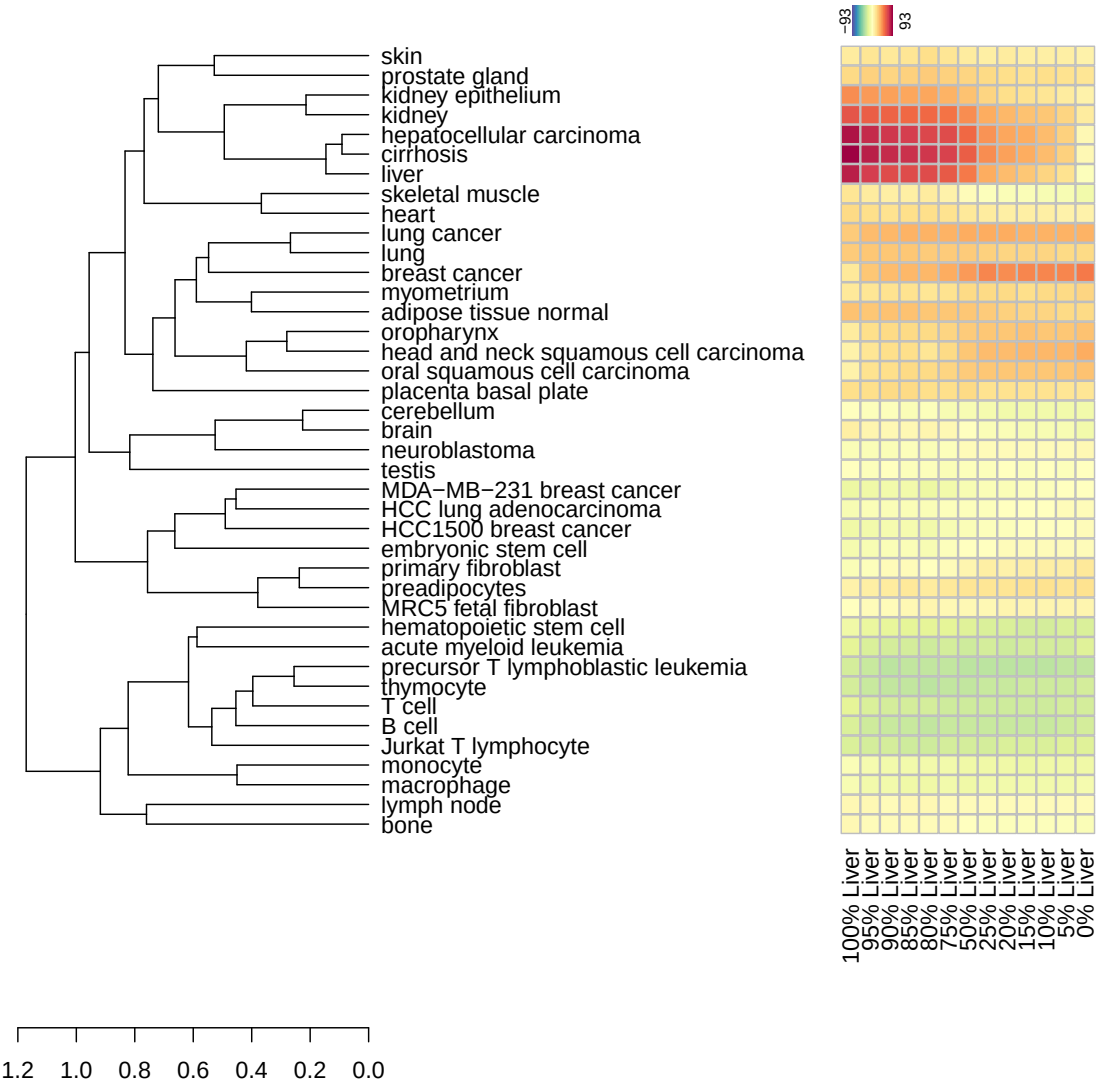


Figure B.8.: Linear mapping of the liver and breast cancer mixtures to the residual space without PCA based decomposition. The breast cancer signature shows higher values than for the two-scale approach, but other cancer signatures, e.g. lung cancer show also relatively high associations, which are better suppressed in the two-scale approach.

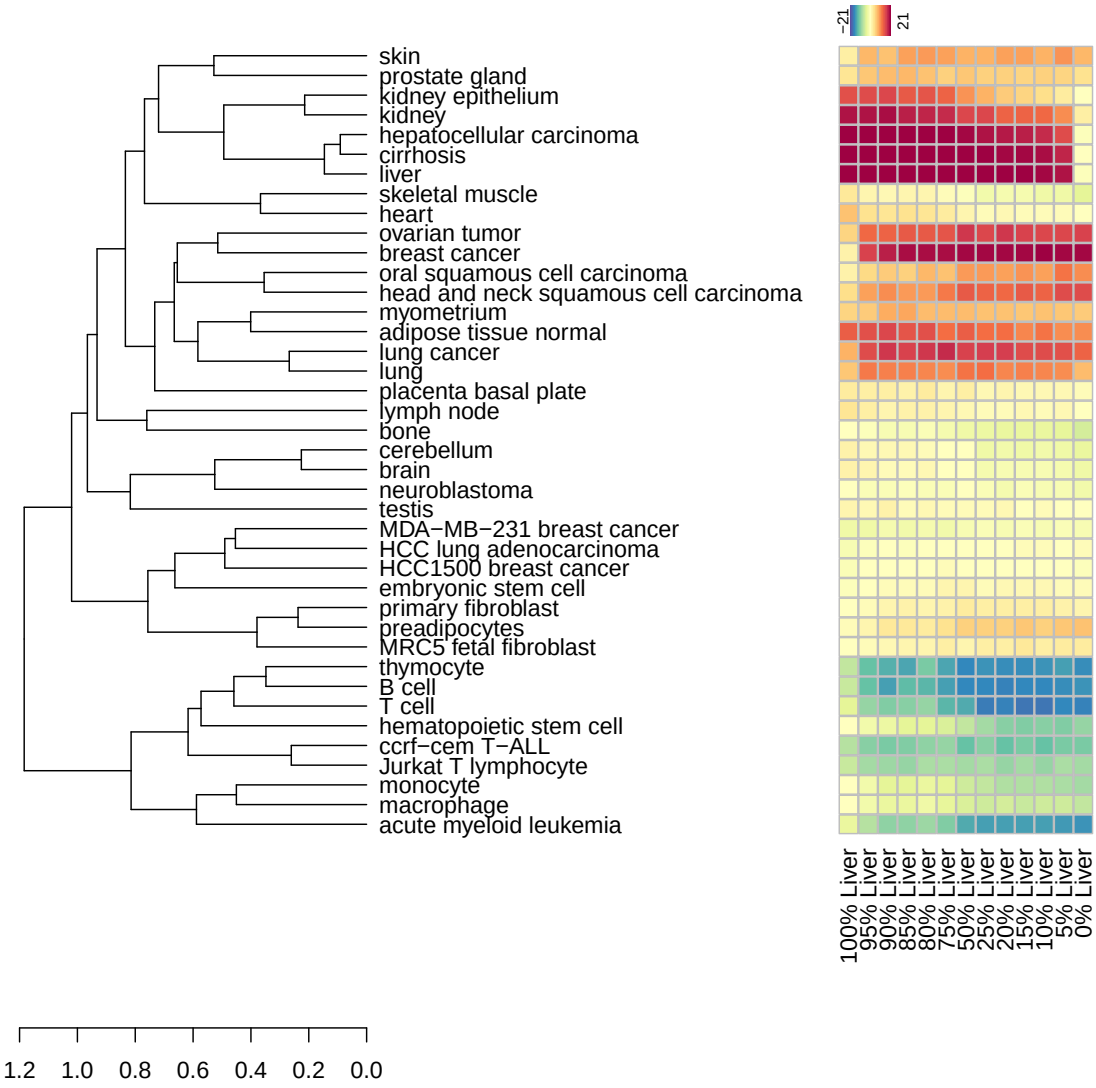


Figure B.9.: Non-linear mapping of the liver and breast cancer mixtures to the residual space without PCA based decomposition. The results are clearly less specific than for the two-scale approach.

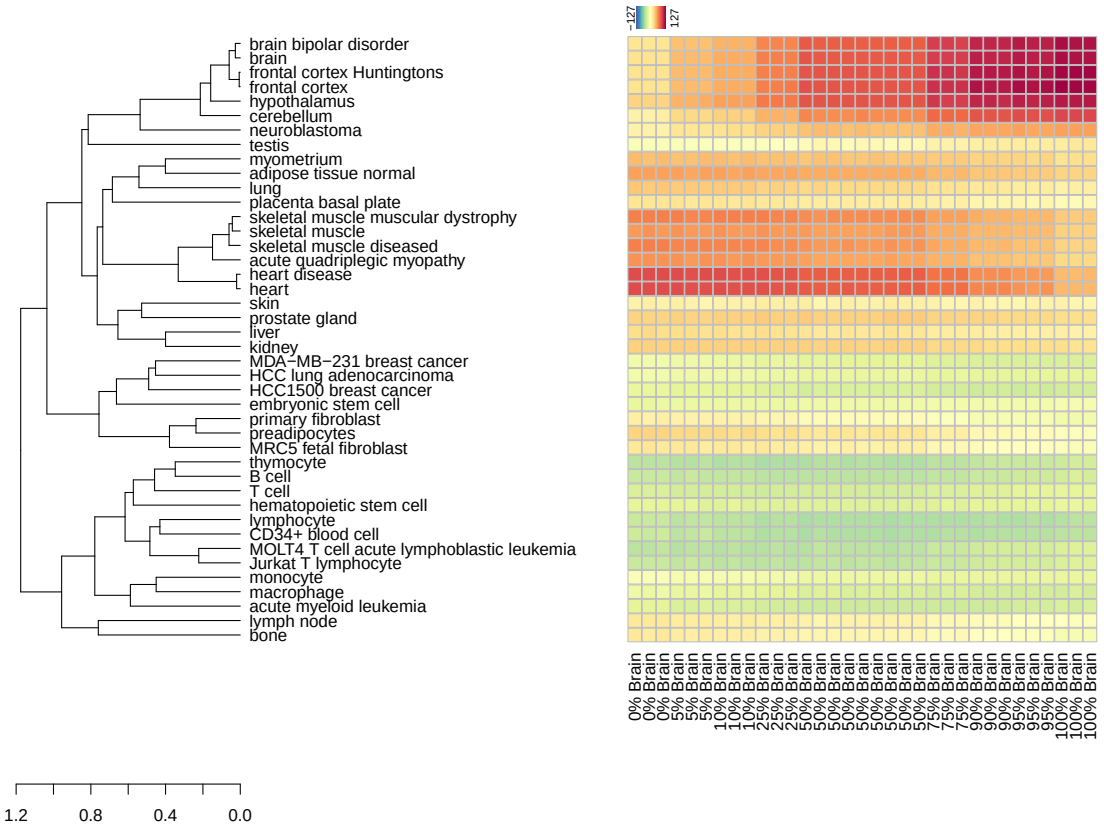


Figure B.10.: Linear mapping of brain and heart mixtures to the tissue specific space without PCA based decomposition. The distinction of different brain regions is significantly less clear as for the two-scale approach (Fig. 4.15).

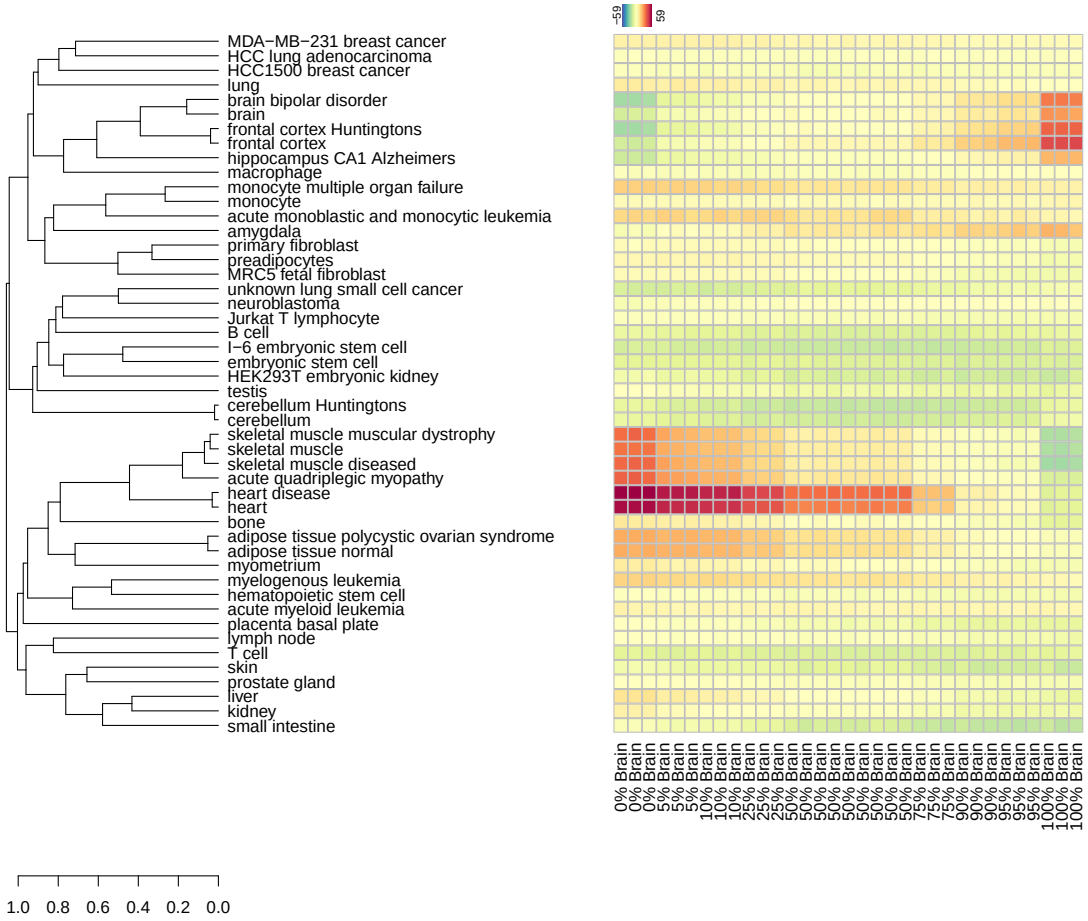


Figure B.11.: Non-linear mapping of brain and heart mixtures to the residual space. The heart score decreases continuously with increasing fraction of brain tissue, but the brain specific scores show a relatively sharp increase at very low and very high brain fractions with a relatively slow increase in-between.

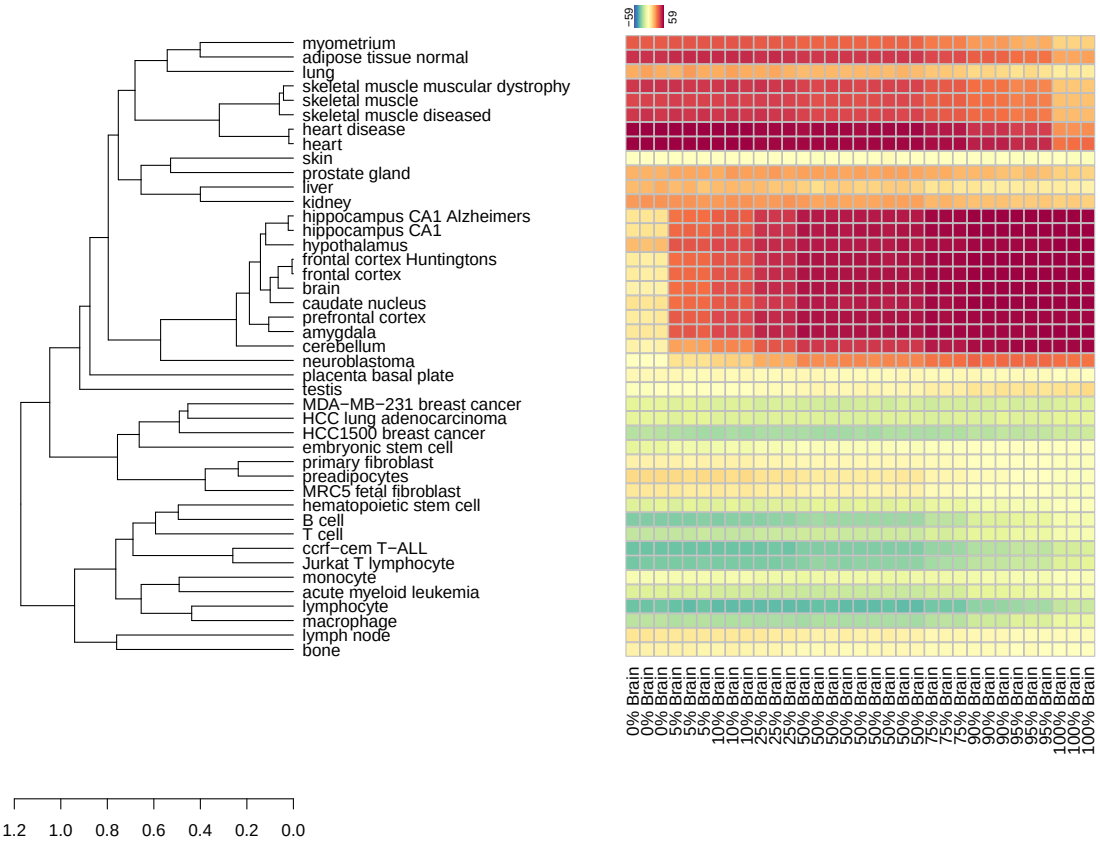


Figure B.12.: Non-linear mapping of brain and heart mixtures to the tissue specific space without PCA based decomposition. The results do not show a clear separation of different brain regions and also no smooth transition for samples with different mixture fractions.

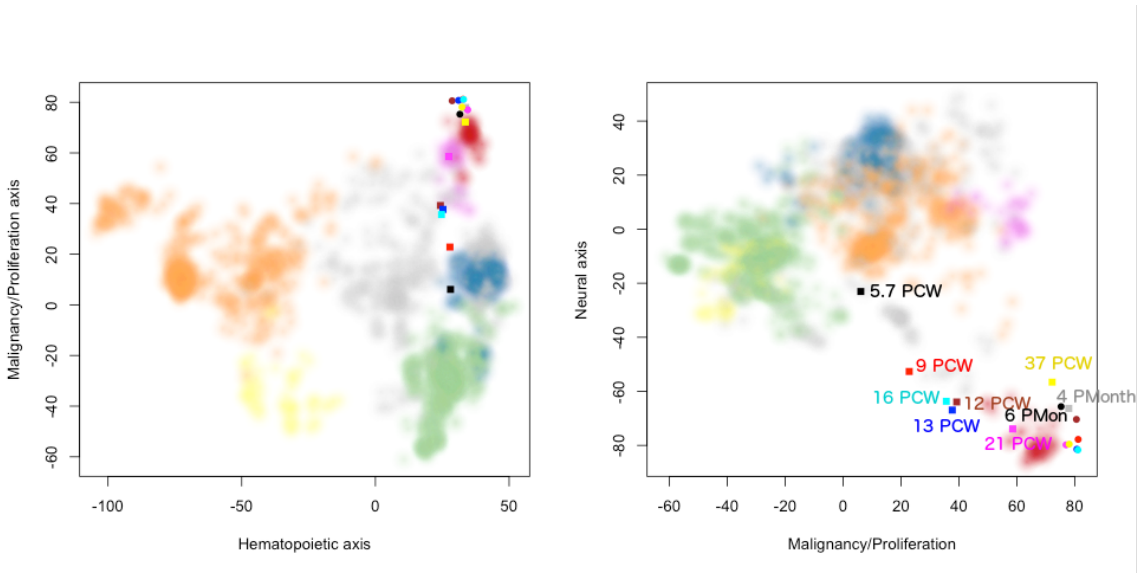


Figure B.13.: Samples from human cortical development mapped onto the PCA space from the Lukk [12] dataset. During brain development, the expression patterns get continuously closer to those of adult brain tissue (red background). Note that the data were generated on the Affymetrix Human Exon 1.0 st-v1 array, while the reference dataset used for the background colours was hybridised to the Human Gene 1.0 st-v1 array. This may explain the slight differences between the adult brain tissues and the red background on PC2.

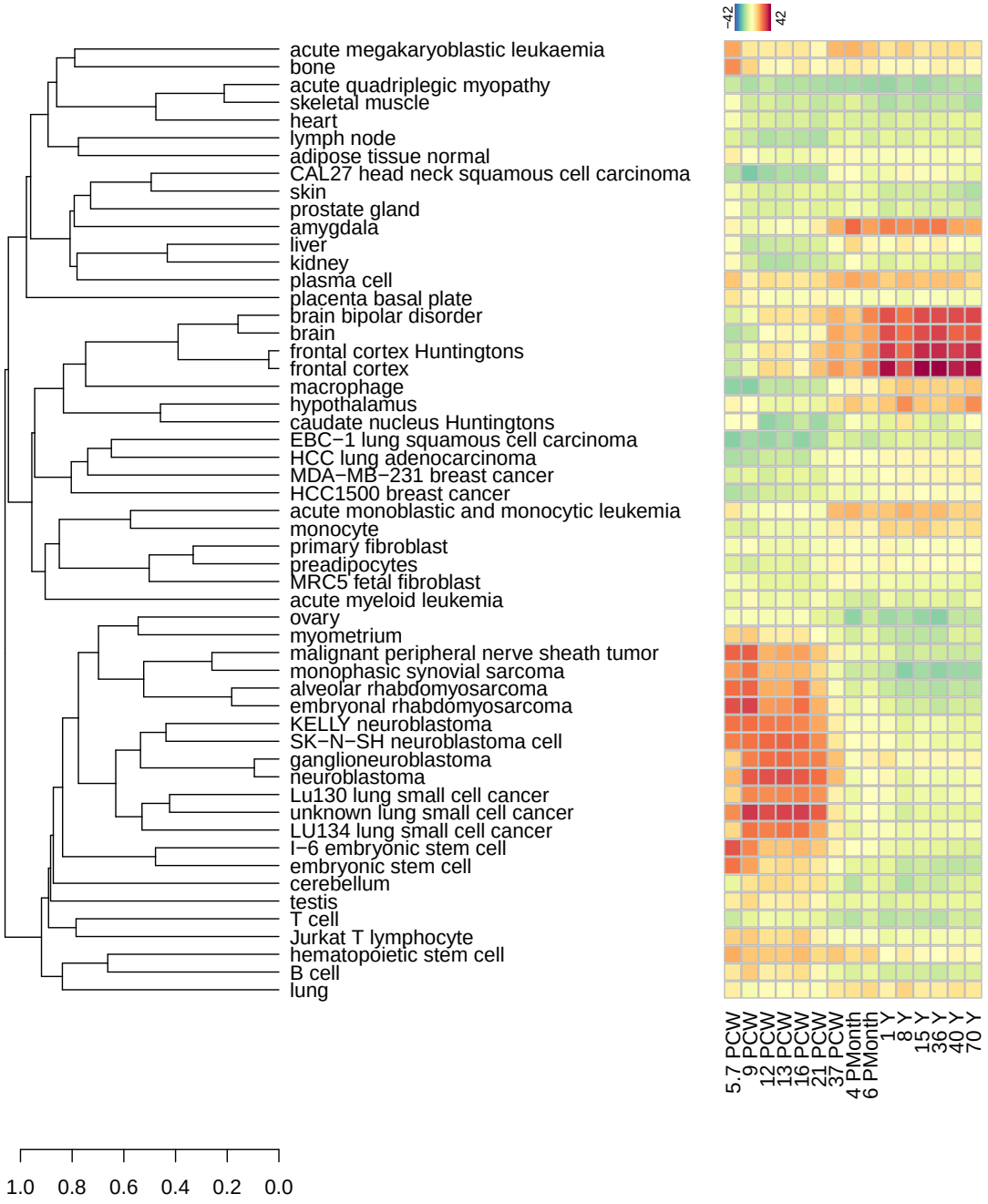


Figure B.14.: Mapping of 15 cortical brain tissue samples from post-conceptional week 5.7 to adulthood (70 years) onto the residual tissue specific space from the Lukk [12] dataset. Unfortunately, the Lukk dataset does not contain samples from fetal brain or neural progenitor cells. Therefore, the highest associations of developing brain samples is with cancer cell lines. This is the reason, why we use the own dataset for the analysis of the *in vitro* transformed neural cells.

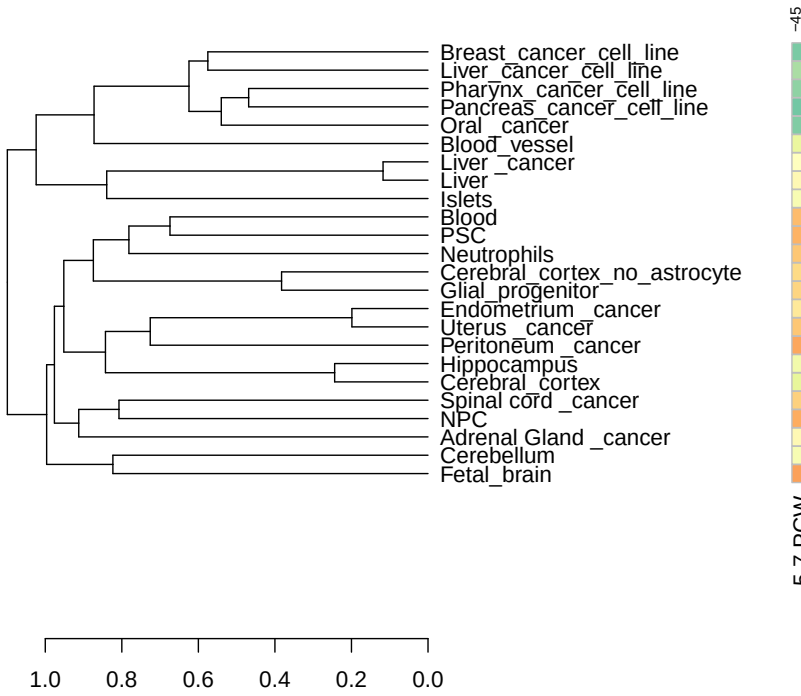


Figure B.15.: 15 samples from the embryonic to adult cerebellum mapped to the residual tissue specific space of the own dataset. The analysis shows that there is a clear association with the cerebellum signature from 4 months to 36 years after birth, indicating that different brain regions can be nicely distinguished. Surprisingly, this association diminishes in older ages.

C. Supplemental tables

Table C.1.: Number of samples per group for all 369 groups in the Lukk [12] dataset.

Tissue/cell type	Number of samples
293t-tva	1
600MPE breast cancer	1
A549 lung adenocarcinoma	31
A673 Ewing tumor	3
ABC-1 lung adenocarcinoma	1
AMO-1 myeloma	4
AU565 breast carcinoma	1
B cell	11
B-cell lymphoma	213
B-cell lymphoma cell line	4
BT20 breast cancer	1
BT474 breast cancer	10
BT483 breast cancer	1
BT549 breast cancer	1
BeWo choriocarcinoma	4
C2C12 mouse muscle myoblast	3
CAL27 head neck squamous cell carcinoma	2
CAMA1 breast cancer	1
CCL-153 fetal pulmonary fibroblast	3
CD138+ plasma cell	8
CD138+ plasma cell monoclonal gammopathy	8
CD138+ plasma cell myeloma	120
CD33+ cells	1
CD34+ blood cell	8
CD34+ blood cell thymus	1
CMK megakaryoblastic	1
Calu-1 lung epidermoid carcinoma	1
Calu-3 lung adenocarcinoma	16
Calu-6 lung anaplastic carcinoma	1
DU 145 prostate carcinoma	1
EBC-1 lung squamous cell carcinoma	1

EKVX lung non small cell carcinoma	1
EW24 Ewing sarcoma cell line	4
FM9514 human embryonic myoblast	30
FR4 multiple myeloma	1
H1 mesenchymal precursor	1
H9 mesenchymal precursor	1
HBL 100 breast cancer	1
HCC lung adenocarcinoma	1
HCC-2998 colon cancer	1
HCC1007 breast cancer	2
HCC1171 lung non small cell lung carcinoma	1
HCC1187 breast cancer	1
HCC1195 lung adenocarcinoma	1
HCC1359 lung large cell carcinoma	1
HCC1428 breast cancer	1
HCC15 lung squamous cell carcinoma	1
HCC1500 breast cancer	1
HCC1937 breast cancer	1
HCC2157 breast cancer	1
HCC2185 breast cancer	1
HCC227 lung adenocarcinoma	1
HCC2935 lung non-small cell lung cancer line	1
HCC3153 breast cancer	1
HCC366 non-small cell lung carcinoma	1
HCC38 breast cancer	1
HCC44 non-small cell lung carcinoma	1
HCC515 lung adenocarcinoma	1
HCC70 breast cancer	1
HCC78 lung non-small cell carcinoma	1
HCC827 lung adenocarcinoma	1
HCC95 lung non-small cell carcinoma	1
HCT-15 colorectal adenocarcinoma	1
HCT116 colorectal carcinoma	6
HEK293 embryonic kidney	18
HEK293T embryonic kidney	6
HMEC S1	1
HMEC184 breast cancer	5
HMT3522S1 breast cancer	1
HMT3522S1 breast epithelium	4
HOP 92 non-small cell lung carcinoma	1
HOP-62 lung adenocarcinoma	1

HS578T breast cancer cell line	1
HeLa cervical adenocarcinoma	32
HeLa cervical adenocarcinoma transfected	10
Hep-2 larynx sqamous cell carcinoma	10
I-6 embryonic stem cell	1
IB3-1 adenovirus transformed bronchial epithelia	11
IGROV1 ovarian cancer	1
JEG3 placental choriocarcinoma	3
Jurkat T lymphocyte	2
KELLY neuroblastoma	1
KM12 colon cancer	1
KM4 multiple myeloma	8
KS Y-1 Kaposis sarcoma AIDS associated	1
KS-IMM Kaposis sarcoma non AIDS associated	1
Kaposi sarcoma	14
LC-1F lung squamous cell	1
LC1/SQ lung squamous cell	1
LC2/AD non-small cell lung adenocarcinoma	1
LK-2 lung squamous cell	1
LNCaP prostate cancer	16
LOXIMVI malignant amelanotic melanoma	1
LU134 lung small cell cancer	1
LU139 lung small cell cancer	1
LU165 lung small cell cancer	1
Lu130 lung small cell cancer	1
Lu65 lung non-small cell adenocarcinoma	1
M14 malignant melanoma	1
M70e hematopoietic	2
MALME-3M malignant melanoma	1
MCF10A breast epithelial fibrocystic disease	2
MCF10A-myc breast epithelial fibrocystic disease	8
MCF12A breast epithelial+M824	1
MCF7 breast epithelial adenocarcinoma	213
MDA-MB-134VI breast cancer	1
MDA-MB-175VII breast cancer	1
MDA-MB-231 breast cancer	26
MDA-MB-361 breast cancer	1
MDA-MB-415 breast cancer	1
MDA-MB-435 melanoma	3
MDA-MB-436 breast cancer	1
MDA-MB-453 breast cancer	1

MDA468 breast cancer	10
MOLT4 T cell acute lymphoblastic leukemia	18
MRC5 fetal fibroblast	1
MS-1 lung small cell lung cancer	2
NC-NC lymphoblastoid B cell	3
NCI-ADR-RES ovarian cancer	1
NCI-H226 lung squamous cell carcinoma	1
NCI-H322M lung small cell bronchioalveolar carcinoma	1
NCI-H460 lung large cell carcinoma	1
NCI-H522 lung adenocarcinoma	1
NCI-H929 myeloma	1
NCU-MM1 multiple myeloma	1
OVCAR-3 ovarian cancer	1
OVCAR-4 ovarian cancer	1
OVCAR-5 ovarian cancer	1
OVCAR-8 ovarian cancer	1
PC-1 lung squamous cell cancer	1
PC9 lung non-small cell adenocarcinoma	1
QG-56 lung squamous cell cancer	1
RERF-LC-AI lung sqamous cell cancer	1
RERF-LC-KJ lung non-small cell adenocarcinoma	1
RERF-LC-MS lung non-small cell adenocarcinoma	1
RJ2.2.5 Burkitts lymphoma	8
RKO colon carcinoma	24
RXF-393 renal cell carcinoma	1
SBC-5 lung small cell cancer	1
SBC3 lung small cell cancer	1
SF-268 glioblastoma	1
SF-539 glioma	1
SK-N-AS neuroblastoma	1
SK-N-MC neuroblastoma	4
SK-N-SH neuroblastoma cell	4
SKBR3 breast adenocarcinoma	9
SKMC fetal skeletal muscle myoblast	3
SKMM1 multiple myeloma	1
SN12C renal cell carcinoma	1
SNB-19 glioblastoma	1
SNB-75 astrocytoma	1
SQ-5 lung sqamous cell cancer	1
SR lymphoma	1
SUM1315MO2 breast cancer	1

SUM149PT breast cancer	1
SUM159PT breast cancer	1
SUM185PE breast cancer	1
SUM190PT breast cancer	1
SUM225CWN breast cancer	1
SUM44PE breast cancer	1
SUM52PE breast cancer	1
SW480 colon cancer primary tumor derived	6
Skgt4 esophageal cancer	5
Su-dhl1 anaplastic large cell lymphoma	6
T cell	47
T cell diseased	43
T47D breast ductal carcinoma	11
TERV	7
TK-10	1
U251 glioblastoma	1
U937 lymphoma	1
UACC-257 melanoma	1
UACC-62 melanoma	1
UACC812 breast cancer	1
UO-31 renal cell carcinoma	1
WI38 fetal fibroblast	1
ZR751 breast cancer	1
ZR7530 breast cancer	1
ZR75B breast cancer	1
a498 renal cell carcinoma	1
acute lymphoblastic leukemia	95
acute megakaryoblastic leukaemia	3
acute monoblastic and monocytic leukemia	6
acute myeloid leukemia	295
acute myelomonocytic leukemia	4
acute promyelocytic leukemia	18
acute quadriplegic myopathy	5
adipose tissue normal	8
adipose tissue polycystic ovarian syndrome	9
adipose-derived adult stem cells	6
adipose-derived adult stem cells cultured	3
airway epithelial cell	1
airway epithelial cell cystic fibrosis	5
alveolar rhabdomyosarcoma	9
amygdala	1

bladder	3
bladder cancer	41
bladder mucosa	4
blood	39
blood transplant rejection	3
bone	7
bone hyperparathyroidis	7
brain	39
brain bipolar disorder	31
brain tumor	82
breast cancer	672
bronchial epithelia	33
bronchoalveolar lavage cell	25
bronchoalveolar lavage cell transplant rejection	4
caco2 colon adenocarcinoma	15
caki-1 renal cell carcinoma	1
caudate nucleus	30
caudate nucleus Huntingtons	38
ccrf-cem T-ALL	1
cerebellum	26
cerebellum Huntingtons	39
chondroblastoma	4
chondromyxoid fibroma	3
chondrosarcoma	6
chordoma	4
chronic myeloid leukemia	44
cirrhosis	5
colo205 colorectal adenocarcinoma	1
colon mucosa bowel disease	8
colorectal cancer	34
conjunctiva	4
connective tissue fibroblast	3
coronary artery smooth muscle cell	9
dedifferentiated chondrosarcoma	2
dendritic cell	1
ea.hy926 cell imortalized endothelial	7
embryonal rhabdomyosarcoma	10
embryonic lung fibroblast	15
embryonic skin fibroblast	23
embryonic stem cell	5
endothelial cell	8

epidermis	5
epidermis dermatitis	12
erythrocyte	6
esophageal adenocarcinoma	6
esophagus Barretts	7
esophagus epithelium	7
ewings sarcoma	29
fetal lung	1
fetal lung fibroblast	20
fibroblast Cockayne syndrome	8
fibromatosis	3
frontal cortex	27
frontal cortex Huntingtons	36
ganglioneuroblastoma	4
ganglioneuroma	1
germ cell tumor	71
gingiva cell periodontitis	3
glioblastoma	2
glioblastoma cell line	5
granulocyte	6
hI60 promyelocytic leukemia	27
head and neck squamous cell carcinoma	8
heart	36
heart disease	51
hematopoietic stem cell	22
hepatocellular carcinoma	60
hippocampus CA1	5
hippocampus CA1 Alzheimers	8
ht-29 colorectal adenocarcinoma	27
human bronchial epithelial cell line	2
hypopharynx	3
hypothalamus	20
k562 myelogenous leukaemia	48
keratinocyte	8
kidney	29
kidney carcinoma	5
kidney epithelium	9
leiomyosarcoma	7
leukocyte	56
leukocyte multiple organ failure	7
ligament cell periodontitis	3

lipoma	3
liposarcoma	2
liver	1
lung	5
lung cancer	77
lung emphysema	1
lymph node	10
lymphocyte	80
lymphocyte asthma	3
macrophage	18
malignant peripheral nerve sheath tumor	2
mcf-7aro breast epithelial adenocarcinoma	23
mesangial cell	8
mesenchymal stem cell	25
microvascular endothelial	6
monocyte	21
monocyte familial hypercholesterolemia	4
monocyte multiple organ failure	5
mononuclear cell	143
mononuclear cell infection	314
monophasic synovial sarcoma	7
multiple myeloma	1
multipotent adult progenitor cell	1
myelogenous leukemia	37
myometrium	12
myxoid liposarcoma	6
neuroblastoma	19
neurofibroma	2
oral squamous cell carcinoma	23
oropharynx	1
osteosarcoma	9
ovarian tumor	93
ovary	4
pc-10 lung sqamous cell cancer	1
pc3 prostate cancer	64
pc6 lung small cell cancer	1
pc7 lung non-small cell adenocarcinoma	1
placenta basal plate	21
plasma cell	1
plasma-cell leukemia	6
preadipocytes	24

precursor T lymphoblastic leukemia	59
prefrontal cortex	2
primary fibroblast	9
primary intervertebral disc	10
progeria syndrome fibroblast	9
prostate cancer	25
prostate gland	11
pterygium	6
regenerating cell periodontitis	3
renal cell carcinoma	11
rpmi-8226 myeloma	1
sarcoma	3
schwannoma	2
sk-mel-28 melanoma	1
skeletal muscle	17
skeletal muscle diseased	50
skeletal muscle muscular dystrophy	33
skin	4
skin basal cell UV treated	3
skmel5 melanoma	21
small intestine	6
smooth muscle	10
spindle cell tumor	5
squamous cell carcinoma of the oral cavity	4
ssMCF7 breast cancer	23
te85 osteosarcoma	8
testis	1
theca	8
theca polycystic ovarian syndrome	5
thymocyte	14
thyrocyte	6
thyroid adenocarcinoma	19
thyroid gland	1
tonsil	10
trabecular meshwork cell	6
ts anaplastic large cell lymphoma	27
u-20s osteosarcoma	2
umbilical vein endothelial cell	42
universal reference	18
unknown gliobastoma	1
unknown lung adenocarcinoma	15

unknown lung cancer	2
unknown lung cell carcinoma	1
unknown lung large cell carcinoma	3
unknown lung mesothelioma	1
unknown lung neuroendocrine	1
unknown lung non small cell cancer	4
unknown lung small cell cancer	17
unknown lung squamous cell carcinoma	2
unknown small cell lung cancer	1
uterine tumor	23

Table C.2.: Number of samples per group for all 191 groups in the own dataset.

Tissue/cell type	Number of samples
Abdominal _cancer	6
Adipose_stem_cell	5
Adipose_tissue	20
Adrenal Gland _cancer	2
Adrenal_gland	8
Amygdala	16
Appendix _cancer	1
Astrocytes	9
B cells	30
Basal_ganglia	113
Bladder _cancer	32
Bladder or Kidney _cancer	1
Bladder_cancer_cell_line	33
Blood	200
Blood_vessel	20
Bone _cancer	4
Bone_marrow	13
Brain _cancer	33
Brain_cancer_cell_line	56
Breast	20
Breast _cancer	371
Breast_cancer_cell_line	224
Bronchus	6
CD11b+	3
CD133+	2
CD235a+	3
CD34+	2
CD4+	2
CD45+	3
CD8+	2
CNS_cancer_cell_line	29
Cerebellum	21
Cerebral_cortex	345
Cerebral_cortex_no_astrocyte	3
Cervix	9
Cervix Uteri _cancer	1
Cervix _cancer	36
Cervix_cancer_cell_line	2

Colon	8
Colon _cancer	358
Colon_cancer_cell_line	174
Common Bile Duct _cancer	1
Corpus_callosum	18
Dorsal_root_ganglia	16
Duodenum _cancer	1
EC	2
Embryo	6
Endometrium _cancer	5
Endometrium_cancer_cell_line	13
Endothelial_cell_line	2
Eosinophils	4
Erythroblasts	8
Esophagus	8
Esophagus _cancer	6
Esophagus_cancer_cell_line	14
Exocrine Pancreas _cancer	1
Fallopian tube _cancer	8
Fallopian_tube	3
Fetal_brain	2
Fetal_liver	1
Fibroblast	68
Fibroblast_cell_line	6
Gallbladder _cancer	2
Gastric _cancer	3
Gastric_cancer_cell_line	24
Glial_progenitor	7
Glioblastoma_neurosphere	8
Granulocytes	2
Granulosa	2
Heart	207
Hematopoietic System_cancer_cell_line	123
Hippocampus	156
Hypothalamus	16
Islets	3
Islets of Langerhans _cancer	1
Keratinocyte	5
Kidney	18
Kidney _cancer	280
Kidney_cancer_cell_line	1

Larynx_cancer_cell_line	2
Leukemia_cancer_cell_line	18
Liver	150
Liver _cancer	125
Liver_cancer_cell_line	32
Lung	7
Lung _cancer	139
Lung_cancer_cell_line	247
Lymph Node _cancer	17
Lymph_nodes	8
Lymphoblastoid_cell_line	35
Lymphoma_cancer_cell_line	6
MPC	2
MSC	13
Medulla	18
Melanoma_cancer_cell_line	51
Mesenchymal cells	1
Microglia	4
Midbrain	64
Monocytes	55
Motor_neuron	8
Myeloid Dendritic cells	5
Myometrium _cancer	1
NK cells	12
NPC	3
Nasal Septum_cancer_cell_line	2
Nasopharynx_cancer_cell_line	2
Neutrophils	6
Nipple	8
Nodose_nucleus	16
Non-Small Cell Lung_cancer_cell_line	26
Oocyte	3
Oral _cancer	13
Oral cavity_cancer_cell_line	4
Oral_mucosa	8
Ovarian_cancer_cell_line	42
Ovary	27
Ovary _cancer	560
Ovary_cancer_cell_line	17
PBMC	3
PSC	473

Pancreas	2
Pancreas _cancer	23
Pancreas_cancer_cell_line	41
Parotid gland _cancer	4
Partial iPSC	8
Pelvis _cancer	1
Penis	6
Penis _cancer	1
Pericardium	1
Peritneum	1
Peritoneum _cancer	22
Pharynx	8
Pharynx _cancer	1
Pharynx_cancer_cell_line	6
Pituitary_gland	14
Placenta	1
Plasma cells	2
Plasmacytoid Dendritic cells	5
Pons	15
Prostate	44
Prostate _cancer	88
Prostate_cancer_cell_line	31
Rectosigmoid _cancer	43
Rectum _cancer	41
Renal pelvis _cancer	15
Renal_cancer_cell_line	32
Rib _cancer	1
Salivary Gland _cancer	1
Salivary_gland	10
Skeletal_muscle	28
Skin	41
Skin _cancer	13
Skin_cancer_cell_line	5
Small intestine _cancer	15
Small_intestine	7
Soft tissue _cancer	34
Spinal cord _cancer	1
Spinal_cord	18
Spleen	10
Spleen _cancer	4
Stomach	24

Stomach (Gastroesophageal junction) _cancer	1
Stomach _cancer	16
Stomach_cancer_cell_line	36
Synovium	14
T cells	95
Testes	10
Testis _cancer	5
Thalamus	24
Thymus	3
Thyroid Gland_cancer_cell_line	4
Thyroid _cancer	36
Thyroid_gland	10
Tongue	18
Tongue _cancer	2
Tongue_cancer_cell_line	12
Tonsil	6
Trachea	8
Trigeminal_ganglia	16
Universal Reference	2
Unknown_cancer_cell_line	5
Ureter _cancer	3
Urethra	8
Uterine Cervix _cancer	1
Uterus	87
Uterus _cancer	231
Vagina	8
Vagina _cancer	1
Vulva	8
Vulva _cancer	11
White_matter	28

Table C.3.: Number of samples per GEO series for all 108 GEO series in the own dataset.

GEO Series	Number of samples
GSE10843	206
GSE1145	37
GSE11882	173
GSE12390	21
GSE12485	10
GSE12486	10
GSE12583	9
GSE13828	10
GSE14711	11
GSE14897	15
GSE14970	12
GSE15175	16
GSE15176	12
GSE15491	3
GSE16093	5
GSE16654	6
GSE16694	8
GSE16963	9
GSE17476	8
GSE18147	10
GSE18180	3
GSE18618	15
GSE18674	24
GSE20033	19
GSE2109	2158
GSE21222	18
GSE21243	4
GSE21610	68
GSE22167	14
GSE22246	13
GSE23402	42
GSE23583	35
GSE23968	14
GSE24223	179
GSE24487	10
GSE24530	8
GSE25090	3
GSE25417	12

GSE26451	6
GSE26455	6
GSE26672	19
GSE27206	9
GSE27280	12
GSE27924	10
GSE28406	16
GSE28490	47
GSE29115	8
GSE29625	14
GSE29783	10
GSE29819	38
GSE30038	8
GSE32474	174
GSE33025	6
GSE33066	6
GSE33109	39
GSE33325	39
GSE33536	9
GSE33903	22
GSE34211	230
GSE3526	353
GSE35373	6
GSE35603	78
GSE35864	72
GSE36098	9
GSE36609	9
GSE36634	24
GSE36667	2
GSE36754	9
GSE37258	18
GSE37842	4
GSE37896	8
GSE37982	3
GSE38069	2
GSE40438	8
GSE40444	8
GSE40709	9
GSE40751	12
GSE40873	49
GSE41804	40

GSE42073	6
GSE42114	19
GSE44841	8
GSE45537	135
GSE49541	72
GSE49910	46
GSE5281	161
GSE57083	627
GSE7179	9
GSE7234	4
GSE7307	677
GSE7624	35
GSE8331	8
GSE8590	17
GSE9171	30
GSE9196	53
GSE9440	8
GSE9510	5
GSE9709	10
GSE9770	34
GSE9832	16
GSE9834	6
GSE9835	6
GSE9843	91
GSE9865	13
GSE9891	285
GSE9894	12
GSE9940	18
GSE9941	8