LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München
Institute of Stem Cell Research
Institute of Computational Biology

Masterarbeit
in Bioinformatik

Analysis of multivariate expression profiles of astrocytes of postnatal and adult mice

Sonja Christina Waldraff

Aufgabensteller:	Prof. Dr. Magdalena Götz
	Prof. Dr. Dr. Fabian Theis
Betreuer:	Dr. Nikola Müller
	Dr. Jovica Ninkovic
	Dr. Giacomo Masserdotti
Abgabedatum:	15.03.2015

Ich versichere, dass ich diese Masterarbeit
selbständig verfasst und nur die angegebenen
Quellen und Hilfsmittel verwendet habe.

15.03.2015 _____

Sonja Waldraff

Acknowledgement

I would like to thank Magdalena Götz and Fabian Theis for providing the interesting and fascinating topic and the possibility to participate in their researches. Giacomo Masserdotti and Jovica Ninkovic I would like to thank for their valuable input, especially giving me feedback on biological questions. Special thanks to Nikola Müller for her great supervision and important hints on bioinformatical issues as well as substantial feedback during the thesis.

Abstract

Astrocytes are cells located throughout the CNS participating in ion concentration regulation and in the creation of the brain environment. Astrocytes help forming synapses and regulate the blood flow. After an injury molecular and functional changes of astrocytes are possible. Reactive astrocytes surround the injured region and interdigitate their processes to separate this region and the surrounding cells. A severe injury can lead to a glial scar mainly consisting of astrocytes.

Microarray experiments are used to measure gene expressions of different cell types or cell states. Analyses of microarrays therefore are a common field in Bioinformatics.

The aim of this master thesis is to compare gene expressions of several cell types of different microarray profiles at once and therefore get new insights into gene regulations of astrocytes. Probes of several different cell types were analyzed with microarrays based on two different microarray platforms. As different platforms contain different genes, we split the work in two parts, one single and one combining dataset. To combine profiles, biological factors, like cell types and known biological features, or non-biological factors, like technical differences between the microarrays. We try to find and correct non-biological factors in the dataset. To avoid removing biological information instead of non-biological bias, detailed knowledge about the biological features is needed.

We used one pipeline for the different microarray datasets. Both datasets were normalized and the dataset that combined several microarray profiles was additionally corrected for non-biological factors. We performed statistical to find similarities and dissimilarities between the cell types. We investigated the relationship between biological features in the dataset, like being astrocytes from the diencephalon or from a lesion side, using a linear regression model to identify differentially regulated genes. This needed broad biological background knowledge of the datasets to choose an appropriate collection of biological factors. To structure the probes of the combined dataset, we used fourteen features for the linear regression model calculating a t-statistic for each factor and all genes. We could observe which genes were expressed statistically significantly in a factor. Using the list of significant genes we started a functional analysis looking for enriched functions and pathways of the genes for each factor and we also could compare the two dataset. Next we used the fact that genes mostly interact in groups and integrated protein-protein interaction information of gene networks to the t-statistic calculation of the regression. We compared the significant genes before and after this network smoothing. For some biological features we identified further genes but in most cases the number of significant genes decreased.

Finally we suggest integrating information across biological features when applying the network smoothing method, additionally as they can depend on each other. Therefore we showcase our ideas of regression smoothing on a mini-dataset with seven example genes.

Zusammenfassung

Astrozyten sind Zellen, die im ganzen zentralen Nervensystem vorkommen und die an der Ionenkonzentrationsregulierung und der Bildung der Hirnumgebung beteiligt sind. Astrozyten helfen dabei Synapsen zu formieren und regulieren den Blutfluss. Nach einer Verletzung ist es möglich, dass Astrozyten sich molekular und funktional verändern. Reaktive Astrozyten umgeben die verletzte Region und verbinden ihre Fortsätze um diese von den umgebenden Zellen zu separieren. Eine ernste Verletzung kann zu einer Glianarbe führen, die hauptsächlich aus Astrozyten besteht.

Microarray Experimente werden zur Messung von Gen-Expressionen verschiedener Zelltypen oder verschiedener Zellzuständen verwendet. Daher werden in der Bioinformatik oft Microarrays analysiert.

Das Ziel dieser Masterarbeit ist die Expressionen von Zelltypen verschiedener Microarray Profile zusammen zu vergleichen und dadurch neue Erkenntnisse über Astrozyten zu erhalten. Dabei wurden Proben von Astrozyten in verschiedenen Zuständen, neuronale Stammzellen, radiale Gliazellen, embryonale Stammzellen und Neuronen zur Verfügung gestellt. Diese Proben wurden mit Microarrays untersucht, wobei zwei verschiedene Microarray Plattformen verwendet wurden. Da auf verschiedenen Plattformen unterschiedliche Gene aufgetragen sind, teilten wir die Daten in zwei Datensätze auf, ein einzelner und ein zu kombinierender, die Expressionsprofile von Astrozyten und anderen Zelltypen beinhalten. Für die Kombination dieser Profile mussten viele Faktoren betrachtet werden. Das konnten biologische Faktoren, wie Zelltypen und bekannte biologische Eigenschaften, sowie nicht-biologische Faktoren, wie technische Unterschiede zwischen den Microarrays, sein. Wir versuchten daher die unterschiedlichen nicht-biologischen Faktoren in dem Datensatz zu finden und zu korrigieren. Dafür wiederum waren detaillierte Kenntnisse über die biologischen Eigenschaften nötig, damit wir nicht tatsächliche biologische Informationen entfernen.

Auf beiden Datensätzen verwendeten wir eine Abfolge verschiedener Microarray Analysetechniken mit denen neue Erkenntnisse über Astrozyten erlangt wurden. Zu Beginn wurden die beiden Datensätze jeweils normalisiert und der kombinierte Datensatz wurde zusätzlich noch auf Unterschiede zwischen den Microarrays korrigiert. Wir führten statistische Analysen wie Hauptkomponentenanalyse und Clustern durch, wobei wir einige Ähnlichkeiten und Unterschiede zwischen den Zelltypen feststellten. Wir betrachteten die Beziehungen zwischen biologischen Eigenschaften, wofür wir ein lineares Regressionsmodell erstellen. Dies erforderte ein breites biologisches Hintergrundwissen der Daten, um eine geeignete Auswahl an biologischen Faktoren zu treffen, wie zum Beispiel der Faktor „erwachsen“ für alle Proben, die aus erwachsenen Mäusen entnommen wurden und nicht aus einer Zellkultur. Zur Strukturierung der Proben des kombinierten Datensatzes verwendeten wir vierzehn Eigenschaften, um das lineare Regressionsmodell zu erstellen.

Dieses berechnete für jeden Faktor eine t-Statistik über alle Gene. Daraus haben wir abgelesen, welche Gene in einem Faktor statistisch signifikant exprimiert waren. Mit den Listen von signifikanten Genen führten wir eine funktionelle Analyse durch. Für jeden Faktor suchten wir häufig auftretende Funktionen und Pathways der Gene. Auf dieser Basis konnten wir auch den separierten mit dem kombinierten Datensatz vergleichen. Als nächstes nutzten wir die Tatsache, dass Gene meist in Gruppen agieren, und binden Protein-Protein Interaktion Informationen von Gennetzwerken in die Berechnung der t-Statistik der Regression ein. Wir verglichen die signifikanten Gene vor und nach diesem Netzwerk-Smoothen. Für manche biologische Faktoren konnten wir weitere Gene finden, aber in den meisten Fällen ist die Anzahl der signifikanten Gene kleiner als vorher.

Abschließend schlagen wir vor zusätzlich auch Informationen über die biologischen Faktoren zu berücksichtigen, da diese abhängig voneinander sein können. Wir entwickelten dafür Ideen, die wir auf einem kleinen Beispieldatensatz mit sieben Genen darstellen und durchführen.

Table of contents

1. Introduction 1
   1.1. Aim of this thesis 2
   1.2. Summary of this thesis 2
2. Background 4
   2.1. Transcriptome 4
   2.2. Cell types and cellular growth factors 5
      2.2.1. Embryonic stem cells 5
      2.2.2. Central nervous system 5
      2.2.3. Cellular growth factors 10
   2.3. Statistical Background 11
      2.3.1. Principal component analysis 11
      2.3.2. Distances measures 11
      2.3.3. Clustering techniques 11
      2.3.4. Linear regression model 12
      2.3.5. Hypothesis testing 12
      2.3.6. Multiple hypothesis testing 13
   2.4. Software 13
   2.5. Bioinformatics Background 13
      2.5.1. Webservers/Databases 14
      2.5.2. Normalization of microarray data 15
      2.5.3. Removing biological biases 15
      2.5.4. Functional analyses 16
      2.5.5. Network smoothing 17
3. Materials and Methods 19
   3.1. Transcriptome datasets 19
      3.1.1. Growth factor stimulation of astrocytes 19
      3.1.2. Combining various cell types 20
   3.2. Normalization of the microarray expression profiles 23
   3.3. Removing non-biological bias 25
   3.4. Statistical analysis 26
   3.5. Functional analysis 27

3.6. Network smoothed t-statistics 28
3.7. Methods proposal for regression smoothing 29
4. Results 32
4.1. Growth factor dataset 32
4.1.1. Statistical analyses 32
4.1.2. Functional analysis 36
4.2. Combined dataset analysis 37
4.2.1. Removing biological bias 37
4.2.2. Investigate the corrected dataset with statistical analyses 38
4.2.3. Functional analyzes 46
4.2.4. Network smoothing using a regression model 47
4.2.5. Novel approaches for regression smoothing 55
4.3. Comparison of functional analyses 57
5. Conclusion 59
6. Appendices 61
7. References 68

Figures

Fig. 1: Procedure of an Affymetrix GeneChip experiment. (Scheme by http://www.dkfz.de/gpcf/24.html) 4
Fig. 2: A schematic overview of different cell types in the CNS, by Clarke, 2003 (Illustration by Cheng-Jung Lai) [18]. 5
Fig. 3: Morphology of a protoplasmic astrocyte. 8
Fig. 4: Images of immunohistochemical staining of GFAP. 9
Fig. 5: Timepoints of taking probes from P6 mice cell cultures. 20
Fig. 6: Pictures of astrocytes provided by the Götz Lab. 20
Fig. 7: Overview of probes from the embryonic stem cell culture. 21
Fig. 8: Graphical representation of the datasets. 22
Fig. 9: RNA degradation of the 45 samples before normalization showing the mean intensity. 23
Fig. 10: Density Histogram for each of the 45 samples after normalization. 24
Fig. 11: Boxplot of the normalized log intensities for each of the 45 sample after normalization. 24
Fig. 12: Illustration of extending a network for regression smoothing. 31
Fig. 13: The first two principal components for the samples are plotted. 32
Fig. 14: PCA rotation matrix of genes and the standard deviations of the principle components. 33
Fig. 15: Global sample clustering using Euclidean distances. 34
Fig. 16: Number of genes which t-statistics are statistical significant with a p-value <0.01 for the coefficients ‘SAF’, ‘SAFE’ and ‘SN’. 34
Fig. 17: Gene regulation comparisons of the coefficients ‘SAF’, ‘SAFE’ and ‘SN’. 35
Fig. 18: Volcanoplots for ‘SAF’ and ‘SAFE’ and comparison of the significant genes of these two coefficients. 35
Fig. 19: Comparison of FGF and EGF using scatterplot and Venn diagram. 36
Fig. 20: Robustness of samples localized in PC1 using ComBat with 50 random design matrices 37
Fig. 21: PCA result for the 43 samples. The first two principal components are plotted. 38
Fig. 22: PCA rotation matrix of genes and the standard deviations of the principle components. 39
Fig. 23: Hierarchical clustering of the Pearson correlation coefficients for the corrected dataset. 40
Fig. 24: Number of significant genes for the coefficients of the combined dataset. 43
Fig. 25: Overlap of significant genes for coefficients ‘P6’, ‘Adult’ and ‘Astrocytes’ on the left. On the right ‘DIEC’ and ‘PastPl’ are added, too. 44
Fig. 26: Volcano plot of the coefficient ‘Astrocytes’ of the linear regression model. 45
Fig. 27: Regulation comparison of the coefficients ‘aNSC’ and ‘Astrocytes, ‘aDIEC’ and ‘aNSC’ as well as ‘aDIEC’ and ‘Lesion’. 45
Fig. 28: Expressions of the 50 most significant genes of ‘P6’ 46

Fig. 29: One cluster as example network with nodes Prelp, Smoc2, Fmod, Bgn, Tnc, Mmp2 and Ptn. 47
Fig. 30: Expression values for the mini-network. 49
Fig. 31: Comparison of the t-statistics for coefficient ‘p48h’ before and after smoothing. 50
Fig. 32: Gene expressions of 89 genes significant after smoothing for coefficient ‘p48h’, which were not significant before. 52
Fig. 33: Neighborhood network of the significant genes of ‘p48h’. 53
Fig. 34: Protein-protein interaction network of the significant genes of ‘Astrocytes’. 54
Fig. 35: Estimations of GO-annotations by MGSA for both growth factor dataset and combined dataset. 57
Fig. 36: KEGG-pathway estimations across the two datasets. 58
Fig. 37: Hierarchical clustering of the Pearson correlation coefficient of the samples of the growth-factor-dataset. 62
Fig. 38: Hierarchical clustering of the Euclidean distances of the samples of the combined-dataset. 62
Fig. 39 : Volcano- and Scatterplots for biological factors of the combined-dataset. 64

Tables

Table 1: Example extract of a design matrix generated for ComBat. 25
Table 2: Design matrix used for linear modeling. 42
Table 3: KEGG pathways for ‘P6’ 47
Table 4: Kernel matrix for the mini-network with nodes “Prelp”, “Smoc2”, “Fmod”, “Bgn”, “Tnc”, “Mmp2” and “Ptn” 48
Table 5: t-statistic matrices divided by the maximum per coefficient before (T) and after smoothing (T). 48
Table 6: p-values for the t-statistics of the linear regression (P) and for the smoothed t-statistics calculated via permutation test (P). 49
Table 7: Number of significant genes before and after smoothing. 51
Table 8: Results of regression smoothed t-statistic using aggregation. 55
Table 9: Results of regression smoothed t-statistic using extended networks. 56
Table 10: Design matrix for ComBat. Columns represent the covariates/coefficients and rows the samples. 61
Table 11: KEGG pathways for features ‘SAF’ and ‘SAFE’ of the growth-factor-dataset 65
Table 12: GO annotations for features of the combined-dataset. 66
Table 13: KEGG pathways for features of the combined-dataset. 67

List of abbreviations

CA	cellular aggregates
CNS	central nervous system
EB	Empirical Bayes
EGF	epidermal growth factor
ESC	embryonic stem cell
FDR	false discovery rate
FGF	fibrous growth factor
FGF2	basic FGF
GFAP	glial fibrillary acidic protein
GFP	green fluorescent protein
GO	gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
NSC	neural stem cell
PCA	principal component analysis
RA	retinoic acid
RMA	robust multi-array average
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVZ	subventricular zone

List of abbreviations for biological samples

SAF	Samples of astrocytes treated with FGF
SAFE	Samples of astrocytes treated with FGF and EGF
SN	Samples of neurospheres
P6	Samples of postnatal six mice cell cultures
p4h	Samples of P6 after four hours
p24h	Samples of P6 after twenty-four hours
p48h	Samples of P6 after forty-eight hours
pDel	Samples of P6 six days after culturing
a1DIEC	Samples of astrocytes taken from the diencephalon of adult mice
a1NSC	Samples of neural stem cells taken from the SVZ of adult mice
a1	Samples of the first adult batch including DIEC and NSC
a2GFPplus	Samples of reactive astrocytes of adult mice (from cortex)
a2GFPminus	Samples of not astrocytic cells at a lesion of adult mice (from cortex)
a2WT	Samples of astrocytes taken from healthy tissue of adult mice (from cortex)
a2	Samples of the second adult batch including GFPplus, GFPminus and WT
eCa4d	Samples of CA of ESC four hours before RA is added
eCa6d	Samples of CA of ESC six hours after RA is added
eCa8d	Samples of CA of ESC eight hours before RA is added (radial glia)
eP14h	Samples of radial glia fourteen hours after plating
eNd7	Samples of cells seven days after plating (neurons)

1. Introduction

The central nervous system (CNS) is a highly complex system composed of various cell types together coordinating information processing and flow throughout the body. The two cell types forming the nervous tissue are glial cells and neurons. Subtypes of glial cells are, for example, oligodendrocytes, microglia, radial glial cells and astrocytes. Astrocytes can appear in different forms. They show differences in healthy and injured tissue. The molecular mechanisms of astrocytes in the CNS are still poorly understood [1]. Furthermore, it is unknown how mechanisms differ between embryonal and adult stage. To understand complex molecular interactions high-throughput screenings are used. Here we examine gene expression profiles across different forms of murine astrocytes in addition with neurons and neural stem cells during development and adulthood.

Astrocytes are located throughout the CNS. They participate in the regulation of calcium ion concentration [2] and are involved in the creation of the brain environment. Astrocytes help forming synapses and regulate the blood flow. Furthermore, they play a role in maintaining the blood brain barrier. After an injury astrocytes can change molecularly and functionally. Thereby, both loss and gain of functions can alter the astroglial activities [3]. Reactive astrocytes surround the injured region and interdigitate their processes to separate this region and the surrounding cells. The activation of reactive astrocytes depends on the severeness of the lesion. Reactive astrogliosis is main part of a glial scar. A severe injury leads to cell death and the generated empty space is filled with a fine network of astrocyte processes [4]. This glial scar is an important part for the protection and repair of neurons [3].

Typically statistical and bioinformatical analysis of high-throughput screens are pairwise comparisons [5, 6]. A t-test, for example, is a method to determine a difference between two groups in a statistical significant way. However, for more complex experimental setups pairwise comparisons cannot resolve dependencies. For example, assuming probes taken from groups like women, men, boys and girls, a paired test of women against boys would not show if differences between them are caused of gender or age. We therefore needed more complex methods to analyze complex setups. One possible approach for this problem was using regression models [7]. Nevertheless, for complex datasets it was not straight forward to derive and design the covariables capturing several biological features and dependencies. In addition functional testing take place on pairwise comparisons usually. Typically, Fisher's exact test or model based ontology analysis like mgsa served to find enriched functional terms of gene sets [8]. Those enriched terms characterized covariables.

As genes function in clusters, an approach to improve the statistical analysis with biological knowledge was the so-called network smoothing using known protein-protein interaction

networks. MATISSE, for example is a tool, which searches for co-expressed subnetworks that are connected significantly [5]. The network based stratification (NBS) approach integrates somatic tumor genomes using the mutation profiles with gene interaction networks [9]. A third technique applies smoothed t-statistic with a random walk kernel. This stSVM approach integrates a paired t-test [10].

1.1. Aim of this thesis

We aimed to comprehensively summarize and describe expression profiles of cell types from the CNS to get new insights in the biology of astrocytes. We searched for new receptors and transcription factors specific for astrocytes. Additionally we wanted to find gene clusters typically for astrocytes and we wanted to investigate if there are more neuronal astrocytes.

The Götz-lab provided five microarray profiles containing seventeen different cell types processed on two distinct Affymetrix microarray platforms. Therefore, we separated the dataset into two parts, the growth-factor-dataset containing probes of astrocytes and neural stem cells (NSC) treated with growth factors and the combined-dataset including samples of post-natal six (P6) cell cultures, direct adult lines and cell cultures of embryonal stem cells (ESC).

1.2. Summary of this thesis

The background chapter summarizes important information about biological knowledge like the different cell types and about statistical methods. Additionally it includes bioinformatical background.

Chapter 3 provides detailed descriptions of datasets and methods used to generate results. The combined datasets underlie batches, which had to be removed, and samples were not matched across experiments. That is why we used the state-of-the-art empirical Bayes method called Combat [11] to remove batch effects and also evaluated by bootstrap in the combined-dataset (3.3).

We applied statistical tests (3.4) to get information about similarities and dissimilarities between the cell types. We therefore used principal component analyses, Pearson correlation coefficients and Euclidean distances. Those analyses showed us some clusters of astrocyte samples and the other cell types. Additionally, we wanted to compare the relationship of biological features and generate a linear regression model for that. We used the three different cell types of the growth factor dataset. For the combined dataset finding a good representation of biological factors was difficult. The dataset was too complex to apply pairwise comparison. So we designed a matrix where the relationships between samples and biological features were indicated. However, many different features existed and not all could be included in the model. Our final matrix was iteratively designed in close

communication with the biological partners (ISF, HMGU), but other combinations would be possible. With this matrix we generated the linear regression model for the combined dataset. This calculated different statistical tests including a B-statistic and a t-statistic for each gene in each biological factor indicating if the gene was expressed statistically significantly for the specific feature [12]. A detailed description of this procedure is given in this section.

With the lists of significant genes we applied functional analysis (3.5). Therefore we used the gene ontology (GO) and KEGG databases to find gene functions and pathways that were enriched in one of the gene lists gaining further characteristics about the cell types. Thereby we were able to compare the results of the growth factor and the combined dataset.
In the next step we tried to integrate the results of the linear regression profile by including information of neighboring genes using a corresponding protein-protein network. Therefore protein-protein network information was regarded. We chose to implement a method based on the stSVM approach [10] where a p-step random walk kernel was used on the gene network, which we used for the analysis. Instead of the paired t-test, we now used the results of the linear regression. For some biological features we could identify new significant genes and for some others we viewed a lower number of significant genes.

Section 3.6 describes the smoothing of the t-statistic using gene networks. The combined-dataset consisted of various cell types with several biological features. Therefore we designed covariables and applied linear regression instead of pairwise tests. We performed network smoothing on the t-statistic. However, the covariables partly depended on each other. Therefore we proposed and developed some ideas (Section 3.7) to improve the results of the linear regression not only by gene network information but also by integrating dependencies between biological features. This we tested only on a mini-dataset with seven example genes. Further researches on such a “regression smoothing” might further refine microarray analysis of multivariate experimental designs.

Chapter 4 summarizes the results of the analyses for the both sets. New insights into astrocytes could be observed. Additionally we compared the results of the functional analysis of the two datasets (Section 4.3).

2. Background

2.1. Transcriptome

DNA is transcribed into messenger RNA (mRNA) which carries its genetic information to the ribosome. At the ribosome mRNA is translated into an amino acid sequence of a protein. Like DNA, mRNA is a nucleotide sequence including the nucleotide uracil instead of thymine. As proteins are difficult to analyze because of their three-dimensional structure, the amount of mRNA is used to evaluate functions of a particular cell. If a gene is active, it produces specific mRNAs. The amount of the mRNA from each gene in a specific cell or tissue is the gene expression.

Microarray is a technology for measuring thousands of gene expressions qualitatively. They can be used to examine gene expression between different cell types or different populations [13]. To measure expression levels of genes of a cell type, the RNA of interest, the target, is extracted and labeled. Particular probe DNA sequences are affixed on a solid matrix. Then the DNA copy (cDNA) of the RNA sample is hybridized to these DNA sequences and the abundance of the nucleotide sequences is measured quantitatively. Sample material is fluorescently labeled such that probe expression can be inferred from imaging of the microarray. Therefore microarray technology is a tool for screening biological samples for changes in mRNA expressions [14].

This master thesis used expression profiles of Affymetrix microarray platforms, which belong to the single-channel microarrays. Affymetrix GeneChips are microarrays that are available for different model species including human and mouse. The GeneChips use glass surfaces containing lots of oligonucleotides, which typically have a length of 25 nucleotides. For each

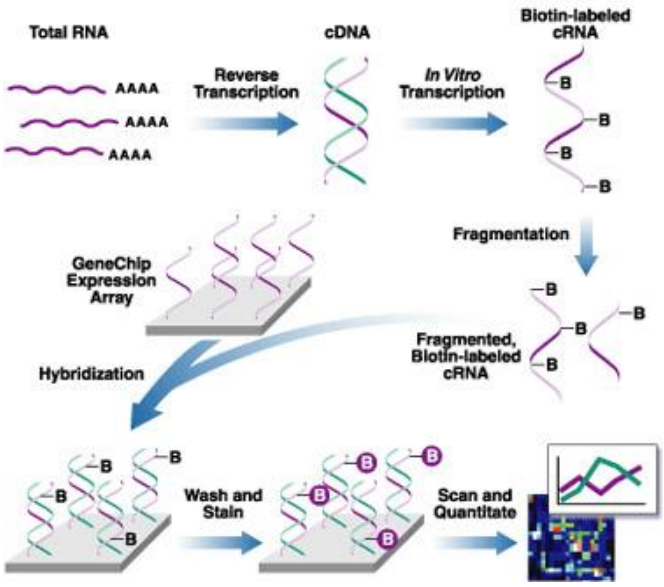


Fig. 1: Procedure of an Affymetrix GeneChip experiment. (Scheme by http://www.dkfz.de/gpcf/24.html)

target (e.g. gene of interest) several probes are designed, for which each one probe (oligonucleotide) is built identical to the target and another with a single base mismatch in the middle of the sequence. The array is synthesized by photolithography and measures the average differences in the intensities of the target and the mismatch sequence for all probe pairs forming an expression index [13]. The resulting expression profile can then be analyzed. Fig. 1 gives a schematic overview. For comparison of the data profile with arrays from other experiments, such a single-channel microarray is easier to handle as there is no competitive hybridization of two samples. In any case, batch effects must be considered.

2.2. Cell types and cellular growth factors

The following segment explains a couple of cell types important in the subsequent analyses. Those are stem cells, radial glia and astrocytes. Additionally growth factors promoting the differentiation of the cells are presented.

2.2.1. Embryonic stem cells

Embryonic stem cells (ESCs) are cells extracted from the blastocyst of an embryo. In general, they are defined as cells that are not differentiated and which have the ability to differentiate into one or several types of more specialized cells. Furthermore ESCs have the capacity of unlimited self-renewal. ESCs can generate body tissues [15].

2.2.2. Central nervous system

In the central nervous system (CNS) a distinction is made between neurons and glial cells, whereby glial cells include all cells in the CNS like oligodendrocytes, astrocytes, microglia and

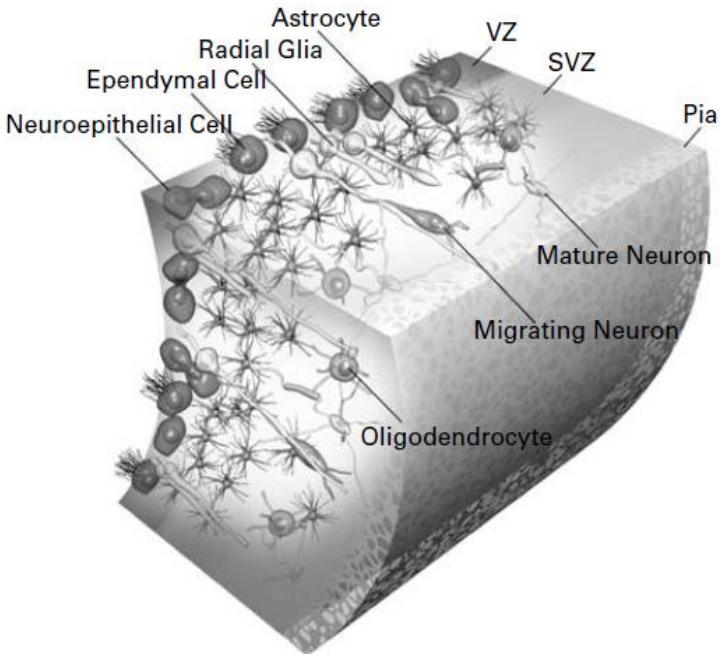


Fig. 2: A schematic overview of different cell types in the CNS, by Clarke, 2003 (Illustration by Cheng-Jung Lai) [18].

radial glia that are not nerve cells [16]. Glial cells have a high heterogeneity. For example, astrocytes and radial glial cells both differ in various brain regions [17]. In the CNS of adult vertebrates mainly four cell types are differentiated. Those are neurons and astrocytes as well as oligodendrocytes and ependymal cells [18]. Fig. 2 shows an illustration of different CNS cell types.

2.2.2.1. Neural stem cells

Neural stem cells (NSCs) can give rise to neurons, astrocytes and oligodendrocytes in the CNS [18]. NSCs are self-renewing, but this is narrowed to a limited cell division number. NSCs divide symmetrically and asymmetrically during development. Thereby, symmetric divisions give rise to two daughter cells whereas asymmetric division results in one cell identical to the mother cell and one other cell which is more differentiated, like a neuron [19]. NSCs differentiate into neurons and glial cells via neurogenesis and subsequently gliogenesis. The multipotency is lost [18, 20]. However, this depends on time and region. The neurons migrate with the help of newly generated glial cells beyond the ventricular zone guided by radial glia [18, 21]. The subventricular zone (SVZ) is built. In this region a high concentration of NSCs is shown in adult and postnatal brains [22]. The remaining neuroepithelial cells of the ventricular zone differentiate more glioblast. Because of some clonal studies, NSCs are assumed to be the main origin of glial cells. They migrate to other locations of the CNS, proliferate and differentiate into oligodendrocytes and astrocytes. Adult neural stem cells (aNSC) are identified in the SVZ [18, 23].

In culture NSCs form highly dynamic and complex structures termed neurospheres. Cells of neurospheres show heterogeneous morphologies. In one neurosphere, for example, cells coexist in different cell cycle phases and various sizes [23, 24]. A bigger sphere has higher cell type heterogeneity [25]. Within these neurospheres the localization of the NSCs influence biological processes like apoptosis or mitosis. NSCs have a high plasticity and therefore have the ability to participate in processes transforming like maturation. They can migrate or differentiate into neurons and glia [22, 23]. A hypothesis is that the clustering of NSCs into neurospheres occurs due to survival issues in culture environment [23]. Neurospheres can be used for NSC assays and are valuable as a system to model neural development and neurogenesis [25]. Neurospheres can be established from NSCs and progenitor cells of the adult CNS as well as from ESCs. The cells are plated on a medium containing fibrous growth factors (FGF) and sometimes additionally epidermal growth factors (EGF) [25, 26]. Those growth factors have an influence on the size of neurospheres [23]. In each neurosphere cells in diverse phases of differentiation are included like neural progenitor cells that proliferate, postmitotic glia, neurons and certainly stem cells [25, 26]. Enlarged neurospheres show an increase in necrotic and apoptotic events of NSCs [23].

2.2.2.2. Radial glial cells

Characterization of radial glial cells ensues via their morphology as well as via their astroglial characteristics. Like NSCs they have long radial processes where the cell body is located in the ventricular zone [17, 21]. The processes are connected to the ventricular zone and to the pial surface. They are frequently connected with blood vessels [27]. Proliferation of almost all radial glial cells occurs throughout neurogenesis.

Radial glial cells are one of the earliest cell types in the CNS that differentiate. During the development of the nervous system they appear in almost all regions of the CNS [28]. NSCs differentiate into radial glia. Then the radial glial cells undergo asymmetric divisions, where on one hand radial glia cells and on the other hand intermediate progenitor cells or neurons arise [2]. Their main role is to guide migrating neurons. Nevertheless, during development of the CNS radial glial cells have a role as precursor cells that can generate neurons and glia [17, 27, 29, 30] and give rise to adult stem cells of the SVZ, too [31].

Additionally radial glia participate in patterning and act as key elements in differentiation of the CNS at specific regions. A role as progenitors of astrocytes is assumed as, after neuronal migration, a transformation of radial glial cells into multipolar astrocytes was observed in the mammalian CNS [17, 21, 29].

Radial glial cells are non-neuronal but are involved in generation and migration of distinct neuronal subtypes and even may be involved in differentiation and specification of those subtypes [17]. Regional differences are observed for radial glia in case of their neurogenic potential [21, 30, 32]. Neuronal migration in the developing brain is limited through boundaries generated by radial glia. Most neurons in the brain are built by radial glial cells and the radial glial neurogenesis is timed differently in diverse regions [30]. Radial glial cells express some characteristic molecules of astrocytes in the CNS like GFAP, the intermediate filament protein [17, 29], and a close relationship between astrocytes and radial glia is suggested. In fact radial glial cells disappear obviously in postnatal state and to this time astrocytes appear [2].

In culture radial glia can be differentiated from ESCs. Therefore the cellular aggregates in the culture are neuronal induced with retinoic acid (RA). ESCs that are treated with RA are known to differentiate into various types of neural cells including neurogenic radial glia and neurons [33, 32].

2.2.2.3. Astrocytes

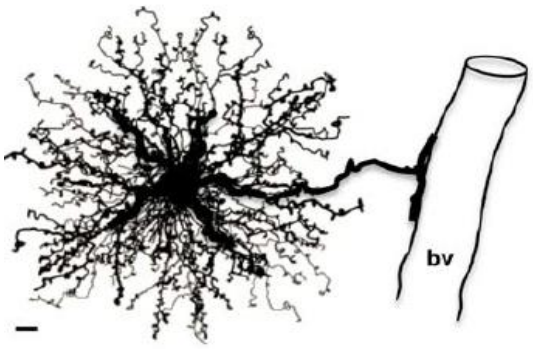


Fig. 3: Morphology of a protoplasmic astrocyte. A large stem branch is connected to a blood vessel and the neuronal cell body is surrounded by a dense network of thin branches. Adopted by **[3]**.

Astrocytes are cells located all over the CNS in both the white matter (fibrous astrocytes) and the gray matter (protoplasmic astrocytes) as a subtype of glia cells. They have a star-shaped structure and are closely related to neuronal synapses. The neuronal cell body is surrounded by a network of fine branches, whereas a long stem branch is connected to a blood vessel (Fig. 3). Furthermore, astrocytes form gap junctions between distant processes of adjacent astrocytes and are linked to synapses, too [16, 3]. Nevertheless, astrocytes are very heterogenic in function and morphology. Like neurons they have potassium and sodium channels. However, action potentials cannot be propagated along astrocyte processes [3]. Similar to radial glia, astrocytes can display fast changes in the intracellular concentration of calcium [2].

Astrocytes are involved in brain environment creation, development of neural cells and during development of gray and white matter. Furthermore they are important for the function and formation of synapses during their development and play a role in the regulation of the blood flow, in homeostasis as well as in energy and metabolism. They also are important for the blood brain barrier. During periods of high neuronal activity and during periods with low glucose in the blood, neuronal activity can be sustained by astrocytic glycogen utilization [3, 4]. At synapses astrocytes help to control the levels of, for example, neurotransmitters and potassium ions in the extracellular space [16].

Three subtypes of astrocytes are astrocytes in healthy central neural tissue, reactive astrogliosis and glial scar formation (Fig. 4, [1]). Astrocytes in healthy tissue show little proliferation and are hardly new generated, whereas reactive astrogliosis happens after an injury or infection. Thereby, astrocytes can show different changes like alterations in molecular expression, cellular hypertrophy or proliferation. Astrocyte activities can be changed through loss and gain of functions. Reactive astrocytes can have pro- or anti-inflammatory potential, which can arouse advantages and disadvantages for the surrounding cells. Depending on the severity of the injury alterations in astrocytes can be different and reactive astrogliosis can appear in a mild to moderate form. In these form the cells have the potential for returning to a similar appearance as in healthy tissue after solving the triggering mechanism [3]. Along with severe reactive astrogliosis compact glial scars can form. The function of reactive astrogliosis and glial scar formation seems to be the protection of neural cells after an injury or infection and therefore play an important role in CNS inflammation

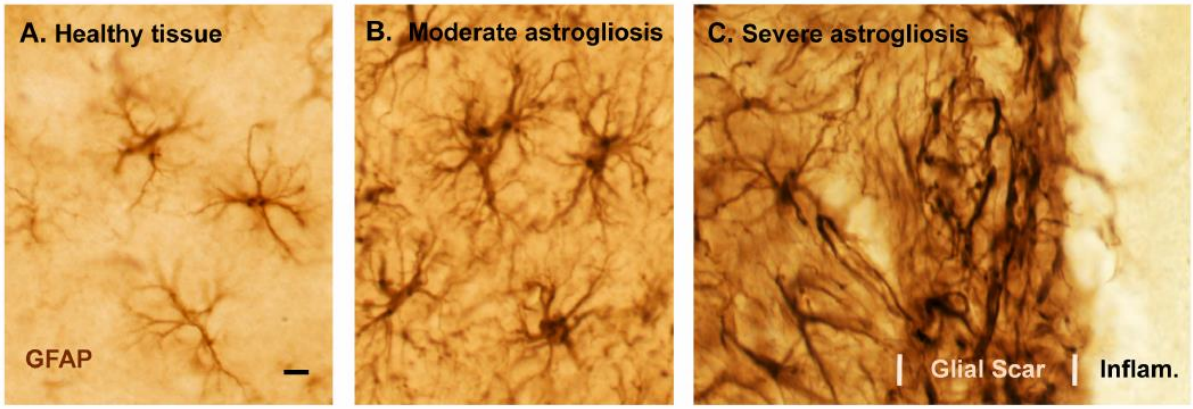


Fig. 4: Images of immunohistochemical staining of GFAP.
A) Astrocytes in healthy tissue is shown. B) shows a moderate form of reactive astrogliosis and in C) severe astrogliosis along with glial scar formation can be viewed. Adopted by Sofroniew, 2009 **[1]**.

regulation [3, 1]. According to Sirko et al. reactive astrocytes "can form self-renewing and multipotent neurospheres in vitro" [34].

Cell death begins directly after a lesion. Thereby, both neuronal and glial cells die including astrocytes. In the beginning this is restricted to the torn region, but after a few days the lesion spreads more and more. Macrophages are the first cells that arrive from the bloodstream and from the surrounding tissue microglia are migrating. A few days later further other cells are recruited from the environment and astrocytes get activated and start proliferating towards the site of injury. Finally, the glial scar consists of a network of predominantly astrocytes interdigitating their processes. This meshwork is entangled and bound together by tight und gap junctions and surrounds the injured region like a wall. Therefore, the glial scar has an inhibitory influence on axonal regeneration [4, 35].

Astrocytes near the lesion upregulate the production of nestin and vimentin [3]. Nestin and vimentin are intermediate filaments which exist in different cell types in the CNS like astrocytes [36]. This characteristic also applies to radial glia cell showing their relationship [17]. Mature astrocytes contain vimentin and another intermediate filament, the "Glial fibrillary acidic protein" (GFAP), but no nestin anymore. Especially in reactive astrocytes in the CNS this is documented [36]. GFAP is essential in reactive astrogliosis and glial scar formation processes and therefore is useful for immunohistochemical identification for targeting astrocytes. Within an astrocytic reaction, an up-regulation of this protein can be observed. In combination with vimentin GFAP characterize the glial scar [4]. Another sensitive chemical marker for astrocytes is the protein Aldh1L1..

In the diencephalon multipotent precursor cells can be found. If EGFs are withdrawn those precursors can be differentiated into neurons and glia. Neuronal cells are built by stimulating the precursors with FGF2 and additionally a medium that is conditioned by a cell line of astrocytes [37]. Proliferation of CNS stem cells is induced by FGF family members. In addition

FGFs can initiate stem cells to differentiate into astrocytes and neurons [37]. Diencephalic astrocytes are glial cells that are differentiated postmitotically. Adult NSCs and hGFAP:GFP+ cells, however, contain progenitor cells and proliferating stem cells [24].

2.2.3. Cellular growth factors

Growth factors are proteins that are transferred as signals from one cell to another and thereby convey information. Beside their function as signal proteins they are important in the development of multicellular organisms.

One important member of the growth factors is the **fibroblast growth factor (FGF)** family belonging to the polypeptide growth factors. In human and mice altogether twenty-two FGFs are known. FGFs are conserved through different species especially in vertebrates for example in zebrafish, mouse and human. Similarities in amino acid sequence and gene structure are revealed [38, 39]. There are four tyrosine kinase receptors, the FGF receptors (FGFR1-4), in mice and humans. FGFs play a role in embryonic development, cell division and cell growth, tissue repair, angiogenesis, cell differentiation, cell migration and proliferation [40]. Originally two FGF could be isolated. One was the acidic FGF (FGF1) and the other the basic FGF (FGF2) [38, 39]. FGF2 seems to play a role in the production of CNS neurons mostly generated during early development [41]. Usually FGF2 has a mitogen activity and promotes the proliferation of NSCs. This is transmitted by the FGF receptors and some signaling pathways to regulate astrocyte specification. Thereby FGF2 influences the NSC potential to undergo gliogenesis. Additionally FGF signaling is involved in radial glia generation. These radial glia cells are later differentiated into astrocytes. For proliferation an FGFR signal is delivered by a MAP kinase pathway [20]. A tight regulation of the FGF signaling is necessary. Thereby many of FGF-regulating factors are controlled by FGFs themselves by a negative feedback loop [40].

Another growth factor is the **epidermal growth factor (EGF)**. Together with FGF2, EGF stimulates proliferation in astrocytes and DNA synthesis. The MAP kinase pathway is activated if astrocytes are stimulated with EGF and other mitogens. This again induces the cells to change their genetic expression pattern required for proliferation. In astrocytes, EGF stimulation leads to the expression of FGF2 [42]. EGF improves tissue regeneration and wound healing in several adult organs. In the CNS EGF is involved in the neuronal development. It has trophic and mitogen actions. EGF participates in cell division and proliferation [41], [43]. Progenitor cells that give rise to glia and neurons respond to EGF [43]. Often EGF and FGF2 are used to promote the production of neurospheres [22]. Together they induce neural precursor cell proliferation from specific brain regions [44]. Furthermore they seem to stimulate embryonic or adult precursors of the CNS to divide. Both EGF and FGF2 interact with tyrosine kinase receptors inducing the activation of the MAP kinase pathway [44] and participate in neurogenesis forming neurospheres as well as in gliogenesis.

2.3. Statistical Background

2.3.1. Principal component analysis

Principal component analysis (PCA) is an unsupervised technique that identifies directions of largest variance within high-dimensional data. It can be used to simplify and to visualize data globally. Using only components, which explain most of the variance in the data, allows reducing the dimensionality. When applying PCA to gene expression dataset, the variations that can be detected between the high-dimensional samples are captured in vectors, the principal components, which are linear combinations of the genes. By design, the second principal component captures less variation then the first, the third less than the second and so on. For visualization usually the first two principal components are used [6]. In this thesis we manually analyzed the number of principal components explaining largest fraction of variance in the data. The standard deviation of the principal components shows how much of the data is described by each principal component.

2.3.2. Distances measures

Distance measures like Pearson correlation coefficient, Euclidean distance or mutual information indicate the degree of similarity.

The Pearson correlation coefficient is defined as a linear correlation measure of two random variables X, Y [6] with means $E(X), E(Y)$ and standard deviations σ_X, σ_Y, respectively.

$$cor(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

with $cov(X,Y) = E[\big(X - E(X)\big) * \big(Y - E(Y)\big)]$. The result is a value between -1 and 1. Thereby -1 represents negative correlation, 0 random correlation and 1 positive correlation [45].

Another distance measure is the Euclidean distance, which determines the distance between two points. The points denote the genes in multidimensional space, whereas each sample represents one axis where each gene has a coordinate according to its gene expression in this sample [6]. The pairwise distances between the observations of the random variables X, Y are then calculated with the following formula:

$$d(X,Y) = |X - Y| = \sqrt{\sum_{i=1}^{n} |X_i - Y_i|^2}$$

2.3.3. Clustering techniques

Clustering approaches can be supervised or unsupervised [6, 46]. They are used to cluster probes with similar expression patterns and to assign dissimilar patterns in different clusters [5]. Supervised clustering algorithms define categories a priori and objects are assigned to

them [46]. Example methods are nearest-neighbor analysis and support vector machines. Unsupervised clustering methods try to find relationships and structure in the data, for example, on the basis of the expression profiles. PCA is an unsupervised clustering technique as well as k-means clustering or hierarchical clustering [6].

For hierarchical clustering, distance measures like Pearson correlation coefficient or Euclidean distance are usually applied. In an iterative procedure high correlated genes (according to their expression measurements) are grouped together until clusters of genes are formed and a dendrogram can visually represent their hierarchical clusters. A dendrogram is like a branching tree, where the leaves are the genes. The longer the branches the higher the dissimilarity is [6, 46]. Here, we used the package 'gplots' and the function *heatmap.2* to cluster the samples according to their distance measurements hierarchically.

2.3.4. Linear regression model

A regression model shows the linear relationship between response variable y and the explanatory variables in X defined as

$$y \sim \beta X + \varepsilon$$

$\boldsymbol{X}$ is a design matrix of $n \times m$ dimensionality with n samples and m covariables. To test gene-wise significance in microarray analysis, y represents the different gene expressions. A lot of different variations choosing the coefficients are possible [12]. $\boldsymbol{\beta}$ is a vector of length m representing the expression coefficient of the factor X corresponding to one biological feature. $\boldsymbol{\varepsilon}$ is called error term and is a vector of length n. It represents all other factors besides the ones indicated in X that influence y. By calculating the linear model the coefficients are corrected for each other.

2.3.5. Hypothesis testing

To determine if groups show relationships 'null hypothesis'-tests can be applied. Therefore the opposite is assumed claiming no relationships between the groups. If the p-value for this test is very low, the null hypothesis can be rejected and the relationships are said to be significant. Otherwise the null hypothesis is assumed to be true [47]. Null hypothesis are statistical tests like significance tests.

An example method to determine a difference between two groups X, Y in a statistical significant way is the t-test [46]. This statistical test tries to identify genes which have a different expression value at different conditions. There is the ordinary t-test which is based on the fold change.

$$t_g = \frac{X_g - Y_g}{\sigma_g}$$

with X_g is the gene expression value of gene g for covariable X and Y_g for covariable Y. σ_g represents the standard deviation of gene g. A similar test is the adjusted t-test (modified t-statistics) where the standard deviation is adjusted.

The moderated t-test is implemented in the R package 'limma' [12]. The expected gene expression of a gene can be defined for pairwise comparison of two subgroups of samples (X, Y) for a gene g. Given a prior distribution, in our case a normal distribution, the standard deviation is calculated. The moderated t-statistic uses a Bayes approach based on the ordinary t-statistic but applying the posterior variance instead of the sample variance [13].

$$\tilde{t}_{jg} = \frac{X_g - Y_g}{\sigma_g \sqrt{v_{jg}}}$$

with v_{jg} is the value of gene g in coefficient j. The p-value is a statistical measure to rate an observation. These rates are compared to a confidence threshold (typically 0.01 or 0.05 are used). If the observed score, for example the t-statistic, has a p-value smaller than the chosen threshold, the observation is said to be statistically significant.

2.3.6. Multiple hypothesis testing

When multiple tests are performed on the same dataset, we drastically overestimate the error rate. If many hypotheses are tested on one dataset, the probability increases that one of them is assumed to be true, even if it is wrong. Therefore a multiple testing correction is needed [47].

This can be done using the Bonferroni correction, which minimizes the family-wise error rate. On the other hand this method could be too strict. A further correction possibility to adjust the p-value is using false discovery rate (FDR). For a set of tests the error rate is controlled by FDR. Therefore it is calculated regarding a collection of different scores [47]. FDR gives the expected proportion of null hypotheses that are rejected and which are false positives. Therefore, a FDR of zero indicates that all null hypotheses are not rejected [46]. FDR can be computed with the Benjamini-Hochberg procedure from the p-values.

2.4. Software

We use the programming language R (version 3.1.2.). Additional packages are downloaded either from the Comprehensive R Archive Network (CRAN, http://cran.r-project.org/) or from the Bioconductor server (http://www.bioconductor.org/).

2.5. Bioinformatics Background

Bioinformatics combines computer science and mathematics to study biological data and is useful especially for big datasets. It is applied in many different fields of biological analyses particularly genetics and genomics. Example fields are sequencing analyses, structural

bioinformatics or system biology. In the following an overview about different bioinformatical tools and approaches is given that are relevant.

2.5.1. Webservers/Databases

Several databases exist to collect, summarize and provide structured information on e.g. genes and their product function. Some common databases which are used later for functional and other analyses are presented here.

Gene ontology (GO) is a database for describing the biological roles of genes and gene products in different species using ontology terms structured in a hierarchical ontology [48]. The database is hierarchically structured with three major ontologies called biological process, molecular function and cellular component. Molecular function refers to the biochemical activity of a gene product, for example 'ligand', 'enzyme' or 'adenylate cyclase'. Biological process is defined as the process the gene product or gene contributes to. Often physical or chemical interactions are involved like 'signal transduction' or 'cAMP biosynthesis'. Cellular component indicates the location of an active cell product in a cell and reflects the cell structure of eukaryotes. The association of an ontology class and a gene product is then defined in the 'GO annotation'. All annotations are evidence supported and the type of evidence can be examined [48, 49]. We used version of September 2014, where a total number of 41,775 GO terms were included in the database as well as 4,185,487 annotations [49]. The webserver can be found at http://geneontology.org/.

The **Kyoto Encyclopedia of Genes and Genomes (KEGG)** is a database established in Japan (http://www.genome.jp/kegg/). Originally KEGG was established to be a reference knowledge base of cellular processes like metabolism. Its goal is to be able to interpret genome sequence data biologically [50]. In principal KEGG is a collection of gene catalogs for mainly completely sequenced genomes which are linked to functions of the cell and organism [51].

The PATHWAY database contains cellular processes like metabolism or cell cycle and is represented in a graphical way. Additionally information about pathway motifs, which are conserved sub-pathways, are included in PATHWAY. This can be used for gene function prediction. Assigning a function in KEGG is a process where a set of genes in a genome is linked with a network of, for example, a complex, a pathway or other molecules in the cell that interact. We used the packages 'mogene20sttranscriptcluster.db' and 'mouse4302.db' for our pathway analyses both based on version of March 2011.

The **Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)** is another public database. This network resource includes both known and predicted interaction data for more than 1100 different species [52]. Known interactions correspond to experimental data as well as to text-mining and transferred-interactions. Thereby about five million proteins

and more than 200 million interactions are stored. Protein interactions often base on the associations of their corresponding genes on the genome. If genes often occur in close proximity to each other, functional interactions tend to be encoded and they may be part of the same metabolic pathway or the same protein complex or others. Physical and functional interactions can be found at STRING, which can be viewed in both the web interface (http://string-db.org/) and in the R package 'STRINGdb' as colored network where each node represents a protein and each edge an interaction with another protein. Evidences for the interactions are given as different scores and combinations with other tools like GO analysis or other databases like KEGG are possible [53, 52].

2.5.2. Normalization of microarray data

Normalization is used to adjust measured intensities from the samples for comparability. For each probeset the relative levels are represented by the normalized data. Different approaches for normalization exist like using the overall brightness of the scanned microarray images, nonlinear techniques or others [6, 54]. Another technique is the robust multi-array average (RMA). For a detailed explanation of the RMA method, see [55]. RMA is based on a log scale model which is linear additive. It corrects perfect match intensities with the background and applies the base-2 logarithm. Then based on a robust average of these corrected and log scaled intensities, RMA estimation is done.

2.5.3. Removing biological biases

If two or more different microarray experiments are performed on the same platform but at different days or even different people, non-biological biases are introduced even for repeated experiments. This can also happen within one experiment when large samples have to be made over some months or even a year [56]. This can be due to several different facts like the environment during sample preparation for example room temperature or like the conditions when the biological samples are stored. In fact, all experimental factors will add some biases that vary between different microarrays [11, 57, 58, 59]. Therefore, batch effects can be defined as all systematic technical differences that occur during processing and measuring in different batches. Furthermore, the differences are not associated to biological variation in the microarray experiment [60]. These differences (batch effects) have to be corrected before the microarrays can be compared to avoid misleading results [56]. Batch effects are typically and easily identified using principal component analysis, visualization techniques or linear models and so on [57].

"Combatting Batch Effects When Combining Batches of Gene Expression Microarray Data", in short *ComBat*, is based on an empirical Bayes (EB) method [11]. This method is part of the Bioconductor package 'sva'. *ComBat* estimates parameters for location and scale adjustment of each batch for each gene independently. There are two methods, one using a parametrical prior by assuming a normal distribution together with an inverse gamma and another using a non-parametrical prior [11, 58, 60]. The distribution of the estimated

parameter is calculated based on the prior. Afterwards the method applies a parametric shrinkage adjustment. Therefore, the data is standardized gene-wise first to get similar mean and variance overall datasets. Standardization is needed because probe sensitivity and mRNA expression level expression values could be different across genes. After that, EB batch effect parameters are estimated using the parametric/non-parametric empirical priors. The procedure estimates the location and scale adjustment parameters of each batch for each single gene independently. Finally the data is adjusted for batch effects, where the batch effects estimated with EB are removed on a similar way like the methods distance weighted discrimination, singular value decomposition and location and scale adjustments [11].

EB methods are very useful as they can handle high-dimensional data like microarray problems very robustly even when sample sizes are small. The estimation uses information across genes and experimental conditions [11, 57, 56]. Beside of the here used EB method, there are several other methods for removing batch effects like simply taking group and time as variables or methods based on singular-value decomposition, distance weighted discrimination and location and scale adjustments. Another method would be the surrogate variable analysis [11, 59, 56]. However, those methods need large batch sizes. The EB method was shown to be more robust as the location and scale batch estimates shrink when combining information across genes [11, 57, 58, 60]. In an evaluation of different batch adjustment methods *ComBat* outperformed five other programs [58].

2.5.4. Functional analyses

After statistically significant enriched genes have been found the question remains about regulated biological functions. Therefore, respective gene products are mapped to their function commonly using GO or KEGG. Like described in 2.5.1 a lot of functional terms exist, captured in ontologies or pathways.

Often the methods are based on gene set enrichment or Fisher’s exact test, which tests for significant enriched genes in each term [61]. Hypergeometric distribution is the basis of Fisher's exact test. A 2x2 contingency table is used and for each configuration the probability for observing it is calculated [62].

However there is a lot of redundancy due to the hierarchical structure of GO which also results in problems at multiple testing corrections as GO terms are dependent on each other [8, 63, 64]. Therefore model-based methods were implemented [64]. Initially such approaches are premised on a combination of penalization and model likelihood. Maximum likelihood approaches maximize the likelihood given a set of parameters and some observed data with respect to the active GO terms set. However, they are not very robust because the calculation ignores alternative solutions only finding a local maximum. Afterwards Bayesian modelling approaches were introduced for optimization [8]. An example approach is the

multi-level ontology analysis (MONA) which uses a Bayesian approach approximating the marginal posteriors of the terms for a set of genes. This is done by applying the expectation propagation algorithm [64]. Another tool is the model based gene set analysis (MGSA) algorithm, which we used in this thesis (3.5). MGSA employs a robust Markov Chain Monte Carlo approach. It estimates marginal posterior probabilities, which determine if categories are active. Therefore the Metropolis-Hastings algorithm was implemented. MGSA makes a gene-category analysis leading to a reduction of the number of redundant categories.

In contrast to Fisher's exact test, which only finds enrichments in each term, MGSA is a global approach. It tries to find a combination of GO terms which can explain the observed biological response in the best way. MGSA approaches tests all terms/categories simultaneously in one test, thus there is no need for multiple testing corrections. MGSA embeds a Bayesian network where a function of biological categories activation is modeled representing the gene response. It can be applied with GO, KEGG and with every other ontology list, too [8, 63].

2.5.5. Network smoothing

Related methods

Expander is a tool for microarray analysis, integrates a graph-theoretic algorithm called "Module Analysis via Topology of Interactions and Similarity Sets" (MATISSE) which looks for co-expressed subnetworks that are connected significantly [5]. Another method is the network based stratification (NBS) approach, which integrates somatic tumor genomes using the mutation profiles with gene interaction networks. Therefore a binary state profile (1,0) represents the mutations, where 1 indicates a gene with a mutation. The profile is smoothed with a network so instead of binary states the proximity to the mutated gene is shown. Therefore a random walk on the given network is simulated. The result is then clustered and the patients can be stratified into subtypes [9].

Network smoothing using stSVM

A third tool called "smoothed t-statistic SVM" (stSVM), implemented in the 'netClass' package, uses a random walk kernel to smooth t-statistic of genes along a network [10]. For network information, originally the PathwayCommon Database and the KEGG database as well as the MicroCosm database is utilized. The *stSVM* algorithm starts with an undirected graph $G = (V, E)$. A is the adjacency matrix of this graph G and D is the diagonal matrix of node degrees with $D = diag\big(dev(v_1), dev(v_2), \dots dev(v_n)\big)$ for node $v_1, \dots, v_n$. Here nodes are genes and edges are interactions like protein-protein interactions between them. Furthermore, the Laplacian matrix is defined as $L = A - D$. A normalized version of L is

$$L^{norm} \coloneqq D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}.$$

With this formula the p-step random walk kernel K used in *stSVM* is defined as

$$K = \alpha I - L^{norm} = ((\alpha - 1)I + D^{-\frac{1}{2}}AD^{-\frac{1}{2}})^p$$

α is a constant, whereas p represents the number of random walks that are used [65].

A random walk is a finite Markov chain. For a given node in a graph a random neighbor is selected to which it is moved. Then for this node a random neighbor is chosen, too. The resulting sequence of nodes and movements between them is then the random walk. Assuming to start in root vertex r the probability can be computed that a random walk ends in node e with a given length for the random walk.

To finally smooth the t-statistic (t-statistics for each gene are considered as vector t) of a paired t-test, the t-statistic is multiplied with the kernel:

$$\tilde{t} = t^T * K \rightarrow \tilde{t}_i = \sum_{j=1}^{n} t_j K_{ij}$$

By applying this scalar multiplication the t-statistic is smoothed with the network information denoted in the kernel. To decide the statistical significance of the smoothed t-statistic a permutation test is applied. Therefore the data is permuted to design some pseudo-datasets based on the original. Afterwards the original observed values are compared to the pseudo-values and the deviation is tested to decide if a value is significant [46].

3. Materials and Methods

In this section we provide a description of the dataset provided by the Götz lab (Institute of stem cell research, HMGU) and a detailed description of the implemented methods used to understand molecular mechanisms of astrocytes.

The provided datasets spanned multiple profiling projects with individual aims, but all centered on the question to understand molecular mechanisms of astrocyte regulation. We initially had to understand and review the data in order to design our analysis strategy accordingly. For example, we had cells from adult mice that were taken from injured brain regions (lesions), where both astrocytes and cells, that were not astrocytes, were sampled. Therefore, many features must be considered for astrocytes like being reactive astrocytes, being taken from lesion and being cells from an adult line and not from a cell culture generated with cells from young mice. Additionally very complicated was that datasets were measured on two microarray platforms. This is hardly feasible to integrate on probeset, or gene level [66]. We decided to analyze two "datasets" separately and then only compared them functionally.

As genes function in clusters, we tried to improve result interpretation by smoothing the t-statistics with gene network information. We used the t-statistics of the linear regression and smoothed it given protein-protein interactions as prior network. As the regression was a very complex setup, we proposed ideas for "regression-smoothing" considering the nested covariables to include dependencies between them.

3.1. Transcriptome datasets

In this section we provide a description of the five projects, which were for the first time jointly analyzed. The datasets consist of mRNA expression profiles of microarray experiments. In here we give an overview over the samples and the different groups of cells. The data represented in the following consists of gene expression profiles which were done on two different platforms of Affymetrix Microarrays. Fig. 8 is a sketch summarizing all datasets and experimental conditions. We grouped the five project datasets into two datasets of different data source given by their microarray platform.

3.1.1. Growth factor stimulation of astrocytes

The growth-factor-dataset was performed on microarray platform "`MOUSE GENE 2.0 st`" and contained probes of astrocytes treated with FGF (SAF) and probes treated with both EGF and FGF (SAFE). Third there is a subset including probes of neurospheres (SN). Astrocytes treated with growth factors should show changes in the activity of astrocytes that might be similar to reactive astrocytes or neural stem cells. It is reported that growth factors can induce astrocytes to proliferate [42].

3.1.2. Combining various cell types

Further microarray experiments were performed on the Affymetrix microarray platform "MOUSE GENOME 430 2.0" including four different batches. Those batches are given by project-based datasets.

Cells of **post-natal six (P6)** mice that were cultured are part of one project dataset. Samples were taken after 4 hours, 24 hours and 48 hours after culturing as well as after six days which is called 24 hours delayed (Fig. 5).

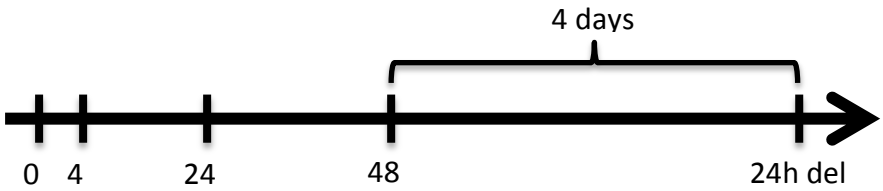


Fig. 5: Timepoints of taking probes from P6 mice cell cultures.
Probes were taken after 4 hours, 24 hours and 48 hours. Furthermore probes where taken 6 days after time point 0 which is called 24 hours delayed.

Fig. 6 shows microscopy images of astrocytes at two different time points, 24 hours and 24 hours delayed.

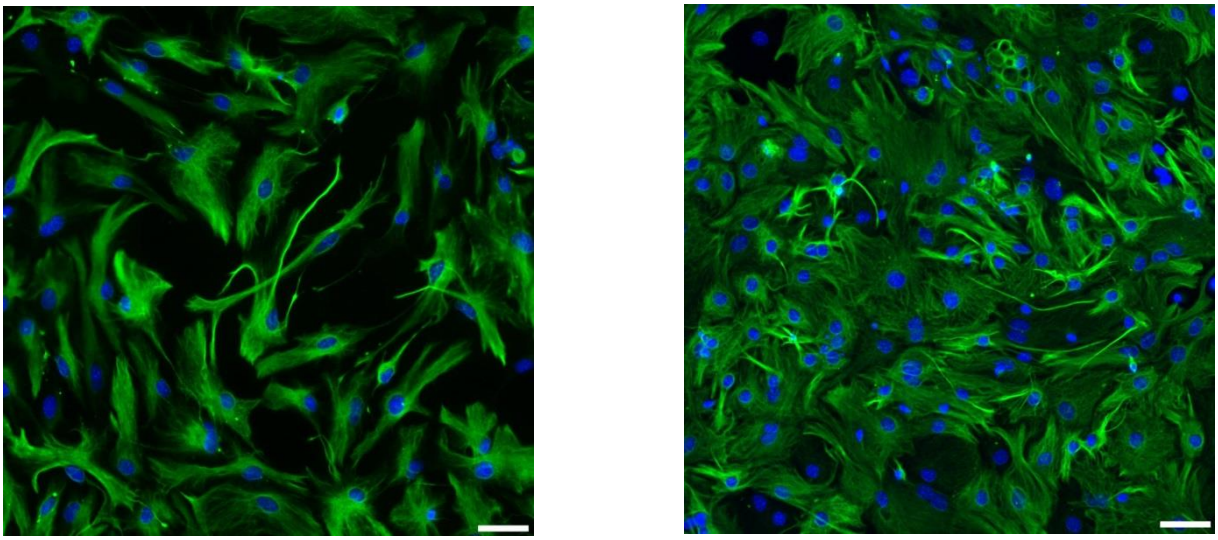


Fig. 6: Pictures of astrocytes provided by the Götz Lab.
Left picture is taken after 24 hours and right after 24 hours delayed.
Green coloring: GFAP (glial fibrillary acid protein, an astrocyte marker)
Blue coloring: DAPI (a marker for nuclei)
Scale bar: The white bar at the bottom right in each picture corresponding to 50µm in the picture

The next two project-derived datasets are based on cells of **adult mice**:

i) One batch consists of samples of diencephalic astrocytes that were isolated from hGFAP:GFP mice and of neural stem cells isolated from the SVZ.

ii) The other batch was divided in three subgroups. First it included astrocytes which were taken from healthy tissue of Aldh1l1:GFP mice. Therefore these cells are non-reactive and represent the wildtype form. Furthermore there were cells isolated from regions of focal lesion of hGFAP:GFP mice. Thereby GFP-plus cells represent reactive astrocytes, where GFP-minus represents cells from injured regions that are negative for GFP. Therefore they are identified as being non astrocytic.

Finally, the last project dataset consists of **cultured embryonic stem cells (ESCs)**. There are

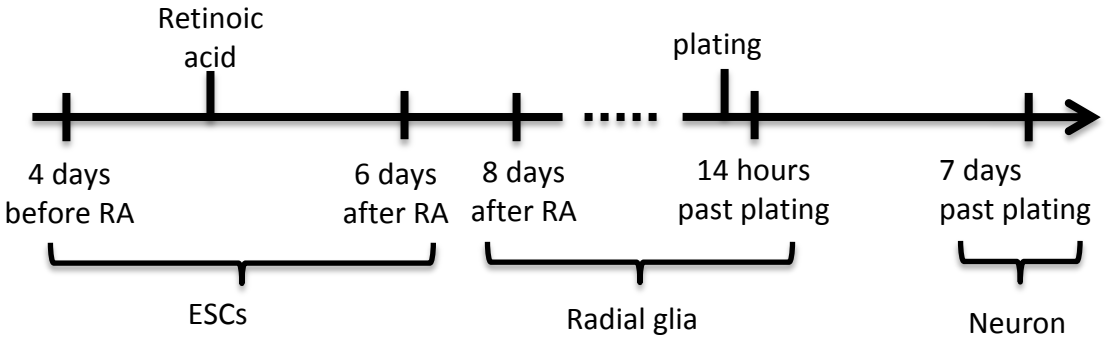


Fig. 7: Overview of probes from the embryonic stem cell culture.

cellular aggregates (CA) four days before retinoic acid (RA) was added and six days after the first add of RA. Both are ESCs. Samples of CAs eight days after adding RA can be considered differentiated radial glial cells. The forth group was taken fourteen hours past plating of the radial glial cells. After seven days past plating those cells differentiated into neurons. Fig. 7 shows a schematic overview of the ESC cultures.

Besides of the NSC of the SVZ and astrocytes of the diencephalon (DIEC) all cells were extracted from the cortex.

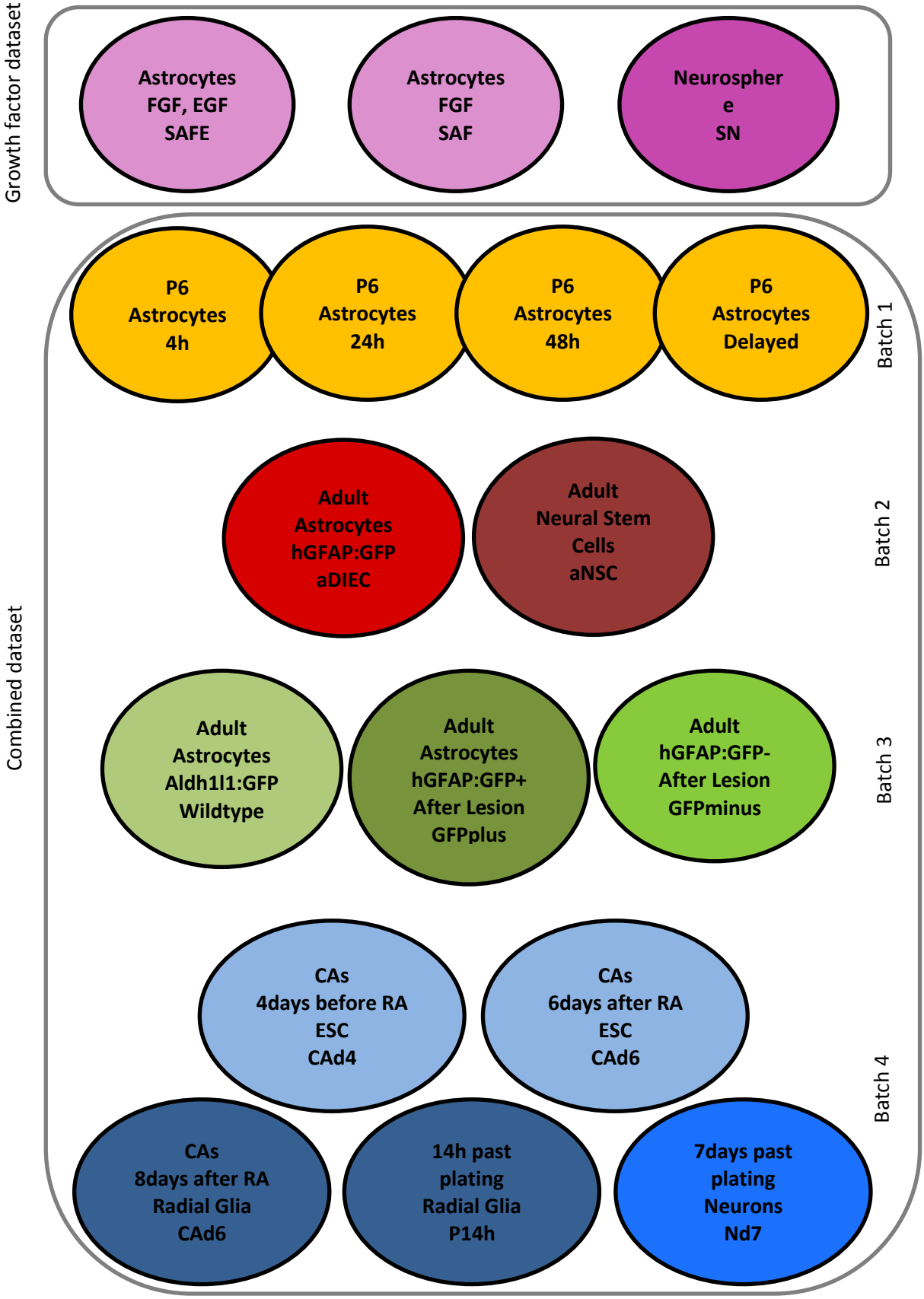


Fig. 8: Graphical representation of the datasets.

The upper three subsets that are encircled belong to the growth factor dataset. The second circle includes all datasets of the combined set. Different colors represent different batches and the color grade indicates known differences in a batch like being ESCs, radial glia or neurons for the five subsets at the bottom.

3.2. Normalization of the microarray expression profiles

For the growth factor dataset we used packages ‘oligo’ and ‘pd.mogene.2.0.st’ for reading raw files and the package ‘mogene20sttranscriptcluster.db’ to obtain probeset annotations. For the combined-dataset we combined raw files before normalization. We added the ESC cell culture samples to separate **‘P6’** and **‘adult’** from cell culture and direct line as the ESC culture consisted of cell culture probes, too, but the cells were not from P6 mice. In the following probes starting with “p” represent the first batch, “a1” the second batch (DIEC and NSC), “a2” the third (GFPminus, GFPplus and Wildtype) and “e” the forth batch belonging to the ESC line. For the “combined”-dataset the packages ‘affy’ and ‘mouse4302.db’ were used.

We utilized the function *rma* (Robust Multi-array Average) to normalize both datasets. The quality control of the microarray expression profiles was performed for the combined dataset in detail. We used different quality control methods, like the density plot or RNA degradation for both the raw and the normalized data. We identified differences in the quality between the two cell culture lines and the two direct lines before normalization (Fig. 9). The RNA degradations from 5’-3’ were different for adult lines than when compared with P6 or ESC lines. Especially two samples, *a2GFPminus_11* and *a2GFPminus_12,* with high 3’ degradation rate indicate a less quality for the cells taken directly from adult mice than for the samples derived from the two cell culture lines.

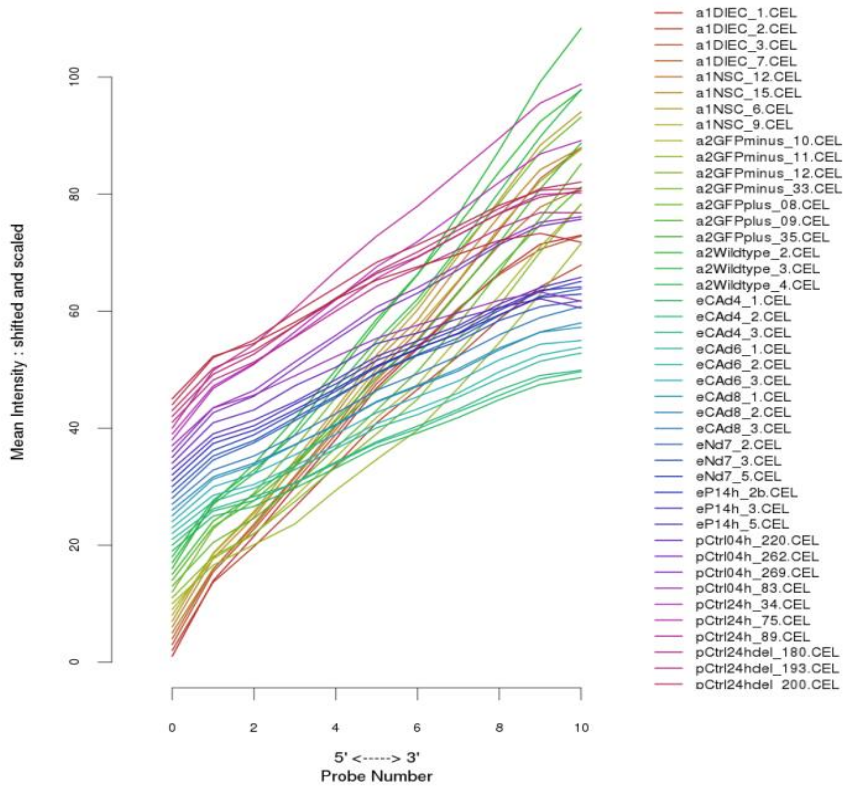


Fig. 9: RNA degradation of the 45 samples before normalization showing the mean intensity.
RNA degradations from 5’-3’ are different for adult direct lines compared to the P6 or ESC cell culture lines.

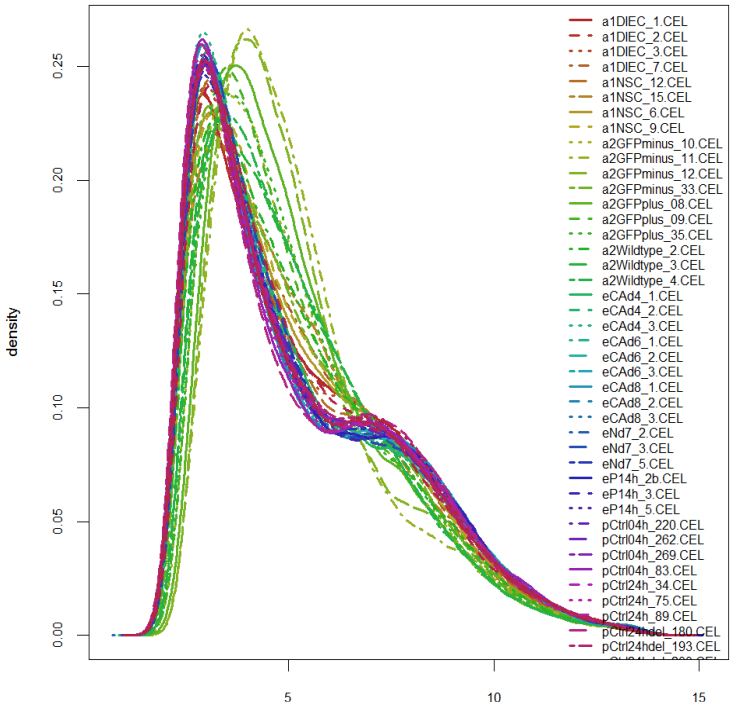


Fig. 10: Density Histogram for each of the 45 samples after normalization.
Differences between the quality of batch 1 and batch 4 to batch 2 and batch 3 can be viewed.

After normalization we still observed differences between those two groups (Fig. 10). From the distribution of the density histogram we identified some diverse distributions especially for the samples of **'a2'**. Above all the samples *a2GFPminus_11* and *a2GFPminus_12* showed a distinction to the other samples. The first peak of the densities was shifted to the right but the second small peak was not visible.

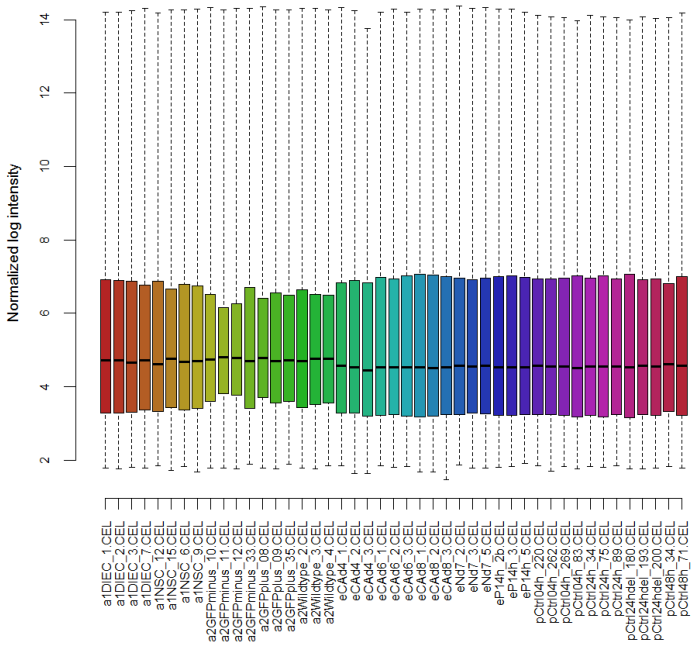


Fig. 11: Boxplot of the normalized log intensities for each of the 45 sample after normalization.

Furthermore, we produced boxplots of the normalized log intensities for each sample shown in Fig. 11. In general the normalization step results in comparable mean and standard deviations of the fold changes. The samples of **'a2'** differed a lot in the boxplots indicating less quality. Again especially the two same samples as in the density histogram are different to the others. As these two arrays had low quality in other quality tests, too, we removed them for further analysis. Therefore 43 samples remained for the combined-dataset.

3.3. Removing non-biological bias

All probesets included for experimental reasons were removed for the expression analysis from the combined-dataset. We aimed to remove batch effects to make the different sets comparable. Therefore, we used the empirical Bayes method *ComBat* [11]. To relay the program which batches were included in the dataset, we generated a sample info file, indicating which samples belong to which of the four project datasets. As we tried to find and remove technical noise in the dataset, we had to be careful not to remove real biological information. For the biological background information we designed a matrix indicating the biological cofactors of the samples. Rows showed all, in this case 43, samples, the columns cofactors. The cofactors represented the biological factors in the dataset that should be tried to remain when removing the batch effects. For *ComBat* we chose the different cell types, which were included in the dataset, as cofactors. Thereby, "1" labeled every member of a cell type in the corresponding column of the matrix and "0" otherwise. Table 1 provides an example extract of the design matrix. The complete matrix is shown in the appendix (Table 10).

	a1DIEC	a1NSC	a2GFPm	a2GFPp	a2WT	…
a1DIEC_1	1	0	0	0	0	…
a1DIEC_2	1	0	0	0	0	…
a1DIEC_3	0	0	0	0	0	…
a1DIEC_7	1	0	0	0	0	…
a1NSC_6	0	1	0	0	0	…
…	…	…	…	…	…	…

Table 1: Example extract of a design matrix generated for ComBat.
Rows show the samples, where columns represent the used cofactors. 1 indicates that a sample belongs to a cofactor otherwise this is denoted with 0. Zero lines (i.e. for a1DIEC_3) were made for computational reasons.

There had to be at least one "zero line" for each batch like for row a1DIEC_3. Otherwise using *ComBat* was not possible as the cofactors and batches were computationally singular. To choose a design matrix we performed a robustness test of *ComBat*. Therefore, we built 50 different design matrices using 20% random "zero lines". Afterwards we ran *ComBat* with each random design matrix to remove the batch effects on the combined-dataset. Results are provided in section 4.2.1. We chose one of the matrices which showed a high similarity in robustness to the others. This matrix contained eight "zero lines" and was used to correct the combined-dataset for further analyses

3.4. Statistical analysis

To perform a principal component analysis (PCA) in R we used the function *prcomp*. We applied PCA to structure and visualize the expression profiles. For each principle component we used standard deviations to assess the percentage of variance explained in the data. We used PCA to examine the robustness of *ComBat*. For each iteration using the random design matrices we calculated the PCA and compared the distributions of PC1 for the design matrices over the samples. In addition we applied PCA to visualize the clustering of the samples as well as the genes of the datasets with the first two principle components.

To perform a hierarchical clustering of the samples we measured similarity by calculating pairwise Pearson correlation coefficients as well as Euclidean distances of respective gene expression measurements. Samples for which gene expressions were more similar than to others clustered together hierarchically. We visualized these clusters representing the similarities and dissimilarities with heatmaps using the package 'ggplot2'. In a heatmap we depicted both the dissimilarity measurements and the dendrogram representing the clusters. Resulting clusters indicated possible similarities between the cell groups.

Relationships between the biological features of the datasets were searched. Therefore, we utilized a linear regression model using the 'limma' package (linear models for microarray data) [7]. To model gene expression, we needed a design matrix that indicated biological covariates and associated samples. For the growth-factor-dataset this design matrix was generated using the three included cell types themselves. So we segmented the design matrix into **'SAF'**, **'SAFE'** and **'SN'**. To differentiate influences of FGF and EGF we created a second design matrix for another linear regression model including these two coefficients for the samples.

For the combined-dataset we had to overcome a couple of difficulties to generate a design matrix. Many biological features were included in this dataset and pairwise comparisons were not possible. We inspected the most meaningful features and designed many different matrices by hand. Example features would be **'adult'**, **'ESC'** or **'p4h'**. Another issue in finding the right matrix was the hierarchical structure of the biological features. All samples of the first batch, for example, were cells from post-natal six (P6) cell cultures. We wanted to include both the factor **'P6'** as well as the different time points P6 after four hours (p4h), twenty-four hours (p24h), forty eight hours (p48h) and after six days (pDel). All time-points together represented **'P6'** once again and at least one of the five features had to be removed for computing the model. After some considerations we excluded **'p24h'** indicating the baseline level of **'P6'** samples. Finally we used fourteen covariates with nested structure (see Table 2 in results chapter (4.2.2). Additional to the design matrix with the biological factors, we constructed a contrast matrix to answer specific biological questions. The contrast matrix defined which comparisons between the samples or factors had to be performed. In this case, the matrix included the features of the design matrix themselves. For the combined set

we also added **'Astrocytes'** as coefficient, which is a combination of all samples belonging to the coefficients **'P6'**, **'aDIEC'**, **'aGFPp'** and **'aWT'**.

We used the function *lmFit* for generating the linear regression model in addition with the factors of the contrast matrix. Finally, we added empirical Bayes parameters to the linear regression and built scatterplots of the pairwise coefficients. Additionally, we utilized the function *topTable* which summarized the results of the linear model and performed hypothesis tests like computing the moderated t-statistic for each gene and each contrast [12]. For each coefficient of the contrast matrix we got a list of genes, which were expressed differentially. The t-statistic of a gene of one coefficient represented if the gene expression of this gene was up-regulated or down-regulated in comparison to all other features. A corresponding p-value showed if the t-statistic of the gene was significant. As multiple comparisons existed, we needed a multiple testing correction and adjusted the p-values with the false discovery rate (FDR). We defined all genes, which had a t-statistic with a p-value smaller than 0.01, as significant and considered only these genes for further functional analysis. We used some graphical representations of the fitted results like scatterplots and volcano plots. Furthermore, we plotted the different numbers of significant genes for the coefficients.

3.5. Functional analysis

To find specific functions for the gene lists suggesting characteristics of the biological features, we mapped the genes to their biological functions using the databases GO and KEGG (2.5.1). Therefore we used the package 'mgsa' [63]. Initially, we downloaded GO annotations from the gene ontology website (http://geneontology.org/page/download-annotations). Those annotations were stored in a "Gene Annotation Format" (GAF). We used the *readGAF* function to read the downloaded data of mouse GO-annotation. With this we could keep the hierarchical structure of the GO database. We computed functional enrichments of the gene lists for biological processes only. For KEGG pathway annotation we simply generated a list with known pathway-annotations of mice. Then we ran *mgsa* on each gene list for both annotations. An active function in a gene set is an annotation which is overrepresented for the genes in the list. For each function an "estimated" value was calculated and we only chose terms as significant with an "estimated" value of at least 0.5. The result was a list for each coefficient with significant enriched GO or KEGG annotations showing their possible characteristics.

The results of the functional analysis were discussed for both datasets separately (section 4.1.2 and 0). Additionally, we performed a comparison of the results to show relationships and differences between the features of the two sets (chapter 4.3).

3.6. Network smoothed t-statistics

A linear regression model is a statistical way to relations different coefficients (strength of covariate influence estimated using regression coefficients) and response variables, in our case the biological features and gene expression, respectively. A gene expression is typically modeled independently of interacting or neighboring genes. Nevertheless genes and proteins interact with each other and usually function in pathways or complexes. Thereby network information like protein-protein interaction networks can improve the analysis. Network modules can be identified, if the expression data is combined with a network, like combining similar expression patterns with connected subnetworks. Therefore the neighborhood of genes in a network is accounted for. Furthermore, results of such a combined analysis are more reliable as there is a higher probability that the function of the genes in a module is linked somehow.

To enhance the statistical analysis on single-gene level with biological knowledge, we added network information of protein-protein-interactions to the t-statistic of the linear regression model based on the *stSVM* approach [10]. We downloaded the network information for mice, which have the species ID 10090, with the 'STRINGdb' package using the STRING version 9.1. Interactions in STRING are scored and we only admitted interactions with a score of at least 400. For using the 'STRINGdb' package the probeset IDs had to be mapped to the STRING IDs. For some probeset IDs and gene symbols no corresponding STRING ID existed and we excluded them from the dataset. Furthermore, some probeset IDs referred to the same gene. In such a case we combined the expression values of the probesets by taking the expression value with the highest variance resulting in one expression per gene. For remaining genes a subnetwork of the mice network was built, where nodes represented the genes and edges the interactions between them.

Then we generated an adjacency matrix of the network. All nodes which had no edges to any other node were removed from the network. Finally, we used the 'netClass' package [65] to calculate the p-step random walk kernel (2.5.5). We determined the kernel with the *calc.diffusionKernelp* function using one iteration and the constant value α=1. The original implementation performed a paired t-test resulting in one single vector of t-statistics for smoothing ($\tilde{t} = t^T * K$), but we applied the linear regression model again. Thereby, we only used genes included in the kernel. We extended the original *stSVM* implementation for our t-statistics for different coefficients. So instead of a vector with one t-statistic per gene, we created a vector of t-statistics for each coefficient separately using linear regression. The matrix T ($n \times p$, with n genes and p coefficients) combined the resulting t-statistics of the different coefficients. We divided the t-statistics of each coefficient by its maximum and calculated the smoothed t-statistic by:

$$\tilde{T} = T^T * K$$

Afterwards, we divided $\tilde{T}$ again by the maximum per coefficient like in the original implementation of *stSVM*.

To decide which of the new t-statistics were significant we calculated p-values using a permutation test, similar to the original implementation of *stSVM*. For each permutation we had to run the linear regression anew by randomizing the samples of the design matrix and smoothed the resulting T_{perm} afterwards, too. This we performed 1000 times. Then we compared the resulting t-statistics $\tilde{T}_{perm}$ with the observed smoothed t-statistics $\tilde{T}$. For each gene entry we tested if $\tilde{T}_{perm}$ was higher than $\tilde{T}$ given an error rate of 5%.

We compared the total number of significant genes before and after smoothing. Chapter 4.2.3 contains our results of a mini-example with seven genes and four coefficients as well as the results of the complete dataset.

3.7. Methods proposal for regression smoothing

So far we used gene network information to improve the regression only across genes within one coefficient. Coefficients were treated independently for smoothing but in our expression analyses many biological features were included in the model, which were partly nested. Therefore, we considered smoothing the t-statistics over the coefficients, too. As not all factors depended on each other combining all coefficients in a single vector for smoothing did not make sense. Instead we thought that only the dependencies between the coefficients must be regarded. For example coefficients like **'P6'** and **'Adult'** were independent, but **'p4h'** or **'p48h'** both depended on **'P6'**. As the linear regression model assumed linear independency of the coefficients, we tried to take this separation with smoothing into account by using dependent coefficients. We performed different attempts to add information across coefficients. This we ran for a small set of depending coefficients using 'P6' and its sub-coefficients **'p4h'**, **'p48h'** and **'pDel'** and a small number of genes.

Aggregated smoothing

For the first two attempts we tried summarizing the t-statistic vectors of the linear regression to combine the dependent coefficients. Vector addition is commonly used for a linear combination of vectors. In the first approach the t-statistic was smoothed with the network like before. The dependent coefficients were combined by taking the smoothed t-statistic vector of the coefficient, which included multiple sub-coefficients, and summed it with the observed, but not smoothed, t-statistic vector of the sub-coefficient. For example, the smoothed t-statistic of coefficient **'p4h'** would be calculated in the following way:

$$\tilde{t}_{P6} = {t_{p6}}^{T} * K$$

$$\tilde{t}_{p4h} = t_{p4h} + \tilde{t}_{P6}$$

However, with this approach all coefficients depending on the same factor were altered in the same way. The network information was not used on all factors, but only on the dependent one. So the "smoothed" t-statistic of **'p4h'** was not smoothed with the kernel itself, but calculated by adding the unsmoothed t-statistic of **'p4h'** with the smoothed t-statistic of **'P6'**.

In the second approach we again summed the dependent t-statistics, but this time we performed this before smoothing with the network. Therefore, we used absolute t-statistic values. This time we tried to treat the two dependent factors equally in contrast to the first approach. For example, the new t-statistic for the coefficient **'p4h'** would be:

$$t_{p4h\_new} = abs(t_{p4h}) + abs(t_{P6})$$

The same was performed on the other coefficients. For **'P6'** the t-statistics of **'p4h'**, **'p48h'**, **'pDel'** and **'P6'** were summed. Then the new t-statistic vectors all were combined in a matrix T_{agg2} and were smoothed with the kernel.

$$\tilde{T}_{agg2} = {T_{agg2}}^{T} * K$$

Using aggregated smoothing the kernel remained like before and vector addition was performed to combine t-statistics of dependent coefficients.

Regression smoothing with extended network

We now tried to use the t-statistic of one gene for a coefficient β_1 and include the t-statistic of the same gene, but for a dependent coefficient β_2. To accomplish using the dependent coefficient for smoothing as well, we can readily extend the original network by adding a second set of nodes, duplicating the network. To illustrate the idea, imagine a network with three genes A, B and C with two dependent coefficients β_1 and β_2. Now assume that we wanted to smooth the t-statistic of β_1 for gene A, which is node A^1 in the network. As β_1 and β_2 were dependent factors we used the information about the t-statistic of β_2 for the same gene (A^2), too. Then we used all nodes A^2, B and C to smooth A^1. Thereby, we enlarged the t-statistic for β_1 by concatenating it with the t-statistic of β_2. Additionally this gene of β_2, A^2, was also included in the network as node interacting with A^1. The original t-statistic matrix was extended for all coefficients to duplicated length for each coefficient. The proposed approach extends the protein-protein interaction network and added a node for one dependent t-statistic.

For each gene of β_1 the corresponding t-statistic of β_2 was concatenated, so the t-statistic of β_1 had double length than before as each gene got a second t-statistic value.

$$t_{new\_\beta_1} = c(t_{\beta_1}, t_{\beta_2})$$

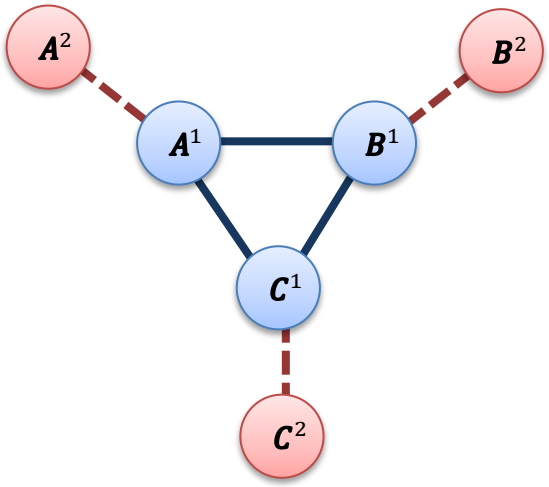


Fig. 12: Illustration of extending a network for regression smoothing.
A, B and C are genes, where 1 indicates the original network for one coefficient and 2 the extension for the dependent coefficient, duplicating the nodes. For example, $\boldsymbol{A^1}$ got a new neighbor $\boldsymbol{A^2}$ to include the t-statistic of the dependent coefficient.

Additionally, we extended the network. For each gene we added a new node which linked gene A of coefficient β_1 to gene A of coefficient β_2. This was performed for all nodes like shown in Fig. 12.
Therefore, the network increased to double size as each gene was included twice. Then we generated the adjacency matrix of the extended network by using the original adjacency matrix and added the duplicated nodes to it. Thereby, an interaction was included for A^1 to A^2 and so on. Then we calculated the kernel for the extended adjacency matrix. With this kernel we finally smoothed the concatenated t-statistics.

$$\tilde{t}_{\beta_1} = t_{new\_\beta_1} * \mathrm{K}_{extended}$$

Note that we now included information of coefficient β_2, when computing smoothing of β_1, since we had the $A^1 - A^2$ gene-gene "interaction" in the network.

Again two approaches were tested on a mini-example. Thereby, we extended **'P6'** with a vector of zeros as it depended on more than one other coefficient and would need an extension with the t-statistics of **'p4h'**, **'p48h'** and **'pDel'** and therefore three additional nodes per gene. First, the t-statistics were concatenated before dividing by the maximum per coefficient and the other time the divided t-statistics were concatenated with each other before multiplying t-statistic and kernel. The first method seemed to depend more on the dividing step. If the dependent t-statistic included the highest value, the t-statistics of the other coefficient, for example **'p4h'**, got lower values than dividing it with the highest value of itself. Therefore, it might be that t-statistic values of **'p4h'** got less important than in the second approach.

4. Results

The first part shows the result for the growth factor dataset including statistical and functional analysis. In the second part we describe how the combined set is changed when removing batch effects and analyze the resulting expression profile statistically and functionally. In the third part, we show the results of network smoothing on the combined set. In the fourth part we investigate the novel approach for regression smoothing on a mini-network. In the final part of this section we compare the results of the functional analyses of both datasets with each other.

4.1. Growth factor dataset

For the growth-factor-dataset we performed statistical and functional analysis to test how similar astrocytes treated with different growth factors are in comparison to neural stem cells (NSC).

4.1.1. Statistical analyses

The principal component analysis (PCA) showed a clear separation between astrocyte cells and cells from the neurosphere in the first principal component (Fig. 13). For the second principal component no clear division of the data was visible. The plot shows a clear clustering of **'SAF'** and **'SAFE'** indicating that adding EGF had no influence on the gene expression.

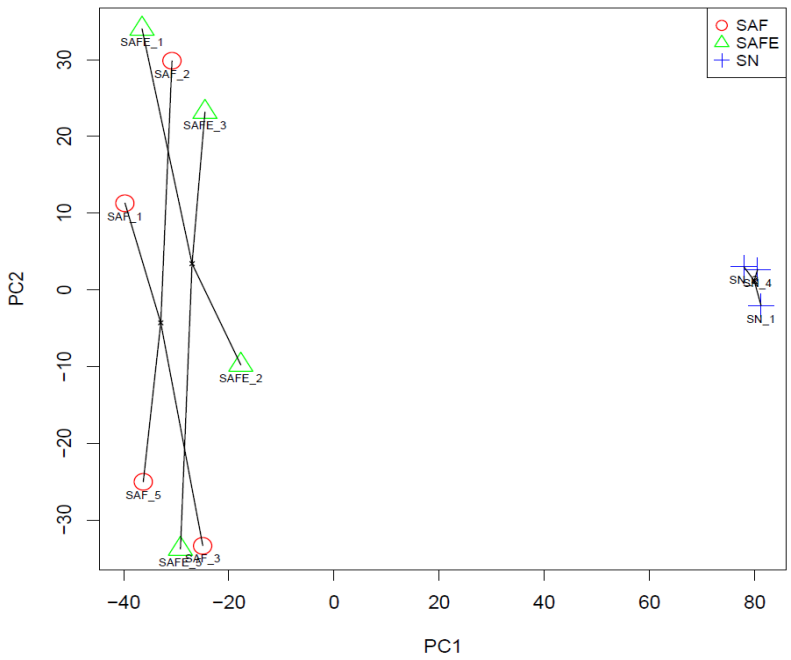


Fig. 13: The first two principal components for the samples are plotted.
Separation into astrocytes and NSCs can be viewed.

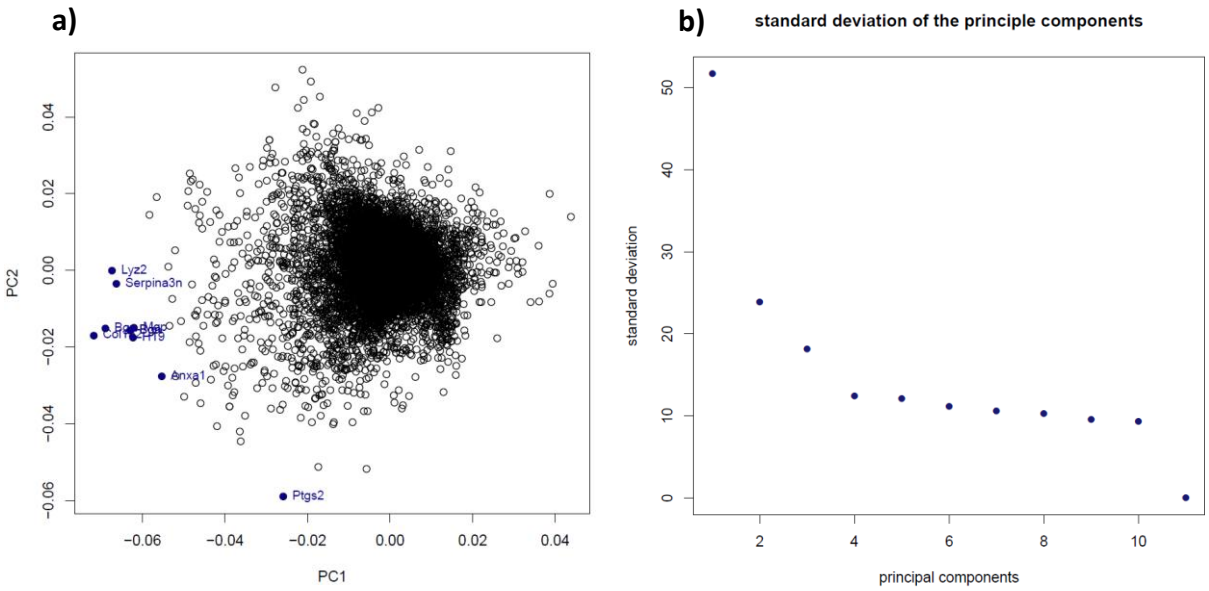


Fig. 14: PCA rotation matrix of genes and the standard deviations of the principle components.
The left plot shows the first two principal components for the genes. Top 10 genes that have largest influence in the respective component are marked.
The right plot shows the standard deviations of the principal components. The y-axis represents how much of the data is explained by the appropriate principal component**.**

Fig. 14 b) shows that only the first principle component divided the dataset like it was already detected in Fig. 13. The standard deviation of the principle components suggested that only the first principle component described more than 50% of the variance of the data. The separation into astrocytes and NSC appeared to be the main expressional difference.

The PCA of the genes (Fig. 14 a)) showed ten genes marked as relevant genes. Those were Col1a2, three probesets referring to the gene Bgn, Lyz2, Serpina3n, Ptgs2, H19, Mgp and Anxa1. Bgn for example is the short form of biglycan and is a small proteoglycan containing two glycosaminoglycan chains.

We analyzed sample similarity using hierarchical clustering of global gene expression using Euclidean distances (Fig. 15). We identified the same division into NSC and astrocytes as for the PCA. Again the method could not separate **'SAF'** and **'SAFE'**.

This indicated that EGF did not change the characteristic of the astrocytes much if at all. Furthermore NSCs from the neurospheres seemed to have a high difference to astrocytes.

We next analyzed gene expression using linear regression modelling. For the coefficient **'SN'** no significant genes were detected. For **'SAF'** and **'SAFE'** more than 3000 genes had a significant p-value for their t-statistic, which also were similar (Fig. 16).

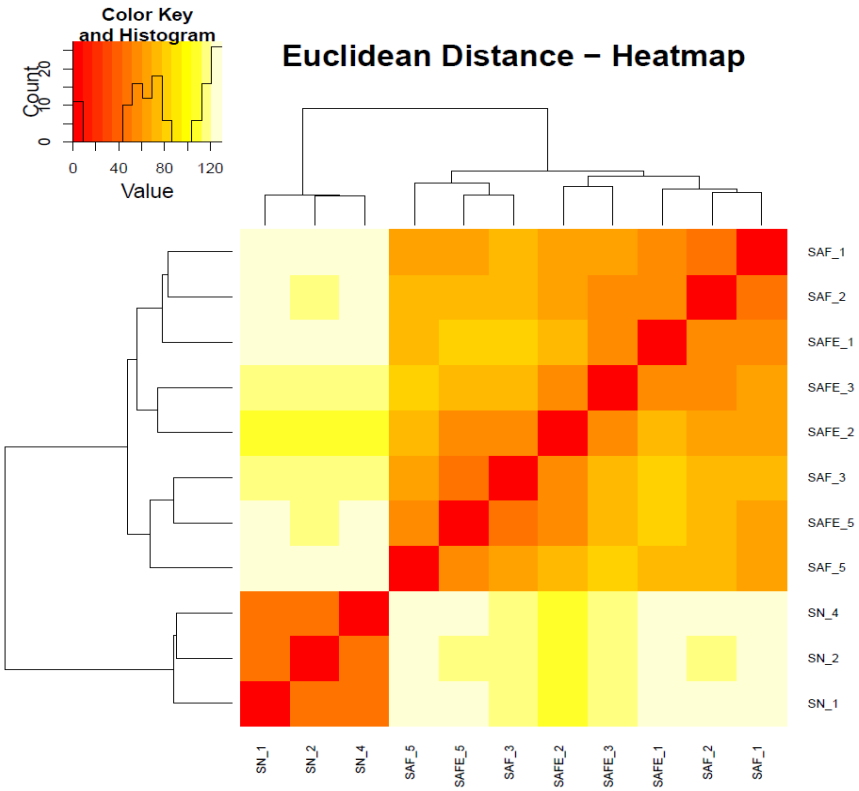


Fig. 15: Global sample clustering using Euclidean distances.
'SAF' and 'SAFE' samples are clustered.

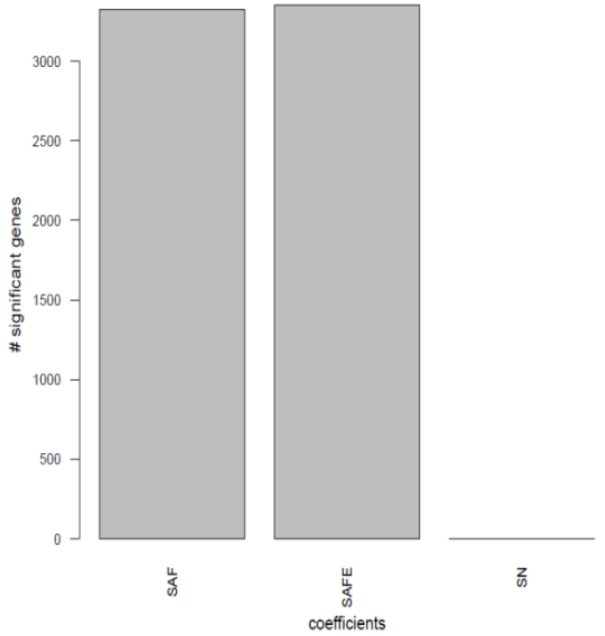


Fig. 16: Number of genes which t-statistics are statistical significant with a p-value <0.01 for the coefficients 'SAF', 'SAFE' and 'SN'.
No genes were found for 'SN' and more than 3000 for the other two coefficients.

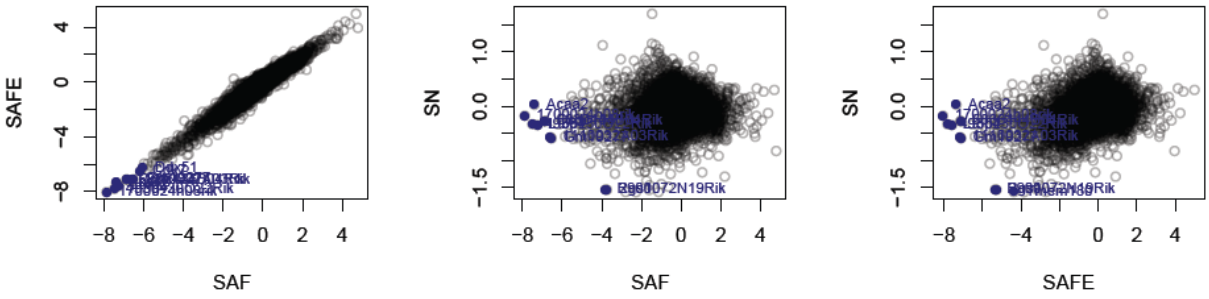


Fig. 17: Gene regulation comparisons of the coefficients 'SAF', 'SAFE' and 'SN'.
Only for 'SAF' against 'SAFE' gene regulations were correlated.

To investigate the correlation between sample-specific gene regulation, we analyzed three coefficients using pair-wise scatterplots (Fig. 17). We could not detect any correlation between **'SAFE'** and **'SN'**. The same applied to **'SAF'** and **'SN'**. **'SAF'** and **'SAFE'**, however, again had same genes being regulated. Nearly all genes seemed to be expressed the same way. In addition the volcano plot of **'SAF'** and **'SAFE'** showed similar fold changes over the logs odds (Fig. 18). Additionally we observed a high overlap between those two coefficients indicated in the Venn diagram sharing more than 2700 genes. Each of them had about 500 genes only significant for the one coefficient.

Altogether this dataset showed a high similarity between the coefficients **'SAF'** and **'SAFE'** meaning that FGF is a predominantly influencing gene regulation and EGF only a bit. However, **'SN'** differed from those. Therefore astrocytes treated with growth factors showed hardly, if any neural stem cell characteristic in all statistical analyses.

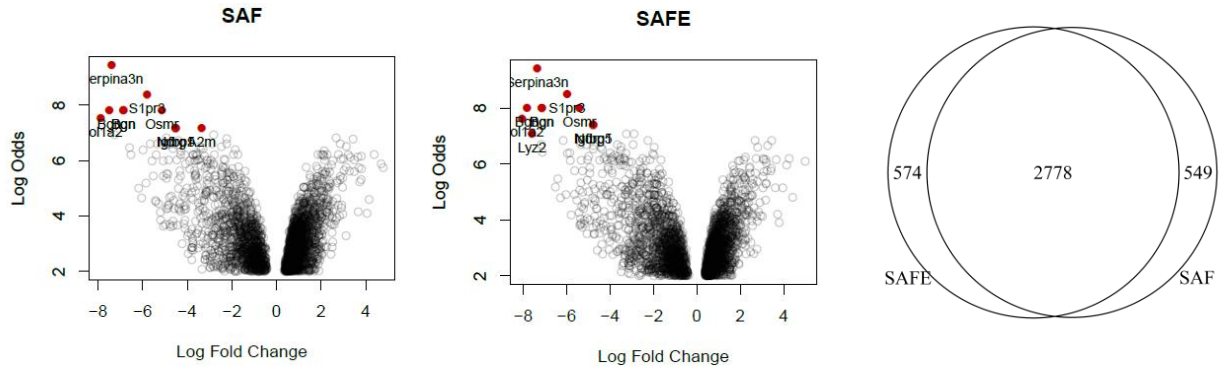


Fig. 18: Volcanoplots for 'SAF' and 'SAFE' and comparison of the significant genes of these two coefficients.
Ten genes with highest t-statistic values are marked in the volcano plots. A high overlap between the significant genes can be viewed.

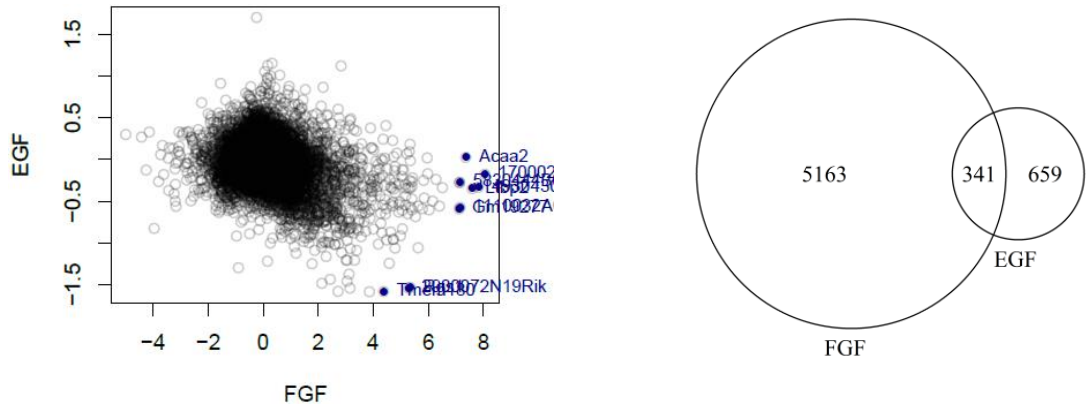


Fig. 19: Comparison of FGF and EGF using scatterplot and Venn diagram.
EGF and FGF gene regulation were uncorrelated. Significant genes of FGF overlap only with a small part of the best 1000 genes of EGF, for which no gene was significant.

Next we investigated the influences of FGF and EGF separately. No significant genes were found for EGF. In fact, all adjusted p-values had a value bigger than 0.9. Nevertheless we examined the best 1000 genes and compared them to the 5163 genes found for FGF (Fig. 19). 341 of the 1000 genes for EGF overlapped with the significant genes of FGF. This indicated that separating **'SAF'** and **'SAFE'** is hardly feasible.

4.1.2. Functional analysis

We next examined characteristics of the coefficients by analyzing gene functions and pathways of the significant genes. As **'SN'** showed no significant genes, we could not search any gene functions or pathways that might be enriched for this coefficient. For **'SAF'**, the method *mgsa* identified six terms using GO biological processes. Those are "intracellular signal transduction", "macromolecule localization", "cellular component organization or biogenesis", "cellular metabolic process", "single-organism metabolic process" and "cell adhesion". For **'SAFE'** exactly the same GO-functions were significant with a slightly different estimation value.

Furthermore, using KEGG we identified twenty-one pathways for **'SAF'** and twenty-eight pathways for **'SAFE'** including "Spliceosome", "ribosome biogenesis in eukaryotes", "cell cycle", "p53 signaling pathway" and twelve other pathways in both features. For **'SAF'** additional four pathways occurred. Those are "aminoacyl-tRNA biosynthesis", "lysine", "degradation", "other types of O-glycan biosynthesis" and "sulfur metabolism". For 'SAFE' we could report eleven pathways that were not significant for **'SAF'**. The appendix contains the complete list (Table 11).

As **'SAF'** and **'SAFE'** both were astrocytes, those functions and pathways indicated possible characteristics of astrocytes.

4.2. Combined dataset analysis

In this part we investigate the expression profiles of the combined-dataset.

4.2.1. Removing biological bias

After normalization we found non-biological difference between the subsets. The problem in combining different datasets was that there were some predominantly due to the fact that we combine data from different projects from the Götz lab. However, those differences showed no real biological differences. As a consequence for the analysis it was necessary to remove these so called batch effects. We used the empirical Bayes method *ComBat* to generate a couple of random design matrices for testing the robustness of the program. For all 50 different design matrices, which randomly included 8 samples with zeros only we analyzed the PCA. Fig. 20 illustrates the distributions of the first principal component (PC1) over the 43 samples for each of the 50 iterations. The additional red line corresponds to the first principal component for the samples of the normalized dataset which is not yet corrected for non-biological biases. For the different iterations we observed similar

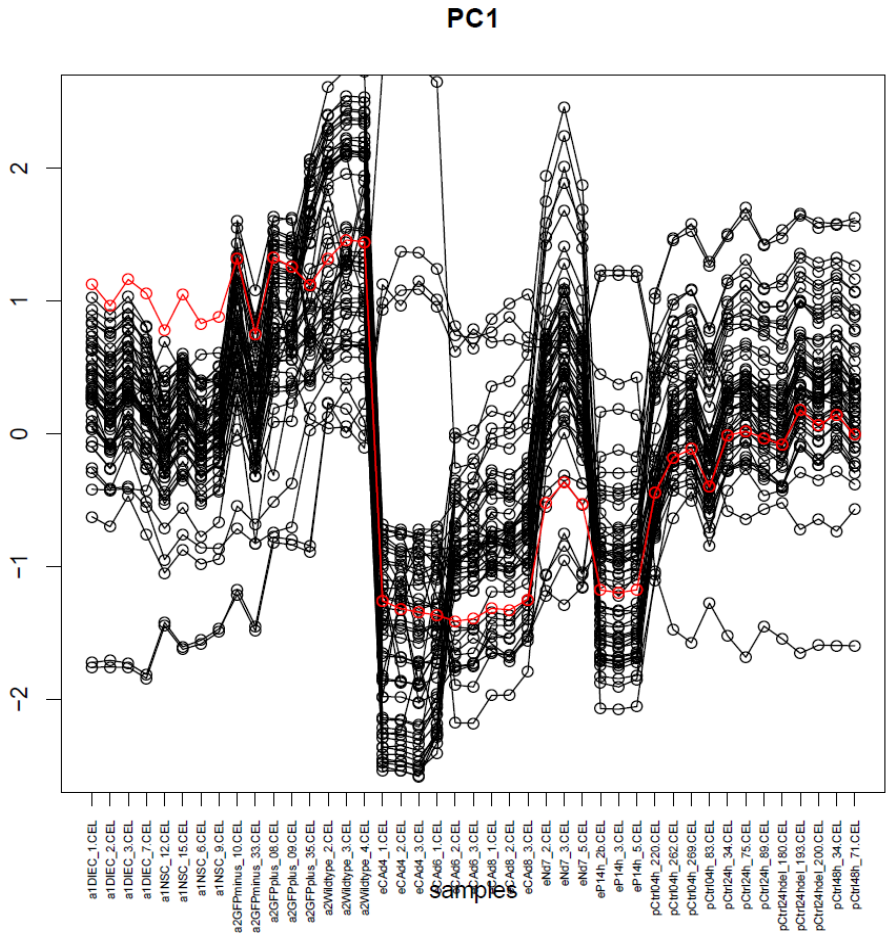


Fig. 20: Robustness of samples localized in PC1 using ComBat with 50 random design matrices
The x-axis indicates all 43 samples for which the first principal component is shown on the y-axis. The red line marks the distribution of the normalized but uncorrected dataset. Iterations show similar distributions indicating that ComBat is robust.

distributions of the first principle components. Therefore, the method appeared to be robust and we chose one design matrix to correct the combined dataset. In the following analyses we used the corrected dataset.

4.2.2. Investigate the corrected dataset with statistical analyses

Besides looking at the distribution of one principal component, we used the first two components to visualize the clustering of the samples. Fig. 21 a) depicts the PCA before correcting the dataset with ComBat. This plot shows the separation into the project-datasets. The right plot, Fig. 21 b), draws the first two principal components, PC1 and PC2, after correcting the combined-dataset. In the following we name the adult line containing diencephalon astrocyte (DIEC) and neural stem cell (NSC) samples **'a1'** (red circles) whereas the other line including GFP-minus, GFP-plus and wild-type astrocyte samples is named **'a2'** (green crosses). **'P6'** (violet quadrats) represents all samples of post-natal six cell cultures and **'ESC'** (blue diamonds) samples of the cultured embryonic stem cell project-dataset.

Obviously, before correcting the dataset for batch effects, the four subsets were clustered into the project-datasets. However, after removing batch effects the clustering was more mixed up indicating that non-biological biases were removed which separated the datasets before. Only the **'P6'** samples still clustered, which indicated a close relationship between them independent of time. Most obvious was the clustering of neurons to NSC. Additionally

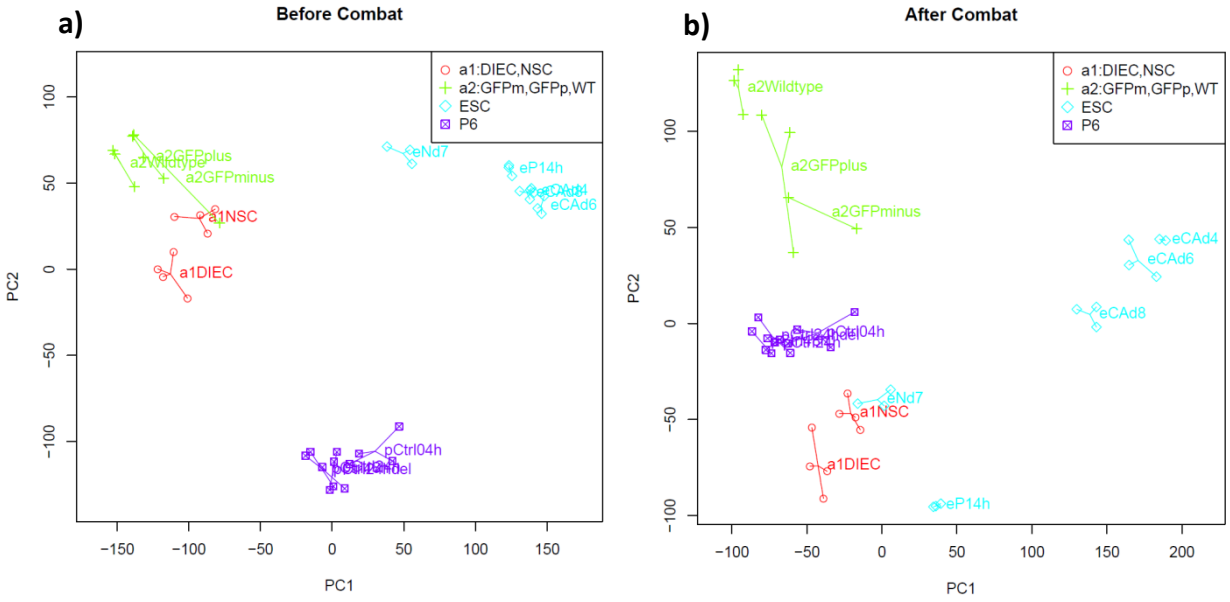


Fig. 21: PCA result for the 43 samples. The first two principal components are plotted.
Red circles and green crosses indicate samples from the adult cell line, where diencephalic astrocytes and NSCs are the red ones, whereas GFP-minus, GFP-plus and Wildtype are shown in green. Blue diamonds show the ESC samples. Violet quadrats mark all samples of the P6 cell line.

a) shows the PCA for the normalized but not-corrected dataset, whereas

b) shows the result for the dataset corrected for non-biological bias.

Diversion of the datasets can be viewed in a) whereas clustering got more mixed up in b) indicating tthat non-biological bias was removed.

diencephalon astrocytes were closer to NSC than to the 'P6' astrocytes of the cortex. This indicated similarities between the neural stem cells and neurons. Furthermore the results showed astrocytes from the diencephalon had more characteristics like neural stem cells. **'a1'** and **'a2'** samples got separated. So although they both came from adult lines the features neural stem cells and astrocytes of the diencephalon differed from normal or reactive astrocytes. Like expected ESCs, which were samples from cellular aggregates four days before and six days after adding retinoic acid, cluster together. Next neighbor were the samples eight days after adding RA. Although these **'eCAd8'** samples were identified as radial glia cells, the plated radial glia showed a closer relationship to neurons and NSC than to **'eCAd8'** samples. Nevertheless plated radial glia were more or less separated to all other clusters. Overall the first principal component distinguished between astrocytes together with NSC and neurons and samples which consisted of ESC or radial glial cells derived from ESC. The second principal component then divided the dataset into different astrocytic conditions.

In Fig. 22 a) we showed the clustering of the genes. Almost all genes clustered together. Relevant genes were marked and named. Those were Cd24a, Zfp711, Atp1a2, A730054J21Rik, Cdh1, Lin28a, Mapt, Pbx3 and Tac1. The right plot (Fig. 22 b)) illustrates that about 90% of the data was already described by the first principal component whereas the second still described about 60%. Starting with the fourth principal component, each component could explain less than 50% of the data.

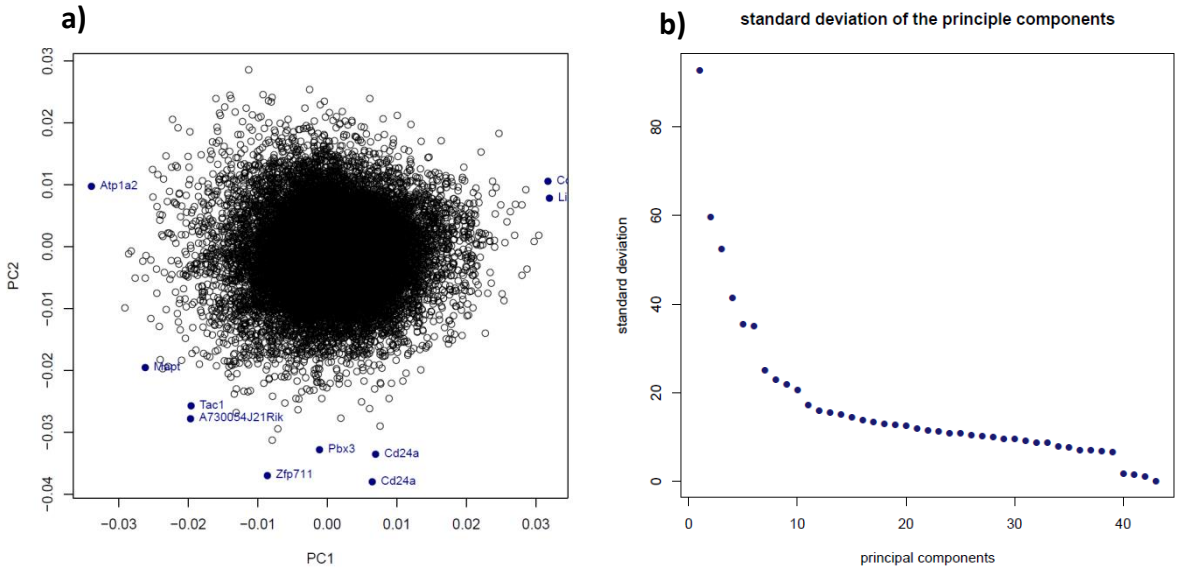


Fig. 22: PCA rotation matrix of genes and the standard deviations of the principle components.
The left plot (a) shows the first two principal components for the genes. Genes that do not cluster central are marked as outliers.
At the right (b) the standard deviations of the principal components are plotted. The y-axis represents how much of the data is explained by the appropriate principal component. The firs principal component describes more than 90% of the data

Now we investigated sample similarities with hierarchical clustering using Pearson correlation coefficients (Fig. 23). Euclidean distances resulted in a very similar plot (appendix, Fig. 38). Additionally to the colored map the dendrogram illustrates the clustering structure of the samples. Thereby we could observe a similar clustering like in the PCA result (Fig. 21). Especially the **'P6'** samples clustered together and **'a1'** samples clustered with neuron samples. Furthermore **'P6'** and **'a1'**-neuron samples were neighbors in the tree. Maybe the astrocytes of the **'P6'** samples therefore had some properties and capacities like neural stem cells or neurons. The cellular aggregate samples formed another cluster. Neighboring samples were the radial glia after plating. This was different to the clustering in the PCA, where the radial glial cells past plating were closer to diencephalon astrocytes or NSC and neurons than to the cellular aggregates. Finally **'a2'** samples formed a cluster which divided into its different cell types GFP-plus together with GFP-minus and wild-type. It is worth mentioning that one GFP-plus sample clustered with the **'P6'** cluster, which might be due to the fact, that this sample was one of the "zero lines" in the design matrix when applying *ComBat*. This showed that batch effects must be treated very carefully.

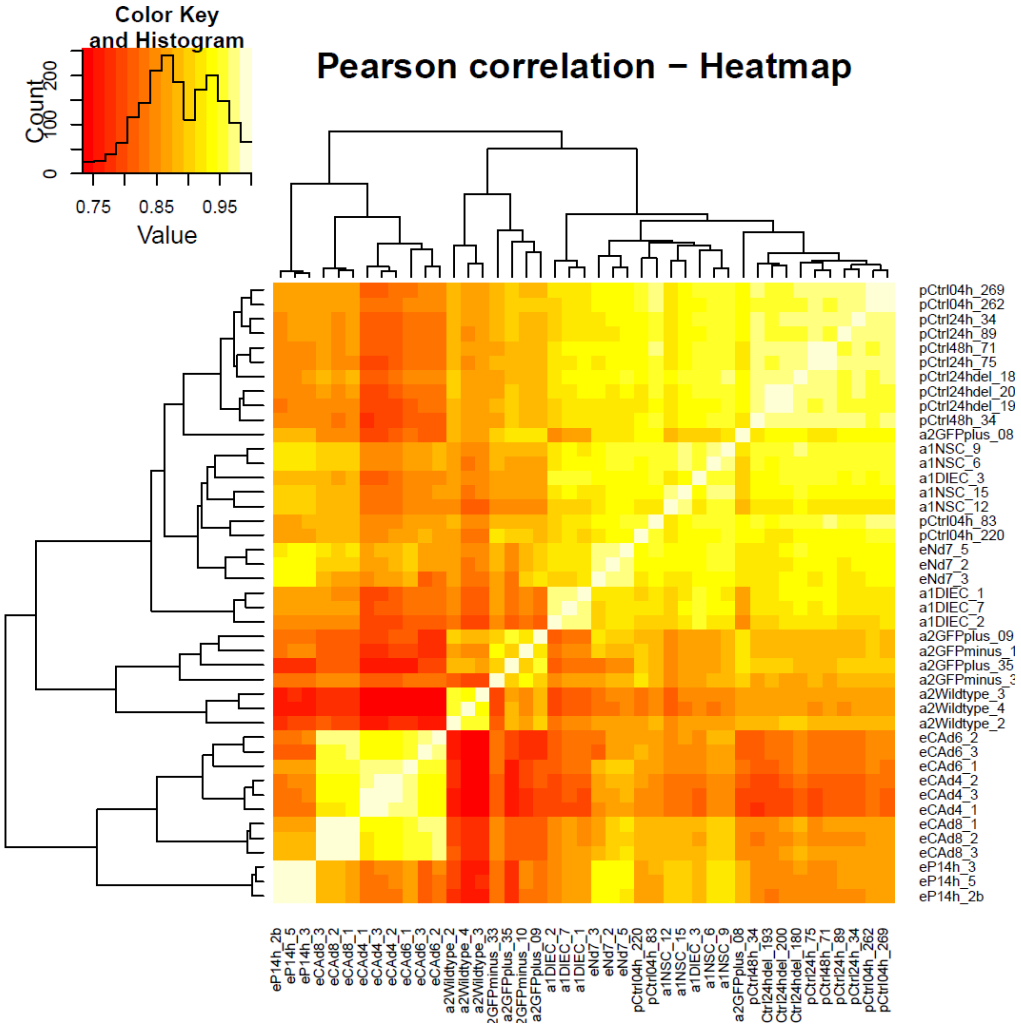


Fig. 23: Hierarchical clustering of the Pearson correlation coefficients for the corrected dataset.
All 43 samples are shown and a dendrogram show the clustering of the samples.

For both PCA and hierarchical clustering astrocytes showed some similarities to NSC and neurons like the astrocytes sampled from the diencephalon. Additionally **'P6'** astrocytes seemed to have some characteristics of NSC.

We next investigated the relationship between the biological features for the combined dataset. Deduced from the biological features we generated a design matrix (Table 2). Rows show all samples, whereas columns show the different biological features we finally chose for this complex dataset. Those were **'p4h'**, **'p48h'** and **'pDel'** for post-natal six cell cultures at different time points and 'P6' as factor they shared. Respectively we included a coefficient **'Adult'** for samples of the two adult lines (**'a1'** and **'a2'**) and the sub-coefficients **'aNSC'** for NSC, **'aWT'** for non-reactive astrocytes in healthy tissue, **'aGFPp'** for reactive astrocytes and **'aDIEC'** for astrocytes taken from the diencephalon. Additionally we added **'Lesion'** to the matrix for all cells taken from a lesion independent if they were astrocytes or not. Finally for the **'ESC'**-line **'eCA4d'** was added for non-treated ESCs, **'PastPl'** for all cell of this cell line that were plated, **'eNd7'** a part of the **'PastPl'** feature indicating neurons and **'radialGlia'** for all samples that included radial glial cells. However other features like **'p24h'** or **'ESC'** would be possible, too.

Sample	p4h	p48h	pDel	P6	aDIEC	aNSC	aGFPp	aWT	Adult	Lesion	eCA4d	PastPl	eNd7	radialGlia
a1DIEC_1	0	0	0	0	1	0	0	0	1	0	0	0	0	0
a1DIEC_2	0	0	0	0	1	0	0	0	1	0	0	0	0	0
a1DIEC_3	0	0	0	0	1	0	0	0	1	0	0	0	0	0
a1DIEC_7	0	0	0	0	1	0	0	0	1	0	0	0	0	0
a1NSC_6	0	0	0	0	0	1	0	0	1	0	0	0	0	0
a1NSC_9	0	0	0	0	0	1	0	0	1	0	0	0	0	0
a1NSC_12	0	0	0	0	0	1	0	0	1	0	0	0	0	0
a1NSC_15	0	0	0	0	0	0	0	0	1	0	0	0	0	0
a2GFPminus_10	0	0	0	0	0	0	0	0	1	1	0	0	0	0
a2GFPminus_33	0	0	0	0	0	0	0	0	1	1	0	0	0	0
a2GFPplus_08	0	0	0	0	0	0	1	0	1	1	0	0	0	0
a2GFPplus_09	0	0	0	0	0	0	1	0	1	1	0	0	0	0
a2GFPplus_35	0	0	0	0	0	0	1	0	1	1	0	0	0	0
a2Wildtype_2	0	0	0	0	0	0	0	1	1	0	0	0	0	0
a2Wildtype_3	0	0	0	0	0	0	0	1	1	0	0	0	0	0
a2Wildtype_4	0	0	0	0	0	0	0	1	1	0	0	0	0	0
eCAd4_1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
eCAd4_2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
eCAd4_3	0	0	0	0	0	0	0	0	0	0	1	0	0	0
eCAd6_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eCAd6_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eCAd6_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eCAd8_1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
eCAd8_2	0	0	0	0	0	0	0	0	0	0	0	0	0	1
eCAd8_3	0	0	0	0	0	0	0	0	0	0	0	0	0	1
eP14h_2b	0	0	0	0	0	0	0	0	0	0	0	1	0	1
eP14h_3	0	0	0	0	0	0	0	0	0	0	0	1	0	1
eP14h_5	0	0	0	0	0	0	0	0	0	0	0	1	0	1
eNd7_2	0	0	0	0	0	0	0	0	0	0	0	1	1	0
eNd7_3	0	0	0	0	0	0	0	0	0	0	0	1	1	0
eNd7_5	0	0	0	0	0	0	0	0	0	0	0	1	1	0
pCtrl04h_83	1	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl04h_220	1	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl04h_262	1	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl04h_269	1	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl24h_34	0	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl24h_75	0	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl24h_89	0	0	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl48h_34	0	1	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl48h_71	0	1	0	1	0	0	0	0	0	0	0	0	0	0
pCtrl24hdel_180	0	0	1	1	0	0	0	0	0	0	0	0	0	0
pCtrl24hdel_193	0	0	1	1	0	0	0	0	0	0	0	0	0	0
pCtrl24hdel_200	0	0	1	1	0	0	0	0	0	0	0	0	0	0

Table 2: Design matrix used for linear modeling.

Rows show the samples, whereas columns represent the coefficients. 0 indicates that the specific sample is not part of the coefficient. For example both the samples of GFPminus and of GFPplus are taken from a region of focal lesion and therefore have 1 at coefficient “Lesion”, whereas all other samples have 0 at this column.

We could report a different number of genes which were statistically significant in the different coefficients. Fig. 24 shows the number of genes for each in form of a barplot. Like expected summarizing features like **'P6'** with 12022 and **'Adult'** with 11758 significant enriched genes had a higher number than more specialized terms like **'p4h'** or **'p48h'**. However, diencephalon astrocytes and also neural stem cells had a high number with 10447 and 7446 genes as well. Additionally past plating cells including radial glia and neurons got 9220 significant genes. For **'p48h'** and for the **'aWT'** samples we only identified a low number of significant genes which were fifteen and seven, respectively. For the additional feature **'Astrocytes'**, which was the combination of the biological features defining astrocyte samples (**'P6'**, **'aDIEC'**, **'aGFP'** and **'aWT'**), 3259 genes showed up. The top five significant genes identified for **'Astrocytes'** were Klrd1, Prmt2, Sall4, Lin28a and Scrn1.

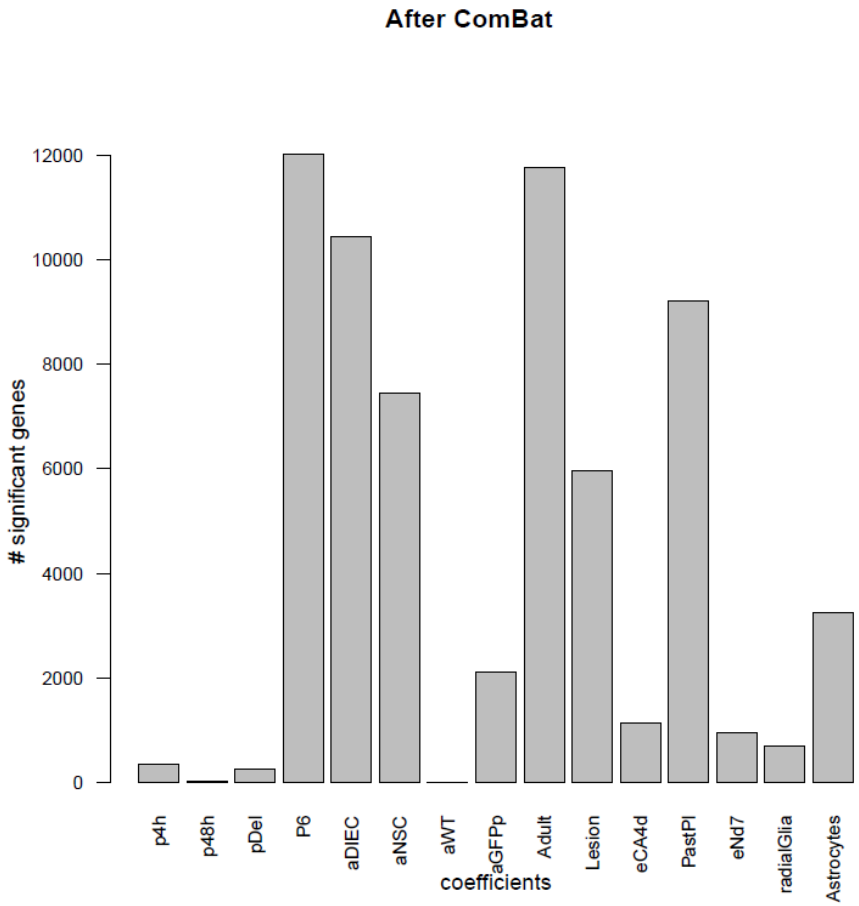


Fig. 24: Number of significant genes for the coefficients of the combined dataset.
On the x-axis all coefficients used at linear modeling are drawn. The y-axis represents the number of genes that are significantly up- or down regulated using the t-statistic of LIMMA and an adjusted p-value < 0.01. For 'p48h' and 'aWT' a few genes are statistically significant.

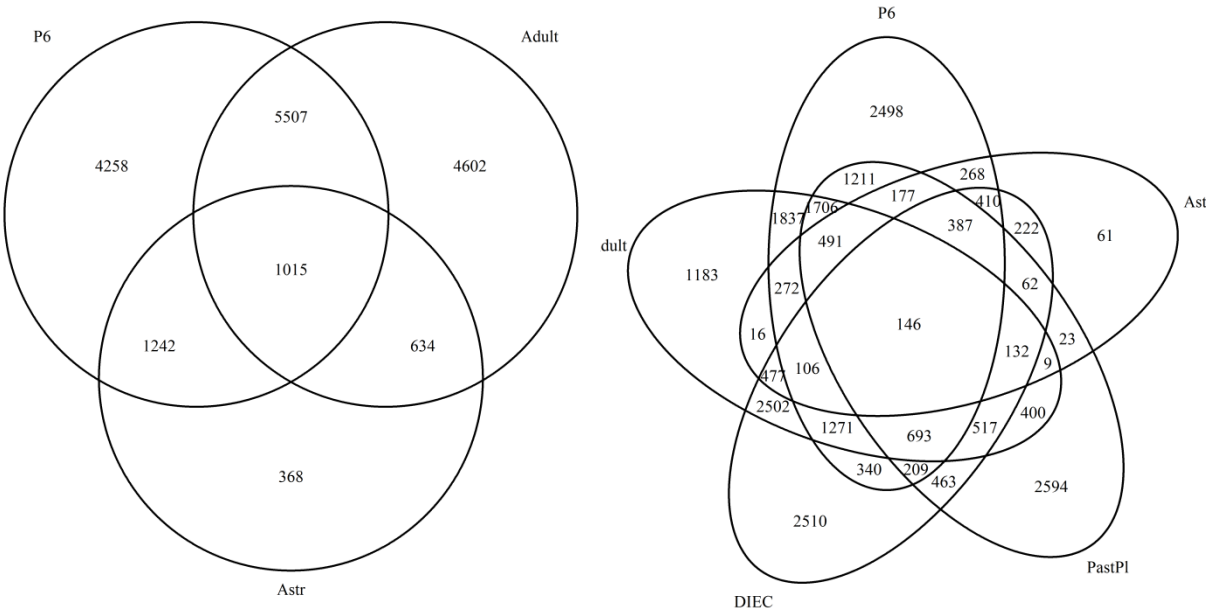


Fig. 25: Overlap of significant genes for coefficients 'P6', 'Adult' and 'Astrocytes' on the left. On the right 'DIEC' and 'PastPl' are added, too.

To investigate how much of the genes overlapped between different features, we generated Venn Diagrams for the two and for the four features that had the highest number of genes in the regression together with the factor **'Astrocytes'**. Fig. 25 showed the two diagrams including **'P6'**, **'Adult'** and **'Astrocytes'** in the left diagram and the same three coefficients together with **'aDIEC'** and **'PastPl'** in the right. We could observe a high overlap between **'P6'** and **'Adult'** sharing about 6500 genes with each other. As both **'P6'** and **'Adult'** contained astrocytes samples the high overlap between the two coefficients and 'Astrocytes' does not surprise. Nevertheless 368 genes were identified for **'Astrocytes'** that did not overlap with **'P6'** or **'Adult'**. For **'aDIEC'** and **'PastPl'** we could observe that many of their significant genes overlapped with another coefficient. For both still more than 2500 genes existed for each of these two coefficients not overlapping with the other coefficients in the plot.

We analyzed highly significant genes which also showed a magnitude fold-change using volcano plots. Fig. 26 shows the result of coefficient **'Astrocytes'**. Ten genes that had the highest t-statistic value were marked in the plot. We observed a large fold change for "Kird1" or "Bai3", too.

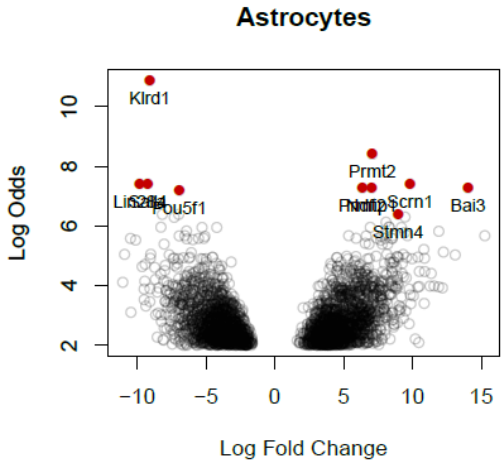


Fig. 26: Volcano plot of the coefficient 'Astrocytes' of the linear regression model.
The x-axis represents the fold change, whereas y-axis shows the log odds. Ten genes with highest t-statistics are marked. Klrd1 and Bai also show a magnitude fold change.

We investigated similarities of regulated gene between the coefficients using scatterplots. Three of the scatterplots are shown in Fig. 27. For **'aNSC'** against **'Astrocytes'** only a small dependency if any could be viewed between the two coefficients. For **'aDIEC'** and **'aNSC'** there was a strong correlation showing the similarities between astrocytes of the diencephalon and NSC of the subventricular zone. Additionally **'aDIEC'** showed a correlation to **'Lesion'**, not identified between **'aDIEC'** and **'GFPp'**. The ten most relevant genes, when plotting **'aNSC'** against **'Astrocytes'**, were Rreb1, 2210016F16Rik2, Gemin8, Vps29, Pcm1, Ubtf, Cdkn2d, Tmem255a, Acsl4, Wnt9b. Further volcano and scatterplots can be viewed in the appendix (Fig. 39). Scatterplots which showed only few if any dependencies are not printed.

For the fifty most significant genes of feature **'P6'** we generated a heatmap of the expression profile (Fig. 28). Thereby we plotted only the nested samples of **'P6'** together with the samples of **'ESC'** before treated with retinoic acid. Those **'ESC'** samples served as outliers. We observed that genes with lower expression values in **'P6'** mostly had higher values in the

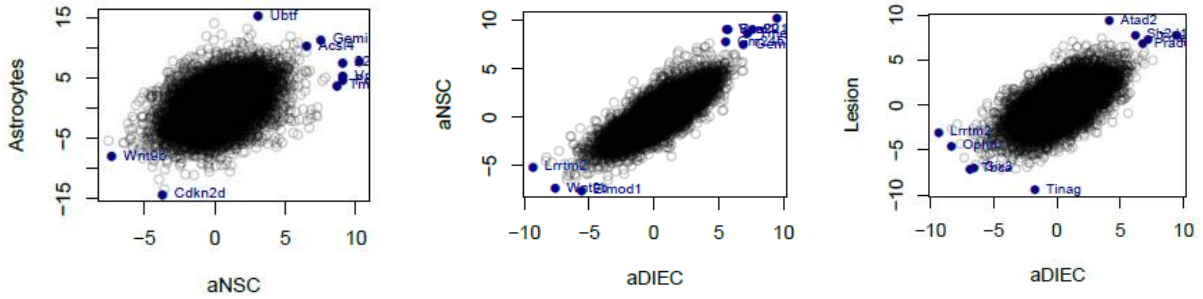


Fig. 27: Regulation comparison of the coefficients 'aNSC' and 'Astrocytes, 'aDIEC' and 'aNSC' as well as 'aDIEC' and 'Lesion'.
Correlations between diencephalic astrocytes and NSC was observed and also with cells from lesion sides. Only a slight correlation could be viewed for 'aNSC' against 'Astrocytes'.

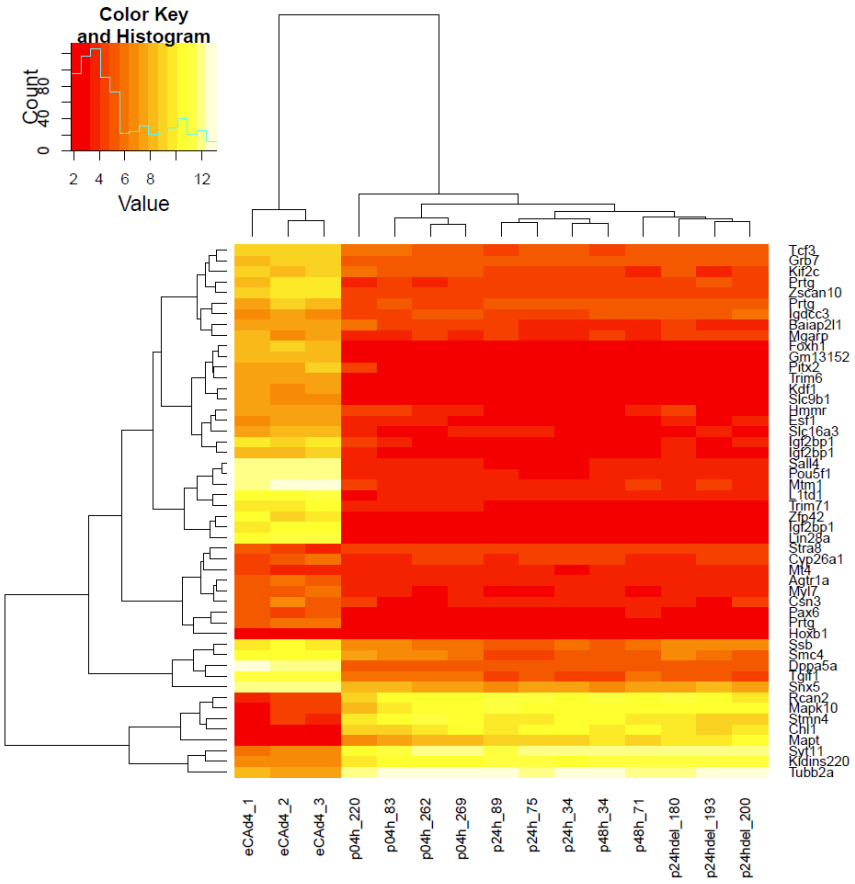


Fig. 28: Expressions of the 50 most significant genes of 'P6'.
Only 'P6' samples are plotted with 'eCAd4' as outgroup showing high similarity between all 'P6' samples.

'ESC' samples and vice versa. Furthermore we identified a high similarity between the **'P6'** samples.

Using linear regression a few similarities between astrocytes and NSC were observed, especially for diencephalon astrocytes. Diencephalon astrocytes also showed some characteristics like cells from injured regions. To other cell types like ESC, radial glia or neurons, astrocytes showed no relationships.

4.2.3. Functional analyzes

We analyzed if the genes shared biological features and searched for enriched gene ontology (GO) terms and KEGG pathways. We identified two terms for feature **'Astrocytes'** in the analysis using *mgsa* with GO of biological processes. These were "organelle organization" and "nervous system development". "Organelle organization" was also observed for **'aNSC'** and **'aDIEC'** as well as for **'Adult'** in general indicating that there might be a relationship between astrocytes and NSC. We found five terms for **'P6'**, too. Its sub-groups **'p4h'**, **'p48h'** and **'pDel'** alone had no significant terms. The five terms for **'P6'** were "synaptic transmission", "intracellular signal transduction", "cellular component organization or biogenesis", "macromolecule localization" and "cellular metabolic process".

Searching KEGG pathways, we observed that **'P6'**-genes participated in the "spliceosome" pathway as well as in "RNA transport" and "cell cycle". We already knew that astrocytes form gap and tight junctions which also appeared in feature **'P6'** with the enriched pathway "tight junction". The complete list of **'P6'** can be viewed in Table 3. In addition we identified two pathways enriched for **'Astrocytes'** which were "Ribosome" and "DNA replication". The appendix lists GO-terms and KEGG pathways of other coefficients (Table 12 and Table 13).

P6	
Spliceosome	Pyrimidine metabolism
RNA transport	Tight junction
Cell cycle	Nucleotide excision repair
Ribosome biogenesis in eukaryotes	Cyanoamino acid metabolism
mRNA surveillance pathway	Purine metabolism
Cysteine and methionine metabolism	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Ribosome	Basal transcription factors
Homologous recombination	

Table 3: KEGG pathways for 'P6'

4.2.4. Network smoothing using a regression model

In the following, we describe the approach of network smoothing on the basis of a small real data example. Therefore, we searched a small group of genes interacting with each other. Fig. 29 shows the interaction network of seven genes Prelp, Smoc2, Fmod, Bgn, Tnc, Mmp2 and Ptn, the mini-example contained.

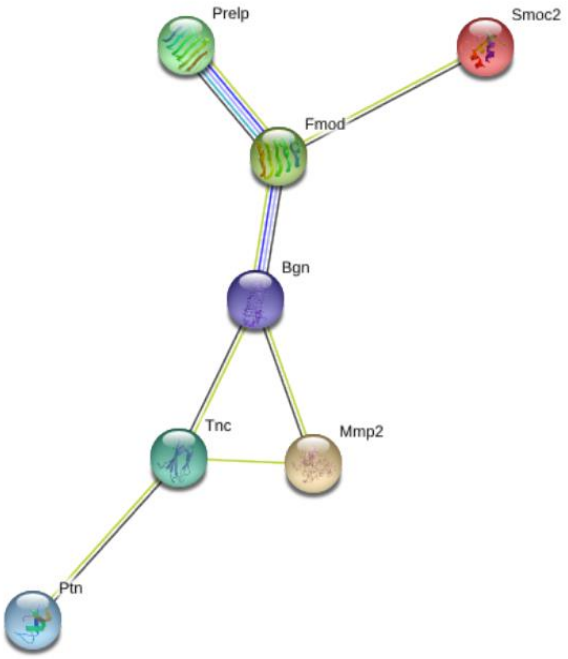


Fig. 29: One cluster as example network with nodes Prelp, Smoc2, Fmod, Bgn, Tnc, Mmp2 and Ptn.

Table 4 prints the kernel that was calculated for this network.

K	Smoc2	Mmp2	Fmod	Prelp	Tnc	Ptn	Bgn
Smoc2	1	.	0.58	.	.	.	.
Mmp2	.	1	.	.	0.41	.	0.41
Fmod	0.58	.	1	0.58	.	.	0.33
Prelp	.	.	0.58	1		.	.
Tnc	.	0.41	.	.	1	0.58	0.33
Ptn	.	.	.	.	0.58	1	.
Bgn	.	0.41	0.33	.	0.33	.	1

Table 4: Kernel matrix for the mini-network with nodes "Prelp", "Smoc2", "Fmod", "Bgn", "Tnc", "Mmp2" and "Ptn"

We built a nested design matrix as an exemplary subset of the original design matrix (Table 2) including the four **'P6'** features as well as the corresponding samples and one outlier sample of the ESC line. The matrix $\boldsymbol{T}$ (Table 5) showed the observed t-statistic after linear regression and the smoothed t-statistic ($\widetilde{\boldsymbol{T}}$) for the small network and the four example coefficients. For example Mmp2 had two neighbors Bgn and Tnc. Looking at one coefficient like 'p4h' we observed that in $\boldsymbol{T}$ the t-statistics of the two neighbors were higher. Therefore the smoothed t-statistic of Mmp2 in $\widetilde{\boldsymbol{T}}$ was also increased.

T	p4h	p48h	pDel	P6
Smoc2	0.13	0.13	0.11	0.13
Mmp2	0.21	0.05	0.01	0.22
Fmod	1	0.6	0.29	1
Prelp	0.47	0.34	0.12	0.48
Tnc	0.44	0.85	0.69	0.94
Ptn	0.16	1	1	0.21
Bgn	0.54	0.41	0.28	0.74

$\widetilde{T}$	p4h	p48h	pDel	P6
Smoc2	0.46	0.3	0.2	0.44
Mmp2	0.4	0.36	0.3	0.57
Fmod	1	0.63	0.37	1
Prelp	0.69	0.43	0.2	0.66
Tnc	0.52	1	0.98	0.87
Ptn	0.27	0.94	1	0.47
Bgn	0.72	0.58	0.44	0.92

Table 5: t-statistic matrices divided by the maximum per coefficient before ($\mathbf{T}$) and after smoothing ($\widetilde{\mathbf{T}}$).

Table 6 represents the two matrices $\boldsymbol{P}$**,** which included the p-values of the linear regression, and $\widetilde{\boldsymbol{P}}$ with the p-values for the smoothed t-statistics of the permutation test. Thereby we marked p-values smaller than 0.05.

P	p4h	p48h	pDel	P6
moc2	0.7	0.79	0.88	0.48
Mmp2	0.7	0.86	0.97	0.3
Fmod	0.08	0.15	0.74	0.001
Prelp	0.38	0.38	0.88	0.04
Tnc	0.38	0.048	0.24	0.001
Ptn	0.7	0.04	0.09	0.3
Bgn	0.38	0.33	0.74	0.003

$\widetilde{P}$	p4h	p48h	pDel	P6
Smoc2	0.58	0.86	0.93	0.61
Mmp2	0.78	0.85	0.9	0.62
Fmod	0	0.8	0.94	0
Prelp	0.29	0.61	0.87	0.37
Tnc	0.78	0	0.31	0.43
Ptn	0.74	0.02	0	0.52
Bgn	0.44	0.73	0.9	0.17

Table 6: p-values for the t-statistics of the linear regression (P) and for the smoothed t-statistics calculated via permutation test ($\widetilde{P}$).

In P we identified six significant values. After smoothing with the kernel, the genes Tnc and Ptn of coefficient **'p48h'** still were significant. For coefficient **'p4h'** gene Fmod and for **'pDel'** gene Ptn were now significant. Both coefficients had no significant gene before. For **'P6'**, however, three of the four significant genes got lost after smoothing. Only Fmod remained.

We analyzed expression profile for the seven genes (Fig. 30). Differences between samples could be hardly observed. Nevertheless the heatmap already showed decreased expression values of Fmod especially for the **'p4h'** samples and the ESC sample. This indicated that we could improve the t-statistic with the network information.

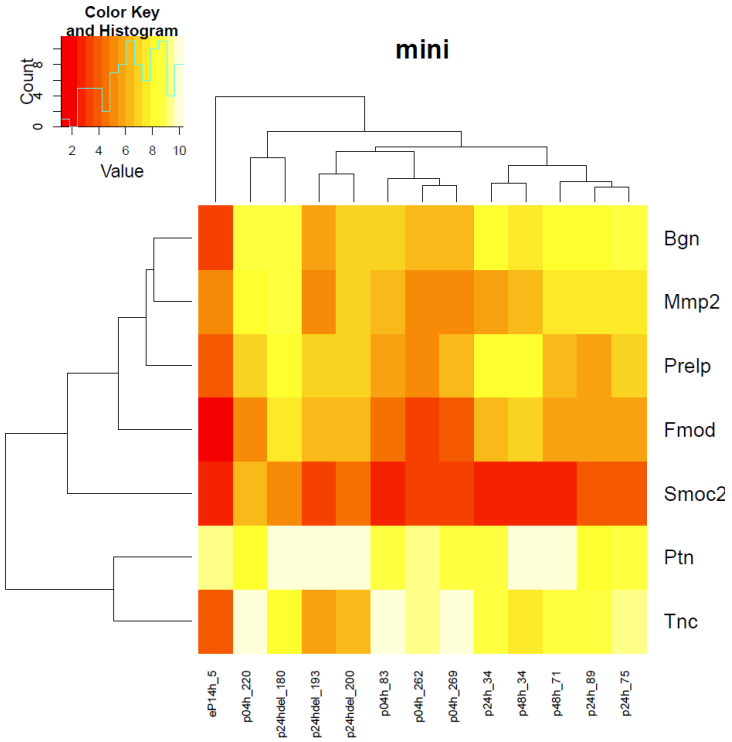


Fig. 30: Expression values for the mini-network.

We next analyzed the results for the network smoothing on the combined dataset. Initially, mapping the probeset IDs and gene symbols to the STRING IDs resulted in 34170 genes. 17038 genes remained after combining probeset-IDs according to the STRING ID and only 15381 genes of those showed an interaction with another gene.

To test the performance of the network smoothing, we performed a scatterplot of the observed t-statistic (t) to the smoothed t-statistic (tx) (Fig. 31) resulting in a plot a volcano-like appearance. Like for **'p48h'**, all coefficients showed such a dependency between the two statistics in the scatterplot suggesting that the original information still is included. Thereby it is worth mentioning that when including all nodes without any interaction this was not the case.

In the next step we investigated the number of genes that were significant before and after smoothing. Table 7 gives an overview of the results. The first column showed the coefficients. For the linear regression model of the mapped and combined expression profile we set the p-value threshold to 0.05. The second column gives the numbers of significant genes before smoothing. Thus the third column represents the numbers of significant genes after smoothing and the last column the number of genes that were significant before and after smoothing. For most coefficients more significant genes were found before smoothing. However, we identified new significant genes for **'p48h'**, **'pDel'** and **'aWT'**. The number of overlapping genes of these coefficients was equal to the number of significant genes before smoothing. For example for **'p48h'** there were 172 genes significant after and 83 before smoothing with 89 new significant genes. Coefficients **'p4h'** and **'eCA4d'** had more

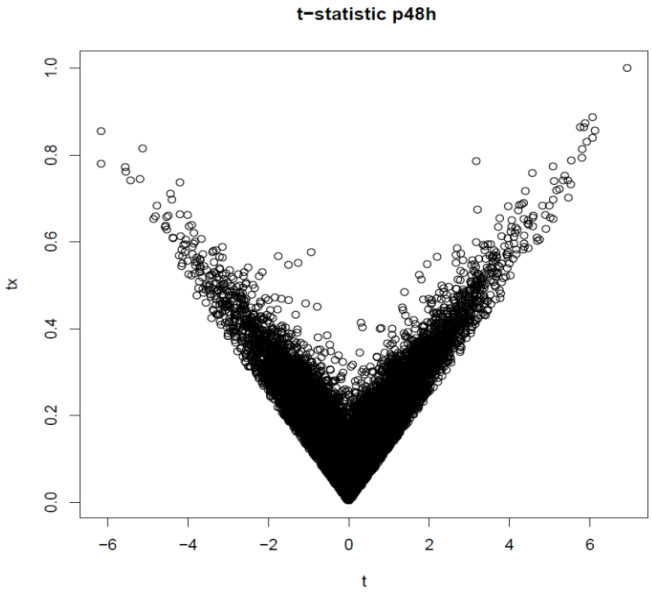


Fig. 31: Comparison of the t-statistics for coefficient 'p48h' before and after smoothing.
The x-axis corresponds to the originally observed t-statistic (t) and the y-axis to the smoothed version (tx). Comparison of the t-statistics show that the original information is still included indicating a good performance of network-smoothing.

significant genes before smoothing, however, they got a few new genes, too. For the other coefficients, the number of genes significant in both variations was identical to the number of significant genes after smoothing. No new genes were added only some genes were not significant after smoothing.

Coefficient	#sig $P < 0.05$	#sig $\widetilde{P} < 0.05$	overlap
p4h	466	170	167
p48h	83	172	83
pDel	12	194	12
P6	8082	101	101
aDIEC	6903	269	269
aNSC	5826	221	221
aWT	19	302	19
aGFPp	2476	4	4
Adult	8002	861	861
Lesion	5007	105	105
eCA4d	1359	19	18
PastPl	4921	217	217
eNd7	2492	36	36
radialGlia	717	46	46
Astrocytes	3643	180	180

Table 7: Number of significant genes before and after smoothing.
Rows indicate the coefficients. The second column shows the number of significant genes before smoothing. The third column represents the number of genes significant after smoothing with the network information and the last column shows the number of genes that are significant in both statistics. Threshold for significance is 0.05.

So far we identified 89 new significant genes for the feature **'p48h'** in the analyses. Fig. 32 plots their expression values. Expressions showed similarities in samples of **'p4h'**, **'p24h'** and **'p48h'**. For **'pDel'** they showed some differences. The outlier samples of **'eCAd4'** exhibited distinctions to the expressions of the other samples.

Fig. 33 shows the neighboring network of those genes. Yellow nodes represented genes that were significant after smoothing but not before, whereas blue nodes showed the original significant values with linear regression only. Many genes were added to the original gene network extending it to a big cluster. Nevertheless also some genes were now significant which did not have any interaction partners in the gene network of **'p48h'**.

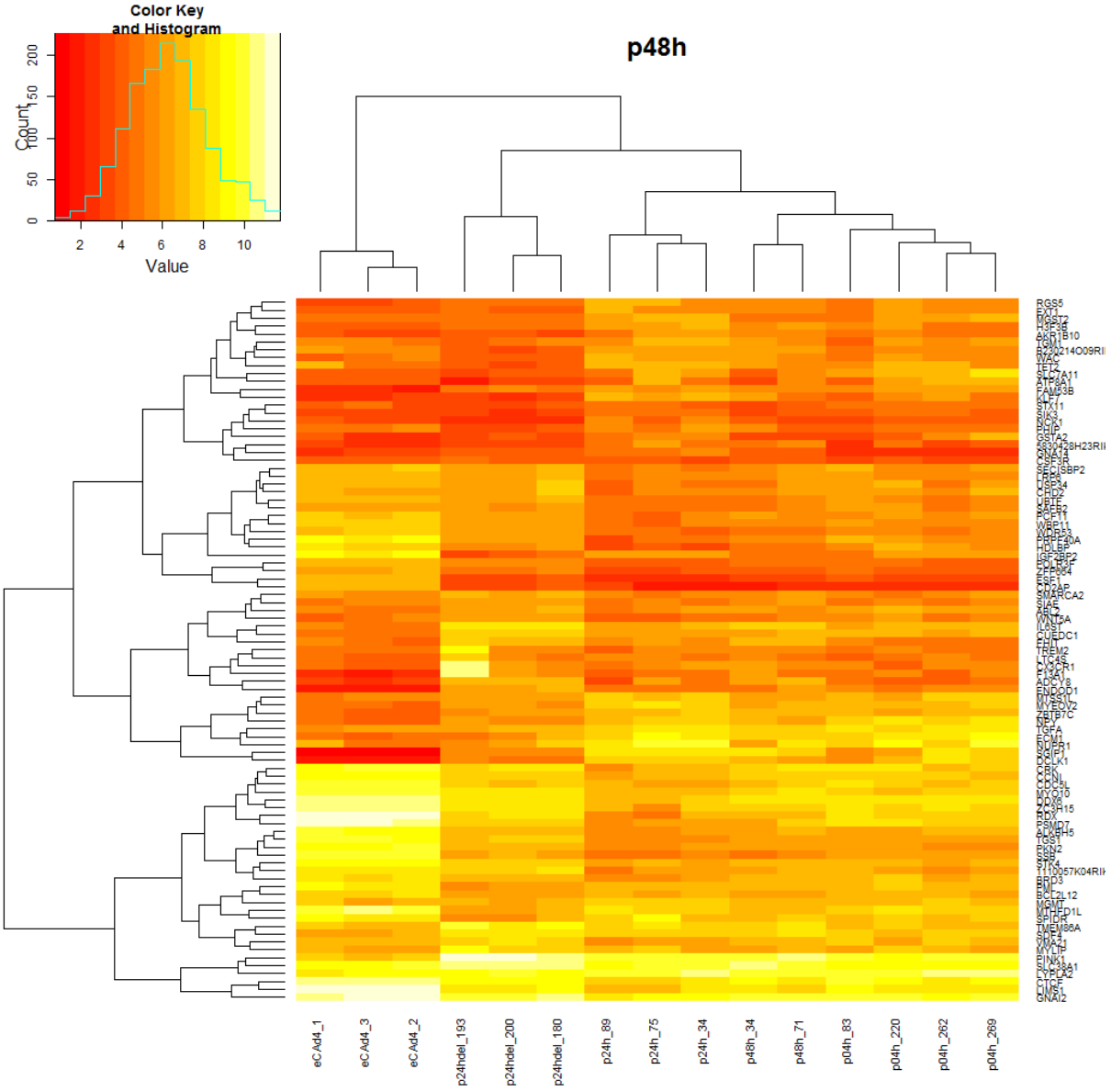


Fig. 32: Gene expressions of 89 genes significant after smoothing for coefficient 'p48h', which were not significant before.

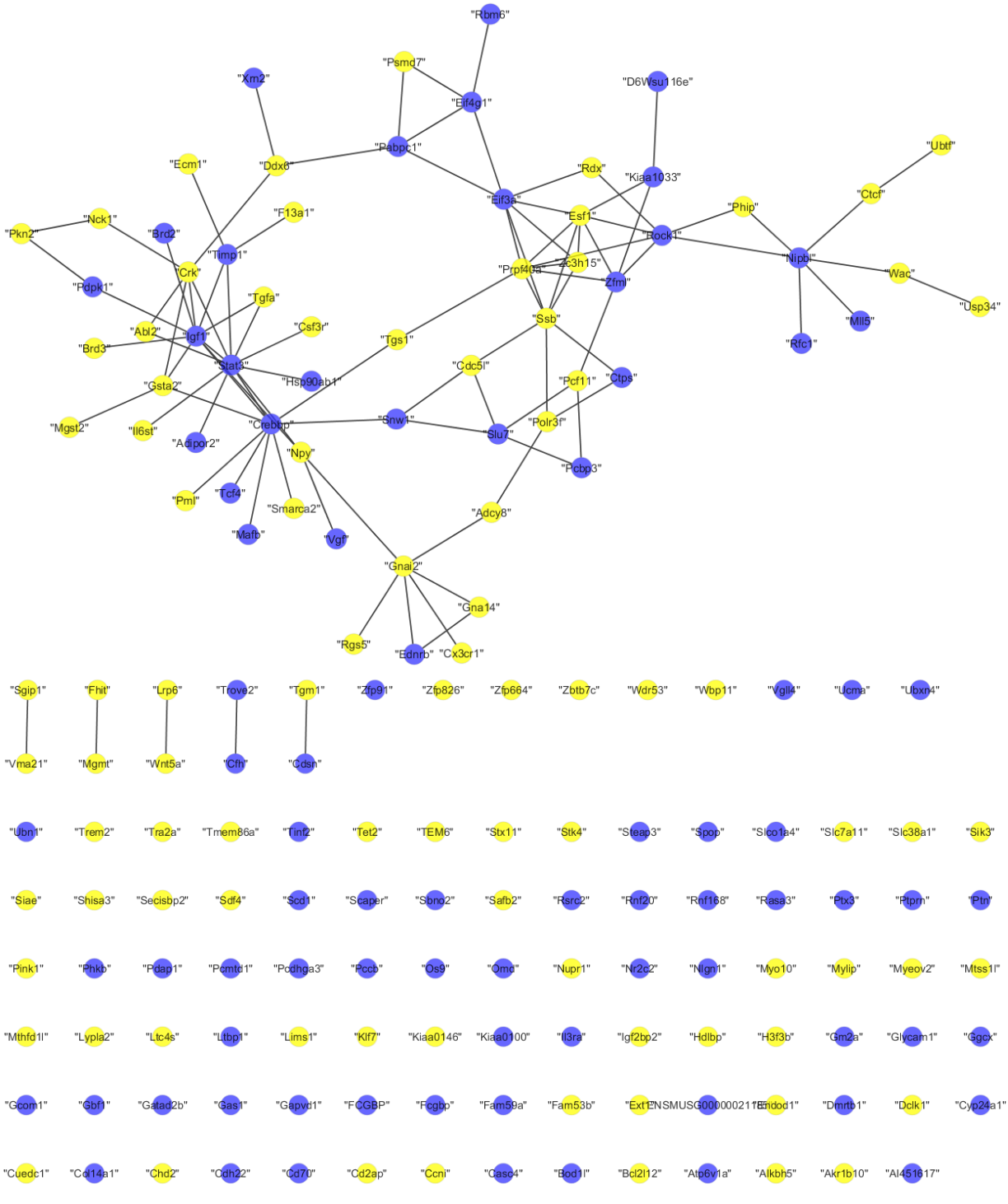


Fig. 33: Neighborhood network of the significant genes of 'p48h'.
In yellow the 89 genes, which were significant after smoothing but not before. The original cluster (blue nodes) is extended.

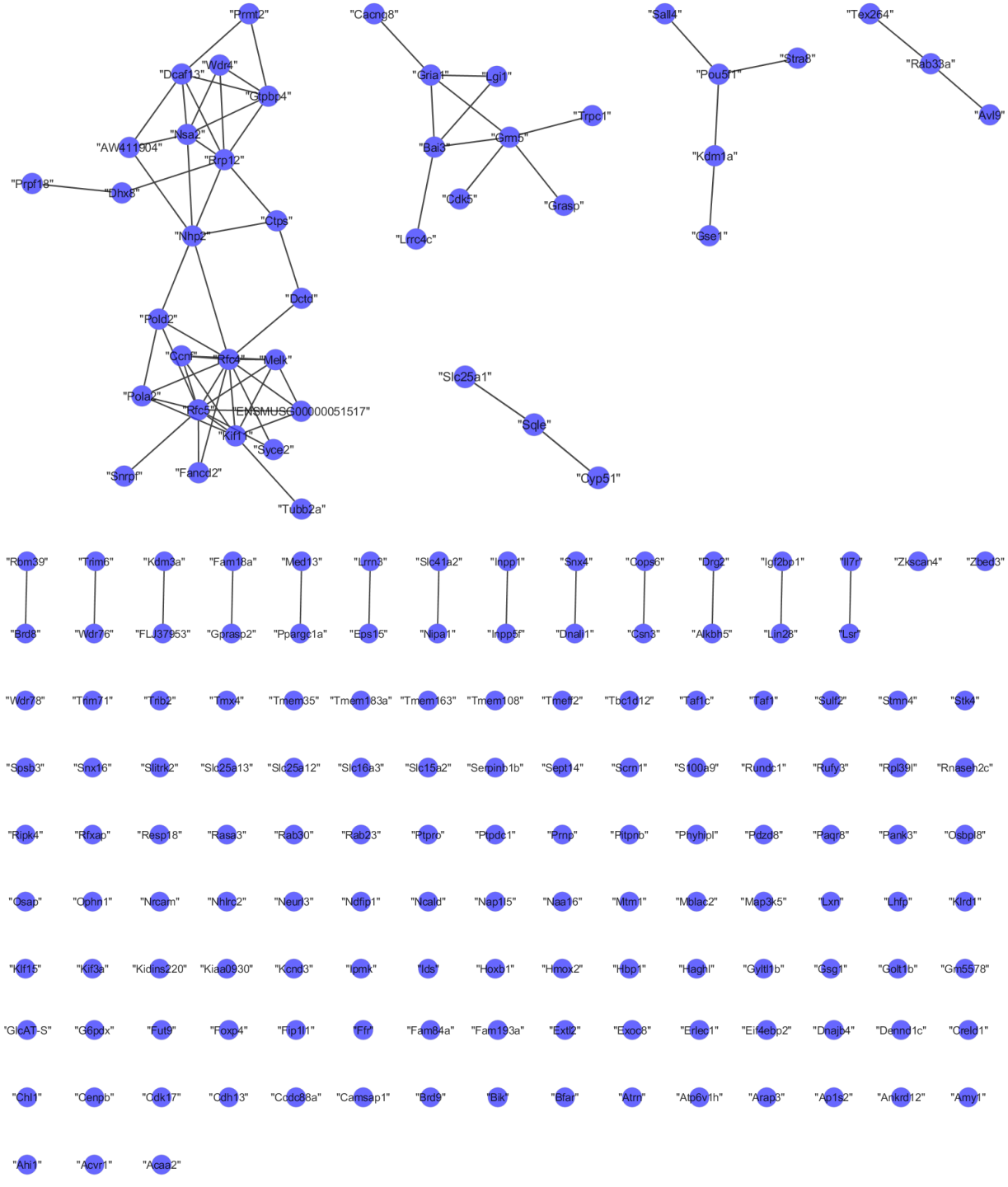


Fig. 34: Protein-protein interaction network of the significant genes of 'Astrocytes'.

One big gene cluster can be viewed and some additional smaller clusters. No genes were significant after smoothing but not before.

Additionally, Fig. 34 shows the protein-protein interaction network with significant genes of coefficient 'Astrocytes' including one big gene cluster and a couple of smaller ones. No significant genes exist after smoothing, which were not significant before. The network was only reduced.

4.2.5. Novel approaches for regression smoothing

We analyzed four methods using the same mini example as in 4.2.4.

First we investigated the two aggregation methods. Table 8 shows the final t-statistics and p-values for those on the same mini-example like before. Next we analyzed our approaches of regression smoothing using extended networks. As the t-statistic of **'P6'** was extended with zeros, they were the same in both approaches. Table 9 shows the results.

T_{a1}	p4h	p48h	pDel	P6
Smoc2	0.32	0.37	0.39	0.31
Mmp2	0.43	0.43	0.44	0.34
Fmod	1	0.98	0.9	1
Prelp	0.59	0.62	0.56	0.57
Tnc	0.71	1	1	0.97
Ptn	0.35	0.78	0.84	0.83
Bgn	0.78	0.84	0.84	0.78

P_{a1}	p4h	p48h	pDel	P6
Smoc2	0.868	0.83	0.76	0.94
Mmp2	0.86	0.87	0.82	0.95
Fmod	0	0.41	0.52	0
Prelp	0.49	0.44	0.55	0.56
Tnc	0.66	0	0	0.36
Ptn	0.69	0.1	0.07	0.07
Bgn	0.39	0.29	0.29	0.39

T_{a2}	p4h	p48h	pDel	P6
Smoc2	0.45	0.42	0.41	0.43
Mmp2	0.51	0.53	0.54	0.51
Fmod	1	0.94	0.89	0.94
Prelp	0.67	0.63	0.57	0.63
Tnc	0.76	1	1	1
Ptn	0.4	0.7	0.7	0.74
Bgn	0.86	0.87	0.86	0.84

P_{a2}	p4h	p48h	pDel	P6
Smoc2	0.73	0.82	0.81	0.86
Mmp2	0.78	0.76	0.76	0.84
Fmod	0	0.62	0.66	0.65
Prelp	0.32	0.41	0.52	0.39
Tnc	0.61	0	0	0
Ptn	0.62	0.13	0.12	0.06
Bgn	0.27	0.27	0.33	0.35

Table 8: Results of regression smoothed t-statistic using aggregation.
$\boldsymbol{T_{a1}}$ and $\boldsymbol{P_{a1}}$ are the smoothed t-statistic results and the corresponding p-value using the first aggregation method and $\boldsymbol{T_{a2}}$ and $\boldsymbol{P_{a2}}$ for the second.

T_{e1}	p4h	p48h	pDel	P6
Smoc2	0.27	0.23	0.18	0.34
Mmp2	0.31	0.3	0.25	0.51
Fmod	0.95	0.81	0.65	1
Prelp	0.61	0.53	0.4	0.59
Tnc	0.66	1	0.93	0.9
Ptn	0.24	0.73	0.7	0.38
Bgn	0.69	0.67	0.57	0.92

P_{e1}	p4h	p48h	pDel	P6
Smoc2	0.91	0.94	0.98	0.72
Mmp2	0.95	0.95	0.97	0.68
Fmod	0.3	0.43	0.65	0
Prelp	0.44	0.54	0.71	0.43
Tnc	0.6	0	0.29	0.37
Ptn	0.83	0.16	0.17	0.6
Bgn	0.5	0.52	0.68	0.22

T_{e2}	p4h	p48h	pDel	P6
Smoc2	0.31	0.24	0.19	0.34
Mmp2	0.34	0.3	0.27	0.51
Fmod	1	0.76	0.59	1
Prelp	0.63	0.5	0.35	0.59
Tnc	0.63	1	1	0.9
Ptn	0.25	0.82	0.87	0.38
Bgn	0.72	0.65	0.57	0.92

P_{e2}	p4h	p48h	pDel	P6
Smoc2	0.89	0.96	0.97	0.72
Mmp2	0.93	0.95	0.98	0.72
Fmod	0	0.57	0.76	0
Prelp	0.49	0.64	0.83	0.41
Tnc	0.67	0	0	0.36
Ptn	0.88	0.12	0.09	0.55
Bgn	0.45	0.59	0.72	0.22

Table 9: Results of regression smoothed t-statistic using extended networks.
$\boldsymbol{T_{e1}}$ and $\boldsymbol{P_{e1}}$ are the smoothed t-statistic results and the corresponding p-value using the first extended network approach and $\boldsymbol{T_{e2}}$ and $\boldsymbol{P_{e2}}$ for the second.

Both aggregation methods and the second extension method showed significant t-statistics for coefficient **'p4h'** gene Fmod and for **'pDel'** gene Tnc. Furthermore, all four approaches showed gene Tnc significant for **'p48h'** and Fmod for **'P6'** was significant in all but the second aggregation approach. Those were the only two genes we could find with linear regression before smoothing (Table 6), but for the second aggregation method additionally Tnc was significant for **'P6'** like before.

For network smoothing without coefficient information the results were more similar. However, significance for Ptn got lost with regression smoothing for both **'p48h'** and **'pDel'**, but we gained Tnc for **'pDel'** as new gene. However, those methods were not yet fully analyzed and a further look on "regression smoothing" is necessary.

4.3. Comparison of functional analyses

So far we analyzed the growth factor and the combined dataset separately. To make a comparison between them we investigated the functional analyses of both sets. Using *mgsa* we got enriched GO or KEGG terms together with an estimated value. Fig. 35 plots the estimated values of all GO-terms that were enriched significantly in at least one of the features. Only 'SN' of the growth factor dataset was not included, because no significant genes were observed for functional investigations. Therefore we were not able to compare the NSC from the neurosphere (SN) to the NSC from the direct adult line (**aNSC**) or any other feature of the combined dataset.

The strongest overlap of **'SAF'** and **'SAFE'** of the growth factor set could be viewed with **'P6'** of the combined set. Four of the six terms significant for **'SAF'** and **'SAFE'** were significant for **'P6'**, too. Additionally one term overlapped with **'Adult'** and one with **'PastPI'**. Altogether we identified different functions for the different features. For some features like **'p4h'** or **'aWT'** no specific functions were observed. The similarity between **'SAF'** and **'SAFE'** to **'P6'** indicated that even if astrocytes were treated with growth factors they were similar to

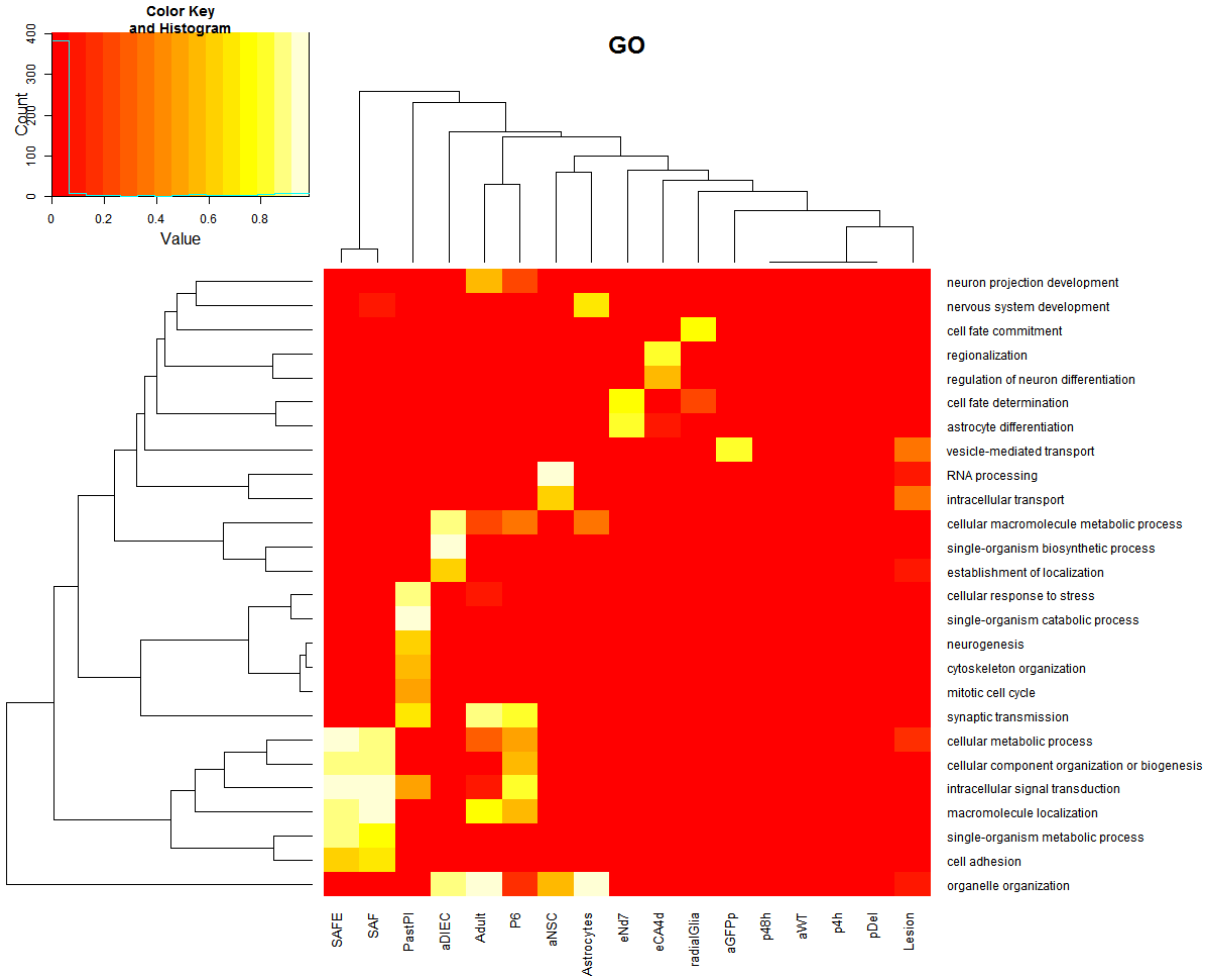


Fig. 35: Estimations of GO-annotations by MGSA for both growth factor dataset and combined dataset. Astrocytes treated with growth factors show similar functions as 'P6' astrocytes.

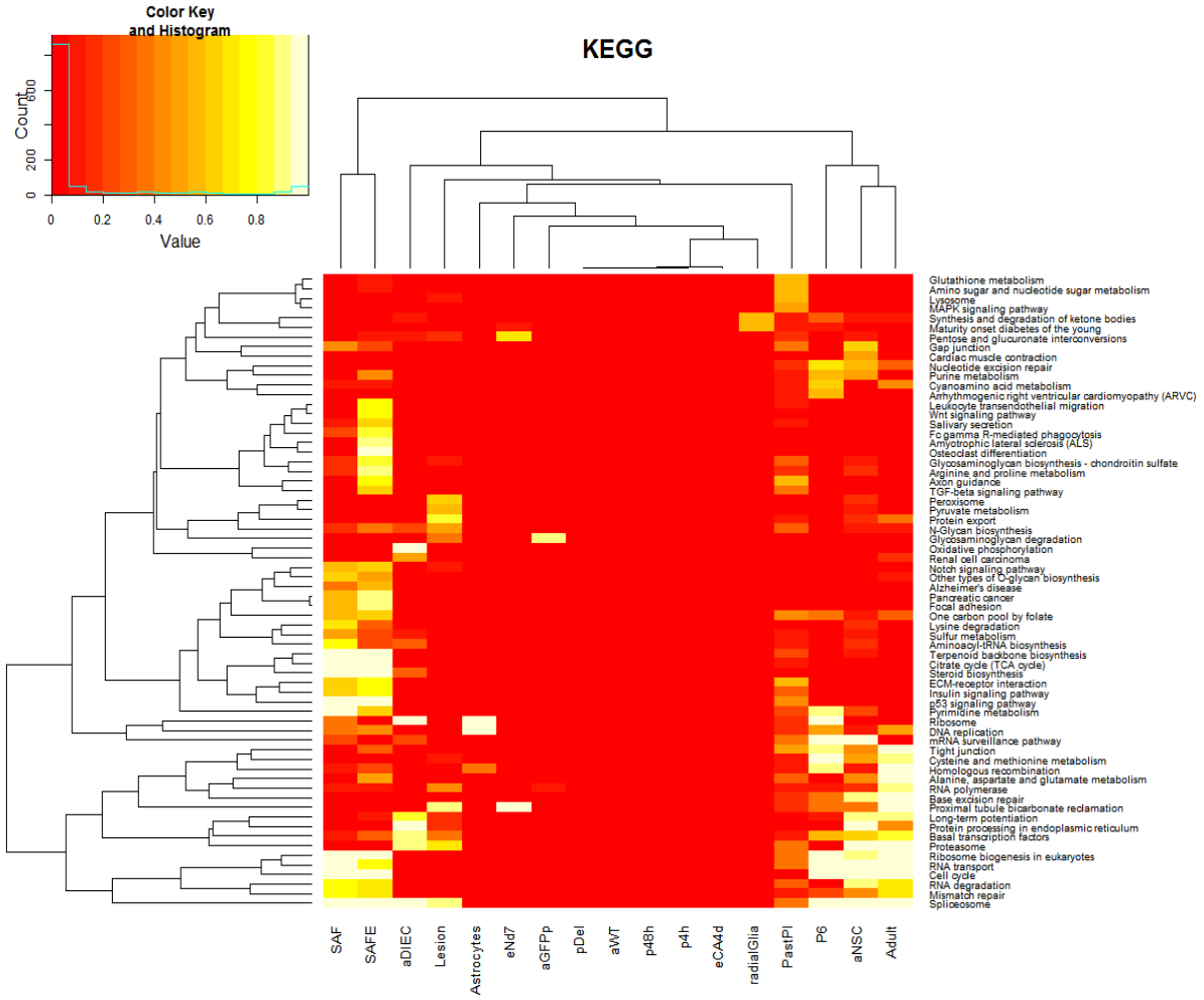


Fig. 36: KEGG-pathway estimations across the two datasets.
Similarities of 'SAF' and 'SAFE' to 'P6', 'NSC' and 'Adult' can be observed.

untreated astrocytes of P6 cell cultures independent of different time-points. The four shared terms were "intracellular signal transduction", "macromolecule localization", "cellular component organization or biogenesis" and "cellular metabolic process".

Besides the relationships shown for astrocytes treated with growth factors and **'P6'** astrocytes, only **'P6'** and **'Adult'** showed similarities, but no coherences between astrocytes and other cell types were observed with functional analyses of GO terms.

With KEGG pathways we could not identify a strong similarity between **'P6'** and growth-factor astrocytes (Fig. 36), where some pathways overlapped between the features and some did not. KEGG pathways of **'SAF'** and **'SAFE'** showed more similarities to a couple of **'aNSC'** and **'Adult'** pathways, but still many differed.

5. Conclusion

In this work we successfully managed to combine various project-datasets for a combined analysis of several cell types getting new insights into astrocytes. We combined four microarray expression projects and dealt with strong non-biological biases.

We implemented a pipeline, which we applied to both the growth and the combined dataset separately. The pipeline included statistical analyses like clustering of the samples showing a relationship between astrocytes of P6 cells taken at different time points and astrocytes from the diencephalon with NSC. A linear regression model was applied, after we captured the complex experimental and biological setup in a design matrix with nested covariates. As a result genes that are significant for the biological features were found. We used functional analysis to find enriched gene functions and pathways for the biological features. Using functional analysis, we were finally able to compare the growth factor and the combined dataset, which revealed a functional similarity between astrocytes treated with growth factors and astrocytes of post-natal six mice.

Additionally the t-statistic calculated in the linear regression model of the combined dataset was corrected using gene network information. In the original method t-statistics were calculated with a paired t-test. Therefore we extended the original method technically using a t-statistic matrix with several cofactors instead of a single factor. Finally, we added gene networks to the calculation. As a result, while several genes were remained to have a significant coefficient (but the number of hits shrank), several genes became significant. The newly identified genes were indeed ranked higher, due to their connections to significant genes. The network of the significant genes showed a big and some small clusters for astrocytes.

As we used several dependent biological factors for the linear regression, we finally suggested improving the t-statistics by adding information across the biological features, additional to the improvement across genes. Therefore we developed a few ideas like aggregation methods or extending the network with one node per gene.

For the growth factor dataset we could not observe any relationship between astrocytes and NSC. In addition **‘SAF’** and **‘SAFE’** could hardly be separated indicating a low influence of EGF. In the combined dataset, we could identify a few similarities between astrocytes and NSC, especially for diencephalic astrocytes. Functional analysis showed similarities of **‘SAF’** and **‘SAFE’** to **‘P6’** in biological processes. In the KEGG pathways regulations were similar between **‘P6’**, **‘Adult’** and **‘NSC’**.

In future it might be interesting to investigate “regression smoothing” further. Additionally a cross-analysis of different Affymetrix microarray platforms would be useful. In such a case it would be possible to make a direct comparison of the astrocytes treated with growth factors and the other astrocytes in the linear regression. Therefore characteristics of astrocytes could be detected more accurate. Additionally, the NSC from direct lines and neurospheres could be compared and a better comparison of astrocytes and NSC would be possible. Furthermore, NSCs and astrocytes of direct lines and cell cultures would be included in one dataset removing the need to add another dataset like the cells from the ESC line, which was done in this thesis.

6. Appendices

ComBat: Design Matrix

Sample	DIEC	NSC	GFPm	GFPp	WT	CAd4	CAd6	CAd8	Nd7	P14h	p4h	p24h	p48h	pdel
a1DIEC_1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
a1DIEC_2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
a1DIEC_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a1DIEC_7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
a1NSC_6	0	1	0	0	0	0	0	0	0	0	0	0	0	0
a1NSC_9	0	1	0	0	0	0	0	0	0	0	0	0	0	0
a1NSC_12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a1NSC_15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a2GFPm_10	0	0	1	0	0	0	0	0	0	0	0	0	0	0
a2GFPm_33	0	0	1	0	0	0	0	0	0	0	0	0	0	0
a2GFPp_08	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a2GFPp_09	0	0	0	1	0	0	0	0	0	0	0	0	0	0
a2GFPp_35	0	0	0	1	0	0	0	0	0	0	0	0	0	0
a2WT_2	0	0	0	0	1	0	0	0	0	0	0	0	0	0
a2WT_3	0	0	0	0	1	0	0	0	0	0	0	0	0	0
a2WT_4	0	0	0	0	1	0	0	0	0	0	0	0	0	0
eCAd4_1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
eCAd4_2	0	0	0	0	0	1	0	0	0	0	0	0	0	0
eCAd4_3	0	0	0	0	0	1	0	0	0	0	0	0	0	0
eCAd6_1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
eCAd6_2	0	0	0	0	0	0	1	0	0	0	0	0	0	0
eCAd6_3	0	0	0	0	0	0	1	0	0	0	0	0	0	0
eCAd8_1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
eCAd8_2	0	0	0	0	0	0	0	1	0	0	0	0	0	0
eCAd8_3	0	0	0	0	0	0	0	1	0	0	0	0	0	0
eNd7_2	0	0	0	0	0	0	0	0	1	0	0	0	0	0
eNd7_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eNd7_5	0	0	0	0	0	0	0	0	1	0	0	0	0	0
eP14h_2b	0	0	0	0	0	0	0	0	0	1	0	0	0	0
eP14h_3	0	0	0	0	0	0	0	0	0	1	0	0	0	0
eP14h_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p04h_83	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p04h_220	0	0	0	0	0	0	0	0	0	0	1	0	0	0
p04h_262	0	0	0	0	0	0	0	0	0	0	1	0	0	0
p04h_269	0	0	0	0	0	0	0	0	0	0	1	0	0	0
p24h_34	0	0	0	0	0	0	0	0	0	0	0	1	0	0
p24h_75	0	0	0	0	0	0	0	0	0	0	0	1	0	0
p24h_89	0	0	0	0	0	0	0	0	0	0	0	1	0	0
p24hdel_180	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p24hdel_193	0	0	0	0	0	0	0	0	0	0	0	0	0	1
p24hdel_200	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p48h_34	0	0	0	0	0	0	0	0	0	0	0	0	1	0
p48h_71	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Table 10: Design matrix for ComBat. Columns represent the covariates/coefficients and rows the samples.

Pearson correlation coefficient and Euclidean distances

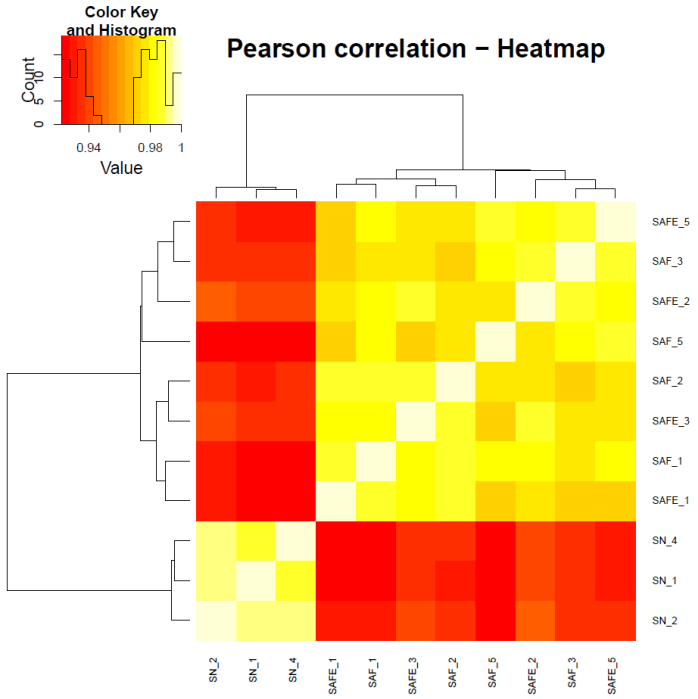


Fig. 37: Hierarchical clustering of the Pearson correlation coefficient of the samples of the growth-factor-dataset.

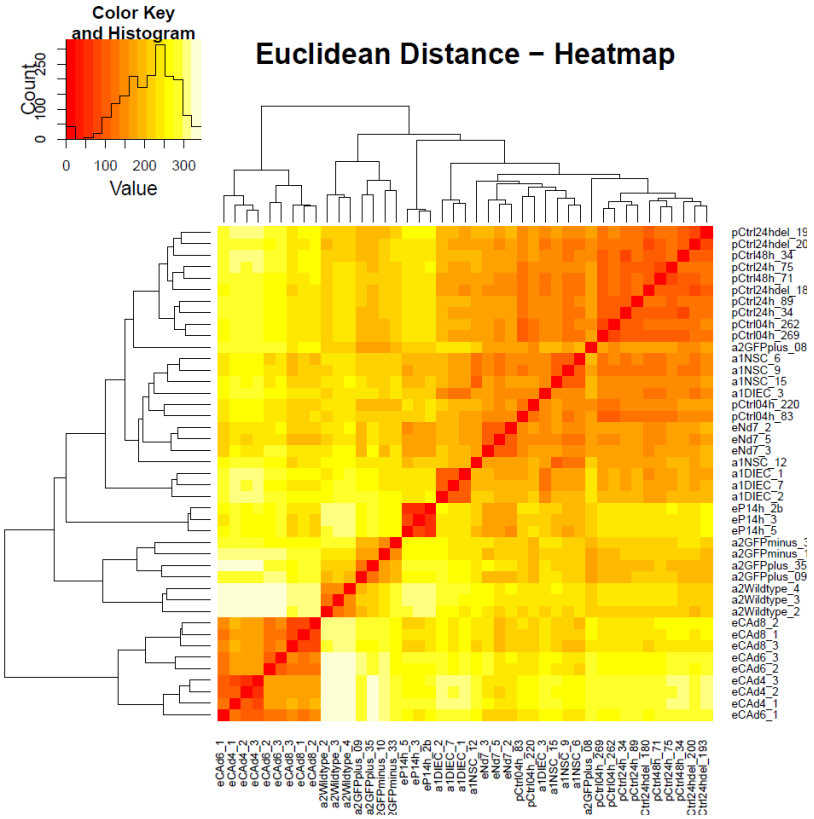
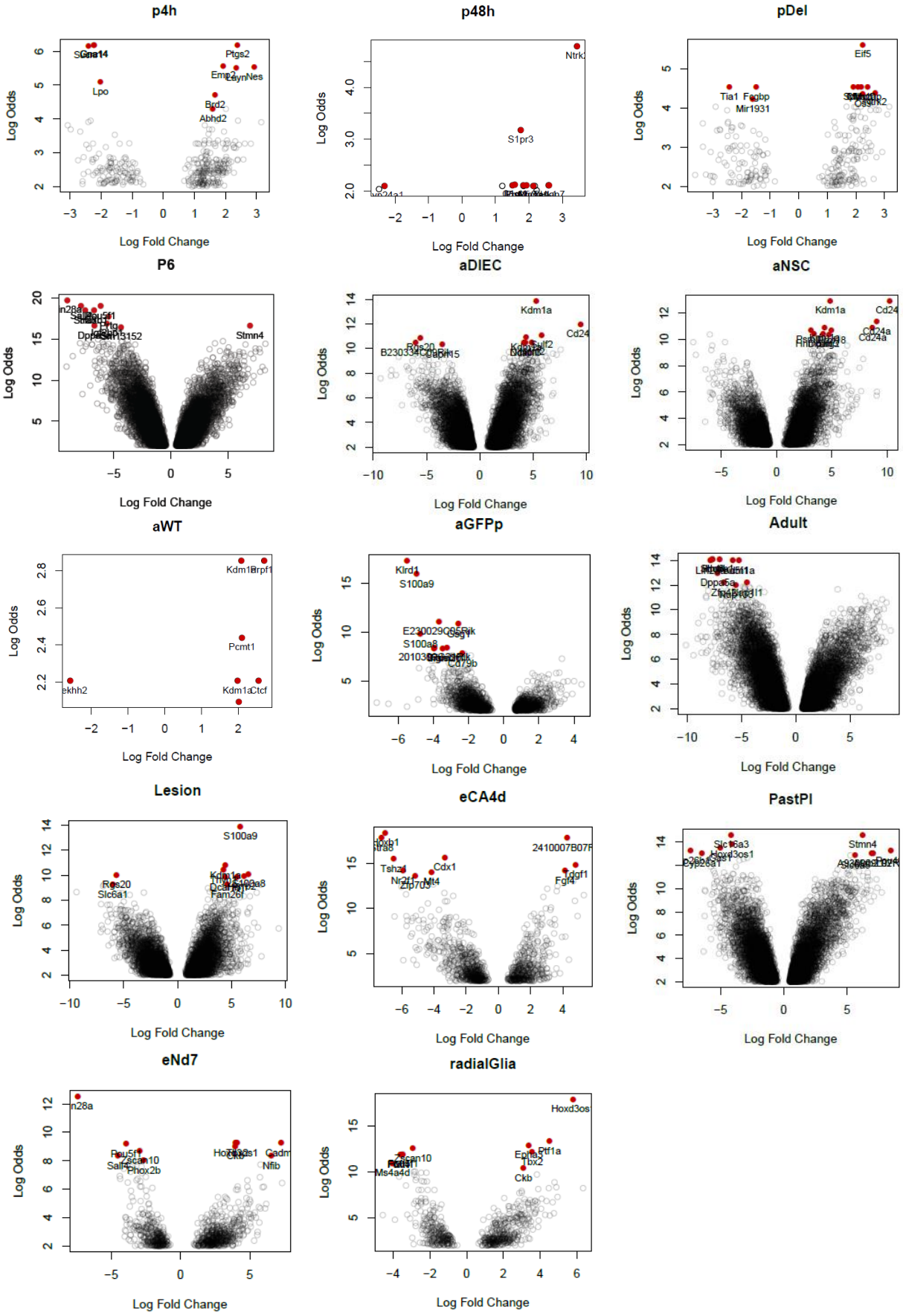


Fig. 38: Hierarchical clustering of the Euclidean distances of the samples of the combined-dataset.

Volcanoplots & Scatterplots of combined Dataset



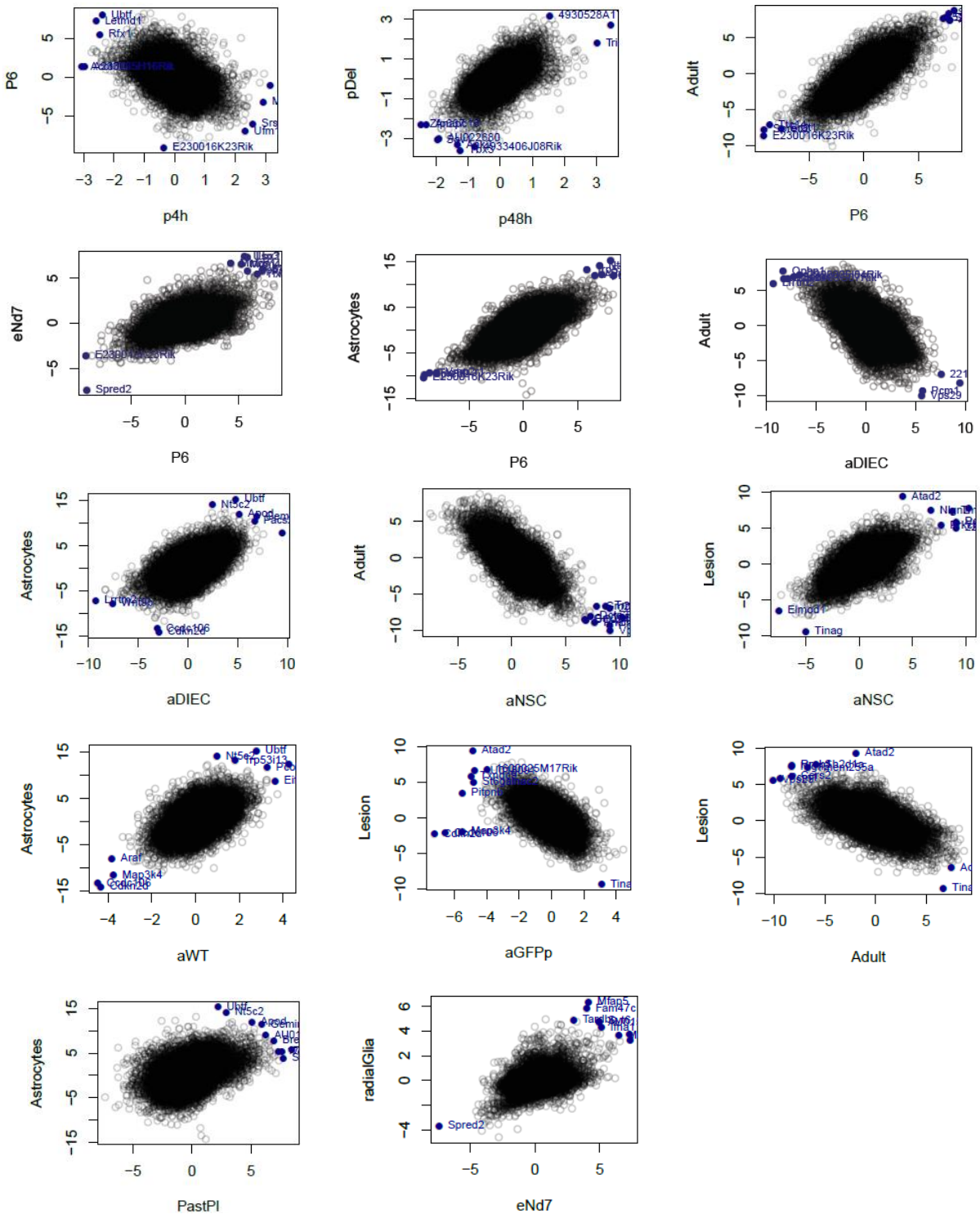


Fig. 39 : Volcano- and Scatterplots for biological factors of the combined-dataset.

MGSA – List of KEGG pathways for the growth factor dataset

SAF	SAFE
Ribosome biogenesis in eukaryotes	Spliceosome
Spliceosome	Ribosome biogenesis in eukaryotes
p53 signaling pathway	Steroid biosynthesis
Steroid biosynthesis	Cell cycle
RNA transport	p53 signaling pathway
Cell cycle	Citrate cycle (TCA cycle)
Pyrimidine metabolism	Osteoclast differentiation
Terpenoid backbone biosynthesis	Terpenoid backbone biosynthesis
Citrate cycle (TCA cycle)	Pancreatic cancer
Aminoacyl-tRNA biosynthesis	Focal adhesion
Mismatch repair	Arginine and proline metabolism
RNA degradation	Amyotrophic lateral sclerosis (ALS)
Lysine degradation	Glycosaminoglycan biosynthesis - chondroitin sulfate
Other types of O-glycan biosynthesis	Fc gamma R-mediated phagocytosis
ECM-receptor interaction	Leukocyte transendothelial migration
Insulin signaling pathway	Axon guidance
One carbon pool by folate	Insulin signaling pathway
Notch signaling pathway	RNA transport
Pancreatic cancer	ECM-receptor interaction
Focal adhesion	Wnt signaling pathway
Sulfur metabolism	RNA degradation
	Mismatch repair
	Salivary secretion
	Notch signaling pathway
	One carbon pool by folate
	Pyrimidine metabolism
	TGF-beta signaling pathway
	Alzheimer's disease

Table 11: KEGG pathways for features ‘SAF’ and ‘SAFE’ of the growth-factor-dataset

MGSA - GO-tables and KEGG-tables of the combined dataset

aDIEC	"estimate"	"term"
"GO:0044711"	0.9337274	single-organism biosynthetic process
"GO:0044260"	0.918305	cellular macromolecule metabolic process
"GO:0006996"	0.9174946	organelle organization
"GO:0051234"	0.6127546	establishment of localization

Adult	"estimate"	"term"
"GO:0006996"	0.9214686	organelle organization
"GO:0007268"	0.8876466	synaptic transmission
"GO:0033036"	0.7746822	macromolecule localization
"GO:0031175"	0.5411174	neuron projection development

GFPp	"estimate"	"term"
"GO:0016192"	0.8232686	vesicle-mediated transport

NSC	"estimate"	"term"
"GO:0006396"	0.9709584	RNA processing
"GO:0046907"	0.6008218	intracellular transport
"GO:0006996"	0.5268798	organelle organization

CA4d	"estimate"	"term"
"GO:0003002"	0.8394006	Regionalization
"GO:0045664"	0.5560274	regulation of neuron differentiation

Nd7	"estimate"	"term"
"GO:0048708"	0.8257066	astrocyte differentiation
"GO:0001709"	0.73124	cell fate determination

Past Plating	"estimate"	"term"
"GO:0044712"	0.9853062	single-organism catabolic process
"GO:0033554"	0.876392	cellular response to stress
"GO:0007268"	0.7010686	synaptic transmission
"GO:0022008"	0.6068308	Neurogenesis
"GO:0007010"	0.590417	cytoskeleton organization
"GO:0000278"	0.5223514	mitotic cell cycle
"GO:0035556"	0.509544	intracellular signal transduction

radialGlia	"estimate"	"term"
"GO:0045165"	0.7262164	cell fate commitment

Table 12: GO annotations for features of the combined-dataset.

aNSC	Adult
Spliceosome	Spliceosome
Protein processing in endoplasmic reticulum	Ribosome biogenesis in eukaryotes
Proteasome	RNA transport
RNA transport	Cell cycle
Cell cycle	Tight junction
mRNA surveillance pathway	Proteasome
Base excision repair	Homologous recombination
Ribosome biogenesis in eukaryotes	Alanine, aspartate and glutamate metabolism
Long-term potentiation	Proximal tubule bicarbonate reclamation
RNA degradation	Base excision repair
Basal transcription factors	Long-term potentiation
Gap junction	Cysteine and methionine metabolism
Nucleotide excision repair	RNA polymerase
Purine metabolism	Basal transcription factors
Cardiac muscle contraction	RNA degradation
Cysteine and methionine metabolism	Mismatch repair
	DNA replication

GFPp	radialGlia	Neuron
Glycosaminoglycan degradation	Maturity onset diabetes of the young	Proximal tubule bicarbonate reclamation
	Synthesis and degradation of ketone bodies	Pentose and glucuronate interconversions

aDiec	Lesion	PastPl
Ribosome	Proximal tubule bicarbonate reclamation	Axon guidance
Oxidative phosphorylation	Spliceosome	Glutathione metabolism
Spliceosome	Protein export	ECM-receptor interaction
Protein processing in endoplasmic reticulum	Proteasome	Lysosome
Proteasome	Peroxisome	Amino sugar and nucleotide sugar metabolism
Basal transcription factors	Pyruvate metabolism	MAPK signaling pathway
Long-term potentiation	N-Glycan biosynthesis	
Renal cell carcinoma		

Table 13: KEGG pathways for features of the combined-dataset.

7. References

[1] M. V. Sofroniew, "Molecular dissection of reactive astrogliosis and glial scar formation," *Trends Neuroscience, Vol. 32(12), 638–647,* 2009.

[2] M. Sild and E. S. Ruthazer, "Radial Glia: Progenitor, Pathway, and Partner," *The Neuroscientist, Vol. 17, No. 3, 288–302,* 2011.

[3] M. V. Sofroniew and H. V. Vinter, " Astrocytes: biology and pathology," *Acta Neuropathol (2010) 119:7-35,* 2009.

[4] J. W. Fawcett and R. A. Asher, "The glial scar and central nervous system repair," *Elsevier: Brain Research Bulletin, Vol. 49, No. 6, pp. 377–391,* 1999.

[5] I. Ulitsky, A. Maron-Katz, S. Shavit, D. Sagir, C. Linhart, R. Elkon, A. Tanay, R. Sharan, Y. Shiloh and R. Shamir, "Expander: from expression microarrays to networks and functions," *Nature Protocols, Vol.5, No.2,* 2010.

[6] A. Butte, "The use and analysis of microarray data," *Nature Reviews, doi:10.1038/nrd961,* 2002.

[7] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research,* 2015.

[8] S. Bauer, J. Gagneur and P. N. Robinson, "GOing Bayesian: model-based gene set analysis of genome-scale data," *Nucleic Acids Research, Vol. 38, No. 11, 3523–3532,* 2010.

[9] M. Hofree, J. P. Shen, H. Carter, A. Gross and T. Ideker, "Network-based stratification of tumor mutations," *Nature Methods, Vol.10, No.11,* 2013.

[10] Y. Cun and H. Fröhlich, "Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics," *PLOS ONE, Vol. 8, Issue 9, e73034,* 2013.

[11] W. E. Johnson, C. Li and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics, 8, 1, pp. 118–127, doi:10.1093/biostatistics/kxj037,* 2007.

[12] G. K. Smyth, "Limma: Linear Models for Microarray Data," *Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397–420.,* 2005.

[13] E. Suárez, A. Burguete and G. J. McLachlan, "Microarray Data Analysis for Differential Expression: a Tutorial," *PRHSJ Vol. 28 No. 2 review article,* 2009.

[14] D. Murphy, "Gene expression studies using microarrays: principles, problems, and prospects," *Advan in Physiol Edu, Vol. 26, 256-270,* 2002.

[15] A. E. Bishop, L. D. K. Buttery and J. M. Polak, "Embryonic stem cells," *J Pathol, Vol. 197, 424-429,* 2002.

[16] B. A. Barres, "The Mystery and Magic of Glia: A Perspective on Their Roles in Health and Disease," *Neuron 60, 430-440,* 2008.

[17] K. Campbell and M. Götz, "Radial glia: multi-purpose cells for vertebrate brain development," *TRENDS in Neurosciences Vol.25 No.5 ,* 2002.

[18] D. Clarke, "Neural stem cells," *Bone Marrow Transplantation 32, S13–S17,* 2003.

[19] M. Götz and W. B. Huttner, "The cell biology of neurogenesis," *Nature Reviews, Molecular Cell Biology, Vol. 6, 777-788,* 2005.

[20] K. Kang and M.-R. Song, "Diverse FGF receptor signaling controls astrocyte specification and proliferation," *Biochemical and Biophysical Research Communications 395, 324–329,* 2010.

[21] E. Hartfuss, R. Galli, N. Heins and M. Götz, "Characterization of CNS Precursor Subtypes and Radial Glia," *Developmental Biology 229, 15–30,* 2001.

[22] E. D. Laywell, P. Rakic, V. G. Kukekov, E. C. Holland and D. A. Steindler, "Identification of a multipotent astrocytic stem cell in the immature and adult mouse brain," *PNAS, Vol. 97, No. 25, 13883–13888,* 2000.

[23] A. Bez, E. Corsini, D. Curti, M. Biggiogera, A. Colombo, R. F. Nicosia, S. F. Pagano and E. A. Parati, " Neurosphere and neurosphere-forming cells: morphological and ultrastructural characterization," *Brain Research 993 18–29,* 2003.

[24] R. Beckervordersandforth, P. Tripathi, J. Ninkovic, E. Bayam, A. Lepier, B. Stempfhuber, F. Kirchhoff, J. Hirrlinger, A. Haslinger, D. C. Lie, J. Beckers, B. Yoder, M. Irmler and M. Götz, "In Vivo Fate Mapping and Expression Analysis Reveals Molecular Hallmarks of Prospectively Isolated Adult Neural Stem Cells," *Cell Stem Cell 7, 744–758,* 2010.

[25] J. B. Jensen and M. Parmar, "Strengths and Limitations of the Neurosphere Culture System," *Molecular Neurobiology, Vol 34,* 2006.

[26] M. A. Caldwell, X. He, N. Wilkie, S. Pollack, G. Marshall, K. A.Wafford and C. N. Svendsen, "Growth factors regulate the survival and fate of cells derived from human neurospheres," *Nature biotechnology, Vol 19,* 2001.

[27] S. C. Noctor, A. C. Flint, T. A. Weissman, R. S. Dammerman and A. R. Kriegstein, "Neuronsderived from radial glial cells establish radial units in neocortex," *NATURE, Vol. 309 (letters to nature),* 2001.

[28] G. Chanas-Sacre, B. Rogister, G. Moonen and P. Leprince, "Radial Glia Phenotype: Origin, Regulation, and Transdifferentiation," *Journal of Neuroscience Research, Vol 61, 357–363 ,* 2000.

[29] P. Malatesta, M. A. Hack, E. Hartfuss, H. Kettenmann, W. Klinkert, F. Kirchhoff and M. Götz, "Neuronal or Glial Progeny: Regional Differences in Radial Glia Fate," *Neuron, Vol. 37, 751–764,* 2003.

[30] T. E. Anthony, C. Klein, G. Fishell and N. Heintz, "Radial Glia Serve as Neuronal Progenitors in All Regions of the Central Nervous System," *Neuron, Vol. 41, 881–890,* 2004.

[31] F. T. Merkle, A. D. Tramontin, J. M. García-Verdugo and a. A. Alvarez-Buylla, "Radial glia give rise to adult neural stem cells in the subventricular zone," *PNAS, Vol. 101, No. 50, 17528–17532,* 2004.

[32] T. Glaser and O. Brüstle, "Retinoic acid induction of ES-cell-derived neurons: the radial glia connection," *TRENDS in Neurosciences Vol.28 No.8,* 2005.

[33] M. Götz and Y.-A. Barde, "Radial Glial Cells: Defined and Major Minireview Intermediates between Embryonic Stem Cells and CNS Neurons," *Neuron, Vol. 46, 369–372,* 2005.

[34] S. Sirko, G. Behrendt, P. A. Johansson, P. Tripathi, M. R. Costa, S. Bek, C. Heinrich, S. Tiedt, D. Colak, M. Dichgans, I. R. Fischer, N. Plesnila, M. Staufenbiel and C. Ha,

"Reactive Glia in the Injured Brain Acquire Stem Cell Properties in Response to Sonic Hedgehog," *Cell stem cell 12, 426–439,* 2013.

[35] J. McGraw, G. Hiebert and J. Steeves, "Modulating Astrogliosis After Neurotrauma," *Journal of Neuroscience Research 63:109–115 ,* 2001.

[36] T.-k. Shin, Y.-d. Lee and K.-b. Sim, "Embryonic Intermediate Filaments, Nestin and Vimentin, Expression in the Spinal Cords of Rats with Experimental Autoimmune Encephalomyelitis," *Journal of Veterinary Science, Vol. 4, No. 1, 9-13,* 2003.

[37] M. Murphy, K. Reid, R. Dutton, G. Brooker and P. F. Bartlett, "Neural Stem Cells," *NEURAL STEM CELLS, Vol 2, No 1 ,* 1997.

[38] D. M. Ornitz and N. Itoh, "Fibroblast growth factors," *Genome Biology, 2(3):reviews 3005.1–3005.12,* 2001.

[39] N. Itoh and D. M. Ornitz, "Functional Evolutionary History of the Mouse Fgf Gene Family," *Developmental Dynamics 237:18–27,* 2008.

[40] B. Thisse and C. Thisse, "Functions and regulations of fibroblast growth factor signaling during embryonic development," *Developmental Biology 287, 390–402,* 2005.

[41] B. A. Reynolds and S. Weiss, "Generation of Neurons and Astrocytes from Isolated Cells of the Adult Mammalian Central Nervous System," *SCIENCE, VOL. 255,* 2014.

[42] S. I. Mayer, O. G. Rössler, T. Endo, P. Charnay and G. Thiel, "Epidermal-growth-factor-induced proliferation of astrocytes requires Egr transcription factors," *Journal of Cell Science 122, 3340-3350,* 2009.

[43] B. A. Reynolds, W. Tetzlaff and S. Weiss, "A Multipotent EGF-Responsive Striatal Embryonic Progenitor Cell Produces Neurons and Astrocytes," *The Journal of Neuroscience, 12(11): 4565-4574,* 1992.

[44] F. Ciccolini and C. N. Svendsen, "Fibroblast Growth Factor 2 (FGF-2) Promotes Acquisition of Epidermal Growth Factor (EGF) Responsiveness in Mouse Striatal Precursor Cells: Identification of Neural Precursors Responding to both EGF and FGF-2," *The Journal of Neuroscience, 18(19):7869–7880,* 1998.

[45] P. Ahlgren and B. Jarneving, "Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient," *Journal of the american society for information science and technology, Vol. 54, No. 6, 550–560,* 2003.

[46] D. B. Allison, X. Cui, G. P. Page and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *NATURE REVIEWS, GENETICS, Vol. 7 ,* 2006.

[47] W. S. Noble, "How does multiple testing correction work?," *Nature Biotechnology, Vol 27, No 12 ,* 2009 .

[48] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics, Vol. 25,* 2000.

[49] The Gene Ontology Consortium, "Gene Ontology Consortium: going forward," *Nucleic Acids Research,* 2014.

[50] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research, Vol. 42, D199–D205,* 2014.

[51] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research, Vol. 28, No. 1, 27-30,* 2000.

[52] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research, Vol.41, D808–D815,* 2013.

[53] B. Snel, G. Lehmann, P. Bork and M. A. Huynen, "STRING: a webserver to retrieve and display the repeatedly occuring neighborhood of a gene," *Nucleic Acids Research, Vol. 28, No. 18,* 2000.

[54] J. Lovén, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee and R. A. Young, "Revisiting Global Gene Expression Analysis," *Cell 151,* 2012.

[55] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research, Vol. 31, No. 4,* 2003.

[56] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, C. Zhao, F. Elloumi, W. Shi, R. Thomas, S. Lin, G. Tillinghast, G. Liu, Y. Zhou, D. Herman, Y. Li, Y. Deng, H. Fang, P. Bushel, M. Woods and J. Zhang, "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data," *The Pharmacogenomics Journal, 10, 278–291,* 2010.

[57] P. Kupfer, R. Guthke, D. Pohlers, R. Huber, D. Koczan and R. W. Kinne, "Batch correction of microarray data substantially improves the identification of genes differentially expressed in Rheumatoid Arthritis and Osteoarthritis.," *BMC Medical Genomics 5:23,* 2012.

[58] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin and C. Liu, "Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch," *Adjustment Methods. PLoS ONE, Volume 6, Issue 2, e17238,* 2011.

[59] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews | Genetics, Volume 11, 733,* 2010.

[60] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y.Weiss-Sols, R. Duque, H. Bersini and A. Nowe, "Batch effect removal methods for microarray gene expression data integration: a survey," *BRIEFINGS IN BIOINFORMATICS. page 1 of 22; doi:10.1093/bib/bbs037,* 2012.

[61] S. Grossmann, S. Bauer, P. N. Robinson and M. Vingron, "Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis," *Bioinformatics, Vol. 23, no. 22, 3024-3031,* 2007.

[62] M. Z. Man, X. Wang and Y. Wang, "POWER_SAGE: comparing statistical tests for SAGE experiments," *Bioinformatics, Vol. 16, No. 11, 953-959,* 2000.

[63] S. Bauer, P. N. Robinson and J. Gagneur, "Model-based gene set analysis for Bioconductor," *Bioinformatics Applications Note, Vol. 27, No. 13, pages 1882–1883,* 2011.

[64] S. Sass, F. Buettner, N. S. Mueller and F. J. Theis, "A modular framework for gene set analysis integrating multilevel omics data," *Nucleic Acids Research, 1–12,* 2013.

[65] Y. Cun and H. Fröhlich, "netClass: An R-package for network based, integrative

biomarker signature discovery," *Bioinformatics,* 2014.

[66] C. Hummert, F. Mech, F. Horn, M. Weber, S. Drynda, U. Gausmann and R. Guthke, "Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays," *Int'l Conf. Bioinformatics and Computational Biology, BIOCOMP,* 2011.