LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Chair M12: Mathematical modelling of biological systems

Bachelor Thesis
in Bioinformatics

Gene signatures of epithelial-mesenchymal transition in embryonic stem cells

Johannes Höffler

Supervisor:	Prof. Dr. Fabian J. Theis
Advisor:	Martin Preusse
Submission Date:	15. April 2015

I assure that this thesis is a result of my personal work and that no other than the indicated aids have been used for its completion. Furthermore I assure that all quotations and statements that have been inferred literally or in a general manner from published or unpublished writings are marked as such. Beyond this I assure that the work has not been used, neither completely nor in parts, to pass any previous examination.

April 13, 2015

Name

Abstract

Scientists have made huge progress in the area of stem cell research, but there are still countless aspects, which remain to be discovered. Cell differentiation takes place in the gastrulation process, forming the three primitive germ layers. The endoderm is the inner primitive germ layer from which cell development will form epithelial cells. The middle germ layer is represented by the mesoderm. The last of the three layers, the ectoderm, is the outer layer of the three primitive germ layers of an embryo. The epithelial-mesenchymal transition is the process during which cells lose their epithelial characteristics, gain a lesser regular appearance and get the mesenchymal migratory properties, by dissolving the epithelial cell-to-cell adhesion. The epithelial and mesenchymal phenotypes are not irreversible, they convert under specific conditions between those two phenotypes. This transition process is called EMT - the epithelial-mesenchymal transition. The reverse EMT process is called the mesenchymal-epithelial-transition process (MET), which also occurs during embryonic stem cell development. These two processes are critical for the appropriate morphogenesis of the organs.
We try to understand the differentiation process, more precisely how the split between endoderm and mesoderm is regulated. This split seems to be a epithelial-mesenchymal respectively mesenchymal-epithelial transition process. During in mouse embryo live imaging, endodermal cells give the appearance to transform from epithelial to mesenchymal cells and than to retransform back to epithelial cells. As this EMT-MET process doesn't express EMT core signaling factors, endodermal cells don't seem to go through the typical EMT followed by MET.

CD24 was used to subclassify whithin mesodermal and endodermal stem cells. Indeed we observe that CD24 clarifies whether the stem cell is in an early or late developmental stage, independently of the time of in vitro measurement. After data normalization and preprocessing, we performed a principal component analysis (PCA), which is a mathematical way of identifying patterns in data and expressing the data, highlighting their similarities and differences. After that, a regression model was designed and calculated, resulting in p-values for every expressed gene, indicating their regulatory activity.

We were able to dissect gene regulation from mouse ESC differentiating to endoderm and mesoderm. We calculated the most regulatory active genes not only for early and late embryonic developmental stages, benefitting from the sub segregation using the CD24 as marker, but also for mesodermal and endodermal stem cells, using Foxa2 und T as markers. We found out, that Foxa2 and T positive samples behave like late endodermal cells and that CD24 indeed distinguish early and late endoderm and not as assumed in vitro measurement time dependent embryonic stem cell development.

Zusammenfassung

Wissenschaftler haben grosse Fortschritte auf dem Gebiet der Stammzellforschung gemacht, aber es gibt immer noch zahllose Aspekte, die es zu entdecken gilt. Zelldifferenzierung, die in dem Gastrulationsprozess stattfindet, bildet die drei primitiven Keimblätter. Das Endoderm entsteht aus dem inneren primitiven Keimblatt, aus dessen Zellentwicklung sich Epithelzellen bilden. Das mittlere Keimblatt wird zum Mesoderm. Das letzte der drei Keimblätter, das Ektoderm, ist die Aussenschicht der drei primitiven Keimblätter eines Embryos. Die epithelial-mesenchymale Transition ist der Prozess, bei dem Zellen ihre epithelialen Charakteristika verlieren, um eine weniger rigide Form zu gewinnen, indem sie die epithelialen Zell-zu-Zell-Verbindungen lösen und ihre mesenchymalen Eigenschaften zur Migration in womöglich andersartiges Gewebe nutzen. Die epithelialen und mesenchymalen Phänotypen sind nicht irreversibel, die Zellen transformieren unter bestimmten Bedingungen zwischen diesen beiden. Dieser Transformationsprozess wird als EMT - der epithelial-mesenchymale Transition beschrieben. Den EMT Vorgang in gegenläufiger Richtung nennt man den mesenchymalen-epithelialen-Übergangsprozess (MET), welcher auch während der embryonalen Stammzellentwicklung auftritt. Diese beiden Prozesse sind kritisch für die entsprechende Morphogenese der Organe.

Wir versuchen die Differenzierung zu verstehen, genauer gesagt, wie die Spaltung zwischen Entoderm und Mesoderm reguliert wird. Diese Spaltung scheint ein epithelial-mesenchymaler bzw. mesenchymal-epithelialer Übergangsprozess zu sein. Live-Bildgebungen im Mausembryo erwecken bei Entodermzellen den Anschein, sich von epithelialen zu mesenchymalen Zellen und wieder in epitheliale Zellen zu verwandeln. Da dieser EMT-MET-Prozess nicht die Kern-Signalfaktoren exprimieret, scheinen die Entodermzellen nicht der angenommenen EMT-MET Umwandlung zu folgen. CD24 wurde verwendet, um innerhalb mesodermaler und endodermaler Stammzellen Subgruppierungen aufzuspüren. Nach der Datennormalisierung und Präprozessierung, wurde eine Hauptkomponentenanalyse (PCA) durchgefürt, einer mathematische Analyse zur Feststellung von Mustern in Daten und zur Hervorhebung ihrer Gemeinsamkeiten und Unterschiede. Danach wurde ein Regressionsmodell konstruiert und berechnet, wodurch die p-Werte für jedes exprimiertes Gens berechnet wurden, die deren regulatorische Aktivität widerspiegeln.

Wir konnten die Genregulation der Differenzierung von embryonalen Stammzellen hin zu Endoderm und Mesoderm isoliert beobachten. Wir berechneten die regulatorische Aktivität der Gene, nicht nur für die frühen und späten embryonalen Entwicklungsstadien, sondern auch für mesodermale und endodermale Stammzellen. Wir fanden heraus, dass sich Foxa2 und T positive Proben genauso verhalten, wie späte Entodermzellen. CD24 unterscheidet in der Tat zwischen frühem und spätem Endoderm und nicht wie ursprünglich angenommen embryonalen Stammzellentwicklung nach in-vitro-Messpunkten sortiert.

Contents

1. Introduction

In 1909 the term 'stem cell' was first used in a scientific way by the Russian histologist Alexander Maksimov at the congress of the hematologic society in Berlin, postulating the existence of hematopoietic stem cells. It was published in the folia haematologica 'The lymphocyte as a stem cell, common to different blood elements in embryonic development and during the post-fetal life of mammals' [1]. Nowadays, the term 'stem cell' describes unspecialized cells, which have the ability to renew themselves indefinitly and become tissue specific cells with their unique functions.

From 1909 to the present day, scientists have made huge progress in this area of research, but there are still countless aspects, which remain to be discovered. Stem cell research is driven by opportunities for developing new medical therapies [2], e.g. a cure of diabetes by studying the pancreas development, new cancer treatments, etc., as well as fairytale-like dreams of an everlasting self-renewing life. The diversity of all species including genetically similar individuals, arises during embryogenesis. Gaining more knowledge about this process is one of the most important aims of modern developmental biology. The term 'adult stem cells', describes undifferentiated cells found in specialized tissue, which have the ability to renew themselves and convert into nearly all cell types, from which the tissue originated [3]. This self-renewal property might be causing the involvement of stem cells in processes of carcinogenesis. Although stem cell research seems to promise a lot of yet unknown benefits, the research itself is controversial, given the different ethical views of society as well as the legal classification of the early embryo. Limiting the research only to mice or other model organisms is not sufficient due to their substantial biological differences.

1.1. Gastrulation

The word gastrulation is derived from the latin word 'gaster' which can be translated as 'guts', implying 'the formation of the guts'. The term 'gastrulation' was first used by Ernst Haeckel 1872-1877 in his 'Studies for the Gastrae Theory'[4]. Primarily, it is a stage of the common embryonic cell development. The overall aim of gastrulation is the manifestation of the three primitive germ layers. In this thesis the mouse 'mus musculus' serves as model organism. Keep in mind that embryonic human development is slightly different to the rodent's one, although main processes are very similar.

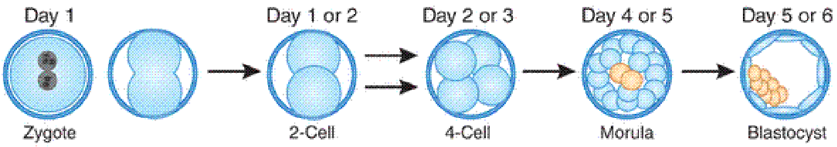


Figure 1.1.: This figure shows the development of a fertilized egg, starting as a zygote and eventually developing into a blastocyst [5].

Beginning as a zygote, the embryo development starts with multiple cleavages until it becomes a 'morula' after four or five days. Morula not only describes the developing zygote, but also the embryonal developmental stage. This cell consists of 8 to 32 blastomeres, emerging from the previous cleavage divisions. The next step of embryogenesis is the stage of blastulation, in which the morula transforms into a 'blastocyst'. Its outer cell layer, also called 'trophectoderm' or 'trophoblast', is a cavity called 'blastocoel' filled with fluid, and in its interior is a cluster of cells called the 'inner cell mass' (ICM), containing approximately 30 pluripotent stem cells. Stem cells with the ability to differentiate into cell types beyond those of the own tissue cell type are referred to as 'pluripotent'. The outer layer counts approximately 70 trophoblastial cells [6].

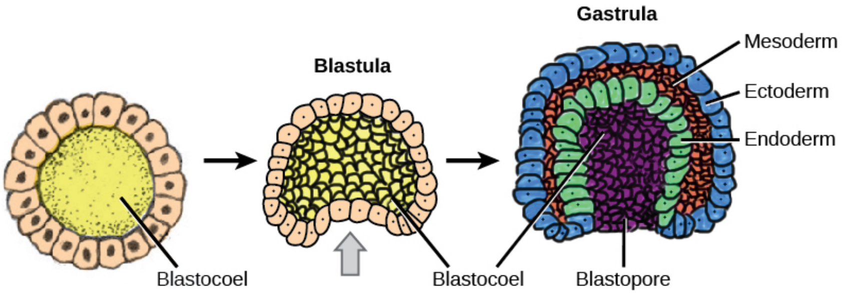


Figure 1.2.: Generalised gastrulation process, showing the main developmental stages [7].

Gastrulation begins shortly after a blastocyst containing the 'blastocoel' implants itself into the uterine wall of the mother, followed by the organogenesis. A 'blastula' emerges from that process. The term 'blastula' originally means a different stage of the embryonically development, but the term is often used to describe both. The blastocoel is composed of the inner cell mass, surrounded by a layer of blastomeres. At about four and a half days post coitum (DPC) the blastula implants itself into the uterine wall [6]. Two days later, at six and a half DPC, the gastrulation process starts. The polar trophectoderm and the ICM convert into the extra-embryonic ectoderm, the epiblast and a layer of visceral endoderm. Gastrulation starts by forming the primitive streak, through which epiblast cells ingress to form the mesoderm and endoderm. Mesoderm, endoderm and the ectoderm, representing descendants of epiblast cells that did not pass the primitive streak constitute the primary germ layers [6]. A part of the cell differentiation process is called the 'epithelial-mesenchymal transition'. Two important genes are 'Brachyury' and 'Foxa2'. They can be fused with fluorescence tags and provide thus an intracellular information regarding the endoderm/mesoderm differentiation, as Foxa2 is expressed in endodermal and Brachyury in mesodermal differentiated stem cells.

1.1.1. Brachyury

In 1927, Nadine Dobrovolskaia-Zavadskaia discovered this gene, located within the T-box complex of genes. She named it Brachyury due to the greek words 'brakhus' meaning short and 'oura', the tail - as it caused short tails or the death of mice, when mutated. Nowadays, Brachyury has the gene name 'T', first cloned in 1990 by B. Herrmann et al., encoding 436 amino acids, which binds with a specific DNA section, called the T-box, consisting of 18 T-box genes [8].

Brachyury is a nuclear tissue-specific transcription factor that is expressed during embryogenesis, mostly in the notochord and also in other embryonal and extraembryonal tissues [9]. Genetic and molecular embryonic development studies have demonstrated its importance in regulating cell fate decisions that establish the early body plan, and in later processes underlying organogenesis [10]. Mutant alleles of Brachyury have been isolated and they have an effect on the development of the mesoderm and its derivatives [11]. T is the founding member of the T-box family of transcription factors, often used as a marker of the primitive streak, nascent mesoderm, and other tissues [11].

1.1.2. Foxa2

The fork-head Foxa family of transcription factors, encoded by three genes, regulates haptic and/or pancreatic gene expression [12] and is selectively expressed in respiratory epithelial cells. It also plays a critical role in suppressing Th2-mediated pulmonary inflammation and goblet-cell metaplasia in the developing murine lung during postnatal development [13].

Foxa2 is the first of the three family members to be expressed in the embryo prior to gastrulation. All are expressed in the embryonic endoderm cells that constitute the precursor cells for all organs forming the guts, and all of them remain present and active in the adult liver [14].

Targeted disruption of Foxa2 resulted in embryonic lethality with defective development of the foregut endoderm, from which the liver and pancreas arise [12]. Foxa2 is required for epithelial differentiation, and its knockout causes goblet-cell metaplasia and Th2-mediated pulmonary inflammation. These mechanisms are not yet fully known [13]. Also, Foxa2 acts as an initiating factor in the earliest stages of liver specification during embryonic development. It plays a major role in glucose and lipid metabolism of adult liver cells and can also open compacted chromatin, allowing for the activation of transcription from silenced genes [14].

1.2. Cell differentiation

Cell differentiation takes place in the gastrulation process, forming the three primitive germ layers. Nowadays it is possible to remove the ICM-cells from embryonic stem cells (ESC) and cultivate them in order to investigate the complex differentiation mechanism. Scientists can force them to differentiate into a specific major germ layer.

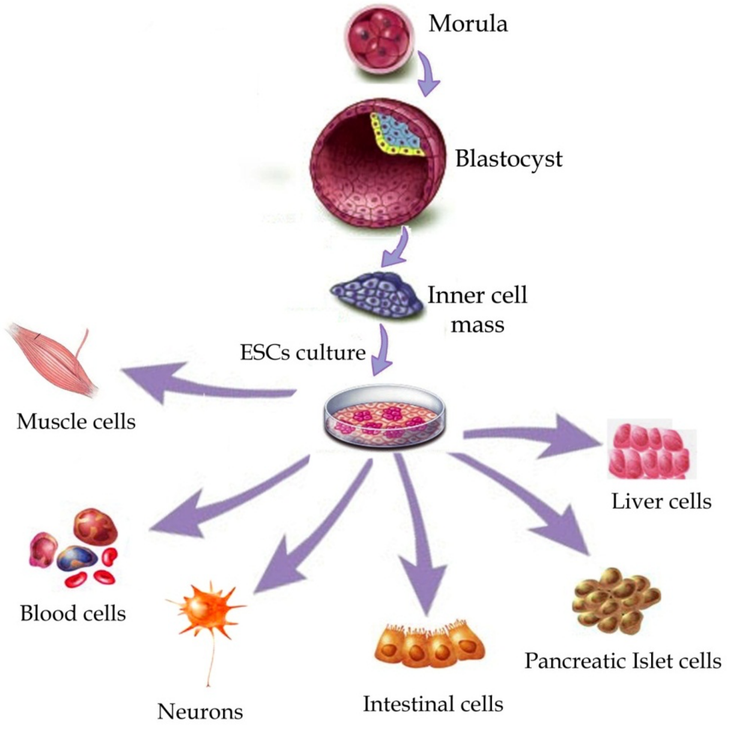


Figure 1.3.: Schematized process of the embryonic stem cell differentiation [15].

The main gastrulation process starts with the formation of the 'primitive node' on the posterior side of the epiblast (Fig. 1.6: C: 'Primitivknoten'). This node is the first indication for the localisation of the head to tail regions (anterior - posterior polarity). The primitive node consists of cells secreting cellular signals, which helps cells to migrate within the embryo during gastrulation. Out of this node the 'primitive streak' emerges. It is a recess from the primitive node towards the ventral side of the embryo. The elongation of this primitive streak induces an ingression of epiblast cells into the primitive streak. Once moved through, the cells become mesendoderm and start to cover the outside of the embryo.

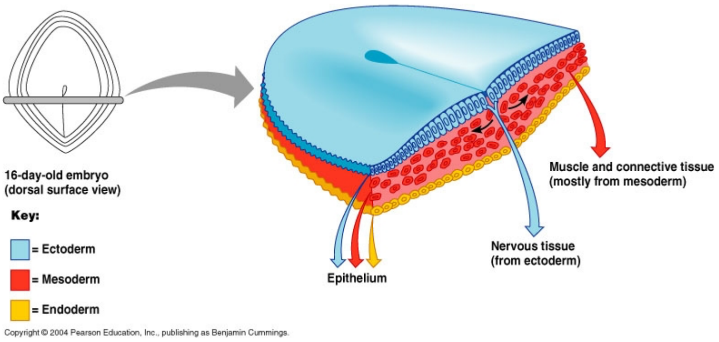


Figure 1.4.: Dorsal surface view of a 16-day-old embryo, showing the composition of the epiblast [16].

The endoderm is the inner primitive germ layer from which cell development will form epithelial cells, giving birth to lungs, the gastrointestinal tract, liver, pancreas and the urinary bladder. The middle germ layer is represented by the mesoderm. It is responsible for the formation of the cardiovascular system, blood cells and bone marrow, the skeleton, muscles and parts of the reproductive and excretory system. The last of the three layers, the ectoderm, is the outer layer of the three primitive germ layers of an embryo. Skin, hair, nails, nerve and brain cells are developed within this layer [17].

1.3. Epithelial-mesenchymal transition

The epithelial-mesenchymal transition is the process during which cells lose their epithelial characteristics, gain a lesser regular appearance and get the mesenchymal migratory properties, by dissolving the epithelial cell-to-cell adhesion.

The terms 'epithelium' and 'mesenchyme' need to be introduced, to understand the process of epithelial-mesenchymal transition: the epithelium is a collective name for glandular tissue, one of the four basic animal tissue types. These cells secrete bodily products. It first forms a sheet of cells that are connected to each other by cell to cell junctions [18]. This sheet is polarised, giving birth to a basal (bottom) and apical (top) region and consists of several layers. As there are several different types of epithelial cells, they are classified by the number of the layers and the shape of the outlining cells on its surface.

The mesenchyme is embryonal connective tissue. It consists of several different cell types. Mesenchymal cells have the ability to develop into tissues of the lymphatic or the circulatory system. Compared to epithelial cells, mesenchymal cells are loosely aggregate and can therefore transform into other cell types very easily, as well as integrate into surrounding or remote tissue. Cells derived from epithelium and mesenchyme are needed for the organogenesis [18].

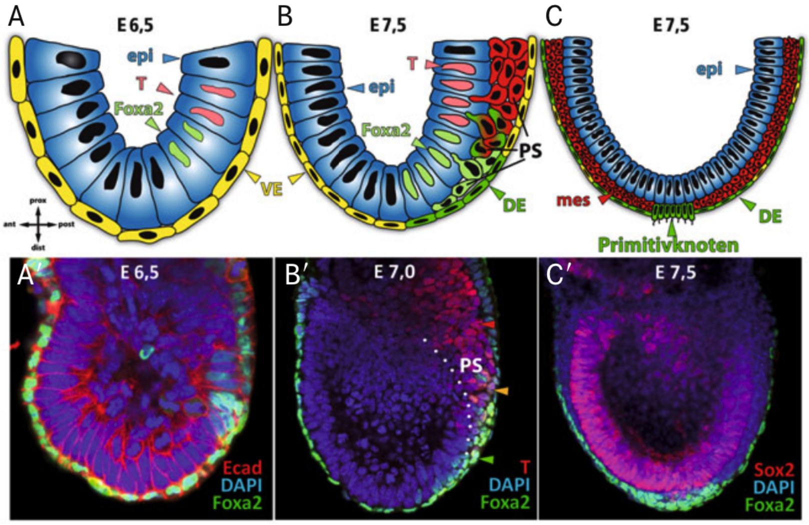


Figure 1.5.: Schematic drawings and fluorescence microscopic images of mouse embryos on day A(E6.5), B(E7.0) and C(E7.5). Day B is labeled incorrectly, instead of E7,5 it should say E7,0 [17].

Right before the gastrulation process begins(A), the posterior epiblast cells(epi) start to express Foxa2(Foxa2, green) and Brachyury (T, red). The outer layer consists of visceral entoderm cells (VE, yellow), which also express Foxa2. During gastrulation (B) the progenitor cells of the mesoderm and the endoderm emigrate from the epiblast and start to form the primitive streak (PS). The Foxa2 positive entoderm progenitor cells intercalate into a new layer, between the epiblast cell layer and the visceral endoderm cell layer. Those new cells, definitively become endoderm cells (DE, green) and gradually displace the outer visceral endoderm cells (VE, yellow). The proximal T positive mesoderm progenitor cells (mes, red) become mesodermal cells and manifest inbetween the epiblast layer and the endoderm cell layer, creating the mesoderm. Ecad stands for e-cadherin, a marker for the cell membranes, DAPI is a cellular nucleus marker and Sox2 strengthens the anterior epiblast cells. In the last step (C), the primitive node (Primitivknoten, green) arises [17].

The EMT process is categorized as followed: the first category takes place in embryonic cell development and organ formation. It is associated with the process of implantation, embryo formation and organ development. Neither causes EMT fibrosis nor does it induce an invasive phenotype resulting in spread via circulation. EMT is responsible for the creation of mesenchymal cells that can generate secondary epithelia [19]. The second category handles the involvement of EMT in the fields of wound healing, tissue regeneration and organ fibrosis. Also, the EMT process is associated with rising and ceasing inflammation. The third and last category treats EMT in the environment of carcinogenesis, cancer invasion, recurrence and metastasis. EMT can also occur in neoplastic cells, that have undergone genetic and epigenetic changes, specifically in genes that favor cloned outgrowth and the development of localized tumors[20].

The epithelial and mesenchymal phenotypes are not irreversible, they convert under specific conditions between those two phenotypes. This transition process is called EMT - the epithelial-mesenchymal transition. There are multiple similar terms, all used to describe EMT. Also, the term itself is often applied to distinct biological events. EMT related processes range in intensity from a transient loss of cell polarity to a total reprogramming of a cell. The epithelial-mesenchymal transition is very fundamental to life, generating morphologically and functionally distinct cell types. EMT has also a big influence at the time of tumor spreading. The reverse EMT process is called the mesenchymal-epithelial-transition process (MET), which also occurs during embryonic stem cell development. These two processes are critical for the appropriate morphogenesis of the organs [19].

1.3.1. CD24

CD24 is widely used as a marker for differentiation of multiple lineages of cells and controls an important genetic checkpoint for homeostasis and autoimmune diseases. Because of its extreme resistance to heat-inactivation, CD24 was originally called the heat-stable antigen (HSA). Since its initial discovery in 1978, CD24 has been used extensively to study differentiation of hematopoietic cells and neuronal cells, in addition to tissue and tumor stem cells. The CD24 gene has a diverse function due to its wide distribution, depending both on its composition and its cellular environment [21]. It encodes a small protein, ranging between 20 and 70 amino acids, that is heavily glycosylated and attached to the cell membrane by a glyco-phosphotidoyl-inistol (GPI) anchor. It is expressed in a broad range of cell types. CD24 plays crucial roles in lymphocyte maturation, neuronal development and tissue renewal homeostasis under physiologic conditions. This molecule has been proposed as a genetic checkpoint in T-cell homoeostasis and pathogenesis of autoimmune disease and has been recently added to the list of putative intestinal stem cell markers [22]. GPI-linked proteins are involved in signal transduction mediated by member of the protein tyrosine kinase (PTK) family [23]. In immune cells CD24 has been implicated in a number of functions: regulation of proliferation and apoptosis of B cell precursors and thymocytes [24] [25].

1.4. Microarrays

Microarray experiments generate vast amounts of data for functional genomics. Nowadays, they are available for almost every organism. The amount of an expressed gene can be measured using the quantity of the existent correspondent mRNA amount in the cell. This expression level can vary from cell type to cell type, also within the same cell, due to physiological influences.

Microarrays can measure the expression of large numbers of different genes simultaneously. There are several different types of microarrays. The underlying principle for all types works as described in the following: Microarrays consist of short DNA fragments attached to a spot on a plate. Depending on the type, the type of fragments may change from DNA to RNA, peptide fragments, antibodies, etc. The position of the spot and the sequence of every fragment is known. After labeling the mRNAs of the sample with fluorescent tags which will transmit a signal, if the mRNA hybridizes with a correspondent DNA fragment. This signal can be detected and measured. The intensity of the signal correlates with the quantity of bound mRNA, in case of DNA fragments. Also depending on the microarray type, the fluorescence can also be tagged on the attached fragment.

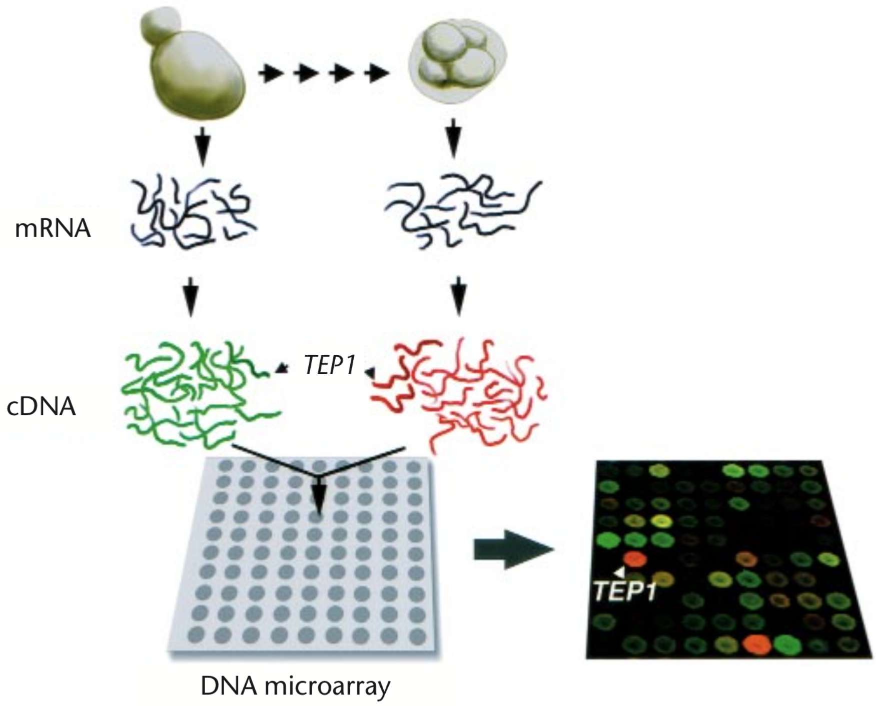


Figure 1.6.: Description of a microarray experiment using cDNA fragments tagged with fluorescent labels and RNA targets [26].

In order to compare different microarray experiments, the intensities need to be preprocessed and normalized. Optimally, all the experiments contain about the same quantity of mRNA. In this case, only the brightness needs to be adjusted. Usually, more complicated steps are needed, involving different factors, as expression levels of known housekeeping genes or assumptions about the behavior of genes.

1.5. Motivation

In this thesis we try to understand the differentiation process, more precisely how the split between endoderm and mesoderm is regulated. This split seems to be a epithelial-mesenchymal respectively mesenchymal-epithelial transition process. Preliminary in vitro studies by Ingo Burtscher from the Institute of Diabetes and Regeneration Research (IDR) at Helmholtz Zentrum Muenchen showed, that mesodermal cells seem to go through the classical epithelial-mesenchymal transition process and stay mesenchymal. During in mouse embryo live imaging endodermal cells give the appearance to transform from epithelial to mesenchymal cells and than to retransform back to epithelial cells. As this EMT-MET process doesn't express EMT core signaling factors (e.g. SNAIL1), endodermal cells don't seem to go through the typical EMT followed by MET. One of the aims of this thesis is to create the gene expression signatures of the endoderm differentiation process regarding the EMT process. CD24 was used to subclassify between early and late stages of embryonic stem cell development. CD24 is used as a cell-surface marker to distinguish stem cells, which express both, Foxa2 and Brachyury, and also to distinguish Foxa2 respectively Brachyury expressing stem cells to identify individual groups within those groups themselves.

Indeed we observe that CD24 clarifies whether the stem cell is in an early or late developmental stage, independently of the time of in vitro measurement.

2. Materials & methods

This chapter contains both the nature of the given information and the statistical methods that were applied. The fast development in wide areas of biotechnology and computation induce new types of data, increasing in storage size, resolution and complexity, it includes not only unexpected conclusions, but also new challenges in terms of data processing. These challenges require improved analysis tools and complex software.

2.1. R & Bioconductor

In the 1990s Ross Ihaka and Robert Gentleman developed a programming language 'R' due to lack of practical statistical software for their needs. It is a programming language requiring the use of a command line. Since its arrival, scientists have provided nearly 6000 packages for all kind of special needs. These packages allow users to just use already existing code, without having to invent the wheel erverytime.

Bioconductor is an open-source, open-development software project for the analysis and comprehension of high-throughput data in genomics and molecular biology. It is based on the statistical programming language R and includes 934 interoperable packages contributed by a global community of scientists [27]. The version 3.1.2 of the programming language R and the Bioconductor version 3.0 (BiocInstaller 1.16.2) were used for the statistical analysis in this thesis.

2.2. Mapping

As a result of incorrectly adapted Affymetrix packages for the microarray type used in this thesis, a json-file containing the latest mapping data from Affymetrix provided by Martin Preusse, was used to map the corresponding gene IDs with their gene names. JSON is an open standard language-independent file format, consisting of attribute-value pairs [28]. The package 'rjson' was used to import the json file.

2.3. Quality control

Quality control analysis needs to be performed in order to detect arrays with lower or poor quality. This analysis normally includes checking the signal intensity, signal variance, border elements, RNA degradation and array-to-array correlation. The overall aim is to eliminate problematic arrays mostly arising from problems during the microarray experiment. These arrays need to be removed from the statistical analysis, one could encounter bias in the results.

Gene expression can vary from cell type to cell type; even within the same cell it may vary due to changes in physiological circumstances. Depending on the type of microarray/data, the raw data must be normalized to be comparable across the given arrays.

Generally spoken, quality control is all about adjusting the overall brightness/intensities of each microarray experiment and checking in a very early stage of the statistical analysis, whether the given data is usable, or if the experiment should be redone. In addition there are several sources of noise in microarray data.

Unprocessed probe intensities

By comparing the raw probe intensities across all arrays, you can determine the overall signal intensities by box-plotting the log_2-intensities of every array. Commonly, this measurement is not very sensitive, as problematic arrays may appear as counterparts of arrays with better quality, but it delivers a first and quick check, if something went totally wrong during the experiment. Additionally you perform a histogram to identify distributions of signal intensities which behave different, like having a higher/lower density or a wider/smaller amplitude. For these tasks, the basic R functions 'boxplot' and 'hist' were used.

RNA degradation plots

RNA has participated at the 'end of it's life' in many protein synthesis. Gradually the RNA become degraded by the synthesizing enzymes. Some experiments may have been using bad or old RNA samples. By plotting the RNA degradation, it is very easy to identify those possible defective experiments. The probes are ordered from the 5' end of the targeted transcript to the 3' end. A strong degradation results in a systematic shift towards lower signal values of the probes closer to the 5' end, as RNA degradation starts from the 5' end of a transcript. The bioconductor packages 'affy' and 'affydata' were used for this analysis. The function 'AffyRNAdeg' calculated the signal values and the function 'plotAffyRNAdeg' plotted the results of the just named method. Different array types may result in array-typical individual results. So it is important to compare all the given arrays. If they have similar results, the gene comparisons may be valid. If the results show different degrees of RNA degradation, pointing out a possible bias in the experiments.

2.4. Preprocessing and normalization

The main goal of this step is to gain reliable gene expression values. Preprocessing and normalization of expression microarrays consist roughly of three major components: background correction, between-array normalization and reporter summarization. Background correction is done by adjusting the intensities for non-specific signals. It has an increasing sensitivity effect on the array's signals. The between-array normalization fits the intensity values between the arrays for technical variability, which is created by differences when handling the experiments, in labeling and hybridization steps or in the scanning process of the microarray. The last component consists of computing a gene expression value in summary for each gene from all the features on the array that targets its transcripts [29].

2.4.1. Preprocessing

Relative log error

The RLE plot shows the relative expression for each gene. Every gene is defined as the difference of an estimated gene expression and the median expression across all given arrays. Problematic arrays are indicated by larger spread out, by significantly shifted location of the boxplot regarding the zero line on the y-axis, or both [30].

Normalized unsorted standard error

NUSE stands for the normalized unsorted standard error. It represents the individual probe error, fitting the Probe-Level Model (PLM). PLM models expression measures, using different estimators and regression models. The median values for each probe are set to 1, so a boxplot of the data can easily detect, whether a probe has a low quality, or not. A boxplot near the one-value on the y-axis, indicates a good quality, whereas a boxplot centered further away from the one-value on the y-axis or a boxplot with a higher spread of NUSE distribution relative to the other probes, are signals for a low quality probe [30].

2.4.2. Normalization

Robust multi-array average

Robust multi-array average (RMA) is an algorithm to create an expression matrix from affymetrix microarray data. RMA calculates background-adjustments, normalization and log-transformation of PM values. PM stands for perfect match. Typically every gene is represented by a 16-20 pairs of oligonucleotides, also called 'probe sets'. These pairs are divided into two groups, which are 'perfect matches' PM and 'mismatches' MM. Mismatches are generated by changing the 13th base, to measure the the non-specific binding property of the probe pair. RMA computes background-corrected PM intensities for every PM cell on every array-spot [31]. After this step, the log_2 of every value computed in the step before is calculated and normalized using a quantile normalization [32]. RMA uses a linear model to fit the normalized data and gains in that way an expression measure for every probe set on every array [31].

2.5. Principal component analysis

Principal component analysis (PCA), invented in 1901 by Karl Pearson [33], is a mathematical way of identifying patterns in data and expressing the data, highlighting their similarities and differences. Mostly used for exploring huge interrelated data sets and for predictive models, this statistical procedure reduces the dimensionality of the data, transforming the given data orthogonally into new data sets of linearly uncorrelated and ordered variables, the so called 'principal components'. The first principal component retains the most of the original variation [34] [35]. For all calculations we used the bioconductor packages 'affy', 'affycoretools', 'limma' and 'simpleaffy'.

2.6. Regression Model

The regression model was calculated using the bioconductor package ’limma’, more precisely the functions ’lmFit’, ’eBayes’ and ’topTable’. lmFit fits a linear model for every gene and eBayes performs an empirical Bayes moderation of the standard errors [36]. The function ’topTable’ lists for a given number n the n most differentially expressed genes.

3. Results

3.1. Data

The data for this thesis has been produced by Ingo Burtscher and Heiko Lickert from the Institute of Diabetes and Regeneration Research (IDR) at Helmholtz Zentrum Muenchen . The microarray data is available in the CEL format, an Affymetrix file format. It is produced by the Affymetrix microarray scanning software. 20 CEL-files representing the different times of measurement and marker, were imported into R using the Bioconductor package.

3.1.1. Experimental data overview

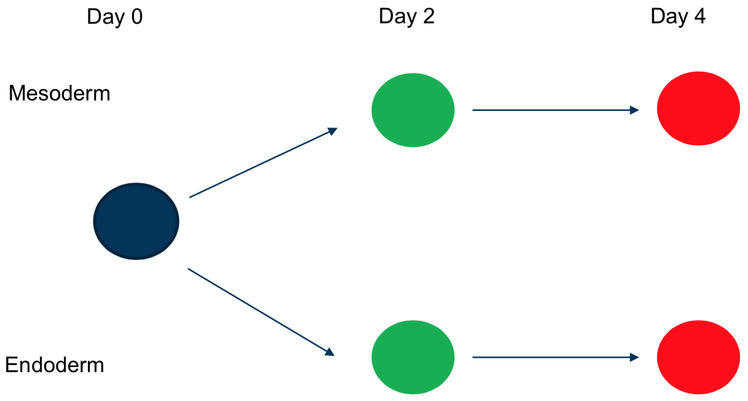


Figure 3.1.: Overview of the different times of measurements.

To reach the goal of exploring in vitro embryonic stem cell differentiation of mice towards endoderm and mesoderem, embryonic stem cell lines were used, expressing Foxa2 and T. Those proteins were fused with flourescence tags in order to report their intracellular expression levels. Additionally, cell surface staining for CD24 was performed, which allows to sort the stem cells for three different markers, 'Foxa2', 'T' and 'CD24'. Protein expression level were measured for three times, resulting in twenty CEL-files, containing a Foxa2 or T positive respectively negative expression states, as well as a CD24 high, low or negative expression states, as listed in the following table.

CEL-file name	time	Foxa2	T	CD24
ES-09.CEL	Day 0	negative	negative	NA
ES-34.CEL	Day 0	negative	negative	NA
F–CD24lowMC-D2-13.CEL	Day 2	positive	negative	low
F–CD24lowMC-D2-25.CEL	Day 2	positive	negative	low
F–CD24highMC-D2-28.CEL	Day 2	positive	negative	high
F–CD24highMC-D2-29.CEL	Day 2	positive	negative	high
F-T-D2-27.CEL	Day 2	positive	positive	NA
F-T-D2-33.CEL	Day 2	positive	positive	NA
T-CD24low-D2-05.CEL	Day 2	negative	positive	low
T-CD24low-D2-22.CEL	Day 2	negative	positive	low
F–CD24lowMC-D4-04.CEL	Day 4	positive	negative	low
F–CD24lowMC-D4-14.CEL	Day 4	positive	negative	low
F-CD24high-D4-10.CEL	Day 4	positive	negative	high
F-CD24high-D4-18.CEL	Day 4	positive	negative	high
F-T-D4-17.CEL	Day 4	positive	positive	NA
F-T-D4-24.CEL	Day 4	positive	positive	NA
T-CD24low-D4-06.CEL	Day 4	negative	positive	low
T-CD24low-D4-23.CEL	Day 4	negative	positive	low
T-CD24neg-D4-03.CEL	Day 4	negative	positive	negative
T-CD24neg-D4-19.CEL	Day 4	negative	positive	negative

Table 3.1.: Overview of all the celfiles, regarding their properties.

To identify the samples, the CEL-file names were used in the following plots.

3.2. Quality control

3.2.1. Unprocessed probe intensities

First we performed a boxplot of the raw probe intensities across all arrays, to determine the overall signal intensities by plotting the log2-intensities of every array.

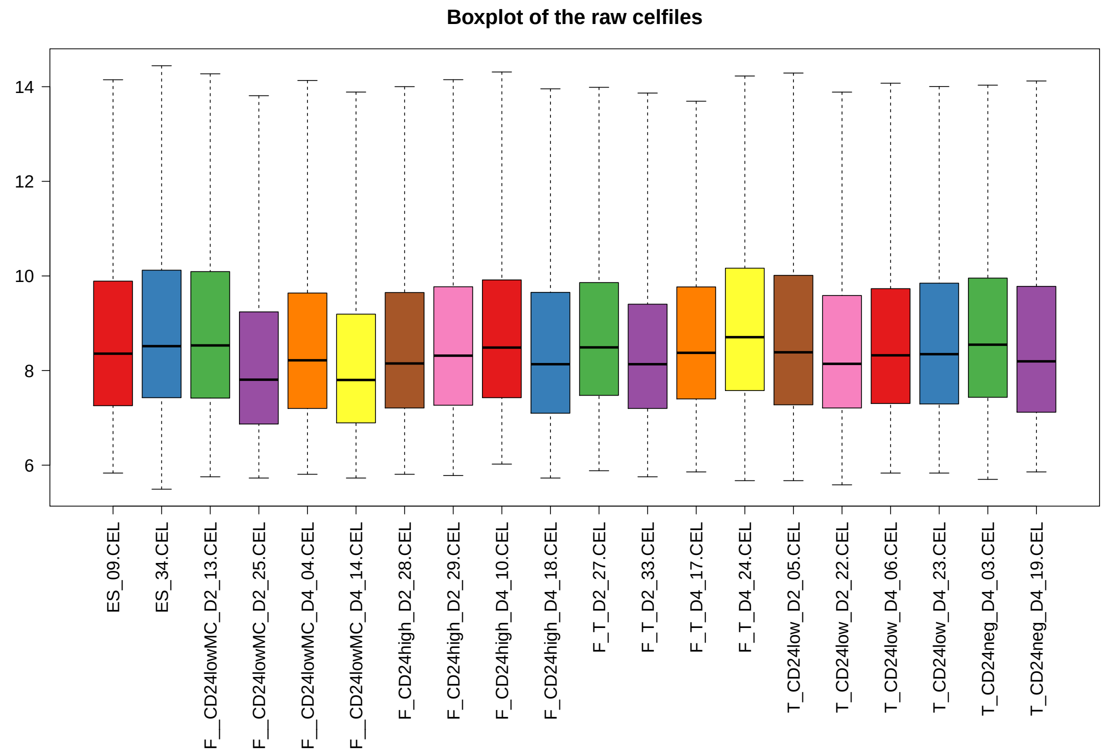


Figure 3.2.: Boxplot of all the raw data sets.

Figure 3.2 indicates, that there are no potentially defective data sets, as all boxplots range in a very similiar value, relative to each other. The next figure shows a boxplot of already normalized data, indicating a nearly perfect adjustment of the intensity values, to make all datasets comparable to each other.

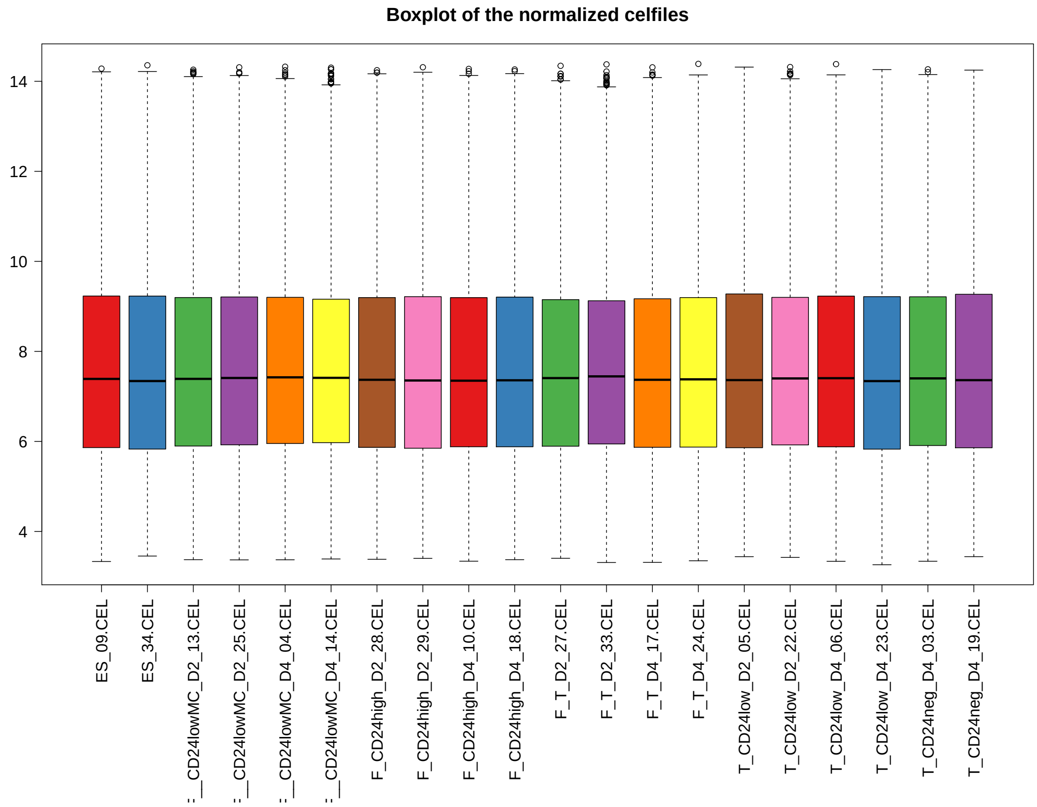


Figure 3.3.: Boxplot of the normalized data.

A histogram of the raw data distributions of signal intensities can identify data sets with different behavior, relative to all other datasets. The histogram demonstrates no abnormal behavior. Especially, the probe and control sets of every data sets correlate strong to each other, indicating well performed experiments.

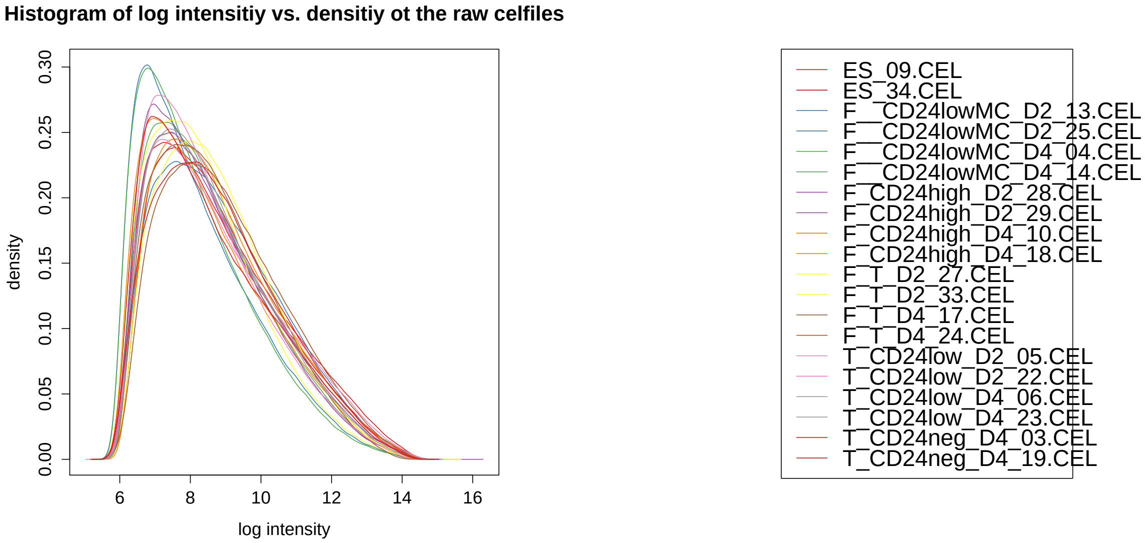


Figure 3.4.: Histogram of the log intensity versus the density of the raw celfiles.

3.2.2. RNA degradation plots

Similar results confirm valid intensity values due to not degraded RNA samples. As Figure 3.5 demonstrates, all microarrays have very similar degrees of RNA degradation compared to its control dataset, indicating proper experimental data sets.

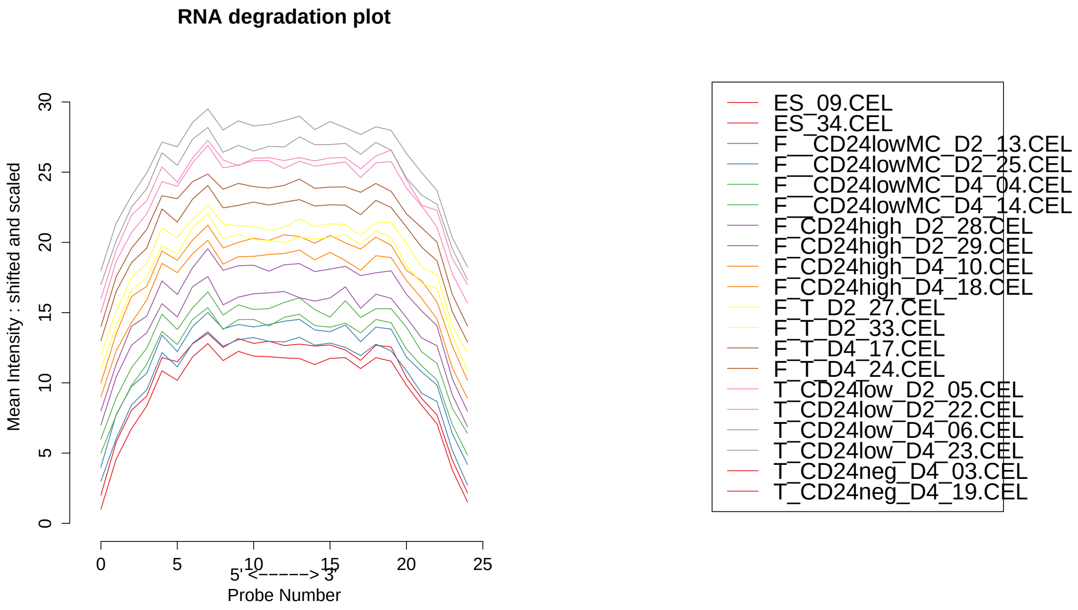


Figure 3.5.: RNA degradation plot, comparing all the datasets.

A plot of the summary of the RNA degradation function shows datapoints, representing the individual datasets. These datapoints indicate a bias in the experimental data, if moved to much away from zero value on the y-axis. Figure 3.6 presents excellent values for every dataset, indicating a valid condition of the used datasets.

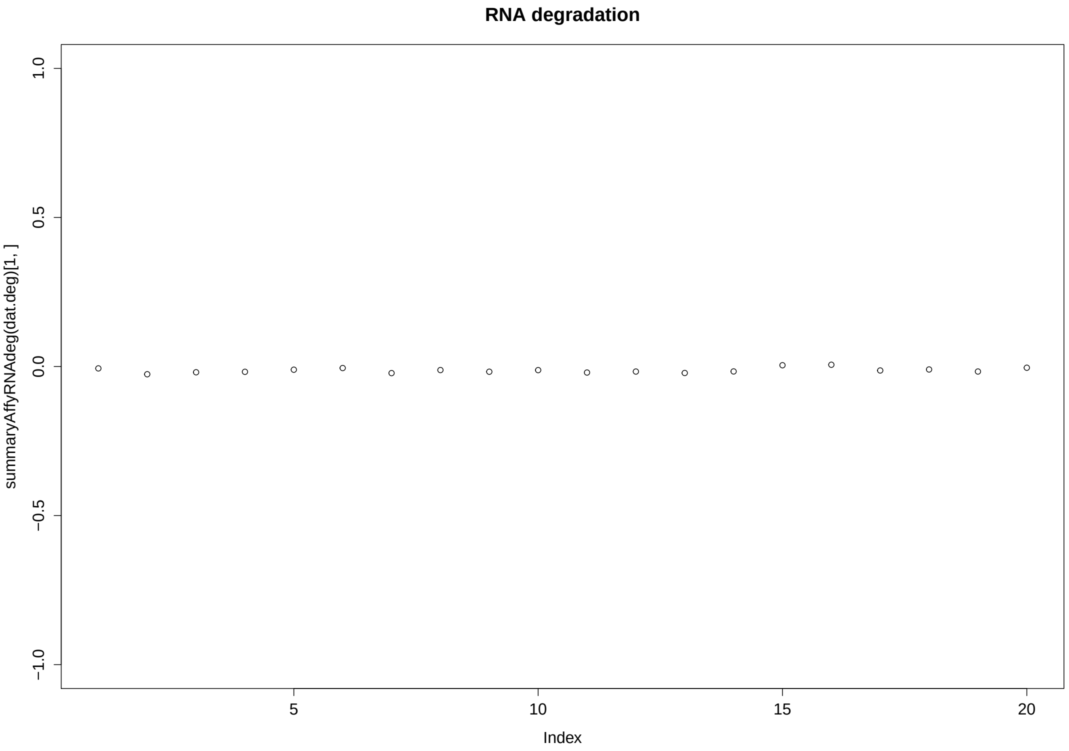


Figure 3.6.: RNA degradation plot, plotting the values computed by the summary-function of the AffyRNAdeg-function.

3.3. Data preprocessing & normalization

RLE - relative log error

The RLE plot shows the relative expression for each gene of every dataset. Every gene is defined as the difference of an estimated gene expression and the median gene expression across all given arrays. Problematic arrays have a larger spread-out or a location significantly different from a zero value on the y-axis, or both. Figure 3.7 shows neither significantly large spread-outs, nor significantly shifted y-values, as they range between about -0,5 to 0,4.

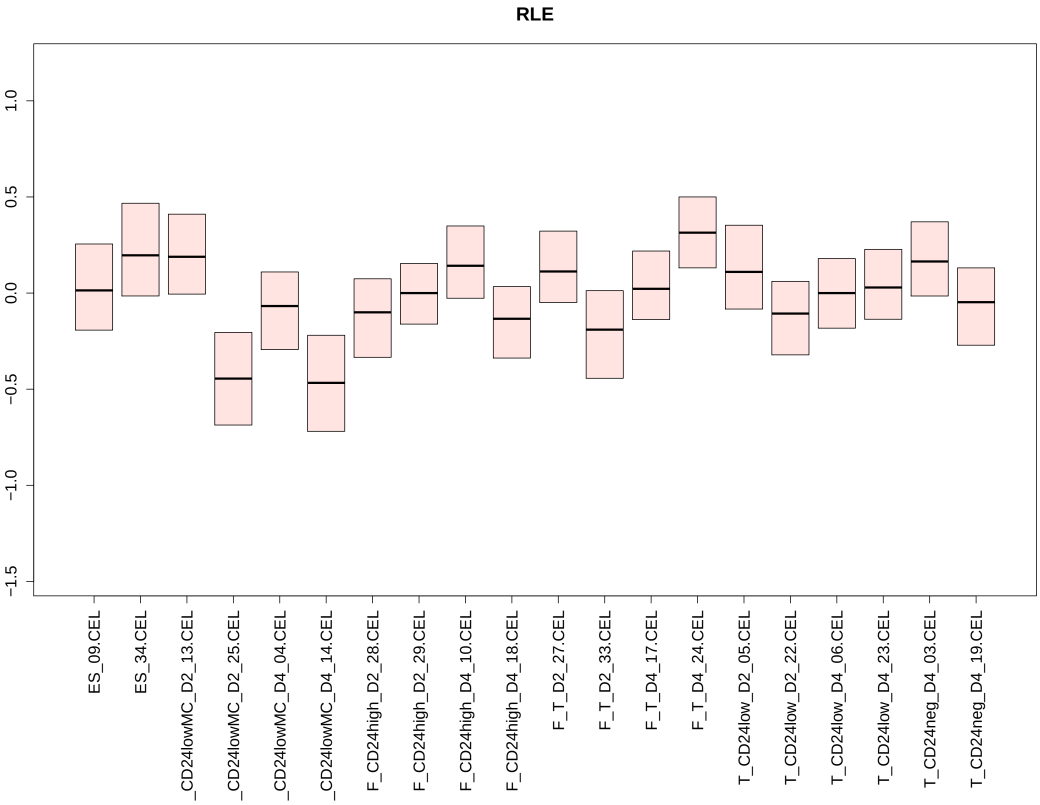


Figure 3.7.: Relative error log plot of the unprocessed datasets.

NUSE - Normalized unsorted standard error

As described in the previous chapter, a boxplot near the one value on the y-axis indicates a good quality, while a boxplot centered further away or a boxplot with a higher spread, relative to the other probes, is a signal for a low quality probe. Figure 3.8 shows again the validity of the datasets.

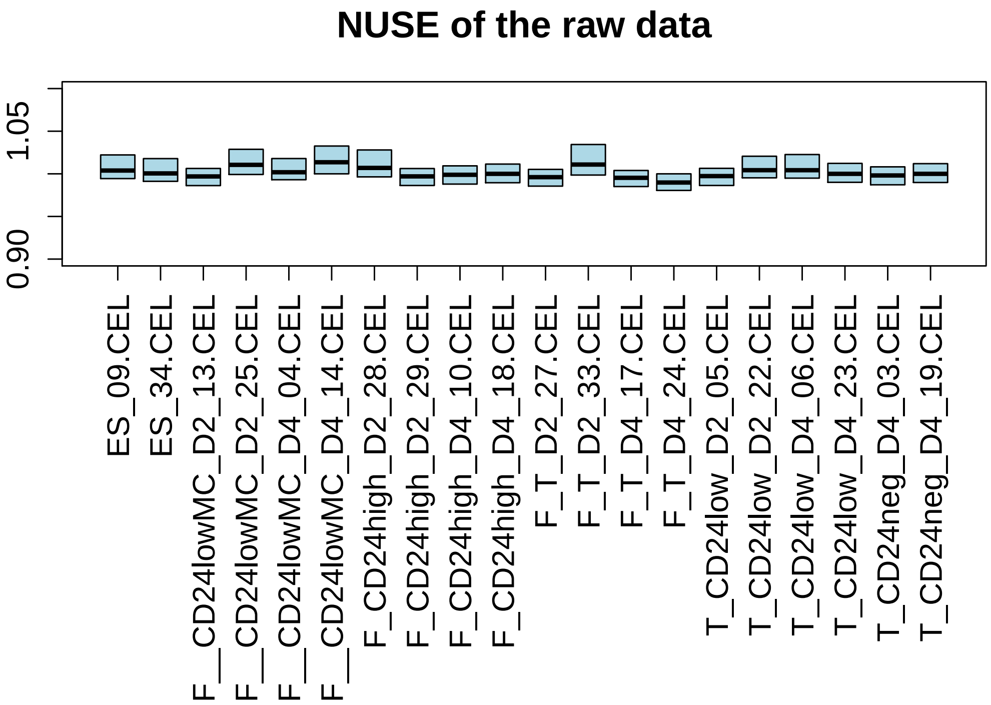


Figure 3.8.: NUSE-plot of the raw data.

3.4. Statistical analysis

3.4.1. Principal component analysis

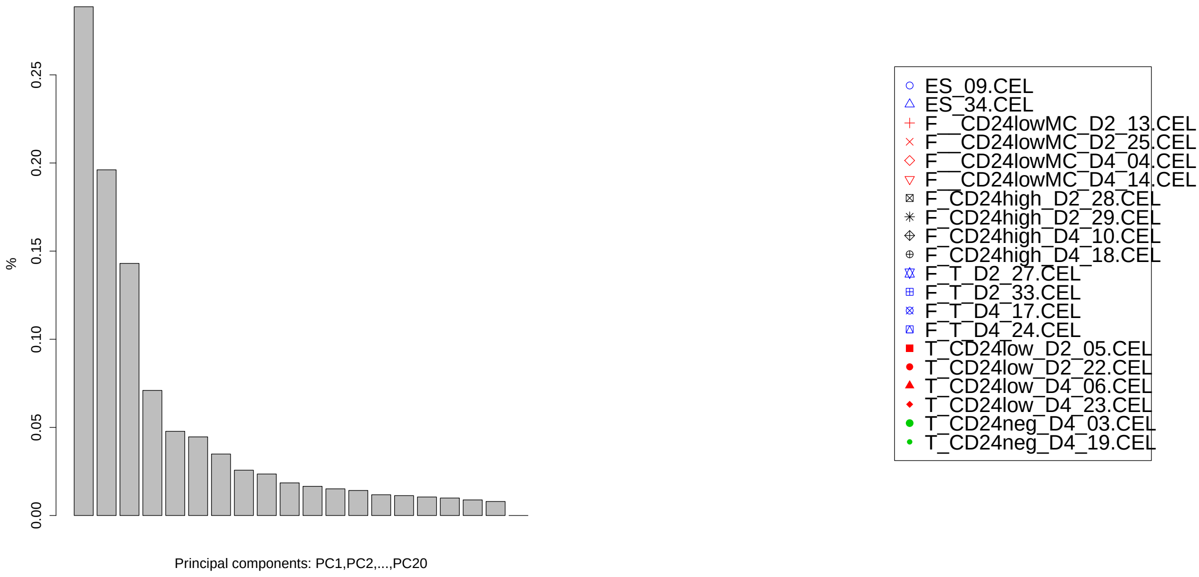


Figure 3.9.: Overview of the percentage distribution of the original variation among the principal components and the PCA plot legend.

The barplot in the right half shows the percentage distribution of the original variation among the principal components. PC1 (35.30 %) , PC2 (29.10%) and PC3 (24.85 %) add up to 89.25% of the total variance and thus they are sufficient for applying the gained knowledge by only regarding PC1, PC2 and PC3, to the entire dataset.

For every variable the principal component analysis calculates a specific value, which has to be multiplied to the standardized original value in order to get the score of the principal component. These specific values are called 'loadings' and were plotted, resulting in the so called rotation matrices. The value indicates how much the value has to be corrected to be part of the principal component. Thus this rotation matrix outliers indicate genes with high changes in variance, allowing a first glance at maybe regulatory important genes.

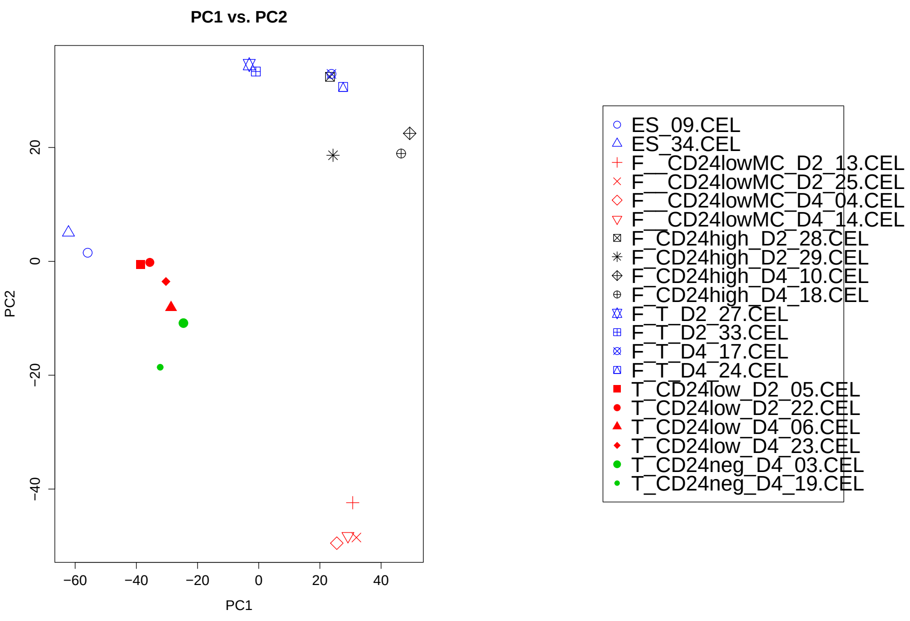


Figure 3.10.: PCA plot showing the principal components 1 and 2.

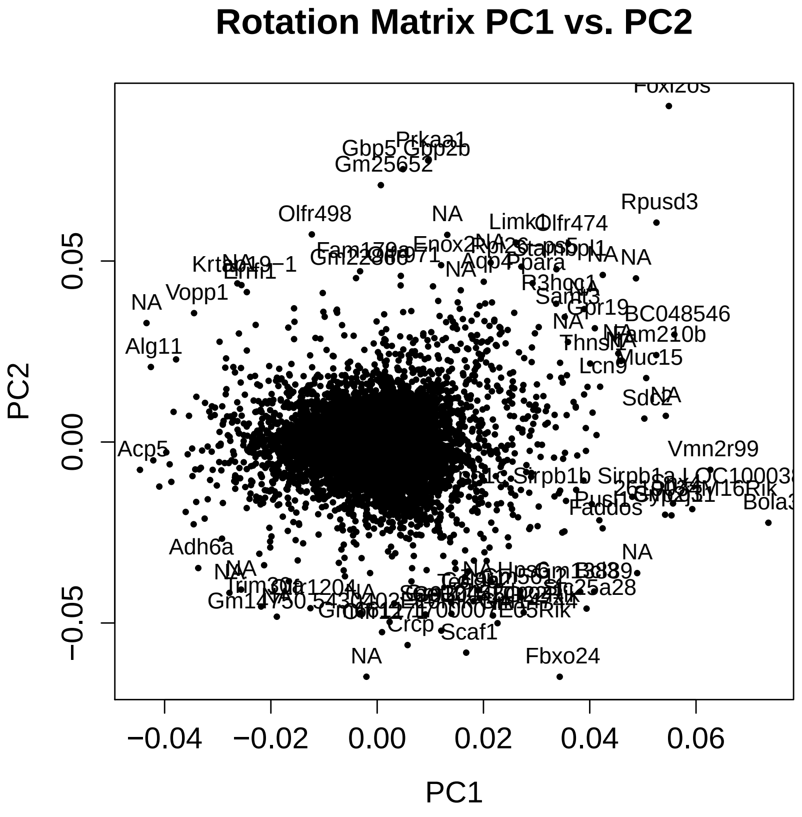


Figure 3.11.: Rotation matrix PC1 PC2

Figure 3.10 indicates that PC1 distinguishes between non Foxa2 positives samples and Foxa2 positive samples, while PC2 separates CD24 high and CD24 negative respectively low samples, pointing out, that double positive samples (Foxa2 positive and T positive), seem to behave like late endoderm (CD24 high) samples. Interestingly, CD24 doesn't cluster concording to the time of measurement (D0/D2/D4), thus we assumed that CD24 segregates by the progress of embryonic development.

The top six genes were 'Foxl2OS', 'Rpusd3', 'Prkaa1', 'Bola3', 'Gbp5/Gbp2b' and 'Fbxo24'. Foxl2OS(located in the top right corner of Figure 3.11) is an unclassified non-coding RNA gene, with no significant coding region. OS stands for opposite strang, as in mouse tow isoforms result from alternative adenylation. Foxl2 is a forkead transcription factor for eyelid development as well as in development and adult function of the ovary in mammals [37]. Due to its position, this gene seems to have an impact in PC1 as well as in PC2. Also found in a similar position on the rotation matrix, the gene Rpusd3 encodes a RNA synthase, involved in the intramolecular conversion of uridine to pseudouridine within an RNA molecule. This post-transcriptional base modification occurs in tRNA, rRNA, and snRNAs. Prkaa1, found in the top center, is an AMP-activated protein kinase. Due to its functionality, it is involved in many regulatory activities, e.g. also in the regulation of neuronal structure in developing neurons [38]. This function is consistent with the PCA, as the genes position indicates a late embryonic developmental stage. Bola3, a protein encoding gene, is found und the very right of the rotation matrix. Its human homolog is involved in the production of iron-sulfur clusters and for the assembly of the mitochondrial respiratory chain complexes [39]. Gbp5 respectively Gbp2b (positioned slightly below Prkaa1) , are guanylate binding proteins, with no further known relation regarding embryonic stem cell development [40]. Fbxo24 is a member of the F-box protein family, localized in the bottom center of Figure 3.11. Its human homologe, constitutes one of the four subunits of the ubiquitin protein ligase complex [41]. Slightly above we find 'Scaf1', a SR-related CTD-associated factor. Its human homolog may be considered as a new prognostic marker for breast and ovarian cancer [42]. 'Alg11' and 'Acp5', which are found on the left side of Figure 3.11 have no notable known function or involvement in embryonic stem cell development, but seem to be important in the early embryonic stem cell development, regarding their position. Slightly above, you find 'Vopp1', which is involved in regulation of transcription and signal transduction and is also over expressed in cancer [43]. Whereas slightly below, you find 'Adh6a', an alcohol dehydrogenase, with no further known thematic involvement.

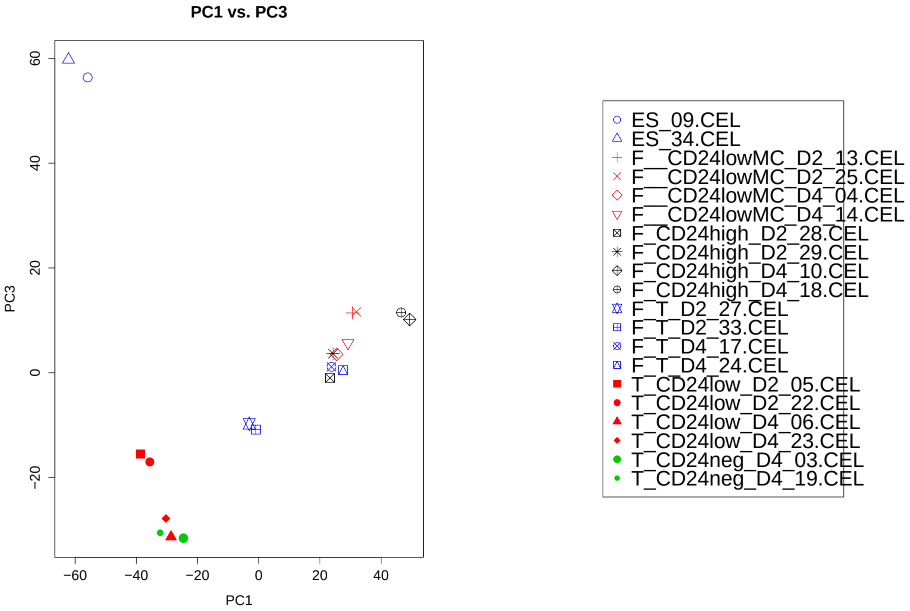


Figure 3.12.: PCA plot showing the principal components 1 and 3.

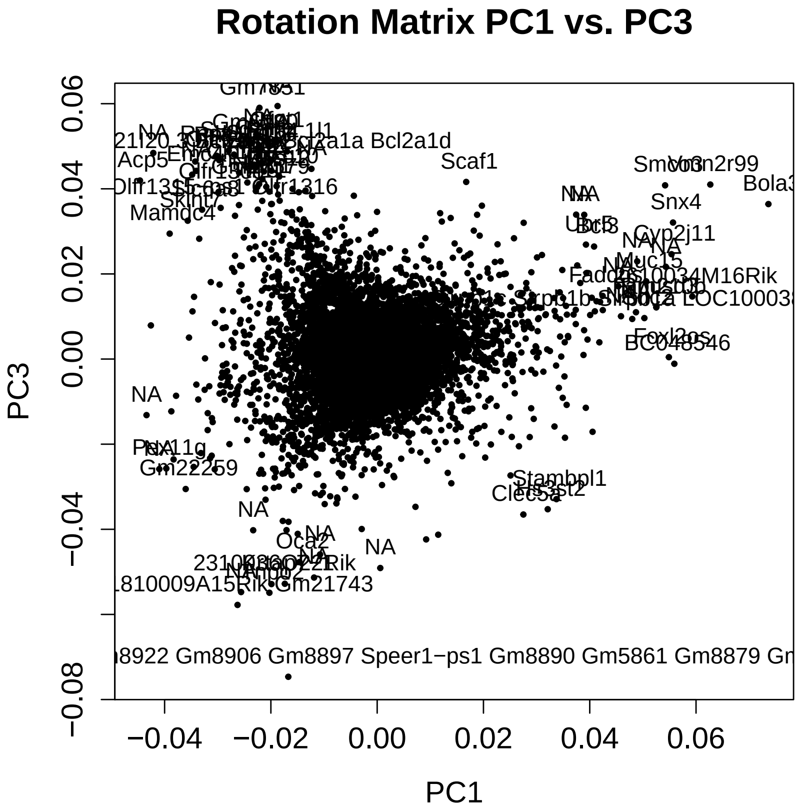


Figure 3.13.: Rotationsmatrix PC1 PC3

Figure 3.12 shows the PCA plot of the principal components 1 and 3. PC3 distinguishes between ESC samples and the rest. PC3 also separates the non ESC samples between endodermal and mesodermal disregarding the CD24 expression level, having the only T positive samples on one side, going over with a center filled with double positive samples (Foxa2 and T) and ending with only Foxa2 positive samples.

Besides some of the already discussed genes, we find in the rotation matrix (Figure 3.13) some other high variance genes: 'Speer1-ps1', 'Vmn2r99', 'Smco3' and 'Snx4'. The pseudogene speer1-ps1 (spermatogenesis associated glutamate (E)-rich protein 1, pseudogene 1) found in the bottom of the rotation matrix, is expressed tissue specifically in the testis of mice [44]. Vmn2r99 is a vomeronasal receptor, needed for primarily chemical detection of pheromones [45]. Smco3 (single-pass membrane protein with coiled-coil domains 3) encodes a transmembrane transfer protein [46]. No indication towards an involvement in embryonic development was found for all discussed genes. Snx4 encodes a protein involved in the endocytic recycling process, but also is found in the process of epidermal growth factor receptor binding [47].
Figure 3.14 shows the PCA plot of PC2 and PC3, confirming the already described properties. PC2 separates early endodermal samples from late endodermal and double positive (Foxa2 and T) samples. Also the described fact, that those double positive samples seem to behave like CD24 high samples, is shown clearly. PC3 behaves also as already described, separating the ESC samples from the other samples. Those samples show also in this plot the gradually distribution within the non ESC samples from Foxa2 positive samples going over to double positive samples and ending with the T positive samples.

The rotation matrix contains mostly already described labeled genes, indicating a consistence within the principal component analysis.

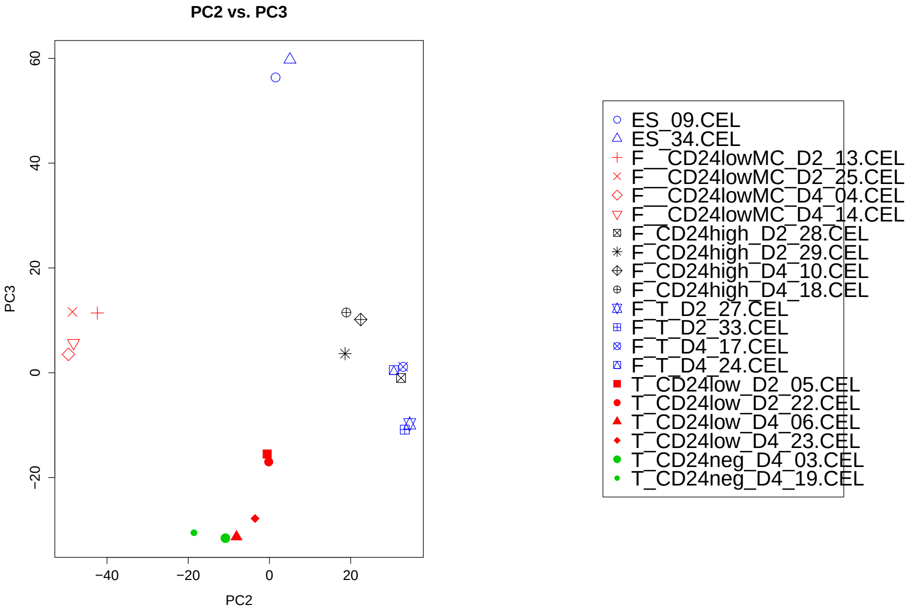


Figure 3.14.: PCA plot showing the principal components 2 and 3.

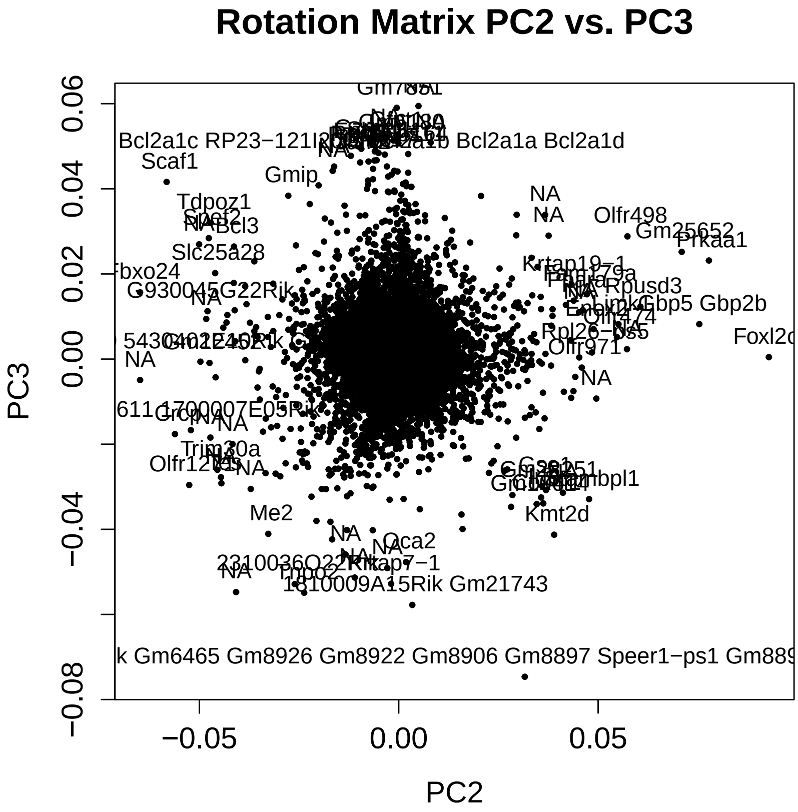


Figure 3.15.: Rotationsmatrix PC2 PC3

3.5. Regression model

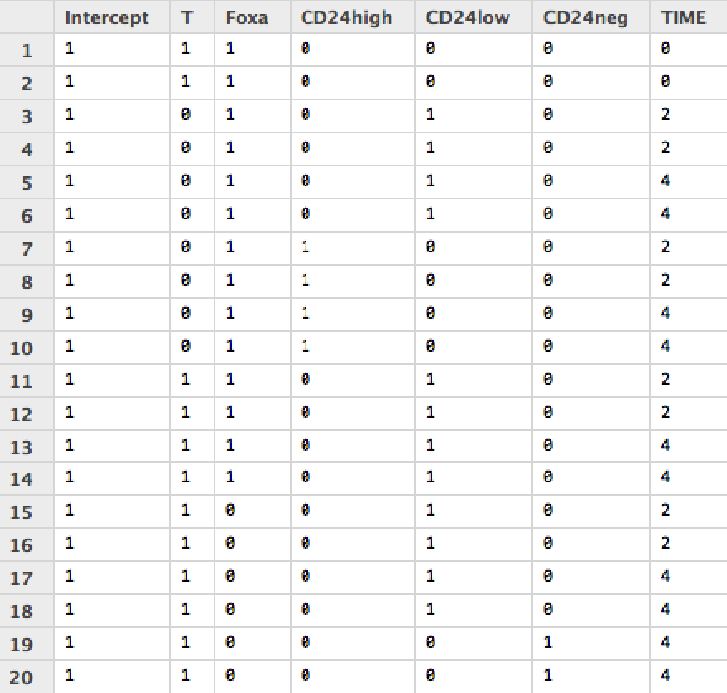
In order to gain for every gene(γ) a time ('TIME' $= D_0, D_2, D_4$) and feature (T positive / Foxa2 positive / CD24 high / CD24 low / CD24 negative) dependent expression value, a regression model was designed as described below:

$$\gamma \sim \beta_0 + \beta_1 Foxa + \beta_2 T + \beta_3 CD24high + \beta_4 CD24low + \beta_5 CD24neg + \beta_6 TIME$$

γ is composed, adding the different gene expression values, if the feature is expressed and represents in that way the overall regulatory activity of this gene on a specific time (Day 0, Day 2, Day 4). The used functions 'lmFit' and 'eBayes' fit a linear model for every gene and perform an empirical Bayes moderation of the standard errors, resulting in p-values for every gene.

Design matrix

The design matrix contains a column 'intercept', to guarantee that every cel-file is being used in the calculation of the gene expression values. A zero value causes a disregard of the correspondent feature.



	Intercept	T	Foxa	CD24high	CD24low	CD24neg	TIME
1	1	1	1	0	0	0	0
2	1	1	1	0	0	0	0
3	1	0	1	0	1	0	2
4	1	0	1	0	1	0	2
5	1	0	1	0	1	0	4
6	1	0	1	0	1	0	4
7	1	0	1	1	0	0	2
8	1	0	1	1	0	0	2
9	1	0	1	1	0	0	4
10	1	0	1	1	0	0	4
11	1	1	1	0	1	0	2
12	1	1	1	0	1	0	2
13	1	1	1	0	1	0	4
14	1	1	1	0	1	0	4
15	1	1	0	0	1	0	2
16	1	1	0	0	1	0	2
17	1	1	0	0	1	0	4
18	1	1	0	0	1	0	4
19	1	1	0	0	0	1	4
20	1	1	0	0	0	1	4

Figure 3.16.: Design matrix representing the experimental data. The design matrix was needed for the calculation of the regression model.

Volcano plots

A volcano-plot is usually used to plot an effect-measurement on the x-axis (in our case the log fold change) and the statistical significance on the y-axis (in our case the log odds). Figure 3.17 shows a volcano plot of the genes most regulatory active, when T is expressed, indicating which genes have the highest statistical relevance and the highest log fold. The 30 most outlying genes were labeled using the latest mapping data from Affymetrix.

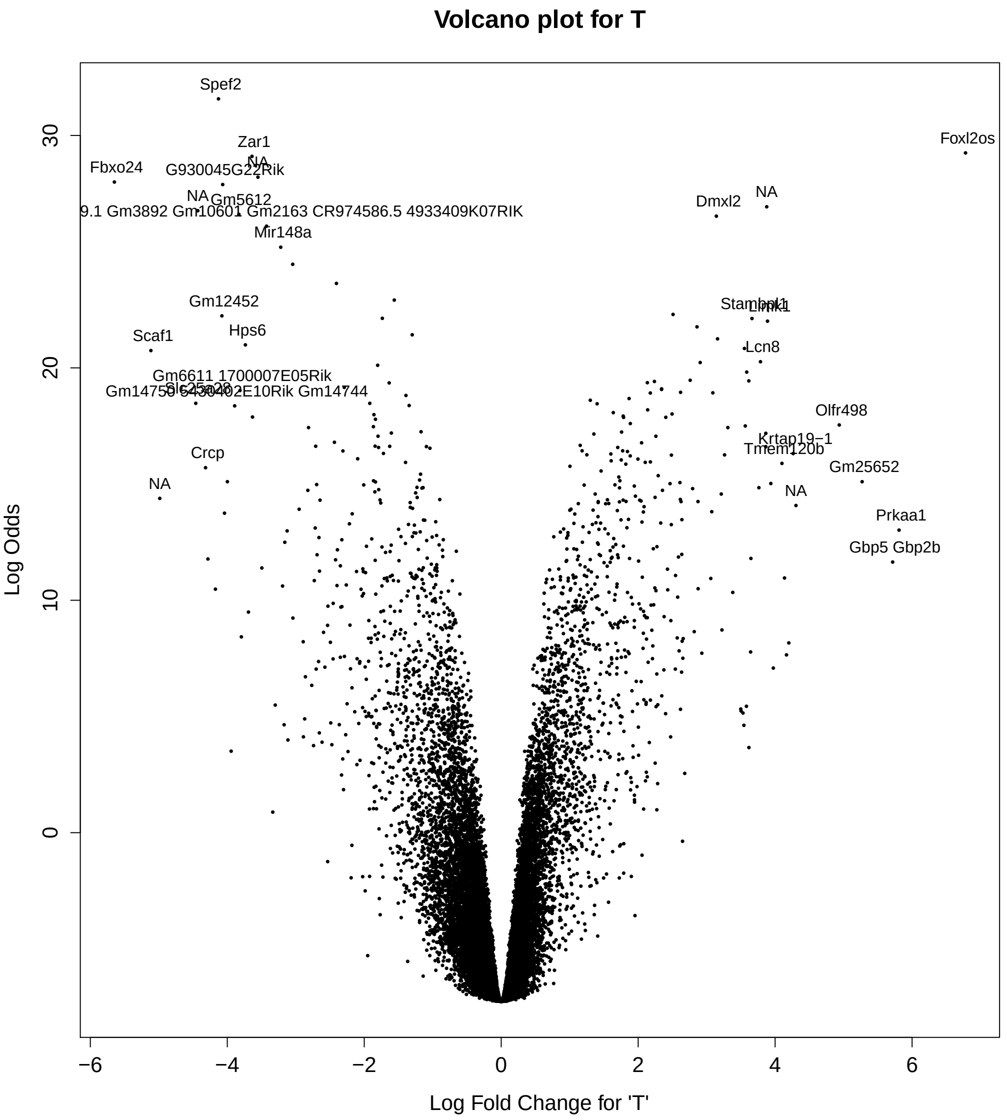


Figure 3.17.: Volcano plot of 'T'.

Figure 3.17 shows some familiar genes as Foxl2OS, Prkaa1, Gbp5/Gbp2b, Scaf1 and Fbxo24, confirming their regulatory activity in samples, in which T was expressed. Some new ones appear, like 'Spef1', a protein coding gene for a sperm flagellar, apparently nothing vital regarding the embryonic stem cell development. Another one is 'Zar1', which encodes the protein 'zygote arrest 1', a gene, critical for the oocyte-to-embryo transition [48].

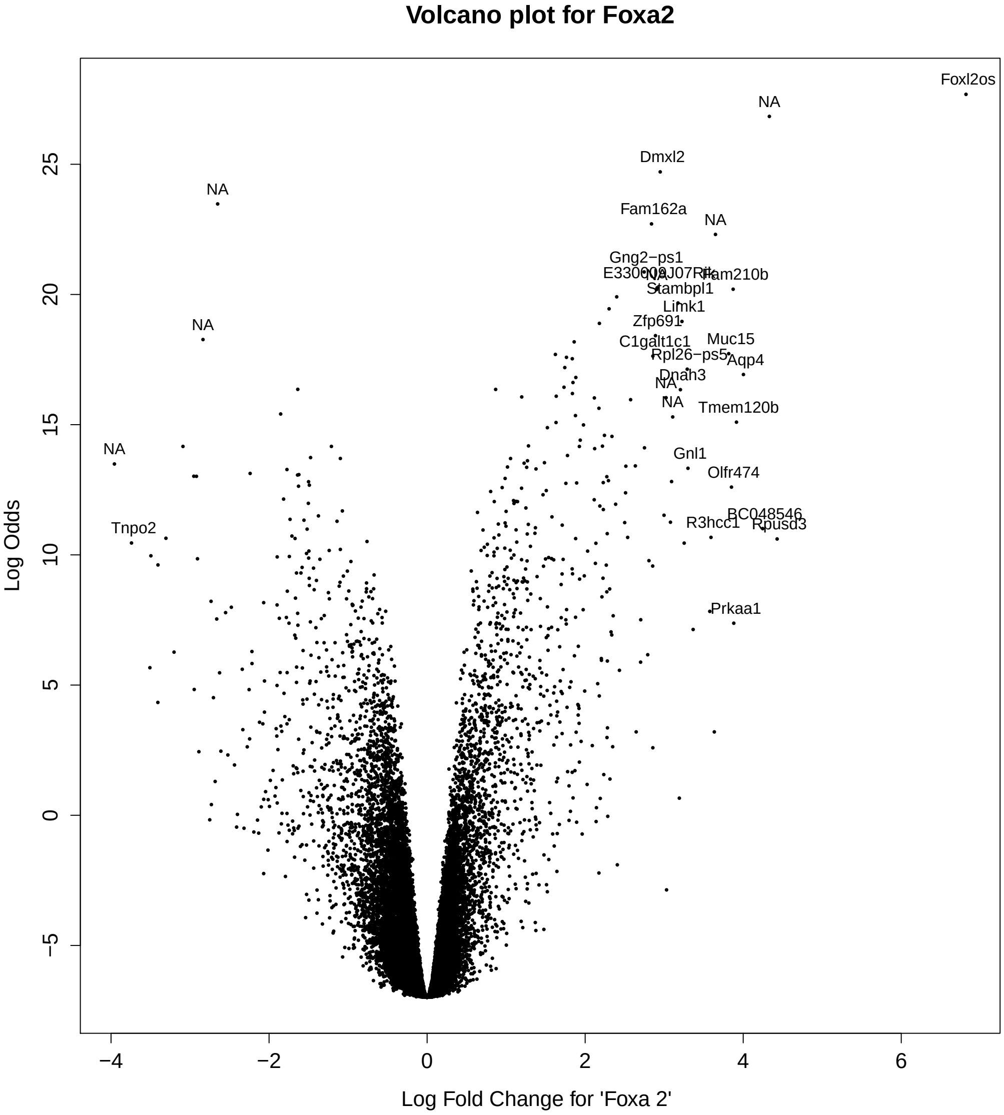


Figure 3.18.: Volcano plot of 'Foxa2'.

This figure (Figure 3.18) contains also some already described genes, as Foxl2OS or Prkaa1. 'Dmxl2' encodes a protein involved in cell junction process, 'Muc15' or 'Tnpo2' on the other hand are GO annotated as a integral component of membrane respectively functioning in intracelllular protein transport.

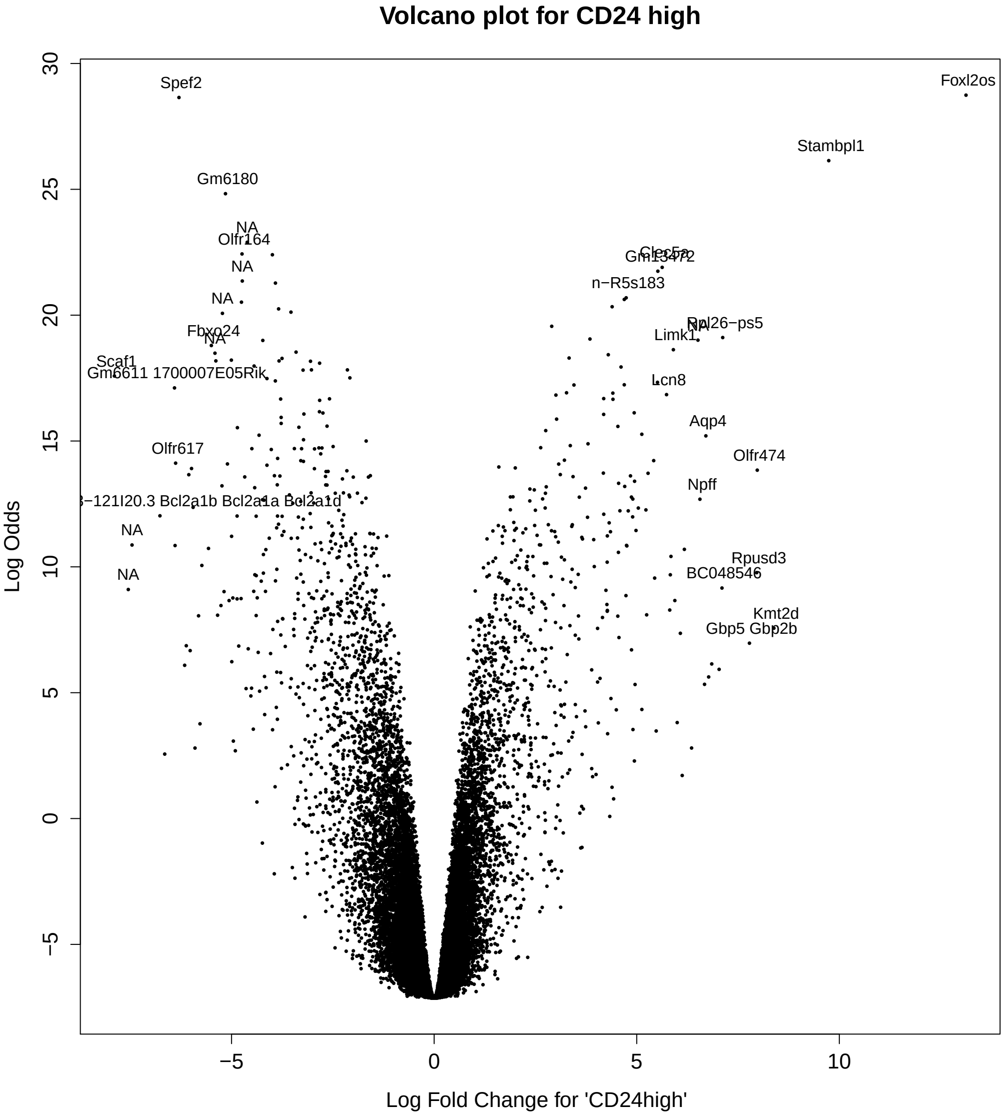


Figure 3.19.: Volcano plot of 'CD24 high'.

In Figure 3.19 we find again Scaf1, Foxl2OS, Rpusd3, Gbp5/Gbp2b, Spef2 and Fbxo24. New highlighted genes are e.g. 'Stambpl1', a STAM binding protein needed in the process of proteolysis according to its GO annotation or also 'Gm6180', a predicted pseudogene with no known function.

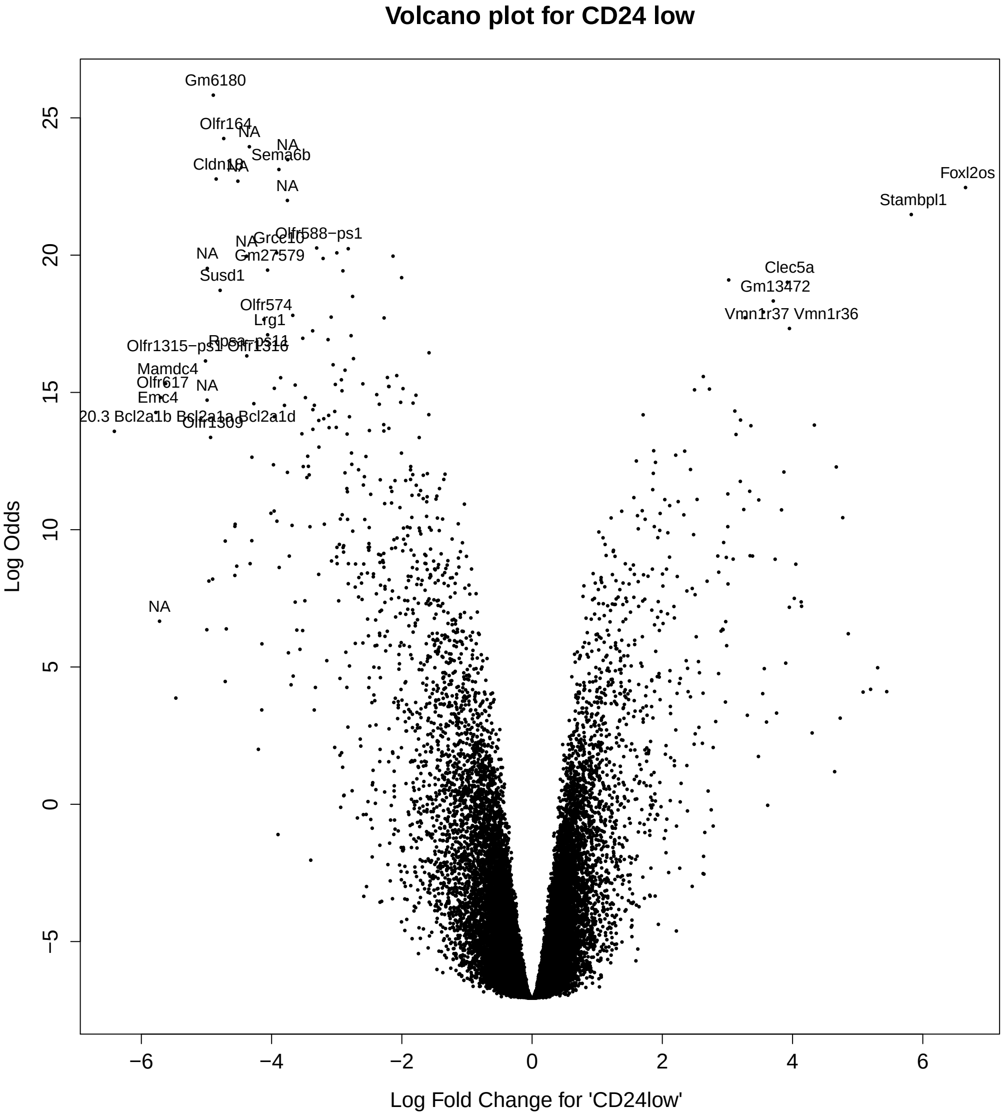


Figure 3.20.: Volcano plot of 'CD24 low'.

The volcano plot for CD24 low (Figure 3.20) contains only one known labeled gene Foxl2os, but the results are very similar to the volcano plot for CD24 negative (Figure 3.21). Several genes are represented in both plots, but also Figure 3.21 contains some genes, represented in Figure 3.19 and not labeled in Figure 3.20. Figure 3.20 and figure 3.19 contain the gene 'Clec5a', involved in the activation of myeloid cells [49]. 'Gm6180', is a predicted pseudogene with no known function [50].

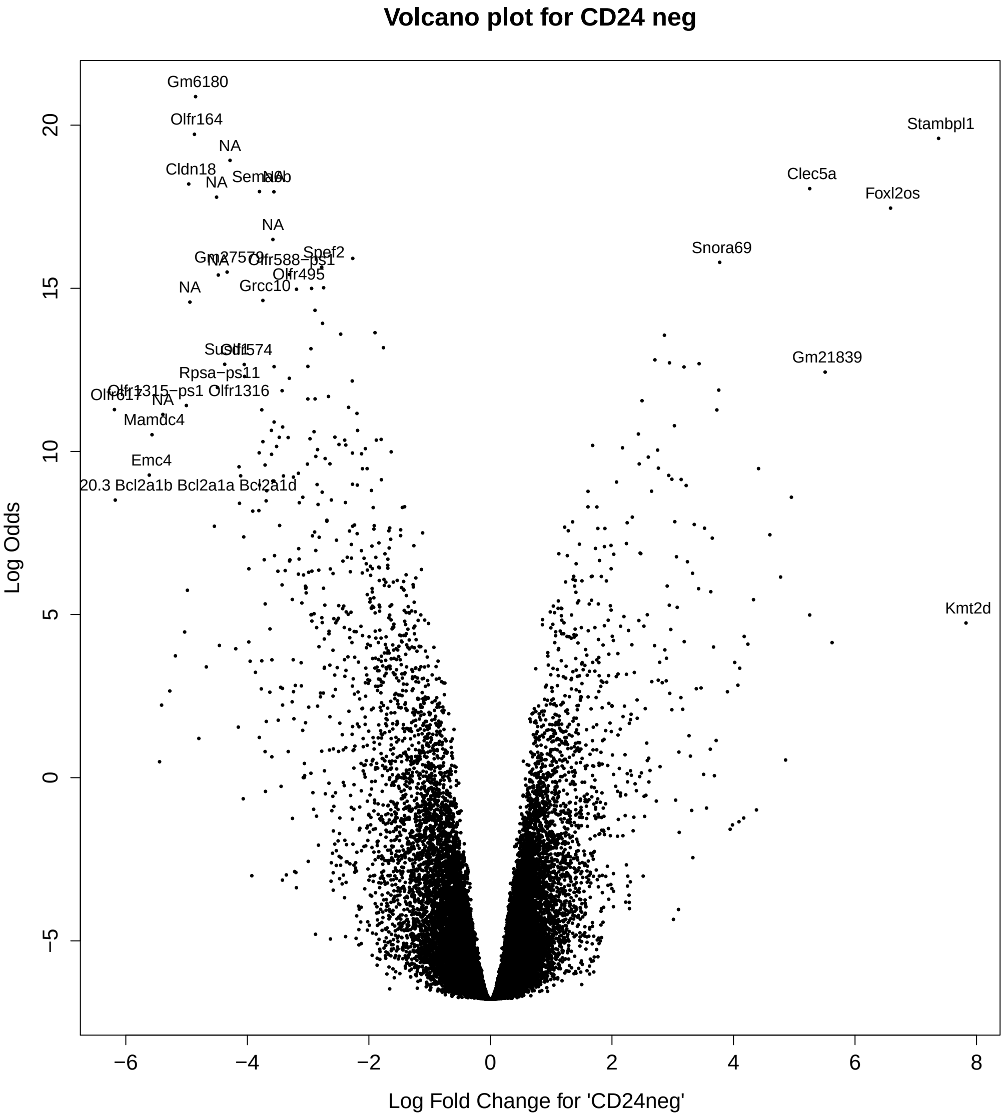


Figure 3.21.: Volcano plot of 'CD24 neg'.

The volcano plot for CD24 negative smaples, contain several already described genes. Mostly all of them are already represented in Figure 3.19 and 3.20.

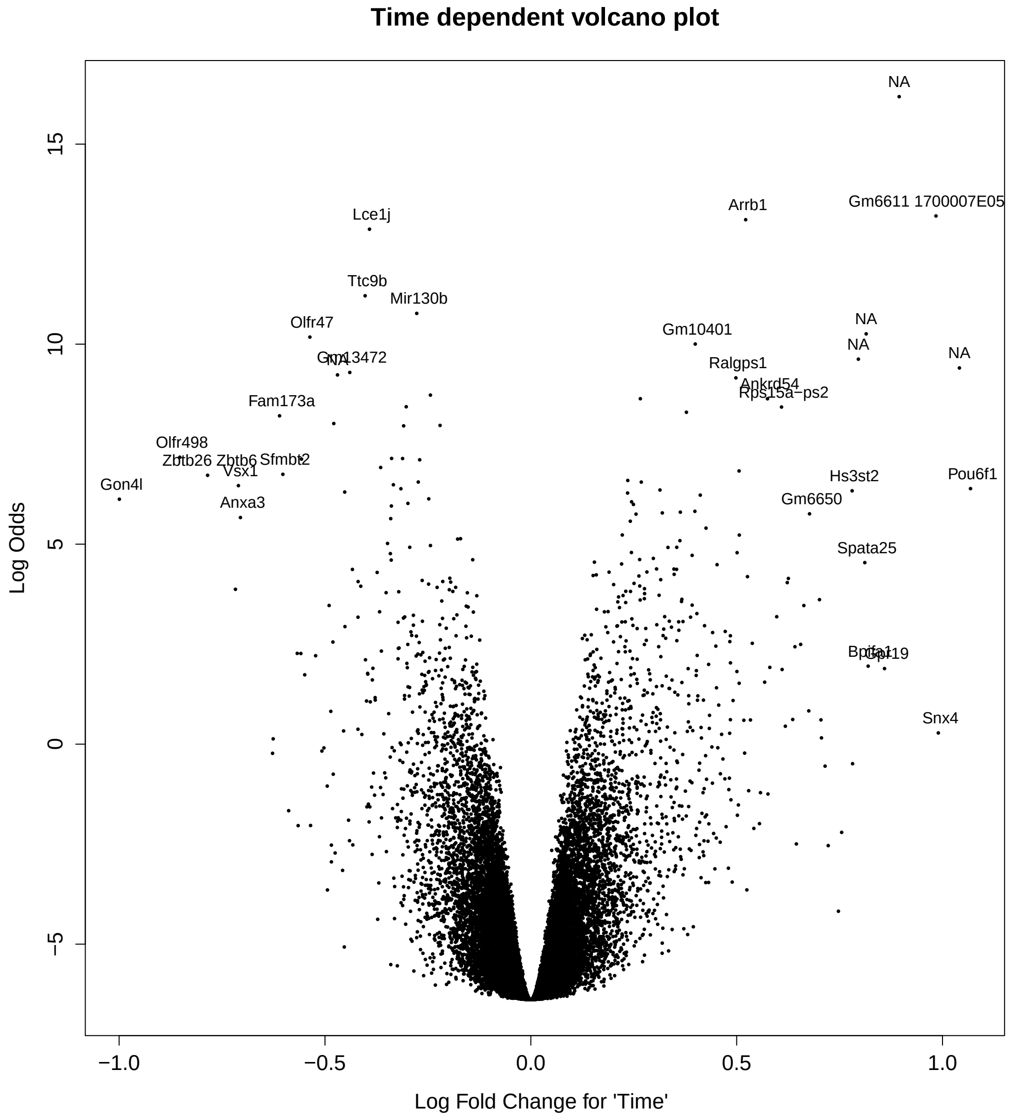


Figure 3.22.: Volcano plot of 'TIME'.

This volcano plot contains nearly only still not discussed genes. 'Ttc9b', a tetratricopeptide protein coding gene, is GO annotated within in the process of chaperone-mediated protein folding. 'Arrb1' encodes β-arrestin, a protein involved in initiation and progression of myeloid leukemia [51], while 'Pou6f1', a POU domain class 6 transcription factor 1 encoding gene is GO annotated within the area of regulation of transcription, functioning as a DNA binding protein [52].

3.6. Conclusion

We were able to dissect gene regulation from mouse ESC differentiating to endoderm and mesoderm. We calculated the most regulatory active genes not only for early and late embryonic developmental stages, benefitting from the sub segregation using the CD24 as marker, but also for mesodermal and endodermal stem cells, using Foxa2 und T as markers. We found out, that Foxa2 and T positive samples behave like late endodermal cells and that CD24 indeed distinguish early and late endoderm and not as assumed in vitro measurement time dependent embryonic stem cell development. The problem with not expressed endoderm differentiation core signaling factors of EMT is still present. We did not find any genes involved in EMT, nor genes that could be regulating the reverse transition process. On the other hand, high regulated genes, don't need to be the most functional important ones. If a biological process is developed in a efficient way, core signaling proteins, don't need to be expressed in a high level. I strongly suggest to investigate the found genes more profoundly. These genes may have a still unknown important regulatory role in embryonic stem cell development, but surely have a still hidden complex relationship. Also a 'reverse search', meaning the search of already known EMT-related proteins within the resulting data, would give better outlook, where to possibly find the interesting genes for the reverse EMT process.

Bibliography

[1] Alexander Maximow. Der Lymphozyt als gemeinsame Stammzelle der verschiedenen Blutelemente in der embryonalen Entwicklung und im postfetalen Leben der Säugetiere. *Folia Haematologica*, 8:125–134, 1909.

[2] John Lynch. Stem Cells and the Future of Regenerative Medicine. *Journal of the National Medical Association*, 97:1041, 2005.

[3] Joseph Dosch, Cheong Jun Lee, and Diane M Simeone. Cancer Stem Cells: Pancreatic Cancer. *Stem Cells and Cancer*, 414(November):105–111, 2009.

[4] Biologiezentrum Linz. Morphologie : HAECKELS Gastraea-Theorie und ihre Folgen. 131(131):147–168, 1998.

[5] `http://ohsu2015.wikispaces.com/file/view/zygote_to_blastocyst.gif/248532519/zygote_to_blastocyst.gif`.

[6] Patrick P L Tam and David a F Loebel. Gene function in mouse embryogenesis: get set for gastrulation. *Nature reviews. Genetics*, 8(5):368–381, 2007.

[7] `http://philschatz.com/biology-concepts-book/resources/Figure_18_02_03.jpg`.

[8] Kohei Ota, Matsui Makoto, Edgar L. Milford, Glenn a. Mackin, Howard L. Weiner, and David a. Hafler. Â© 19 90 Nature Publishing Group. *Letters To Nature*, 346:183–187, 1990.

[9] Afra Hadjizadeh and Charles J Doillon. Directional migration of endothelial cells towards angiogenesis using polymer fibres in a 3D co-culture system. *Journal of tissue engineering and regenerative medicine*, 4(7):124–128., 2009.

[10] Chris Showell, Olav Binder, and Frank L Conlon. T-box genes in early embryogenesis. *Developmental dynamics : an official publication of the American Association of Anatomists*, 229(1):201–218, 2004.

[11] Kimberly E. Inman and Karen M. Downs. Localization of Brachyury (T) in embryonic and extraembryonic tissues during mouse gastrulation. *Gene Expression Patterns*, 6(8):783–793, 2006.

[12] Haiyan Wang, Benoit R. Gauthier, Kerstin a. Hagenfeldt-Johansson, Mariella Iezzi, and Claes B. Wollheim. Foxa2 (HNF3??) controls multiple genes implicated in metabolism-secretion coupling of glucose-induced insulin release. *Journal of Biological Chemistry*, 277(20):17564–17570, 2002.

[13] Xiaoju Tang, Xiaojing J. Liu, Cuijie Tian, Qiaoli Su, Yi Lei, Qingbo Wu, Yangyan He, Jeffrey a. Whitsett, and Fengming Luo. Foxa2 regulates leukotrienes to inhibit Th2-mediated pulmonary inflammation. *American Journal of Respiratory Cell and Molecular Biology*, 49:960–970, 2013.

[14] Elizabeth D. Wederell, Mikhail Bilenky, Rebecca Cullum, Nina Thiessen, Melis Dagpinar, Allen Delaney, Richard Varhol, Yongjun Zhao, Thomas Zeng, Bridget Bernier, Matthew Ingham, Martin Hirst, Gordon Robertson, Marco a. Marra, Steven Jones, and Pamela a. Hoodless. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Research*, 36(14):4549–4564, 2008.

[15] `http://www.intechopen.com/source/html/18233/media/image2.jpeg`.

[16] `http://apbrwww5.apsu.edu/thompsonj/Anatomy%20&%20Physiology/2010/2010%20Exam%20Reviews/Exam%201%20Review/04-13_EmbryoTissue_1.JPG`.

[17] Ingo Burtscher, Silvia Engert, Stefan Hasenöder, Daniela Padula, and Heiko Lickert. Embryonalentwicklung: Steuerungsmechanismen der entodermentwicklung in der maus. *BioSpektrum*, 17:520–523, 2011.

[18] Jian Xu, Samy Lamouille, and Rik Derynck. TGF-beta-induced epithelial to mesenchymal transition. *Cell research*, 19(2):156–172, 2009.

[19] Raghu Kalluri. EMT: When epithelial cells decide to become mesenchymal-like cells. *Journal of Clinical Investigation*, 119(6):1417–1419, 2009.

[20] Raghu Kalluri and Robert a Weinberg. Review series The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation*, 119(6):1420–1428, 2009.

[21] Yang Liu and Pan Zheng. CD24: a genetic checkpoint in T cell homeostasis and autoimmune diseases. *Trends in Immunology*, 28(7):315–320, 2007.

[22] Lina-Marcela Diaz-Gallo, Luz María Medrano, María Gómez-García, Carlos Cardeña, Luis Rodrigo, Juan Luis Mendoza, Carlos Taxonera, Antonio Nieto, Guillermo Alcain, Ignacio Cueto, Miguel a. López-Nevot, Elena Urcelay, and Javier Martin. Analysis of the influence of two CD24 genetic variants in Crohn's disease and ulcerative colitis. *Human Immunology*, 72(10):969–972, 2011.

[23] J a Zarn, S M Zimmermann, M K Pass, R Waibel, and R a Stahel. Association of CD24 with the kinase c-fgr in a small cell lung cancer cell line and with the kinase lyn in an erythroleukemia cell line. *Biochemical and biophysical research communications*, 225(2):384–391, 1996.

[24] Niko P. Bretz, Alexei V. Salnikov, Kai Doberstein, Natalio Garbi, Volker Kloess, Safwan Joumaa, Inna Naumov, Louis Boon, Gerhard Moldenhauer, Nadir Arber, and Peter Altevogt. Lack of CD24 expression in mice reduces the number of leukocytes in the colon. *Immunology Letters*, 161(1):140–148, 2014.

[25] Elena Israel, Joseph Kapelushnik, Tikva Yermiahu, Itai Levi, Isaak Yaniv, Ofer Shpilberg, and George Shubinsky. Expression of CD24 on CD19-CD79a+ early B-cell progenitors in human bone marrow. *Cellular Immunology*, 236(1-2):171–178, 2005.

[26] P O Brown and D Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genet.*, 21:33–37, 1999.

[27] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael a Irizarry, Michael Lawrence, Michael I Love, James Macdonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Publishing Group*, 12(2):115–121, 2015.

[28] Ecma International. The JSON Data Interchange Format. (October):1–14, 2013.

[29] Florian Hahne, Wolfgang Huber, Robert Gentleman, and Seth Falcon. *Bioconductor Case Studies*. Springer New York, New York, NY, 2008.

[30] Genevestigator Documentation. 2015-03-30.

[31] Rafael a Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, 2003.

[32] Speed T P Bolstad, Irizarry. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[33] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572, 1901.

[34] I. Jolliffe. Principal component analysis, second edition. page 518, 2002.

[35] Brian S Everitt Hothorn and Torsten. A Handbook of Statistical Analyses Using R. pages 1–207, 2006.

[36] Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. 3(1):1–26, 2004.

[37] Julie Cocquet, Maelle Pannetier, Marc Fellous, and Reiner a. Veitia. Sense and antisense Foxl2 transcripts in mouse. *Genomics*, 85(5):531–541, 2005.

[38] S. Ramamurthy, E. Chang, Y. Cao, J. Zhu, and G. V. Ronnett. AMPK activation regulates neuronal structure in developing hippocampal neurons. *Neuroscience*, 259:13–24, 2014.

[39] Gene Id. `http://www.ncbi.nlm.nih.gov/gene/388962`. pages 1–7, 2015.

[40] More Resources, Submit Data, Analysis Tools, and Contact Us. `http://www.informatics.jax.org/marker/MGI:2429943`. (24):4–5.

[41] C. Cenciarelli, D. S. Chiaur, D. Guardavaccaro, W. Parks, M. Vidal, and M. Pagano. Identification of a family of human F-box proteins. *Current Biology*, 9(20):1177–1179, 1999.

[42] a Scorilas, L Kyriakopoulou, D Katsaros, and E P Diamandis. Cloning of a gene (SR-A1), encoding for a new member of the human Ser/Arg-rich family of pre-mRNA splicing factors: overexpression in aggressive ovarian cancer. *British journal of cancer*, 85(2):190–198, 2001.

[43] More Resources, Submit Data, Analysis Tools, and Contact Us. `http://www.informatics.jax.org/marker/MGI:2141658`. (40):4–5.

[44] Andrej-Nikolai Spiess, Norbert Walther, Nadine Müller, Marga Balvers, Christoph Hansis, and Richard Ivell. SPEER–a new family of testis-specific genes from the mouse. *Biology of reproduction*, 68(6):2044–2054, 2003.

[45] I Rodriguez, P Feinstein, and P Mombaerts. Variable patterns of axonal projections of sensory neurons in the mouse vomeronasal system. *Cell*, 97(2):199–208, 1999.

[46] More Resources, Submit Data, Analysis Tools, and Contact Us. `http://www.informatics.jax.org/marker/MGI:2443451`. (27):4–5.

[47] More Resources, Submit Data, Analysis Tools, and Contact Us. `http://www.informatics.jax.org/marker/MGI:1916400`. pages 15–16.

[48] Xuemei Wu, Maria M Viveiros, John J Eppig, Yuchen Bai, Susan L Fitzpatrick, and Martin M Matzuk. Zygote arrest 1 (Zar1) is a novel maternal-effect gene critical for the oocyte-to-embryo transition. *Nature genetics*, 33(2):187–191, 2003.

[49] a B Bakker, E Baker, G R Sutherland, J H Phillips, and L L Lanier. Myeloid DAP12-associating lectin (MDL)-1 is a cell surface receptor involved in the activation of myeloid cells. *Proceedings of the National Academy of Sciences of the United States of America*, 96(17):9792–9796, 1999.

[50] `http://www.informatics.jax.org/marker/MGI:3643972`. page 3643972, 2015.

[51] M. Fereshteh, T. Ito, J. J. Kovacs, C. Zhao, H. Y. Kwon, V. Tornini, T. Konuma, M. Chen, R. J. Lefkowitz, and T. Reya. Â -Arrestin2 mediates the initiation and progression of myeloid leukemia. *Proceedings of the National Academy of Sciences*, 109(31):12532–12537, 2012.

[52] More Resources, Submit Data, Analysis Tools, and Contact Us. `http://www.informatics.jax.org/marker/MGI:102935`. pages 20–21.

List of Figures

A. Appendix