LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München
Institut für Bioinformatik und Systembiologie

Diplomarbeit
in Bioinformatik

Single-cell analysis of multipotent hematopoietic progenitor cells

Michael Schwarzfischer

Aufgabensteller:	Prof. Dr. Dr. Fabian Theis
Betreuer:	Jan Krumsiek, Carsten Marr
Abgabedatum:	15.12.2009

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

15.12.2009 ______________________________
Michael Schwarzfischer

Abstract

Hematopoiesis is the process of generating distinct blood cells in higher eukaryotes and keeping them in balance. This procedure is accomplished by hematopoietic stem cells which can differentiate into more specialized progenitor cells which further differentiate in fully functional blood cells. In this work, we focus on a particular differentiation step and try to elucidate molecular mechanisms behind it. The transcription factor PU.1 plays a major role in the process of myeloid progenitors differentiating into granulocyte/macrophage-progenitors or erythrocyte/megakaryocyte-progenitors. In our experiments we monitor common myeloid progenitor cells of mice with YFP-tagged PU.1 proteins. With fluorescence image time-lapse microscopy it is possible to observe the process of cell differentiation in individual cells. In this thesis we present a normalization method of fluorescence images in order to the clean cellular expression signal of PU.1. A custom software is developed which applies a cell detection algorithm and offers an easy way to correct for automatic detection errors in order to get most accurate expression data. Furthermore, we present several methods to normalize and visualize single-cell time courses and analyze correlations in general properties like protein production rates and cell life times. We investigate if a certain amount of PU.1 is maintained over several cell generations (memory effect). A specific surface marker which only appears in granulocyte/macrophage progenitor cells and their offspring allows to investigate whether a particular expression pattern can lead to a lineage commitment. Finally, we show that myeloid progenitor cells divide symmetrically and estimate the absolute protein amount of PU.1 in the cells.

Zusammenfassung

Unter Hämatopoese versteht man die Blutbildung in höheren Eukaryoten, in welcher das Gleichgewicht zwischen verschiedenen Blutzellen gehalten werden muss. Dafür verantwortlich sind Blutstammzellen, die im Knochenmark angelagert sind und über mehrere Reifungsschritte zu verschiedenen Vorläuferzellen und letztendlich zu den Blutzellen werden. In dieser Arbeit konzentrieren wir uns auf einen bestimmten Teil der Blutzelldifferenzierung und untersuchen die zugrunde liegenden molekularen Mechanismen. Der Transkriptionsfaktor PU.1 spielt eine maßgebliche Rolle in der Differenzierung der myeloiden Vorläuferzellen und führt bei hoher Expression zur Differenzierung in Granulozyt-/Makrophagenvorläufern, während bei geringer Expression Erythrozyt/Megakaryozyt-Vorläufer entstehen. In unseren Experimenten beobachten wir myeloide Vorläuferzellen in transgenetisch veränderten Mäusen, die ein fluoreszenzmarkiertes PU.1 Protein exprimieren. Die Vorgängerzellen können mittels zeitaufgelöster Mikroskopieverfahren über längere Zeit beobachtet und der Reifungsprozess von individuelle Zellen explizit verfolgt werden. In dieser Arbeit untersuchen wir zunächst die Eigenschaften von Fluoreszenzbildern und stellen eine Normalisierungsmethode vor, die Lampenflackern und Ausleuchtungsungenauigkeiten korrigiert. In einem von uns entwickeltem Programm, wird im weiteren Verlauf eine automatische Zellerkennung angewandt, die eine einfache manuelle Nachkorrektur der automatischen Erkennungsprozesse ermöglicht. Wir stellen unterschiedliche Normalisierungs- und Visualisierungsmethoden zum Vergleich von Expressionsverläufen einzelner Zellen vor und untersuchen grundlegende Eigenschaften wie Lebensdauer oder Protein-Produktionsraten. Wir beschreiben einen charakteristischen Expressions-Verlauf über den Zellzyklus einer Zelle hinweg und ermitteln, ob ein bestimmtes PU.1 level über mehrere Generationen hinweg erhalten bleibt (memory effect). Im Hinblick auf eine spezifische Zelloberflächenmarkierung, die erstmals in Granulozyt-/Makrophagen-Vorläufern auftritt, untersuchen wir, ob spezifische Expressionsmuster in den Zellen Hinweise auf die Linienentscheidung geben. Schließlich zeigen wir, dass sich myeloide Vorläuferzellen symmetrisch teilen und wenden daraufhin eine statistische Methode zur Abschätzung der PU.1 Proteinanzahl in den Zellen an.

Acknowledgments

Thanks to Jan Krumsiek, Carsten Marr and Fabian Theis for being great supervisors. Thanks to Philipp Hoppe and Timm Schroeder for the experiments and valuable feedback. Thanks to my roomie and mathematical advisor Daniel Schmidl and the whole CMB group for the inspiring working atmosphere.

Special thanks to Jan, my friends and my family for the abundance of patience and support.

Contents

Chapter 1

Introduction

Hematopoiesis is the process of building all blood cell types and keeping them in balance in order to continuously maintain the blood system [35]. The word hematopoiesis is derived from Ancient Greek (haima = blood, poiesis = to make). The complicated process of maintaining a defined composition of blood cells in homeostasis as well as in stress situations has to be managed. The continuous destruction of millions of blood cells in every second must be countervailed by a sophisticated procedure of regenerating new cells [46]. There is great effort in this research field due to the biomedical impact of hematological diseases like leukemia or anemia. One major goal is the understanding of the regulatory processes of blood stem cell differentiation. Over the last 50 years hematopoiesis has become one of the best studied mammalian systems [14]. Despite this long period of research there is still incomplete knowledge about the molecular mechanisms of differentiation [9].

New technologies allow to analyze single-cells of a specific cell type over time on a molecular level. Time-lapse microscopy and protein labeling techniques make it possible to investigate the expression of certain proteins [54]. Creating movies of a cell population permits tracking of individual cells, leading to observations of their properties like cell fate or lifetime. Additionally, measurements of protein levels give information of the real-time protein expression which could explain cell-cycle dynamics and decision making in differentiating cells [49].

1.1 The blood system

During embryonic development a very small population of hematopoietic stem cells (HSCs) is built up in different anatomical sites which later colonize the bone marrow [32, 7]. HSCs do not fulfill directly the request of the blood system but are responsible for the constant renewal of mature blood cells [52]. HSCs can be furthermore distinguished into two subclasses: cells in the first class remain in a steady-state, have a very low self-renewal rate, divide only once in around 57 days and are therefore called long-therm HSCs (LT-HSCs). A LT-HSC can become more specialized by differentiating into a short-therm HSC (ST-HSC) which has lost the feature of self renewal and will differentiate into a more specific cell. ST-HSCs have the capability to differentiate into all mature cell types and are also called multipotent progenitor cells (MPPs). Differentiated blood cells can be categorized into myeloid and lymphoid lineages. T, B and natural killer (NK) cells are known as lymphoid cells whereas erythrocytes, megakaryocytes, granulocytes (which can be distinguished into neutrophils, eosinophils and basophils) and macrophages belong to the myeloid lineage [22]. The development of MPPs into specific mature cells is guided by the specific regulation of several transcription factors [25]. The complete commitment to a specialized cell type passes through several intermediate steps [37]. A first distinction is given by differentiating into common myeloid progenitors (CMPs) or common lymphoid progenitors (CLPs). These again can become more specialized by differentiating one step further. CMPs will give rise to megakaryocyte/erythrocyte-progenitors (MEPs) or granulocyte/macrophage-progenitors (GMPs) [2]. CLPs differentiate into T, B and NK progenitor cells (Figure 1.1).

To understand the mechanisms of the blood system all kinds of blood cells including HSCs are required for experiments. Since there are only few HSCs in the bone marrow (about 1 HSC in 10.000 bone marrow cells) and even less in the blood stream (about 1 HSC in 100.000 blood cells) the research on the human blood system is challenging and expensive [1]. Investigating hematopoiesis in model organisms like zebrafish or mouse provides easier experimental conditions. Furthermore, results of these organisms are transferable on the human blood system since the process of hematopoiesis is generally conserved throughout vertebrates [37].

1.2 Lineage-specific surface markers

Every specific cell type ranging from hematopoietic stem cells to mature blood cells has different morphological and phenotypic properties and can readily be identified by these characteristics. Cells committed to GM lineage will develop FCγ receptors [15, 47] whereas CD150 is a highly expressed surface molecule in HSCs and the erythroid lineage but repressed in the myeloid lineage [41]. FC receptors are required to fulfill the role of monocytes or macrophages in the immune system [21].

These proteins are accessible from the surface and bind to certain antibodies. Therefore, the surface proteins can be used to visualize the cell lineage by adding colored antibodies into the culture medium. The anti-

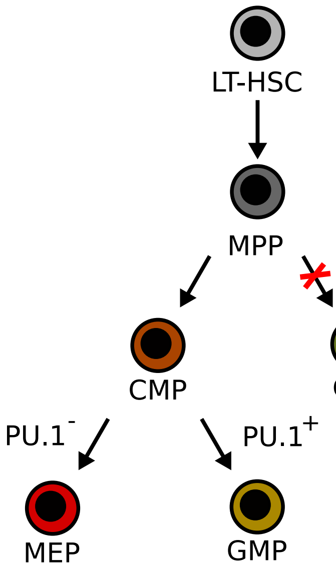


Figure 1.1: Illustration of the differentiation process of hematopoiesis. Long-term hematopoietic stem cells (LT-HSCs) become more specialized by differentiating into multipotent progenitor cells (MPPs). These can further differentiate into common myeloid progenitor cells (CMPs) or common lymphoid progenitor cells (CLPs). CMPs can give rise to megakaryocyte/erythrocyte progenitors (MEPs) or granulocyte/macrophage progenitors (GMPs). In this project we focus on the PU.1 expression in MPPs and CMPs which plays a major role in the lineage decision between MEPs and GMPs. The lineage of CLPs cannot be investigated due to experimental conditions.

bodies will colonize around the corresponding membrane proteins and and radiate if excited with a specific wavelength.

1.3 The myeloid lineage decision

The transcription factors PU.1 and GATA-1 play major roles in the lineage decision of CMPs. The interaction of these two factors is a highly discussed topic in the research field [9, 19]. The DNA binding protein GATA-1 is required for the development of erythrocytes and megakaryocytes [36]. It has been shown that over-expression of GATA-1 induce the expression of erythroid-megakaryocyte-affiliated lineage markers and furthermore repress the monocytic and granulocytic markers. GATA-1 competes with the Ets family transcription factor PU.1 which is required for non-lymphoid leukocyte cell development [13]. A high expression of PU.1 even in erythroid-megakaryocyte progenitors induce a conversion into monocytic lineage by repressing GATA-1 [53, 16]. PU.1 and GATA-1 interact physically and inhibit each other's transcriptional activity [57]. In addition both factors also control their own expression, forming autoregulatory loops [48, 6]. As long as the balance of these two players is kept the cell stays in CMP state otherwise a winner arises which will make the the cell differentiate either into ME or the GM lineage [34].

There exist several different mathematical approaches which attempt to model the characteristics of this lineage decision. A model formalizes the interaction and represents a simplified system which can easily be modified. Several combinations of inhibitions or activations can be simulated in silico in order to gain a model which describes the biological observations best possible [44, 20, 8]. Most of the models describing genetic switches are based on differential equations suitable if many molecules (transcription factors) are present in a cell [17]. But there are also other approaches which include probabilities and noise leading to stochastic modeling since many lineage decisions are based on stochastic processes [29].

1.4 The PU.1-YFP mouse

In this study we focus on the transcription factor PU.1. We study mice where PU.1 is tagged with a fluorescent marker, namely the yellow fluorescent protein (YFP). YFP is a mutant of the green fluorescent protein (GFP) and is a widely used reagent [12, 4]. The technique of tagging proteins is a long biochemical process where the protein of interest is separated from the DNA by several restriction enzymes. After combining it with the DNA of an fluorescent protein and amplifying the tagged protein its natural function has to be tested. Since fluorescent proteins are known to be not toxic, the technique of fluorescence tagging allows to express a labeled protein in living cells [30]. After assuring that a cell with the reinserted tagged protein still remains its natural function it is possible to express this protein even in animals without changing the common phenotype. From the measured fluorescent signal one can now directly infer the expressed amount of tagged protein as we assume the intensity grows proportional to the number of proteins.

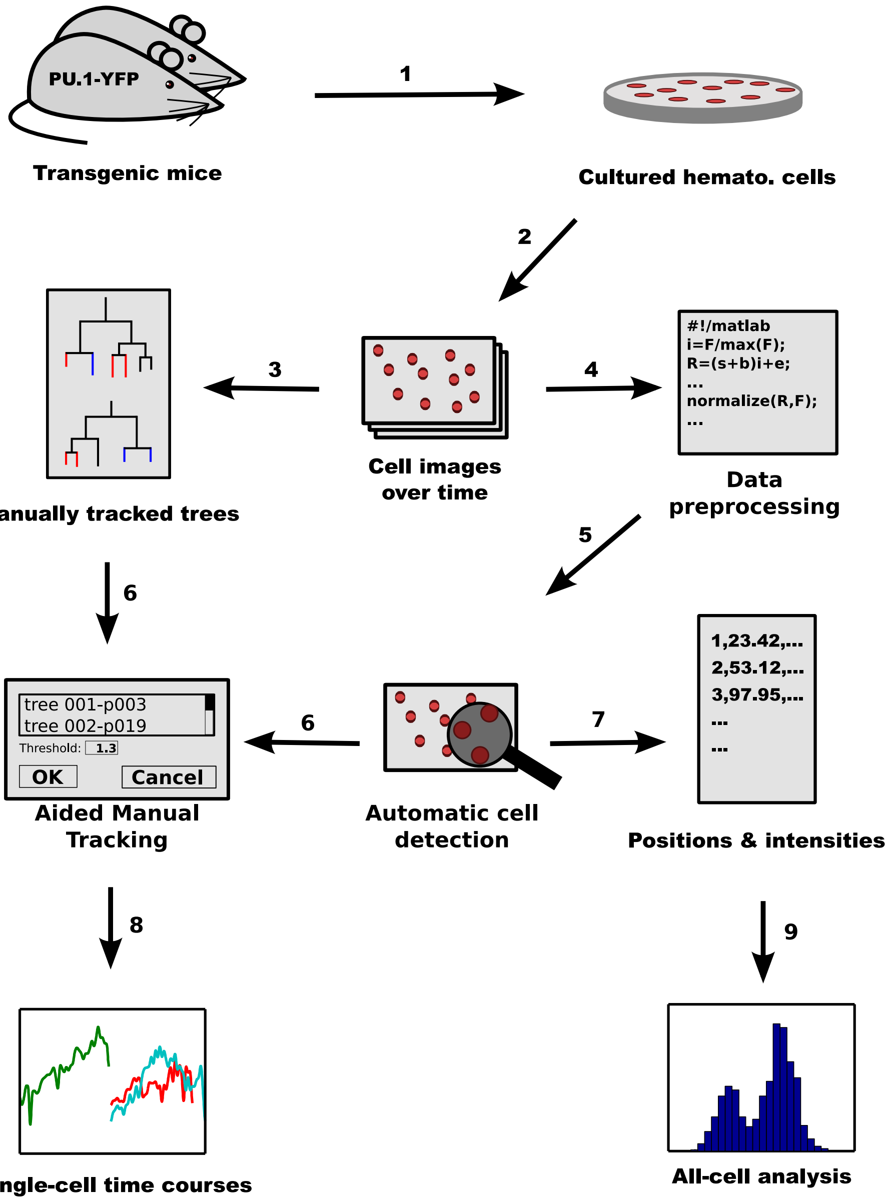
In our experiments the YFP tagged PU.1 transcription factor has been inserted into the model organism mouse which is healthy and has no strange phenotype [24]. The mice are bred over several weeks and their hematopoietic cells are extracted from the femoral bone-marrow. These cells are sorted in order to gain only MPPs.

The surface proteins described above are also used in fluorescence-activated cell sorting (FACS) technique where different cells can be sorted in a high-throughput manner by flow cytometry [55]. Colored antibodies are added to the cell medium which bind to cell specific surface markers which can then be excited. Ranges of intensities as well as cell size and granularity are defined which can be used to separate the different healthy cell types.

1.5 General workflow

We focus only on a small part of the whole differentiation process and concentrate on the event of MPPs differentiating into MEP or GMP cells. The technology of live-cell imaging uses fluorescently tagged PU.1 MPPs to investigate this particular differentiation process. This work is a continuation of a former diploma thesis performed by Jan Krumsiek [26]. There, a qualitative model of a regulatory interaction network containing the main myeloid development players as well as a quantitative model of the PU.1-GATA-1 switch was presented. A first examination of the early live-cell imaging experiments was discussed including an image analysis pipeline. Here, we developed better methods to improve the imaging processing and cell detection technique and present new results.

The different parts of the present work are incorporated in a workflow illustration (Figure 1.2). Step 1 is the sample preparation already described in this Chapter whereas Chapter 2 discusses the techniques of live cell imaging and tracking (step 2 and step 3) and several normalization methods (step 4 and step 5). The later process of automatic cell detection and measurement tuning (step 6 and step 7) is described in Chapter 3. The analysis and results of population wide-statistics and single-cell time courses (step 8 and step 9) are finally shown in Chapter 4. We present distributions of cell populations over time and several single-cell time course representations. We investigate the cell-cycle as well as cell lifetime statistics. Different methods are shown of finding particular FCγ dependent PU.1 expression profiles and a number of analyses on the tree level are presented. The final Chapter 5 gives a summary of the work we presented and discusses further investigations.



```
#!/matlab
i=F/max(F);
R=(s+b)i+e;
...
normalize(R,F);
...
```

```
tree 001-p003
tree 002-p019
Threshold: 1.3
OK    Cancel
```

```
1,23.42,...
2,53.12,...
3,97.95,...
...
...
```

Figure 1.2: The general workflow of this thesis. Step 1 to 3 are performed by the Schroeder group at the Institute of Stem Cell Research. First the transgenic mice are bred and their hematopoietic stem cells are extracted (1). These cells are cultured in medium and imaged by time-lapse microscopy (2). The cell images are tracked by a custom software leading to cell trees (3). The first step of our work is processing the images by several normalization steps (4). We then apply a detection algorithm (5). Our results are combined with the cell trees and used in a self developed toolbox (Aided Manual Tracking) to correct for detection errors (6). After cleaning up the data, single-cell time courses can be investigated (8). From the automatic detection of cells alone population wide statistics can be achieved (7 and 9).

Chapter 2

Processing of cell microscopy data

In the following chapter we present the experimental procedures (performed by the Institute of Stem Cell Research (ISF)) and processing pipeline (performed by our group). We highlight general issues of fluorescence images and present a correction method tailored to the experimental settings. Furthermore, we compare this correction method against previously published methods and present a mathematical model which estimates bleaching rates of fluorescence proteins.

2.1 Time-lapse microscopy

After the hematopoietic cells are extracted from the mouse bone marrow and sorted by FACS, the cells are grown in medium and observed with the new arising technology of time-lapse microscopy (Figure 1.2 step 2) [11, 28]. Images of the cells are taken in defined time intervals giving the opportunity to follow individual cells over time and create specific statistics on single-cell level [43]. It is also possible to integrate fluorescent images to measure the expression of fluorescently tagged proteins at single timepoints in order to create single-cell expression profiles. Therefore a tracking of single-cells is needed which is a highly discussed topic and many approaches and algorithms are developed for this purpose [33].

2.1.1 Live cell imaging at the ISF

In our experimental setup a small robot moves the growth medium under a camera initially containing multipotent progenitor cells (MPPs). The plate is divided into up to 39 overlapping zones (so called positions). Every position has its own coordinates depending on the pixel resolution where the cells later can be well identified. Every two minutes a brightfield image

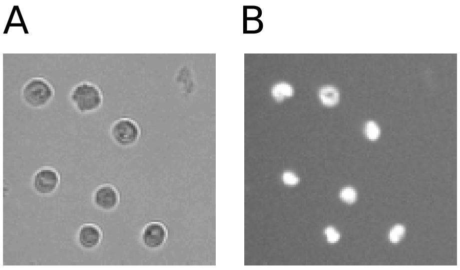


Figure 2.1: Small region from a brightfield image (A) and the corresponding fluorescence image (B) showing the intensity of YFP tagged PU.1 of several cells in the raw imaging data.

of each position is taken (for a detailed view, see Figure 2.1 A). Brightfield microscopy is an optical microscopy technique where the probe is illuminated by white light from behind. A limitation of this method is the low contrast in the taken pictures, but it is still sufficient to see the cells moving over time.

Fluorescence images are taken in a longer time interval since the irradiation with a specific wavelength, depending on its energy, might damage the cells. Therefore the YFP of the tagged PU.1 is only excited every 30 minutes to assure cell health leading to fluorescent pictures of every position (Figure 2.1 B).

There are more channels for measuring on different wavelengths. In our experiment two additional channels are used to identify colored surface markers. On a certain wavelength the surface marker FCγ can be observed which, if active, can be taken as are sure evidence for GMP lineage commitment. The other wavelengths of this experiment should show another surface marker CD 150 which is the equivalent player for the MEP lineage. This surface marker could not be used in our current study since there are still unsolved experimental issues. These surface marker images are only taken every $\approx$2 hours, still sufficient to capture the appearance of a surface marker.

The resulting data of a movie experiment leads to about more than 50 gigabytes of image files, containing 3000 to 5000 brightfield images, 250 fluorescence images and 65 images for each surface marker for every position. Each image has a resolution of about 1300x1100 pixels depending on the experiment. For this thesis eleven different experiments were conducted which are described in detail in Table 2.1. Due to the collaboration of the Institute of Stem Cell Research (ISF) with our group many technical and experimental issues have been identified and eliminated from experiment to experiment. In this thesis we focus on experiment three and eight since most of the tracking data is based on these experiments.

Experiment	Date	Description	Time	Positions	trees	size
1	29.12.08	initial movie	6.5d	36	20	47 GB
2	11.05.09	crashed after 2 days	2d	39	0	39 GB
3	28.05.09	new focusing method	4.5d	39	35	55 GB
4	29.05.09	fluorescine experiment	-	1	0	18 MB
5	16.07.09	bead experiment	4.5d	9	0	23 GB
6	28.07.09	fluorescine experiment 2	3.5d	1	0	3.4 GB
7	31.07.09	colibri experiments	5d	1	0	5.2 GB
8	01.09.09	latest movie	7d	39	34	91 GB
9	07.09.09	same as Exp 8	7d	39	0	96 GB
10	14.09.09	background movie	2d	19	0	24 GB
11	16.11.09	population specific images	-	39	0	600 MB

Table 2.1: All experiments available for this thesis. The table shows the date, a short description, the duration, the number of positions, the number of tracked trees and the size of each experiment. The first movie led to first impressions and showed some technical issues. A new focusing method was invented and used in experiment 2 and 3. Experiment 2 could not be used for further analysis as there were technical problems after two days. Experiments 4 to 7 were used to test for different illumination and light source techniques. The movies 8 and 9 combine all the improvements derived from the earlier experiments. Movie 10 gave evidence of bleaching of the background due to some auto fluorescence in the plastic of the experimental setup. Experiment 11 includes three population specific images of just one timepoint. The different cell types MPPs, MEPs and GMPs were sorted by FACS and immediately imaged in order to verify the imaging and detection technique. The main focus of this work lies on experiment 3 and 8 since the most tracking was performed on these experiments.

2.1.2 Tracking data

From the raw movie data further processing steps are performed. The ISF has implemented a custom tracking software called Timm's Tracking Tool (TTT, step 3). Researchers at the ISF play back the experiment movie and accurately follow the cells over their lifetime. Important facts like the accurate location in the movie described by the position (referring to the 39 sections of the probe), the position-specific coordinates (pixel coordinates of the position image) and the timepoint as well as additional annotations like the status of surface markers are captured. All occurring events like cell division, cell death or differentiation are also written down. The resulting tracking of the first *mother cell* and all its progeny (*daughters*) leads to cell *trees*, also called *genealogies* [18]. At every timepoint each cell is well-defined by its position and the position-specific pixel coordinates. Since living cells

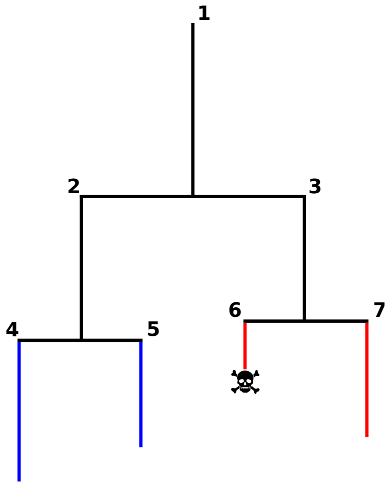


Figure 2.2: The tree illustration of a cell with all its progeny and their cell fates over time (y-axis). Different colors indicate different cell lineages, the skull stands for cell death. Numbers are assigned by giving left cells $n \cdot 2$ and right cells $n \cdot 2 + 1$ whereas n represents number of mother cell. The first cell is defined as cell 1.

always divide into two descendants such trees can easily be represented as a binary tree (see Figure 2.2). The cell numbers are assigned as follows: The first mothercell is defined as cell one. Every daughter cell on the left will get the number of its mothercell (n) multiplied by two ($n \cdot 2$), every right cell will get $n \cdot 2 + 1$. This method is commonly used in informatics and assures that every cell gets a unique number. The first mothercell is also called generation zero (cell 1), its daughters are generation one (cell 2 and 3), its grandchildren generation two (cell 4 to 7) and so on.

The amount of tracked generations, defining the tree depth, varies depending on several events like e.g. cell death. The most limiting factor is the simple case if a cell is already committed to a certain linage which can be determined by looking at the aforementioned surface markers. Further tracking is not necessary since the interesting part of lineage decision is completed.

The tree files are exported as CSV files and used for our own tools in a later process corresponding to step 6 in the workflow. Every tree file contains cell number, timepoint, position, the coordinates and the annotation informations.

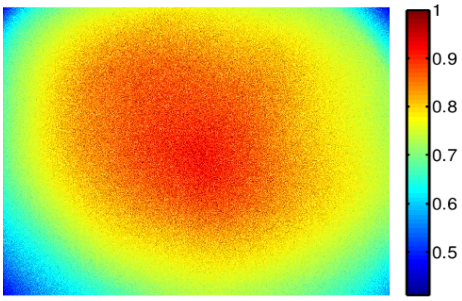


Figure 2.3: Illumination of a fluorescence image derived by imaging a fluorescine dilution of experiment 6. The color represents the measured intensity ranging from blue to red. The image should show a homogeneous intensity over the whole area. This image is used to correct for the obvious uneven illumination in the other experiments.

2.2 Processing of fluorescence images

Measuring the exact fluorescence intensity of a single-cell is a commonly known task [50, 10] where one has to regard several difficulties. First the quality of the image depends on the right focus of the microscope to record an exact measurement of the fluorescent proteins of the cell. As the cell is a three dimensional element and an image will only present the two dimensional mapping there will always be some information loss. We assume linearity of fluorescence which will countervail this effect on fluorescence images. We expect that a longish cell should glow approximately the same as if the cell lies or stands in the taken picture. Thus the two dimensional fluorescence signal appearing on the image corresponds to the concentration of the whole cell.

Another problem on the experimental side is to create an even illumination in order to make measured fluorescence intensities comparable over the whole field of view. A conventional lightsource will create an uneven illumination unless a laser or LED technique is used. In our experiment a strongly uneven illumination can be seen which has to be corrected for (Figure 2.3 and 2.4). Furthermore we have to verify that the light source will continuously emit with the same intensity. It should not flicker or decrease over time.

In the culture medium of growing cells there are many chemical substances which also emit light upon excitation and influence the resulting image. As the medium is a fluid and its reagents diffuse quickly, we assume that the medium (background signal) illuminates with a constant value on every position at a given timepoint. If this background signal can be deter-

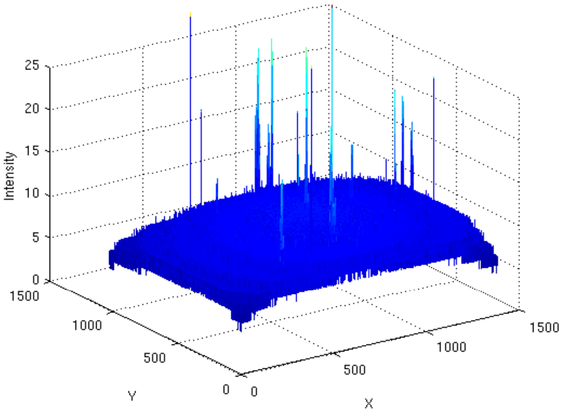


Figure 2.4: A raw fluorescent image containing cells. The x- and y-axis represent the image width and height whereas the fluorescence intensity can be observed on the z-axis. Issues like uneven illumination, the background signal and noise can be seen. The peaks in the image represent cellular signal.

mined one can easily subtract the factor from the entire image in order to normalize the real cell signal.

Working with experimental data will always include a certain amount of noise. This noise can be further distinguished into biological and technical noise. A certain amount of technical noise cannot be avoided and is caused by the imaging and detection techniques which is described in detail in Section 4.1. The biological noise cannot be avoided but this can also deliver conclusions of biological mechanisms.

Another problem of fluorescence imaging over time is the effect of photobleaching. Irradiating fluorescence molecules will always bleach some of them leading to less fluorescence in the next exposure step. The irradiation irreversibly changes the conformation of the fluorescence protein which can never be excited again. Imaging living cells with tagged proteins makes it more complex to compare different cells over time since there will always be a production of new molecules, so the signal will not bleach continuously until a steady state of the bleaching rate and protein decay against the production rate is achieved. The bleaching rate strongly depends on the exposure time and the time interval between images. In order to recalculate the real proportion of fluorescence proteins a model has been developed described in Section 2.5.

In order to measure fluorescence appropriately, all these issues have to be corrected for (compare Figure 2.4). In the later process automatic detection and tracking algorithms will be applied which obviously fail on unnormalized data due to the mistaking of background for cells.

In the later analysis steps it becomes even more important to have clean data, as a following example shows: When calculating an expression fold change the result would completely differ if the background level was not subtracted. An arbitrary expression data of two timepoints increasing from 3 to 6 represents a two-fold change. But if theres a background level of 2 in the data the fold change is from 1 to 4 which is four-fold. In the following Section we will present several methods to correct for these issues. These include our own methods as well as already existing methods from other working groups.

2.2.1 Uneven illumination

The illumination in our experimental setup is uneven due to the usage of an mercury-vapor lamp. The illumination distribution of this lamp can be shown in a simple experiment: Fluorescine is a commonly used fluorophore which can be excited and will always emit with a known intensity [51]. After diluting it with water, the feature of a fluid and the resulting diffusion of the molecules lead to a constant signal over a whole probe. A single static fluorescine image demonstrates the illumination distribution of the lamp (see Fig 2.3). Several test with different concentrations and different exposure times were examined (experiments 4,5 and 6) in order to get the right dilution of fluorescine which accurately estimates the illumination. In experiment 4 three different concentrations (1:100, 1:1.000 and 1:10.000) each with four different exposure times (1ms, 10ms, 100ms and 1000ms) were imaged. It turns out that a similar exposure time as of the live cell movies (1500ms) and a concentration of 1:10.000 accurately captures the illumination.

In experiment 5 a different approach to estimate the illumination was tested using fluorescent beads instead of fluorescine. These also have a constant fluorescence but could be added into the medium and could directly be imaged along with the cells. Therefore, the normalization factor of the illumination of a cell can be estimated by looking at the nearest bead. However, the beads rapidly bleach and have to be detected additionally, so this method is not very applicable for estimating the illumination.

The movie of experiment 6 imaged fluorescine which does not show fast bleaching effects over a longer time period. The constant concentration and constant exposure time leads to the relative illumination and furthermore to another important result. The overall shape of the illumination of the lamp over time stays the same and the difference of a late image against an early image is about 2%. This was validated by comparing the ratios of a small block of every corner against a block in the middle of the images.

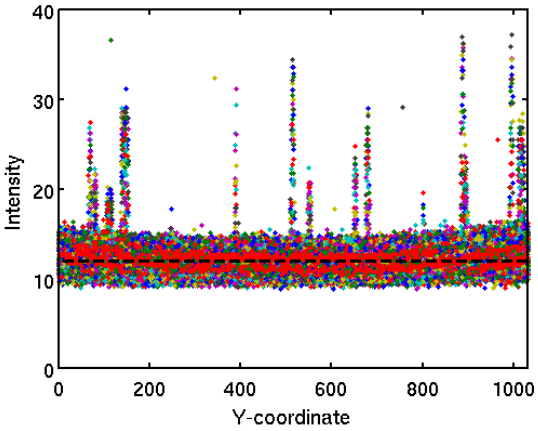


Figure 2.5: A projection onto the y-axis/intensity plane of an illumination corrected cell image, where each dot represents the intensity belonging to a row of the original image. The median of this image lies within the background whereas the cell signal peaks deviate from the band. The dots are arbitrarily colored for a better visualization.

Based on these results another experiment is necessary in order to give a proof of fluorescence linearity. This could be proven with a constant concentration and more exposure times between 500ms and 1500ms or with constant exposure time and different concentrations.

2.2.2 Determining the background signal

As mentioned before we assume an identical background signal over the whole image due to the background medium fluid. The measured background signal only depends on the illumination factor. Assuming that we corrected for that effect, we focus on estimating the background level. There are several possibilities to determine this level. An intuitive solution would be to search for background pixels in each image. If all cells were already detected we could scan within a small grid for an area without cells. This represents the background signal and one could subtract the mean or median of this area to eliminate the background level. An experimental approach is to create an entire movie under the same conditions as every other movie but without cells. This will only record the characteristics of the background and the culture medium. However, since the experiments depend on many variables (e.g. the age of the light source) this method would not be practically applicable.

We correct for the background signal with a different approach. Given a single illumination corrected image we assume that the median of this picture lies within the background. As long as there is less cellular signal than background signal in the image the median represents the background. This holds true in our experiment even up to the latest timepoints (later shown in Figure 3.4). The fact that the median is robust for outliers justifies taking the median in order to accurately estimate the background (Figure 2.5). Once the value of the background level is obtained, the value is subtracted from every pixel in the image, therefore all images are normalized to the same ground level.

Another big advantage of this method is that other influences such as flickering or fading of the lamp as well as bleaching of the background over time are eliminated as well. The lamp flickering and also the decrease of lamp intensity or bleaching of the background over time will cause the median value of each image to flicker or bleach accordingly. The median is calculated independently for each image so this method should always correct all images to the same ground level.

2.3 Correction method

The resulting normalization procedure for every single picture of every position is now composed of following steps: First we correct for the uneven illumination via an estimation with fluorescine. Afterwards the background is estimated and subtracted of the images.

The raw image is denoted as I_{xy} (Figure 2.6 A) where the range of x represents the image width and the range of y represents its height. Each image consists of the cell signal s_{xy}, a constant background signal b, both depending on the illumination effect f_{xy}, and some technical noise ϵ.:

$$I_{xy} = (s_{xy} + b)f_{xy} + \epsilon \tag{2.1}$$

The dependency on the illumination can be interpreted as a multiplicative factor since we assume a linear relationship between the measured intensities per pixel and the corresponding illumination. The noise is assumed not to be dependent on the illumination effect for simplification. It is evenly distributed over the whole position.

The fluorescine image F_{xy} with constant fluorescence signal s' and the same constant background b can be accordingly written as:

$$F_{xy} = (s' + b) \cdot f_{xy} + \epsilon = c \cdot f_{xy} + \epsilon \tag{2.2}$$

The sum of s' and $b$is here represented by a constant factor c since the decomposition is not possible. From F_{xy} we can infer the relative illumination factor f'_{xy} by dividing the whole image by the median of the 30 highest values. This should be a robust estimator of the maximum against the technical noise:

$$f'_{xy} := \frac{F_{xy}}{\max(F)} \approx \frac{c \cdot f_{xy}}{\max(c \cdot f)} = \frac{f_{xy}}{\max(f)} \tag{2.3}$$

The relative illumination factor f'_{xy} (Figure 2.6 B) should now precisely describe the pixel-specific illumination. The values theoretically range from 0 to 1, whereas in our case they range from ≈0.4 in the corners to 1 in the middle. This indicates that the signal in the corners is approximately more than two times darker than the signal in the middle of each position.

Now it is possible to normalize the illumination of the raw images by a pixelwise division with the relative illumination factors:

$$I'_{xy} = \frac{I_{xy}}{f'_{xy}} \tag{2.4}$$

I'_{xy} (Figure 2.6 C) represents the illumination-corrected image still containing the background signal. But all cell signals within one image are now on a comparable level.

The next step is to eliminate the background signal b. We estimate its value by taking the median of the I'_{xy} picture (Figure 2.6 D):

$$b \approx median\left(\frac{I_{xy}}{f'_{xy}}\right) \tag{2.5}$$

Combining the equations above the resulting cellular signal (Figure 2.6 E) can be derived as:

$$s_{xy} \approx I'_{xy} - b = \frac{I_{xy}}{f'_{xy}} - b \tag{2.6}$$

This calculation is done for every single fluorescence image of every position. Since we are working with MATLAB, which provides fast matrix manipulation methods, the normalization can be performed in ≈0.1 seconds per image. Therefore the normalized images are not stored separately and the calculation is processed on-the-fly.

2.3.1 Comparison to other correction methods

Sigal et al, 2006

In the work of Sigal et al [49], endogenous proteins in the nuclei of lung cell carcinoma cells are randomly tagged and cell-cycle dependencies of nuclear protein levels are investigated. Also time-lapse microscopy is used and the same issues have to be corrected for. They estimated the illumination also by different experiments using fluorescent fluids, here a dilution of 1:1000 GFP. For each pixel they performed a linear regression of the gray levels from different exposure times (ranging from 0 to 600 milliseconds). They calculated the offset and the gray level per millisecond of exposure time called the *gain*, which is the estimate for the illumination pattern. The normalized gain pattern was derived by dividing the gain pattern by its average over all pixels. Each image was *flat-field* corrected by subtracting the offset

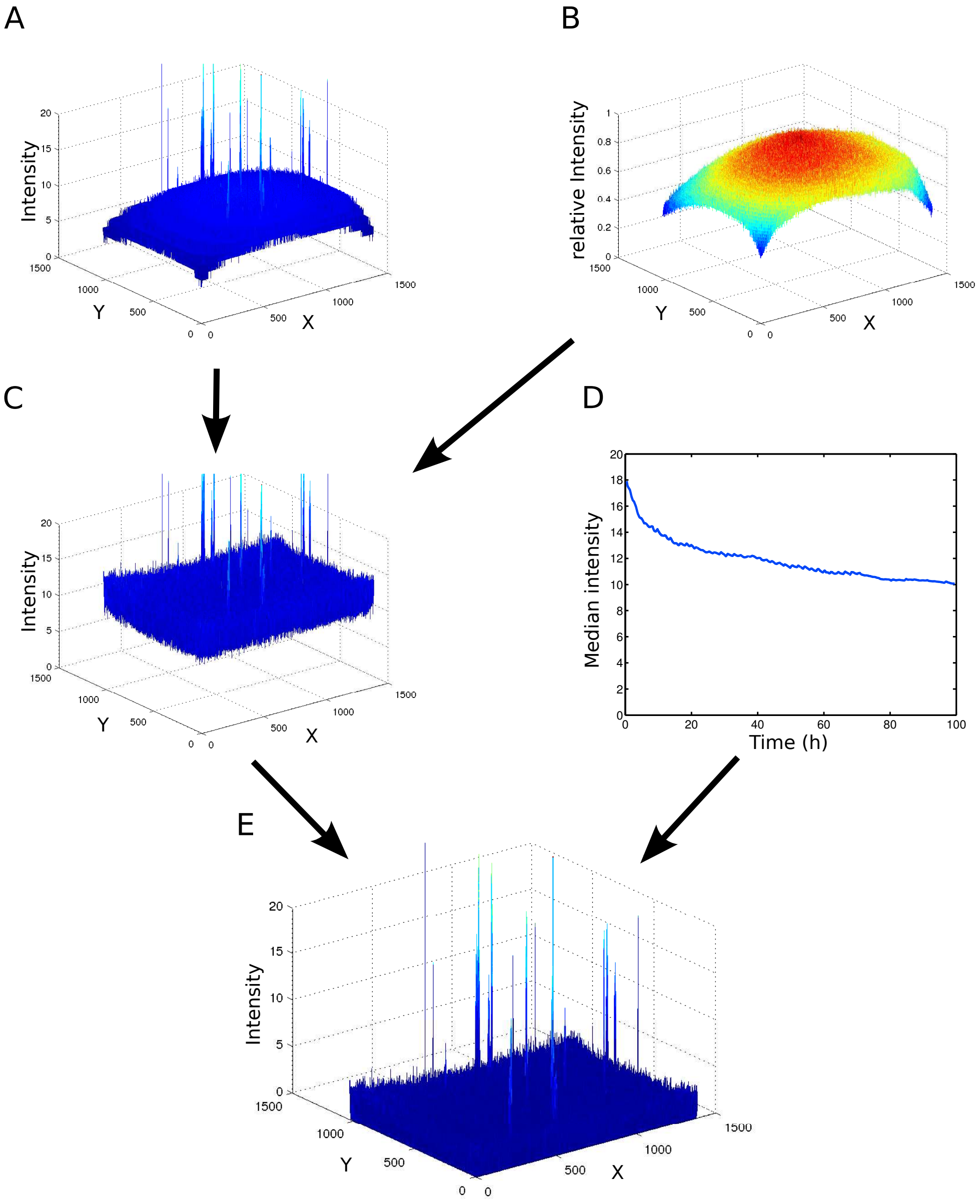


Figure 2.6: An illustration of the normalization workflow: A) The raw image containing illumination affected cellular and background signal and technical noise. B) Relative illumination derived from the fluorescine experiment. C) The illumination corrected image derived by a pixelwise division of the raw image by the fluorescine image. D) The developing median over the whole experiment 3. The bleaching of background, the fading of the lamp as well as the flickering can be seen. E) The resulting normalized image containing only cell signal and technical noise.

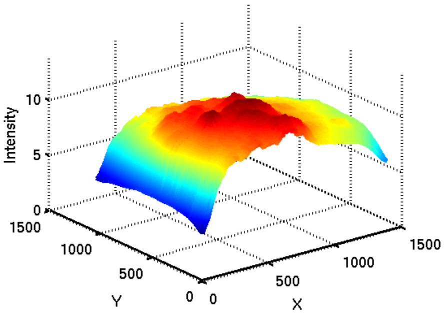


Figure 2.7: Background and illumination estimation via moving average over a raw image containing background as well as cell signal performed by Michel et al [31]. The disadvantage of this method can be seen as the development of small hills due to high cellular signal. Correcting with this image will lead to errors in the resulting cell intensities. A moving average with a higher two dimensional window size is less susceptible to this type of error but also estimates the illumination and background level less accurate.

of each pixel and by dividing by the normalized gain. They also assumed the linearity of fluorescence and that only the illumination has some effect on the measured intensity with different exposure times. Since our experiments are always performed with a constant exposure time, one estimation is sufficient to normalize every timepoint since the shape of illumination does not change. The method Sigal et al use could not yet be tested on our data since the necessary experiment with different exposure times has not been performed yet, but is planed.

After the normalization the background is estimated by segmenting the flat image into small blocks and by taking the histogram of appearing values. The gray level of the 10^{th} percentile is extracted and only these values remain in the block. They interpolated between them to refill the block which is then subtracted from the flat image. Our method of taking the median of the whole image seems to be slightly more robust as long as there is more background than cell signal. It could happen that the block is unfortunately chosen so that there are so many cells or contaminations that the estimated background is wrong. Spot testing on some images lead to no significant difference in resulting measurements. In the case of Figure 3.4 B the median overlaps with the 10th percentile.

Michel et al, 2007

A completely different approach to identify the uneven illumination was proposed by Michel et al [31]. It is possible to estimate the illumination along with the background signal by applying an two dimensional moving average with a given window size (Figure 2.7). This is only applicable if there are very few cells and accordingly very few corresponding intensity peaks in the image. Depending on the amount of cells in the images, a different window size for the algorithm has to be chosen. After subtracting the calculated illumination shape the background signal within the image should also be eliminated.

Again this method would not be applicable in the later movie phase where the cell signal amount increases (see Figure 3.4). In the beginning of the movie the results again do not differ from our method (data not shown). Furthermore the computing the moving average takes much more time than our method.

2.4 Experimental setup changes

At the very beginning of this project we did not expect how severely the illumination and the technical noise affect the movie data. The close collaboration between our group and the ISF led to important setup changes and technical methods to countervail these effects. The first improvement was the illumination correction method via diverse experiments with fluorescine (Table 2.1 experiment 4, 5 and 6 described in 2.2.1).

Another improvement was the adding of small plastic beads on which the microscope focuses in order to get the same image quality at every position and experiment (this is performed since experiment 3).

The ISF tested another light source for a few weeks called the *colibri* which is based on LED technique. With this system a more even illumination and a better signal to background ratio should be achieved. After some testing the illumination could not be improved significantly. Only in future projects when a higher time resolution of fluorescent images is necessary, this equipment could be required since LED technique is less damaging to the cells. Another experimental change was inferred from the median intensity over time depending on the auto fluorescence of the medium. (Figure 2.6 D). The sharp drop of intensity at the very beginning has to belong to the phenol red in the medium which acts as an pH indicator and bleaches rapidly. To countervail this effect, the medium is preirradiated leading to an almost flat line in the newest movie (Figure 2.8).

2.5 Bleaching model

Before the aforementioned method of normalization was derived and the background could not be corrected for, simply because we did not consider the effect of auto-fluorescence in the background. Therefore one could see

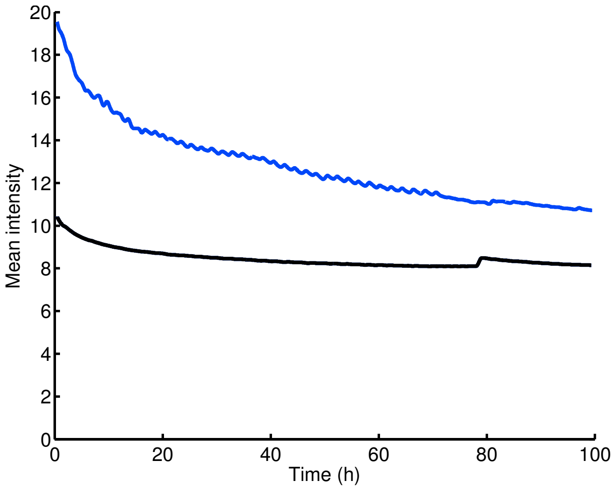


Figure 2.8: Mean intensity over all positions in experiment 3 (blue) and experiment 9 (black). The drop of intensity is almost gone in experiment 9 due to preirradiating the culture medium. Furthermore the flickering of the lamp is eliminated. The small peak at time 80 cannot be explained and needs further investigations.

the bleaching of the background in the foreground as well and we misleadingly inferred that there is bleaching in the cellular signal. It is known from the literature that bleaching has visible effects in living cells leading to incomparable measured intensities [3]. We developed a simple model which simulates the bleaching of the fluorescent proteins in a living cell in order to estimate the parameters for protein production, decay and the bleaching rate in order to normalize the single-cell time courses to their real protein amount. This effect was only visible in the first tracked cells since the *steady-state* of bleaching against new protein production should be reached in a few hours.

The model consists of three equations with active protein A, bleached protein B, a bleaching rate β, a decay rate γ and a protein production α :

$$A \xrightarrow{\beta} B$$
$$A \xrightarrow{\gamma} \emptyset$$
$$\emptyset \xrightarrow{\alpha} A$$

We assume that bleached proteins cannot switch back to active proteins and furthermore that bleached proteins do not emit any light. Furthermore it is not possible to distinguish between protein decay and bleaching rate and only the sum b of them can be observed:

$$b := \gamma + \beta$$

From these equations one can infer an ordinary differential equation (ODE) which describes the change of A over time:

$$\frac{dA}{dt} = \alpha - (\beta + \gamma) \cdot A = \alpha - b \cdot A$$

The indeterminacy of the two parameters of bleaching and protein decay can be solved by two different experiments with different exposure times where the two variables α and γ are constant as the same cells are used. But different experiments with a known factor x between the exposure times lead to:

$$\beta_2 = x \cdot \beta_1 \tag{2.7}$$

The parameter α and β can be determined from the data using a least-square fitting procedure. This results in the same α and two different b. Substituting equation 2.7 we get:

$$\begin{aligned}
b_1 &= \gamma + \beta_1 \\
\Rightarrow \beta_1 &= b_1 - \gamma \qquad (2.8) \\
b_2 &= \gamma + \beta_2 \\
b_2 &\overset{(2.7)}{=} \gamma + x \cdot \beta_1 \\
b_2 &\overset{(2.8)}{=} \gamma + x(b_1 - \gamma) \\
b_2 &= xb_1 - x\gamma + \gamma \\
b_2 - xb_1 &= (1 - x)\gamma \\
\gamma &= \frac{b_2 - xb_1}{1 - x} \qquad (2.9)
\end{aligned}$$

Therefore the biological protein decay which is constant for both experiments can be determined by the two fitted parameter b_1 and b_2 and a known multiplicative factor of exposure times x as the bleaching rate linearly grows with the exposure time. The protein production can be estimated from both experiments independently and as described in later chapter 4.5.1 where the real protein amount can be calculated. With this method the real biological protein production rates could be estimated.

Our method solves indeterminacies of two parameters by a second experiment which does not care about the cell health. Images with higher exposure time or smaller time interval between excitations lead to the real protein production and decay rates. However, this model is not applicable to our data since we do not observe a bleaching effect in the cells after subtracting the background. In future experiments where fluorescence images are taken at a higher time resolution, this model might become more relevant.

Chapter 3

Cell detection and tracking

There are three main approaches of expression analysis in the literature. First, there is a large-scale approach which measures the expression of many cells over time and the characteristics of a whole population are studied. Looking at the expression of a population of cells at a specific timepoint one only measures the mean instead of each individual cell expression. This is done in many common methods like microarray analysis or western blotting. Just monitoring the mean of many individuals can have a certain disadvantage. Imagine cells that are forming two different populations, one with a less intense expression, the other one with a higher expression, but with the same mean as a homogeneous population (Figure 3.1). In such cases one misses important events with this large-scale approach.

Therefore, other methods like FACS sorting measure many individual events separately. It is possible to observe the population wide characteristics as well as the single-cell expression level. The only feature these methods are missing is a time variable since only a snapshot of a certain timepoint can be investigated. It is possible to repeat this experiment at different timepoints and one will again obtain a snapshot of the current population on a single-cell resolution. But in the case of FACS sorting experiments over more timepoints would be very time-consuming and expensive and there is no way keep track of individual cells between these snapshots e.g. in order to estimate rates of cell death or differentiation.

The third main approach of expression analysis can be compared to the results of a FACS sorting but with an essential additional feature. Live cell imaging gives the possibility to follow many single cells along with their expressions over time. This techniques allows two different approaches since the movie can be regarded as a large-scale experiment over many timepoints and, with additional effort of tracking cell, as a small-scale approach for the analysis of single-cell characteristics. After managing this challenge one can determine cell events such as cell-fate decisions, cell-cycle behavior or cell-cell contact and can create according time dependent expression profiles.

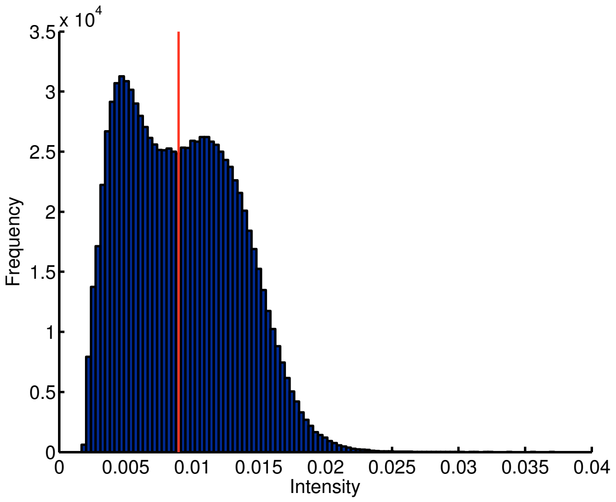


Figure 3.1: Histogram of measured PU.1 intensities of one specific timepoint from our dataset. The red line indicates the mean value measured by common profiling methods like micro-array analysis which would not dissect the two different cell populations.

In this thesis we are performing both, an imaging large-scale approach and a single-cell analysis with a strong focus on the later approach. The manually created tracking of cells annotated by the ISF makes it possible to analyze many individual cells simultaneously and therefore to raise hypotheses of single-cell behavior with biological relevance.

3.1 Detecting cells in fluorescence images

In order to handle the huge amount of data of live cell imaging, it is essential to automatize the cell detection and measurement processes. The accurate and reliable segmentation of living cells is essential for our analysis and therefore it is necessary to look into this subject more carefully. The fundamental process of cell detection in fluorescence images can be reduced to the following problem: A gray-scale image of fluorescent cells (Figure 2.4) consists, as mentioned before, of the background, the cell signal and a certain amount of noise. Ideally the background is close to zero after normalization (Figure 2.6 E). The first step is to find a threshold which separates the background from the real cell signal (Figure 3.2). Values above this threshold are defined as signal whereas every other value is considered to be background.

A variety of algorithms is known to handle the thresholding problem. For example, the well known algorithm of Otsu et al [38], which is based

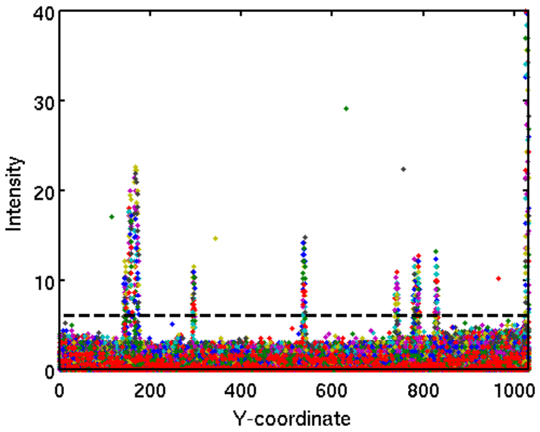


Figure 3.2: A projection of normalized picture onto the y-axis/intensity plane of a normalized image where each dot represents the intensity belonging to a row of the image. The line represents a hypothetic threshold separating background from cell signal. The dots are arbitrarily colored for a better visualization.

on statistical mechanics, investigates the distribution of all occurring grayscale values. The procedure attempts to maximize the variance between the two classes (back-/foreground) and to minimize the variance within both of them, leading to a fixed threshold for the whole image. Connected areas of foreground signal which exceed a size threshold are now defined as cells. Obviously, this method is only applicable to corrected images. A color gradient due to uneven illumination would severely disrupt the thresholding methods since a static threshold, as used in our approach, radically forces every higher value to be a cell. This again points out how important the steps of image preprocessing and normalization are.

Even on corrected images automatic algorithms might fail as they cannot distinguish between two touching objects due to the static threshold. A second step has to be performed: the correct separation of these so called clumped objects can be done by applying a watershedding algorithm afterwards (described later).

The whole detection procedure can be divided into five parts. First, the raw image (Figure 3.3 A) is used to calculate the threshold separating background from foreground. Generally the thresholding performs better on blurred images, so a Gaussian filter is applied on the image in order to smooth the values (Figure 3.3 B). Taking all connected values above the threshold delivers the cell signal and creates a binary image where signal gets the value 1 and the background is represented by 0 (Figure 3.3 C). In the case of the example of Figure 3.3 C, the thresholding performs well by means of differentiating background and real cell signal, but it fails to separate the detection into two distinct cells.

Now the watershedding algorithm takes the binary image (Figure 3.3 C) and calculates a distance matrix which is computed by the euclidean distance of every white pixel to the nearest black pixel. The result is shown in Figure 3.3 D displaying two different maximums of the distance matrix in the center of each cell. Based on these two maxima a dilation procedure begins leading to growing areas which are restricted by a simple rule. A pixel is not added to an area if it is already assigned to a different one. The tricky part of this algorithm is that the dilation is based on the calculated distance matrix using a two dimensional neighborhood approach, so a bigger cell will grow faster than a smaller one leading to the correct intersecting line (Figure 3.3 E).

This separating line is now combined with the binary image of the thresholding algorithm which leads to two separated cells (Figure 3.3 F). Overlaying the detected cell boundaries over the raw image will lead to the final cell detection (Figure 3.3 G). The gray-scale values in between this boundaries are integrated to assign the specific expression intensity to the cell.

In order to obtain a more accurate detection (Figure 3.3 F) for each individual cell a manual inspection and adjustment of the threshold and watershedding parameters is necessary which is discussed in detail in Section 3.3.

3.2 Large-scale approach

In the large-scale approach of this thesis we attempt to detect all cells over a whole movie and to create statistics over the whole cell population. The complexity of this approach lies in the tuning of parameters for a best possible cell segmentation delivering the same quality of cell detection ranging from the beginning to the end of the movie. The image conditions vary over time just by the fact that the cells are proliferating and thus more and more cell signal arises. After four days there are so many cells that it is difficult to distinguish clumps of overlapping cells even by eye (see Figure 3.4). One set of parameters could perform well at the beginning of the movie, where just a few cells have to be separated from the background, but the algorithm will mistake cells for background or vice versa in the end of the movie. For detecting all cells over the whole movie, we are using the freely available CellProfiler toolbox for MATLAB [5]. It allows to create a fully automated processing pipeline which can be applied to every single image. This pipeline provides several computational steps and allows for the integration of our own normalization method. The pipeline we used in the analysis is composed of the following tasks:

1. load a single fluorescence image

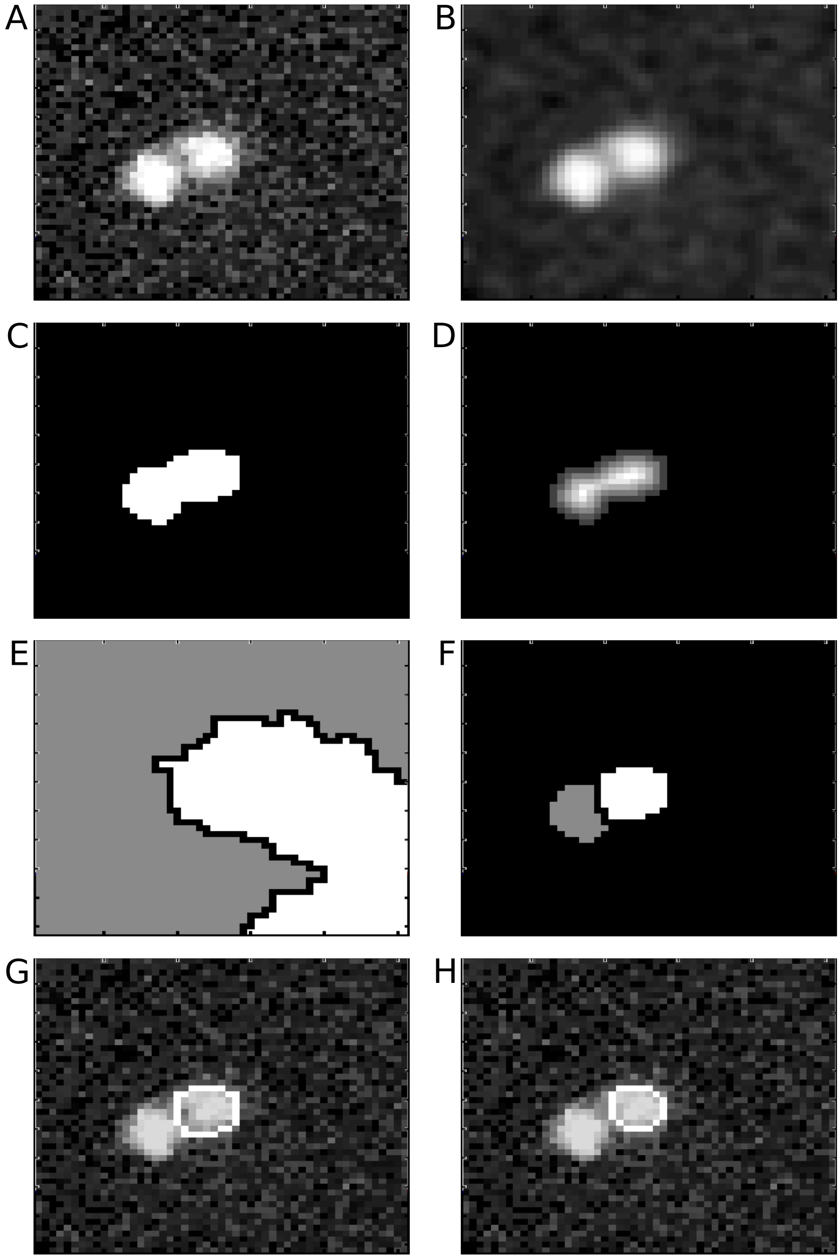


Figure 3.3: General workflow of cell detection algorithms on gray scale fluorescence images. The raw image (A) is smoothed by a Gaussian filter for better performance (B). An automatically calculated threshold is applied and every signal above this threshold is referred to cell signal and will get value 1, whereas every other pixel gets 0, leading to a binary image (C). Based on this image a distance matrix is calculated assigning every foreground pixel the distance to the nearest black pixel (D). The matrix is used by watersheddding algorithm to dilate the two maximums based on their two dimensional-neighborhood (E). Overlaying the binary image with the two calculated separate areas (white and gray) results in two separate cell nuclei (F). Projecting the detected cell boundaries onto the raw image shows the detected cell (G). Integrating over the values in between gives the assigned cell intensity. After adjusting the threshold a more accurate cell detection can be obtained (H).

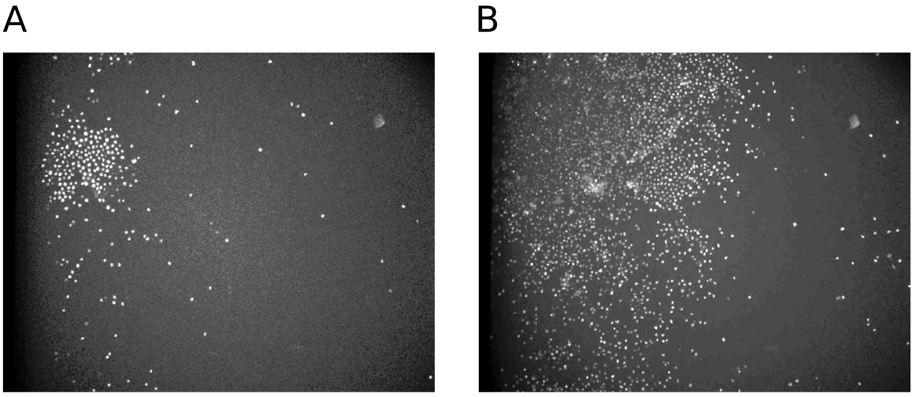


Figure 3.4: Fluorescence image of an early (A, day 1) and late timepoint (B, day 4). The median of picture (B) is still within the background assuring that the normalization described in 2.3 still holds.

2. apply illumination and background correction method
3. detect nuclei by thresholding and watershedding, afterwards filter by size
4. measure intensity
5. save raw image with detected outlines
6. save measurement

The CellProfiler also allows to overlay the detected cell boundaries over the original images and save the results in separate image files. These images are used afterwards to create movies with a higher contrast in order to gain a better insight of what the program detects and whether it makes mistakes (Figure 3.5). As we want to detect as many false positives as possible, a set of parameters is now chosen by evaluating the outline-movies via manual inspection.

The analysis of a whole movie by CellProfiler takes around one day on a quadcore machine creating a file with all detected nuclei (about 700MB) and about 3.5GB of outlined images. After converting the detection file into a MATLAB readable format every line represents a detection consisting of the experiment position, the timepoint, the position specific coordinates X and Y and two quantifications: the measured intensity and the measured size.

By manually inspecting the outline-movies at every position, we detected some artifacts and some spurious detections on the edges of the experimental plate (Figure 3.5). Therefore we implemented a simple position-specific filtering which allows us to exclude such artifacts.

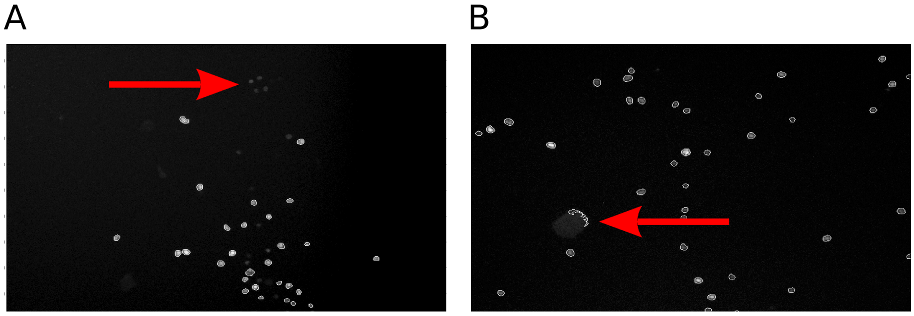


Figure 3.5: Two outline images of the automatic CellProfiler detection. (A) A high threshold did not detect all cells. (B) A low threshold detects artifacts. The detection of CellProfiler can be manually inspected by contrasted movies, which can help to adjust the parameters or to filter the results afterwards.

For every single image of the experiment there exists an XML file which records the exact position of the robot as well as the exact time. The final step of the large-scale data preprocessing is a mapping algorithm which assigns the exact real time to every single image based on the timepoint and the position.

3.3 Aided Manual Tracking

The tracking trees of ISF deliver position-specific coordinates but they will not deliver detections or measurements of the intensity of the nuclei. We use these coordinates and attempt to identify the correct cell boundaries and measure the intensity of each cell. Since the automatic detection regularly fails one has to look manually into the detected data. The recorded cell movie still relies on a live biological experiment where some artifacts like contamination always appear. We developed a MATLAB toolbox called Aided Manual Tracking (AMT) which implements several thresholding methods which have been evaluated on manually set thresholds for several trees (results not shown). The purpose of this tool is to refine the automatic detection of manual created trackings for several artifacts or misdetections in order to get accurate expression data. The continuous development of that toolbox and the implementation of more and more methods was a major task of this thesis.

3.3.1 Single-cell data preprocessing

The program is basically based on two layers. First an automatic detection algorithm is applied which measures the intensity and size of each cell over

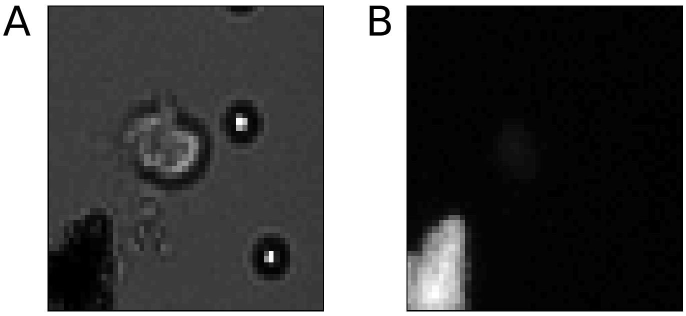


Figure 3.6: Brightfield (A) and contrasted fluorescence image (B) of intensely glowing artifact. Small dots represent the focusing beads. The intensity of the contaminants is so high that the cell signal virtually disappears.

time. It takes as input an ISF tracking tree and uses the manually set coordinates to extract a small window of the normalized cell image around theses coordinates. Then it automatically detects cells in the given window and takes the nearest cell to the center as the cell to be measures. It is obvious that this cannot be performed manually as there are usually more than 300 different detections for each tree. For the AMT it is as important, as for the CellProfiler, that the thresholding parameters are set correctly. It is possible to test just only a few of the trees with a parameter set before computing the whole set in order to tune the parameters to perform for the current movie conditions. It is also possible to recalculate cell boundaries for a whole cell life or tree with modified parameters.

After adjusting the parameters, the automatic detection can be applied to the whole set of trees. The resulting data now serves as the base for the second part of the program which from now on provides a graphical user interface.

3.3.2 Identifying and correcting detection errors

As already said there are many cases where automatic detections fail. An example is given in Figure 3.6 where some undefinable artifacts disturb the detection. Some of these problems are uncovered very late in the processing pipeline because at the manual tracking step on the brightfield images the contamination does not look that fatal (Figure 3.6 A).

It frequently happens that in the tracking data the tracking position is not accurately over the cell. It is considerably difficult to precisely keep track of a whole tree and every individual cell. Our tool is designed to take the nearest cell center as the cell to measure. However, in the late movie it

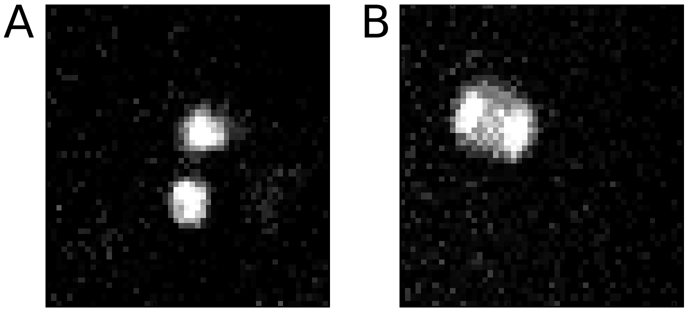


Figure 3.7: (A) Two cells in close proximity at a specific trackpoint. The movie of brightfield images can help to select the right one to measure. (B) A cell during devision process.

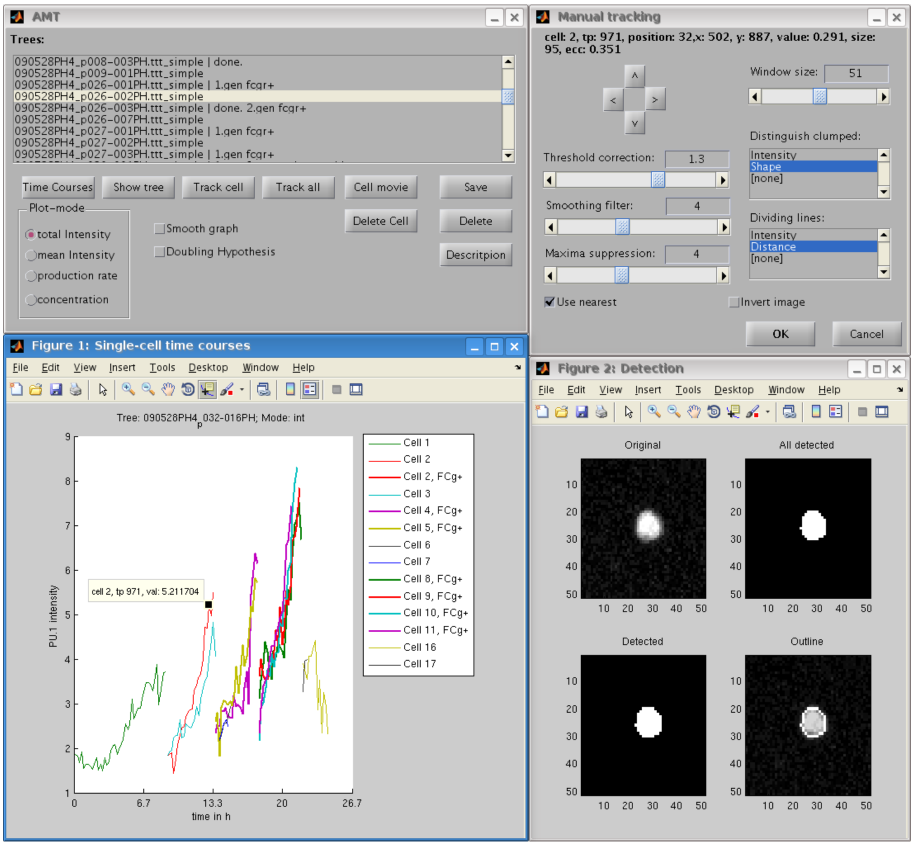
happens that taking the nearest cell is not the appropriate method and the algorithm chooses the wrong cell.

A problem due to the low time resolution of the fluorescence images is that an image is taken as the cell is still dividing (Figure 3.7 B). There are already two cell nuclei but the cell has not finished the mitosis. One has to decide whether this is allocated to the mothercell or whether both nuclei are assigned to the daughters.

All problems listed before can easily be detected and corrected via the AMT tool (Figure 3.8). The AMT shows the single-cell time courses to the user (Figure 3.8 lower left) and serious problems attract immediate attraction (Figure 3.10 left). The program allows to select a specific timepoint and cell. A dialog will appear displaying the extracted window of the cell image and all detected nuclei in this picture. One can adjust the parameters of the thresholding and watershedding algorithm or choose the correct cell (Figure 3.8 two boxes on the right). Instead of just selecting one specific cell and timepoint it is also easily possible to look through the detections of a cell over its whole lifetime.

The program also allows to investigate the raw movie zoomed in and centered around the cell of interest. The movie can be played in both direction and examined image by image. The playback function of the movie allows a deeper insight into the cell movement and makes it easy e.g. to choose the correct cell if two or more cells are in close proximity (Figure 3.7). All three channels can be chosen: the brightfield movie with a higher time-resolution, the detection or the measurement channel of the fluorescence signal images.

In the case of PU.1 movie detection and measurement rely on the same channel. The tool is also applicable to other projects and other data types than PU.1 movies. In another collaboration with the ISF, dealing with fluorescence movies of differentiating embryonic stem (ES) cells, we also use the AMT to clean up the expression data. In this project two different



Trees:

```
090528PH4_p008-003PH.ttt_simple | done.
090528PH4_p009-001PH.ttt_simple
090528PH4_p026-001PH.ttt_simple | 1.gen fcgr+
090528PH4_p026-002PH.ttt_simple
090528PH4_p026-003PH.ttt_simple | done. 2.gen fcgr+
090528PH4_p026-007PH.ttt_simple
090528PH4_p027-001PH.ttt_simple | 1.gen fcgr+
090528PH4_p027-002PH.ttt_simple
090528PH4_p027-003PH.ttt_simple | 1.gen fcgr+
```

Plot-mode: ● total Intensity ○ mean Intensity ○ production rate ○ concentration

□ Smooth graph

□ Doubling Hypothesis

cell: 2, tp: 971, position: 32,x: 502, y: 887, value: 0.291, size: 95, ecc: 0.351

Window size: 51

Threshold correction: 1.3

Smoothing filter: 4

Maxima suppression: 4

Distinguish clumped: Intensity / **Shape** / [none]

Dividing lines: Intensity / **Distance** / [none]

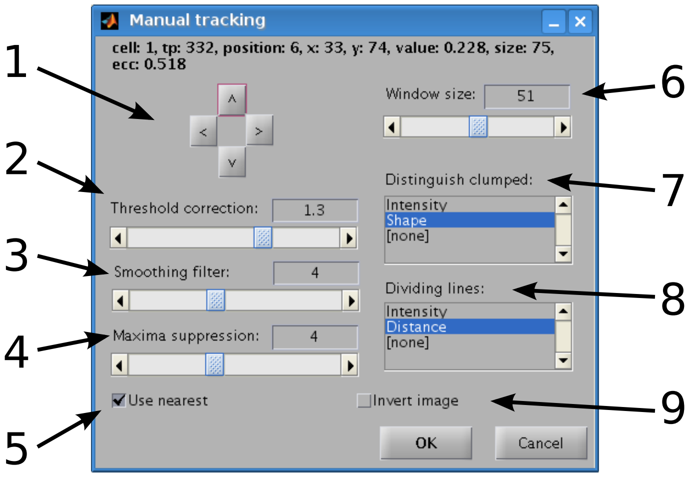
☑ Use nearest

□ Invert image

Tree: 090528PH4_p032-016PH; Mode: int

cell 2, tp 971, val: 5.211704

Figure 3.8: The Aided Manual Tracking (AMT) tool interface. The user can choose a tree to be analyzed in the tree selection window (upper left). The corresponding time courses are shown below, where the user can interactively choose a specific cell and timepoint to be inspected in more detail. The dialog on the lower right shows the detected cell with its surrounding area. All possible parameters at the upper right can be adjusted for a better detection (compare Figure 3.9). The detection with new parameters is updated on-the-fly and the result can be seen in the cell detection dialog.



Manual tracking

```
cell: 1, tp: 332, position: 6, x: 33, y: 74, value: 0.228, size: 75,
ecc: 0.518
```

Window size: 51

Threshold correction: 1.3

Smoothing filter: 4

Maxima suppression: 4

Distinguish clumped: Intensity / Shape (selected) / [none]

Dividing lines: Intensity / Distance (selected) / [none]

☑ Use nearest

☐ Invert image

Figure 3.9: All parameters which can be tuned to correct for errors. The first line shows the information of the detected cell depending on the parameters. The cell number of the cell tree, the timepoint, the position of the experiment as well as the position-specific coordinates x and y, the calculated intensity, the nucleus size and eccentricity. The following parameters can be adjusted:

1. **Position:** Move center of the image in order to select the correct cell

2. **Threshold correction:** A correction factor to manipulate the calculated threshold

3. **Smoothing filter:** The size of the Gaussian filter for blurring

4. **Maxima suppression:** Suppresses maxima of objects and affects whether objects close to each other are considered a single object or multiple objects. Should be smaller if two clumped object are still connected

5. **Use nearest:** Take the nearest cell if no cell is in the center of the subimage

6. **Window size:** Change size of the window around a cell

7. **Distinguish clumped:** Choose different methods to distinguish two clumped objects either by their *intensity* if a dim dividing line between both is visible or by *shape* where a distance matrix is calculated for the watershedding algorithm

8. **Dividing lines:** Draw the dividing line between two clumped object either based on their *intensity* or on their *distance* matrix

9. **Invert image:** Detecting e.g. on phase contrast images sometimes works better on inverted images

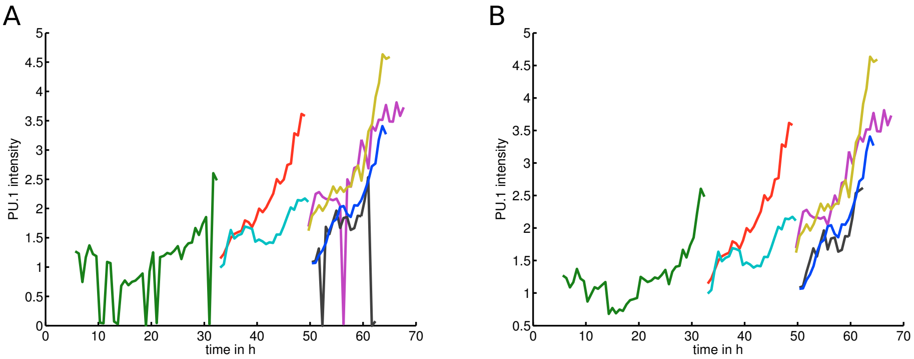


Figure 3.10: The time courses showing measured PU.1 intensity over time after automatic detection (A) and after manual correction of detection (B). Sharp peaks indicate problems of the automatic detection.

proteins are tagged with fluorescent markers. The ES cells tend to build up colonies which makes it impossible to track the cells by eye on brightfield images (Figure 3.11). Therefore the nuclear membrane protein *nucmem* is labeled which is always expressed in the nuclear membrane and which is used to detect the cell nucleus. In a different channel the actually relevant transcription factor *nanog* is quantified depending on the detected nucleus of the nucmem.

All correction steps implemented in the AMT are essential to clean up the data as much as possible and the program allows all this in a very simple and intuitive way. Once the tool prepared the data, it also provides basic functions to run analysis methods directly on the chosen data (which will be described in detail in the next Chapter). In the future the program will be extended for the work with all movie types at the ISF and, in addition, will be transferred to the computer system at the institute.

3.4 Combining experiments

For the next chapter it is of a great interest to combine the tracking data of different movies. Since the biological conditions throughout the experiments should be constant there are only few variables left we have to assure. The varying background signal, different flickering and unequal bleaching effects are eliminated due to our normalization method. Although the exposure time is not changed over all cell movies, one important variable left is the light source intensity which might change with an increasing life time of the light source. This variable leads to a multiplicative factor which affects all measured intensities. The boxplot of all measured intensities of the two

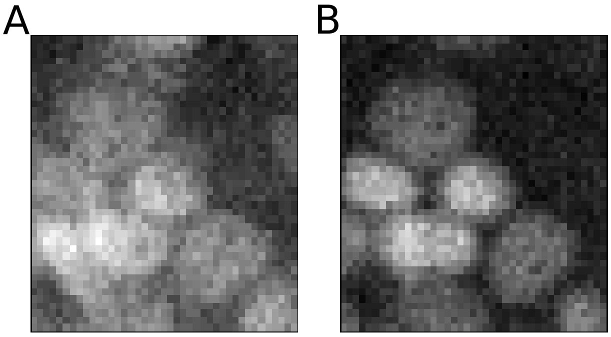


Figure 3.11: Two fluorescent images of the same timepoint but different wavelengths. The *nanog* signal is used to quantify the intensity but since the detection or even the cell tracking is not possible on this channel due to the colonies (A). Another fluorescently tagged membrane protein *nucmem* is used to detect the cell nucleus which is then projected on the *nanog* signal (B) to get the intensity.

experiments 3 and 8 show a multiplicative shift which causes the newer movie to appear darker (Figure 3.12 A).

The first attempt is to estimate the difference of intensities via examining the fluorescine line which is measured for every experiment. Almost all attributes of the experiments are equal, but the newer movie was preirradiated in order to eliminate the autofluorescence of the background (caused by the phenol red in the medium). This leads to extreme differences in the first images of the experiments and prevents an appropriate comparison of the fluorescent images.

Therefore any raw image will have incomparable mean intensities and only the following method seems convenient. Looking at the measured intensities of all cells of each experiment will identify the multiplicative factor. The boxplot Figure 3.12 A shows a shift of 1.56 of their medians. After multiplying the newer movie with the factor an almost overlapping histogram of the relative intensity distribution confirms this procedure and makes the measured data comparable (Figure 3.12 B). Experiment 8 (green) has more zero outliers than experiment 3 (blue) since the newer data has not yet been sufficiently corrected by manual inspection and only the important trees described in the next chapter were revised.

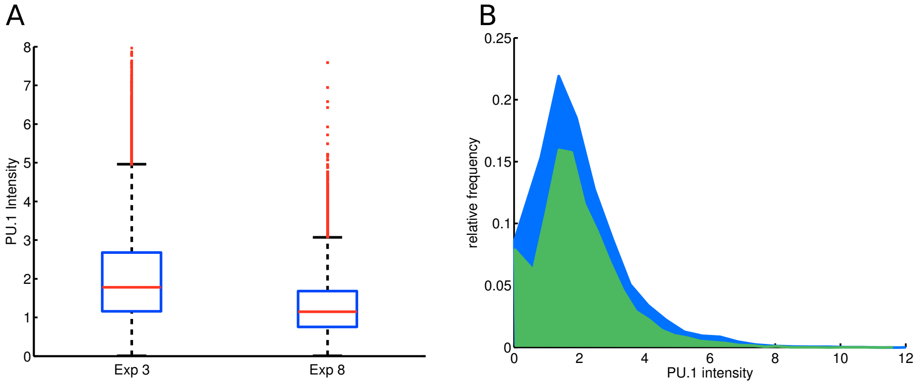


Figure 3.12: (A): Boxplots of all measured PU.1 intensities of experiment 3 and 8. There are still some outliers in the data since only several interesting trees have been inspected manually. The shift of the median of both distributions is 1.56. (B): Multiplying experiment 8 with this shift leads to almost the same distribution of intensity values allowing to combine these two experiments. Experiment 8 (green) has a greater amount of zero outliers than experiment 3 (blue) due to less manual inspection.

Chapter 4

PU.1 expression analysis

The data preprocessing and data consolidation outlined in the previous chapters leads to cleaned fluorescence intensity levels for the YFP tagged PU.1 in MPPs. This allows for the large-scale analysis of intensity distributions and enables us to study molecular properties and estimate noise of single cells over a whole genealogy. We present different expression normalizations, correlations of cell lifetimes as well as cell-cycle expression profiles. Furthermore, due to single-cell tracking, the experimental data includes annotations of the state of the FCγ surface marker indicating a GMP lineage commitment, if active. This allows us to investigate specific PU.1 expression profiles of GMP committing cells and their progenitor cells. At last we perform some further tree analyses and an estimation of PU.1 protein.

4.1 Estimation of detection quality and noise strength

A quantification of the imaging and detection quality can be given by comparing the measured results against the quantification result from well-established methods. This is possible with experiment 11 (compare Table 2.1) where hematopoietic cells were sorted by FACS and imaged after sorting. Three different population-specific images of separate populations containing either MPPs, MEPs or GMPs were obtained. The intensities were measured during the FACS sorting process and then compared to our detection and measurement methods. The histograms of Figure 4.1 A and B show the distributions of population-specific intensities from FACS and from our imaging and detection technique. The sharp cuts of the histograms in Figure 4.1 A are due to assigning upper and lower boundaries for the PU.1 intensity during the sorting process. Nevertheless, both methods can be compared on relative scales. The difference is 3-fold from MEP to MPP and 4-fold from MPP to GMP on linear scale. The barplot in Figure 4.2 A shows the mean intensity of each population with its standard deviations

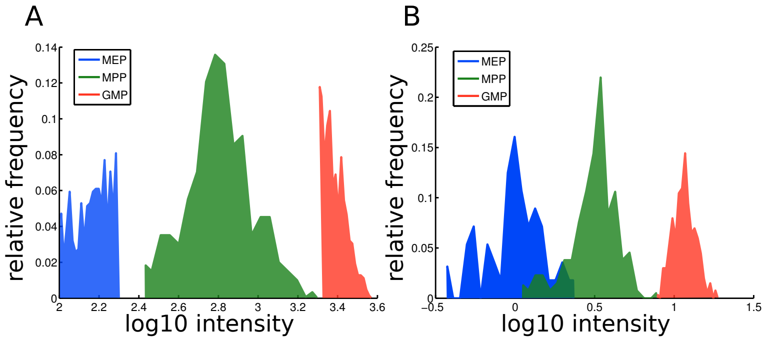


Figure 4.1: (A) FACS intensities of all three cell populations. (B) Intensities of all three cell populations measured by our normalization and detection method. Both show an fold-change of 3 from MEP to MPP and a fold-change of 4 from MPP to GMP in linear scale.

with two different scalings on the y-axis. The measured FACS intensities belong to the left y-axis whereas the imaging values belong to the right y-axis. This indicates highly comparable results between FACS and the normalized imaging method which is a confirmation of the cell detection and quantification. Furthermore a comparison of the measurements without the correction method described in Section 2.3 against the FACS intensities highlights the importance of normalizing and gives a proof that our correction method is correct (Figure 4.2).

The noise introduced by the thresholding and segmentation process described in Section 3.1 can be estimated by the following approach: Ordinarily a cell is manually corrected and the threshold is adjusted for an optimal detection within ±20%. We now vary the threshold from 80 to 120% of its supposedly correct value. With the altered thresholds the detection algorithm is reapplied on the cell. The resulting errorbars in Figure 4.3 show that the introduced noise is relatively small ($\approx 3.1\%$) derived by the coefficient of variation. This is calculated by dividing the mean standard deviation by the mean. The remaining fluctuations of single-cell time courses can be attributed to biological noise probably emerging from transcription and translation. It is a general understanding that these biological events go through bursts and do not produce protein with a constant rate [42]. An experiment where fluorescent images are taken with a higher time resolution could confirm this hypothesis and is already planned.

The technical noise due to inaccuracies of the imaging technique can be estimated by looking at experiment 10 where only the background is imaged. We assume that the technical noise in a pixel of a measured cell is the same as in the background pixels or in fluorescine images. The noise can easily be measured in this experiment since the signal in these images

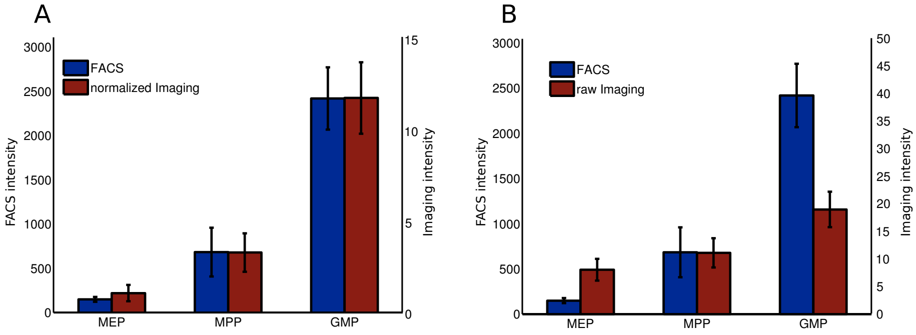


Figure 4.2: (A) Comparison of FACS intensities against normalized imaging intensities showing almost the same ratios. (B) The bars of the raw imaging data indicate that our normalization method is necessary. The FACS values belong to the left y-axis and the imaging values belong to the right y-axis.

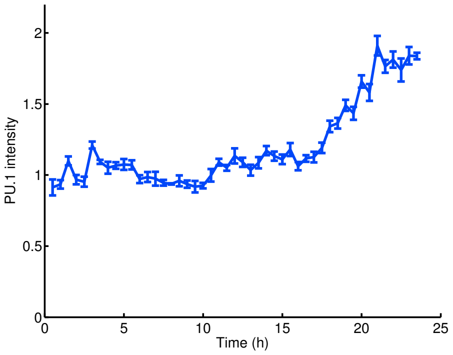


Figure 4.3: Time course of a single-cell detected with different thresholds ranging from 80% to 120% of the manually set parameter. The noise introduced by threshold is estimated to be about 3.1% derived by the coefficient of variation.

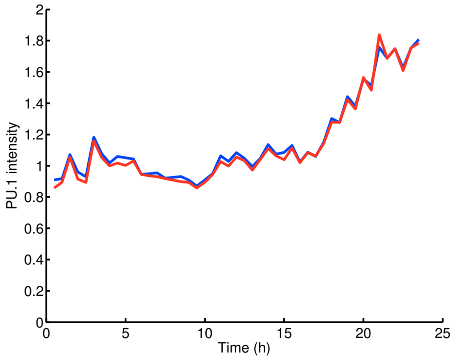


Figure 4.4: Comparison of two different normalization methods. The blue line refers to the method of Sigal et al [49] whereas the red line is calculated using our method described in Section 2.3 resulting in almost identical single-cell time course.

is constant. Again calculating the coefficient of variation by dividing the mean background standard deviation by the mean cellular signal intensity per pixel leads to a technical noise of 1% (result not shown).

A quantification of our normalization method can be given by comparing it with the well-established method of Sigal et al [49], which was already described in Section 2.3.1. Single-cell time courses calculated with both methods show that both perform very similar (see Figure 4.4). They only differ in ≈ 2% of total measured intensity and the general characteristics of the line is retained.

4.2 Large-scale analysis

The cleaned up large-scale data contain the measured intensities and sizes of all cells at every timepoint over a whole experiment, leading to about two million individual cell detections in experiment 8 (compare 2.1). Histograms showing the distributions of intensities for each day of the experiment are given in Figure 4.5 A. At the beginning the experiment contains only MPP population whose intensity distribution can be seen in the histogram of the first day. Looking at the second day shows no change in the distribution but a lot of more cells appear due to proliferating MPPs. Starting with day three the distribution begins to shift its mean towards higher intensities which can probably be accounted to a shift towards the GMP lineage. Due to the experimental conditions we should observe a dominating population of GMPs at later timepoints (knowledge from earlier experiments). The further development of the populations shows a continuation of this trend in the fourth day of the experiment. However, the distributions of the last two days differ from this expectation since the mean decreases again and a large population of low intensity cells comes up and proliferates.

Interestingly, no intensities higher than 7 are observed. As already discussed in Section 4.1 our detection method detects GMPs with a mean intensity of 11. To exclude any technical problems, we re-measured the cells at different timepoints and positions by manual inspection which led to the same results as shown in the histograms (data not shown). Looking at the surface marker FCγ gives convincing evidence that GMP cells do appear in our experiments. Therefore we can only assume that there is some bleaching effect in the movie which cannot be seen in single-cells and is only visible at larger time scales. Since the experiment discussed in Section 4.1 consists of only one timepoint, further investigations are needed to clarify if a bleaching effect exists in the cells.

Assuming that there is some bleaching, the shift of intensities from day two to three can be explained by cells committing to GMP lineage. In day four one could assume two different population, possibly representing MEP and GMP, respectively. Attempts of fitting two separate Gaussian distributions lead to no clear separation of the distributions (data not shown). The second shift of intensities can be explained by the lineage commitment and subsequent PU.1 repression in the further differentiation process. To validate this assumption, again further investigations are needed.

In Figure 4.5 B, we show the time development of the mean and the standard deviation of the fluorescence intensities and the detected number of cells. Since the experiment examined here had some technical problems in the first twelve hours, the mean as well as the standard deviation are meaningless during this phase. After this timepoint the mean intensity slightly increases for about 100 hours and only decreases in the middle of day five. The standard deviation obviously decreases after 100 hours due to the quickly proliferating cells leading to a large population of presumably homogeneous cell types (this observation cannot be explained so far). The number of cells constantly increases over four days until a strong onset at ~95h, reaching a final level of about 50.000 cells at the end of the movie. The detected cell number after 120 hours of movie time will not be representative since the detection quality decreases due to the large amount of cells on the movie (compare Figure 3.4).

4.3 Single-cell time course analysis

Instead of looking at populations we now investigate the measured fluorescence intensities of single cells. The additional knowledge of tracking leads to expression data time courses which allow deeper insight into precise differentiation mechanisms.

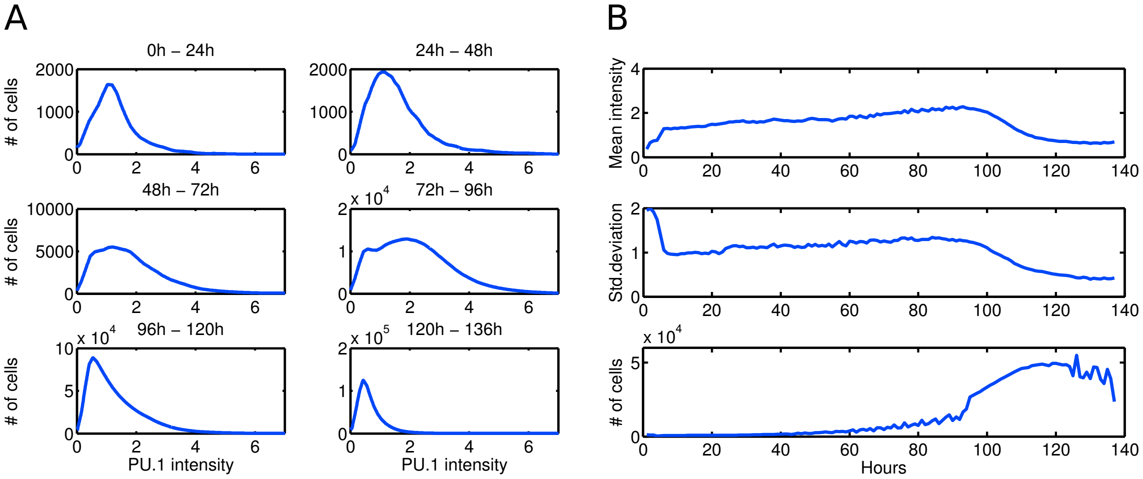


Figure 4.5: Analysis of the large-scale expression data where about 2 million individual cells of experiment 8 (compare Table 2.1) have been detected. (A) Histograms of measured PU.1 intensities calculated for each day. In the first two days a population of proliferating MPPs can be observed. Day three and four highlight a increasing shift of the measured intensities, possibly due to the differentiation into the GMP lineage. A second shift can be seen in day five and six where the intensities decrease again possibly due to PU.1 not playing a role in the further differentiation process. (B) The development of the mean and standard deviation show some arbitrary behavior due to technical problems in the experiment in the first twelve hours. The mean and the standard deviation slightly increase over 100 hours and drop rapidly afterwards due to the large increase of cell numbers. The cell numbers show a constant increase over 90 hours. After that the cells rapidly proliferate to a population of about 50.000 cells. Measurements after 120 hours are inaccurate due to decreasing detection quality.

4.3.1 Absolute intensities

The measured fluorescence intensity against time of single-cells is given in Figure 4.6 A. Every line represents the fluorescence intensity of a single cell over its lifetime. As expected, the intensity grows with lifetime, after a cell division two new lines represent each daughter cell. Since MPPs predominantly divide symmetrically with respect to PU.1 (as discussed later in Section 4.5.1), every daughter starts with about one half of the mother's intensity.

4.3.2 Net production rate

The *net production rate* r is defined as the first derivative of the absolute intensities. (Figure 4.6 B, [45]) We calculate it as:

$$r(t_i) := \frac{f(t_{i+1}) - f(t_i)}{t_{i+1} - t_i}$$

with f representing the absolute PU.1 intensities. This numerical derivative represents the change of protein over time and includes both protein production and decay. In this view differences of intensity slopes can be made more visible. Moreover, r is independent of the absolute fluorescence intensity, allowing to compare cells of different trees or different absolute levels. After cell division the production rate can change because there is only half of the DNA left leading to less transcription and translation. An increase should be seen after the DNA is replicated in S phase.

4.3.3 Normalization based on a doubling hypothesis

In order to make cell time courses of different generations or trees more comparable, we introduce a normalization based on a *doubling hypothesis*: A proliferating cell doubles up its volume and protein amount during each cell-cycle. The intensity is recalculated in the following way: (1) we take the median of the first five timepoints giving a robust estimate of the starting intensity at cell birth. (2) A line over the whole cell lifetime from the initial value to the doubled starting intensity is subtracted from the single-cell time course. This should normalize ordinarily proliferating cells to a straight line and highlights cells deviating from this line, possibly indicating a lineage decision. Figure 4.6 C shows time courses normalized by this method. The first cell does not satisfy the straight line but reaches zero again at the end of its life time whereas cells 2 and 3 clearly differ from a horizontal line. Time courses with a positive ending line indicate highly expressed PU.1 levels where commitment to the GMP lineage would be expected. Checking the cell fate of cell 2 and 3 shows that both commit to GMP lineage. Cells which commit to the MEP lineage should show negative values at the end of their life time. The later cells do not fulfill the assumption as they are not

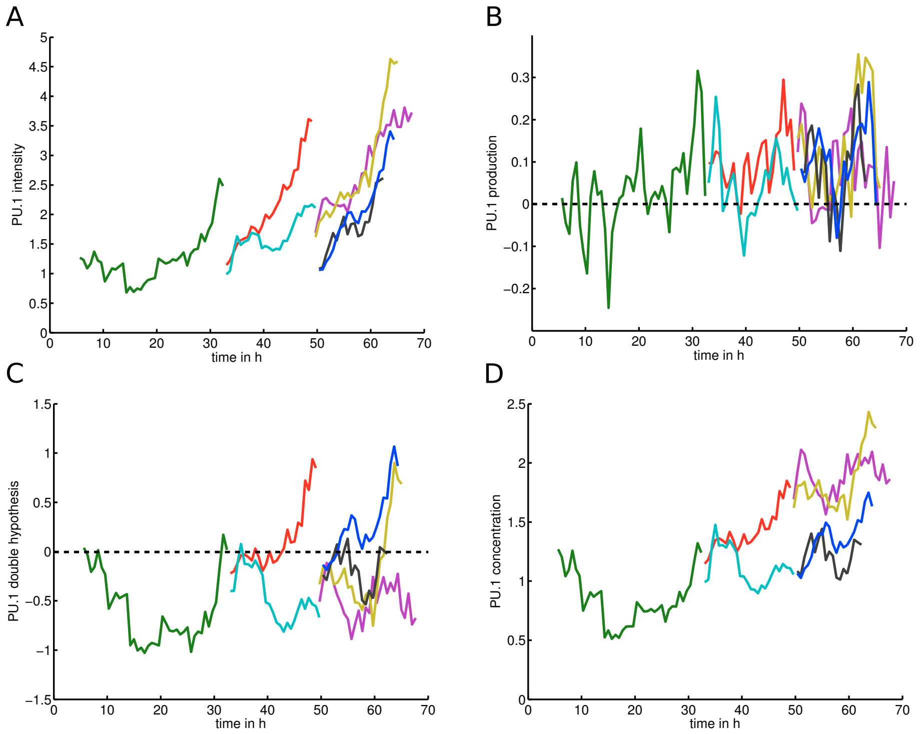
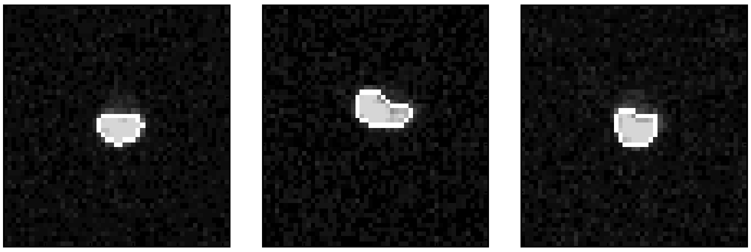


Figure 4.6: Different representations of single-cell time courses of one complete genealogy where each line describes the time course of a cell: (A) The absolute measured intensity over time shows increasing intensities as expected. After a division, two new lines appear both starting with approximately one half of the amount of the last timepoint of their mothercell. (B) The *net production rate* calculated by the first derivative of absolute values where some cell cycle dependent pattern can be emphasized against plot (A). At the end of each cell-cycle a positive derivative indicates cell growth for the pending cell division. (C) The time courses modified by the doubling hypothesis where a line from the first intensities to the doubled amount over the lifetime of each cell time course is subtracted. A cell which produces constantly protein over its lifetime will therefore represented by a straight line. Every cell deviating from this hypothesis will possibly have made a lineage decision. Due to the cell-cycle kinetics the cells do not build a straight line but become zero at the end of their lifetime (compare first cell). (D) A representation of the protein concentration derived by an approximation of the cell volume described in Section 4.3.5.



Time (min)	1374	1404	1434
Intensity	3.23	3.35	3.52
Size (pixel)	58	77	67

Figure 4.7: Detected cell nuclei in a continuous time series with constantly changing shape but increasing intensity. Fluorescent images are taken at a time interval of 30 minutes.

tracked over their whole lifetime since the lineage decision is already made by their mothercells.

4.3.4 Mean pixel intensity

The output of the detection algorithm also delivers the measured size. So a *mean pixel intensity* could be computed by dividing every measured intensity by the measured size in order to estimate the concentration in a cell. However this method is not very meaningful as the cell and the nucleus are three dimensional objects and the measured two dimensional area does not represent the volume of the cell. Figure 4.7 shows three images of a continuous time series in which one can clearly see the cell nucleus. The measured intensity constantly increases over time but the measured size is strongly varying. Therefore it is obvious that the measured cell area cannot be used to create a concentration time course.

Another approach to outline the varying of cell sizes is given in Figure 4.8 A showing the correlation between cell size and cell intensity. The color represents the cell lifetime spanning from black (cell birth) to red (a cell about to divide). One can see that there is a correlation between cell-cycle and size, as well as between cell-cycle and PU.1 intensity. The black dots are mainly on the left bottom whereas the red ones are spread around the figure. In this analysis we observe a pairwise correlation between size and intensity of only 0.19, which might be due to looking at very different cells.

The relationship between size and intensity can be studied in more detail by plotting a trajectory of a single cell over time (Figure 4.8 B). Starting at the bottom left corner the cell moves around over time and ends up in the right upper corner. The trend of this curve is as expected and the size as well as the absolute intensity are growing with the cell lifetime. At first,

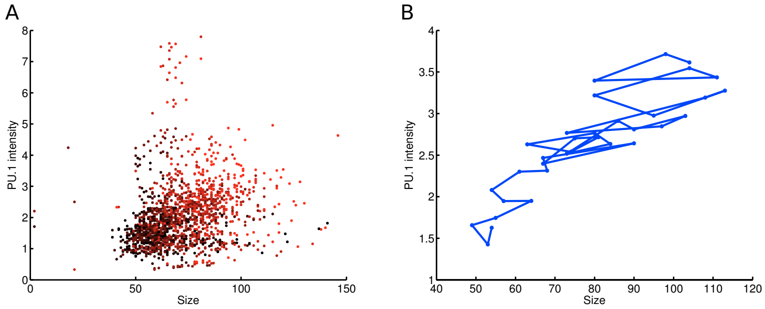


Figure 4.8: (A) Measured size against measured intensity of cells, where the color code represents the lifetime spanning from black (cellbirth) to red (a cell about to divide). The small pairwise correlation between size and intensity of 0.19 highlights the fluctuating measured cell size. (B) Single-cell trajectory showing developing size against intensity over time starting in the left bottom. The intensity rapidly increases in the first eight timepoints. After the cell reaches a stable level of PU.1 (≈2.5) the strong size fluctuations can be observed. At the end of the cell-cycle the cell size varies between 80 and 115 pixel.

the intensity rapidly increases and after eight timepoints it remains at the constant level.

4.3.5 Protein concentration

When analyzing the regulation kinetics of a transcription factor, a substantial question is whether to investigate the absolute or the relative concentration of the protein. As the amount of DNA (or transcription factor binding sites) is supposedly constant during the cell cycle, we assume the relative concentration to be the relevant quantity. We introduce a method to estimate the protein concentration $C(t)$ of a cell defined as the fraction of protein amount $P(t)$ and its volume $V(t)$:

$$C(t) = \frac{P(t)}{V(t)} \tag{4.1}$$

where t describes the relative cell life time ranging from 0 to 1. Therefore, we need an estimation of the volume. As already discussed the previous methods will not satisfyingly estimate the real cell volume.

We make the assumption that the cell can be described by a sphere and its two dimensional projection can be described by a circle. A first indication of this feature can be seen in Figure 4.9 showing all measured cell sizes (about 9000) centered and summed up resulting in a circle.

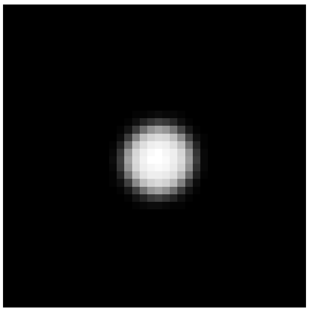


Figure 4.9: All detected cell areas (about 9000) centered and summed up resulting in a circle. This gives a first indication that the cell volume can be estimated by a sphere.

Relations between volume and area of a sphere

The cell volume V with radius r can be calculated by $V = \frac{4}{3}\pi r^3$. The cell area A represented as a circle is derived by $A = \pi r^2$. A volume V_0 represents the initial cell volume, A_0 the corresponding initial area, both depending on the same initial radius r_0. Another volume V_x is the initial volume multiplied by a factor x which has a corresponding radius r_x and an unknown area A_x. We derive the relation of the area to the volume scaling factor x by:

$$V_x = x \cdot V_0 = x \cdot \frac{4}{3}\pi r_0^3 = \frac{4}{3}\pi \cdot (\underbrace{\sqrt[3]{x} \cdot r_0}_{r_x})^3 \tag{4.2}$$

$$A_x = r_x^2\pi \overset{(4.2)}{=} \left(\sqrt[3]{x} \cdot r_0\right)^2 \pi = \sqrt[3]{x^2} \cdot r_0^2\pi = \sqrt[3]{x^2} \cdot A_0 \tag{4.3}$$

From this follows that multiplying a given volume with a factor x will lead to an increase of $\sqrt[3]{x^2}$ of the area.

As the next step, we take all quality-filtered cells of all movies and normalize their lifetime to the same length 1. We excluded the first and the last cells in each tree since their effective lifetime is unknown. Averaging over the measured size of every cell leads to an interesting result. The cell area is growing linearly as Figure 4.10 A shows.

Furthermore, we take the assumption that a cell approximately doubles its volume over its live time ($x = 2$). Normalizing the measured cell areas shows that they exactly grow to a total of $\sqrt[3]{2^2} \approx 1.59$ times their starting size (Figure 4.10 B). Looking at some individual cell area time courses and their linear fit shows that the residuals are small. This indicates that the fit over the mean cell sizes is representative (data not shown). This is a further strong indication that the cell can be approximated by a sphere.

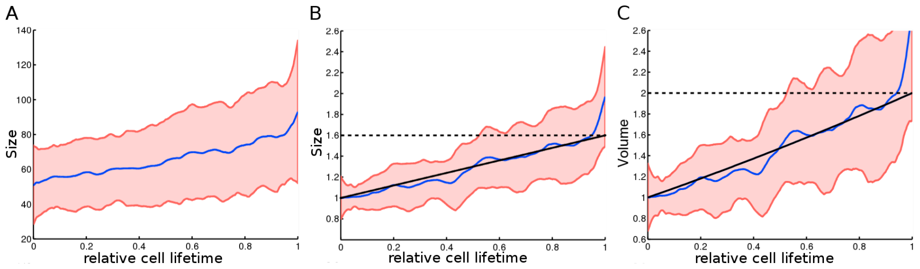


Figure 4.10: Average cell size over normalized cell lifetime. (A) The mean of measured cell sizes of all cells with its standard deviation. (B) Normalized size increases from 1 to 1.6 linearly. (C) The estimated normalized cell volume ranging from 1 to 2. The large deviation at the end of each line is due to an effect of the cell division (compare Figure 3.7 B)

Time development of volume

Including the observation of the linearly increasing area along with the relation of volume and area of a sphere ((4.2) and (4.3)) allows us to calculate the development of the volume in between $t = 0$ and $t = 1$:

$$V(t) = \sqrt{(m \cdot t + n)^3} \tag{4.4}$$

For simplification, we set the initial cell volume to 1 ($V(0) = 1$) and we know that the cell doubles its volume over the cell-cycle ($V(1) = 2$). Therefore we can determine the two variables m and n:

$$V(t) = \sqrt{((\sqrt[3]{2^2} - 1) \cdot t + 1)^3} \tag{4.5}$$

Thus the hypothetical volume of a cell grows between linear and quadratic depending on t resulting in a bent line from 1 to 2 which is shown in Figure 4.10 C. This method gives us the opportunity to normalize every single-cell time course by equations (4.1) and (4.5). $P(t)$ can be estimated by the measured PU.1 intensity and the result can be regarded as the PU.1 concentration in the cell. It is also possible to calculate the estimated volume depending on the real values of the measured area of the first timepoints. But it is sufficient to assume that all cells in a tree have the same volume at their first cell timepoints since MPPs and its offspring should have approximately the same volume. However, calculating the real volume from the measured area would only be a multiplicative factor. As already shown in Section 4.3.4 the measured size is inexact and would introduce unnecessary noise.

The advantage of this method against the aforementioned doubling hypothesis is that we now have comparable concentrations throughout all cells highlighting regulatory mechanisms. The resulting Figure 4.6 D shows the concentration over time.

4.3.6 Life times and birth time

Figure 4.11 A shows a histogram of cellular lifetimes with an average lifetime of 11.6 hours. There are only few cells with a higher lifetime than 15 hours or lower than 10 hours. To investigate if the lifetime changes during the movie we analyze the lifetime of each cell against its birthtime. In Figure 4.11 B we find that the lifetime decreases with increasing birthtime in the movie. This is in accordance with the general knowledge that progenitor cells like MEPs or GMPs proliferate faster then stem cells [40].

Plotting the median net production rate of each cell against its birthtime reveals developing subpopulations (Figure 4.11 C). Between t=0 and t= 35h cells proliferate with positive production rate. After about 35 hours one subpopulation with a positive production rate and one with a negative rate appear, possibly representing GMP and MEP lineage, respectively.

Figure 4.11 D displays the median production rate against the cell lifetime. We find that cells with shorter lifetimes differ in their production rate of PU.1 level strongly, representing differentiating cells. Cells with a higher lifetime represent one cell type of an earlier state in the blood cell differentiation.

The dominance of positive production rates in Figure 4.11 C and D is due to the fact that cells tend to differentiate towards the GMP lineage in our experiment. Red dots stand for cells with active FCγ surface marker indicating GMP commitment. All cells show positive production rates indicating a correlation between high PU.1 concentrations and GMP commitment.

4.3.7 Cell Cycle

In order to compare cells with different lifetimes, we normalized every single-cell time course to the same length. This method does not take into account the diversity of cell phase timescales. It assumes that every cell phase has the same length in every single cell. This makes it possible to create a mean cell cycle time course of PU.1 expression (Figure 4.12 A). Calculating the *net production rate* leads to the mean PU.1 production over the cell life time where three different regions appear (Figure 4.12 B). This analysis was also performed in the former thesis of Jan Krumsiek and the results are almost equal [26]. The derivative of the absolute intensities can be compared and both studies show the same development. The strongly varying expression strengths between the supposed phases can also be observed in our plot. The first section is described by a decreasing slope indicating a reduced production of PU.1 which can be referred to the G1 phase. In the second phase the PU.1 production is very low since a cell in the S phase concentrates on DNA replication. After the DNA is doubled the PU.1 production again increases indicating cell growth of the G2 phase for the pending cell division.

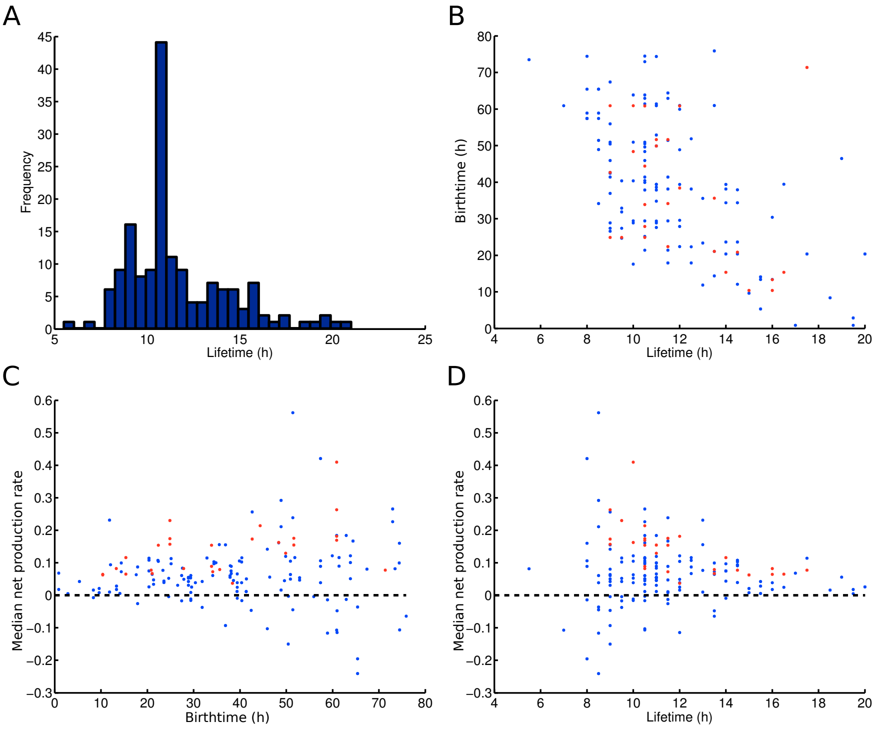


Figure 4.11: Various properties of 162 cells of experiment 3. Red dots indicate cells which activate FCγ. (A) Histogram of cell lifetimes with a mean of 11.6 hours. (B) Cell lifetime against birthtime showing an decreasing proliferation rate in the later movie phase. (C) The birthtime of all cells against their median net production rate, showing two arising populations with a high and a low net production rate after 35 hours. (D) The cell lifetimes against their median net production rates, showing similar results as the cells with longer lifetime. Presumably this indicates that MPPs have a homogeneous production rate whereas cells with shorter lifetime again build up two populations possibly standing for GMP and MEP lineages.

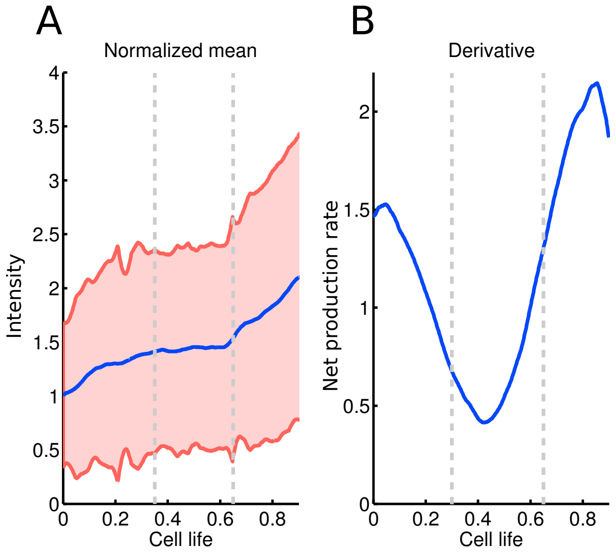


Figure 4.12: (A): Cell cycle of 211 cells of experiment 3 and 8 (compare Table 2.1) normalized to the same length. The blue line represents the mean fluorescence intensity, red lines the standard deviation. (B): The first derivative shows that the slope constantly decreases in the first cell phase and increases again after half of the cell lifetime. Vertical gray lines indicate the three supposed cell phases.

4.4 FCγ marker

FCγ is a surface marker indicating GMP lineage commitment. We try to link the onset of FCγ with the PU.1 expression profiles to give insights into the kinetics of blood cell differentiation. The challenge in this part consists of comparing many cells with the same annotation (FCγ positive) but different onsets and different mother cells, even from different experiments (see 3.4).

4.4.1 Correlation with PU.1 and onset

As known from the literature high PU.1 expressions lead to GMP commitment and thus FCγ onset [23, 47]. This can also be observed in our data: Figure 4.13 A shows all measured intensities of FCγ positive and FCγ negative cells. As expected, cells with an active GMP marker have a higher expression of PU.1.

When does the FCγ switch on in our cells? In Figure 4.13 B the frequency distribution over the generations in which an onset of FCγ is observed showing that even early generations commit to the GMP lineage. Figure 4.13 C shows the onset with respect to the movie time, again showing that even in the early time of the movie GMP commitment is noticeable. However, there is no onset before 11 hours giving an indication of the timescale needed from MPPs to complete GMP lineage commitment. The distribution over the whole movie shows no specific maximum. In later analyses, only cells which activate FCγ at later timepoints should be examined since the FACS MPP sorting might contain already committed cell populations. Excluding the first GM committing cells assures that the we observe a whole populations beginning with MPPs.

To check if the FCγ switches on in a specific cell phase we calculated the relative onset by dividing a cell life time by the timepoint of the first appearance of the marker. Figure 4.13 D shows that most cells tend to activate FCγ in the later lifespan.

4.4.2 Protein memory

The next step is to include cell genealogies into this analysis in order to see whether certain PU.1 levels are maintained over generations. Looking at the first intensities after division over the whole tree can give a hint when the cell fate decision was made. Cells which activate FCγ in a generation ≥ 3 are examined and the median of its first three intensities are compared to the first three intensities of earlier generations. The difference of the FCγ activating cell against its mother cell is referred as δ^{-1}, to its grandmother δ^{-2} and the difference against the grand-grandmother is δ^{-3}, shown in Figure 4.14 A. The boxplot of δ^{-1} is close to zero, indicating that these two generations have approximately the same PU.1 level at their first timepoints. The delta of earlier generations increases slightly indicating that the former cells are still MPPs and no lineage decision has been made. Looking at the median net production rate instead of just the first intensities shows no significant difference in the δ values (Figure 4.14 B).

Taking more than only single values for each generation into account and plotting the whole single-cell lifetime of a FCγ activating cell along with its mother cell leads to Figure 4.15 A. A selection of twenty time courses is shown where the mother cell as well as the daughter cell is normalized to the same lifetime of 1. Sigal et al [50] developed a method to check if a certain protein shows a memory effect over more than one cell cycle meaning that a high expression of the mother cell will lead to high expression in its daughters and vice versa. The plots of Figure 4.15 A are used to create a matrix containing ranks of absolute values of every single-cell time course at each timepoint. The development of the ranks is given in Figure 4.15 B where the colors indicate the initial rank of each time course ranging from blue (lowest rank) to red (highest rank). One would expect that in the end of the time courses the distribution of the colors is maintained if the cells have a high memory effect. In our case the colors are mixed up even after a short time interval and only few cells retain their ranks. To illustrate this effect, the autocorrelation based on the ranking matrix is shown in

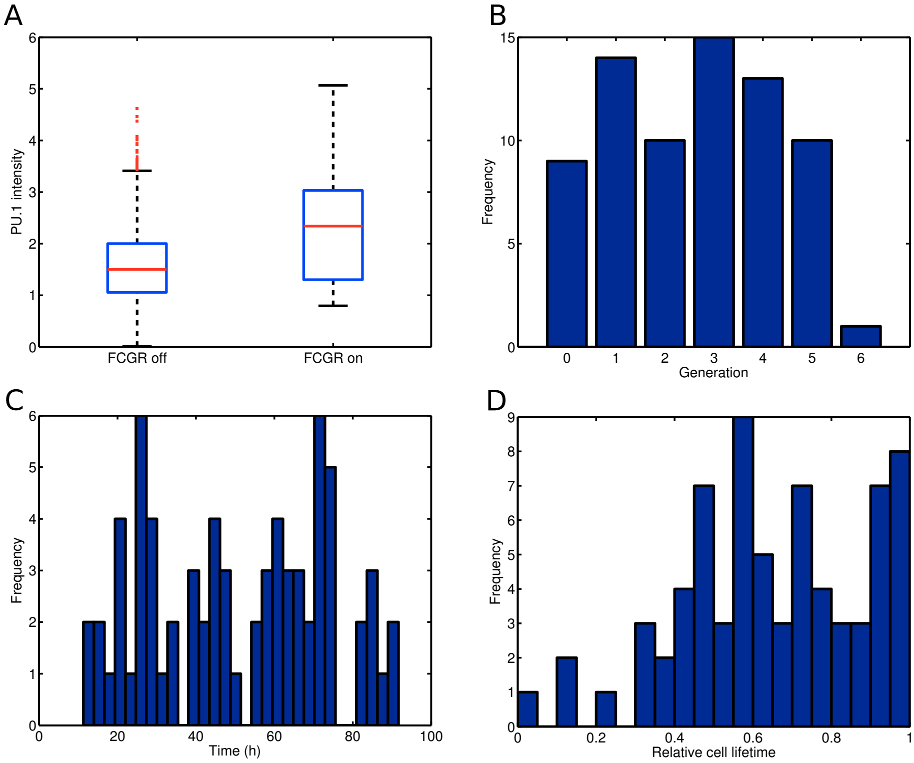


Figure 4.13: (A) Comparison of measured PU.1 intensities of FCγ positive and negative cells showing a higher PU.1 level for cells with active FCγ marker. (B) The distribution of GMP lineage decision with respect to the generation shows that even early cells commit to the lineage. (C) The distribution of FCγ activating cells with respect to the movie time can give an indication for the timescales needed for a MPP committing towards GMP. (D) The relative onset of all cells which commit to the GMP lineage, indicating a rather late onset during the cell-cycle.

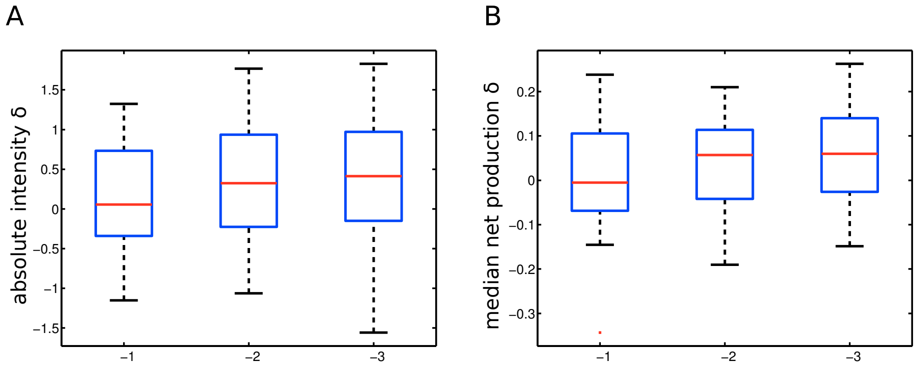


Figure 4.14: (A) The δ is derived by taking the first measurements of cell which activates FCγ and subtracting the first measurement of the mothercell (-1). The difference to the grandmother is indicated by -2, the difference to the grand-grandmother is indicated by -3. The difference increases over generations indicating that an equal PU.1 level is only maintained over one generation and that former cells have a lower PU.1 level. (B) Here, the δ is derived by taking the median net production rate instead of only the first measurements leading to the same results.

Figure 4.15 D. The sharp drop indicates that the initial ranks are rapidly lost. Calculating the autocorrelation on the real intensity values shows an appearing correlation for $\tau = 1$ as the values reach approximately the same amount after one cell cycle (Figure 4.15 C). The ranking of the full set of 102 cells does not show any difference (data not shown) which confirms that cells exhibits a very short memory with respect to PU.1 expression.

4.4.3 Shifted time courses

Since we assume that a specific expression profile of PU.1 leads to GMP commitment we investigate FCγ activating time courses in more detail. But the different onsets of FCγ in the cell-cycle (Figure 4.13 D) make it difficult to compare the expression time courses of GMP committing cells.

Therefore, we present an attempt where cell time courses which activate FCγ are shifted in time in order to set the activation event at time zero. Additionally, we take all mother cells in front of the FCγ positive cells (Figure 4.16 A). In this graph some mother cells are plotted twice but shifted differently because both daughters switch the surface marker on but at different timepoints. Taking the first derivative and averaging over the cells the mean along with its standard deviation can be plotted (Figure 4.17 A).

This plot shows that there is a higher PU.1 production about 10 hours in front of the FCγ onset. Due to the differently shifted time courses this should not be an effect of the cell-cycle. The first and the last timepoints

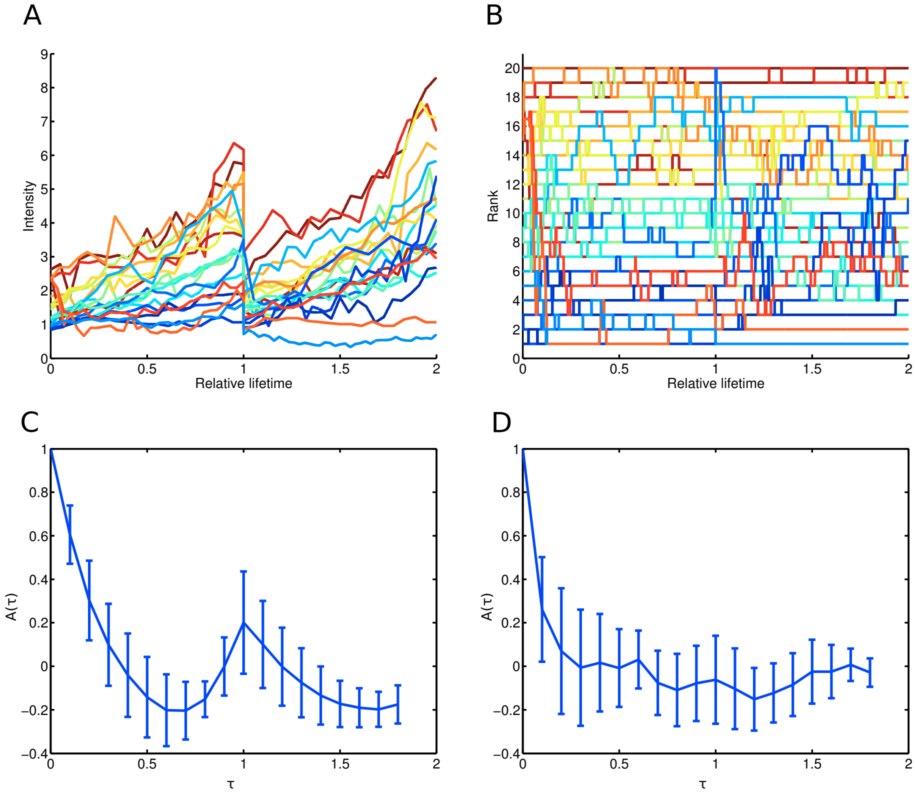


Figure 4.15: (A) Several single-cell time courses over two generations are taken and each cell is normalized to the same length of 1. They are colored by their first intensity values ranging from blue (low) to red (high). (B) The time courses are ranked at each timepoint and the colors by the initial rank from blue to red as in (A). The color scheme should be obtained if PU.1 has a protein memory over generations but mixes up in our case. (C) The autocorrelation function of absolute intensities showing a peak at $\tau = 1$ since similar intensities are maintained after one cell cycle. (D) The autocorrelation of the ranking matrix indicating no correlation of the ranks.

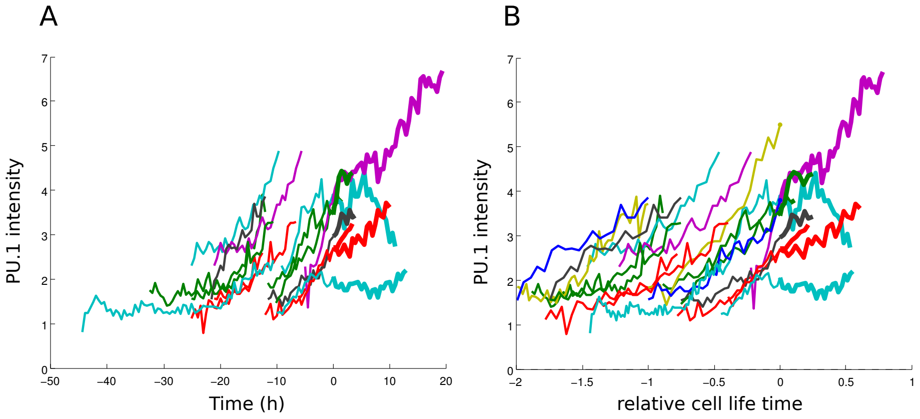


Figure 4.16: Cells of different trees each with a different onset of the FCγ marker which is shifted to zero. The time courses after the FCγ activation are indicated by bold lines. For every cell its mother time course is shown right before it. (A) The time courses are shown in absolute timescale. (B) The cells are normalized to the same cell lifetime of length 1.

have no meaning since there are only few measurements. But creating the same plot with an FCγ negative set and a randomly chosen shift shows a similar development. There are only a few FCγ positive cells which could be used in this method resulting in huge errorbars. More tracked cells are needed to investigate if a significant expression profile can describe a GMP lineage decision.

A different idea is that the visible PU.1 differentiation decision initiation cannot be measured in absolute time but in cell cycles. We assume that normalizing the cells to the same length, a peak would be visible if this is the case. Figure 4.16 B shows the same cells of Figure 4.16 A but normalized to the length 1. Again all cells are shifted such that the onset is on timepoint zero. All mother cells are added in front of every cell again normalized to the same length. Looking at the average of the first derivative neither shows any significant expression pattern (data not shown).

Conclusion

Combining the results of this section, we can affirm that a higher PU.1 level can be observed in GMP committed cells as well as right before the onset of the FCγ marker. A significant lineage decision in the PU.1 expression profile could not be observed so far and needs further investigations. An approximate timescale can be estimated of taking about 10 hours from observing a lineage decision in MPPs until finished cell commitment to GMPs. This

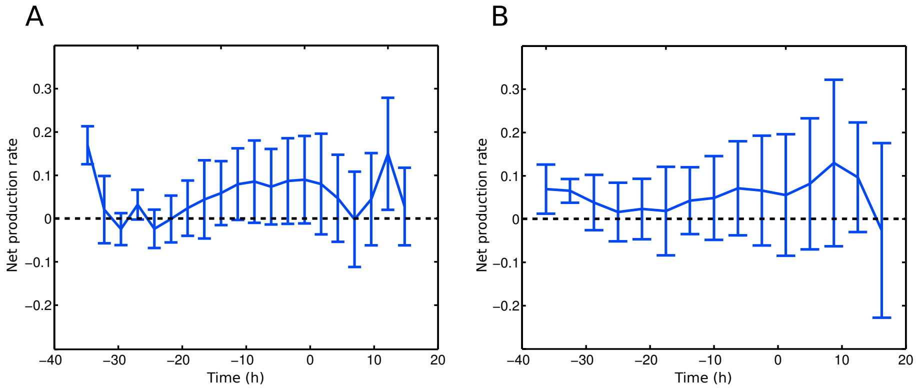


Figure 4.17: (A) The values of Figure 4.16 A are taken and the average net production rate with its standard deviation is shown. (B) A set of FCγ negative cells is processed by the same method. The huge errorbars and the same shape of the two plots indicate that there is no significant PU.1 expression pattern which would conclude a GMP commitment.

leads to the assumption that the decision has to be made in the mothercell of a GMP.

4.5 Further tree analysis

4.5.1 Division statistics

Assuming that that every PU.1 protein in a cell has the same chance to end up in either of the two daughter cells. Therefore this process can be described by a binomial distribution but instead of having the real protein amount we can only observe an arbitrary intensity value. The binomial distribution can be estimated by a normal distribution for large n which is applicable for our data since there should be more than $n > 100$ proteins in a cell [27, 39].

The histogram Figure 4.18 A shows the distribution of the offspring percentage according to the intensity of the respective mother cell. A test against a normal distribution can be illustrated in a *Q-Q-Plot* (Figure 4.18 B). This method divides two distribution into quantiles and plots the values per quantile of each distributed against each other leading to a bisecting line for two equal distributions. Our plot shows some deviations at the ends of the line due to few outliers at the edges of the histogram. But overall, the bisecting line is reproducible. Another test to validate that our data behaves like a normal distribution is derived by calculating the 95% confidence intervals by a maximum likelihood estimator. This results in significantly small intervals for the mean illustrated in Figure 4.18 A by red

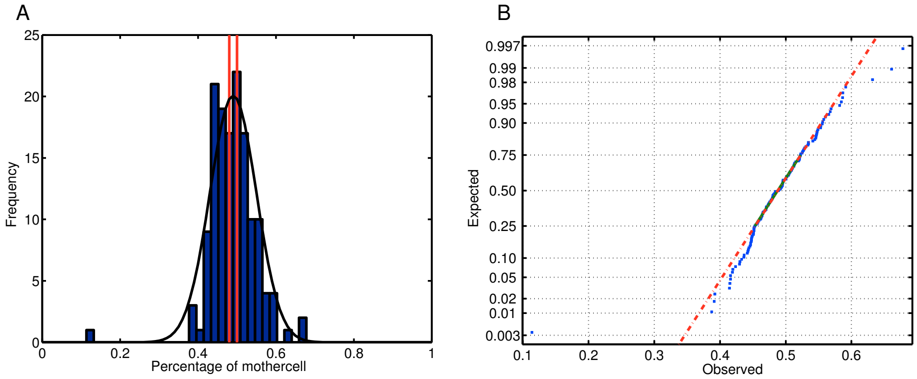


Figure 4.18: (A) Distribution of relative intensity of a daughter cell after cell division. The red bar indicate an maximum likelihood estimation of the mean. (B) A quantile-quantile plot of the percentage of a daughter cell against a normal distribution. Both plots indicate that MPPs and their offspring divide symmetrically.

lines. The maximum likelihood estimator also gives confidence intervals of the standard deviation which is also very small

By combining these results we can infer that the distribution of the relative mothercell intensity is normal distributed with a mean of 0.5. Therefore we conclude that MPPs and their progeny divide symmetrically.

4.5.2 Estimation of protein amount

In the work of Rosenfeld et al. [45], a method for the estimation of absolute protein amounts from the distribution of cell divisions was proposed. As shown above, cells divide symmetrically and thus every daughter cell has approximately half of the intensity of its mother. Ideally, the difference of both daughters will be close to zero. Assuming now that the proteins will be distributed based on a stochastic process and every protein has the same probability of $p = 0.5$ to end up in either of the daughter cells, then the daughter cell proteins follow a binomial distribution.

Since we assume linearity between the fluorescence signal S and the real protein amount N,

$$S = N \cdot c \tag{4.6}$$

with a constant factor c. The protein amount in the mothercell N_{tot} is given by:

$$N_{tot} = N_1 + N_2$$

where $N_{1,2}$ represent the daughter cell protein numbers. The expectation value and the variance of a daughter cell following from the binomial distribution are [56]:

$$E(N_1) = \bar{N}_1 = N_{tot} \cdot p = \frac{N_{tot}}{2}$$

$$\sigma^2 = N_{tot} \cdot p \cdot q = \frac{N_{tot}}{4} \tag{4.7}$$

Furthermore:

$$\begin{aligned}\sigma^2 &= \left\langle (\bar{N}_1 - N_1)^2 \right\rangle \\ &= \left\langle \left(\frac{N_{tot}}{2} - N_1 \right)^2 \right\rangle \\ &= \left\langle \left(\frac{N_1 + N_2}{2} - N_1 \right)^2 \right\rangle \\ &= \left\langle \left(\frac{N_2 - N_1}{2} \right)^2 \right\rangle \end{aligned} \tag{4.8}$$

Combining equations (4.7) and (4.8) leads to

$$\left\langle (N_2 - N_1)^2 \right\rangle = N_{tot}$$

Adding equation (4.6) :

$$c^2 \cdot \left\langle (N_2 - N_1)^2 \right\rangle = c^2 \cdot N_{tot}$$

$$\left\langle (S_2 - S_1)^2 \right\rangle = c \cdot S_{tot} \tag{4.9}$$

This gives us the opportunity to estimate the fluorescence intensity factor c by fitting the total intensities of the mother cell against the difference of the two daughter cells in order to conclude the protein amount.

The last timepoint of every mother and the first timepoints of both daughters is taken resulting in Figure 4.19 A, showing the difference of both daughters on the y-axis and their corresponding mother intensity on the x-axis. A linear function through the origin is fitted into the data cloud. The resulting gradient denotes the fluorescence factor c which is used to calculate the protein amount of the intensity of the mother by equation 4.10. The resulting distribution is shown in the histogram in Figure 4.19 B ending up with an estimation of approximately 214 proteins.

Intuitively this number sounds rather small because the transcription factor PU.1 is assumed to play a major role in the myeloid lineage decision and one would suggest a greater number to fulfill all of its functions.

The absolute amount of protein is essential for a general interest and certifies stochastic modeling which can be applied if the protein amount is that low.

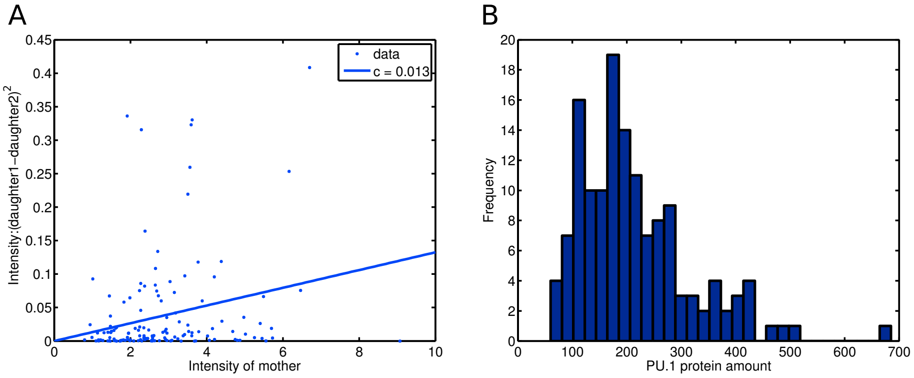


Figure 4.19: (A) Fitting a line through the intensity of mother against the squared difference of its daughter intensities leading to the fluorescence factor c. (B) The resulting histogram shows the calculated protein amount in each mother cell with a mean of about 214 proteins.

4.5.3 Tree visualization

Since the representations of single-cell time courses of a whole genealogy introduced so far will rapidly become confusing with higher tree depth, we introduce an innovative perspective of a cellular tree (compare Figure 2.2 and 4.20). On the y-axis, the real time in hours is plotted. It shows the mothercell with all their children and grandchildren as already shown in Figure 2.2. In this view the thickness of the life lines and the color of it can be used to represent more information. In the example shown in Figure 4.20 A, the same information is plotted twice. The thickness as well as the color represent the absolute measured intensity of the cell and illustrate in an intuitive way that there are two distinct sections of the tree. The left part shows less PU.1 intensity whereas the right part indicates high PU.1 levels, and finally the whole progeny switches to FCγ positive cells indicated by black bars. Interestingly, one cell on the left part also switches to GMP lineage but only a small increase in PU.1 level can be observed. It is also possible to allocate the two dimensions with other methods we already introduced previously. Figure 4.20 B shows the absolute intensity in x-axis whereas the color scheme is based on the first derivative highlighting a higher PU.1 production in cell 8 which could explain the final GMP commitment. It takes further investigation if the cell had some contact with other GMP cells which could lead to this commitment. But cell-cell contact could not be explored in this thesis. Furthermore cell 19 shows also a higher PU.1 production but further tracking would be necessary to confirm final GMP commitment.

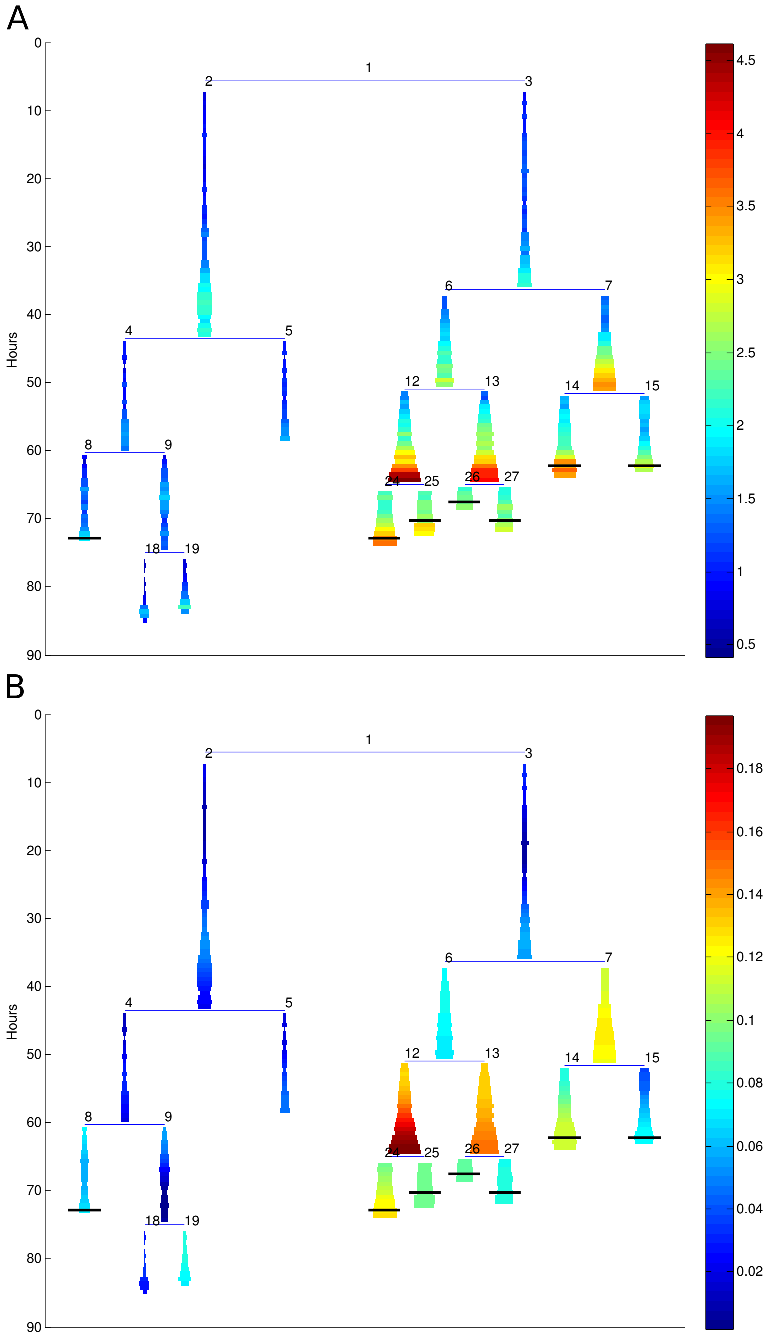


Figure 4.20: Cell tree over time (y-axis) with two additional features. (A) The information of absolute intensity is plotted twice. The thickness as well as the color represents the measured PU.1 level. (B) The color scheme is coded by the production rate and the absolute PU.1 values are on the x-axis providing a better visualization. The black bars indicate an onset of the FCγ surface marker.

Chapter 5

Summary and outlook

Single-cell analysis of time-lapse microscopy is a increasing popular field which provides novel insights of hematopoietic differentiation by delivering accurate expression profiles on a single-cell level. One has to deal with many issues when using fluorescent images, such as background signal or uneven illumination. Wrong or incomplete correction methods can corrupt protein measurements and expression analysis. We developed a fast and effective method to correct for all theses issues and compared it to methods developed by other research groups leading to almost identical expression measurements. However, our method has the advantage of being more robust against outliers such as contaminations and is tailored to the experiment settings. The accurate examination of the experiment data and our close collaboration with the Institute of Stem Cell Research led to several important improvements of the experimental setup and normalization methods.

After normalizing the fluorescence signal, two approaches have been applied. In a large-scale approach we detected all cells in a whole experiment by a fully automatic detection pipeline providing the opportunity to investigate population wide expression profiles on a single-cell level. The complex process of detecting cells on fluorescence images has been implemented into a custom software toolbox. This software allows to perform the second, a small-scale approach by additionally including tracking data of single-cells. With this information it is possible to analyze individual single-cell expression profiles of YFP tagged PU.1 proteins over time and over a whole progeny of cells. A powerful detection algorithm is necessary in order to achieve accurate cell intensity measurements. Our tool highlights potential errors of the automatic cell detection process and provides easy and intuitive ways to correct these failures by manual inspection. It is applicable to other data and has been used in other projects. For example, imaging experiments of *nanog* labeled embryonic stem cells can be measured, although they build up colonies, by detecting them on a different fluorescently tagged membrane protein.

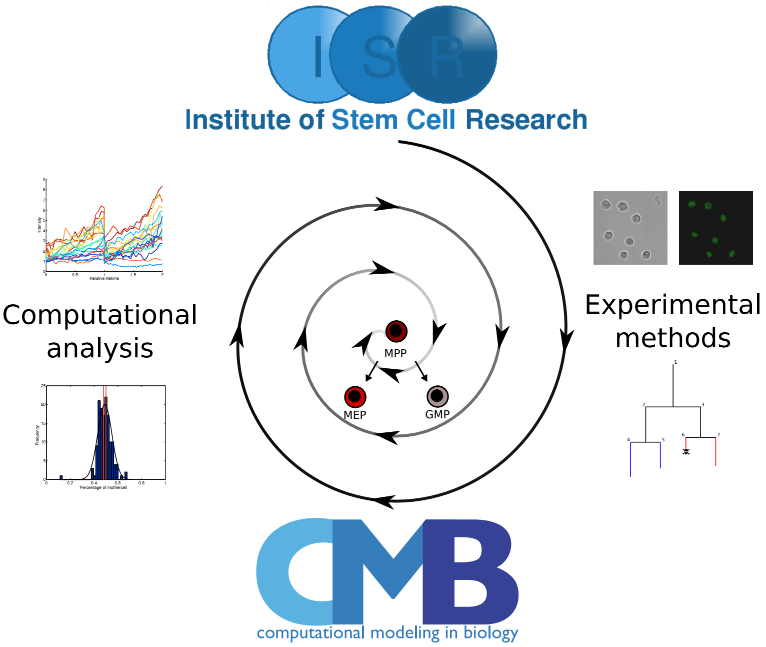

ISR
Institute of Stem Cell Research
Computational analysis
Experimental methods
MPP
MEP
GMP
CMB
computational modeling in biology

Figure 5.1: The close collaboration of the Institute of Stem Cell Research with our group led to important experimental and methodical improvements in order to obtain most accurate single-cell expression profiles. The results of this work as well as the further collaboration provide promising new insights into the myeloid differentiation process.

After correcting for extrinsic noise (like background signal, lamp flickering, etc.) it is important to estimate the remaining noise in the data, especially when investigating biological mechanisms on a molecular level. We showed that after our correction procedure and manual curation of single-cell data only a small variance can be assigned to the measurement techniques.

We developed a new method to analyze single-cell time courses by normalizing the measured intensity by an estimated cell volume which results in the relative PU.1 concentration in each cell. In order to represent time courses of a whole cell with its progeny we created a new tree visualization. This illustration is used to display the information of cell annotations along with expression signal and expression development in a single figure providing intuitive understandings of processes during cell differentiation.

Additional annotations describing the status of cell-specific surface markers which indicate finished cell commitment allow to investigate lineage specific commitment decisions in PU.1 expression signals. It is known that high levels of PU.1 indicate granulocyte/macrophage (GM) lineage commitment [16]. A first confirmation could be given by basic analysis methods which showed that cells with active FCγ marker representing GM lineage have high PU.1 levels. We discussed that multipotent progenitor cells (MPPs) can commit to the GM lineage after about eleven hours and activate the FCγ marker even in the first generations. The distribution of FCγ activation in the relative cell lifetime showed that MPPs tend to activate FCγ in the late cell-cycle.

At last we showed that MPPs and their offspring divide symmetrically. Taking the relative intensity of every daughter with respect to the intensity of their mother a histogram with mean 0.5 was obtained. Based on this result a statistical method could be applied and the number of proteins in each cell could be estimated. An unexpectedly small number of only 100 - 500 proteins in each cell turned out which will be experimentally validated in the near future. Small protein abundances suggest the usage of stochastic modeling and could explain the fluctuations which are observable in the single-cell time courses. Another experiment is planned where a higher time resolution of fluorescence images could lead to new insights of the single-cell variances.

The vast amount of data which is produced by every imaging experiment cannot entirely be inspected by eye. Therefore it is of great interest to automatize the data preparation. There are many automated tracking approaches which would give the opportunity to enrich the data we get from an experiment so far. Ideally this would lead to an accurate tracking of cells on brightfield images which have a higher time resolution than fluorescent images. First attempts showed that this goal is not trivial and needs further investigation since the detection of cells on brightfield images essentially differs from fluorescence images. The analysis of PU.1 expression, especially before an FCγ activation, would strongly benefit from more time courses and could lead to more significant results. A fully tracked experiment would allow to automatically investigate cell-cell contact which could clarify if such interactions have an impact on the myeloid differentiation. Analyses could show if lineage decisions can be transferred by cell-cell signaling and if this signal forces both cells to commit to the same or to different lineages.

Several new experiments are planned based on the results we outlined in this study which can give evidences for the assumptions we made. Having a functional MEP marker and therefore having sure evidence for the both major differentiation cell types of MPPs the single-cell analysis could be extended showing more distinct PU.1 expression profiles. A new transgenic mouse with a fluorescently tagged GATA-1 transcription factor, which is the main competitor of PU.1 in common myeloid progenitors, is explored. If this mouse is healthy and a crossbreeding together with the PU.1 mouse will be successful new opportunities to investigate the myeloid cell differentiation would arise. Single-cell time courses of the two major transcription proteins expressed in the same cell could give new insights into the mechanisms of lineage decision of myeloid progenitor cells. The readout of another transcription factor would also allow to test current mathematical models which describe the PU.1-GATA-1 switch. Having real time courses also allows to establish own models and to fit these to the data which hopefully describes the myeloid lineage decision in a more accurate way.

The results of this work promise new insights into the fundamental processes of blood cell differentiation and, ultimately, lead to new clinical applications in order to improve treatment of severe hematopoietic diseases.

Bibliography

[1] *Stem Cells: Scientific Progress and Future Research Directions.* Department of Health and Human Services, 2001.

[2] Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774):193–197, 2000.

[3] Benson, D.M., Bryan, J., Plant, A.L., Gotto, A.M., and Smith, L.C. Digital imaging fluorescence microscopy: spatial heterogeneity of photobleaching rate constants in individual cells. *J Cell Biol*, 100(4):1309–1323, 1985.

[4] Boisnard-Lorig, C., Colon-Carmona, A., Bauch, M., Hodge, S., Doerner, P., Bancharel, E., Dumas, C., Haseloff, J., and Berger, F. Dynamic analyses of the expression of the histone::yfp fusion protein in arabidopsis show that syncytial endosperm is divided in mitotic domains. *Plant Cell*, 13(3):495–509, 2001.

[5] Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P., and Sabatini, D.M. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*, 7(10):R100, 2006.

[6] Chen, H., Ray-Gallet, D., Zhang, P., Hetherington, C.J., Gonzalez, D.A., Zhang, D.E., Moreau-Gachelin, F., and Tenen, D.G. Pu.1 (spi-1) autoregulates its expression in myeloid cells. *Oncogene*, 11(8):1549–1560, 1995.

[7] Cheshier, S.H., Morrison, S.J., Liao, X., and Weissman, I.L. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proc Natl Acad Sci U S A*, 96(6):3120–3125, 1999.

[8] Chickarmane, V., Enver, T., and Peterson, C. Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol*, 5(1):e1000268, 2009.

[9] Choe, K.S., Ujhelly, O., Wontakal, S.N., and Skoultchi, A.I. Pu.1 directly regulates cdk6 gene expression, linking the cell proliferation and differentiation programs in erythroid cells. *J Biol Chem*, 2009.

[10] Cohen, A.A., Kalisky, T., Mayo, A., Geva-Zatorsky, N., Danon, T., Issaeva, I., Kopito, R.B., Perzov, N., Milo, R., Sigal, A., and Alon, U. Protein dynamics in individual human cells: experiment and theory. *PLoS One*, 4(4):e4901, 2009.

[11] Eilken, H.M., Nishikawa, S.I., and Schroeder, T. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature*, 457(7231):896–900, 2009.

[12] Falk, M.M. and Lauf, U. High resolution, fluorescence deconvolution microscopy and tagging with the autofluorescent tracers cfp, gfp, and yfp to study the structural composition of gap junctions in living cells. *Microsc Res Tech*, 52(3):251–262, 2001.

[13] Fisher, R.C. and Scott, E.W. Role of pu.1 in hematopoiesis. *Stem Cells*, 16(1):25–37, 1998.

[14] Foster, S.D., Oram, S.H., Wilson, N.K., and Göttgens, B. From genes to cells to tissues-modelling the haematopoietic system. *Mol Biosyst*, 2009.

[15] Fridman, W.H. Fc receptors and immunoglobulin binding factors. *FASEB J*, 5(12):2684–2690, 1991.

[16] Friedman, A.D. Transcriptional control of granulocyte and monocyte development. *Oncogene*, 26(47):6816–6828, 2007.

[17] Gardner, T.S., Cantor, C.R., and Collins, J.J. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[18] Glauche, I., Lorenz, R., Hasenclever, D., and Roeder, I. A novel view on stem cell development: analysing the shape of cellular genealogies. *Cell Prolif*, 42(2):248–263, 2009.

[19] Graf, T. and Enver, T. Forcing cells to change lineages. *Nature*, 462(7273):587–594, 2009.

[20] Huang, S., Guo, Y.P., May, G., and Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol*, 305(2):695–713, 2007.

[21] Indik, Z.K., Park, J.G., Hunter, S., and Schreiber, A.D. The molecular dissection of fc gamma receptor mediated phagocytosis. *Blood*, 86(12):4389–4399, 1995.

[22] Iwasaki, H. and Akashi, K. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity*, 26(6):726–740, 2007.

[23] Kastner, P. and Chan, S. Pu.1: a crucial and versatile player in hematopoiesis and leukemia. *Int J Biochem Cell Biol*, 40(1):22–27, 2008.

[24] Kirstetter, P., Anderson, K., Porse, B.T., Jacobsen, S.E.W., and Nerlov, C. Activation of the canonical wnt pathway leads to loss of hematopoietic stem cell repopulation and multilineage differentiation block. *Nat Immunol*, 7(10):1048–1056, 2006.

[25] Koschmieder, S., Rosenbauer, F., Steidl, U., Owens, B.M., and Tenen, D.G. Role of transcription factors c/ebpalpha and pu.1 in normal hematopoiesis and leukemia. *Int J Hematol*, 81(5):368–377, 2005.

[26] Krumsiek, J. *Computational modeling of regulatory networks in hematopoietic differentiation*. Master's thesis, Ludwig-Maximilians-Universität München, Technische Universität München, 2009.

[27] Laplace, P. *Théorie analytique des probabilités*. Courcier, 1820.

[28] Larson, D.R., Singer, R.H., and Zenklusen, D. A single molecule view of gene expression. *Trends Cell Biol*, 19(11):630–637, 2009.

[29] Losick, R. and Desplan, C. Stochasticity and cell fate. *Science*, 320(5872):65–68, 2008.

[30] Malhotra, A. Tagging for protein expression. *Methods Enzymol*, 463:239–258, 2009.

[31] Michel, R., Steinmeyer, R., Falk, M., and Harms, G.S. A new detection algorithm for image analysis of single, fluorescence-labeled proteins in living cells. *Microsc Res Tech*, 70(9):763–770, 2007.

[32] Mikkola, H.K.A. and Orkin, S.H. The journey of developing hematopoietic stem cells. *Development*, 133(19):3733–3744, 2006.

[33] Mosig, A., Jger, S., Wang, C., Nath, S., Ersoy, I., Palaniappan, K.P., and Chen, S.S. Tracking cells in life cell imaging videos using topological alignments. *Algorithms Mol Biol*, 4:10, 2009.

[34] Nerlov, C., Querfurth, E., Kulessa, H., and Graf, T. Gata-1 interacts with the myeloid pu.1 transcription factor and represses pu.1-dependent transcription. *Blood*, 95(8):2543–2551, 2000.

[35] Orkin, S.H. Diversification of haematopoietic stem cells to specific lineages. *Nat Rev Genet*, 1(1):57–64, 2000.

[36] Orkin, S.H., Shivdasani, R.A., Fujiwara, Y., and McDevitt, M.A. Transcription factor gata-1 in megakaryocyte development. *Stem Cells*, 16 Suppl 2:79–83, 1998.

[37] Orkin, S.H. and Zon, L.I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644, 2008.

[38] Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[39] Papoulis, A. and Pillai, S. *Probability, random variables and stochastic processes*. McGraw-Hill Education (India) Pvt Ltd, 2002.

[40] Passegu, E., Wagers, A.J., Giuriato, S., Anderson, W.C., and Weissman, I.L. Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates. *J Exp Med*, 202(11):1599–1611, 2005.

[41] Pronk, C.J.H., Rossi, D.J., Mnsson, R., Attema, J.L., Norddahl, G.L., Chan, C.K.F., Sigvardsson, M., Weissman, I.L., and Bryder, D. Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell*, 1(4):428–442, 2007.

[42] Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.

[43] Rieger, M.A., Hoppe, P.S., Smejkal, B.M., Eitelhuber, A.C., and Schroeder, T. Hematopoietic cytokines can instruct lineage choice. *Science*, 325(5937):217–218, 2009.

[44] Roeder, I. and Glauche, I. Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors gata-1 and pu.1. *J Theor Biol*, 241(4):852–865, 2006.

[45] Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, 2005.

[46] Schroeder, T. Asymmetric cell division in normal and malignant hematopoietic precursor cells. *Cell Stem Cell*, 1(5):479–481, 2007.

[47] Shibuya, A. and Honda, S.I. Molecular and functional characteristics of the fcalpha/mur, a novel fc receptor for igm and iga. *Springer Semin Immunopathol*, 28(4):377–382, 2006.

[48] Shimizu, R., Trainor, C.D., Nishikawa, K., Kobayashi, M., Ohneda, K., and Yamamoto, M. Gata-1 self-association controls erythroid development in vivo. *J Biol Chem*, 282(21):15862–15871, 2007.

[49] Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Alaluf, I., Swerdlin, N., Perzov, N., Danon, T., Liron, Y., Raveh, T., Carpenter, A.E., Lahav, G., and Alon, U. Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nat Methods*, 3(7):525–531, 2006.

[50] Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–646, 2006.

[51] Song, L., Hennink, E.J., Young, I.T., and Tanke, H.J. Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy. *Biophys J*, 68(6):2588–2600, 1995.

[52] van der Wath, R.C., Wilson, A., Laurenti, E., Trumpp, A., and Li, P. Estimating dormant and active hematopoietic stem cell kinetics through extensive modeling of bromodeoxyuridine label-retaining cell dynamics. *PLoS One*, 4(9):e6972, 2009.

[53] Warren, L., Bryder, D., Weissman, I.L., and Quake, S.R. Transcription factor profiling in individual hematopoietic progenitors by digital rt-pcr. *Proc Natl Acad Sci U S A*, 103(47):17807–17812, 2006.

[54] Worth, D.C. and Parsons, M. Live cell imaging analysis of receptor function. *Methods Mol Biol*, 591:311–323, 2010.

[55] Yeung, J. and So, C.W.E. Identification and characterization of hematopoietic stem and progenitor cell populations in mouse bone marrow by flow cytometry. *Methods Mol Biol*, 538:301–315, 2009.

[56] Yule, G. *An introduction to the theory of statistics.* C. Griffin and company, limited, 1911.

[57] Zhang, P., Behre, G., Pan, J., Iwama, A., Wara-Aswapati, N., Radomska, H.S., Auron, P.E., Tenen, D.G., and Sun, Z. Negative cross-talk between hematopoietic regulators: Gata proteins repress pu.1. *Proc Natl Acad Sci U S A*, 96(15):8705–8710, 1999.