Technische Universität München
Ludwig Maximilians Universität München

Helmholtz Zentrum München
Institut für Bioinformatik und Systembiologie

Bachelorarbeit
in Bioinformatik

Supervised lipid raft cluster identification in hematopoietic stem cells

Sven Punga

Aufgabensteller: Prof. Dr. Dr. Fabian Theis
Betreuer: Michael Schwarzfischer
Abgabedatum: 15.3.2013

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

15. März 2013 _____

Sven Punga

Abstract

Haematopoiesis is the lifelong process of replenish dead blood cells out of hematopoetic stem cells (HSC) and is a key component for the life of vertebrate animals. HSC are able to differentiate into the full variety of mature blood cells and reside in the bone marrow of adult mammals. Fluorescence activated cell sorting (FACS) allows us to isolate and purify HSC's from a cell population which has been extracted from mice bone marrow. The purified cells can be used to research their behavior and gain insight into yet unknown intracellular processes.

Lipid raft microdomains are cholesterol- and glycosphingolipid-enriched patches in the plasma membrane into which various functional molecules are distributed. LRC's act as platforms of cellular functions and play a huge role in the regulation of HSC dormancy. A deeper understanding of the cellular relevance of LRC's within HSC's during differentiation would help to improve treatments of a variety of genetic disorders, acquired states of bone marrow failure, and cancers like leukemia.

Time lapse microscopy and continuous imaging of cell cultures treated with fluorescent markers is a technique that has already proven to give great insight into cellular processes. Vannini et al. presented a semi-automated image analysis tool to map the presence or absence of lipid raft clusters (LRC) in HSC's in vitro. In this work we want to reproduce and evaluate the existing approach and develop an improved method for LRC prediction in fluorescence cell microscopy images. We designed a 'cluster annotation tool', for an fast and easy annotation of the occurrence of LRC in HSC, to create a 'gold standard' that can be used in the evaluation process. The manually created test set allowed us to use supervised machine learning techniques to determine the occurrence of LRC. We got time lapse microscopy experiment data that were performed under different conditions, providing us a reliable positive and negative control. Since LRC's occur as bright spots within the cell image, we used clustering techniques like the kmeans algorithm and gaussian mixture models to divide the cell into disjunctive subsets. The clustering was performed on the cell intensities corresponding to the distribution of the fluorescent marker in the cell. Based on the clustering we determined self defined cell and cluster features. The distributions of the features of cells with and without LRC showed differences which indicates the usability of this feature for a successful LRC prediction. These features we were used to train a decision tree, a machine learning technique that classifies the cell images based on their cell and cluster features. As the last step we performed a ten fold cross validation to gain a measure of the quality for the different prediction approaches. Especially the gaussian mixture model showed a great performance in the evaluation. Further improvements have been achived by including weighted x-y coordinates of the intensities into the data. We were able to improve the existing approach in sensitivity and specificity and provided a fully automated LRC prediction pipeline. With this pipeline we provide a tool for LRC prediction on time lapse microscopy data that helps researchers to prioritize their research targets.

In this work, we established a basis which combined with tracking information allows to investigate asymetric cell division. It further allows to correlate LRC dynamics to later cell fates. We hope that the pipeline helps to gain new insights into the understanding the relevance of LRC's.

Zusammenfassung

Als Hämatopoese wird der lebenslang andauernde Prozess des Ersetzens toter Blutzellen aus hämatopoetischen Stammzellen (HSC) bezeichnet. Hämatopoetische Stammzellen kommen im Knochenmark von Säugetieren vor und können sich in alle Blutzelltypen entwickeln. Mit Hilfe von Techniken wie 'fluorescence activated cell sorting (FACS)' ist es möglich HSC's zu isolieren und aufzureinigen. Lipid rafts sind mit cholesterol und glycosphingolipiden versetzte Gebilde und beinhalten wichtige funktionale Moleküle. Man kann sie als Plattformen für zelluläre Funktionen betrachten und spielen eine wichtige Rolle für den Zellzyklus. Tiefere Einblicke in die Bedeutung der Lipid raft cluster (LRC) wärend der Differenzierung der Zelle würde zu der Behandlung von diversen Erbkrankheiten und Leukämie beitragen. Zeitaufgelöste Beobachtung von fluoreszierenden Zellkulturen mit Hilfe eines hochauflösenden Mikroskops haben bereits neue Einsichten in intracelluläre Prozesse gegeben. Vannini et al. haben ein halbautomatisches Bildanalyseprogramm entwickelt um die Präsenz von LRC in hS in vitro zu bestimmen. In dieser Arbeit versuchen wir diesen bereits vorhandenen Ansatz zu reproduzieren sowie zu verbessern und anschließend zu auszuwerten. Wir haben ein 'cluster annotation tool' entwickelt welches eine einfache und schnelle Möglichkeit bietet die Anzahl der in HSC's vorkommenden LRC zu annotieren und somit einen 'Gold Standard' zu bekommen. Ein per Hand erstellter Trainigsdatensatz ermöglicht uns das trainieren von einer 'supervised machine learning' Methode, welche für die Vorhersage für LRC benutzt wird. Als Basisdaten haben wir zeitaufgelöste Mikroskopbilder von Experimenten die unter verschiedenen Bedingungen durchgeführt wurden. Wir haben sowohl eine positiv als auch eine negativ Kontrolle. Da LRC in den Mikroskopbildern als helle Punkte zu erkennen sind, benutzen wir Auftrennungsmethoden wie den kmeans Algorithmus und Gaus'sche Mischmodelle um die Zellen in verschiedene Cluster aufzuteilen. Durch die Aufteilung basierend auf Pixelintensitäten haben wir verschiedene Eigenschaften definiert. Die Verteilungen dieser Eigenschaften zeigen Unterschiede auf, woduch sie für eine Verwendung in der Vorhersage eignen. Mit diesen Eigenschaften haben wir einen 'decision tree' trainiert, welcher eine Klassifikation der Bilddaten vornehmen kann. Für die Evaluierung der Methoden haben wir eine zehnfache Kreuzvalidierung vorgenommen. Vorallem die Mischverteilungen, sowie die Datenerweiterung um die x-y Koordinaten der Pixelintensitäten für Auftrennungszwecke erzielten gute Ergebnise. Wir konnten eine Verbesserung in Sensitivität und Spezifität des existierenden Ansatzes erzielen, sowie eine automatisierte Pipeline für die Vorhersage von LRC in HSC Bildern. Mit dieser Arbeit haben wir die Grundlage dafür geschaffen asymetrische Zellteilung näher zu untersuchen. Basierend auf dieser Arbeit können neue Erkenntnisse auf dem Gebiet der LRC Forschung erzielt werden kann.

Contents

Chapter 1

Introduction

1.1 Hematopoiesis

The blood system holds a key position in higher organisms. It consists of a mixture of many cells such as the red blood cells that are responsible for the transport of oxygen to all organs. Mature blood cells only live for a short time and dieing permanently. They need to be replaced in order to maintain the blood system. This lifelong process is called hematopoiesis and is conserved throughout vertebrate evolution. To replenish the lack of blood cells, they were regained from hemaopoetic stem cells (HSCs) that reside in the bone marrow of adult mammals [1]. Long-term HSCs (LT-HSCs) are capable of renewing themselfs and the production of additional HSCs by dividing over long periods. They are in a multipotent state, meaning that they have the ability to differentiate to all blood cell linages[2]. The LT-HSCs sit atop of a hierarchy of progenitor cells that progressivly differentiate to cell types like erythrocytes, megakaryocytes or B / T-lymphocytes. LT-HSCs can become Short-term HSCs (ST-HSCs) which mark the beginning of the differentiation process. ST-HSCs loose their self renewal and reproduction ability, but raise to multipotent progenitor cells wich will differentiate into mature blood cells. To study HSCs they need to be identified and seperated from other cells. With the help of monoclonal antibodys directed to surface markers or considering their metabolic properties it is possible to identify HSC. Flourescence-activated cell sorting (FACS) is a method to seperate HSCs from other more commited progenitor cells [3]. With these modern techniques it is possible to highly purify HSCs for further research interests. Usually HSCs from immunodeficient mice were used to analyse the human blood system. Although the experimental system of mice is not entirely ideal for research of the human blood system, it is considered that the systems are transferable. It is known that the microenvironment of HSCs has a huge influence on the differentiation process [4]. However after years of research the exact effects of the microenvironment during the differentiation process remains unknown. Also the influence of whether intrinsic, extrinsic or both signal types determine cell fate decisions is a controversial question and still needs a lot of research[4]. The fact that asymmetric cell division as a regulatory mechanism has a influence on cell fate is well accepted. However in the past years it was not possible to prove a direct link of asymmetric cell division to the cell fate. It still remains an open field of research [4]. A good reason to explore the secrets of HSCs are the magnificent possibilities in the medical sector. For example the transplantation of marrow in human patients is a current therapy for a variety of genetic disorders, acquired states of bone marrow failure, and cancers like leukemia [1].

1.2 Lipid Raft Cluster

Lipid raft microdomains are cholesterol- and glycosphingolipid-enriched patches in the plasma membrane into which various functional molecules are distributed [5]. The components of proteins occuring in lipid raft clusters (LRC) include transmembrane antigens/receptors, GPI-anchored proteins,cytoskeletal proteins, Src-family kinases, G-proteins, and other proteins that are involved in signal transduction [6]. LRCs act as platforms of cellular functions and play a huge role in cytokine signaling pathways, membrane trafficking and cytoskeleton organization. Experiments by Yamazaki et al. have shown that the formation of LRC is a key event in the regulation of HSC dormancy without affecting the HSC potential of self-renewing and to differentiate into the full range of hematopoietic cell lineages. LRCs are also crucial for the proliferation of HSCs and their progenitor cells.They also assumed that LRCs induced by cytokines are essential for HSCs to reenter into the cell cycle by modulating cytokine signals. The size of LRCs tend to play an important role as well, since larger rafts have a greater potential for concentration of transducers and the exclusion of negative regulators [6]. The transforming-growth-factor-beta ($TGF-\beta$) as well as Methyl-beta-cyclodextrin are known to inhibit the formation of LRCs by depleting plasma membrane cholesterol. The todays research relies on chemicals like them to gain more knowledge about the role of LRCs in HScs. Unfortunately the more we purify HSCs the less cells we can obtain, what makes an application of conventional biochemical analysis to HSCs almost impossible. This is a reason why there is only little knowledge about the intracellular signaling events in HSCs [7]. One technique to overcome this problem is the time lapsed microscopy of cell cultures, which gives the opportunity to observe single cells over time and only a few cells are required. By using fluorescent protein markers is possible to measure their expression levels whereby we gain a better understanding of intracellular processes. This can also be used for the analysis of LRCs. Lipid raft clusters can be identified in time lapse microscopy data by bright spots as illustrated in Figure 2.3a.

1.3 Technical background

Most of the HSCs can be found in a nondividing quiescent state in the bone marrow (BM) niche of mammalian animals [6]. Unfortunately large scale experiments can not be performed on human HSCs, so 12 to 16 weeks old mice were used to extract their bone marrow. This marrow were used to isolate and purify fresh HSCs by fluorescence activated cell sorting (FACS). Of these isolated cells approximately 40% exhibit long-term repopulation (LTR) of hematopoiesis [5]. After the purification procedure the cells were cultured on a plastic slide in serum-free medium.

The Cholera toxin subunit B (CTB) is a sensitive neuroanatomical tracer commonly used in brightfield microscopy [8]. It is non toxic and binds to cholesterol a component that can be found in lipid membranes and thus can be used as a LRC marker. YFP is a yellow emission variant of the green fluorescent protein (GFP) from the jellyfish Aequorea victoria, which have been found useful in a variety of applications in biological systems [9]. However, the slow maturation remains a big obstacle to the use of GFP variants for visualization. Especially when cells were cultured at 37 degrees C [10]. Therefore mutagenesis studies trying continuously to improve the folding properties of GFP to make it usable to more specialized applications. VENUS is a new YFP variant with improved maturation and brightness properties, as well as a reduced environmental dependence [11]. CD63 is a leucocyte surface glycoprotein that is known to occur in membranes [12]. The CD63 gene was fusioned with the VENUS gene, so it can be used as an efficient LRC marker. Van Gogh-Like 2 (VANGL2) is a planar cell polarity protein that is essential for collective migration during embryonic development [13]. It should be evenly distributed in the cell, so it is a candidate for a negative control, since it shows no affinity to LRCs.

After the extraction from the mouse bone marrow the isolated and purified HSCs were cultured and florescence markers like CTB, CD63VENUS or VANGL2VENUS were virually introduced . These with florescence markers treated cell cultures were observed for several hours under a microscopy. The fluorescence markers were excited so the distribution of the markers within the cell can be observed through the florescence emission. The plastic slides with the cell cultures were divided into several positions also known as field of view and pictures in specified time intervals were taken. As a result we gain a time lapsed 'movie' (all pictures consecutively) of the distributions of the fluorescence markers. For our purpose we assume that the binding efficiency of the fluorescence markers is at 100%, although we known that such a value can not be achieved.

1.4 Motivation

1.4.1 Existing approach

The availability of technique like the time lapsed microscopy is a huge progress in HSC research, but produces a lot of data that can not be handled with manual inspection alone. Biologist rely more and more on automated data analysis tools to facilitate their research and manage the enormous amount of data. An automated prediction of LRCs would help biologists to prioritize their targets. Up to our knowledge there is only one existing approach to address this problem so far. In the paper of Vannini et al. [14] they presented a semi-automated image analysis tool to map the presence or absence of lipid raft clusters (LRC) in live hematopoietic stem cells (HSC's) cultured for one hour in serum-free medium supplemented with stem cell factor. They screened the ability of 19 protein candidates to alter lipid raft dynamics and identified six factors that induced either a marked decrease (Wnt5a, Wnt3a and Osteopontin) or increase (IL3, IL6 and VEGF) in LRC. They also hypothesized that the distribution of lipid rafts on idividual HSC's could be used as a very early generic 'reporter' of the cellular state and could indicate whether selected extrinsic signals instruct HSC quiescence or activation. A general work flow of the vannini paper is presented in Figure 1.1. They performed several experiments where they treated cells with different stem cell factors and fluorescent markers. They observe the cells over an hour under a microscope that can measure the fluorescent glow. As a result they gained images of these cells. In order to predict LRC's, they used an approach, that uses the k-means algorithm for clustering the images. They perform their clustering on the pixel intensities of the cell. The pixel intensity can be seen as the photon count at a spot that was measured in a microscopy experiment. After the clustering they gain two clusters of different size. For each cluster they calculate two features. For their first feature they fitted a major and a minor axes through the cell and the brighter cluster and use he difference of these axes. They called it delta chords analysis and is shown in Figure 1.1. When the difference is below a certain threshold they predicted the cell to have no LRC. But if it exceeds the threshold, they used the difference of the mean intensity of the complete cell to the mean intensity of the smaller cluster as their second prediction feature. An arbitrary threshold is used for the delta chords analysis and the intensity difference to determine the prediction. The problem with an arbitrary threshold is that it is not applicative for different data sets since the overall brightness and noise within the images of different data sets can vary heavily. In this work we reconstructed this approach and developed several improvements.

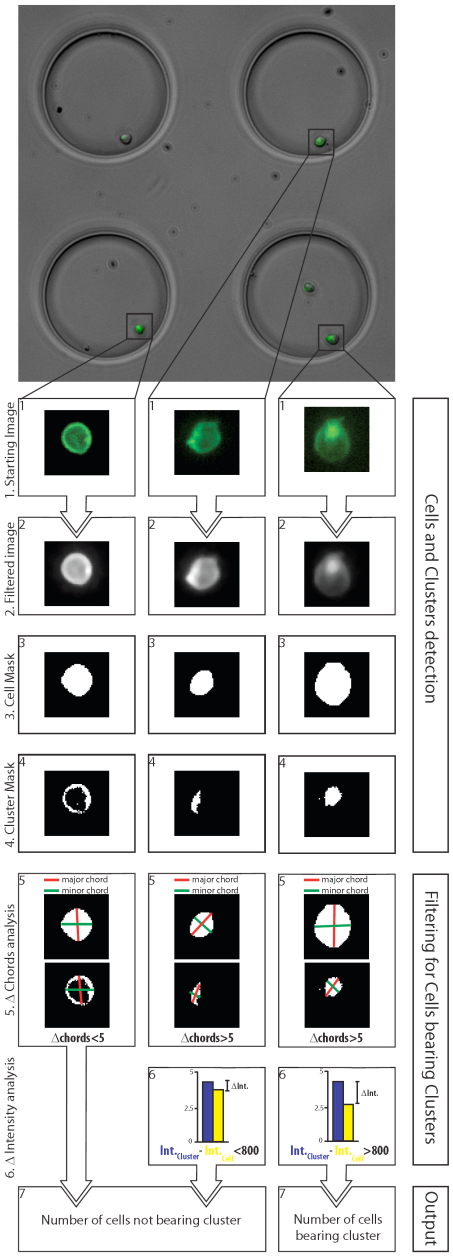


Figure 1.1: Image analysis work flow of the existing approach by Vannini et al. to detect LRC in live cells. A mask for cell detection is created on a thresholded image. Subsequently they used the kmeans clustering in order to identify LRC. LRCs can be identified by bright spots within the cell mask. Two parallel processes based on pixel intensity and cluster sizes are used to evaluate cells with or without clusters. Image taken from [14]

1.4.2 Aim of this work

The field of LRC research is relatively new and therefore only little is known about their effects on the cellular level. It turned out in several publications that the purification of HSCs and analysis of LRCs is not a trivial task and still needs to be improved. Vannini et al. presented a semi automated image analysis tool to predict the occurrence of LRC in time lapsed microscopy data. They used the kmeans clustering method to cluster the intensities within the cell image. They used a chain of fixed threshold cutoffs on cluster size difference and intensity difference to determine their prediction. In this work we want to reproduce and evaluate their approach. We want to go a step further by applying different clustering and prediction procedures in order to improve the existing approach. The group of Timm Schroeder (Research Unit Stem Cell Dynamics) provides us with time lapse microscopy HSC experiments under different conditions. To test and train our methods we needed to generate a reliable data set. We developed an annotation tool that helps us to create a gold standard that is needed for the evaluation of our methods. We will use clustering methods on image data to detect possible LRCs. For the prediction of LRCs we will use a machine learning technique that we train on a small subset of our data set. The goal of this work is to evaluate the method used by the existing approach, compare it to our methods and provide a basis for the development of an easy manageable pipeline and tool for an automated LRC prediction. With this work we hope to improve the identification of LRCs in large scale experiments and thus help to understand more about the formation of LRCs and their effects on the differentiation process of HSCs.

Chapter 2

Materials & Methods

2.1 Data

We got 3 different data sets form the experiments 111012DL6(exp1), 120810DL2(exp2) and 120820DL2(exp3). In these experiments hematopoietic long term stem cell cultures were treated with a fluorescent marker and were observed for several hours under a microscope that is able to measure the intensity of the fluorescent glow. The microscope took in regular intervals pictures of the cell culture. These pictures are the data images we used in this work. All experiments were performed by Dirk Löffler.

The exp1 contains cells that were treated with TGF-beta, a cytokine signal molecule that tend to form LRC's. It contains 84 positions and were measured over 4111 timepoints. In this case positions refer to snapshot locations also known as field of view on the experiment. Such a position snapshot is shown in Figure 2.1. The exp2 experiment consists of 21 positions and were measured over 4238 timepoints at different wave lengths. By considering that one image has the size of 500kB - 900kB, the whole experiment has a size of approximately 120GB. The positions 1-12 were treated with the fluorescence marker HSC-Vangl2VENUS and at the positions 13-21 with the fluorescence marker HSC-CD63VENUS.
Since HSC-Vangl2VENUS should be distributed in the whole cell, it should be a negative control with no visible LRC's, whereas the positions treated with CD63 should show LRC's. The exp3 provided us a reliable negative control. It was treated with an unregulated fluorescent marker that, so it should be evenly distributed in the cell. This experiment worked well, since nearly no LRC is present in the data. It consists of 21 positions measured over 2812 timepoints.

During the work 2 different data structures of the cell movies were provided. These cell movies contain images of cells that were treated with a florescent marker. On the one hand we got data where the cells are completely tracked and quantified, but on the other hand we got a data structure that provides no cell tracking. The tracking was manually created by us and Dirk Löffler at the research unit stem cell dynamics using TTT [16]. It is also possible to track the cells automatically [17] with the TTT program but for high reliable data, a manual creation is unavoidable. Although tracking is a challenging task, it provides the option of creating lineage trees that are necessary for the analysis of asymmetric cell division. An example of a linage tree is illustrated in Figure 2.2. Cells can be quantified with the snapshot QTFy program. It is included in the QTFy program [15] developed by Michael Schwarzfischer but is also available as a stand alone version. With the QTFy program it is possible to navigate through the experiments structure and display several time-resolved cell features of quantified cells. The snapshot QTFy program can normalizes the background images of the the position image (Figure 2.1) and identifies cells automatically in order to extract small

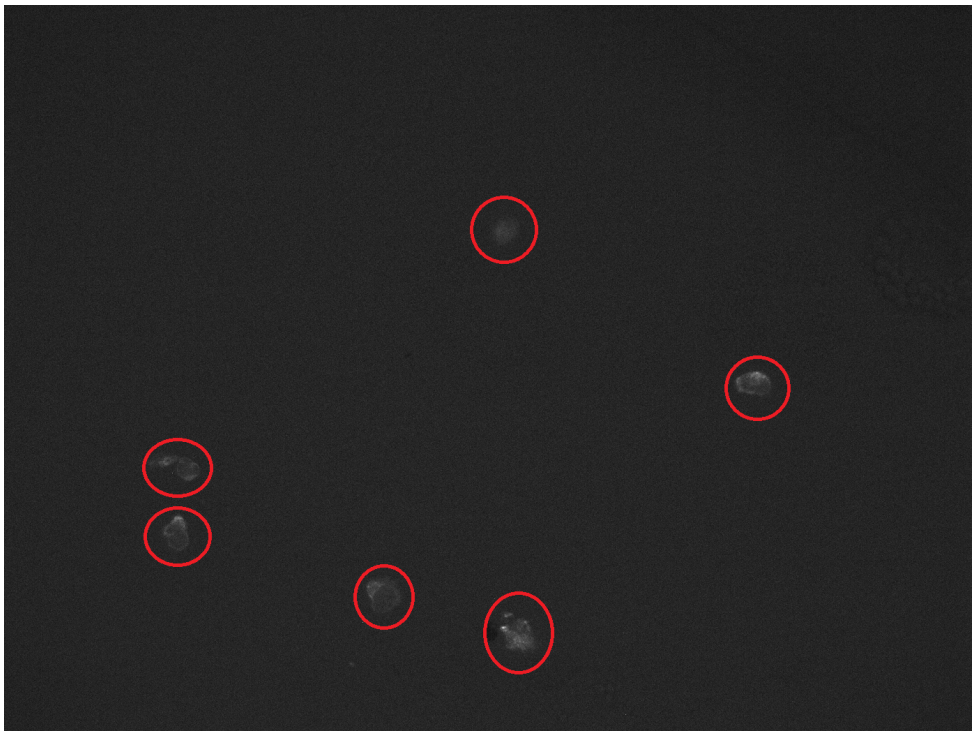


Figure 2.1: Image of a field of view (FOV) of the exp2. In order to highlight the cells they were marked with red circles. The FOV contains cells with and without LRC's. These images are the input for the QTFy(quantification) and TTT(tracking) program. These images need to get background normalized to gain compareable, since they are brighter in the center and darker at the edges [15].

images of the selected cells. A manual selection is possible as well and provides more reliable data. In the best case the small images show only the cell of interest but often cells are very close to each other especially after a cell division (compare Figure 2.3a). The snapshot QTFy program uses the matlabs internal function 'graythresh' a global image thresholding algorithm called Otsu's method [18] to create a cell mask. A cell mask is a binary matrix of the images size consisting of ones where the cell is and zeros where it is not. Thus it is possible to extract only the pixels that belong to the cell from the image. A manually inspection and parameter tuning for the cell mask creating algorithm is essential since a precise and correct cell mask is key for later predictions. If the algorithm can not find the right cell mask it is possible to draw it free handed. Very close cells can cause trouble during the quantification and annotation step, since it is hard for the user and the cellmaks algorithm to determine which cell should be annotated or quantified.

2.2 Annotation Tool

One task to evaluate our prediction results, was to create a gold standard containing the exact amount of spots for every cell. Therefore we developed an annotation tool for stem cells named "CAT"(cluster annotation tool). This program can load and visualize a stem cell experiment. The CAT can be used for annotating tracked cells and cells with no tracking. Figure 2.4 shows a screenshot of the GUI of CAT. The right window and the window in the middle are for navigating through the experiments structure. The left window is for visualization of the selected cells image. It is possible to annotate the cell using keyboard shortcuts on the num pad. This allows a fast annotation of many cells. By pressing any button on the num pad, the

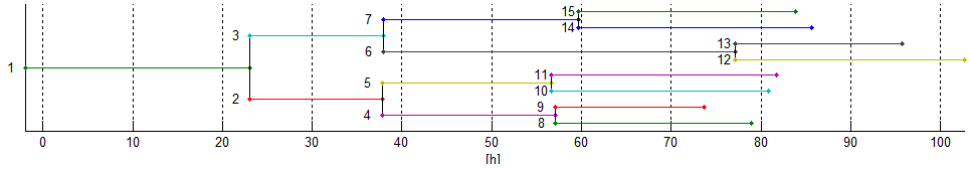
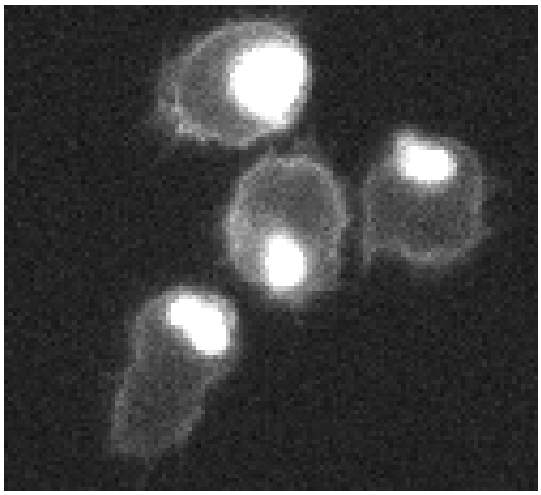
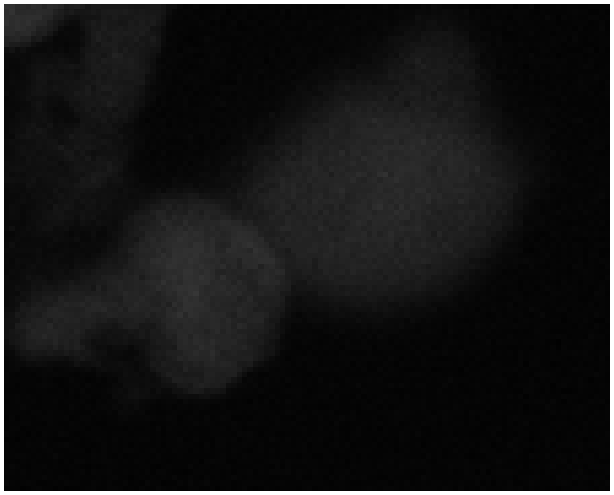


Figure 2.2: This Figure illustrates an example of a linage tree. It is a time resolved tree of the precursor cell one (green). The x axis represents the time in hours. Several 'pictures' of the cell were taken in specified time intervals and were combined to this linage tree. Thereby one image correspond to one time point on the x axis. In hour 23 the first precursor cell divides into two daughter cells two and three.



(a) A small image extraction from exp2. The LRC's are clearly visible.



(b) A small image extraction from exp3 (negative control). A LRC can not be observed in this image, since the fluorescent marker is evenly distributed within the cell. That does not imply that the cell does not have a LRC.

Figure 2.3: Small extracts from the raw data of cells. Figure A shows cells with a clearly visible LRC. In cells of Figure B no LRC can be observed, since the fluorescent marker is evenly distributed within the cell. In both Figures the cells are very close to each other, what causes trouble for the user to determine the cell that should be annotated or quantified.

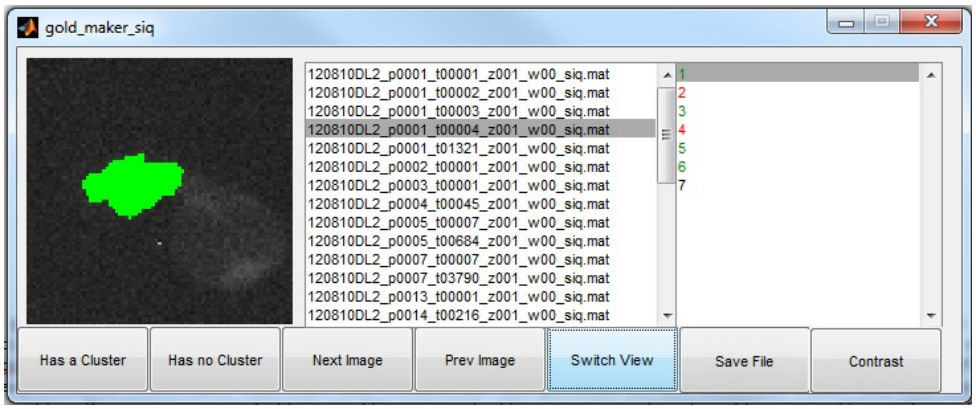


Figure 2.4: Snapshot of the 'cluster annotation tool'. The middle and right windows are for navigating through the experiments structure. The left window is for displaying the selected cell. The 'switch view' button is activated for a better identification of the selected cell. Cells containing at least one spot appear in green color in the right box. Cells without a LRC appear in red color.

program automatically saves the number of spots and the cell ID at a list and visualize the next cell. It also highlights the cell in the list with green color for one or more spots and red color for no spot. With the 'switch view' button you can highlight the cell mask in the image as shown in Figure 2.4. This is useful for images (Figure 2.3a)where cells are closely together and its unclear which cell should be annotated. With the "contrast" option it is possible to adjust the contrast of the image for a better visual identification of the LRC's. The annotation list can be saved with the "save" button or by closing the tool, it automatically ask whether it should save the list.

Michael Schwarzfischer and i annotated independently from each other the set of cells we introduced in 2.1 with the annotation tool. A comparison of the two annotations showed that 81.45% were annotated equally and 18.55% were annotated differently. This shows that even the manual annotation is a difficult task, since it is often unclear whether the cell contains a LRC and thus a 100% correct annotation is not possible. Of course a change in the annotation has a huge influence on the prediction outcome and evaluation. In this work we used my annotation for further processes. The resulting annotation list is shown in Figure 2.5 and represents our gold standard. It is the basis for every evaluation in this work.

2.3 Clustering Methods

2.3.1 k-means

The k-means algorithm is an iterative procedure that splits a set of data points into k clusters. The parameter k must be set before the algorithm starts. In the first step, the algorithm defines k randomly distributed centers and assign every data point in the set to its nearest center. Therefore it uses the squared euclidean distance measure. The most commonly used kmeans algorithm is the Lloyd algorithm that tries to find a global minimum of the function $\sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$, where $x_1 ... x_n$ are the data points that were split into k disjunctive sets. μ_i is the center that is assigned to cluster C_i [19]. Based on that assignment the algorithm recalculates the centers and continues the procedure iteratively until either the maximum number of iterations has been reached or the centers do not alter anymore [20]. The k-means algorithm is a secure method to gain guarantied k clusters.

	1	2	3
1	'Pic'	'Center_Nr'	'Is_Cluster'
2	'120810DL2_p0001_t00001_z001_w00_siq.mat'	7	0
3	'120810DL2_p0001_t00001_z001_w00_siq.mat'	1	1
4	'120810DL2_p0001_t00001_z001_w00_siq.mat'	2	1
5	'120810DL2_p0001_t00001_z001_w00_siq.mat'	3	1
6	'120810DL2_p0001_t00001_z001_w00_siq.mat'	4	1
7	'120810DL2_p0001_t00001_z001_w00_siq.mat'	5	0
8	'120810DL2_p0001_t00001_z001_w00_siq.mat'	6	1
9	'120810DL2_p0001_t00002_z001_w00_siq.mat'	1	1
10	'120810DL2_p0001_t00002_z001_w00_siq.mat'	2	1
11	'120810DL2_p0001_t00002_z001_w00_siq.mat'	3	1
12	'120810DL2_p0001_t00002_z001_w00_siq.mat'	4	1
13	'120810DL2_p0001_t00002_z001_w00_siq.mat'	5	0
14	'120810DL2_p0001_t00002_z001_w00_siq.mat'	6	1
15	'120810DL2_p0001_t00002_z001_w00_siq.mat'	7	1
16	'120810DL2_p0001_t00003_z001_w00_siq.mat'	1	1
17	'120810DL2_p0001_t00003_z001_w00_siq.mat'	2	1
18	'120810DL2_p0001_t00003_z001_w00_siq.mat'	3	1
19	'120810DL2_p0001_t00003_z001_w00_siq.mat'	4	1
20	'120810DL2_p0001_t00003_z001_w00_siq.mat'	5	0
21	'120810DL2_p0001_t00003_z001_w00_siq.mat'	6	1
22	'120810DL2_p0001_t00003_z001_w00_siq.mat'	7	1
23	'120810DL2_p0001_t00004_z001_w00_siq.mat'	1	1
24	'120810DL2_p0001_t00004_z001_w00_siq.mat'	2	1
25	'120810DL2_p0001_t00004_z001_w00_siq.mat'	3	1
26	'120810DL2_p0001_t00004_z001_w00_siq.mat'	4	1
27	'120810DL2_p0001_t00004_z001_w00_siq.mat'	5	0
28	'120810DL2_p0001_t00004_z001_w00_siq.mat'	6	1

Figure 2.5: This Figure shows the output of the 'cluster annotation tool'. It contains the location of the quantification data structure, the CellID and the annotated amount of clusters. This list represents our 'gold standard' an is used in the prediction pipeline for the evaluation of the different prediction methods.

2.3.2 Gaussian mixture model (GMM)

A Gaussian mixture model [21] is a multivariate distribution that consists of a mixture of one or more multivariate Gaussian distribution components. Typically a mixture model consists of k independent distributions. In case of a GMM all distribution need to be Gaussian distributions. Each distribution has its own weight value. These weight values are probabilities and sum up to one. They were also known as the mixing proportions. The overall GMM consists of a linear combination out of every Gaussian distribution multiplied with its mixing proportion. By using a GMM on pixel intensities, it provides for every pixel the probability to be assigned to a cluster. Our intention for using GMM's was to gain the best possible clustering for the cell. The GMM will provide us the cluster count that represents the data in the best way. we hoped by using the GMM for clustering that in the best case this would make a prediction redundant, since a cell with no spot should return only one cluster.

2.3.3 BIC (Bayesian information criterion)

The GM distribution class provides a BIC(Bayesian information criterion) [22] score for fitted data. The BIC score is calculated by -2*ln(L)+m*log(n), where n is the number of observations, m is the number of estimated parameters and L represents the maximized value of the likelihood function for the estimated model. Since the BIC score punishes additional parameters, the best BIC score would be the the lowest. It provides the best representation of the data without having the disadvantage of over fitting.

2.3.4 Spatial scan statistic

In the basian scan statistic paper from Neill et al. [23] they used a method for finding spatial regions where some quantity is significant higher than expected. They use this method to detect clusters of disease cases. They compared the null hypothesis H0 of no clusters to the set of alternative hypothesis H1(S), each representing a cluster on some region S.

Their method seemed suitable for finding bright spots in a stem cell image, so we modified some parameters to match our problem. They proposed a method that relies on previously observed data D. It is assumed that the priors representing the observed data is Gamma distributed. Unfortunately the direct way to calculate the possibility of observing a LRC at location S given the data ($P(H1(S)|D)$)is not a trivial task. In order to solve this problem Neill et al. presented the equations 1 - 7. They presented a way to approximate $P(D|H0)$ equation(2) and $P(D|H1(S))$ equation(3), where D is the given data. In this case the given data are the intensities, x and y coordinates of the pixels in the cell image. Since it is possible to calculate $P(D|H0)$ and $P(D|H1(S))$, the bayes' rule can be used to gain $P(H0|D)$ and $P(H1(S)D)$. The probability of the data can be calculated with equation(1). Alpha and beta are priors were "in" means inside the region S, "out" means outside region S and "all" means all data points. With the help of equation(4)and(5) it is possible to calculate the the priors alpha and beta with the equation (6) and (7), where C is the number of intensities greater than a threshold in the specified area and B is the number of pixels in the area. C/B was the original disease rate and was drawn from a gamma distribution.

We calculated the expected value of the cell (E_{all}) of C/B by the sum of intensities greater than a threshold divided by the number of intensities of the cell. The mean size of cells in exp2 is 1023 pixels. Depending on the threshold the amount of intensities greater than the threshold can vary. Assumed that 20% of the intensities is above the threshold ($C_{all} = 204.4$)that would lead to an expected value of 0.2. As the variance of C/B we decided to take the variance of the intensities. The mean variance of the intensities of exp2 is 0.0183. With the help of equation (6) and (7) we can calculate the priors α_{all} ($(0.2^2)/0.0183 = 2.185$) and β_{all} ($0.2/0.0183 = 10.928$). In order to calculate $P(D|H0)$ we need to calculate $\Gamma(\alpha_{all} + C_{all}) = \Gamma(2.185 + 204.4)$. Since C-in/C-out/C-all are high values, the Gamma function produces very high values, resulting in

infinity since it is not possible for our machines to represent such high values. The approach to use the loggammadistribution also resulted in numerical problems.

$$P(D) = P(D|H0) * P(H0) + \sum_{S}^{allregionsS} P(D|H1(S))P(H1(S)) \tag{2.1}$$

$$P(D|H0) \propto \frac{(\beta_{all})^{\alpha_{all}} * \Gamma(\alpha_{all} + C_{all})}{(\beta_{all} + B_{all})^{\alpha_{all}+C_{all}} * \Gamma(\alpha_{all})} \tag{2.2}$$

$$P(D|H1) \propto \frac{(\beta_{in})^{\alpha_{in}} * \Gamma(\alpha_{in} + C_{in})}{(\beta_{in} + B_{in})^{\alpha_{in}+C_{in}} * \Gamma(\alpha_{in})} X \frac{(\beta_{out})^{\alpha_{out}} * \Gamma(\alpha_{out} + C_{out})}{(\beta_{out} + B_{out})^{\alpha_{out}+C_{out}} * \Gamma(\alpha_{out})} \tag{2.3}$$

$$\frac{\alpha_{all}}{\beta_{all}} = E_{sample}[\frac{C_{all}}{B_{all}}] \tag{2.4}$$

$$\frac{\alpha_{all}}{\beta_{all}^2} = Var_{sample}[\frac{C_{all}}{B_{all}}] \tag{2.5}$$

$$\alpha_{all} = \frac{(E_{sample}[\frac{C_{all}}{B_{all}}])^2}{Var_{sample}[\frac{C_{all}}{B_{all}}]} \tag{2.6}$$

$$\beta_{all} = \frac{E_{sample}[\frac{C_{all}}{B_{all}}]}{Var_{sample}[\frac{C_{all}}{B_{all}}]} \tag{2.7}$$

2.4 Decision trees

A decision tree is a directed hierarchic tree structure with labeled decision nodes. It is used for an automated classification of data. The leafs of the tree represents the desired classification. Each internal node, as well as the root represent a binary decision rule. In our case the rules are a binary decision, whether a variable is bigger or smaller then a threshold that is learned during the creation step. Decision trees are a supervised machine learning method that takes matrix of predictor values and a vector of classes as input. The resulting decision tree represents the best way of splitting the predictor values to represent the classification. To gain a classification, you need to walk the tree using the decision rules at each node until you have reached a leaf. This leaf represents is the best classification that can be achieved with the available data. An example of a decision tree is shown in Figure 2.6.

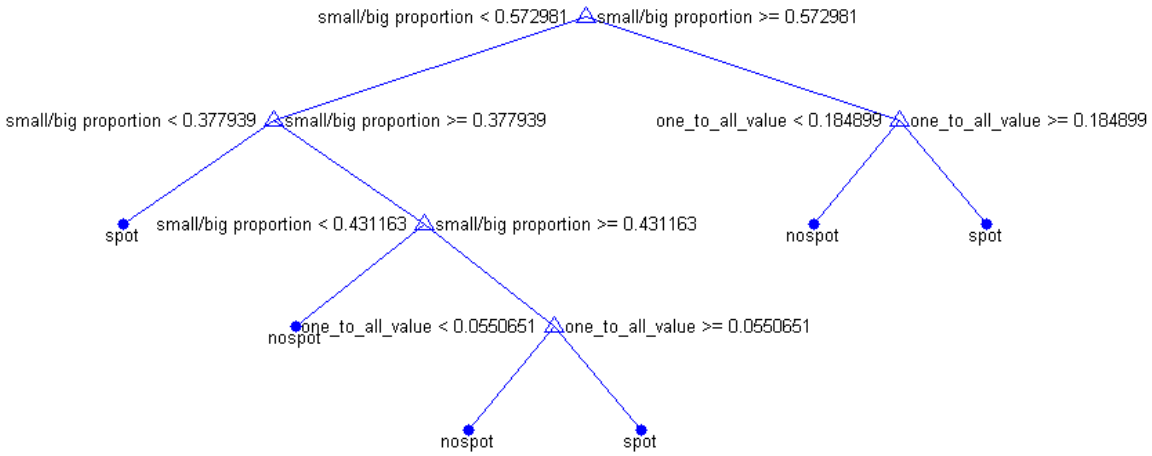


Figure 2.6: A decision tree created out of a small subset of cells after a clustering with kmeans, k=2 on intensities only. The features that were used to create this tree are the 'one-to-all-value' and 'small/big proportion'. In this case the 'one-to-all-value' is the greater value of the mean intensity of both clusters. The 'small/big porportion' feature is the cluster area of the smaller cluster divided through the cluster area of the bigger cluster. The decision tree represents the best way to split the measured features to classify the input data. The knowledge of the classification were gained from the 'gold standard' that can be created with the cluster annotation tool we presented in section 2.2.

Chapter 3

Results & Application

3.1 Data overview

We created 383 cell masks from the movie data of exp2 and exp3 manually with the QTFy program [15] to gain a high accurate and reliable set of cells. Unfortunately these cells have no tracking. The set consists of 383 cells, where 290 cells are from exp2 and the 93 cells from exp3. Figure 3.1 illustrates the distribution of cells containing and missing a LRC. It shows that 159 ells have no spot (42%) and 224 cells have at least one spot or more (58%).

We analyzed a control sample out of exp2 of about 140 cells treated with HSC Vangl2Venus and 150 cells treated with HSC CD63Venus. The cells were chosen randomly. The Figure 3.2a shows that 74% of the cells treated with Vangl2 formed LRC's and as shown in Figure 3.2b 77% of the cells treated with CD63 formed LRC's. The experiment provided us a useful positive sample, since the LRC's are clearly visible as illustrated in Figure 2.3a. The exp3 provided us a useful negative example. We analysed a randomly chosen sample of 93 cells, where 90 cells showed no sign of a LRC. Only 3 cells showed signs to have a LRC.

3.2 Pipeline

3.2.1 Data gathering

The work flow of the data gathering step is illustrated in Figure 3.3. We started our process with the data from the three experiments(exp1, exp2, exp3) we introduced in section 2.1. These experiments consists of time resolved images of hematopoietic stem cells. A very important step that need to be done is the quantification of cells that builds a basic set. Therefore we use the

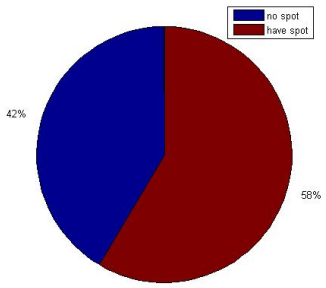
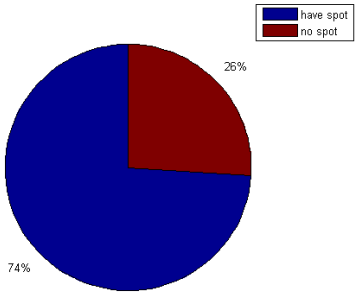
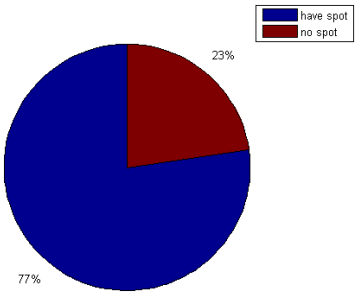


Figure 3.1: The Figure shows the distribution of cells containing and missing of LRC of our data set. The data set contains 383 cells from exp2(290 cells) and exp3(93 cells). 159 of the cells have no spot (42%) and 224 cells have at least one spot or more (58%).



(a) Distribution of LRC's of cells treated with HSC Vangl2Venus. The sample consists of 140 randomly chosen cells. 74% of the cells showed a sign of containing at least one spot or more. Since the cells treated with HSC Vangl2Venus were initially intended to be our negative control, but 74% cells conatined a cluster, the experiment failed. However the LRC were clearly visible and provided us a good positive sample.



(b) Distribution of LRC's of cells treated with HSC CD63Venus. The sample consists of 150 randomly chosen cells. 77% of the cells showed a sign of containing at least one spot or more.

Figure 3.2: Distribution of LRC's from quantified cells of exp2. The cells were treated with HSC Vangl2Venus (Figure a) and HSC CD63Venus (Figure b).

QTFy program to extract the cell images and create a cell mask for each cell we were interested in (Figure 3.3 (2)). The QTFy program is able to normalize the background images. This is necessary to make the data comparable, since the data images show an uneven illumination. Cells in the center of the image glow brighter than cells at the border. As a side effect we also gain a better contrast of the cell. This is important, since some cells are only visible under a certain contrast. It also makes the identification of spots much easier. In order to gain high quality data we segmented and quantified the cells manually. As a result we gain a structure containing the image, the position of the cell in the image and the cell masks. We create a list of each cell and the location on the hard drive (Figure 3.3 (4)). This list will be the input for later tools. This data list will be extended in later steps and will become the basis of our prediction. The data list defines the order in which the cells were clustered and predicted. The order in which the cells were clustered is irrelevant, but by sorting them we can achieve a small performance boost. Loading all images before the start of the algorithms would lead to a huge improvement in run time, however we have way to many images to load them all into the ram. Nonetheless some boost in speed issues could be made, since an image can include more then one cell and only need to be loaded one time.

A cell tracking with the TTT program for the creation of a cell linage tree can be applied but is not necessary for further steps of the prediction (Figure 3.3 (1)). Nevertheless it is necessary for the analysis of asymmetric cell division.

To be able to perform any validation we need to know the exact amount of spots for every cell. Therefore we use the 'CAT'(for more details about CAT see 2.2) to annotate the earlier quantified cells manually (Figure 3.3 (3)). After annotating the cells we join the knowledge of the spot count with the data list we gained earlier.

Before the actual clustering can happen we need to extract a smaller image of the size of its corresponding cell mask. During the quantification we got the positions of the cell in the image. These positions can be used for the extraction (Figure 3.4 (1)).

The method of clustering can be chosen by setting a parameter. The options are a clustering by k-means or with the GMM method. Also the maximum amount of clusters the clustering

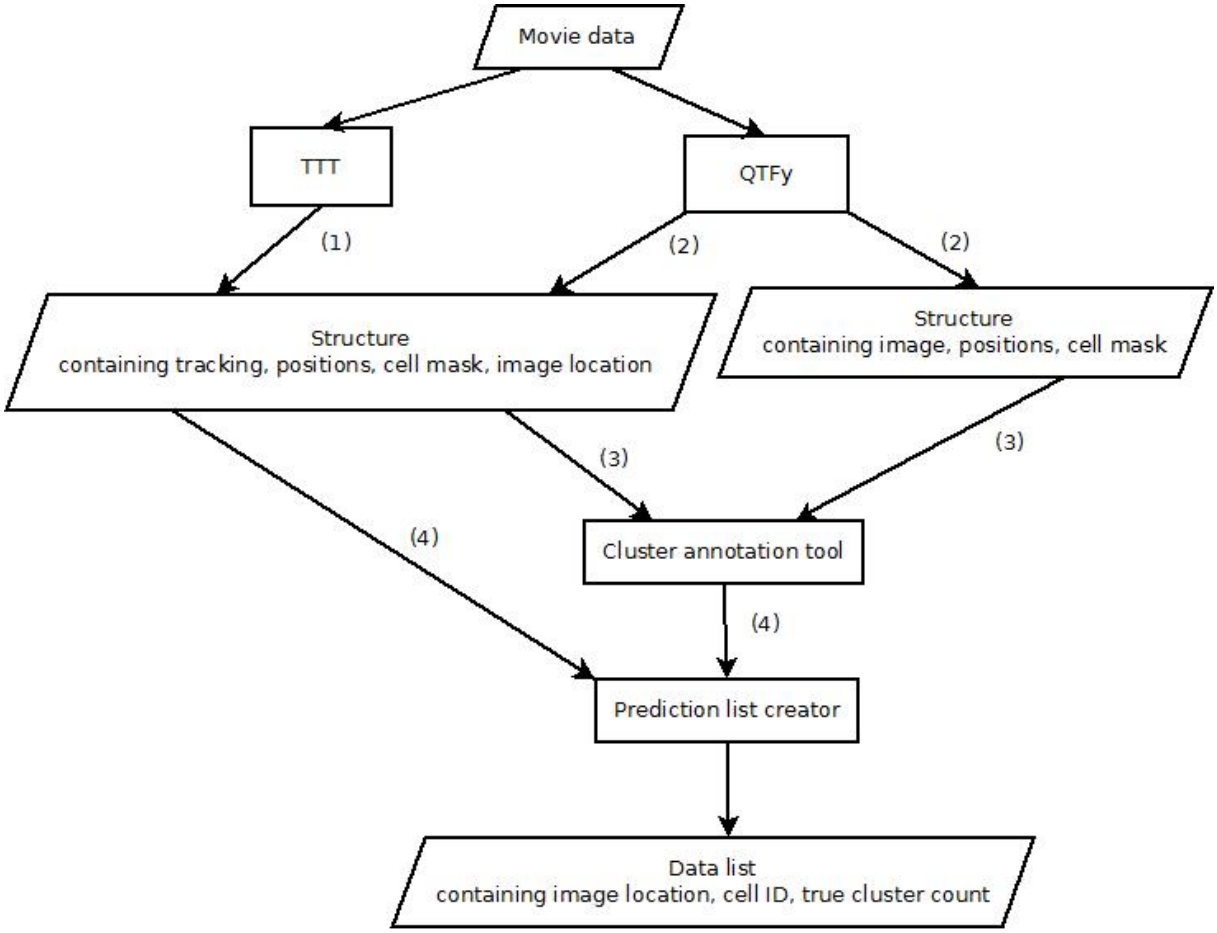


Figure 3.3: This Figure shows the workflow of the data gathering process. The raw movie data can be tracked with the TTT program (1) and must be quantified with the QTFy program (2). The tracking as well as the quantification must be done for each cell image separately. One cell image represents one time point. The resulting data can be used as input for the 'cluster annotation tool' (CAT) program. CAT can be used to perform a manuall annotation (3) to gain a 'gold standard', that is required for the validation process. The result of the data gathering process (4) is a data list containing the Cell ID, its data structure location and the true amount of LRC the cell contains. This data list is the input for the prediction pipeline.

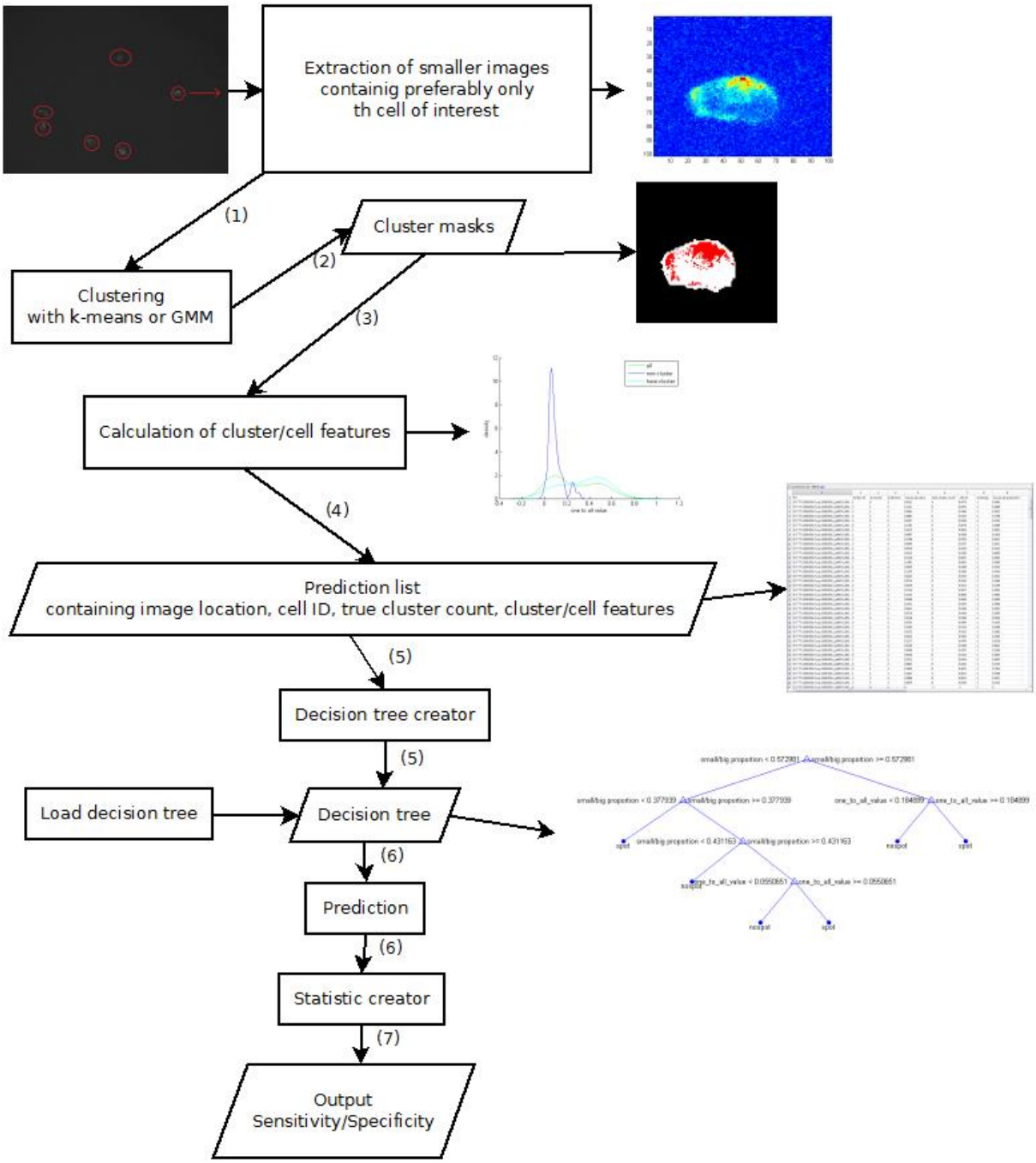


Figure 3.4: This Figure shows the workflow of the prediction pipeline. After the extraction of a smaller image (1) of the cell, we gain its cell mask. With these cell masks the we can perform a clustering with either the kmeans or the GMM method (2). The clustering can be done on either intensities or a combination of x-y coordinates and the intensity. The resulting cluster masks were used for calculation of cell and cluster features (3). These features were saved in a prediction list (4), an extension of the data list we gained from the data gathering process (see Figure 3.3). Out of the features and the 'gold standard' of a small subset it is possible to generate a decision tree (5). This tree is used for the prediction of LRCs (6). The sensitivity and specificity is provided as an output of the prediction pipeline (7).

method should search for can be set as a parameter. The features that should be used in the prediction can be set as well. With the help of the cell mask we can extract the intensities and their x and y coordinates from the image and the clustering can be performed (Figure 3.4 (2)). Depending on the cluster method this step can take different amount of time. The GMM method tend to take more time for calculation, caused through the multiple clustering of different spot counts.

3.3 k-means clustering

We applied the k-means algorithm to the set of cells we introduced in section 2.1. The kmeans algorithm needs a previously defined amount of clusters, parameter k. The algorithm splits the data (either intensities only or intensities combined with their x and y coordinates) into k disjunctive sets. These sets were our cluster masks we used in later steps for the calculation of cluster features.

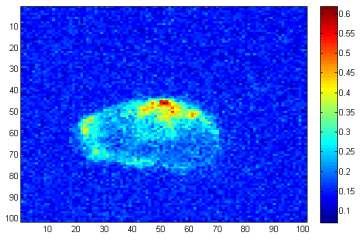
Figure 3.5c shows one example where the clustering fails. It identifies the pink cluster as one cluster, but the original image Figure 3.5a shows that the cell has two different spots. The cell was clustered with k-means, where k=2 and the clustering was performed on intensities only. Since the cell has two spots that are much brighter than the rest of the cell they were clustered together falsely. The x and y coordinates of the pixels were taken into account in order to avoid that false clustering. The Figures 3.5d show that this extension does not solve the issue of false clustering. Since the cell has two LRC's it is impossible to find all three clusters with a k-means with k=2. The solution was to run the k-means with k=3 and illustrated in Figure 3.5e it came close to a correct identification of all clusters. Nevertheless the method has its drawback since you do not know for how many spots we are looking for.

3.4 GMM clustering

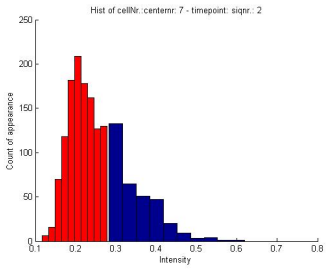
We modeled the data consisting of the intensities and the x and y coordinates of the pixels to a GMM. As a result we gain a 3-D mixture model of out data. In our approach we used the matlabs internal gmdistribution class for clustering. With this class it is able to fit a GM distribution onto given data. The fitting procedure needs the amount of clusters it should search for. We run the gmdistribution fitting procedure multiple times with different cluster counts. We fitted all cluster counts from one to an upper limit of allowed clusters. The upper limit of the cluster count is a predefined parameter. Of course for high numbers of maximum allowed clusters the calculation can take time and is not practical for a fast automated prediction pipeline. We observed the best results with 8 maximum allowed clusters. For each run the BIC(Bayesian information criterion see 2.3.3) score were memorized. We identified the best cluster count based on this criteria.

3.5 Calculation of cell and cluster features

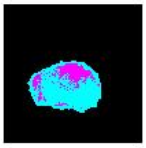
Since the clustering alone does not solve determination of conatining a LRC, we defined some cell and cluster features for the prediction. They were measured after the clustering and are the basis of our prediction. After the clustering of a cell we gain the cluster masks. With these cluster masks and the cell masks, we got from the quantification, we can calculate several clusters and cell features for each cell (Figure 3.4 (3)). We extend the data list with the measured freatures to the prediciton list that we use for prediction (Figure 3.4 (4)). A small extract of a feature list is shown in Figure 3.7. We created distributions for each feature of the cells containing a LRC and those which do not. With these plots it is possible to see if it is a useful feature for prediction. For a meaningful feature the distributions should not overlap.



(a) Image of the original cell. This cell is from exp2 an has two LRC's. One LRC is located at the top of the cell and the second LRC is at left edge of the cell.



(b) Histogram of the intensities. A clustering was performed with kmeans, k=2 on intensities only. The first cluster contains the low intensity compartment and is in red color. The second cluster contains the high intensity compartment and is in blue color.

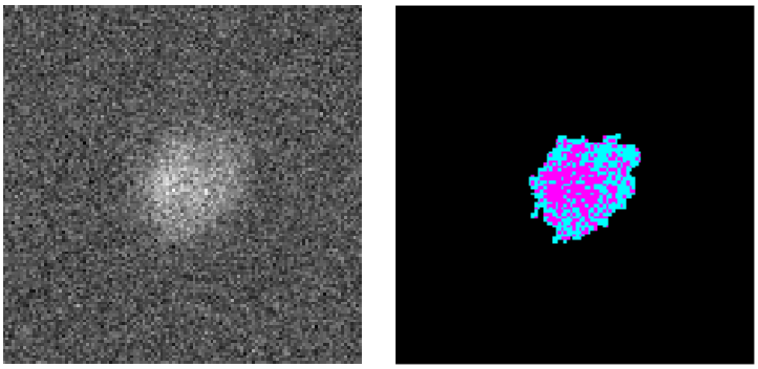


(c) This Figure shows two clusters that were gained after a clustering with kmeans, k=2 on intensities only. The kmeans clustering method can not distinguish between the two LRC's, since itseareches for only two clusters. Both LRC's were approximately equaly bright, so the kmeans clustering identify them as one cluster.

(d) This Figure shows two clusters that were gained with a clustering with kmeans, k=2 on x-y-intensity data. The resulting clusters does not match the acutal LRC's and deliver unfeasible data.

(e) This Figure shows three clusters that were gained after a clustering with kmeans, k=3 on x-y-intensity data. All three clusters were correctly identified with the knowledge of how many spots the cell has. Unfortunatelly this prior knowledge is unknown during the prediction step.

Figure 3.5: Positive example from exp2 of a stem cell containing two LRC's. It also shows the problem of clustering with kmeans. A clustering with kmeans, k=2 is not able to identify two LRC's, since three clusters were needed.



(a) Image of the original cell. This cell is from exp2 and does not have a LRC.

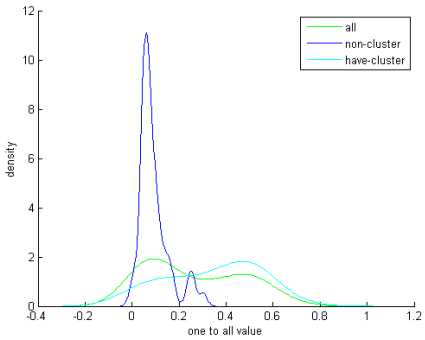
(b) This Figure shows the two clusters that were gained with a clustering with kmeans, k=2. The clustering appears disordered, since the cell has no bright spot that could indicate a LRC.

Figure 3.6: Negative example from exp3. Figure (a) shows a stem cell where no LRC can be observed. Figure (b) shows the result of the clustering process. The resulting clusters were incoherent, indicating an evenly distribution of the florescence maker within the cell.

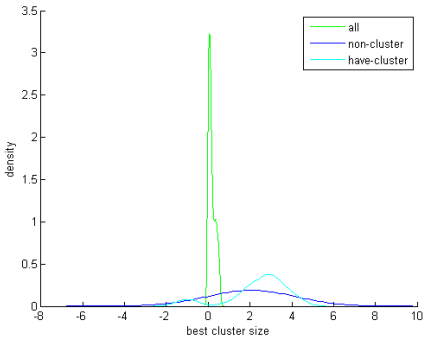
prediction_list <384x9 cell>

	1	2	3	4	5	6	7	8	9
1	'Pic'	'Center_Nr'	'Is_Cluster'	'prediction'	'one_to_all_value'	'best_cluster_count'	'cell_sd'	'small_big'	'one_to_all_proportion'
2	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	1	0	-1	0.0522	4	0.0473	-1	0.2065
3	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	2	1	-1	0.1311	6	0.0591	-1	0.4889
4	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	3	1	-1	0.0661	7	0.0366	-1	0.2913
5	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	4	1	-1	0.0883	6	0.0415	-1	0.3708
6	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	5	0	-1	0.0257	5	0.0295	-1	0.1233
7	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	6	1	-1	0.1292	8	0.0679	-1	0.4089
8	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	7	2	-1	0.1475	8	0.0683	-1	0.5053
9	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	1	1	-1	0.0697	5	0.0568	-1	0.2926
10	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	2	2	-1	0.1769	8	0.0576	-1	0.6127
11	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	3	1	-1	0.0646	4	0.0357	-1	0.2812
12	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	4	1	-1	0.0539	6	0.0410	-1	0.2413
13	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	5	0	-1	0.0143	3	0.0290	-1	0.0690
14	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	6	1	-1	0.1126	8	0.0647	-1	0.3845
15	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	7	2	-1	0.1065	8	0.0691	-1	0.3869
16	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	1	0	-1	0.0836	5	0.0458	-1	0.3335
17	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	2	1	-1	0.1295	8	0.0599	-1	0.4794
18	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	3	1	-1	0.0423	4	0.0353	-1	0.1913
19	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	4	1	-1	0.0714	8	0.0454	-1	0.2966
20	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	5	0	-1	0.0479	5	0.0311	-1	0.2109
21	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	6	1	-1	0.1110	8	0.0627	-1	0.3603
22	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	7	2	-1	0.1143	8	0.0666	-1	0.3964
23	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	1	0	-1	0.0448	3	0.0465	-1	0.1856
24	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	2	1	-1	0.1431	8	0.0593	-1	0.5068
25	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	3	1	-1	0.0441	5	0.0349	-1	0.1957
26	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	4	1	-1	0.0719	8	0.0439	-1	0.3085
27	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	5	0	-1	0.0129	4	0.0289	-1	0.0618
28	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	6	1	-1	0.0393	6	0.0494	-1	0.1455
29	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	7	2	-1	0.1395	7	0.0714	-1	0.4599
30	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	1	1	-1	0.0333	5	0.0262	-1	0.1982
31	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	2	1	-1	0.0424	6	0.0282	-1	0.2363
32	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	3	1	-1	0.1217	7	0.0476	-1	0.5136
33	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	4	0	-1	0.0149	7	0.0308	-1	0.0811
34	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	5	0	-1	0.0489	5	0.0287	-1	0.2569
35	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	6	1	-1	0.0349	6	0.0314	-1	0.1945
36	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	7	1	-1	0.1331	7	0.0543	-1	0.5035
37	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	8	1	-1	0.0803	8	0.0362	-1	0.3745
38	'E:\TTT\120810DL2\sig\120810DL2_p0001\1208...	9	1	-1	0.2004	8	0.0857	-1	0.7020
39	'E:\TTT\120810DL2\sig\120810DL2_p0002\1208...	1	1	-1	0.1815	5	0.0823	-1	0.5569
40	'E:\TTT\120810DL2\sig\120810DL2_p0002\1208...	2	2	-1	0.0996	8	0.0612	-1	0.3671
41	'E:\TTT\120810DL2\sig\120810DL2_p0002\1208...	3	3	-1	0.0879	8	0.0536	-1	0.3103
42	'E:\TTT\120810DL2\sig\120810DL2_p0002\1208...	4	0	-1	0	-1	-1	-1	0

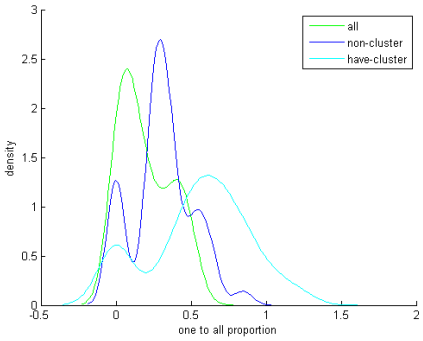
Figure 3.7: Feature List containing the 'one to all value', 'best cluster count', 'cell standard deviation', 'small/big value' and the 'one to all proportion value'.



(a) Distributions of the 'onetoallvalue' for cells with LRC, without LRC and both. The clustering was performed on the big set (exp2 and exp3) with GMM and 4 maximal allowed clusters. The 'onetoallvalue' is calculated by the maximum difference of the mean intensity to all other clusters. The Figure shows that the 'onetoallvalue' can be used as a feature for prediction of LRC's, since the distribution of the 'onetoallvalue' of cells containing a LRC(turquoise distribution) differs from the distribution of cells with no LRC (blue distribution).

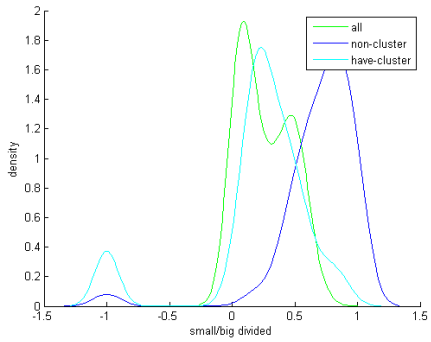


(b) Distributions of the 'best cluster count' for cells with LRC, without LRC and both. The clustering was performed on the big set (exp2 and exp3) with GMM and 4 maximal allowed clusters. It shows that the 'best cluster count' can not be used as a feature for prediction of LRC's.

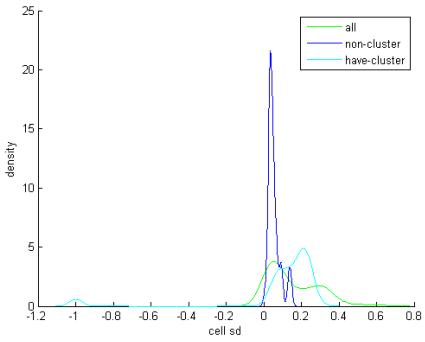


(c) Distributions of the 'one to all proportion value' for cells with LRC, without LRC and both. The clustering was performed on the big set (exp2 and exp3) with GMM and 4 maximal allowed clusters. It shows that the 'one to all proportion value' can be used as a feature for prediction of LRC's.

Figure 3.8: After the clustering we calculated several cluster and cell features. These features were used in later steps to determine our prediction of LRC. These Figures show the distribution of the features of cells where a LRC can be observed, of cells where no LRC can be observed and both. The more the distributions of features with LRC distinguish from those without LRC, the more meaningful the feature gets in regard to LRC prediction.



(a) Distributions of the 'small/big value' for cells with LRC, without LRC and both. The clustering was performed on the big set (exp2 and exp3) with kmeans, k=2. It shows that the 'one to all proportion value' can be used as a feature for prediction of LRC's.



(b) Distributions of the 'cell sd value' for cells with LRC, without LRC and both. The clustering was performed on the big set (exp2 and exp3) with kmeans, k=2. It shows that the 'one to all proportion value' can be used as a feature for prediction of LRC's.

Figure 3.9: After the clustering we calculated several cluster and cell features. These features were used in later steps to determine our prediction of LRC. These Figures show the distribution of the features of cells where a LRC can be observed, of cells where no LRC can be observed and both. The more the distributions of features with LRC distinguish from those without LRC, the more meaningful the feature gets in regard to LRC prediction.

The first feature we defined was the difference of the mean intensity of the whole cell to the mean intensity of the smaller cluster. This feature was proposed by Vannini et al. in their paper that we have introduced in section 1.4.1. As a variation we measured the difference of the mean intensity of the bigger cluster to the mean intensity of the smaller cluster. The problem with these two features is that they are not applicative for different data sets since the data sets we used vary in brightness. To overcome the problem of variation in brightness we used the proportion of the two mean intensities instead of the difference. However these features cant be used for more than two clusters. But since we use the GMM method or the k-means with k >= 2 for clustering in order to find multiple LRC's, these methods can provide more than two clusters. Therefore we developed the 'one to all value' feature. This feature calculates for every single cluster its mean distance to the mean intensity of all other clusters. We kept the order of the subtraction, so negative values can occur. The cluster with the biggest difference to all other clusters should represent the brightest spot in the image. Figure 3.8a shows that is can be used as a feature to predict LRC's. However this feature also has the problem of brightness variation, so we decided to define another feature we called 'one to all proportion'. It uses the same approach as the 'one to all value' but instead of subtraction we used the proportion. In order to gain comparable values the proportions were in the logarithmic space. However as Figure 3.8c shows that it does not improve the prediction compared to the 'one to all value'.

To remodel the approach of Vannini et al. [14] we also fitted the major and minor axes through the cell and the clusters. Therefore we used the matlabs internal function 'regionprops'. With this function we also gain the size of the cells and clusters as well as the eccentricity. We decided to reject the eccentricity property, since a LRC can occur in any shape. A feature we called 'small/big' is the proportion of the cluster sizes. Therefore we divided the smaller cluster through the bigger cluster. This feature is only available for clustering with maximal two allowed clusters. Figure 3.9a shows the distributions of the small/big feature. The distributions of cells containing a LRC and cells not having a LRC differ from each other, so the feature could be used for prediction.

We measured the standard deviation of the cell intensity and used it as a feature. We assumed that the standard deviation of cells containing a LRC should be higher. Figure 3.9b confirms that assumption.

As a result of the GMM clustering we gain the 'best cluster count'. We hoped that we can use this feature for our advantage, but as Figure 3.8b illustrates this property delivered no meaningful prediction.

3.6 Creation and evaluation of decision trees

A prediction of the exact amount of spots is desirable, but it turned out to be a difficult task. Thus we focused on the prediction of spot or no spot. To provide a proper solution, we created decision trees with the matlabs internal function 'classregtree' [24] (Figure 3.4 (5)). The classregtree class takes a matrix of predictor values and a vector of classes as input and automatically creates a decision tree as shown in Figure 2.6. After the clustering we used the resulting list of feature values and the knowledge from the 'gold standard'(see section 2.2 for more details) to create decision trees. The classes are 'spot' and 'no spot'. These decision trees were created out of smaller subsets and were used to evaluate the feature value list with the matlabs internal 'eval' function and assign a class to each cell. The advantage of using decision trees over a fixed threshold is that they provide a more flexible and accurate handling of the measured feature values.

When the clustering is completely done for every cell, we evaluate the prediction list with a decision tree (2.4). The tree can either be loaded from existing trees that were generated in earlier runs with a different data set or in case of the 10-fold cross validation the trees will be generated during the procedure.

3.7 Evaluation

The last step is the evaluation of the predicted classes. Therefore we use a script that compares the predicted classes to the 'gold standard' we created with the annotation tool(2.2). As the output of the whole prediction procedure we gain a true-positive, a true-negative, a false-positive and a false negative value (Figure 3.4 (7)).

The true-positive value represents the amount of cell that were annotated to have a spot and it is consistent with the prediction, whereas the false-negative value represents the amount of cells that are not consistent with the prediction [25]. The true-negative value represents the amount of cells that were annotated to have no spot and it is consistent with the prediction, whereas the false-positive value represents the amount of cells that are not consistent with the prediction. With these values it is possible to calculate the sensitivity (equation 8) and specificity (equation 9) of the method, as well as an F1score. The sensitivity or true-positive rate represents the percentage of the correctly as positive predicted objects from all as positive predicted objects. The specificity or true negative rate represents the percentage of the correctly as negative predicted objects from all as negative predicted objects. Together these two values can be seen as the quality of a prediction.

$$sensitivity = \frac{true_{positive}}{true_{positive} + false_{negative}} \tag{3.1}$$

$$specificity = \frac{true_{negative}}{true_{negative} + false_{positive}} \tag{3.2}$$

The F1 score is calculated by $2 * \frac{precision*recall}{precision+recall}$ and is a measure between 0 and 1 that combines the precision and the recall of the method with the harmonic mean[26]. The precision is calculated by $\frac{tp}{tp+fp)}$, whereas the recall is calculated by $\frac{tp}{tp+fn}$ also known as the sensitivity or true positive rate.

3.7.1 10 fold cross validation

To compare different prediction methods we evaluated each method with a 10 fold cross validation. We used the matlabs internal function 'crossvalind' [27] to split our data set into 10 distinct smaller sets. The 'crossvalind' function assures that each set has an equal amount of cells containing and missing a LRC. We run the prediction pipeline for ten times. With every run we used an other set to create a decision tree and predicted the other nine sets with that tree. We calculated the sensitivity, specificity and the F1 score for every run and calculated the mean of all runs. Figure 3.10 illustrates how the different methods have performed. The best method related to the F1 score was method 25. It used the GMM clustering method with maximal two allowed clusters. The method used the intensities as well as their x and y coordinates for the clustering. The prediction was performed with a decision tree that was trained on a subset and consists of combination of the 'one to all value' and the 'cell standard deviation'. No other features were used for this prediction. The best method that used the kmeans as clustering was method 21 and is the method with the best sensitivity. It used the intensities only for the clustering process and the prediction was performed with a decision tree that was trained on a subset and consists of a combination of the 'one to all value' and the 'small cluster area / big cluster area'. Method 15 represents the approach that was used by Vannini et al.. They used an approach for prediction of LRC that relies on a chain of fixed threshold cutoffs on cluster size difference and intensity difference. The best thresholds were previously determined by a simple ROC analysis (data not shown). All prediction results were very close together. Nevertheless it turned out that the GMM methods showed better results. The best two prediction methods used the gmm clustering. The methods that used the same data and prediction parameters were method 8 and method 3, that showed worse prediction results compared to all other methods. The methods that included the x-y coordinates of the intensities in the clustering showed also better results. The methods that used the same prediction parameters as the top two methods but used the intensities only for the clustering were method 18 and 1.

MethodNr	Cluster method	Prediction parameter Nr.	Data used	Sensitivity	Specificity	F1 Score
1	gmm 2	1	nonspace	0.74067	0.56515	0.71276
2	kmeans 2	5	space	0.74067	0.56515	0.71276
3	kmeans 2	2	space	0.75113	0.59972	0.73102
4	kmeans 2	4	space	0.74406	0.64771	0.7554
5	gmm 4	5	nonspace	0.70794	0.66367	0.75666
6	gmm 4	4	nonspace	0.77594	0.6261	0.76024
7	gmm 8	1	nonspace	0.76499	0.64973	0.76682
8	kmeans 2	1	space	0.74579	0.67207	0.76941
9	kmeans 2	3	nonspace	0.76314	0.66832	0.77241
10	gmm 8	1	space	0.82849	0.65134	0.77506
11	kmeans 2	5	nonspace	0.77433	0.65974	0.77777
12	gmm 4	1	nonspace	0.73569	0.67977	0.78046
13	kmeans 2	2	nonspace	0.76416	0.70882	0.79467
14	gmm 2	4	nonspace	0.75957	0.70767	0.80024
15	kmeans 2	6	nonspace	0.79034	0.70636	0.80313
16	kmeans 2	3	space	0.75384	0.72739	0.80425
17	kmeans 2	6	space	0.68845	0.73996	0.80509
18	gmm 2	2	nonspace	0.76285	0.73821	0.80962
19	kmeans 2	1	nonspace	0.76901	0.75298	0.81899
20	gmm 4	4	space	0.81691	0.74594	0.82914
21	kmeans 2	4	nonspace	0.84188	0.74938	0.83876
22	gmm 4	1	space	0.78952	0.77832	0.84134
23	gmm 2	4	space	0.85972	0.77249	0.84736
24	gmm 2	1	space	0.80173	0.78528	0.8493
25	gmm 2	2	space	0.81366	0.79724	0.85855

Table 3.1: Result table of the 10 fold crossvalidation. The table is sorted by the F1 score. The first row shows the method number that accord with the method numbers from Figure 3.10. The second row shows the method as well as the maximal amount of allowed clusters that was used to cluster the cell images. The third row shows the data that was clustered on. 'nonspace' means that the clustering was performed on intensities only where as 'space' used the intensities and their scaled x-y coordinates. In the fourth row shows the Sensitivity of the method, the fifth row the Specificity and the sixth row the F1score.

Prediction parameter Nr	Parameters used method
1	'one to all value'
2	'one to all value' + 'cell standard deviation'
3	'one to all value' + 'best cluster count' + 'cell standard deviation'
4	'one to all value' + 'small cluster area/ big cluster area'
5	'one to all proportion'
6	Threshold on cluster size difference and intensity difference

Table 3.2: This table lists the prediction parameters that were used. The first row shows the parameter number that accord to the parameter numbers in table 3.1. The prediction of parameter Nr. 1 to 5 stands for a prediction that used a trained decision tree. Parameter 6 represents the approach used by Vannini et al. They used a chain of fixed threshold cutoffs on cluster size difference and intensity difference to determine their prediction. The best thresholds were determined by a ROC analysis (data not shown).

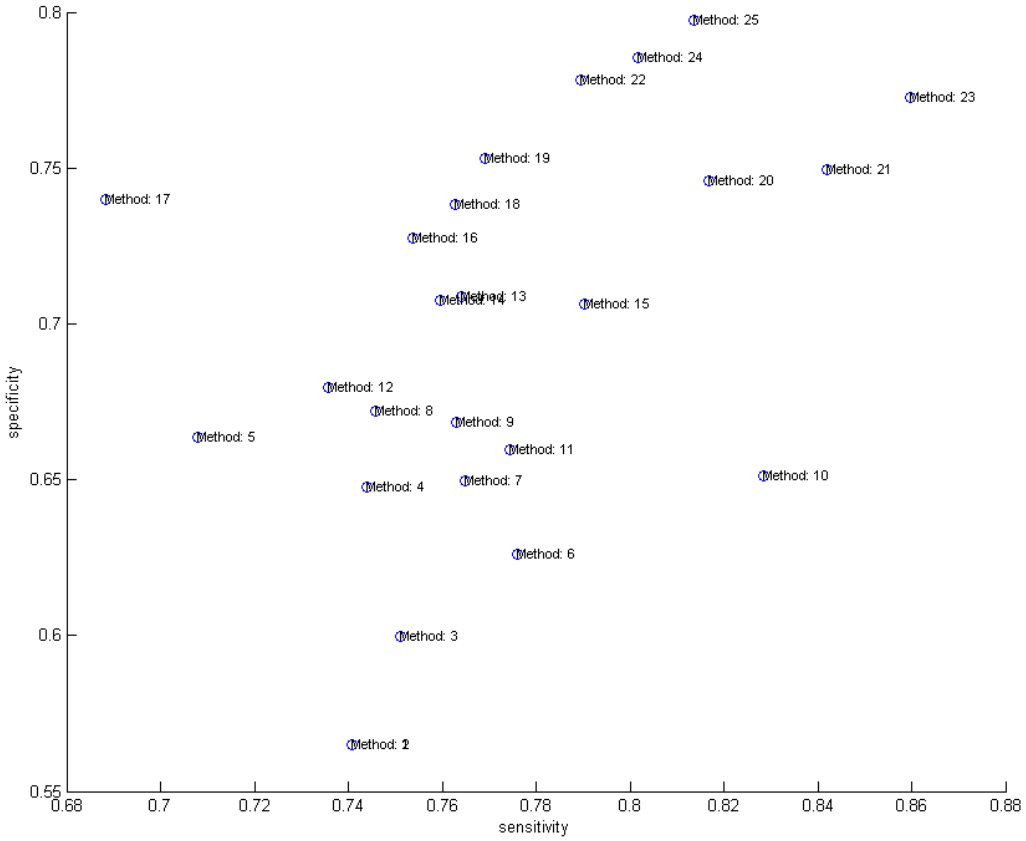


Figure 3.10: This plots shows the result of the 10 fold crossvalidation. The x axis represents the sensitivity, whereas the y axes represents the specificity. This plot shows that especially an improvement in the specificity could be be made over the already existing approach by Vannini et al. [14] (method 15). An improvement in sensitivity was achieved by the methods 25 and 23. The methods that used the GMM as clustering method showed very good results (method 22 to 25). They could achieve an improvement in sensitivity and specificity over method 15.

Chapter 4

Summary & Outlook

We got three data sets of lime lapsed microscopy experiments. The exp1 provided fully quantified and tracked data, but the LRC's were not clearly visible. Half of the positions of exp2 were treated with the fluorescent marker HSC-CD63VENUS that is known to be a LRC marker and provided us with images where the LRC is clearly visible. The second half of the positions of exp2 were treated with HSC-VANGL2VENUS and should be our negative control. However the LRC's were visible, so they provided us with even more positive example cell images. Exp3 provided us with a reliable negative example. During this work we quantified a set of cells from exp2 and exp3 with the QTFy program. As a result we gained cell masks that can be used to extract the pixel intensities as well as their belonging x and y coordinates of the cell images. We created a tool for the manual annotation of time lapse microscopy data. With this tool we were able to annotate the amount of LRC for each cell of the set we have quantified. This annotation is our gold standard that was used to evaluate our prediction methods. We applied two different clustering methods, the kmeans and the GMM method to the cell intensity data. We revealed some drawbacks of the kmeans clustering method, but it turned out that these did not have a huge effect on the prediction quality. An approach to use a basian spartial scan statistic for clustering purposes failed due to numerical problems. The resulting clusters were used to calculate some self defined cluster and cell features. These features were used in combination with the gold standard from the manual annotation to train a decision tree on a small subset of our data set. This decision tree was used to predict the occurrence of a LRC for each cell of our data set excluded the cells the tree was trained on. To evaluate the prediction performance we used the 10 fold cross validation. We split our data set into ten independent data sets, keeping the class balance. We trained our decision tree on one set and used it to predict the other nine sets. We repeated that procedure for all ten subsets. As a result of the evaluation process we gained a sensitivity, specificity and the F1 score for each method. As table 3.1 illustrates we were able to improve the already existing approach by Vannini et al. (method 15). It turned out that the GMM clustering method delivered better prediction results. Also the data extension with the x and y coordinates for the clustering step turned out to deliver better results. We constructed an automated pipeline for the prediction of LRC in HSC on time lapsed microscopy data. A GUI for an easy usage of the prediction pipeline is planed for the future. It should be designed to be intuitionally usable and able to start the prediction pipeline with select the selected parameters. We also planed the creation of labeled linage trees. These trees should look like the tree in figure 2.2, but with a label for each timepoint whether the cell has a LRC. The labeled linage tree feature could be integrated into the GUI. In this work we developed a automated pipeline for an large scale prediction of LRC in time lapse microscopy data that helps biologists to prioritize their research targets and aids in the research of asymmetric cell division. We hope that the pipeline helps to gain new insights into the understanding the relevance of LRC's.

Acknowledgements

Tanks to Michael Schwarzfischer who was a great supervisor. Tanks to Dirk Löffer who supported us with the experiment data. Tanks to the whole CMB group for the inspiring working atmosphere. Special tanks goes to my friends and family who supported me during the work.

Bibliography

[1] S. H. Orkin and L. I. Zon, "Hematopoiesis: an evolving paradigm for stem cell biology.," *Cell*, vol. 132, pp. 631–44, Feb. 2008.

[2] "What are the unique properties of all stem cells? [Stem Cell Information] - http://stemcells.nih.gov/info/basics/pages/basics2.aspx Online Ressource."

[3] T. U. At, "Helmholtz Center Munich Computational Modeling in Biology Masterarbeit," 2012.

[4] T. Schroeder, "Tracking hematopoiesis at the single cell level.," *Annals of the New York Academy of Sciences*, vol. 1044, pp. 201–9, June 2005.

[5] S. Yamazaki, A. Iwama, S.-i. Takayanagi, Y. Morita, K. Eto, H. Ema, and H. Nakauchi, "Cytokine signals modulated via lipid rafts mimic niche signals and induce hibernation in hematopoietic stem cells.," *The EMBO journal*, vol. 25, pp. 3515–23, Aug. 2006.

[6] S. Yamazaki, A. Iwama, S.-i. Takayanagi, K. Eto, H. Ema, and H. Nakauchi, "TGF-beta as a candidate bone marrow niche signal to induce hematopoietic stem cell hibernation.," *Blood*, vol. 113, pp. 1250–6, Feb. 2009.

[7] S. Yamazaki, A. Iwama, Y. Morita, K. Eto, H. Ema, and H. Nakauchi, "Cytokine signaling, lipid raft clustering, and HSC hibernation.," *Annals of the New York Academy of Sciences*, vol. 1106, pp. 54–63, June 2007.

[8] W. L. Conte, H. Kamishina, and R. L. Reep, "The efficacy of the fluorescent conjugates of cholera toxin subunit B for multiple retrograde tract tracing in the central nervous system.," *Brain structure & function*, vol. 213, pp. 367–73, Sept. 2009.

[9] T. Nagai, K. Ibata, E. S. Park, M. Kubota, K. Mikoshiba, and A. Miyawaki, "A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications.," *Nature biotechnology*, vol. 20, pp. 87–90, Jan. 2002.

[10] K. R. Siemering, R. Golbik, R. Sever, and J. Haseloff, "Mutations that suppress the thermosensitivity of green fluorescent protein.," *Current biology : CB*, vol. 6, pp. 1653–63, Dec. 1996.

[11] A. Rekas, J.-R. Alattia, T. Nagai, A. Miyawaki, and M. Ikura, "Crystal structure of venus, a yellow fluorescent protein with improved maturation and reduced environmental sensitivity.," *The Journal of biological chemistry*, vol. 277, pp. 50573–8, Dec. 2002.

[12] H. V. and V. C., "Novel structurally distinct family of leucocyte surface glycoproteins including CD9, CD37, CD53 and CD63.," *FEBS letters*, vol. 288, pp. 1–4, Aug. 1991.

[13] V. A. Cantrell and J. R. Jessen, "The planar cell polarity protein Van Gogh-Like 2 regulates tumor cell migration and matrix metalloproteinase-dependent invasion.," *Cancer letters*, vol. 287, pp. 54–61, Jan. 2010.

[14] N. Vannini, A. Roch, A. Griffa, S. Kobel, and M. Lutolf, "Identification of in vitro HSC fate regulators by differential lipid raft clustering Do not distribute . © 2012 Landes Bioscience .," pp. 1535–1543, 2012.

[15] M. Schwarzfischer, C. Marr, J. Krumsiek, P. S. Hoppe, T. Schroeder, and F. J. Theis, "Efficient fluorescence image normalization for time lapse movies," No. x, pp. 1–5, 2011.

[16] M. A. Rieger, P. S. Hoppe, B. M. Smejkal, A. C. Eitelhuber, and T. Schroeder, "Hematopoietic cytokines can instruct lineage choice.," *Science (New York, N.Y.)*, vol. 325, pp. 217–8, July 2009.

[17] T. U. Oliver Hilsenbeck, "Automated construction of cell lineage trees from time-lapse microscopy data Oliver Hilsenbeck," 2011.

[18] "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[19] E. W. Weisstein, "K-Means Clustering Algorithm – from Wolfram MathWorld."

[20] G. A. F. Seber, ed., *Multivariate Observations*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., Apr. 1984.

[21] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., Sept. 2000.

[22] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, pp. 461–464, Mar. 1978.

[23] D. B. Neill, A. W. Moore, and G. F. Cooper, "A Bayesian Spatial Scan Statistic,"

[24] A, "Construct classification and regression trees - MATLAB - MathWorks Deutschland - http://www.mathworks.de/de/help/stats/classregtree.html Online Ressource trkzjdflgjdfighdfjlgjdfiogjdfio." P, C Y. ER.

[25] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.3325 online ressource,"

[26] S. M. Beitzel, "On understanding and classifying web queries - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.127.634 online ressource,"

[27] "Generate cross-validation indices - MATLAB - MathWorks Deutschland - http://www.mathworks.de/de/help/bioinfo/ref/crossvalind.html Online Ressource."