March 2009

Systematic Evaluation of Image Analysis for Cell-Matrix Adhesion Studies in Cytomics.

Student:
Di Zi
Student number: 0631353
Computer science ~ Bioinformatics
Thesis project

Project location:
Imagery & Media
Section Imaging & Bioinformatics
Leiden Institute of Advanced Computer Science (LIACS)
University Leiden
Niels Bohrweg 1
2333 CA Leiden

Supervisors:
Dr.Ir.F.J.Verbeek — Section Image and Bioinformatics, LIACS
Sandra Zovko — Division of Toxicology, LACDR

Project period:
September 2008 – March 2009

Abstract

The cytomic study of cell-matrix adhesions on"*in-vitro*" cell systems requires automation in order to accomplish analysis on the right scale; that is large scale throughput of image data. These image data need to be properly preprocessed, segmented and subsequently the relevant features need be extracted from the cell-matrix adhesions. The image data represent experiments in which the cells are exposed to various conditions and cell-matrix adhesion kinases are knocked down by different siRNA duplexes; the goal of the measurements and feature extraction is to identify the role of different kinases and different types of cell-matrix adhesions under various experimental conditions and treatments.

The aim of this thesis project is to

1: establish a robust image analysis protocol that can be used for automation
2: estimate influence of different image processing methods and different imaging requiring conditions (microscope settings) on the measurements.
3: identify important kinases which have high information in regulation of cell-matrix adhesion ----- also called hits extraction.
4: Given a number of classifier, a good one is applied to learn different types of cell-matrix adhesion's behavior separately, given the features and experimental setup

Various image processing methods give rise to some questions that need to be solved. Which noise suppression and which segmentation should be used to get stable measurement of the object of interests? At the same time the segmentations should be subject to a critical evaluation in which we try to estimate how segmentation results are influenced by image conditions. Once an idea on robustness has been developed the features are measured. Before using these features, evaluation of features is done to check whether features are truthfully measured. New context-sensitive and valid features must be added to the family of features that is now common in this field. Several classifiers are tested and one is selected in the classification of cell-matrix adhesions. In the end, a robust image analysis is employed in the workflow of the Cytomic screens as these are applied in the Division of Toxicology department of Leiden Amsterdam Center for drug research (LACDR), Leiden University.

Contents

1 Introduction

1.1 Context of this project

The project is a cooperation project between a research group "Imaging and Bioinformatics" and Division of Toxicology department of Leiden University. "Image and Bioinformatics" is a one of the research groups of Leiden Institute of Advanced Computer Science (LIACS). It is lead by Dr. Ir. Fons Verbeek. The research focus of this group is on bio-imaging and the relation of the analysis of image information to other bio-molecular information resources. At present the bio imaging has its emphasis on microscopy modalities, in particular light microscopy.

The Division of Toxicology is part of the Leiden/Amsterdam Center for Drug Research (LACDR) and situated at Leiden University. "Cell-matrix adhesion signaling and tumor/metastasis formation" is one of specific research areas within the Division of Toxicology. The goal of this research is to study the dynamics of cell matrix adhesions in relation to cell migration and cancer cell metastasis. In this project, the Division of Toxicology provides experimental image data, which are obtained by high-through put screening using confocal microscope.

1.2 Problem definition

Cell-matrix adhesion is a subject that has been studied by many research groups in the past years [9, 18] (Cf. paragraph 2.5). However, there are several problems to be solved. The first problem is that no previous research established an evaluation system for segmentation methods according to different image conditions. Most of research uses global segmentation or watershed algorithm combined with global segmentation. The limitation of this method is that its performance is not reliable when it is applied on rather high noise level of images, especially on uneven illuminated images. Previous research [18] states image using objective equal to or better than 60x/0.9 numerical aperture (NA) are required. A robust image analysis should not be limited by image quality to such a big level.

Secondly, most of researches used normally distributed statistical parameters, for instance mean value or standard deviation of each feature, to compare siRNA treated group with control group. However the distributions of most of features are not normally distributed. In this case, the use of normally distributed statistical parameters is clearly inapplicable.

Thirdly, there is no research established an evaluated classification system which allows us to learn different type of cell-matrix adhesion separately. Previous research [18] used 90^{th} percentile values to distinguish different types of adhesion. However, it is not clear on which basis this percentile has been chosen and not for example 80^{th} percentile.

The goals of this project were to address thee questions:

1: we want to build a robust image processing system which is independent of image conditions, based on an evaluation system of their performance.
2: Instead of using normally distributed statistical parameters, we would compare treated to control cells based on non-parametric statistical tests.
3: From the available classifier, a good adhesions classification method is expected to be established given features, which are evaluated given experiment setup.

2 Biology background: Cell-Matrix Adhesion

2.1 Introduction:

In cell biology, cell-(extracellular) adhesions (also called cell-matrix adhesions) are specific types of large macromolecular assemblies formed at the membrane of the cultured cell and the underlying extracellular matrix (ECM). These structurally defined adhesion sites were initially described about 30 years ago in studies using interference-reflection microscopy and electron microscopy (Abercrombie and Dunn, 1975; Abercrombie et al., 1971; Izzard and Lochner, 1976; Izzard and Lochner, 1980). In more than 30 years of research, cell-matrix adhesions are proved that they play an essential role in important biological processes including cell motility, cell proliferation, cell differentiation, regulation of gene expression and cell fate, as the mechanical linkages to the ECM, and as sub-cellular macromolecules that mediate the regulatory effects (e.g. cell anchorage) of ECM adhesion on cell behavior [1]. In following paragraphs, more detail about cell-matrix adhesions are described from morphology, diversity and dynamic of cell-matrix adhesions.

2.2 Morphology:

To date approximately 150 different proteins have been identified as participating in the control of adhesion formation, stability and dynamics [2]. The heterodimeric transmembrane receptor family of integrins forms the main direct connection point between cell-matrix adhesions and proteins of the extracellular matrix [3]. The outer domain of the integrin binds to extra-cellular proteins like collagen, laminun and fbronectin. Within the cell, the intracellular domain of integrin binds to the cytoskeleton via adapter proteins. Talin, α-actinin, filamin, paxillin and vinculin are examples of such adapter proteins [5, 6, 7]. Each confers their own signaling properties and in turn leads to recruitment of many other intracellular signaling proteins such as focal adhesion kinses (FAK) and Src, which form the basis of the adhesion signaling cascade (Cf. Figure 1)[8, 9]

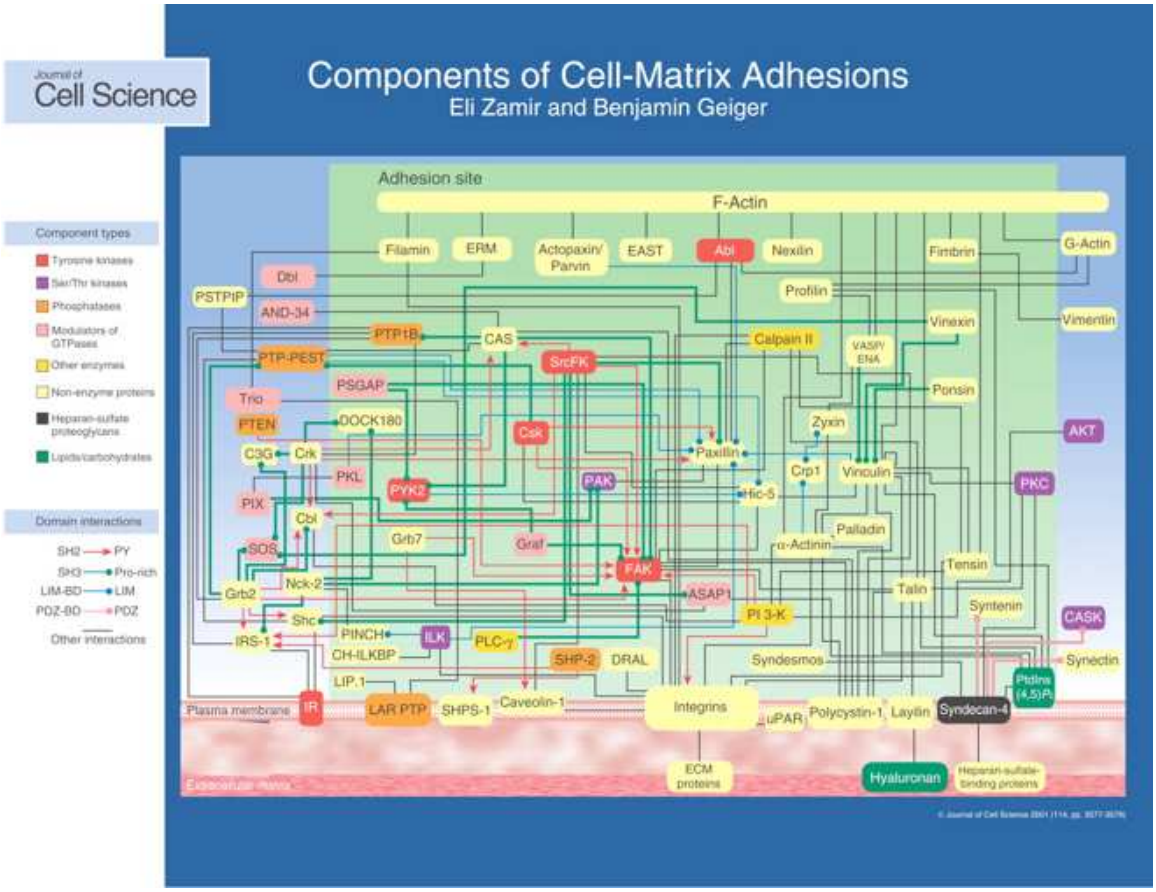


Figure 1: ***A scheme summarizing known interactions between the various constituents of cell-matrix adhesions.*** Components that were found to be associated with cell-matrix adhesion sites are placed inside the internal green box, whereas additional selected proteins that affect matrix adhesions but were not reported to stably associate with them are placed in the external blue frame. The general property of each component is indicated by the color of its box, and the type of interaction between the components is indicated by the style and color of the interconnecting lines, as indicated at the legend. For further details about this scheme see Cell Science at a Glance in this reference issue [9].

2.3 Diversity of cell-matrix adhesions

Observations from very early studies on adhesion structures in fixed cells identified the presence of three types of adhesions within a single cell [10]. (Cf. Figure. 2): focal complexes (FC), focal adhesions (FA) and fibrillar adhesions (FB). Terminology for different types of cell-adhesion structures is sometimes inconsistent, for instance focal adhesions are sometimes termed as focal contacts [9]. In our report, we use the terminology as proposed by [35], which is presented in Figure 2.

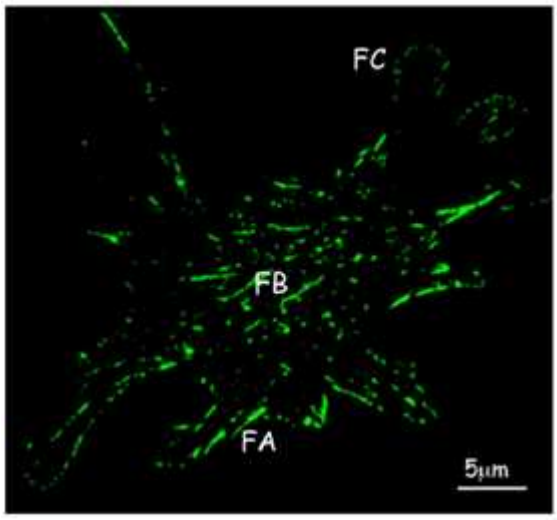


Figure 2: Focal adhesion types and composition. Cartoon depicting the three predominant types of adhesion typically found in an adherent cell plated on extra-cellular matrix; focal complex (FC), focal adhesion (FA) and fibrillar adhesions (FB) are shown. Cartoon schematics of the typical protein composition defining each adhesion type are also shown. Scale bar 5μm.

Focal complexes are small, dot-like, transient structure, which are usually located at the leading edge of lamellipodia [11]. Compared with focal complexes, FAs are larger, more mature structures, which are in part formed from the maturation of FCs. These adhesions are normally oval-shaped and usually located at periphery of the cell. FBs are thought to be derived from a subset of FA [12, 13]. They are long, highly stable complexes and are close to the cell centre (Figure 2).

These three types of matrix adhesion appear to differ not only in their shapes and molecular composition, but also in their functions. The detail of how they function differently beyond the scope of this project., therefore it will not be discussed here.

2.4 Dynamics

The dynamic assembly and disassembly of focal adhesions plays a central role in cell migration. At the beginning of cell migration, focal complexes are formed at the leading edge of the cell in lamellipodia. Many of these focal complexes fail to mature and are disassembled as the lamellipodia withdraws. However, some focal complexes mature into larger and stable focal adhesions, and recruit many more proteins. Once in place, a focal adhesion remains stationary with respect to the extracellular matrix, and the cell uses this as an anchor on which it can push or pull itself over the ECM. During maturation, focal complexes and focal adhesions can also form fibrillar adhesions. The mechanism involved in fibrillar adhesions forming is still poorly understood. As the cell progresses along its chosen path, a given focal adhesion moves closer and closer to the trailing edge of the cell. At the trailing edge of the cell the focal adhesion must be dissolved.

2.5 Previous study on Cell-Matrix adhesions

As cell-matrix adhesions play very important role in cell migration, signal transmitting, regulation of extracellular-matrix assembly, cell proliferation, cell differentiation and cell fate, a lot researches have been already done or are being performed on the its molecular composition, function, and mechanism how it effects cell. Two famous studies which related to this project are presented in [9] and [18]. Hereby the brief introduction and discussion of these two researches are shown in the following paragraphs.

In [9], it was examined for molecular heterogeneity of cell-matrix adhesions and the involvement of actomyosin contractility in the selective recruitment of different plaque proteins. Global segmentation and Watershed segmentation was performed to automatically identify cell-matrix adhesions, followed by quantitative immunofluorescence and morphometric analysis in which axial ratio, area and average intensity of cell-matrix adhesions are measured. Particularly informative was fluorescence ratio imaging, comparing the local labeling intensities of different plaque molecules, including vinculin, paxillin, tensin and phosphotyrosine-containing proteins. Ratio imaging revealed considerable molecular heterogeneity between and within adhesion sites.

The research in [18] combined high-resolution light-microscopy and high-throughput screening to test detailed molecular and cellular responses to multiple perturbations. They developed an application of a screening microscope platform that automatically acquires and interprets sub-micron resolution images at fast rates. The analysis pipeline was based on the quantification of multiple sub cellular features and statistical comparisons of their distributions in treated vs. control cells. The segmentation method used in this research is Watershed segmentation. The features used for the comparisons are axial ratio, area and average intensity of cell-matrix adhesions.

2.6 siRNA and mechanism of gene knockdown by siRNA

Small interfering RNA (siRNA), sometimes known as short interfering RNA or silencing RNA, is a class of double-stranded RNA molecules, 20-25 nucleotides in length, that play a variety of roles in biology. Most notably, siRNA is involved in the RNA interference (RNAi) pathway, where it interferes with the expression of a specific gene [36].

The mechanism of gene knockdown by siRNA is that the double-stranded siRNA is synthesized with a sequence complementary to a gene of interest and introduced into a cell or organism, where it is recognized as exogenous genetic material and bind to messenger RNA of targeted gene so that prevent this messenger RNA from producing a protein. Therefore the expression of the targeted gene would be drastically decreased. Studying the effects of this decrease can show the physiological role of the gene product. Since siRNA may not totally abolish expression of the gene, this technique is referred as a “knockdown” [37].

3. Method

In this chapter, biological experiment protocol and microscope setting will be explained in paragraph 3.1 and 3.2. Paragraphs 3.2 will briefly introduce the workflow of the whole project. Start from paragraph 3.3, important steps of the workflow would be introduced in more detail.

3.1 Screening protocol.

Since cell-adhesion plays very important role on cell migration, through the study of high-throughput screen, focal adhesion kinases (Cf. Appendix Four) which are involved in regulation of the dynamics of cell-matrix adhesions are expected to be found. Those kinases could be targeted to prevent the process of cancer cell metastasis (Cf. Appendix Four). In this study, MCF7 (Cf. Appendix Four) human breast cancer cells are used and 779 genes of kinases contained in the Human kinases siRNA library (Dharmacon, "SMARTpool" siRNA) were tested. The role of the kinases was investigated by performing siRNA mediated knockdown (Cf. chapter 2.6): mixture of 4 siRNAs targeting the same gene.

To be able to investigate the dynamics of cell-matrix adhesions in "*in-vitro*", we developed an assay in which the assembly of the cell-matrix adhesion is decoupled from the disassembly. In this assay, MCF7 human breast cancer cells cultured with Dimethyl sulfoxide (**DMSO**) was used as control treatment. In order to induce the assembly of cell-matrix adhesions, cancer cells were also exposed to the microtubule (MT) depolymerizing agent nocodazole (**Noco**) for 4 hours. MT depolymerization causes adhesions to grow in size. Cell-matrix adhesion disassembly was induced by a 2 hour recovery period or wash-out (**WO**) after the Noco exposure. In this period microtubules were allowed to re-grow into the cell periphery, which normally leads to disassembly of adhesions. In this assay with three conditions, we can look at the role of the kinases in the two separate processes.

The assay was performed in a glass-bottom 96-wells plate and each plate contained following group of cells:

1. no-siRNA treated cells
2. control #2 siRNA treated cells: siRNAs which do not target any kinase genes of interest are introduced to the cells.
3. paxillin siRNA treated cells: siRNA which only targets paxillin is introduced to the cells to check the knock down efficiency.
4. siRNAs treated cells for different kinases

where group 1 and group 2 are control group used to compare with siRNA treated group. Each group of cells was prepared in three conditions: DMSO, Noco, and WO in duple wells (Figure 3). The whole screen includes 10 different experiments which are coded as number, for example the 4^{th} experiment (Figure 3). Each experiment consists of 6 different 96-wells plates (60 plates in total). During screening, the microscope starts at

the first well (A_1 in Figure 3), and continues to A_12, then going to well B_12 towards B_1 (S –shape image acquisition). For each well 5 images are acquired on different randomly chosen positions, with Perfect Focus System (PFS) turned on.

		1	2	3	4	5	6	7	8	9	10	11	12
4_1	A	1				9				15			
	B	2				10				16			
	C	3				11				17			
	D	4				12				18			
	E	5				13				19			
	F	6				14				20			
	G	7				21				23			
	H	8				22							

Figure 3: Experiment scheme of one 96 wells plate. This plate is the first plate of the 4th experiment. Every index from 1 to 20 represents one experiment in which one siRNA knock down one gene. 21 represents no-siRNA control group, 22 – control #2 siRNA group and 23 is paxillin siRNA control group. Each siRNA or control group takes up 4 wells, of which first two wells are DMSO condition and last two wells are NOCO condition. The Washout condition is in plate 5 which is not shown here.

The screen readout (Figure 4) consisted of the imaging of adhesion structures, after fluorescently staining the fixed MCF7 cells on

1. phospho-paxillin by Cy3 (red),
2. vinculin by GFP (green),
3. nuclei by Hoechst (blue).

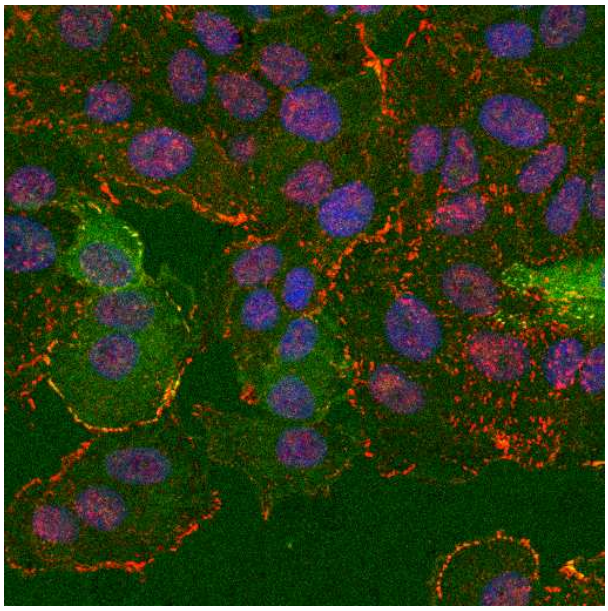


Figure 4: Example of an image obtained with high throughput screening microscopy. It shows MCF 7 cells expressing GFP-vinculin. Nuclei are visible as blue domains, while the adhesion structures are visible as green and red dot-like structures, GFP-vinculin and Cy3-paxillin, respectively. Both vinculin and paxillin are important adapter proteins of cell-matrix adhesions.

3.2 Microscope setting.

Images were automatically acquired using confocal Nikon 1 and 2 microscopes. They are provided by Division of Toxicology department. Imaging conditions used on Nikon 1 microscope were:

- 20x/0.75NA objective (air immersion) – Cf. Appendix Four
- Zoom 4
- Image real size: 151.9 μm
- Pixel length: In images of pixel size 512x512, pixel length is 0.311μm
  In images of pixel size 1024x1024, pixel length is 0.155μm
- No averaging
- Modular Confocal Microscope System: Digital Eclipse C1 plus
- Perfect Focus System
  (These settings were used for experiments 1,2,4,6,7,8,9 and 10)

On Nikon 2 microscope following imaging settings were used:

- 20x/0.75NA objective (air immersion) – Cf. Appendix Four
- Zoom 4
- Image real size: 151.9μm
- Pixel length: 0.311μm
- Image pixel size: 512x512
- Averaging 4x for red and green channels, 2x for blue
- Modular Confocal Microscope System: Digital Eclipse C1 plus
- Perfect Focus System
  (These settings were used for experiments 5 and 3)

3.3 Project Workflow.

This project was divided into three parts:

1: Optimize the segmentation method for different image conditions;
2: Optimize the microscope settings for image analysis;
3: Establish a robust image analysis protocol that can be use to identify hits.

The pipeline of these three parts is shown in Figure 5, 6 and 7. The more detailed explanation was given from paragraph 3.4.

Step 1:

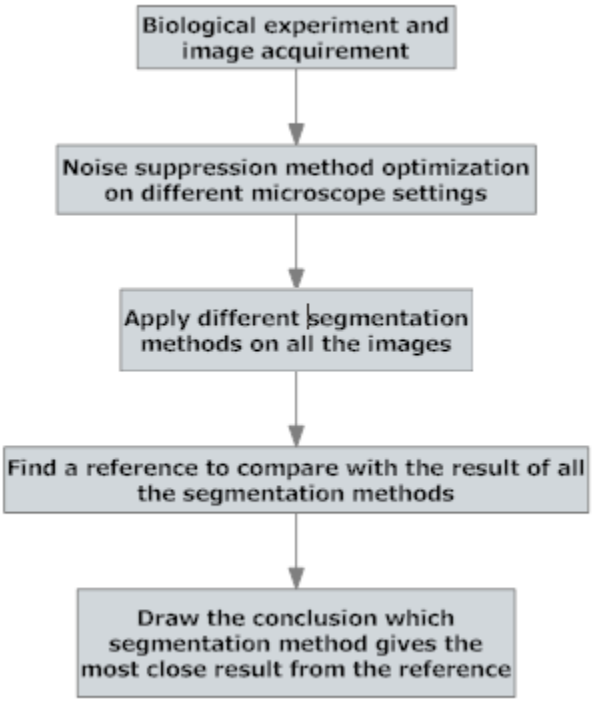


Figure 5: The workflow of optimization of the segmentation method for different image condition

Step 2:

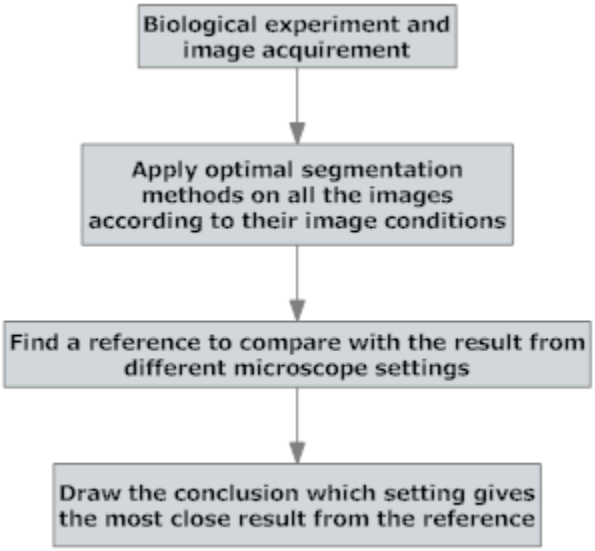


Figure 6: The workflow of optimization of the microscope setting

Step 3:

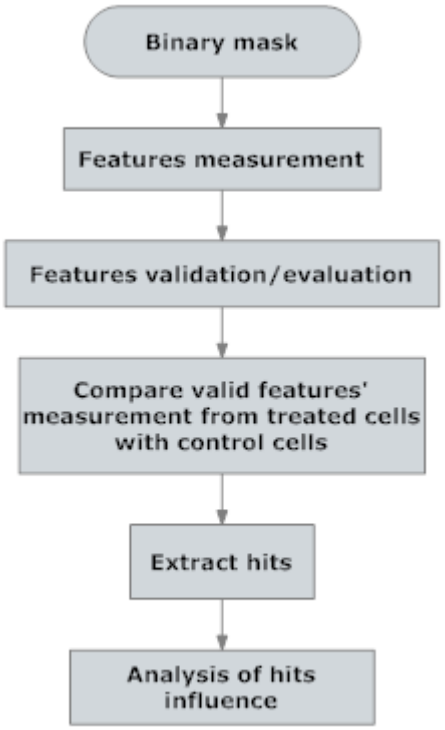


Figure 6: The workflow of image analysis

3.4 Segmentation Optimization

In this project, segmentation is considered as two steps: the first step is the preprocessing of the image which aims to remove the image noise; the second step is the segmentation which produces the binary images. Two steps are optimized separately. The optimization of noise reduction methods will be introduced briefly in paragraph 3.4.1. Paragraph 3.4.2 explains how optimization will be executed on segmentation methods.

3.4.1 Image Noise Reduction Methods Optimization

Image noise reduction is performed before segmentation. The purpose is to correct for imperfections in the frame images. Generally there are two different noises accounted for:

1. Long wavelength modulations of the background intensity due to non-uniform sensitivity among the camera pixels or uneven illumination,
2. Discretization noise from the CCD camera.

For the first type of noise, uneven background can be subtracted. There are several choices for background estimation as following [19]:

1. **Smoothing filte**r: Background is taken as the locally averaged intensity calculated over a region larger than the typical size of the object of interests.
2. **Masked smoothing filter**: As above, but the average excludes the segmented signal regions inside the mask.
3. **Local minimum value**: Instead of averaged intensity, background is taken as the locally minimum intensity over a region larger than the typical size of the object of interests. The advantage of this estimate is its being independent on segmentation masks, and that background subtraction never produces negative values.
4. **Rolling ball** [20]: A local background value is determined for every pixel by averaging over a very large ball around the pixel. This value is hereafter subtracted from the original image, hopefully removing large spatial variations of the background intensities. The radius should be set to at least the size of the largest object. Larger values will also work unless the background of the image is too uneven.

Smooth filter is a good approximation if the object size within the region is small. However, in our experimental images, some local regions have bigger object than background. Therefore this filter is unsuitable for those images. For masked smoothing filter, the mask of object is required, which in our project is not provided before segmentation. Local minimum value would underestimate the background levels so it could not be used neither. Therefore in this project rolling ball is used to subtract background in this project.

The radius of rolling ball is required to be at least the size of the largest object. Experimentally, we found that the radius of objects is less than 6 pixels for images of size 512 x512, less than 10 pixels for images of size 1024x1024. Accordingly we use these two values as radius of rolling balls to remove the uneven background meanwhile keep the shape of our cell-matrix adhesions.

We put our focus on reducing the second noise. Two types of methods are widely used [21]:

1. **Linear smoothing filter**: The most common filter is Gaussian filter. It removes the noise by convolving the original image with a mask of certain size. The weights of mask are determined by a Gaussian function.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad \text{Equation (1)}$$

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution. Since the value of each pixel is also depended on its neighbors: the closer neighbor has bigger weight, which means it

gives more contribution to the current pixel. Through the convolution, the value of each pixel is brought into closer harmony with the values of its neighbors.

2. **Nonlinear methods**: This category has median filter, maximum filter and minimum filter. They assign the median, maximum or minimum value of all pixels within a local region of an image to the middle pixel of this local region. In this project, we concern the local region as a square box (2r+1)*(2r+1), with the span size r. The disadvantage of maximum filter and minimum filter is that the maximum filter could enlarge the noise on the background while the minimum filter could not remove the relatively low intensity noises which are inside the object. However both kinds of noise are existed in our images. Therefore this project only test median filter among nonlinear methods.

For Gaussian filter and median filter, we observed from the experiment that they both affected the size of focal adhesion (Cf. Section 3.6.1) while the size is one of the most important morphological features we need to analyze: Gaussian filter tended to blur the boundary objects which induce the increase of size; median filter decreased the size of all adhesions. Which filter influences the size less and how they influence them is crucial for us. Unfortunately, there was no research about the comparison on their level of influence. Moreover, filters may remove the noises with the compensation of loss of some information in the form of image detail. For some low level noise of images, it might be not necessary to apply any filter, since the segmentation itself would remove noises which have relatively lower intensity than objects. To make a decision, we wanted to find a reference binary mask (Cf. Appendix Four) which could represent the real size of focal adhesion. Then it could be compared with the segmentation result from Gaussian filter processed images, Median filter processed images or no filter processed images. The one which is gives the closest result to reference is the best solution.

3.4.1.1 Reference Binary Mask:

The reference binary mask was set by biologist manually drawing the contour of cell-adhesion on images. Then the contours were refined by locally adjusting threshold. According to biologists' professional view, we could distinguish the cell-matrix adhesion from noise, and briefly recognize the size, shape of cell-matrix adhesion. However, because of sensitive drawing equipment, it was impossible for them to draw the contour precisely. Therefore locally segmentation and local threshold adjusting were applied to refine these contours. The segmentation was performed on each individual cell-matrix adhesion region. All the threshold values within certain scale were tried on each region. The masks obtained from different thresholds were compared with original image. Then the mask which matched adhesion's shape best is picked up to be the reference (Figure 18-B). Though this is a manual, time consuming way to set the reference, it gives a reasonable estimation of real size of cell-matrix adhesion.

3.4.1.2 Test images:

From each microscope setting, 1 image was randomly picked out. Then two sub parts of image were subtracted from each image on random position. These sub-images were used as test images. For convenience, the index is assigned to each test image. Table 1 shows indexes of images and corresponding microscope setting parameters. The reference binary mask of each image was prepared. All test images would be processed by Gaussian filter, Median filter and no filter separately.

Image Index	Pixel length	Image size (pixels x pixels)	Objective: magnification	Objective: NA	Digital zoom-in	Averaging
1	0.155µm	1024x1024	40x	0.75 NA	6x	4x
2	0.155µm	1024x1024	40x	0.75 NA	6x	4x
3	0.155µm	1024x1024	40x	0.75 NA	4x	4x
4	0.155µm	1024x1024	40x	0.75 NA	4x	4x
5	0.311µm	512x512	40x	0.75 NA	6x	4x
6	0.311µm	512x512	40x	0.75 NA	6x	4x
7	0.311µm	512x512	40x	0.75 NA	4x	4x
8	0.311µm	512x512	40x	0.75 NA	4x	4x
9	0.311µm	512x512	20x	0.75 NA	4x	no
10	0.311µm	512x512	20x	0.75 NA	4x	no

Table 1: 10 test images with relative image conditions.

3.4.1.3 Segmentation:

After applying filters on each test image, segmentation was performed to get the binary mask; subsequently we compare them with reference binary mask. Since the best segmentation method is unknown, one segmentation method would be applied for all test images. This is based on the assumption that all segmentation methods have no bias to any one of filters and give the same winner in the comparison. Here we concern Otsu segmentation (Cf. 3.4.2.1).

The reason why the Otsu segmentation is selected is:

1: Otsu segmentation is global segmentation, so it works much faster than other types of segmentation.
2: Otsu segmentation is more appropriate than other global segmentation methods in our case. Isodata method (Cf. 3.4.2.1) [22] only works when the intensity variance of objects and background are almost equal; and experimentally, entropy method [23] overestimate the threshold.

3.4.1.4 Kolmogorov–Smirnov test:

The Kolmogorov–Smirnov test (KS test) is a nonparametric estimation of minimum distance between two independent one-dimensional probability distributions.

Once the size of cell-matrix adhesions from binary mask was measured, the probability distribution of size (also called histogram) and its corresponding cumulative distribution function (CDF) $F_n(x)$ can be drawn easily:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{x_i \leq x} \qquad \text{Equation (2)}$$

where n is the total number of cell-adhesions, x is the size of cell-matrix adhesion, and $I_{x_i \leq x}$ is the indicator function, equal to number 1 if $x_i \leq x$ and equal to 0 otherwise. In other word, $F_n(x)$ is a monotonically increasing function which defines the proportion of cell adhesions of which size is smaller or equal than x.

The distance between CDF obtained from filters processed images (or no filter processed images) $F_n(x)$ and CDF obtained from reference binary mask F(x) can be calculated by KS test, represented as D_n- Kolmogorov–Smirnov statistic

$$D_n = \sup_x |F_n(x) - F(x)| \qquad \text{Equation (3)}$$

where sup is the supremum of set S. If two CDF are closer to each other, D_n converges to 0. Not like T-test which requires normal distribution for sample, KS test is a nonparametric test. In our experiment, the histogram of cell-matrix adhesion size does not show normal distribution. Thus KS test is better to be used in our study than T test.

The *p*-value of KS test is the probability of obtaining a result at least as extreme as the sample given that the sample is drawn from the reference distribution (in the one-sample case). Generally, the sample is not considered as being drawn from the reference distribution if the *p*-value is smaller than or equal to the significance level α, often set as 0.05.

One-tailed KS test can not only estimate the distance between two distributions, it can also predict the direction of the difference, for instance, focal adhesion size from Gaussian filter processed images are significant bigger than that from reference. Here

$$D_n = \sup_x \{F_n(x) - F(x)\} \qquad \text{Equation (4)}$$

which is the maximum negative difference or maximum positive difference for one-tailed KS test.

By using D_n and *p*-value, we would quantify the difference between reference distribution and distribution obtained from filters or no filter processed images.

3.4.1.5 Workflow of optimization of noise reduction methods

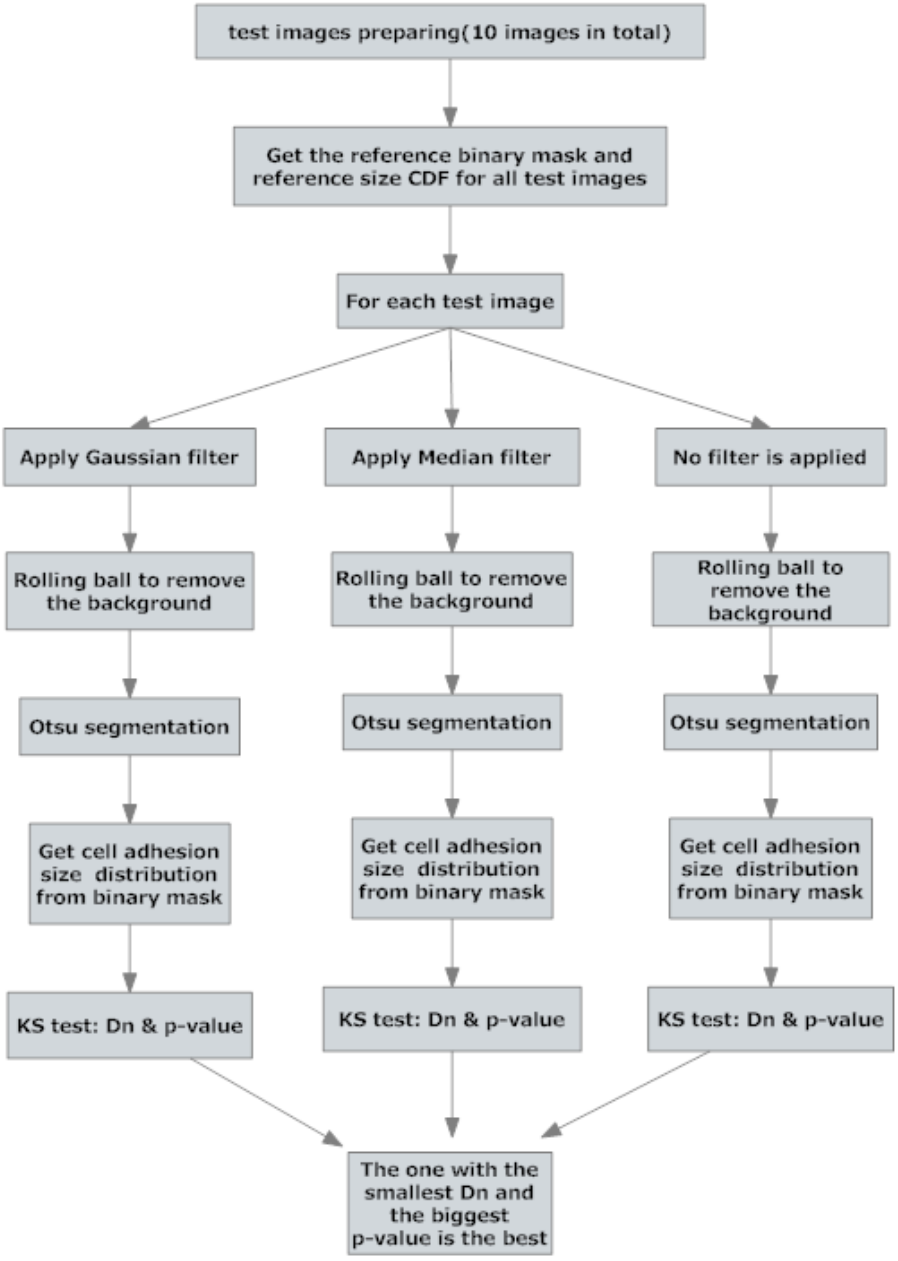


Figure 7: The workflow of Image Noise reduction methods optimization

3.4.2 Image segmentation methods optimization

The image segmentation is a process to separate the objects of interest with background by assigning different label: 1 for objects, 0 for background. Since it is the first step for all the image analysis, the accuracy is rather important. There are several mature segmentation methods applied widely in different situations. Find out which one is the best for certain images requiring conditions is one of our tasks. Here we firstly give a brief introduction of the segmentation methods which are categorized as following:

1. Global segmentation.
2. Local Adaptive segmentation
3. Watershed segmentation
4. Edge detection methods
5. Region growing methods
6. Watershed masked segmentation

3.4.2.1 Global segmentation

In last several years, global segmentation is the most common method applied on the cell-matrix adhesion study [9, 18]. The key of this segmentation is to define a threshold value for the whole image: Pixel with intensity higher than threshold will be labeled as foreground, otherwise background. This value can be set by user manually, or automatically computed by a thresholding algorithm, which is known as automatic thresholding. Two automatic thresholding are popularly used: Isodata segmentation and Otsu segmentation. Here we briefly introduce these two segmentation methods.

Isodata segmentation is an iterative method. The process is following:

1. An initial threshold (t) is chosen; this can be done randomly or according to any other method desired.
2. The image is segmented into object and background pixels as described above, creating two sets:
   G_1 = {I(m,n):I(m,n)>t} (object pixels)
   G_2 = {I(m,n):I(m,n)<=t} (background pixels)
   Where I(m,n) is the intensity of the pixel located in the m'th column, n'th row.
3. The average intensity of each set is computed.
   m_1 = average value of G_1
   m_2 = average value of G_2
4. A new threshold is created that is the average of m_1 and m_2
   $t' = (m_1 + m_2)/2$
5. Go back to step two, now using the new threshold computed in step four, keep repeating until the new threshold matches the one before it

Otsu segmentation exhaustively searches for the threshold that minimizes the within-class variance σ_w^2, defined as a weighted sum of variances of the two classes (1: background and 2: foreground):

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \qquad \text{Equation (5)}$$

Weights ω_i are the probabilities of the two classes which are separated by a threshold t and σ_i^2 is variances of these classes.

Otsu shows that minimizing the intra-class variance is the same as maximizing between-class variance σ_b^2:

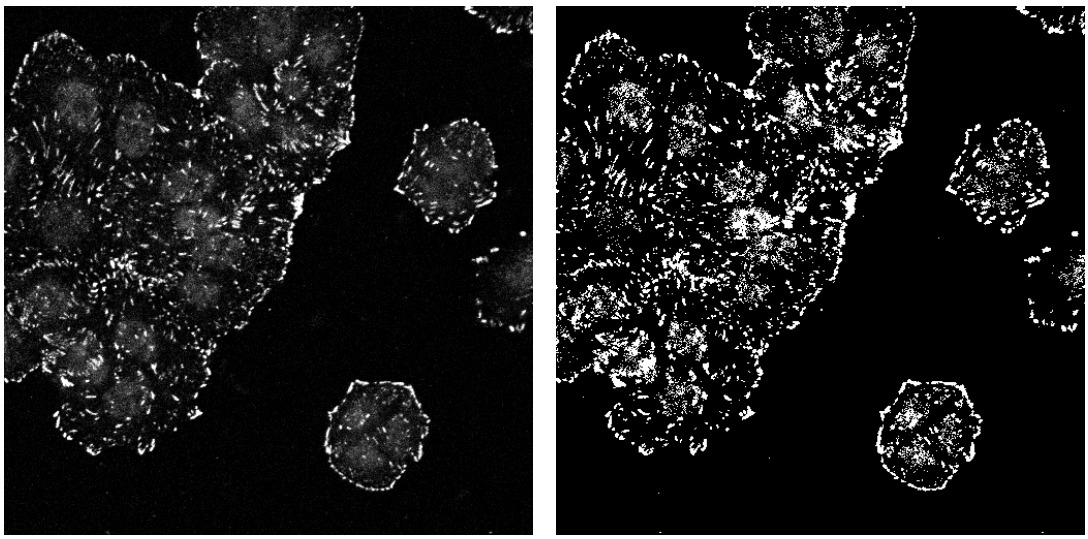
$\sigma_b^2(t) = \sigma^2 - \sigma_\omega^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2$ Equation (6)

which is expressed in terms of class probabilities ω_i and class means μ_i. They can be updated iteratively.

Algorithm

1. Compute histogram/ probabilities of each intensity level
2. Set up initial $\omega_i(0)$ and $\mu_i(0)$
3. Step through all possible thresholds t=1,2,3,....., maximum intensity
   1). Update ω_i and μ_i
   2). Compute $\sigma_b^2(t)$
4. Desired threshold corresponds to the maximum $\sigma_b^2(t)$.

The disadvantage of global segmentation method is that it requires even illumination of the image and that intensity of noise is lower than intensity of all objects. Both paper [9] and [18] used global segmentation since their images may have relatively more faint noise than adhesions. However, our Human Kinases screening images and other available images data did not reach this standard. In our images, the nucleus region and discretization noise were clearly visible (Figure 8). Their intensity was even much higher than some cell-matrix adhesions' intensity. Therefore global segmentation does not seem to fit our images.



A B

Figure 8: A: One example of Human Kinase screening image (red channel). B Binary mask obtained from A by Otsu segmentation.

3.4.2.2 Local Adaptive segmentation

Instead of finding global threshold, local adaptive segmentation performs thresholding in local regions. The local regions are usually square boxes of size (2n+1)x(2n+1). n is span size. Threshold value is calculated like the global thresholding, for instance, Isodata or Otsu, but in each region individually.

Compared with global segmentation, the computational time for local adaptive segmentation is enormous. The computational time would also increase by enlarging the span size n. But if the span size is set too small, some weak foreground would be assigned as background when that local region does not include any background pixels. Thus the optimal span size should be at least as the size of the biggest objects.

3.4.2.3 Watershed Segmentation

The watershed segmentation [24] splits an image into areas, based on the topology of the image. Grey value image can be considered as a 3D topographical map; the height of a point on this map is its intensity value. In the first step, pixels which have local minimum intensity is marked and considered as a start point for flooding. During the successive flooding of each marked region, some adjacent catchments basin would merge on a ridge. This ridge is defined as watersheds (Figure 9) or watershed lines. They are defined as background. When the flooding reaches a predefined height, it stops. The areas water coves are labeled as foreground. The pseudo code and one example result of watershed algorithm are shown below and in figure 10.

Algorithm:

1. Sort all the pixels in the image according to their intensity, forming a list in an increasing order.
2. Scan pixels in the sorted order until meet the pixels <= predefined threshold. At each scanning, the pixel(i, j) either:
   Case 1: Forms a new basin, if it is not touching any existing basin.
   Case 2: Assigned to an existing basin, if it is touching one.

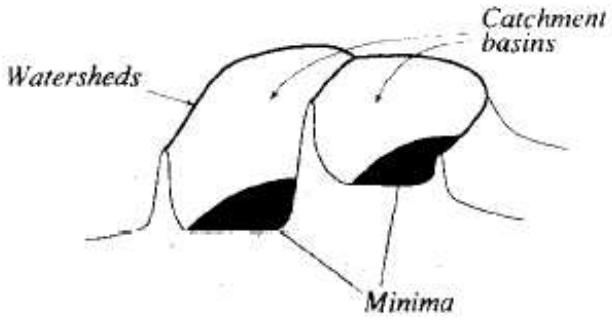


Figure 9: 3D gray-level map of image data.

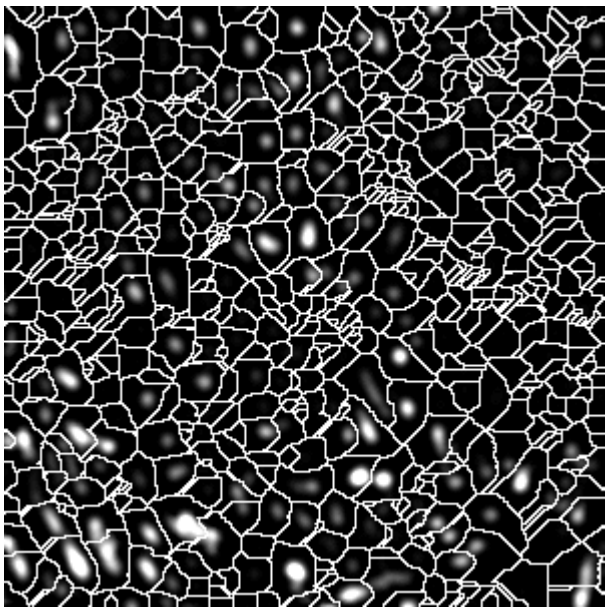


Figure 10: One example of applying watershed algorithm on a test image. The test image has already processed by Gaussian filter with σ=2.

Another popular use of watershed algorithm is to separate objects on given binary mask. Since watershed algorithm still leans on a global predefined threshold, which is similar to global segmentation. It can not give a precise binary mask on images which have noise brighter than some region of objects. But it can separate the objects which are blurred together or not completely separated, as long as there is an intensity valley between them. This is impossible for other segmentation methods. For this reason, watershed algorithm is usually used to produce watershed line to refine the binary masked after other segmentation methods.

3.4.2.4 Edge detection method

An edge is the boundary between two regions with relatively distinct gray-level properties. Basically, the idea underlying most edge detection techniques is the computation of a local derivative operator.
Gradient operator calculates the first derivative vector – gradient vector of image. The first derivative assumes a local maximum at an edge. For a continuous image I(x, y), where here x and y are the row and column coordinates respectively, we typically consider the two directional derivatives ∂xI(x,y) and∂yI(x,y). Of particular interest in edge detection are two functions that can be expressed in terms of these directional derivatives: the gradient magnitude and the gradient orientation. The magnitude of the gradient is defined as

$$|\nabla I(x,y)| = \sqrt{(\partial x I(x,y))^2 + (\partial y I(x,y))^2} \qquad \text{Equation (7)}$$

and the gradient orientation is given by

$$\Theta = \text{ArcTan}\left(\frac{\partial y I(x,y)}{\partial x I(x,y)}\right) \qquad \text{Equation (8)}$$

Local maxima of the gradient magnitude identify edges in I(x, y).

When the first derivative achieves a maximum, the second derivative is zero. For this reason, an alternative edge-detection strategy is to locate zeros of the second derivatives of I(x, y). The differential operator used in these so-called zero-crossing edge detectors is the Laplacian

$$L(x,y) = \frac{\partial^2 I(x,y)}{\partial x^2} + \frac{\partial^2 I(x,y)}{\partial y^2} \qquad \text{Equation (9)}$$

However, the derivatives enhance noise. Therefore an edge detector with smoothing effect is required. This can be implemented by convolve the original image with Gaussian filter before calculating two directional derivatives. Actually it is identical to convolving original image **I** with 1st derivative of Gaussian filter. Here comes to canny edge detector [25] which is based on this principle. The 1st derivative of Gaussian filter is defined as

$$G'(r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{r^2}{2\sigma^2}} \qquad \text{Equation (10)}$$

Where $r = x^2+y^2$ and x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis.

Algorithm:

1. Computing the gradient in X direction Gx by convolution with 1st derivative of Gaussian.
   $Gx = I * 2xG'(x)$ Equation (11)
2. Computing the gradient in Y direction Gy by convolution with 1st derivative of Gaussian
   $Gy = I * 2yG'(y)$ Equation (12)
3. Computing the magnitude and direction of gradient.

   $$G = \sqrt{Gx^2 + Gy^2} \qquad \text{Equation (13)}$$
   $$\theta = \text{ArcTan}\left(\frac{Gy}{Gx}\right) \qquad \text{Equation (14)}$$

4. Non –maximum suppression: The edge points determined in step 3 give rise to ridge in the gradient magnitude image. The algorithm then tracks along the top of these ridges and sets to zero all pixels that are not actually on the ridge top to give a thin line in the output, a process known as non-maximal suppression.

3.4.2.5 Region growing method

Region growing method assumes a certain level similarity among pixels belonging to same object. The algorithm will start from one prior pixel and grows until predefined similarity criterion is no longer fulfilled [26]. The similarity criterion (also called as homogeneous criterion) is regularly defined on pixel value considerations or on the anticipated size or shape of the object. There are different similarity criterions:

1. Intensity difference
2. Intensity gradient
3. Probability threshold.

The intensity difference based similarity criterion assumes that all pixels in interesting region share a range of intensity value. Any pixel connected to initial region will be given the same label if its intensity is in the predefined range.

Intensity gradient is a hybrid technique relies on edge detection and intensity information. Significant gradient change suggests a nature boundary of interesting region. Usually the range of intensity gradient is also predefined. Any pixel connected to initial region will be given the same label if its intensity gradient is in range.

Probability threshold cuts off pixels with probability estimation lower than threshold given density estimation in feature spaces. A predefined initial region is required. In addition, a density kernel must be defined.

3.4.2.6 Masked Watershed Segmentation

As explained in chapter 3.4.2.2, the choice of span size n is quite import. Too big span size would increase the computational time. Too small span size would assign weak foreground to background. Ideally each local region should contain one and only one cell-matrix adhesion. In 2008, a new method is proposed by Kuan Yan [27], which combines local adaptive segmentation with watershed algorithm. Since every local maximum on the grey value image represents a point within a cell-matrix adhesion, each region split by watershed methods has one and only one cell-matrix adhesions. Then the local adaptive segmentation can be applied in those pre-separated regions.

3.4.2.7 Workflow of segmentation methods optimization:

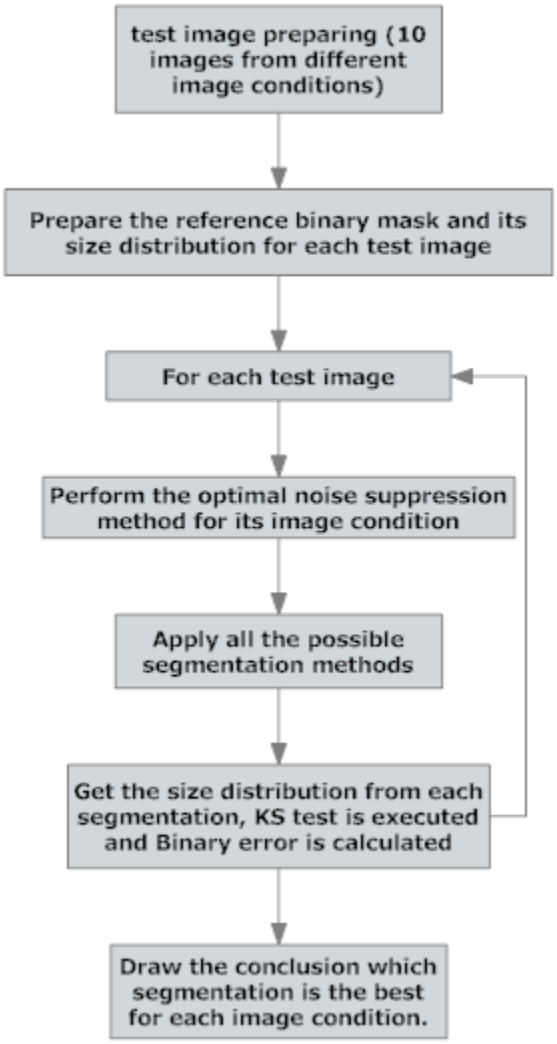


Figure 11: The workflow of Image Segmentation Methods Optimization

3.5 optimization of Microscope setting

Different microscope settings would have different level of influence on the measuring. Generally speaking, higher standard microscope setting, for instance the smaller pixel length and better objective, would keep more detail information of real image and decrease the measurement error. However, the heterogeneous distribution of fluorescence within a cell-matrix adhesion [9] makes it possible that higher standard would even increase the over segmentation rate because it keeps too much unnecessary information. In figure 8, one big focal adhesion is presented. This is acquired on microscope setting 100x /1.3 NA objective and pixel length 0.118 μm. We could even clearly see the dot-like small complexes distributed inside this big focal adhesion. For analysis, we definitely don't want to separate them into every dot-like complex. Unfortunately, the watershed segmentation method would over segment them. To avoid this over segmentation happening, some preprocessing is required, like what we proposed in section 4.1.2. On

the other hand, we can adjust the microscope setting, so that less fine detail can be visualized.

Figure 12: a large structure with a granular texture is found in paper [9]. The objective of microscope to obtain this image is 100x/1.3 NA. Pixel length is 0.118 μm.

The question for this problem is what microscope setting is the best for our measurement of adhesions morphology. Here we systematically study this with several groups of images.

3.5.1 Important microscope settings

As very import microscope settings which will influent segmentation accuracy, pixel length, averaging and zoom will be discussed in this section. Since only one objective is provided from experimental images, the optimization of objective is not discussed here.

3.5.1.1 Pixel length

Pixel length is the length of each pixel which is a unit of image display system. It describes the detail an image holds: Larger pixel length means fewer pixels are used to describe an image – therefore contains less detail than smaller pixel length which needs more pixels to describe the same size of image.

3.5.1.2 Averaging

Consider a noisy image G(x, y) formed by the addition of noise η(x, y) to an original image I(x, y); that is,

$$G(x,y) = I(x,y) + \eta(x,y) \qquad \text{Equation (15)}$$

where the assumption is that at every pair of coordinates (x, y) the noise is uncorrelated and has zero average value. The objective of the following procedure is to reduce the noise effects by adding a set of noisy image, $\{G_i(x, y)\}$.

If the noise satisfied the constraints just stated, it is sample problem to show that if an image $\bar{g}(x,y)$ is formed by averaging M different noisy images,

$$\bar{g}(x,y) = \frac{1}{M}\sum_{i=1}^{M} g_i(x,y) \qquad \text{Equation (16)}$$

then it follows that

$$E\{\bar{g}(x, y)\} = I(x,y) \qquad \text{Equation (17)}$$

$$\sigma^2_{\bar{g}(x,y)} = \frac{1}{M}\sigma^2_{\eta(x,y)} \qquad \text{Equation (18)}$$

where $E\{\bar{g}(x, y)\}$ is expected value of $\bar{g}$, and σ and ση are the variances of $\bar{g}$ and η, all at coordinates (x, y). The standard deviation at any point in the average image is

$$\sigma_{\bar{g}(x,y)} = \frac{1}{\sqrt{M}}\sigma_{\eta(x,y)} \qquad \text{Equation (19)}$$

This equation indicates that, as M increases, the variability of the pixel values at each location (x, y) decreases. Because $E\{\bar{g}(x, y)\} = I(x, y)$, this condition means that $\bar{g}(x, y)$ approaches I(x, y) as the number of noisy image used in the averaging process increase.

However, the more times of averaging, more time would be taken for microscope imaging. Moreover too much averaging would also induce blurring in the output image. So we can not do the unlimited averaging. Finding which averaging is the best for different resolution or objective lens is one goal of our project.

3.5.1.3 Zoom

A zoom lens is a mechanical assembly of lens elements with the ability to vary its focal length. Figure 9 gives an example photo of using zoom lens. They are often described by the ratio of their longest to shortest focal lengths. For example, a zoom lens with focal lengths ranging from 100 mm to 400 mm may be described as a 4:1 or "4×" zoom. The most common zoom lenses are 4x zoom and 6x zoom.

Figure 13: A photograph taken with a zoom lens, in which the focal depth was varied during the course of the exposure.

Test images

For testing how image conditions affect measurement and which condition gives the most stable measurement, three groups of test images were evaluated. Like the test images in "Segmentation Optimization" chapter, they were also sub images and their reference binary masks were prepared by method mentioned in section 3.4.1.1. For each group of test images, we only changed one microscope setting; other settings were still kept the same. The description of each group of image is presented in the table 2:

image group	image number	Pixel length	Averaging	Zoom
1	1	0.155µm	4	6
	2	0.155µm	4	6
	3	0.311µm	4	6
	4	0.311µm	4	6
2	5	0.155µm	0	6
	6	0.155µm	0	6
	7	0.155µm	4	6
	8	0.155µm	4	6
3	9	0.155µm	4	4
	10	0.155µm	4	4
	11	0.155µm	4	6
	12	0.155µm	4	6

Table 2: Test images and their relative image conditions.

In group 1, image 1 and 3 are imaged from the same position but with different pixel length; images 2 and 4 are also imaged from the same position with different pixel length. In group 2, image 5 and 7, 6 and 8 are imaged from the same position except the averaging is different. In group 3, 9 and 11, 10 and 12 they are from the same position of imaging but their zoom lens are different.

3.6 Image analysis

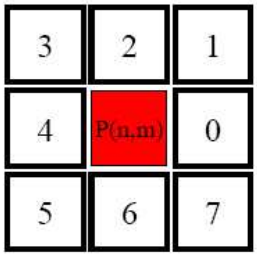
The image we need to analyze is the high-throughput screening of human kinases knockdown experiments. The screen protocol is already introduced in paragraph 3.1. From our study we hope to find important kinases (hits) which are probably involved in cell-matrix adhesion formation, individual signaling mechanisms; and also learn how they affect those mechanisms.

3.6.1 Feature measurement

After segmentation, one binary mask can be obtained from each image. For each 8-connected object, a unique label is assigned. The morphology measurement can be done on each labeled object. The morphology features we measured are following:

1. **Size(or Area):** It is calculated as the number of pixels presented in the binary mask of each object.

2. **Perimeter** [28]: The length of contour of the objects. The perimeter is derived from the boundary pixels. The boundary pixels are firstly converted to a chaincode with connectivity code as following:



8-connected

Ne represents the number of pixels with even chaincodes; No is the number of pixels with odd chaincodes, while Nc indicates the number of corners which represents the chaincode changes direction. The most aware way of calculating perimeter is proposed in 1982 by Vossepoel & Smeulders :

$$Lvs = (0.980)*Ne + (1.406)*No - (0.091)*Nc \quad \text{Equation (20)}$$

3. **Center of gravity** ($\bar{x}$, $\bar{y}$) also called centroid: Informally, it is the "average" of all points of each object. The center of gravity can be derived from raw moment

$$M_{ij} = \sum_x \sum_y x^i y^j (x,y) \quad \text{Equation (21)}$$

where I(x, y) is pixel intensity on the coordinate (x, y) and (x, y) is in the range of current object. Centroid is

$$\{\bar{x}, \bar{y}\} = \{M_{10}/M_{00}, M_{01}/M_{00}\} \quad \text{Equation (22)}$$

4. **Orientation** from moment [29]: The orientation of the major principal axis with respect to the x-axis of the image is given by:

$$\theta = \frac{1}{2}\tan^{-1}\left(2\frac{\mu_{11}}{\mu_{20}-\mu_{02}}\right) \quad \text{Equation (23)}$$

where μ is centralized moments calculated as following:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q I(x,y) \quad \text{Equation (24)}$$

5. **Long axis α and short axis β** [28] are major and minor axis of best fitted ellipse of the object. They are given by

$$\alpha = 2*\text{sqrt}\ \{\frac{2*[\mu_{20}+\mu_{02}+\sqrt{(\mu_{20}-\mu_{02})^2+4*(\mu_{11})^2}]}{\mu_{00}}\}$$ Equation (25)

$$\beta = 2*\text{sqrt}\ \{\frac{2*[\mu_{20}+\mu_{02}-\sqrt{(\mu_{20}-\mu_{02})^2+4*(\mu_{11})^2}]}{\mu_{00}}\}$$ Equation (26)

6. **Extension, dispersion and Elongation** [28]. Extension measures how much the shape differs from the circle. When it is circular shape, the value equals 0. Extension can increase without upper limit as the shape becomes less compact. Dispersion is the minimum extension that can be attained by uniform compression of the shape. To minimize its extension the shape must be compressed along long axis of the shape. Elongation measures how much the shape must be compressed along its long axis in order to minimize the extension. It is never less than 0 and never greater than extension.

   Three features can be derived from normalized moment.

   $$\eta_{ij} = \mu_{ij}/\mu_{00}^{(1+\frac{i+j}{2})}$$ Equation (27)

   and first 2 rotation invariant moments [29]

   $$I_1 = \eta_{20} + \eta_{02}$$ Equation (28)

   $$I_2 = (\eta_{20}-\eta_{02})^2 + (2\eta_{11})^2$$ Equation (29)

   The extension (E), dispersion (D) and elongation (L) are calculated as following:

   $$E = \log_2\lambda_1 = c\ln\lambda_1 \approx \ln\lambda_1$$ Equation (30)

   $$D = \log_2\sqrt{\lambda_1\lambda_2} = \frac{1}{2}c*\ln\lambda_1\lambda_2 \cong \frac{1}{2}\ln\lambda_1\lambda_2$$ Equation (31)

   $$L = \log_2\sqrt{\frac{\lambda_1}{\lambda_2}} = \frac{1}{2}c*\ln\frac{\lambda_1}{\lambda_2} \cong \frac{1}{2}\ln\frac{\lambda_1}{\lambda_2}$$ Equation (32)

   where $\lambda_1 = 2\pi*(I_1+\sqrt{I_2})$ and $\lambda_2 = 2\pi*(I_1-\sqrt{I_2})$

7. **Compactness shape factor** [28]: is a ratio of square perimeter length over surface area. It is a descriptor for irregular boundaries. The more regular object is, the higher the value compactness shape factor has. One example of compactness shape value related to object shape is shown below:

low compactness **compactness=0.764** **compactness=0.668**

Figure 14: One example of compactness shape factor value for different shape of objects.

8. **Average intensity of each project**:

$$\bar{I} = \frac{1}{size} \sum_{x,y} I(x,y) \qquad \text{Equation (33)}$$

where size is the number of pixels representing current object. I(x, y) is pixel intensity on the coordinate (x, y) and (x, y) in the range of current object.

9. **Nucleus distance** is the distance between centroid cell-matrix adhesion and the centroid of nucleus which this cell-matrix adhesion belongs to. Since the cell is not stained, from visualization we could not know which nucleus cell-matrix adhesion belongs to. Therefore we assign the cell-matrix adhesion to the closest nucleus.

10: **InNucleus** is a binary descriptor which represents whether current cell-matrix adhesion is within the nucleus region or not.

11: **Closest FA** is a distance measurement which computes the distance between current cell-matrix adhesion and its closest neighbor.

In addition, number of nuclei and number of cell-matrix adhesion are calculated for each image. This allowed clear identification of toxic effects and cell death. Toxicity or anti-adhesive activity would result in massive loss of cells in some wells. This could reflect large statistical error due to its small number of segmented objects. We therefore do not take into account the images with few nuclei and adhesions (Nuclei: <5, Cell-matrix adhesion: <20).

3.6.2 Feature evaluation

The image quality of human kinases siRNA screening plates is too low; consequently not all the features can be measured truthfully. For this reason, an evaluation system is necessary. For this purpose, we use two groups of images with higher quality as reference to evaluate the measurement on human kinases siRNA screening. These two groups of images are acquired from the microscope setting: pixel size is 0.155µm, objective 40x /0.75 NA, zoom 6x, and averaging 4. They are the images of control wells on one plate:

the first group is the images of nosiRNA control wells; and the second group is from #2 siRNA control wells. Both groups of images have DMSO, Noco and WO exposure condition.

We assume the higher quality of images would have more accurate measurement. For each feature measured in reference images, if its distributions are different under different exposure conditions, the same difference should be detected in the human kinases siRNA screening images. The difference between two distributions can be quantified as the distance between two distributions and the direction of difference. Hereby one- tailed KS test is applied which can predict the direction of change of distributions. All the features are measured in the reference images and the difference of distribution between different exposure conditions are calculated by one-tailed KS test. The images of the same control wells from human kinases siRNA screening are used to compare with reference. For each feature, if it shows the same difference under different exposure conditions as in reference images, this feature is considered as valid feature.

3.6.3 Hits

This section focus on the method of identifying hits which have high information in regulation of cell-matrix adhesion

3.6.3.1 Control group quality evaluation

As a reference to evaluate the variation of cell-matrix adhesions from treated wells, the quality of control wells image is quite important. Before applying statistical comparison test for treat wells against control wells, the control wells' image quality evaluation is necessary.

From our observation of plates, some control wells' images are not properly illuminated or the background noises significantly affect the quality of images which may induce large statistical error like segmentation result. All those images are considered as invalid images which should not be taken into account in later comparison test.

Moreover, from experimental evidence it has been shown Noco would increase the size of focal adhesion so that the size distribution would shift to left compared with DMSO situation (Figure 21); and WO would significant decrease the adhesion size contrast to NOCO (Figure 21). Thus the valid control groups should also have the same character.

For each plate, we check two groups of control wells: no-siRNA and #2 siRNA (Cf. Appendix Four) separately. One-tailed KS test is applied to calculate the distance between CDF of focal adhsion size from DMSO wells and that from Noco wells, and also between Noco wells and WO wells. Only the Noco wells which give significant size increasing is considered as valid control Noco for this plate, and the WO wells which shows significant size decreasing compared with valid Noco wells is valid for corresponding plate.

3.6.3.2 Discovery of hits

The score of each kinase can be calculated by comparison of treated wells to control wells using KS test. The *p*-value is set as score to describe the changes between the treated cells and control cells. Since we have three exposure conditions in control group, each condition of treated cells are compared with the same condition of control cells: DMSO (treated cells) vs valid DMSO (control cells), Noco (treated cells) vs valid Noco(control cells), WO (treated cells) vs valid WO(control cells). Top 10 siRNA with smallest *p*-value for each features (validated by feature evaluation) are selected as hits.

From this analysis, we expect to learn which siRNA-targeted genes severely affect adhesions on which feature, compared with its control group. Meanwhile, we also want to know which siRNA would block the affect of NOCO or WO on the cells.

To achieve this purpose, we apply one-tailed KS test for comparison of treated wells under DMSO against the same siRNA treated wells under Noco condition, and also for comparison of treated wells under Noco condition against the same siRNA treated wells under WO condition. The siRNAs which show no significant size increasing or even significant size decreasing in Noco condition compared with its DMSO condition are selected as hits. The siRNAs which give no significant size decreasing or even significant size increasing in WO condition compared with its valid Noco condition are also considered as hits.

The scheme of hits identification is concluded as following:

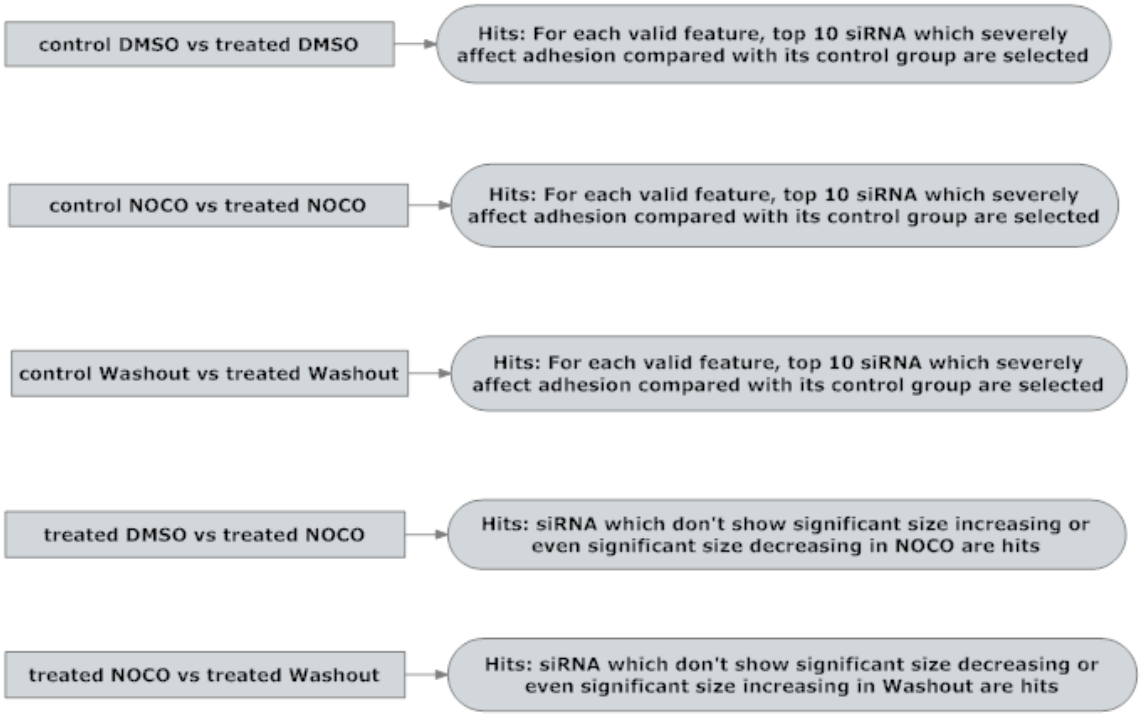


Figure 15: The scheme of hits extraction

3.5.4 Classification of three types of cell-matrix adhesion

In cells, three types of cell-matrix adhesions are distinguished: FC, FA and FB (Figure 2). They have different characteristic on morphology: FCs are small, dot like objects; FAs are bigger than FAs, oval shaped; FBs are fibrillar shaped thus more elongated. We hope that we can classify them into three groups according to their elongation and size so that we can learn how they are affect by siRNA individually. The challenge of classification is that we don't have training data with adhesions type labeled. Unsupervised clustering methods are needed.

Two major unsupervised clustering methods are hierarchical clustering [30, 31] and K-mean clustering [31, 32]. Hierarchical clustering refers to the formation of a recursive clustering of the data points. At the beginning, a distance matrix D (Table 3) is built up for all the data points. D_{ij} represents the distance between data point i and j. The distance can be Euclidean distance, Manhattan distance, maximum norm, Mahalanobis distance and Hamming distance [33]. The most common distance is Euclidean distance. The traditional representation of this hierarchy is a tree with all the single data point as leaves. Each time, hierarchical method finds the closest pair of elements and they are merged to one higher element. New distance between all pairs' of elements is recomputed and distance matrix D is updated. The recursive clustering would build up the hierarchical tree from leaves.

	x_1	x_2	x_3	x_4	x_5
x_1	0.00	1.58	1.76	5.22	4.53
x_2		0.00	0.74	5.50	5.10
x_3			0.00	4.81	4.48
x_4				0.00	1.12
x_5					0.00

Table 3: One example of Distance matrix of five data points. Red marks the most similar pair of objects in current distance matrix.

The K-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

The algorithm steps are [32]:

1. Choose the number of clusters, *k*.

2. Randomly generate *k* clusters and determine the cluster centers, or directly generate *k* random points as cluster centers.
3. Assign each point to the nearest cluster center.
4. Recompute the new cluster centers.
5. Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It minimizes within-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement of data convex clusters ("round shape"). In our data, the shape of each cluster is unknown. For those reasons, we have disregarded K-mean clustering in our approach.

Before we apply hierarchical clustering, cluster validation is performed. The purpose of cluster validation is to check whether grouping is really present. In our study we used Hierarchical Davies-Bouldin index score (DBI score) [33], which is based on within and between group scatter. This score system defines that for a good clustering, it should hold that

1: objects are compactly organized within a cluster and
2: clusters are far apart from each others.

For each pair of cluster, paired cluster criterion R is calculated as the ratio of between cluster variance and within cluster variance

$$R_{ik} = \frac{\sigma j + \sigma k}{||\mu j - \mu k||} \qquad \text{Equation (34)}$$

Where j and k represent cluster labels. μ and σ are calculated as following

$$\sigma_j = \sqrt{\frac{1}{nj}\sum_{xi \in Cj} ||xi - \mu j||^2} \qquad \text{Equation (35)}$$

$$\mu_j = \frac{1}{nj}\sum_{xi \in Cj} xi \qquad \text{Equation (36)}$$

Where xi is a feature vector of data point i and nj means the number of data points in cluster j.

The worst R_{jk} value for each cluster j is

$$R_j = \mathbf{max}_{k=1,\ldots g; k \neq j} R_{jk} \qquad \text{Equation (37)}$$

where g is the number of clusters. Then DBI score is given by

$$I_{DB} = \frac{1}{g}\sum_{j=1}^{g} R_j$$ Equation (38)

which is the average of worst case of all clusters. Thus the lower the DBI score, the better the clustering is.

From calculation of DBI score on 2D feature space [size, elongation], the optimal number of clusters are two instead of three (Figure 16) since the data points are cluttered together in this 2D feature space. According to the characteristic of different types of cell-adhesions, which FC is much smaller than other two types and FB is more elongated. Two layers of clustering are concerned here. On the first layer, we clustered the adhesion into two groups according to the size. From this clustering we expected to separate FC from FA and FB. From our observation of size histogram from control DMSO wells on different plates, two peaks are quite obvious with the valley around 3-5 pixels (Figure 22-A). When we draw the size histogram on control group images with objective 40x/0.75NA, which is two times of magnification of human kinases siRNA screening images, this valley also shifts to 5-9 (Figure 21-A). Moreover we checked the DBI score based on clustering adhesion only on size. The optimal number of groups is two. This indicates two separated groups indeed exist, and the valley is independent on the imaging condition. We thus set this valley as a threshold to segment FC from FA and FB. The rest of adhesions (FA + FB) are clustered by elongation. The DBI score validates that clustering them into two groups gives the lowest score (Figure 17). By this method, we clustered adhesions of all 10 experiments (only in control group) into three groups, plates by plates. The elongation threshold is between 0.8396-1.033. This small variance proves this is a stable threshold.

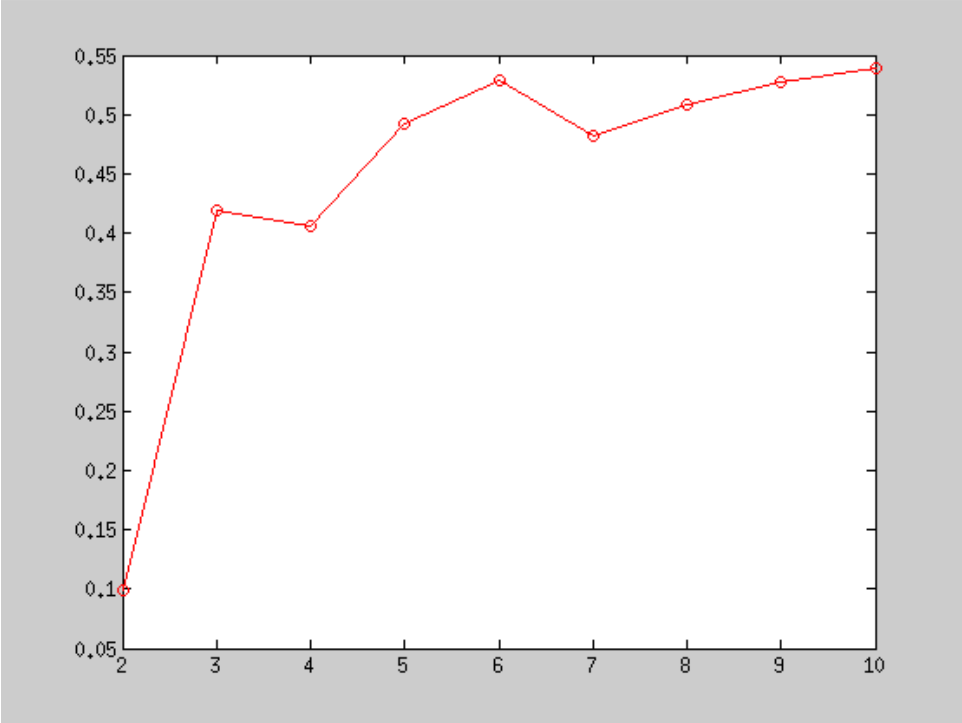


Figure 16: DBI score diagram of classification on both size and elongation: The x-axis indicates the number of group; and y axis is the DBI score of clustering corresponding number of groups. This diagram shows there are two classes instead of three one

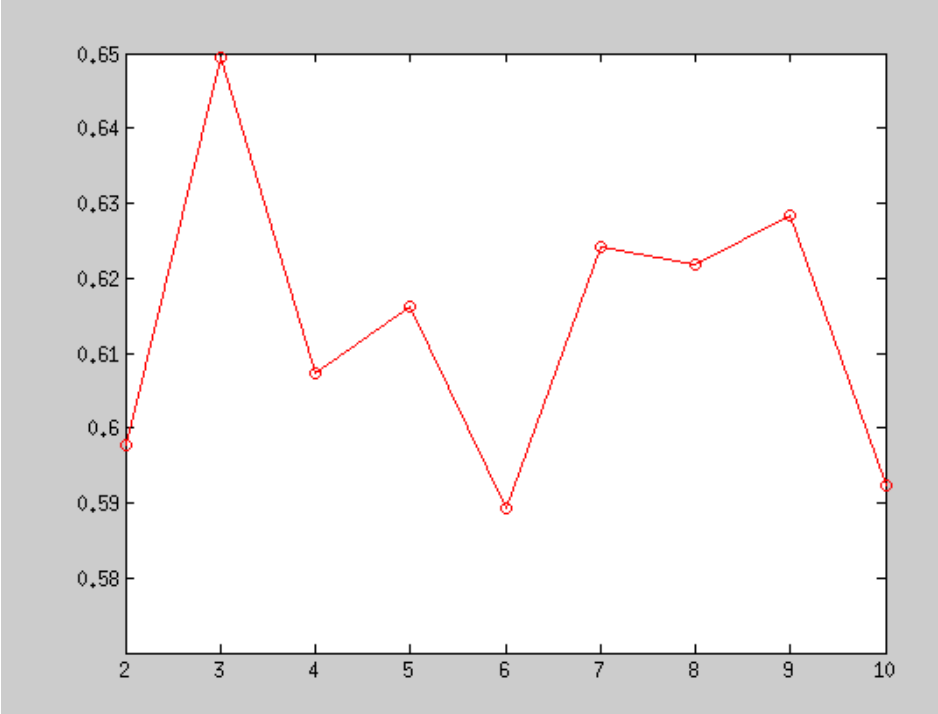


Figure 17: DBI score diagram of classification of (FA +FB) on elongation: The x-axis indicates the number of group; and y axis is the DBI score of clustering corresponding number of groups. This diagram shows the optimal number of groups is 2.

4 Result

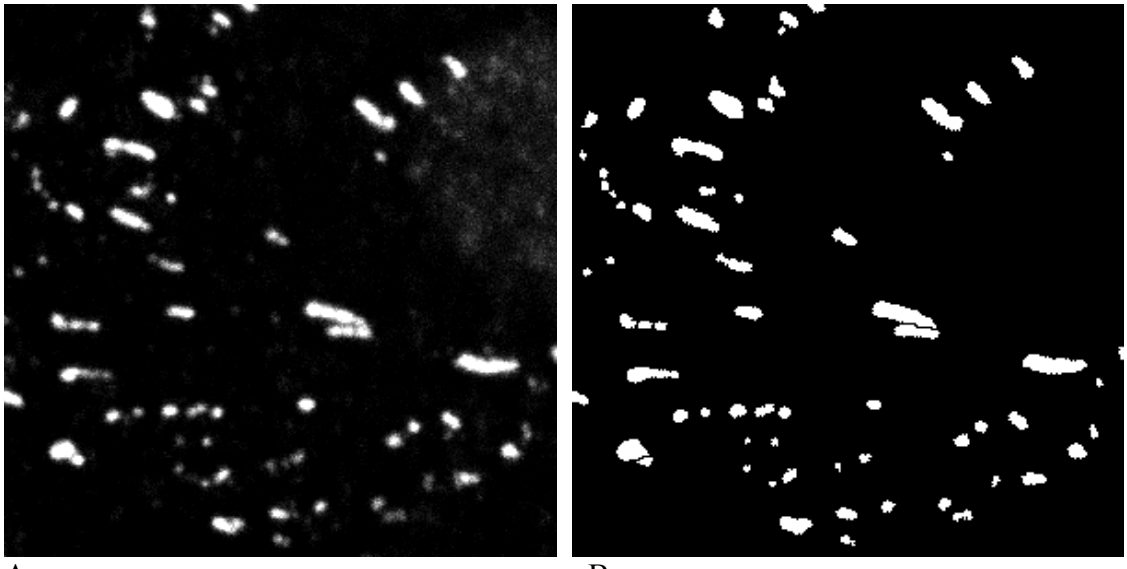
This chapter shows the result of optimization of segmentation methods in section 4.1, optimization of microscope setting in section 4.2 and image analysis in section 4.3

4.1 The result from Segmentation optimization

This section consists of two parts, of which the first part illustrates the result from noise reduction methods optimization and the second part shows the optimization of segmentation methods.

4.1.1 The result from image noise reduction optimization

In our test, we firstly tried to find the best span size n for Median filter. The smallest span size is n = 1 pixel, of which the local region is defined as 3x3. The experimental result shows the bigger span size we selected, the more it influenced the size of adhesions (Figure 18). Median filter with span size 2 and 3 would largely decrease the size of adhesions on all test images. Especially the small adhesions could be smeared by Median filter (Figure 18-D). From the study on the histogram of size, we know these small adhesions count a big percentage in the whole population of adhesions. Therefore in the subsequent comparison of different noise reduction filters, we compared the Median filter of only span size 1 with Gaussian filter or no-filter used method.



A B

C D

Figure 18: Test image 1 from table 1. B: the reference binary mask. C: The binary mask from Median filter (span size=1) + Otsu Segmentation. D: The binary mask from Median filter (span size=2) + Otsu Segmentation. Compared with C, amount of small objects are removed. These small objected are indicated in the reference binary mask.

The first 8 images used the same confocal microscope with different settings, but the last two images were acquired from different microscope which produced images with much higher level of noise. Since optimization of σ for Gaussian filters is also depended on the noise level. Therefore we divided test images in two groups, and applied both Median filter and Gaussian filter separately.

The first 8 images:

We applied both Median filter and Gaussian filter on the test images. For Gaussian filter, we varied the σ from 1.0 to 0.0 and evaluate the result through comparing its CDF with reference by KS test. From the result we noticed the CDFs from Median filter are always below reference CDF, since Median filter filters out small objects. On the other hand, when we set the σ as 1.0, the CDF lines from Gaussian filter are below both reference CDF and CDF from Median filter (Figure 19), since it significantly increase the adhesions size. As we are tuning down the σ gradually from 1.0 to 0.0, the CDF line from Gaussian filter would rise up from the position under the reference; then it would coincide with the reference CDF and continue rising until arriving CDFs from no filter used image. In this manner we expect that by adjusting the σ we could find a best value for Gaussian filter that gives closer CDF to reference CDF than Median filter and no filter.

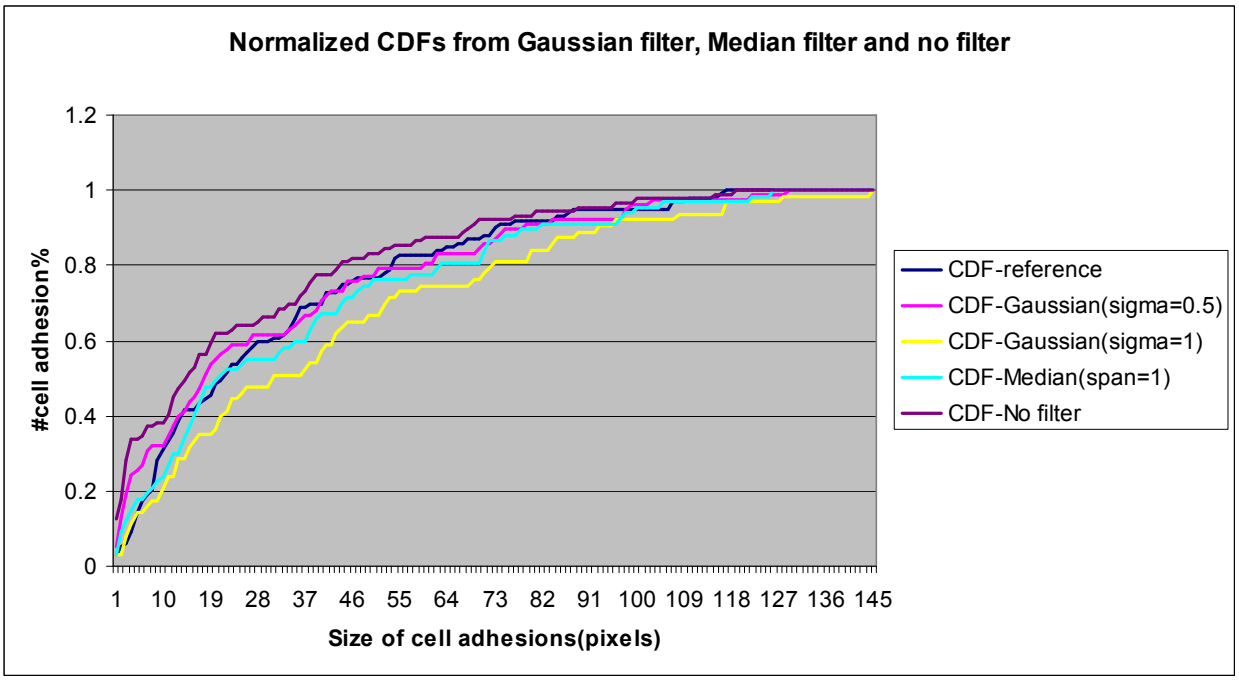


Figure 19: This is a Normalized CDFs chart from test image 3.

From the KS test, we proved that the Gaussian filter with σ = 0.5 is the best, compared with other σ value, Median filter and using no filter. Table 4 represents the statistics result of comparing Gaussian filter (σ=0.5), Median filter (span =1) and using no filter with reference on D_n and *p*-value. In this table, except image 2, 7 out of 8 images show that Gaussian filter gives closer result to the reference, and their corresponding *p*-values are almost all above significant level α = 0.05, except image 2 and 6. It means after Gaussian filter with sigma=0.5, the size distribution would not be influenced significantly. On the contrary using no filter shows the worst performance among these three methods, and only one *p*-value is larger than 0.05. It illustrates that even for those relatively lower noise level of images (compared with image 9 and 10), noise reduction is quite necessary.

Image name	D_n - Gaussian	*p*-value - Gaussian	D_n - Median	*p*-value - Meidan	D_n - No filter	*p*-value - No filter
1	0.2728	0.054	0.3110	0.013	0.3884	0.001
2	0.2392	0.001	0.1610	0.089	0.2705	0.000
3	0.2149	0.056	0.3567	0.014	0.3462	0.000
4	0.2642	0.096	0.2945	0.083	0.2968	0.084
5	0,2039	0,135	0,3088	0.000	0,3209	0.000
6	0,2710	0.000	0,2818	0.000	0,3232	0.000
7	0,2011	0,192	0,2214	0,162	0,3575	0,001
8	0,2517	0,137	0,2707	0,010	0,2766	0,001

Table 4: The comparison of D_n and p-value from Gaussian filter, Median filter and no filter processed images with reference. The image name is the same as the image index in table 1.Red masks mean bigger D_n and smaller *p*-value, which is the better candidate in corresponding test image.

Image 9 & 10

The statistics in table 4 shows that noise reduction is rather necessary. As we demonstrated before, in image 9 and 10 the noise level is even higher than the noise in previous 8 images. Therefore we did only consider the solution with filter.

It is quite likely that the Gaussian filter with $\sigma = 0.5$ could not blur away relatively big noise. Hereby we update the scale of trial σ to 0.0 ~ 3.0 (0.5 for one step). For Median filter we still decide to use span size 1 from experimental result. After comparing the CDF charts and applying KS test, Gaussian filter with $\sigma = 1.0$ shows better result on both images. The comparison of Gaussian filters (σ =1.0) and Median filter is following:

Image name	D_n - Gaussian	*p*-value - Gaussian	D_n - Median	*p*-value - Meidan
9	**0.2991**	**0.077**	**0.3086**	**0.026**
10	**0.2941**	**0.041**	**0.3482**	**0.006**

Table 5: The comparison of D_n and p-value from Gaussian filter and Median filter with reference. The image name is the same as the image index in table 1.Red masks mean bigger D_n and smaller *p*-value, which is the better candidate in corresponding test image.

4.1.2 The result of comparing segmentation methods

In this test, we still use the same test images presented in Table 1. Following segmentation methods are tested:

1. Global Otsu segmentation,
2. local isodata segmentation,
3. local isodata segmentation combined with watershed algorithm,
4. local Otsu segmentation,
5. local Otsu segmentation combined with watershed algorithm,
6. masked watershed segmentation,
7. region growing segmentation of which similarity criterion is intensity gradient,
8. region growing combined with watershed algorithm,
9. canny edge detection
10. canny edge detection combined with watershed algorithm.

For all the segmentation methods, we set a minimum intensity threshold as 10. This value is obtained experimentally. Pixels with intensity above this threshold participate the segmentation, otherwise are considered as background directly.

Due to heterogeneous distributed staining in the cell-matrix adhesion (Figure 12) and noises presented in the object, there are more than one intensity maximum in an adhesion. Consequently, watershed algorithm would over segment cell-matrix adhesions (Figure 20). To avoid this, we develop a method which uses two layers of image smoothing

1. Image smoothing for binary mask: This is for image noise reduction. Gaussian filter with optimized σ is used to smooth the raw image.
2. Image smoothing for retrieval of watershed line: The purpose of this smoothing is to get the watershed line to refine the binary mask obtained from step1. As stated in last paragraph, this step is required to blur heterogeneous distributed staining in the cell-matrix adhesion, smoothing of raw image is applied by Gaussian filter with bigger σ.

The procedure for two layers of watershed segmentation is listed as following:

1: Apply optimized noise suppression filter on original image I, we get more clear images I'.
2: Apply segmentation on I', we get binary mask B.
3: Use Gaussian filter with higher σ to blur I, after "rolling ball" we get very smoothed image I''.
4: Apply watershed algorithm on I'', we get images L with only watershed lines.
5: Use the lines in L to separated objects in B, in the end we get binary mask B', from which we do feature measurement.

For two layers watershed masked segmentation, the procedure is:

1: Apply Gaussian filter with higher σ and "rolling ball" on original image I to get smooth image I'.
2: Apply optimized noise suppression filter on raw image I to get low level noise of image I''.
3: Watershed algorithm is applied on I', subsequently watershed lines is used to cut I'' into regions.
4: Local adaptive segmentation is performed on each region. One binary mask image B can be obtained.

Figure 20: The binary mask from applying local Otsu segmentation combined with watershed algorithm on image in figure 18-A.

Here comes a question: which value of σ is the best for image smoothing to retrieve the watershed lines. Too big σ would have a lot connected objects; too small σ will induce over segmentation. For finding the best σ, we use the first 4 images from table 1 to test it. For each image, we applied different two layers segmentation methods with different σ. In the end we compared the result to reference by KS-test. The test result is following:

Image name	Statistics	Local Isodata+watershed(1)	Local Isodata+watershed(2)	Local Isodata+watershed(2.5)
1	D_n	0.2848	0.1829	0.1975
1	*p*-value	0.0030	0.1540	0.1030
2	D_n	0.1673	0.1857	0.2076
2	*p*-value	0.0130	0.0060	0.0000
3	D_n	0.1715	0.1551	0.1610
3	*p*-value	0.0780	0.1790	0.1080
4	D_n	0.1402	0.1316	0.1494
4	*p*-value	0.3900	0.4050	0.2640

Image name	Statistics	Local Otsu+watershed(1)	Local Otsu+watershed(2)	Local Otsu+watershed(2.5)
1	D_n	0.3009	0.1717	0.1654
1	*p*-value	0.0020	0.2380	0.2840
2	D_n	0.1534	0.1647	0.1724
2	*p*-value	0.0310	0.0200	0.0090
3	D_n	0.1743	0.1452	0.1689
3	*p*-value	0.0700	0.2360	0.1170
4	D_n	0.1373	0.1016	0.1221
4	*p*-value	0.3350	0.7450	0.4460

Image name	Statistics	Masked Watershed(1)	Masked Watershed(2)	Masked Watershed(2.5)
1	D_n	0.2294	0.1574	0.1648
1	*p*-value	0.0340	0.3450	0.3080
2	D_n	0.1647	0.1952	----------
2	*p*-value	0.0180	0.0050	----------
3	D_n	0.2424	0.1533	0.2020
3	*p*-value	0.0050	0.2020	0.0490
4	D_n	0.1945	0.1384	0.1648
4	*p*-value	0.0660	0.3840	0.2140

Image name	Statistics	Region growing+watershed(1)	Region growing+ watershed(2)	Region growing + watershed (2.5)
1	D_n	0.1468	0.1266	0.1378
1	*p*-value	0.3790	0.6310	0.5370
2	D_n	0.2810	0.3200	0.3780
2	*p*-value	0.0000	0.0000	0.0000
3	D_n	0.1873	0.1818	0.1984
3	*p*-value	0.0390	0.0670	0.0110
4	D_n	0.1287	0.0955	0.1010
4	*p*-value	0.4140	0.8290	0.6840

Table 6-9: Each table is a KS test on certain segmentation method combined with watershed segmentation, which is mentioned in the table titles. The number in parentheses is the value for σ. The red marks label the smaller distribution distance D_n and larger *p*-value, which means the better result in relative row. ------ means KS test is not performed with corresponding parameter.

The KS-test for global Otsu, canny edge detection is not presented here. Tables 4 and 5 have already shown global Otsu segmentation give much worse result than other segmentation methods. For canny edge detection, the *p*-values for all the σ values are 0. It supposed that canny edge detection is still too sensitive to the noise and the heterogeneous distributed staining. Especially for heterogeneous distributed staining, it could produce inner edge within an object.

From table 6-9, we can easily tell that the Gaussian filter with σ=2 for watershed method is the best choice for most of cases. σ=2.5 would already induce under segmentation problems. Thus we decided to set σ=2 for the later test. Now we use all 10 test images. We apply different segmentation methods on each image. KS test is still used to compare their performance of which criterion is the distance between size distribution got from segmentation and size distribution of reference. Another criterion is brought into this test, which is called binary error. The formula is:

Binary error = XOR (reference binary mask, binary mask from segmentation)/pixel_number (reference binary mask)

The purpose is to calculate how many pixels labeled differently between two binary masks. More precisely speaking, it tests how precisely the binary mask from

segmentation covers the reference binary mask. KS test can only test how much one operation changes the distribution. In some cases, the distribution is not affected significantly, but the binary masks are totally different. Based on those two tests, we expect we can find the best segmentation method for each microscope settings.

The test result is presented in the table 10. Firstly the segmentation region growing and region growing combined with watershed are discussed. In some images, for instance number 4 and 9, even though region growing shows very small distance between its obtained size distribution and reference, and very high *p*-value, the binary error is rather high. In other images like image 2, both binary error and KS test proves that region growing does not give a good segmentation result. Especially *p*-value, for some images, it reaches 0. For all those reasons, we supposed that the region growing method does not work stable on our images.

We also found in the first 4 images, Local Otsu segmentation gives the best performance compared with others. Both KS-test and binary error demonstrate its higher accuracy. Masked watershed segmentation show slightly better result than Local Isodata segmentation, and has almost same accuracy as Local Otsu segmentation. Especially on the binary error, image 3 and image 4 all show that masked watershed has lowest error among all other segmentation methods. From computational time of view, in our test, masked watershed algorithm was much faster than local Otsu segmentation. Therefore, for this group of images, we supposed that masked watershed segmentation is one of the best segmentation methods.

For last 6 images of which pixel length is larger than previous four images, masked watershed methods shows more significant advantage than other methods, even better than Local Otsu segmentation. All 6 test images testify that masked watershed segmentation effects size distribution at least and it also has smallest binary error. Moreover, except images 6, all other 5 *p*-values from masked watershed segmentation are higher than 0.05, which means even on relatively higher level noise of images masked watershed segmentation still gives good estimation of size of cell-matrix adhesion. According to this we could draw a conclusion that masked watershed segmentation has very strong ability and high robustness on noisy images. Oppositely, the performance of other segmentations is influenced by the noise level of images more severely than masked watershed segmentation.

Image name	Statistics	2	3	4	5	6	7	8
1	D_n	0.2083	0.1829	0.1414	0.1717	0.1574	0.1705	0.1266
1	*p*-value	0.0870	0.1540	0.5080	0.2380	0.3450	0.2970	0.6310
1	BE	0.3330	0.3430	0.3100	0.3400	0.3220	0.4510	0.4630
2	D_n	0.1908	0.1857	0.1643	0.1647	0.1952	0.2904	0.3200
2	*p*-value	0.0060	0.0060	0.0270	0.0200	0.0050	0.0000	0.0000
2	BE	0.3930	0.3910	0.3630	0.3930	0.3890	0.7010	0.6590
3	D_n	0.1604	0.1551	0.1535	0.1452	0.1533	0.1946	0.1818
3	*p*-value	0.1650	0.1790	0.1980	0.2360	0.2020	0.0580	0.0670
3	BE	0.2580	0.2680	0.2640	0.2560	0.2480	0.5240	0.5220
4	D_n	0.1590	0.1316	0.1394	0.1016	0.1384	0.1260	0.0955
4	*p*-value	0.2320	0.4050	0.4010	0.7450	0.3840	0.5310	0.8290
4	BE	0.4300	0.4410	0.3890	0.3690	0.3090	0.5810	0.5910
5	D_n	0.2127	0.1790	0.2182	0.1827	0.1789	0.2394	0.2424
5	*p*-value	0.1220	0.2210	0.0990	0.1950	0.2290	0.0540	0.0340
5	BE	0.3680	0.3710	0.4380	0.4110	0.3240	0.6690	0.6580
6	D_n	0.2637	0.2121	0.2256	0.2288	0.2034	0.4114	0.3852
6	*p*-value	0.0000	0.0020	0.0020	0.0010	0.0030	0.0000	0.0000
6	BE	0.4150	0.4780	0.3970	0.3830	0.3020	0.7090	0.6110
7	D_n	0.1811	0.1767	0.1850	0.1692	0.1599	0.2348	0.2819
7	*p*-value	0.2340	0.3000	0.2900	0.4610	0.5200	0.0700	0.0010
7	BE	0.3980	0.4170	0.3600	0.3810	0.3250	0.5190	0.5520
8	D_n	0.2117	0.2386	0.1986	0.1984	0.1779	0.2901	0.3114
8	*p*-value	0.1700	0.1640	0.2200	0.2500	0.3180	0.0010	0.0000
8	BE	0.4200	0.4670	0.4120	0.3840	0.3770	0.4620	0.5000
9	D_n	0.1824	0.1600	0.1923	0.1790	0.1555	0.2890	0.3088
9	*p*-value	0.1000	0.1120	0.0700	0.1050	0.2300	0.0240	0.0180
9	BE	0.4800	0.4410	0.4290	0.3800	0.3320	0.4910	0.5200
10	D_n	0.1109	0.1283	0.0941	0.1040	0.0888	0.0943	0.0812
10	*p*-value	0.5410	0.4900	0.8300	0.7210	0.8950	0.8000	0.8900
10	BE	0.4210	0.4140	0.3980	0.3920	0.3330	0.7520	0.6070

Table 10: The KS test result and Binary Error (BE) when we compare the binary mask from different segmentations and reference binary mask. In the table title, from the third column, we use the number presented in page 41 to represent segmentation methods.

4.2 The result from microscope setting optimization

In this test, KS statistic D_n and *p*-value of the size distribution are still the criterion to evaluate which image condition would give the smallest influence on our measurement. As we have already demonstrated that masked watershed segmentation gives best accuracy and highest robustness on all available image condition, all images are segmented by this method.

The first group of images is shown in table 2. The purpose of comparing these images is to test which pixel length gives most close cell-matrix adhesion size distribution to the reference. The test result is shown as following (Table 11). Notice the image pairs 1 and 3, 2 and 4 are actually from same position but from images with different pixel length. Both pairs of image show pixel length = 0.155μm gives better result compared with the reference, even though the difference is not very big.

The second group of test is to test which averaging time gives more accurate size distribution to reference distribution. Both pairs of images (5 and 7, 6 and 8) testify that the higher averaging gives more accurate measurement. However when we calculated the distance between size distribution from no averaging image and from 4 time averaging, the D_n for the first pair images (5 and 7) is 0.1102 and corresponding *p*-value is 0.658; the D_n for the second pair of images , 6 and 8, is 0.1302, and the *p*-value is 0.141. Both *p*-value are higher than significant level 0.05. This could illustrates that 4 times of averaging would not improve the size distribution to a significant level. 0 or 2 times averaging are sufficient enough to give a reliable size distribution.

The third group of test focuses on the effect of zoom. To our surprised, the 4x zoom show significant difference compared with 6x zoom. The *p*-value for 4x zoom are even 0, which means both images are out of focus with 4x zoom lens.

Image name	D_n	*p*-value
1	0.1574	0.345
2	0.1952	0.005
3	0.1789	0.229
4	0.2034	0.003

Table 11: the KS test result of different pixel size. Image 1 and 2's pixel size is 0.155μm, the rest images' resolution is 0.311 μm.

Image name	D_n	*p*-value
5	0.2208	0.053
6	0.2211	0.002
7	0.1574	0.345
8	0.1952	0.005

Table 12: the KS test result of different averaging. First two images have no averaging, and last two images have 4 times averaging.

Image name	D_n	*p*-value
9	0.3990	0.000
10	0.3668	0.000
11	0.1574	0.345
12	0.1952	0.005

Table 13: the KS test result of different zoom. The zoom lens for the first two images is 4x, and the zoom lens for last two images is 6x.

4.*3 The result of image analysis*

This paragraph illustrates the result of morphological analysis of adhesion and the identification of hits. Paragraph 4.3.1 focus on the result of features evaluation; hits identification and hits analysis are shown in paragraph 4.3.2.

4.3.1 Evaluation of Features

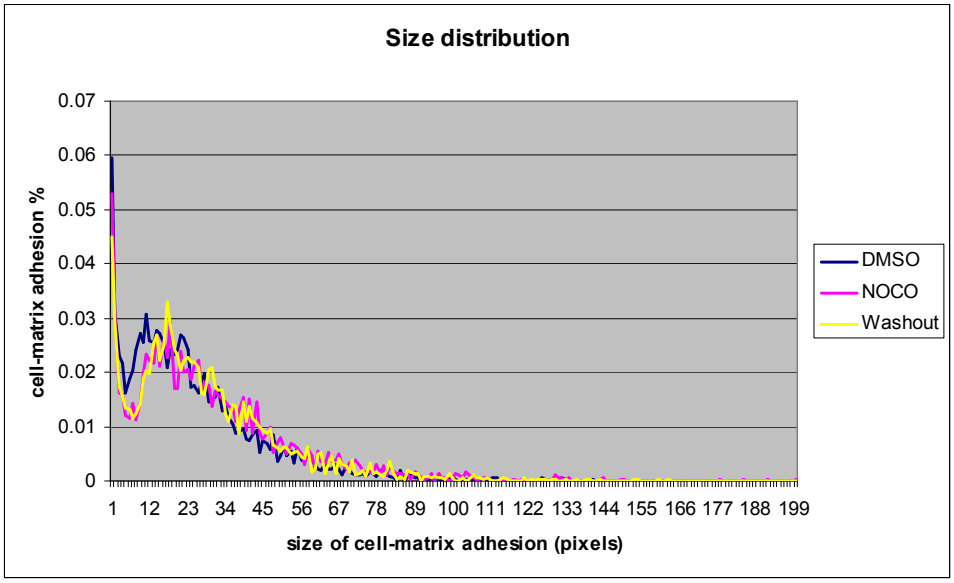
From learning the distribution of each feature in both relatively higher quality images and human kinase siRNA images, a table about how these distribution changes according to different exposure conditions is built up (Table 14). We found that in higher quality of images Noco condition would increase the size of cell-matrix adhesion and WO could drop back the size distribution to certain level (Figure 21). This is already tested experimentally. In chapter 3.1, it is explained that Noco would induce the depolymerizing of MT. Consequently the adhesion would grow in size. After washing out Noco, MT forms again and targets focal adhesion which leads on their disassembling.

Features	Image group	K (DMSO vs Noco)	K (Noco vs WO)
size	Higher quality	1	-1
	Human Kinase	1	-1
perimeter	Higher quality	1	-1
	Human Kinase	1	-1
extension	Higher quality	1 or -1	-1
	Human Kinase	0	----------
dispersion	Higher quality	1 or 0	1 or -1
	Human Kinase	1 or -1	0 or 1
elongation	Higher quality	-1	1
	Human Kinase	-1	1
Compactness	Higher quality	1	-1
	Human Kinase	1	-1
Average Intensity	Higher quality	0	0
	Human Kinase	1	----------
Nucleus distance	Higher quality	-1	0
	Human Kinase	0	----------
Closest FA	Higher quality	0	0
	Human Kinase	-1	----------

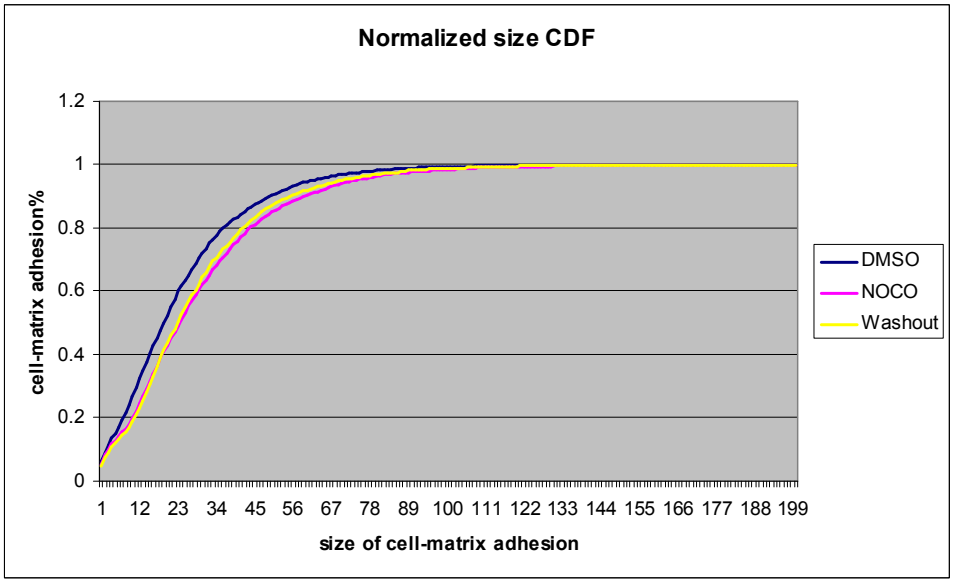
Table 14: The distribution change under different exposure conditions are learned on both higher quality image group and samples from human kinases siRNA screening plates. ------- means that we did not test it because the same row of KS (DMSO vs Noco) is already different from that of higher quality images.

The direction of change is predicted by one-tailed KS test. K = 0 means there is no significant change between two CDF (α=0.05); K=1 presents that the second CDF is lower than the first CDF in a significant level; K=-1 indicates that the second CDF is higher than the first CDF in a significant level. K(DMSO vs Noco) means the one tail KS test result between corresponding feature's CDF from DMSO (the first CDF) and that from Noco; When there is only one number as KS test result in one blank, it means that

two image groups have the same test result. If there are two numbers, for instance "extension"-"higher quality"-"K (DMSO vs Noco)", the first number is from no-siRNA control group and the second number is the test result on #2 siRNA control group.



A

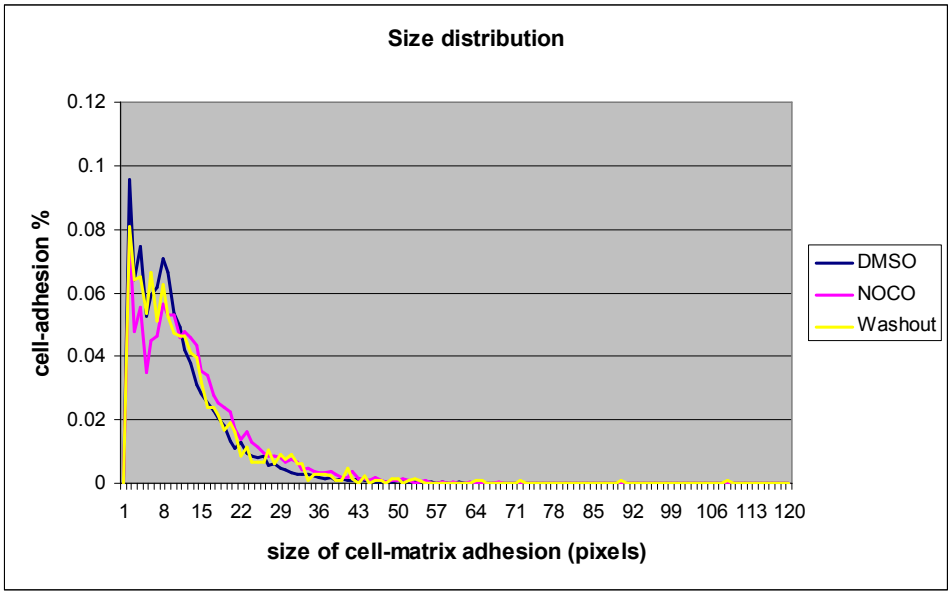


B

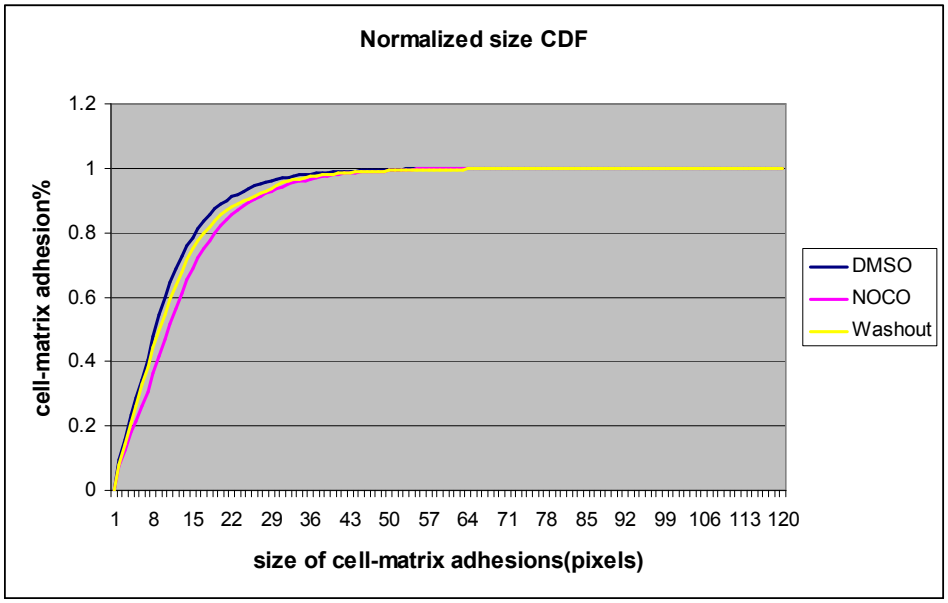
Figure 21: Both size distribution (A) and size CDF (B) of siRNA #2 control group (from higher quality images) are presented. The unit of size is pixel. The dark blue line represents the distribution of size obtained from DMSO exposure condition. The pink line describes the distribution from Noco condition. The yellow line corresponds to WO condition.

In the human kinases siRNA screening plates, the phenomenon of size increasing in Noco condition and it is pulled back in Washout were both observed from both control groups of images (Figure 22). This demonstrates that size measurement was not influenced

according to these two image condition significantly. Therefore it is valid to be used to discover the hits.



A

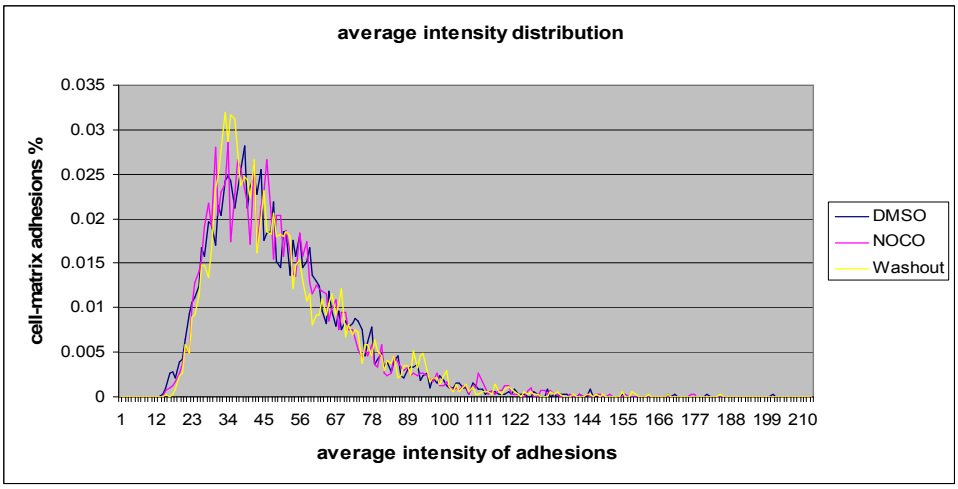


B

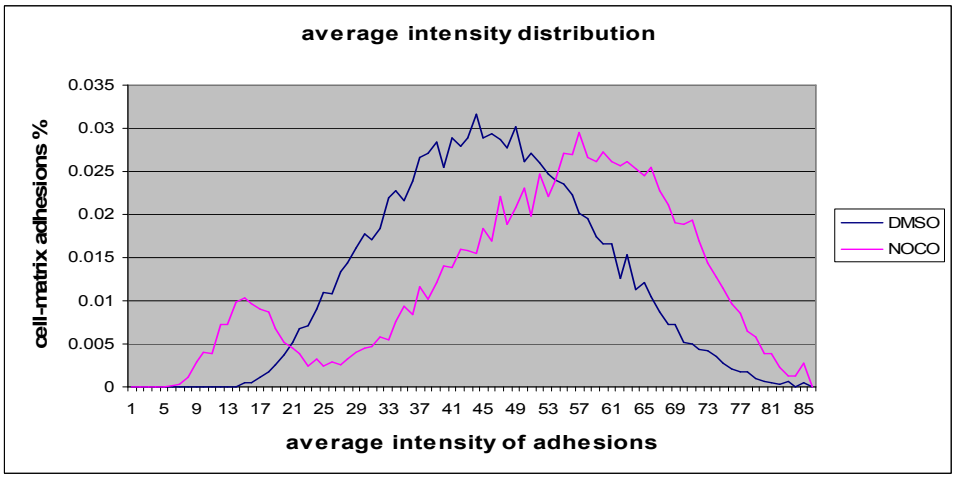
Figure 22: size distribution (A) and size CDF (B) of siRNA #2 control group (from Human Kinase siRNA screening) are presented. The unit of size is pixel. The dark blue line represents the distribution obtained from DMSO exposure condition. The pink line describes the distribution from Noco condition. The yellow line corresponds to WO condition.

Same as size, elongation, compactness and perimeter all show the same difference of distribution on both higher quality image groups and Human Kinase screen. In Noco situation, the distribution of compactness and perimeter are all shifted to the right compared with its corresponding DMSO condition. The distribution of elongation shifts to the left. Then WO pulls them back closer to DMSO condition.

In higher quality images, the distribution of average intensity of adhesions was not changed significantly between different exposure conditions (Figure 23-A and table 14). However, the distributions of intensity in Human Kinases siRNA screening images were influenced significantly by exposure conditions (Figure 23 -B).



A



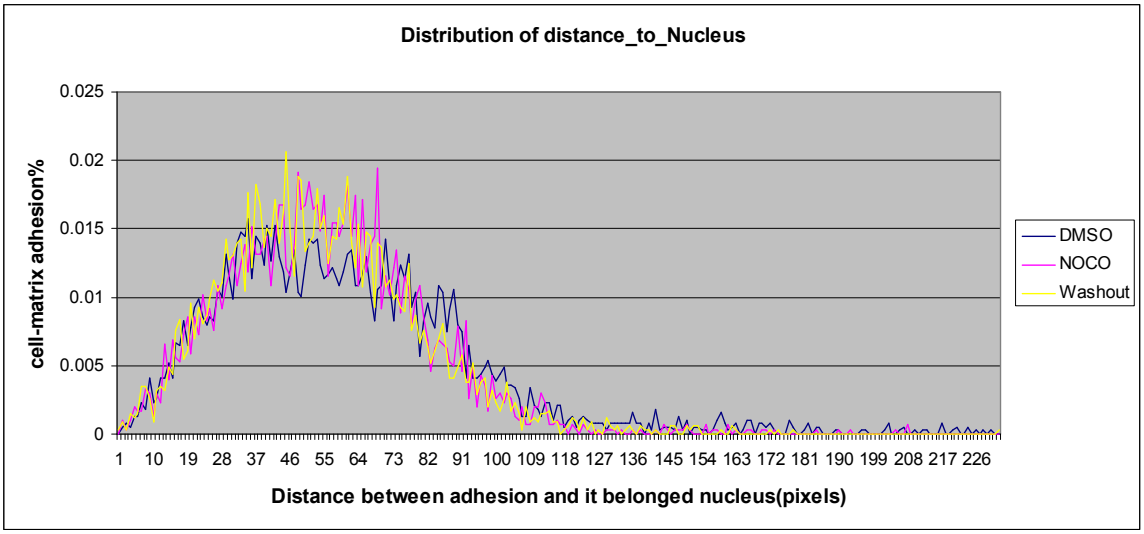
B

Figure 23: ***A is the distribution of adhesions' average intensity of control group #2 siRNA according to different exposure conditions.*** They are from higher quality image. ***B is distribution of average intensity from control group siRNA #2 of Human Kinases siRNA screening.*** Dark blues is distribution under DMSO condition and pink line is from Noco condition. The distribution of Washout is not presented here.

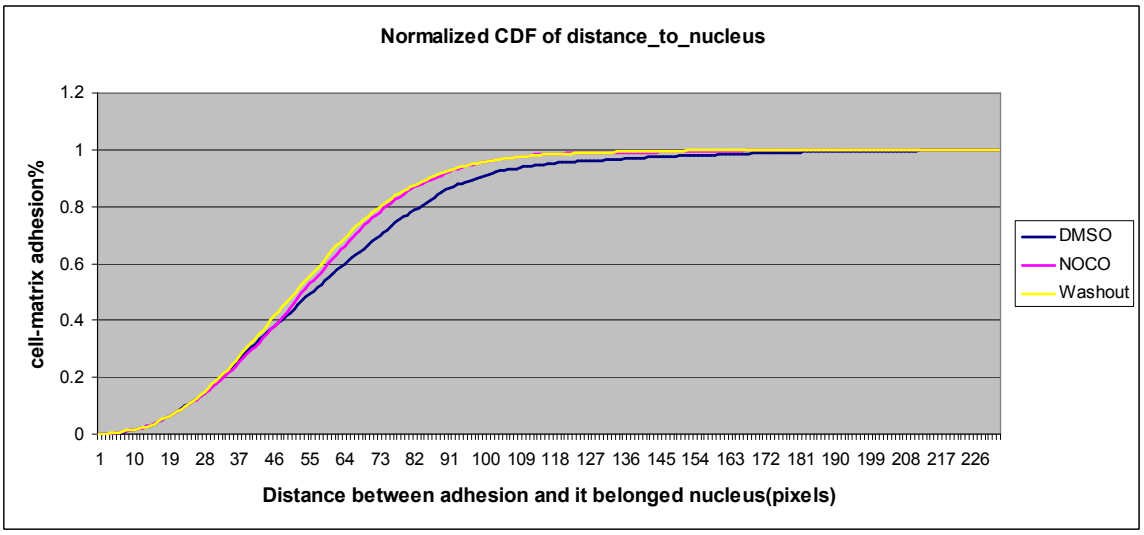
Same as average intensity of adhesions, feature ClosestFA was not affected a lot by different exposure condition in higher quality images but showed significant changes in human kinases siRNA screening images.

The situation in Nucleus distance is opposite. In higher quality images, the distribution of distance_to_nucleus is more uniform distributed in DMSO, compared with it is in Noco. (Figure 24). In the experiment, we observed that in DMSO culture adhesions distributed uniformly inside the cell. But in Noco condition of culture, adhesions are more located at

the edge of cells. We supposed that in the Noco condition, depolymerizing of MT increases the tension inside the cell which pushes the adhesions to the edge of cell. By contraries, in human kinases siRNA screening plates' control group, the change in distribution between DMSO and Noco is not obtained (Figure 25).



A



B

Figure 24: Both distribution of distance_to_nucleus (A) and CDF of distance_to_nucleus. (B) of #2 siRNA control group (from higher quality images) are presented.

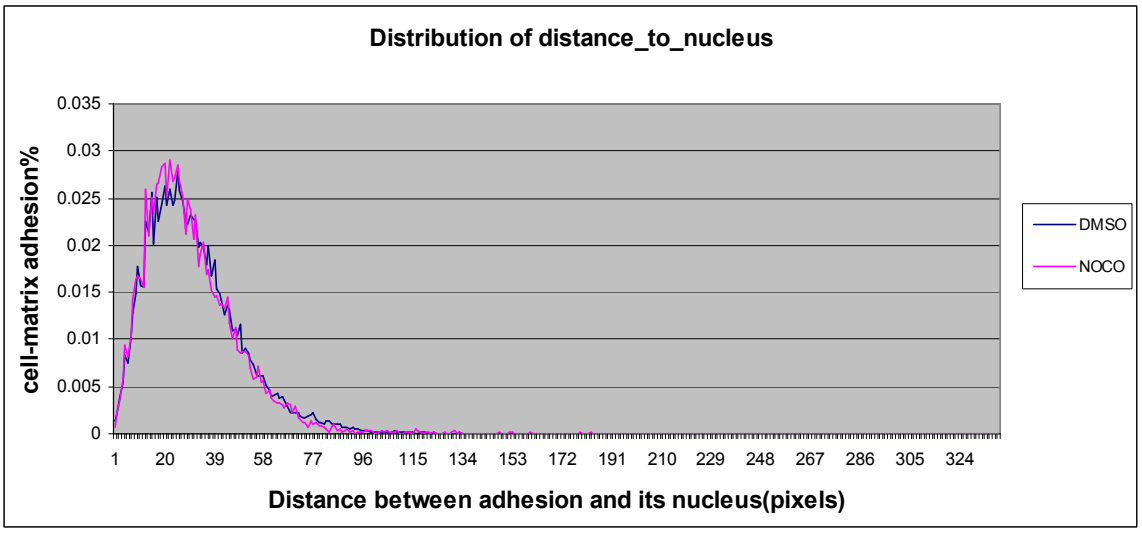


Figure 25: The distribution of distance_to_nucleus from control group siRNA #2 of Human Kinase library plates.

For other features, orientation, gravity center of object they are all depended on the imaging orientation, position. Thus we only used them for hits analysis not for hits discovery.

From feature evaluation, we could draw a conclusion that measurement of size, perimeter, compactness and elongation is more reliable on our Human Kinase screening. Therefore they could be used to explore the hits.

4.2.2 Hits

Control group quality evaluation

From control group quality evaluation we find in all the plates of 10 experiments, no-siRNA gave more stable behavior than #2 siRNA control group. In some plates, #2 siRNA control Noco wells did not give significant size increasing in contrast to its control DMSO wells of the same plates, while no-siRNA control Noco shows significant size increasing. However, when the no-siRNA control Noco was not valid on certain plate, the #2 siRNA control NOCO is not valid neither. For this reason, we used no-siRNA control wells against treated wells.

In experiment 7, the no-siRNA control wells in all the plates are empty, so we use siRNA #2 control wells against treated wells.

Hits discovery

The automated analysis was applied, based on KS test for control cells against treated cells. For each validated feature (size, perimeter, elongation, and compactness), top 10 siRNA which severely affect cell-matrix adhesions were selected as hits. The hits from size, perimeter and compactness were largely overlap due to these features are closely

correlated to each other. To our surprise, in some experiments, the hits of size from "control DMSO vs treated DMSO" and "control NOCO vs treated NOCO" (Cf. Figure 15) were overlapped to a big extent. The direction they change the distribution were also consistent. Here we show the statistics of comparing control DMSO wells with siRNA treated DMSO wells and the statistics of comparing control Noco wells with siRNA treated Noco wells on experiment 4 (Table 15).

control DMSO vs treated DMSO	control NOCO vs treated NOCO
41	41
18	42
52	10
58	58
53	11
5	53
13	56
42	52
3	51
46	13

Table 15: This table shows part of hits of experiment 4. Those hits are identified on the feature size. The number indicates the index of siRNA. See Appendix Two. In 10 hits, 6 hits are shown in both columns. The hits presented in both comparisons are marked as light brown.

Hits analysis

After finding the fits, we want to analysis how these hits affect cell-matrix adhesion. The first important field is that we want to learn how the location of cell-matrix adhesions changes. Section "Evaluation of Features" has already indicated that the measurement of distance_to_nucleus is not reliable on our human kinases screen. Thus the only measured feature which has the information of location of adhesions is a binary descriptor: InNucleus. For each hit, we calculated the percentage of adhesions which are in the nuclei on both control cells and treated cells, for instance number 46 siRNA in the table 15, which represents siRNA FLJ10842. It is a hit obtained from comparison of control DMSO cultured cells and treated DMSO cultured cells. The one tail KS test reveals that the cell-matrix adhesions are significant smaller in the treated cells compared with control cells (Appendix two). We calculated InNucleus in control DMSO images, the percentage of adhesions which are inside the nuclei is 18.86%. Meanwhile this percentage increase to 29.232 % in FLJ10842 treated DMSO cells. We supposed this siRNA would speed up the disassembling of big focal adhesions which are located on the edge of cell. The separated complexes would move in the direction of nuclei.

For hit which is from comparing its DMSO condition and its Noco condition or comparing itself Noco condition with corresponding WO condition, the percentages of inside nuclei adhesion are computed for both conditions. For example MASTL is a siRNA hit when we compared the size CDF between MASTL treated Noco condition and MASTL treated WO condition. After washout of Noco, the size of adhesion was not decreased to the contrast of Noco. Instead, the size was even increased by this siRNA. It could predict that MASTL would affect WO processing, MTs are probably avoided to re-

grow or target focal adhesions and this siRNA may even boost up the growth of focal adhesion. Comparing the percentage of In-Nucleus adhesions, in MASTL treated Noco condition the percentage is 31.299% and this number is 24.772% in treated WO condition, which means the percentage of adhesions inside nuclei gets lower in the WO condition. Subsequently we can presume that the compounds from nuclei are assembling continually to bigger focal adhesions and moving to the edge of the cells..

Except for the change of location, the influence of siRNA on the different type of cell-matrix adhesions was learned. We classified the adhesion into three groups based on two layers of classification. The size and elongation threshold were trained from each plate's control DMSO condition, thus varied on different plates. The distribution of each type of cell-matrix adhesions is defined as Di = percentage of type i adhesion among the whole distribution, where i=1, 2, 3, which represent FC, FA, FB respectively.

Here we give an example how we analyzed the influence of hits on different type of cell-matrix adhesions; Still use hits 46 FLJ10842 (Cf. Appendix Two) as an example. After classifying all the cell-adhesion from FLJ10842 treated wells into three groups, we found 26.741% of cell-matrix adhesions are FC. Compared with control DMSO wells in the same plate, which has only 13.99% of FC in the whole population, the number increases significantly. Oppositely, focal adhesion (FA) number is much smaller than that from control group: 58.516% vs 71.771%n. This phenomenon can explain why one tail KS test shows that this hit decreases the adhesion size to a significant level: More focal adhesion disassemble into smaller focal complexes. However, FB is not affect a lot by FLJ10842. In control group, we get 14.239% of FB. In FLJ10842 treated group, the distribution of FB is 14.743% which is nearly the same as from control group. For all these observation, we could presume this hit could enhance the focal adhesion disassembling ability, but not affect FB. Subsequently we could presume the mechanism involved into regulation of FA and FB are different.

5 Discussion & Conclusion

In this chapter, important topics of this project are discussed in paragraph 5.1. In paragraph 5.2, the major conclusion of this project will be presented.

5.1 Discussion

5.1.1 Minimum area of cell-matrix adhesions

The previous study of cell-matrix adhesions set a minimum area for adhesion. For instance research in [18] defined the range of area is: area > 3.33μm. The previous study of Division of Toxicology's on human kinases siRNA screening also set 5 pixels as the minimum threshold. Generally speaking, this threshold could help to remove small discrete noises. However we should notice that the setting of this threshold is based on two assumptions:

1. The size of noises must be smaller than the size cell-matrix adhesions.
2. There is a biological definition of minimum size for adhesions.

In our study due to low magnification and resolution (Cf. Appendix Four) we indeed observed the cell-matrix adhesions smaller than 5 pixels and it actually takes up a big percentage of whole population. In some images it even reaches 30%. Thus we did not set a minimum threshold for adhesion area.

5.1.2 Masked watershed segmentation vs global segmentation

In previous study [9] and [18], Global segmentation and Watershed segmentation were applied on images acquired from microscope with objective 60x/1.3 NA and microscope with higher NA which is 60x/0.9 NA. It is supposed that those two segmentation methods would give reliable binary mask on images of high standard objective. Here we compare the performance of masked watershed segmentation with those two segmentation methods on images of even better objective. One experimental image is acquired from microscope with objective 60x/1.4 NA. Subsequently we applied both watershed masked segmentation and these two segmentation methods on this image. The histogram of size obtained from each segmentation method is shown in figure 26.

From this figure, we can see the histograms from three segmentation methods are significantly different. The histogram from global segmentation and global + watershed segmentation are exponential- like distribution. But the watershed masked segmentation got smoother distribution with two peeks. This is consistent with the histogram shape from images of lower imaging quality.

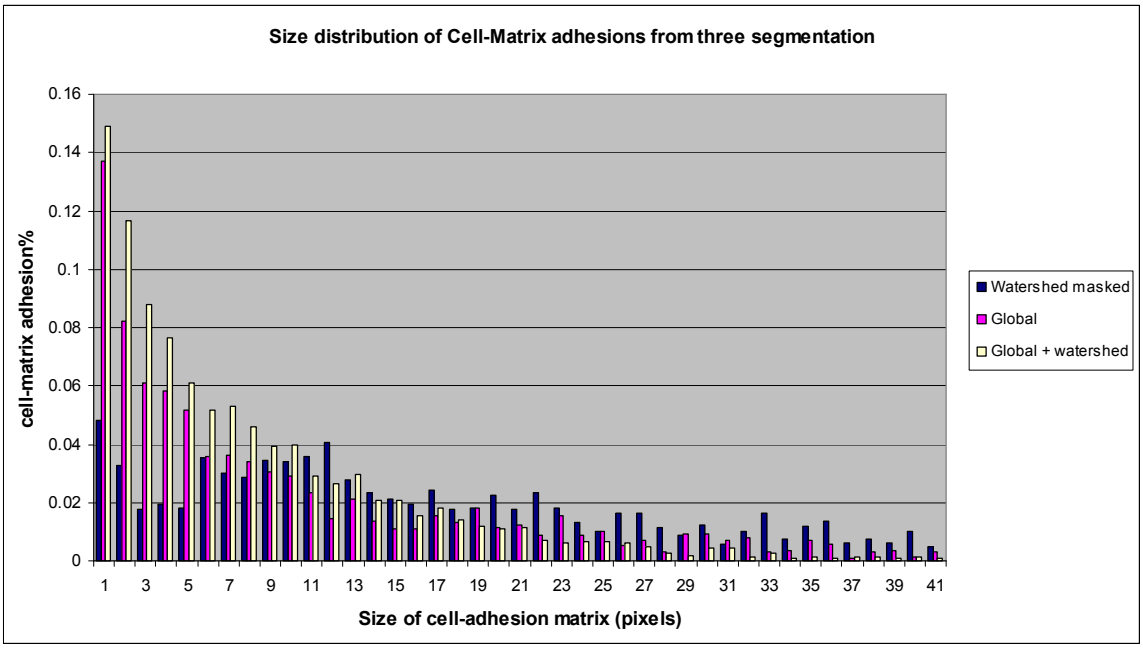


Figure 26: One part of size histograms from three segmentation method.

Actually in this high resolution (Cf. Appendix Four) of image, the heterogeneous distribution of fluorescence is more obvious (Figure 12). Therefore global + watershed segmentation method will seriously over segment the big focal adhesion. This explains the exponential- like distribution from this segmentation method. From our observation of image, we noticed that the discretization noises are still observed. Their intensity is higher than the weak and faint adhesions. In this case, global threshold would mislabel them as foreground. This is the reason why it has very high percentage of small objects. By contraries, masked watershed methods has already be tested that it is a stable segmentation methods on noisy data. The different shape of distribution from different methods indicates that even for images with objective 60x/1.4 NA, masked watershed segmentation is still a better option than global segmentation or global segmentation combined with watershed algorithm.

5.2 Conclusions

In this project, an image analysis protocol is established for automation of high-throughput cytomic screening analysis.

Step 1: Image preprocessing aims to remove the image noise
Step 2: Segmentation is performed to get the binary mask of cell-martix adhesion
Step 3: Morphological features are measured on the binary mask
Step 4: Features are evaluated and only valid features are used for the identification of hits
Step 5: Quality of control group is evaluated and only valid control groups are used for the identification of hits.
Step 6: Based on each valid feature, hits are extracted by KS test which gives the score of comparison of cell-matrix adhesions from treated wells and valid control wells.
Step 7: Hierarchical clustering is applied to cluster cell-matrix adhesion into FC, FA, and FB, based on two layers of clustering: size and elongation. How their distributions are influenced by hits is analyzed.

Moreover this project systematically evaluated the influence of segmentation method and microscope setting on measurement of cell- matrix adhesions. Different segmentation methods were applied on images under different microscope settings and their performance were compared with a reference. From the result it is shown:

- Global segmentation could not give a satisfied result.
- Masked watershed methods performed the best on both high noisy level of images or images with high resolution and high standard of objective.

This project also contributes to find optimal imaging conditions for cytomic screening. We apply watershed masked segmentation on images under different microscope settings. The cumulative distribution (CDF) of size from each image condition is compared with reference. The conclusions are:

- Even though there is heterogeneous distribution of fluorescence within cell-matrix adhesions, pixel length 0.155μm stills shows better performance than pixel length 0.311μm.
- 4 x averaging improves the result slightly and there is no significant difference between CDF from 4x averaging and from no averaging. However 4x averaging increases screening time 4 times as no averaging screening.
- 6x Zoom gives much better result than 4x zoom.

Appendix One: Software tool

SCIL_Image 1.4.1

SCIL_Image is an extensive multiple layered system for image processing and for the development of applications in the image processing domain. It combines user front-end environment SCIL and image processing libraries under the name of images. Figure 1 shows an overview. Compiled functions are interfaced through the library handler, which is practically invisible to the user. The first user visible layer consists of a C-interpreter. In the second layer a command expander working on image processing commands known to the system opens the way to efficient interactive (image processing) design, hiding the C-level from the novice and allowing shorthand typing for the experienced user. In the third layer, a window management system generates menus and dialogs for command selection, now hiding the command level from the naïve user. All interface layers are generated from a Command Description File (CDF) which insulates the interface layers from the application. From the C-interpreter, via the command expander level, to the menu and dialog generator, the system offers increasingly more user-friendliness but looses on flexibility and speed of interaction.

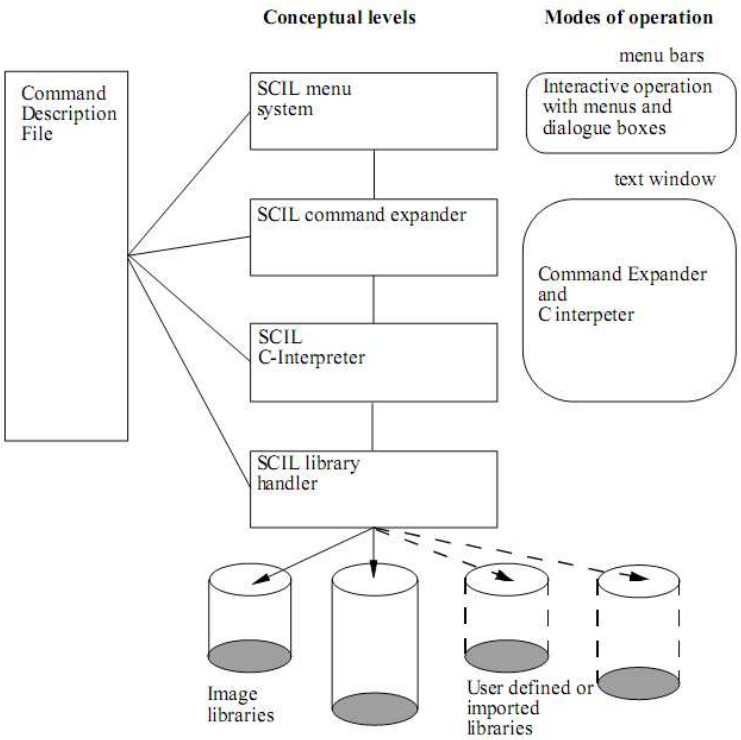


Figure 1: The interface layers of SCIL

ImageJ

ImageJ is a public domain, Java-based image processing program developed at the National Institutes of Health. ImageJ was designed with an open architecture that provides extensibility via Java plugins and recordable macros. Custom acquisition, analysis and processing plugins can be developed using ImageJ's built-in editor and a Java compiler. User-written plugins make it possible to solve many image processing and analysis problems, from 3-dimensional live-cell imaging to radiological image processing, multiple imaging system data comparisons to automated hematology systems. ImageJ's plugin architecture and built in development environment has made it a popular platform for teaching image processing. A lot plugins have been made available via http://rsbweb.nih.gov/ij/plugins/index.html.

ImageJ can display, edit, analyze, process, save and print 8-bit, 16-bit and 32-bit images. It can read many image formats including TIFF, PNG, GIF, JPEG, BMP, DICOM, FITS, as well as raw formats. ImageJ supports image stacks, a series of images that share a single window, and it is multithreaded, so time-consuming operations such as image file reading can be performed in parallel with other operations. ImageJ can calculate area and pixel value statistics of user-defined selections and intensity thresholded objects. It can measure distances and angles. It can create density histograms and line profile plots. It supports standard image processing functions such as logical and arithmetical operations between images, contrast manipulation, convolution, Fourier analysis, sharpening, smoothing, edge detection and median filtering. It does geometric transformations such as scaling, rotation and flips. The program supports any number of images simultaneously, limited only by available memory. Figure 2 shows a screenshot of ImageJ.

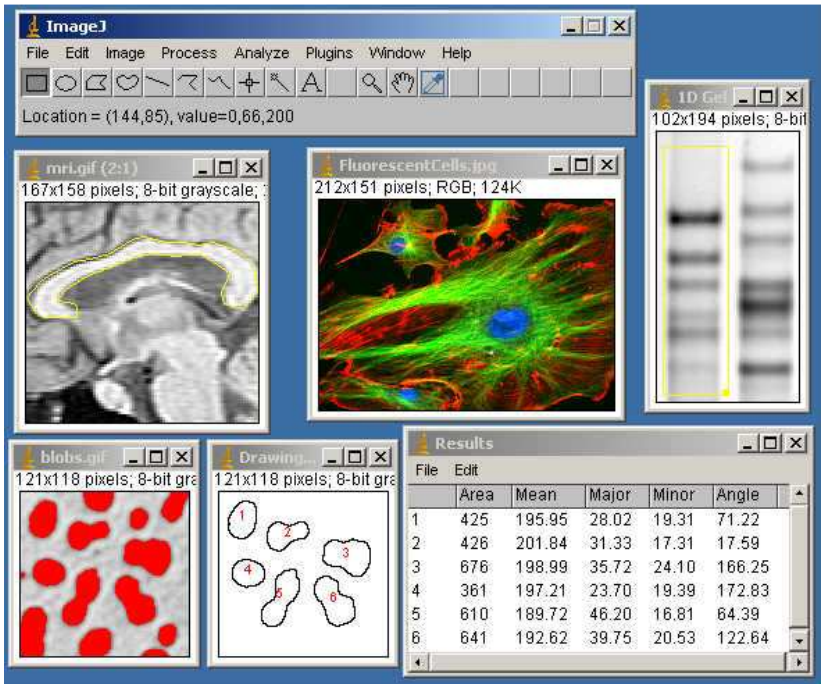


Figure 2: Screenshot of imageJ

MATLAB 7.0.1

MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. MATLAB can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran.

MATLAB can be used in a wide range of applications, including signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology. Add-on toolboxes (collections of special-purpose MATLAB functions, available separately) extend the MATLAB environment to solve particular classes of problems in these application areas.

MATLAB provides a number of features for documenting and sharing work. The MATLAB code can be integrated with other languages and applications, and distribute new MATLAB algorithms and applications is allowed and convenient.

Key Features

* High-level language for technical computing
* Development environment for managing code, files, and data
* Interactive tools for iterative exploration, design, and problem solving
* Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration
* 2-D and 3-D graphics functions for visualizing data
* Tools for building custom graphical user interfaces
* Functions for integrating MATLAB based algorithms with external applications and languages, such as C, C++, Fortran, Java, COM, and Microsoft Excel

Appendix TWO: Hits and their analysis

Due to the data rights, here we only show an example of hits retrieval and hits analysis which is from experiment 4.

Experiment 4:

1. Index of siRNA

Index	siRNA	Index	siRNA
1	ERBB2	41	FASTK
2	ERBB3	42	FER
3	ERBB4	43	FES
4	FGFR4	44	FLJ12476
5	FGFR3	45	RFK
6	FGFR2	46	FLJ10842
7	MASTL	47	FLJ25006
8	ULK4	48	FLJ32685
9	THNSL1	49	C9ORF98
10	FLT3	50	FRK
11	FLT1	51	FRDA
12	TPRXL	52	FRAP1
13	FYB	53	GCK
14	FYN	54	GFRA2
15	GAK	55	GK
16	GNE	56	GSK3A
17	GMFG	57	GSG2
18	GMFB	58	GRK7
19	GTF2H1	59	HAK
20	HIPK3	60	STK32B
21	ERK8	61	GUCY2C
22	ERN1	62	HCK
23	EVI1	63	FGFR1
24	FLJ10761	64	FLJ13052
25	FLJ10074	65	HSMDPKIN
26	FGR	66	HIPK4
27	FLJ23074	67	GUCY2D
28	LRRK1	68	HIPK1
29	FLJ23356	69	FLJ34389
30	FN3KRP	70	FUK
31	FN3K	71	HUNK
32	FLT4	72	HK1
33	GALK1	73	GUK1
34	GALK2	74	HIPK2
35	GAP43	75	GK2
36	GRK6	76	GSK3B
37	GRK5	77	ITGB1BP1
38	GRK4	78	HK3
39	GUCY2F	79	HSPB8
40	HK2	80	HRI

2. The top 10 hits from control DMSO vs. treated DMSO comparison

'area'	'perimeter'	'elongation'	compactFactor'
41	41	79	41
18	18	52	18
52	52	61	52
58	58	59	53
53	53	42	58
5	13	58	13
13	5	7	5
42	42	14	3
3	3	12	62
46	7	65	61

- Red masks: compare with control, the result from treated wells are significantly decreased by one tail KS test.

3. The top 10 hits from control NOCO vs. treated NOCO comparison

'area'	'perimeter'	'elongation'	'compactFactor'
41	41	41	41
42	10	10	10
10	42	11	11
58	11	3	42
11	58	47	58
53	53	45	53
56	56	55	56
52	51	42	51
51	52	56	55
13	55	52	52

- Red masks: compare with control, the result from treated wells are significantly decreased by one tail KS test.

4. The top 10 hits from control Washout vs. treated Washout comparison

'area'	'perimeter'	'elongation'	'compactFactor'
60	3	2	3
3	60	43	60
19	19	45	2
44	2	58	19
45	45	5	44
2	44	44	45
41	58	10	58
58	41	51	8
50	8	8	15
15	15	6	41

- Red masks: compare with control, the result from treated wells are significantly decreased by one tail KS test.

5. From comparing the treated DMSO with corresponding treated NOCO conditions, the hits which don't show significant size increasing or even significant decreasing in NOCO.

Index
11

6. From comparing the treated NOCO with corresponding treated Washout conditions, the hits which don't show significant size decreasing or even significant increasing in Washout.

Index
1
2
3
7
8
9
10
11
12
14
15
16
17
19
41
58
4
6
50
51

7. Analysis on the hits (based on parameter 'area') from 2: The percentage of adhesions which are in the nuclei on both control cells and treated cells.

Index	#FA (control) in nucleus	# FA(treated) in Nucleus
41	0.1886	0.37648
18	0.3168	0.26924
52	0.1886	0.21722
58	0.1886	0.26541
53	0.1886	0.27629
5	0.3168	0.29911
13	0.3168	0.28620
42	0.1886	0.41601
3	0.3168	0.25981
46	0.1886	0.29232

8. Analysis on the hits (based on parameter 'area') from 3: The percentage of adhesions which are in the nuclei on both control cells and treated cells.

Index	#FA (control) in nucleus	# FA(treated) in Nucleus
41	0.2545	0.33166
42	0.2545	0.33252
10	0.3188	0.49774
58	0.2545	0.22936
11	0.3188	0.51845
53	0.2545	0.21964
56	0.2545	0.26586
52	0.2545	0.17769
51	0.2545	0.29302
13	0.3188	0.17423

9. Analysis on the hits (based on parameter 'area') from 4: The percentage of adhesions which are in the nuclei on both control cells and treated cells.

Index	#FA (control) in nucleus	# FA(treated) in Nucleus
60	0.18990	0.20595
3	0.26160	0.29111
19	0.26160	0.25670
44	0.18990	0.17508
45	0.18990	0.18760
2	0.26160	0.32355
41	0.18990	0.21340
58	0.18990	0.33333
50	0.18990	0.17323
15	0.26160	0.23390

10. Analysis on the hits (based on parameter 'area') from 5: The percentage of adhesions which are in the nuclei on both treated DMSO cells and treated NOCO cells.

Index	#FA (DMSO) in nucleus	#FA (NOCO) in nucleus
11	0.30366	0.37660

11. Analysis on the hits (based on parameter 'area') from 6: The percentage of adhesions which are in the nuclei on both treated NOCO cells and treated Washout cells.

Index	#FA (NOCO) in nucleus	#FA (Wahsout) in nucleus
1	0.29072	0.29224
2	0.33899	0.32355
3	0.16493	0.29111
7	0.31299	0.24772
8	0.40690	0.35492
9	0.37660	0.33401
10	0.49774	0.27964
11	0.51845	0.20540
12	0.40275	0.30526
14	0.27912	0.26105
15	0.22407	0.23390
16	0.24111	0.32625
17	0.27881	0.25679
19	0.21707	0.25670
41	0.33166	0.21340
58	0.22936	0.33333
4	0.31967	0.30958
6	0.33703	0.18670
50	0.24924	0.17323
51	0.29302	0.121500

12. Analysis on the hits (based on parameter 'area') from 2: The percentage of different types of adhesions on both control cells and treated cells.

Index	FC% in control wells	FC% in treated wells
41	0.13990	0.15431
18	0.18294	0.15431
52	0.13990	0.15431
58	0.13990	0.19195
53	0.13990	0.21658
5	0.18294	0.21658
13	0.18294	0.14659
42	0.13990	0.14659
3	0.18294	0.23619
46	0.13990	0.26741

Index	FA% in control wells	FA% in treated wells
41	0.71771	0.67816
18	0.69412	0.70867
52	0.71771	0.67816
58	0.71771	0.66045
53	0.71771	0.63624
5	0.69412	0.66008
13	0.69412	0.74491
42	0.71771	0.72457
3	0.69412	0.64087
46	0.71771	0.58516

Index	FB% in control wells	FB% in treated wells
41	0.14239	0.16753
18	0.12294	0.13702
52	0.14239	0.16753
58	0.14239	0.14760
53	0.14239	0.14718
5	0.12294	0.12335
13	0.12294	0.10850
42	0.14239	0.12884
3	0.12294	0.12293
46	0.14239	0.14743

13. Analysis on the hits (based on parameter 'area') from 3: The percentage of different types of adhesions on both control cells and treated cells.

Index	FC% in control wells	FC% in treated wells
41	0.093053	0.12575
42	0.093053	0.12575
10	0.138540	0.12575
58	0.093053	0.14768
11	0.138540	0.30708
53	0.093053	0.12931
56	0.093053	0.20553
52	0.093053	0.12310
51	0.093053	0.13822
13	0.138540	0.20947

Index	FA% in control wells	FA% in treated wells
41	0.73975	0.69319
42	0.73975	0.69319
10	0.72292	0.73196
58	0.73975	0.70182
11	0.72292	0.58519
53	0.73975	0.71134
56	0.73975	0.64632
52	0.73975	0.73745
51	0.73975	0.68840
13	0.72292	0.66750

Index	FB% in control wells	FB% in treated wells
41	0.16719	0.18106
42	0.16719	0.18106
10	0.13854	0.14230
58	0.16719	0.15050
11	0.13854	0.10773
53	0.16719	0.15935
56	0.16719	0.14815
52	0.16719	0.13945
51	0.16719	0.17337
13	0.13854	0.12303

14. Analysis on the hits (based on parameter 'area') from 4: The percentage of different types of adhesions on both control cells and treated cells.

Index	FC% in control wells	FC% in treated wells
60	0.17127	0.12426
3	0.13880	0.12426
19	0.13880	0.12426
44	0.17127	0.13985
45	0.17127	0.12983
2	0.13880	0.12983
41	0.17127	0.15042
58	0.17127	0.11719
50	0.17127	0.13201
15	0.13880	0.11275

Index	FA% in control wells	FA% in treated wells
60	0.68895	0.69169
3	0.73624	0.72741
19	0.73624	0.72741
44	0.68895	0.68794
45	0.68895	0.68623
2	0.73624	0.72119
41	0.68895	0.66295
58	0.68895	0.69120
50	0.68895	0.69812
15	0.73624	0.64216

Index	FB% in control wells	FB% in treated wells
60	0.13978	0.18405
3	0.12496	0.14833
19	0.12496	0.14833
44	0.13978	0.17221
45	0.13978	0.18395
2	0.12496	0.14898
41	0.13978	0.18663
58	0.13978	0.19161
50	0.13978	0.16987
15	0.12496	0.24510

15. Analysis on the hits (based on parameter 'area') from 5: The percentage of different types of adhesions on both treated DMSO cells and treated NOCO cells.

Index	FC% in treated DMSO wells	FC% in treated NOCO wells
11	0.31489	0.30708

Index	FA% in treated DMSO wells	FA% in treated NOCO wells
11	0.57654	0.58519

Index	FB% in treated DMSO wells	FB% in treated NOCO wells
11	0.10857	0.10773

16. Analysis on the hits (based on parameter 'area') from 6: The percentage of different types of adhesions on both treated NOCO cells and treated Washout cells.

Index	FC% in treated NOCO wells	FC% in treated Washout wells
1	0.15837	0.14332
2	0.13626	0.11719
3	0.10086	0.11334
7	0.20553	0.12747
8	0.22452	0.13201
9	0.20947	0.14623
10	0.29558	0.12606
11	0.30708	0.12744
12	0.22204	0.15865
14	0.13006	0.10242
15	0.12002	0.11553
16	0.13660	0.13475
17	0.13822	0.12584
19	0.12062	0.11263
41	0.19287	0.10918
58	0.12931	0.11275
4	0.14571	0.12911
6	0.14954	0.10193
50	0.12044	0.08858
51	0.12601	0.15421

Index	FA% in treated NOCO wells	FA% in treated Washout wells
1	0.71192	0.73684
2	0.71069	0.72468
3	0.79736	0.76652
7	0.68023	0.74892
8	0.64266	0.73155
9	0.66750	0.71905
10	0.59129	0.72253
11	0.58519	0.71139
12	0.63251	0.68872
14	0.72905	0.75893
15	0.72105	0.75487
16	0.72565	0.72428
17	0.71978	0.74250
19	0.71975	0.74583
41	0.66932	0.71960
58	0.71134	0.61275
4	0.71807	0.73853
6	0.69869	0.70708
50	0.71800	0.78543
51	0.6979	0.64019

Index	FB% in treated NOCO wells	FB% in treated Washout wells
1	0.12971	0.11985
2	0.15305	0.15813
3	0.10177	0.12015
7	0.11424	0.12362
8	0.13282	0.13643
9	0.12303	0.13472
10	0.11313	0.15141
11	0.10773	0.16117
12	0.14545	0.15263
14	0.14089	0.13865
15	0.15893	0.1296
16	0.13776	0.14097
17	0.14199	0.13166
19	0.15964	0.14153
41	0.13782	0.17122
58	0.15935	0.27451
4	0.13622	0.13236
6	0.15176	0.19099
50	0.16156	0.12598
51	0.17609	0.20561

Appendix Three: List of abbreviation

1. **NA** – Numerical aperture
2. **ECM** – Extracellular matrix
3. **FAK** – Focal adhesion kinase
4. **FC** – Focal complex
5. **FA** – Focal adhesions
6. **FB** – Fibrillar adhesions
7. **siRNA** – Small interfering RNA
8. **DMSO** – Dimethyl sulfoxide
9. **Noco** – Nocodazole
10. **WO** – wash-out after Noco exposure
11. **MT** – Microtubule
12. **KS test** – Kolmogorov–Smirnov test
13. **CDF** – Cumulative distribution function

Appendix Four: Explanation of related biological terminology

1. **Kinases** is particular type of proteins that through modification – phosphorylation to activate or inactivate other proteins.

2. **Cell Metastasis** is is the spread of a disease from one organ or part to another non-adjacent organ or part. Only malignant tumor cells – cancer and infections have the established capacity to metastasize; Cancer cells can break away, leak, or spill from a primary tumor, enter lymphatic and blood vessels, circulate through the bloodstream, and be deposited within normal tissue elsewhere in the body. Metastasis is one of three hallmarks of malignancy.

3. **MCF7** is a breast cancer cell line was isolated in 1970 from a 69-year-old Caucasian woman.

4. **Objective** is the lens or mirror in a microscope that gathers the light coming from the object being observed, and focuses the rays to produce a real image. Microscope objectives are characterized by two parameters, namely, magnification and numerical aperture. For example objective 60x/1.4NA means the magnification is 60 times and the numerical aperture is 1.4.

5. **Binary mask** is the mask with all its pixels represented as 1 in object of interest. Pixels in Background us 0.

6. **Control #2 siRNA treated cells** are injected siRNAs which do not target any kinase genes of interest.

7. **Paxillin siRNA treated cells** are injected siRNA which only targets paxillin. They are used to check the knock down efficiency.

8. **Resolution** of an optical microscope is defined as the shortest distance between two points on a specimen that can still be distinguished by the observer or camera system as separate entities. It can be derived as:

   Resolution (r) = λ /(2NA) (1)
   Resolution (r) = 0.61 λ /NA (2)
   Resolution (r) = 1.22 λ /(NA(obj) + NA(cond)) (3)

   Where λ is the imaging wavelength

Reference:

[1]: Chen CS, Alonso JL, Ostuni E, Whitesides GM and Ingber DE, 2003. Cell shape provides global control of focal adhesion assembly. Biochemical and Biophysical Research Communications, 307(2):355–61.

[2]: Zaidel-Bar R, Itzkovitz S, Ma'ayan A, Iyengar R, Geiger B. Functional atlas of the integrin adhesome. Nat Cell Biol 2007;9:858–67.

[3]: Hynes RO. Integrins: bidirectional, allosteric signaling machines. Cell 2002;110:673–87.

[4]: http://en.wikipedia.org/wiki/Focal_adhesion

[5]: Calderwood DA, Zent R, Grant R, Rees DJ, Hynes RO, Ginsberg MH. The Talin head domain binds to integrin beta subunit cytoplasmic tails and regulates integrin activation. J Biol Chem 1999;274:28071–4.

[6]: Calderwood DA, Huttenlocher A, Kiosses WB, Rose DM, Woodside DG, Schwartz MA, et al. Increased filamin binding to beta-integrin cytoplasmic domains inhibits cell migration. Nat Cell Biol 2001;3:1060–8.

[7]: Critchley DR, HoltMR,Barry ST, Priddle H,HemmingsL,NormanJ. Integrinmediated cell-matrix adhesion: the cytoskeletal connection. Biochem Soc Symp 1999;65:79–99.

[8]: Zaidel-Bar R, Cohen M, Addadi L, Geiger B. Hierarchical assembly of cellmatrix adhesion complexes. Biochem Soc Trans 2004;32:416–20.

[9]: Zamir E, Geiger B. Molecular complexity and dynamics of cell-matrix adhesions. J Cell Sci 2001;114:3583–90.

[10]: Izzard CS, Lochner LR. Formation of cell-to-substrate contacts during fibroblast motility: an interference-reflexion study. J Cell Sci 1980;42:81–116.

[11]: Clark, E. A., King, W.G., Brugge, J. S., Symons, M. and Hynes, R. O. (1998). Integrin-mediated signals regulated by members of the rho family of GTPases. J Cell Biol 142, 573-586.

[12]: Cukierman E, Pankov R, Stevens DR, Yamada KM. Taking cell-matrix adhesions to the third dimension. Science 2001;294:1708–12.

[13]: Cukierman E, Pankov R, Yamada KM. Cell interactions with three dimensional matrices. Curr Opin Cell Biol 2002;14:633–9.

[14]: Pankov R, Cukierman E, Katz BZ, Matsumoto K, Lin DC, Lin S, et al. Integrin dynamics and matrix assembly: tensin-dependent translocation of alpha(5)beta(1) integrins promotes early fibronectin fibrillogenesis. J Cell Biol 2000;148:1075–90.

[15]: Zamir E, Katz M, Posen Y, Erez N, Yamada KM, Katz BZ, et al. Dynamics and segregation of cell-matrix adhesions in cultured fibroblasts. Nat Cell Biol 2000;2:191–6.

[16]: Wegener KL, Partridge AW, Han J, Pickford AR, Liddington RC, Ginsberg MH, et al. Structural basis of integrin activation by talin. Cell 2007;128:171–82.

[17]: 11. Katz, B. Z. et al. Physical state of the extraxellular matrix regulates the structure and molecular composition of cell-matrix adhesions. Mol. Biol. Cell.

[18]: Paran Yael; Ilan Micha; Kashman Yoel; Goldstein Sofee; Liron Yuvalal; Geiger Benjamin; Kam Zvi. High-throughput screening of cellular features using high-resolution light-microscopy; application for profiling drug effects on cell-matrix adhesion. Journal of structural biology 2007;158(2):233-43.

[19]: Nurit Lichtenstein, Benjamin Geiger, and Zvi Kam: Quantitative Analysis of Cytoskeletal Organization by Digital Fluorescent Microscopy. Weizmann Institute of Science, Department of Molecular Cell Biology, Rehovot, Israel
[20]: Stanley Sternberg. "Biomedical Image Processing", IEEE Computer, January 1983.
[21]: http://en.wikipedia.org/wiki/Noise_reduction
[22]: Nargess Memarsadeghi, David M. Mount, Nathan S. Netanyahu, and Jacqueline Le Moigne. A Fast Implementation of the ISODATA Clustering Algorithm.
[23]: David L. Donoho Department of Statistics Stanford University. On Minimum Entropy Segmentation. September 1993 Revised March 1994.
[24]: Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, Num. 6 (1991), pages 583-598* .
[25]: Drs, William E. Green, Canny Edge Detection Tutorial, Mechanical Engineering and Mechanics, 3141 Chestnut Street, MEM Department, Room 2-115, Drexel University, Philadelphia, PA 19104, 2002
[26]: Rafael C. Gonzalez, Rechard E. Wood, "Digital Image Processing"
[27]: Kuan Yan, "Masked Watershed Segmentation", Imaging & Bioinformatics, Leiden University, LIACS, The Netherlands
[28]: Dr. Ir. Fons. Verbeek, Dr. Nies Huijsmans, Lecture Notes for Image Analysis in Microscopy, Leiden University, LIACS, The Netherlands
[29]: http://en.wikipedia.org/wiki/Image_moments
[30]: S. C. Johnson (1967): "Hierarchical Clustering Schemes" *Psychometrika*, 2:241-254
[31]: Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides".
[32]: J. MacQueen, 1967
[33]: http://en.wikipedia.org/wiki/Cluster_analysis
[34]: D.L. Davies and D.W. Bouldin, IEEE Transactions on Pattern Analysis and Machine Intelligence 1, pp. 224-227, 1979
[35]: Daniel C.Worth, Maddy Parsons, "Adhesion dynamics: Mechanisms and Measurements", Randall Division of Cell and Molecular Biophysics, Kings College London, New Hunts House, Guys Campus, London SE1 1UL, UK.
[36]: http://en.wikipedia.org/wiki/SiRNA
[37]: http://en.wikipedia.org/wiki/RNA_interference