Contents

1 Introduction

1.1 ENLIGHT

ENLIGHT (ENhanced LIGase based Histochemical Techniques) is an EU (European Union) project intended to utilize proximity ligation and padlock probe techniques to develop assays for *in situ* detection of proteins and genetic markers in tumor cells. In addition, there is a need to develop image analysis tools to detect these *in situ* assay signals and clinically investigate these assays for the diagnosis and management of cancer [1].

> "When combined with fast and precise detection by image analysis, the resulting assay methods will become invaluable tools in biological research, routine diagnosis, and drug development."
> (Björn Ekström, CEO of Olink and coordinator of the ENLIGHT project)

1.2 Problem formulation and project aim

Prof. A. K. Raap and his research team at Leiden University Medical Center, as a part of the ENLIGHT project, is working to understand mtDNA segregation and mitochondrial DNA (mtDNA) mutation accumulation. Their research focuses on mtDNA segregation patterns in *in vitro* cultured cells. Experimentally, this requires the determination of the mutation load in hundreds of individual cells in multiple serial cell culture passages of a cloned heteroplasmic founder cell (i.e., a single cell carrying both mutant and wildtype-mtDNA molecules). *In situ* genotyping mtDNA with the padlock/rolling circle method [2] provides an elegant approach for detection of mtDNA sequences variants at the (sub-) cellular level. This detection method results in a large amount of images of cells with different mutation loads. Estimation of mutation load requires a fast, accurate, automated, and user friendly image analysis application.

In single cell analysis it is crucial to assign signals to a specific cell. An image of a cell culture often contains different cells that all possess different characteristics. Taking the average of such an image will not reveal the differences and variations between the cells. In cases like these, a single cell analysis is the only option in order to observe the dissimilarities among the cells.

Many different image analysis software can perform a single cell analysis as described above, but they are often not suited for easy use by persons lacking extensive knowledge in image analysis , e.g. , laboratory personnel. A windows based software that uses only a few parameters and little manual input, yet still produces accurate and easily handleable data, is required for this purpose. In this master thesis already existing functions from Matlab and IMP (section 2.3.1 and 2.3.2) were implement in C++ code as an Add-In to a windows based user friendly software called VIS (Visiopharm Integrator System) (section 2.3.3). In addition, an evaluation of its performance on mDNA mutation loads was made.

2 Background theory

2.1 Biological background

2.1.1 Mitochondrial DNA

Mitochondrial DNA (mtDNA) is a circular molecule and unlike most DNA of eukaryotic organisms it is located in the mitochondria. The genetic information found in the ∼16kbp human mtDNA is essential for a major energy-generating process of the cell called oxidative phosphorylation. Mitochondrial inheritance is non-Mendelian in contrast to nuclear DNA which presumes that half the genetic material of a zygote is inherited from each parent (Mendelian inheritance). Instead, all mtDNA is inherited from the mother. In humans, mtDNA is present at 100-10 000

copies per cell, consisting of 37 genes and 16,569 nucleotides, coding for 13 proteins (polypeptides), 22 transfer-RNA (tRNA) and two ribosmal RNA (rRNA)[3]. The mtDNA consists of one heavy and one light strand carrying 28 and 9 genes respectively. 8 of the 9 genes on the light strand code for mitochondrial tRNA molecules. The human mtDNA is regulated by only one regulatory region, containing the origin of replication of both heavy and light strands. All DNA mutates and so does mtDNA. However, the mutation rate of mtDNA is approximately 10 times greater than that of nuclear DNA. The high mutation rate leads to a high variation not only among different species but also within the same species [3] .

When the mutation is pathogenic it needs to accumulate to relative large amounts (>80% of all mtDNAs) for the cells energy provision to become so subverted that cell functions are lost and cells die. Such mutation accumulation leads to devastating diseases if the mutation is inherited from the mother or to normal aging phenomena if it is acquired somatically. A major factor in determining cellular mutation loads is the process of mitotic segregation. To understand mtDNA segregation and mtDNA mutation accumulation, one approach is to study the mtDNA segregation patterns in *in vitro* cultured cells. Experimentally, this requires the determination of the mutation load in hundreds of individual cells in multiple serial cell culture passages of a cloned heteroplasmic founder cell (i.e., a single cell carrying both mutant and wildtype-mtDNA molecules).

2.1.2 Rolling-circle amplification of padlock probes

Detection of single molecule variability, interactions and mechanisms may go undetected at the level of populations of molecules. A widely used method such as fluorescence *in situ* hybridization (FISH) is able to detect large mutations such as duplications, translocations and deletions, but is not sensitive enough to detect single nucleotide sequence variations. Padlock probes with rolling-circle amplification (RCA) provide a method for detecting individual nucleic acid molecules with excellent specificity [2].

The principle for detecting single nucleotide polymorphism (SNP) using padlock probes and target primed RCA, rolling-circle amplification, is to first cut the target strand downstream (3'-end) of the SNP with restriction enzymes. After the restriction digestion the target DNA is irreversibly made single stranded by a strand-specific 5'-3' exonucleolysis. The padlock probes are then hybridized to the target DNA and if there is a perfect match to the SNP the ends of the padlock probe will be ligated and circular DNA is formed. If the padlock probe is not a perfect match to the SNP it will not be able to ligate its ends and no circular DNA will be formed. The forming of circular DNA is essential for the RCA to work. After ligation, the RCA is initiated by Φ29 DNA polymerase by turning the target molecule into a primer and the padlock probe then serves as a template for DNA synthesis. The RCA product is then hybridized with fluorescent-labeled oligonucleotides so that the product can be detected (Fig. 1) [2].

The advantage with padlock-probes and RCA is that even a difference in only one nucleotide can be detected. Also, the amplification step creates a higher signal which in turn leads to an easier detection of signals in the analysis. One drawback with this method is that even unmatched padlock probes will hybridize with the target DNA. Although an unmatched padlock-probe will not ligate and produce a signal, it will still affect the result since it will occupy a site and prevent binding of other padlock probes. The detection efficiency will therefore always be less than 50% if two different padlock probes are used to find a SNP.

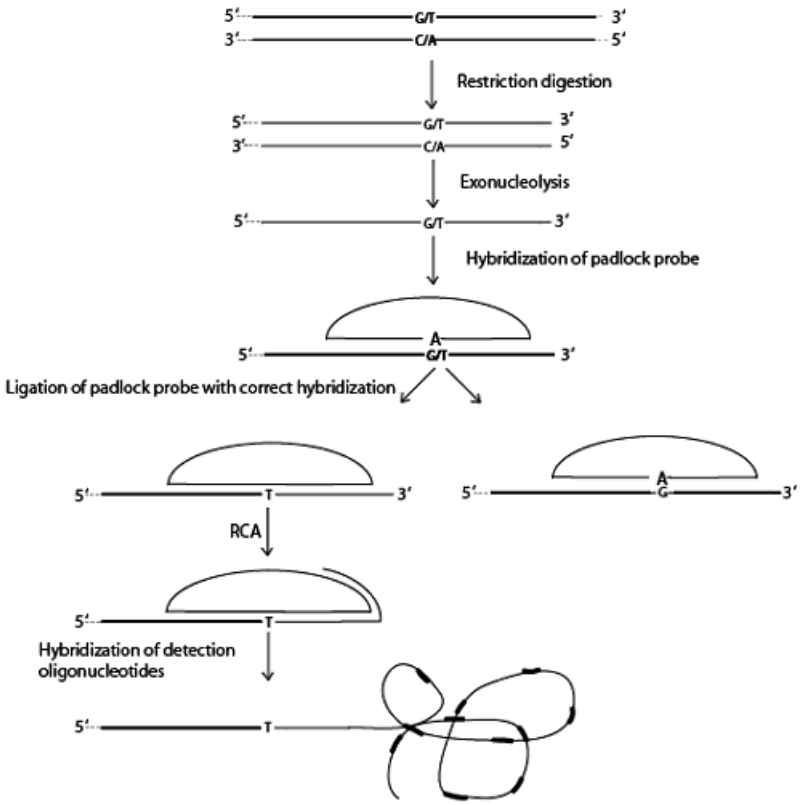


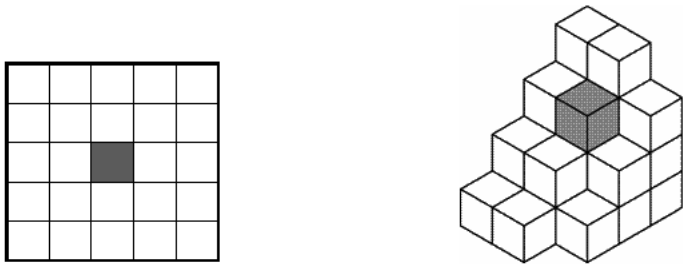
Figure 1: Detection of SNP with padlock probes and target-primed rolling circle amplification, used with permission from Larsson et al. [2]. The target strand is made single stranded by restriction digestion and exonucleolysis. A padlock probe is then hybridized to the target strand. If the padlock probe is a perfect match for the target strand it will ligate and form a circular DNA. The circular DNA serves as a template for the RCA. The amplification product is hybridized with flourescent-labeled oligonucleotides for visualization of the SNP.

2.2 Digital images analysis

2.2.1 Digital images

A digital image is built up of elements called pixels (picture elements), for volume images the elements are called voxels (volume picture elements) (Fig. 2). An image can be represented as a mathematical function of five variables $f(x, y, z, t, b)$. x, y and z are all spatial variables t is the time variable and b represents the spectral band in a multispectral image. Color images have three spectral bands, each one of these can be seen as a grey-scale image. One of the most common ways of representing a color image is by the RGB model. The RGB color model is an additive model in which red, green and blue are combined in various ways to reproduce other colors. Each color band also has a bit-depth representing the amount of grey-levels in that color band. A color band with 8-bit has $2^8 = 256$ different grey-levels which gives an RGB image, 8 bit in each color band, a total of 256^3 or 16777216 colors [4]. A higher bit-value will include more information in the image while a lower bit-value, for example 1, will include only 2^1 levels (black and white), a binary image.

Connectivity
In a digital image neighbors must be defined, the neighbors can be defined in different ways depending on the context. A pixel in 2D has four edge neighbors and four vertex neighbors



(a) $f(x, y)$ (b) $f(x, y, z)$

Figure 2: (a) Pixels (b) Voxels

(Fig. 3). If a 4-connectivity is used only the four edge neighbors are counted as neighbors. In an 8-connectivity all the edge and vertex neighbors are included in the neighborhood (Fig. 4) [4].

(a) (b)

Figure 3: (a) Edge neighbors (b) Vertex neighbors

(a) (b)

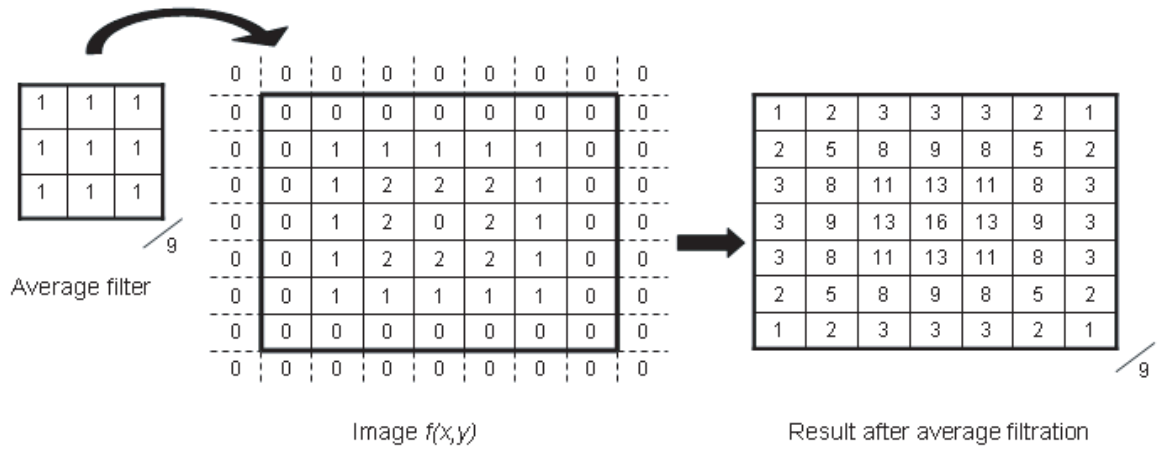
Figure 4: (a) 4-connectivity (b) 8-connectivity

2.2.2 Pre-processing

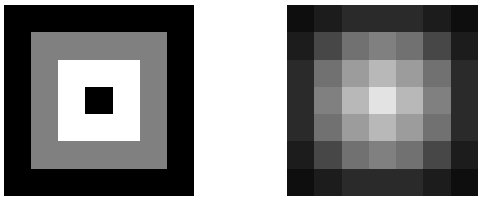
Pre-processing are operations done on an image before an analysis is performed; both input and output are intensity images. The aim of pre-processing is to improve the image ,e.g., suppress noise present in the image or enhance some image features important for further processing.

Filtering
Filtering operations performed directly on the pixels of an image, spatial filtering, is an important aspect of image analysis as it provides a possibility of removing noise, sharpening blurred images, enhancing areas of a specific characteristic, etc. The operation consists of multiplying each pixel in a neighbourhood by a corresponding coefficient and summing the results in point (x, y) in the

image $f(x, y)$. The neighbourhood containing the coefficients are referred to as masks, filters, or filtermasks and can be of any size $m \times n$. The filtermask is moved over the image and for every pixel (x, y) the sum of the neighbourhood multiplied by the mask coefficients is set at the position (x, y)[5]. In Fig. 5 the result from an average filtration with a 3×3 mask is shown.



(a) Average filtration of image $f(x, y)$



(b) (c)

Figure 5: (b) Image $f(x, y)$ before filtration. (c) Image $f(x, y)$ after average filtration.

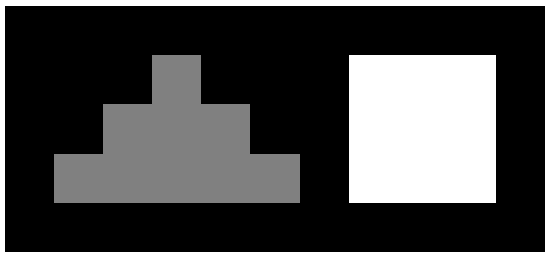
2.2.3 Segmentation

Segmentation is the process of dividing a digital image into specific regions. The aim of the segmentation is to locate and separate different objects in an image. Each object is given a specific value (label) representing that object. In Fig. 6 an example of a segmented image is shown, the background is labelled as 0, the triangle is labelled 1, and the square is labelled as 2.

Thresholding
Thresholding is the easiest way of segmenting an image. In a greyscale image, pixels with a value higher (i.e. brighter pixels) than a certain threshold value are marked as objects often given the value 1. Values lower than the threshold are considered as the background and are given the value 0 (Eq. (1)). After the threshold a binary image of only values 0 and 1 is produced. The easiest way of thresholding an image is to manually supply a threshold value T. As a guide when choosing the threshold value an image histogram is used. The image histogram $p(f)$ from the image $f(x, y)$ is the probability function which gives the frequency of the different grey-scale

0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	2	2	2	0
0	0	1	1	1	0	0	2	2	2	0
0	1	1	1	1	1	0	2	2	2	0
0	0	0	0	0	0	0	0	0	0	0

(a)

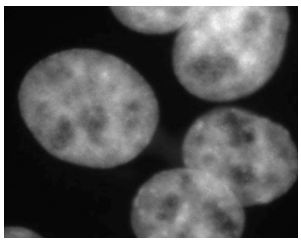


(b)

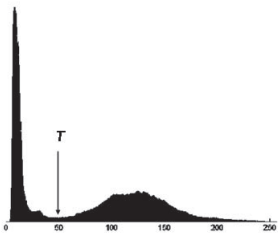
Figure 6: Segmented image: Background label 0, triangle label 1 and square label 2.

levels in $f(x, y)$. The result from a grey-level thresholding $b(x, y)$ of $f(x, y)$ is seen in Fig. 7.

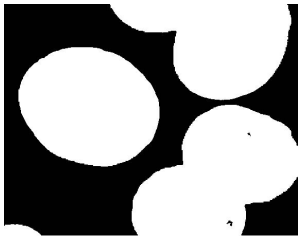
$$b(x,y) = \begin{cases} B_1 & \text{if } f(x,y) \geq T \\ B_2 & \text{if } f(x,y) < T \end{cases} \tag{1}$$



(a) $f(x, y)$



(b) $p(x, y)$



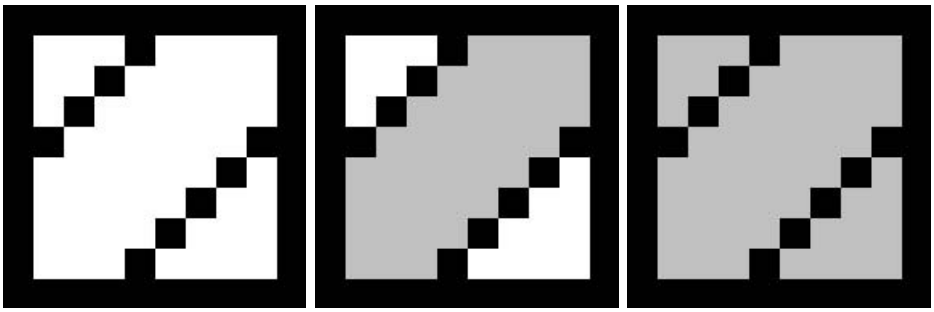
(c) $b(x, y)$

Figure 7: (a)Original image. (b) Histogram. (c) Image after threshold.

Often, a manual threshold is not the most suitable way of thresholding an image. Manual input is unwanted, as it may vary depending on the user, and should be minimized. An automated threshold detection algorithm can therefore be used. One automated method is Otsus method of thresholding. This method is designed to separate foreground from background. From the histogram Otsu selects a threshold that maximizes the between-class variance of the background and foreground. The method relies on the assumption that all the pixels in the image belong to either the background or the foreground (i.e. objects) [6].

Flood fill

The flood fill algorithm is often seen as the bucket in paint programs where it is used to fill a connected area with a new color. The algorithm takes three parameters, the starting pixel, a target value, and a replacement value. The algorithm starts at the starting point (pixel) locates, in a prior defined connectivity, all pixels connected to it by a path of the target value and substitute their value with the replacement value.The results from two types of flood fill can be seen in Fig. 8.



(a) (b) (c)

Figure 8: (a) Before flood-fill. (b) Result from flood fill in 4-connectivity initiated in the middle compartement. (c) Result from flood fill in a 8-connectivity initiated in the middle compartement

0	0	0	0	0	0	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

(a)

0	0	0	0	0	0	0
0	1	1	1	1	1	0
0	1	2	2	2	1	0
0	1	2	3	2	1	0
0	1	2	2	2	1	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

(b)

Figure 9: (a) Binary image. (b) Image after applied distanc transform.

2.2.4 Morphological operations and geodesic transformations

Morphological operations correspond to operate on the shape and form of objects ,e.g. , erosion and dilation. Morphological operations include operations on one input image together with a structuring element. In contrast, a geodesic transformation uses two input images of the same size where a morphological operation is performed on the first image and it is then forced to remain either above or below the second image [7].

Distance transform

A distance transform applied to an image supplies each pixel with a distance to the nearest pixel. The boundary pixel can be either the background or the foreground. If a distance transform is applied to the object in Fig. 9(a) each grey level pixel value will represent the nearest distance to the background (value 0), Fig. 9(b).

There are several ways of representing the distance in an image. In city block distance metric, the path between pixels is based on a 4-connected neighbourhood. Pixels whose edges touch are 1 unit apart and pixels diagonally touching are 2 units apart. The city block distance between two points is defined as:

$$d = |x_i - x_j| + |y_i - y_j| \tag{2}$$

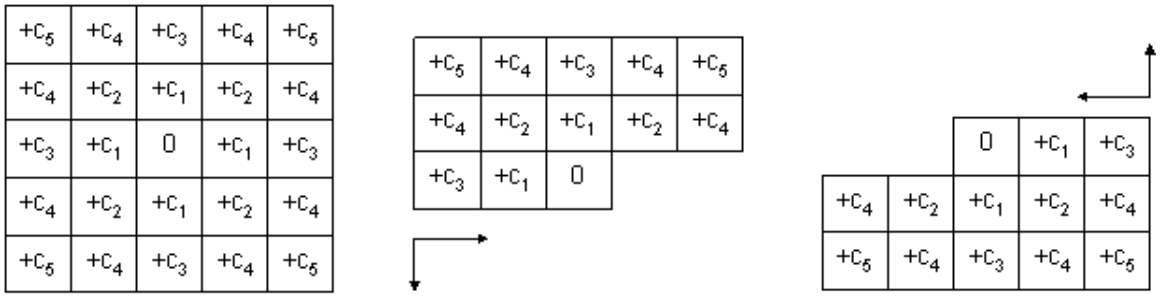
The chessboard distance metric measures the distance based on an 8-connected neighbourhood. Here both edge pixels and corner pixels are considered as neighbourhood and therefore have the same distance, 1 unit, to the centre pixel. The chessboard distance between two points is defined as:

$$d = max(|x_i - x_j|, |y_i - y_j|) \tag{3}$$

In Euclidian distance metric, the distance between two pixels is the length of a straight line connecting the pixels. The Euclidian distance metric is the most accurate distance representation but also the one that has the highest computational cost. The Euclidian distance between two points is defiend as:

$$d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4}$$

Computing the distance from a pixel to a set of boundary pixels is a very costly operation since it is a global operation. As a result, algorithms that consider only a small neighbourhood are necessary for efficiently creating a distance map. These algorithms can calculate the Euclidean distance in an image, but more often an approximation is used. The basic idea for these algorithms is that the global distances in the image are approximated by propagating local distances, i.e., distances between neighbouring pixels. The propagation can be done with either a parallel or sequential algorithm [8]. The masks used in these algorithms can vary in neighbourhood size, see Fig. 10 for an example of a 5x5 mask for parallel and sequential algorithm propagation.



(a) Parallel mask (b) Forward mask (c) Backward mask

Figure 10: Mask describing the distance transformations. (a) Mask used in parallel computations. (b) and (c) The parallel mask is split into two masks used for sequential computations.

In the parallel algorithm the mask in Fig. 10(a) is placed over each pixel in the image. The local distances in each mask pixel c_n is added to its corresponding pixel value in the image. The new pixel value is the minimum of all of these sums. The process is repeated until no pixel value changes in the image.

$$v_{i,j}^m = \min_{(k,l) \in mask} (v_{i+k,j+1}^{m-1} + c(k,l)) \tag{5}$$

$v_{i,j}^m$ is the value of the pixel in position (i,j) at the iteration m. (k,l) is the position in the mask, $(0,0)$ being the center, and $c(k,l)$ is the local distance from the mask.

In the sequential algorithm the symmetrical parallel mask is split into two masks; shown in Fig. 10(b) and Fig. 10(c). Each of the masks are passed over the image once; the forward mask from left to right, and top to bottom, and the backward mask from right to left, and bottom to top. The centre pixel value is, as in the parallel algorithm, the minimum of the sum of c_n and its corresponding pixel value in the image. After the passages of the masks a distance map

is produced in the image. The equation for the sequential algorithm for forward and backward masks is seen in Eq. (6).

$$v_{i,j} = \min_{(k,l)\in mask}(v_{i+k,j+l} + c(k,l)) \tag{6}$$

The notations are the same as in Eq. (5). If no border expansion for the image is done, the initial and final position of the masks have to be adjusted so that none of the mask elements are positioned outside the image border.

The local distances c_n in the masks differ depending on the distance metric used. Integer numbers are often preferred in digital image processing. A good approximation of the Euclidian distance approximation with integer numbers can be achieved by using Chamfer 3-4 for a 3x3 neighbourhood or chamfer 5-7-11 for a 5x5 neighbourhood, they produce maximum errors of 0.0809 and 0.0202, respectively, compared to the Euclidean distance [8]. These masks can be seen below in Fig. 11.

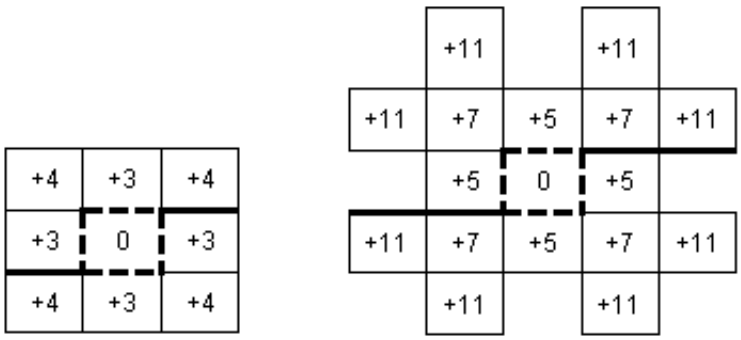


Figure 11: Chamfer 3-4 and Chamfer 5-7-11 for integer approximations of Euclidean distances.

Fig. 12 shows an example of distance transform computed by City block, Chessboard, Chamfer 3-4, Chamfer 5-7-11, and Euclidean distance metrics.

H-extrema

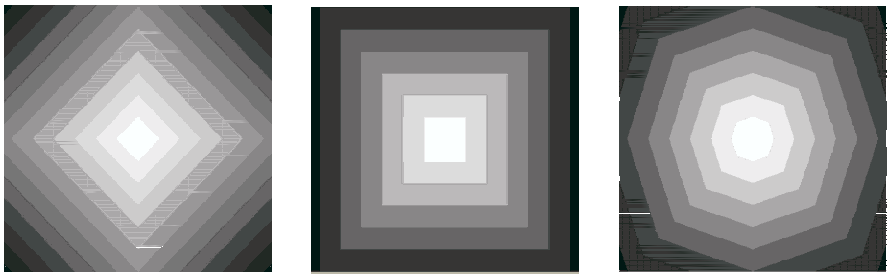
Regional maxima are connected pixels with the same intensity value, t, whose external boundary pixels (in a defined connectivity) all have a value less than t. An image of regional extrema will show both relevant and irrelevant image features. A geodesic transform called h-extrema is able to filter the image by using a contrast criterion. The h-maxima transform is able to suppress all maxima whose depth is smaller than a given threshold level t [7].

H-extrema is an iterative converging method. The image f is first subtracted by a constant value t creating the image g_i. A scan over g_i finds the maximum value of a neighborhood of each pixel, MAXg_i. g_{i+1} is created from the minimum of f and MAXg_i. The operation is performed again with f and g_{i+1} until g_i equals g_{i+1}.

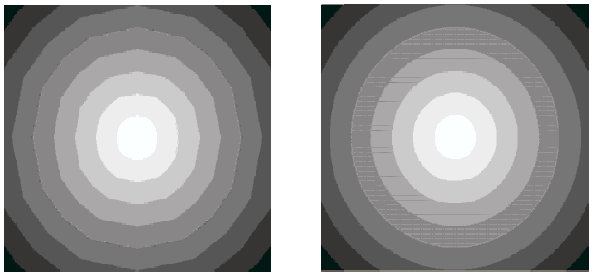
2.2.5 Combining morphological operations with segmentation

Watershed

The watershed segmentation is a region based method of segmenting an image. Watershed segmentation can separate connected objects ,e.g. , clustered cells. The watershed segmentation



(a) City Block (b) Chessboard (c) Chamfer 3-4



(d) Chamfer 5-7-11 (e) Euclidean

Figure 12:

can be understood as seeing the image as a landscape. The gray level intensity represents the differences in elevation in this landscape. Water enters through the local minima and starts to rise. A lake around a local minima is created and referred to as catchment basin. Water stops rising when it reaches a pixel at the same geodesic distance from two different catchment basins. As two catchment basins meet they form a dam or watershed that separates the two objects. All that is left after the watershed segmentation are these watershed lines separating the objects see Fig. 13, [9]. In some versions of the watershed segmentation water enters through the maxima instead of the minima.

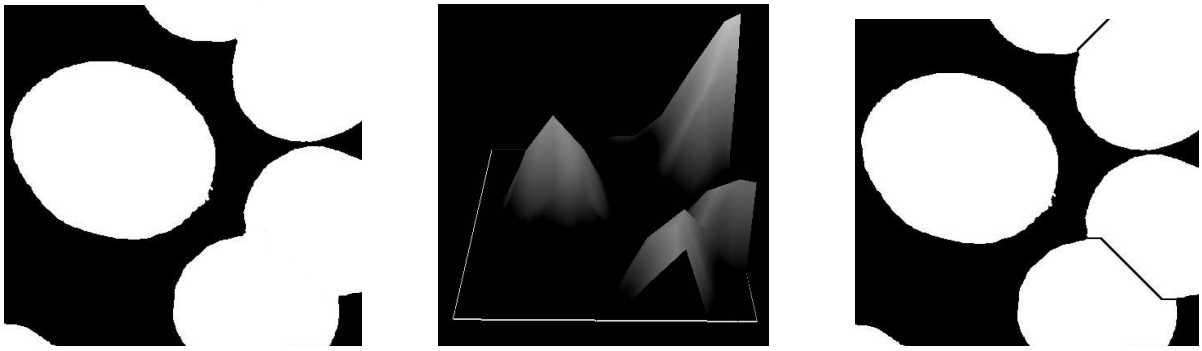
2.3 Image analysis environments

2.3.1 Matlab

Matlab, created by The MathWorks (Natick, Massachusetts, USA, mathworks.com), is a numerical computing and programming language. Matlab allows easy manipulation of matrices, implementation of algorithms, creation of user interfaces and interfacing with programs in other languages. Many addable functions specialised in image analysis are available making matlab a good tool for developing image analysis algorithms. Some disadvantages of matlab are that it is fairly expensive if only needed for one type of analysis application and quite some experience and knowledge is needed in order to benefit of all matlab's functionalities.

2.3.2 IMP (IMage Processing)

IMP is a C/C++ based image processing software for Linux developed at CBA (Center for Image Analysis at Uppsala University). It possesses many image processing features and is excellent in performing a vast variety of image analysis. Knowledge in C/C++ programming provides the



(a) (b) (c)

Figure 13: (a) Binary image before watershed. (b) Landscape like representation of the image after distance transform. (c) Image after watershed.

ability to add new functions to the software. The disadvantage of IMP, from, e.g. , laboratory personnel's point of view, is that it is a locally used (at CBA) software, it runs under Linux, and requires knowledge in image processing.

2.3.3 VIS (Visiopharm Integrator System)

Visiopharm Integrator System (VIS, Visiopharm, Hørsholm, Denmark, visiopharm.com) is a windows based image analysis software for medical applications. It includes a lot of features, e.g. , automated image acquisition, different segmentation approaches, stereology and a developers kit called IU-SDK (Imaging Utilities Software Development Kit). The IU-SDK is a C++ based library (DLL) for advanced image analysis and it is the foundation of VIS and also for the Visopharm Plug-In modules for developers. Users can, by using IU-SDK, make their own algorithms and analyses and use as a Add-In to VIS. Among other things, VIS also has two special features called masks and labels, both visualized as layers over an image. The masks define different regions of interest (ROIs) in an image. When an analysis is performed it can be set to include only parts inside a ROI. The labels can be used to label areas of specific characteristics in an image. Both the masks and the labels can be drawn by the mouse right on the image, which makes it very easy to define and edit regions and labels. In conclusion, one can say that VIS is a user friendly software that can perform a vast variety of image analyses. In addition, advanced users have the possibility to add their own functionality to the software. For a screenshot of VIS see Fig. 19.

Some disadvantages of VIS are that it uses mainly pixel - based classification for segmentation and it focuses on area count rather than single event count. In certain situations neither of these two approaches are the best choice. Furthermore, VIS is currently not able to count signals per cell, making single cell analysis impossible. One objective of the ENLIGHT project is to promote further developement of VIS for applications such as single cell analysis.

2.4 Programming languages and environments

2.4.1 Visual C++

Microsoft Visual C++ (also known as MSVC) is an integrated development environment (IDE) product for C, C++ and C++/CLI programming languages. It consists of tools for developing and debugging C++ code. C++ is an extension of C and it uses object oriented programming.

Visual C++ has features such as syntax highlighting (according to the category of terms), IntelliSense (a coding autocompletion feature) and advanced debugging functionality. The compile and build system features precompiled header files and "minimal rebuild" functionality which all significantly shortens the time to edit, compile and link the program.

2.4.2 Matlab

> "Jack Little and Cleve Moler, the founders of The MathWorks, recognized the need among engineers and scientists for more powerful and productive computation environments beyond those provided by languages such as Fortran and C. In response to that need, they combined their expertise in mathematics, engineering, and computer science to develop MATLAB, a high-performance technical computing environment."[10]

Matlab is a mathematical scripting language similar to C++. It uses efficient matrix and vector computations with easy creation of scientific and engineering graphics. It can be used as a an application development with graphical user interface building. Matlab uses object-oriented programming. File I/O functions, string processing and extensibility (tool boxes) are some more features of the matlab programming language. A matlab-code does not need to be compiled before execution which makes it very easy to use when developing new algorithms. Matlab can be run under windows, Unix and Mac OS X.

3 Materials and methods

For single cell analysis of signal counts VIS is limited and can not complete the task. Functions from Matlab, IMP, and IU-SDK were utilized to make an Add-In for VIS that is able to perform single cell analysis of padlock probed mtDNA.

3.1 Cell staining and image acquisition

The nuclei are stained with DAPI (blue). Padlock probes for mutated mtDNA are detected with Cy5 stain (red) and padlock probes for wild-type DNA are detected with FITC stain (green). For visualization of the cytoplasm tubulin is detected with mouse anti-tubulin antibodies and two different secondary antibodies, rabbit anti-mouse FITC (green) and goat anti-mouse Alexa 594 (red). Cytoplasmic stains can not be used together with padlock probes detected with the same color due to spectral overlap. Images are acquired using an epiflourescent microscope (Leica, Leica Microsystems GmbH, Wetzlar, Germany) equipped with a cooled monochrome CCD camera (Quantix, Photometrix, Melbourne, Australia). From the image acquisition three grayscale images are produced representing the red, green and blue channel. The images are 16-bit tiff with a size of 1308x1020 pixels. All images were acquired at the Department of Molecular Cell Biology at Leiden Medical University.

3.2 Deliniation of nuclei

The cell segmentation is initiated by a segmentation of the image channel representing the nuclear stain (Fig. 14A). Otsu's method of thresholding, which minimizes the variance of the foreground and the background, separates the nuclei from the background [6] (Fig. 14B). Otsu's method of thresholding works well here due to the assumption that only background and foreground is present in the image. Sometimes, dark areas inside the nuclei appear as holes after

the threshold. These holes are filled using a flood fill algorithm. The algorithm floods the background from all background border pixels. All pixeles reached by the flood fill will be set as background while all pixels not reached by the flood fill will be set as objects.

The binary image representing the nuclei is transformed to a landscape-like image using distance transformation (Fig. 14C). The distance image is produced using the 5-7-11 Chamfer distance transform on the binary image [8]. The chamfer distance transform was preferred over the Euclidian distance transform due to lower computational cost and yet sufficient result. In addition, the chamfer distance transform yields integer results which makes computation even easier.

Seeds, entry points for the water in watershed segmentation, representing the different nuclei are needed in order to separate clustered nuclei into different objects with the watershed segmentation. Due to imperfect circularity of the nuclei distance transform may lead to multiple seeding points or local maxima, for the same nucleus. This will result in over-segmentation. The h-maxima transform is able to suppress maxima whose depth is smaller than a given threshold t [7]. A low value of t in contrast to a high t value will result in more seed points, hence more over segmentation. By suppressing all small maxima several adjacent local maxima are merged into one regional maximum, i.e., one seed point for each nucleus is achieved.

Clustered nuclei are separated by watershed segmentation [9]. For our version of the watershed segmentation water raises from the maxima, i.e., the local maxima are the seeds. Water rises and floods until two catchment basins meet and generate a watershed that separates the nuclei. Given that flooding only starts from the seeds, every seed will produce one object. The implementation of the watershed can be done with sorted pixel lists, therefore the segmentation can be done very fast [11]. Incorporated into the watershed is an area count of each object label. This count is used for removing objects that are smaller than a user defined area minimum, i.e. objects not considered to be true nuclei. The value t from h-maxima transform is directly proportional to the radius of an object and can therefore also be used to remove objects that are too small to be true cell nuclei, and thus have a radius less than a specified value. However, the area object count uses less computational time than the h-maxima transform and was therefore preferred for removing small objects. The result from the watershed segmentation can be seen in Fig. 14D.

3.3 Deliniation of cytoplasm

In single cell analysis it is crucial to assign signals to a specific cell. To do this, each cell compartment (cytoplasm) has to be delineated. A common approach is to use a cytoplasm stain with the intention of using the staining as a guide when delineating the cytoplasm [12],[13]. Another approach is to use a membrane stain, which binds to the cytoplasmic or nuclear surface. For many molecular detection systems a blue stain is used for the nucleus, and red and green stain for molecular detection. Due to fluorescence spectral overlap this limits the possibility of using a unique color for a cytoplasmic stain. Also, due to the antigen destructing nature of the padlock/RCA procedure cytoplasmic staining (e.g. tubiln) in a fourth flourecensce color is currently not an option. As a consequence, another approach for delineating the cytoplasm is required in this case. One way to approximate the outline of the cytoplasm, for each cell, without the help of cytoplasmic stain is by using a fixed radial distance from the nucleus border [14].

To validate the fixed radius method of delineation two more approaches to the delineation of the cytoplasm were performed as a comparison; an automated method that utilizes a cytoplasmic stain and another method based on manual delineation. The three methods were compared to see whether the use of a fixed radius was sufficient enough for this type of single cell analysis. Cytoplasms that have a nucleus cut by the image border were excluded since a big portion of

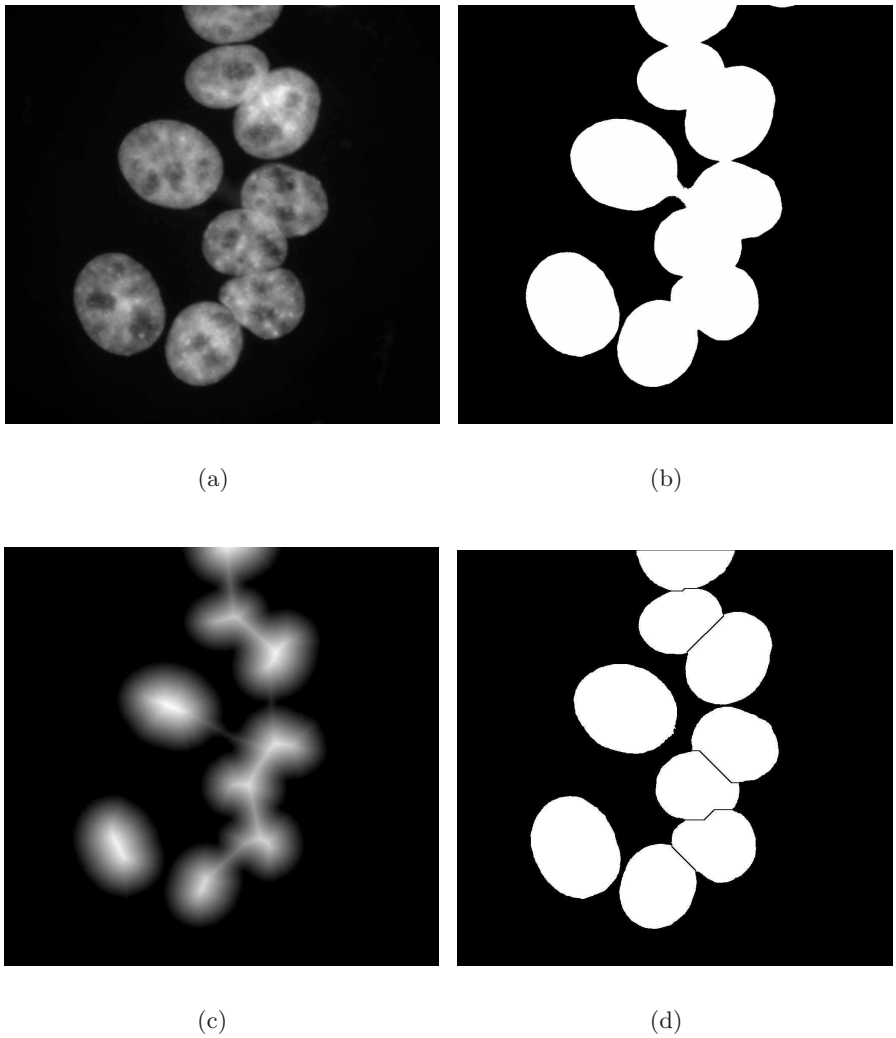


Figure 14: (a) Nuclear stain. (b) Resulting binary image after thresholding. (c) Distance transformation of (b). (d) Final segmentation result.

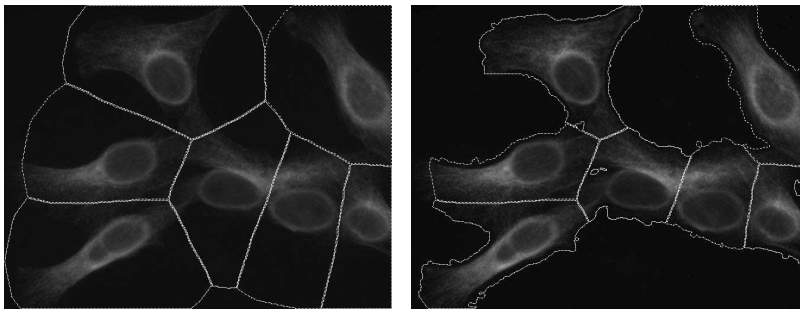
the cell lies outside of the image.

3.3.1 No cytoplasmic stain (NCS)

If no cytoplasm staining is present, the delineation of the cytoplasm is purely based on a fixed distance from the nucleus. A distance transform is applied to the background of the binary image of the nuclei. This results in an image that represents the distance to the nearest nucleus for each pixel in the image. A user defined threshold, corresponding to the maximum radius of the cytoplasm, is applied to the distance transformed background. A watershed is again used to define the borders of the objects. In order to use the same watershed as previously, i.e., with water rising from image maxima, the distance transformed image is inverted. Water rises from the maxima in the image and rises until water from two catchment basins meet and a watershed line, separating two cytoplasms, is formed, see result in Fig. 15A.

3.3.2 Cytoplasmic stain (CS)

The second approach to delineation of the cytoplasm makes use of a cytoplasm staining (tubulin stain). Tubulin is present throughout the whole cytoplasm and can therefore be used as a marker for the cytoplasm. A variance filter is applied to the channel representing the tubulin and areas of high intensity variation (tubulin areas) are enhanced. Thereafter, an average filter is applied to even out areas in close proximity that have varying variance. The variance image is thresholded by Otsu's method, but to include all of the cytoplasm the threshold is adjusted by multiplying it with 0.25. This may be avoided by using a different thresholding method. A watershed transformation seeded by the nuclei and restricted to the binary image of the cytoplasm is there after applied, see result in Fig. 15B.



(a) (b)

Figure 15: **A:** Result of cytoplasmic segmentation not making use of the cytoplasmic stain (NCS). **B:** Result when cytoplasmic stain is included (CS).

3.3.3 Manual delineation (O)

The third method for delineation of the cytoplasm is a manual segmentation. Here, two observers used the software VIS to outline the cytoplasm of the tubulin stained images manually. The cytoplasm was outlined by using different masks (ROI) in VIS.

3.3.4 Comparison of segmentation methods

In the evaluation of the different methods for cytoplasm segmentation, accuracy (agreement with truth), precision (reproducibility) and efficiency (time) are considered, based on the ideas by Udupa et. al. [15]. Define S as being the result of the segmentation method being compared to S_t, the true segmentation. The accuracy makes use of three definitions; False Negative Area Fraction (FNAF), False Positive Area Fraction (FPAF) and True Positive Area Fraction (TPAF). FNAF is the fraction of S_t that was missed by S. FPAF denotes the area that is falsely identified by S as a fraction of S_t. In the current case, the parts of the S that overlap with the image background, as defined by S_t, are not counted as falsely identified because the background does not give rise to any signals and will not affect the calculation of signals per cell. TPAF describes the total amount of cytoplasm defined by S that coincides with S_t as a fraction of S_t.

Precision is the ability to reproduce the same result. Naturally, a fully automated method will always reproduce the same result when applied to the same digital image. With manual delineation the result will most likely not be fully reproducible, there will be inter- and intra-observer variation.

-9	-6	-6	-6	-9
-6	5	15	5	-6
-6	15	28	15	-6
-6	5	15	5	-6
-9	-6	-6	-6	-9

Figure 16: Filter for enchancing areas of local maxima

Two factors must be considered when comparing the efficiency of a segmentation method; the computational time and the human operator time required to complete the segmentation.

3.4 Localization of padlock signals

The image channels containing the padlock signals are filtered with a 5x5 kernel that enhances areas of local maxima Fig. 16. A varying background can be a problem in some of the images, and background reduction by subtraction of an average filtered image is applied prior to signal detection. The average filtration is done three times with a 7x7 mean filter. The signals are separated from the image background by a user defined threshold set to a default value that localizes the major proportion of the signals. However, a lower threshold may be used in images with less background noise. The same threshold was used for all images when comparing methods of delineating the cytoplasm. In order to evaluate the influence of the threshold on the final measure of mutation load the thresholds were increased and decreased by 20%. The variation caused by these changes is shown as error bars in Fig. 18A. The binary image representing the signals is further reduced to single pixels by distance transformation and detection of local maxima. Thus, each signal event is represented by a single pixel.

3.5 Counting signals per cell

The counting of the signals per cell is the last step in the analysis. Since every signal event is represented by only one single pixel the counting is done fairly easy with just one scan over the image. Also, each cytoplasm has its own label representing that specific cell. The scan over the image starts in the top left corner and moves from left to right and top to bottom. When it reaches a green or red signal it checks in what cell it is located and add one red or green signal-count to that cell. When the scan is done all cells have two number representing the count of green and red signals in that particular cell. These data are shown in an excel sheet as a pop-up window with the use of *Microsoft Office Web Components*. The first column shows the name of the image analyzed, the second column shows the ID of the cell, the third and fourth column is the amount of red and green, respectively, signals present in that cell, the fifth and sixth column shows the fraction of red and green signals over the total amount of signals in that particular cell.

4 Results

The results consist of three parts. The first part is a comparison between three different methods of delineating cytoplasms. The second part, image based measurements of single cell mutation load are compared to measurements based on single cell PCR-RFLP (Polymerase Chain

Reaction-Restriction Fragment Length Polymorphism), a biochemical method that measures mutation load in single cells by quantifying DNA-fragment length variation This comparison was performed to validate the image based method of analysis. The third part is the result of the added functionality to VIS, the new tool for single cell analysis referred to as the Add-In.

4.1 Comparison of segmentation methods

For a full comparison of segmentation methods, accuracy, precision, and efficiency should be considered. The comparative study of methods for cytoplasm segmentation was performed on 9 images containing a total of 56 cells. Two fully automated image based segmentation methods, one using information from a cytoplasmic stain (referred to CS), and one not using information from a cytoplasmic stain (referred to as NCS) were compared to each other and to manual segmentation (refered to as O) of the same cytoplasms. Both automated methods are seeded from the same image of the cell nuclei, and both method used the same threshold t for the h-maxima transform. As no gold standard or ground truth is possible to produce, it is assumed that the manual segmentation method (O) results in the true delineation, defined as S_t^O. Manual segmentation was performed three times by two different persons to provide measurements of precision (reproducibility) in terms of inter- and intra- observer variability (refered to as O_{1a}, O_{1b} and O_2). The results can be seen in Table 1.

First of all, considering the accuracy, NCS and CS is significantly (α=0.05) less accurate than O. Between CS and NCS no significant (α= 0.05) difference can be seen in terms of accuracy. Furthermore, method O has noticeably lower precision than the other methods, as the computer based methods will reproduce the same result if re-run on the same image data, i.e., 100% precision, while manual segmentation varies both between observers (inter-observer precision is 79%) and for the same observer assessing the data at different times (intra-observer precision is 84%). Finally, the efficiency of NCS and CS is approximately 30 times higher than that of the manual segmentation O when using a 2.53 GHz Intel Pentium 4 processor and O requires human operation while NCS and CS are fully automatic.

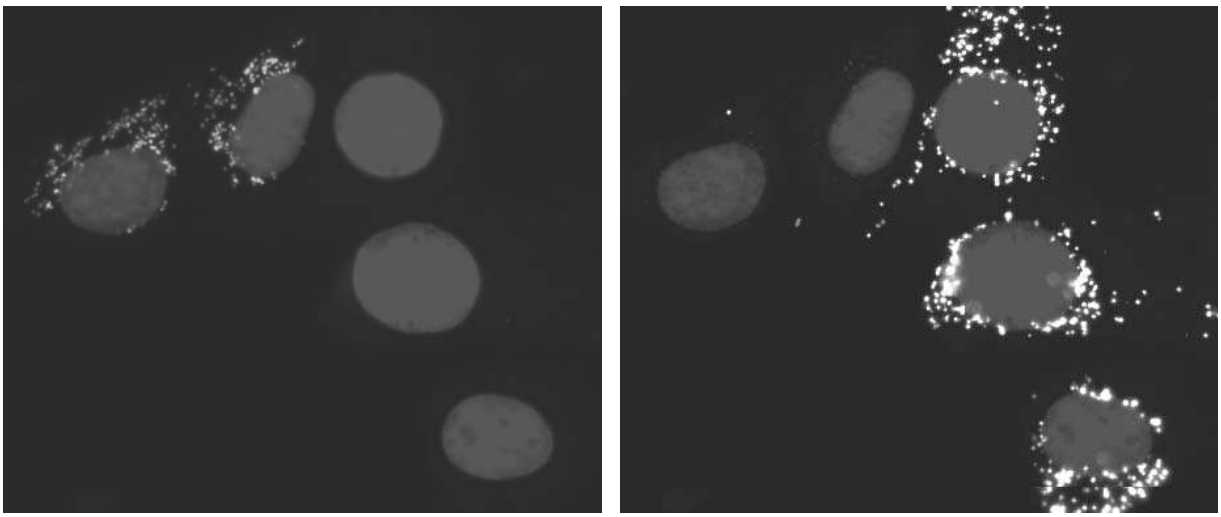
Comparison of methods for cytoplasmic segmentation shows that the addition of a cytoplasmic stain does not result in a significant increase in accuracy. This may seem strange as a cytoplasmic stain will guide the segmentation mask to the true edges of the cytoplasm. However, in the presented analysis, inclusion of parts of the image background does not affect the measurement of mutation load, as no signals are present in the background. Therefore, the inclusion of background as part of the false positive area fraction (FPAF) was not used. For other applications, e.g., if cytoplasmic area is to be measured, a segmentation method making use of the information from a cytoplasmic stain may be necessary. It is also worth mentioning that the agreement between manual cytoplasm segmentation and either of the fully automated methods is about the same as the agreement between manual cytoplasm segmentation performed by two different persons. The automated method not including a cytoplasmic stain turned out to be a sufficiently accurate and fast method for analysis of single cell mutation load. The fact that no cytoplsmic stain was included also allows the use of two different colors for mutant and wild type mtDNA without problems with overlapping fluorescence spectra. With all of these results in mind the NCS is sufficient enough to be used in the single cell analysis by the Add-In. For further verification, images containing cells with known ratios of mtDNA mutaion loads were analyzed by the Add-In using NCS in 4.2.

4.2 Image based measurement of mutation load vs. PCR-RFLP

Mutation load is the proportion of mutated mtDNA (number of red padlock signals) compared to wild type mtDNA (number of green padlock signals) per cell. The Add-In analysis was first performed on a padlock probed co-culture, meaning that cells with 100% wild type mtDNA

Table 1: Comparison of segmentation methods

Method	Accuracy				Precision	Efficiency
	vs.	TPAF	FNAF	FPAF	(%)	Cells/min
NCS	O_{1a}	0.87±0.03	0.14±0.03	0.12±0.04	100	30
CS	O_{1a}	0.85±0.03	0.16±0.03	0.11±0.03	100	30
O_2	O_{1a}	0.84±0.02	0.16±0.02	0.02±0.01	79	1
O_{1b}	O_{1a}	0.90±0.02	0.10±0.02	0.03±0.01	84	1

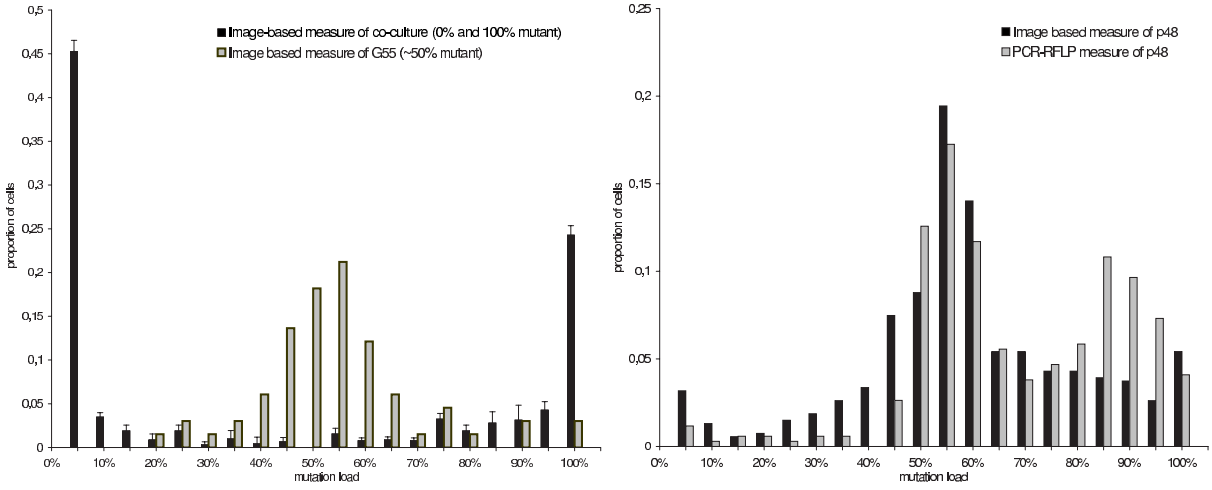


(a) (b)

Figure 17: Two views of the same image of co-cultured cells where padlock probes are seen as small spots and cell nuclei are shown in darker gray. (a) Image channel R and B, showing padlock probes against mutated DNA. (b) Image channel G and B, showing padlock probes against wild type DNA. In this data set, cells should either be 100% mutant or 100% wild type.

were mixed and cultured together with cells having 100% mutated mtDNA, see Fig. 17. This data set consisted of 29 images containing a total of 178 cells. A histogram of mutation load per cell measured from image data is shown in Fig. 18(a). The data from the co-culture shows distinct distributions at the extremes, i.e., cells with 100% and 0% mutation load. The Add-In performed very well considering hardly any intermediate levels were found. A large amount of intermediate levels would have been an indication of a high degree of error in the analysis method. Second, an analysis of a padlock probed culture (G55) of cells with a ∼50% mtDNA mutation load was made on 66 cells in 10 images. As predicted, the analysis from G55 has a clear peak close to 50% mutation load Fig. 18(a).

To study the segregation of mtDNA (originating from a heteroplasmic founder cell) different passages in the progression must be analyzed. Mutation loads of single cells from one passage were measured by the Add-In and compared with a PCR-RFLP based analysis of the same passage Fig. 18(b). The image analysis was done on 536 cells in 58 images.



(a) (b)

Figure 18: (a) Histogram of proportion of cell population against mutation load as achieved by image based measures. Error bars show variation caused when varying the threshold for signal detection. (b) Histogram of image and PCR-RFLP based measure of mutation load in cells from passage 48 of a clone known to be heteroplasmic for the A32n3G mtDNA mutation.

4.3 Analysis tool for VIS

The images intended for analysis are imported into the VIS by the import wizard to the image database (Fig. 19). There is no limit to the number of images being imported. Often color images are imported as separate images for each color channels (red, green and blue). When the images have been imported they can be visualized in VIS. The launch of the Add-In is done in the VIS workspace and a pop up window for the added functionality will appear (Fig. 20).

In Fig 20, the parameters for the Add-In are shown, they are all set to default values that will perform well in most cases. The first parameter, *Radius*, is the fixed radius of the cytoplasm delineation; a higher value will create a larger cytoplasm. The second parameter, *Min Nucleus Area*, is the smallest area (in pixels) that a nucleus can have, objects with a smaller area than this value will be removed. The third parameter, *Blob Threshold*, is the threshold value for finding the padlock-probe signals. The default value is 500 and finds most of the signals without including any background noise. If the image has less background noise a lower value can be chosen and more signals will be identified without falsely identifying signals in the background. The last three checkboxes decides what the Add-In should do. The first option, *merging of color bands*, is used if the image is separated into three grayscale images for each color band. With this function the three grayscale images will be merged into one RGB image. The second check box is for, finding the nucleus, delineation of cytoplasm, and identifying the padlock-probe signals. And the last checkbox is for counting the signals for each cell.

The analysis can be performed on one image or a set of images. If an analysis of several images is preferred one should choose the folder containing the images destined for analysis in the image database. After that, launch the Add-In and set the parameters preferred and click the batch process button in VIS (Fig. 19). The analysis will be performed on all the images at once. When the analysis is done the delineation of the cytoplasms will be seen as different masks and the detected red and green signals will have one label each (see section 2.3.3 regarding masks and labels in VIS). The maks (cytoplams) and labels (signals) can easily be edited with the mouse if the result is not satisfying enough. After the changes to the masks and labels the

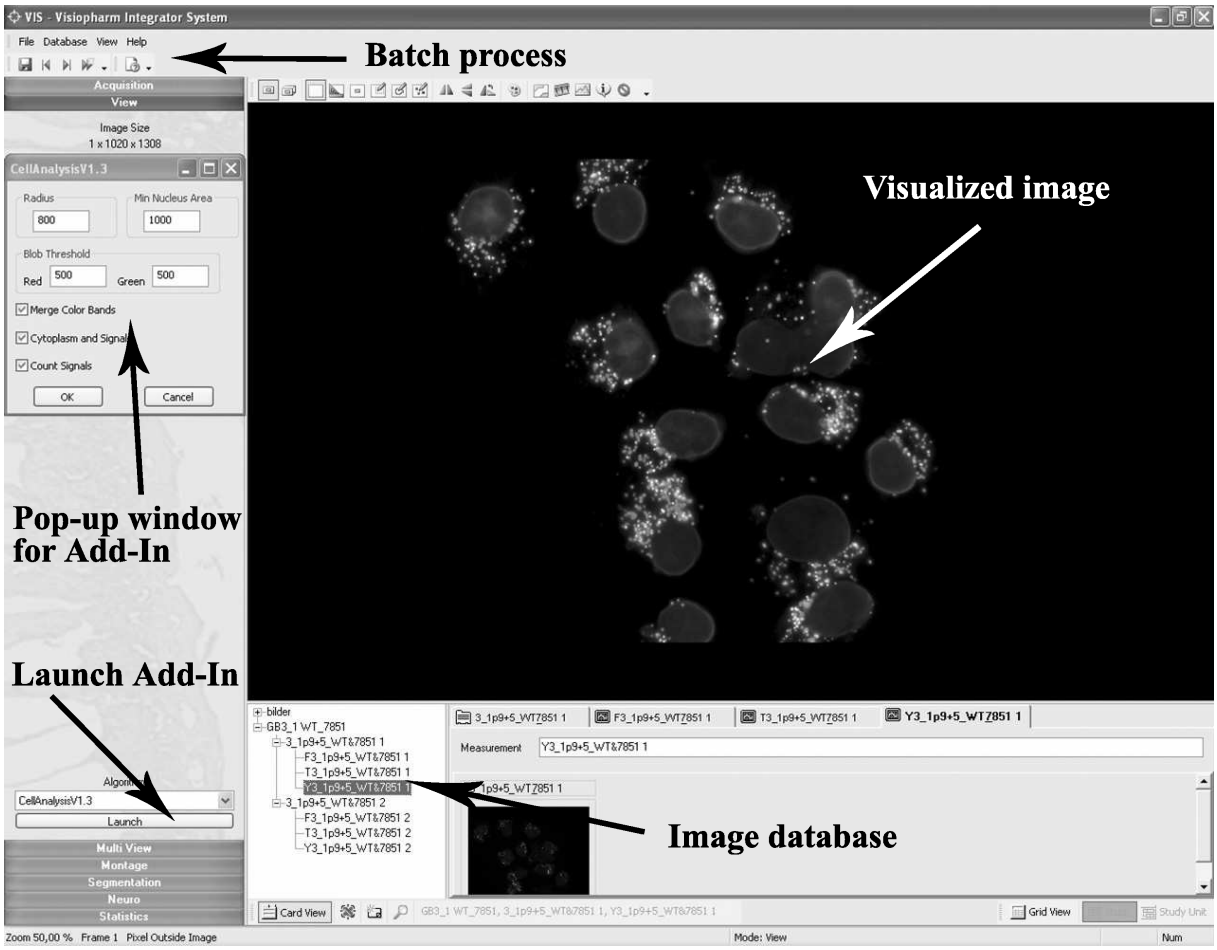


Figure 19: Workspace in VIS, arrows showing; image database, launch of Add-In, Add-In pop-up window, visualization of image and batch process button.

count signals checkbox can be checked and a new count of signals per cell will be performed. The data from the analysis is viewed in an excel sheet as a pop-up window where further statistical analysis can be done. A result from a typical analysis is seen in Fig. 21. An analysis of 50 images, with all the steps included, takes approximately 15 min (20 sec/image) and user input is only required before the initiation of the analysis and, if needed, after the analysis is done.

5 Future work

Most of the signals are detected in an image but some of the weaker signals are not detected when using the default settings of the Add-In. Lower settings can be chosen with the risk of falsely detecting signals from background. This is something that probably can be improved by trying several other approaches to the signal detection.

The delineation of the cytoplasm is a crucial step in single cell analysis and may be the biggest contribution of errors in this type of analysis. If cells lie close together and are very irregular in shape, the error of the fixed radius method would increase. In these cases another approach may be needed in order to get accurate results. Moreover, if cytoplasmic area is to be measured, a segmentation method making use of the information from a cytoplasmic stain may be necessary to include. One approach would be to try to find a cytoplasmic stain that is compatible with the padlock probe and RCA. Also the problem of spectral overlap must be

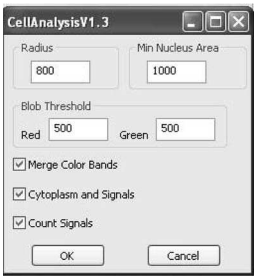


Figure 20: Add-In pop-up window.

solved, one approach could be by doing several stainings with the same color and washing the sample between the stainings.

The Add-In takes approximately 20 sec to analyze an image which might be too much when many images need to be analyzed. Not much time has been spent on optimizing the execution time of the software and a thorough optimization could probably decrease the execution time significantly.

The Add-In has been tested and developed for the detection of padlock probes in cultured cells. Images of ,e.g ,. tissue cells may not be as clear and defined as these cultured cells and the analysis may therefore not work as well in these types of cases. The Add-In could therefore be extended and improved to work for a wider spectrum of applications. As the ENLIGHT project mainly focuses on tissue samples this type of improvment would be of great interest.

Another thing that could be of interest in future work is to add the possibility of saving and loading different Add-In parameter configurations. As of now, the default values appear at startup and changes made to them will be deleted the next time the application is started. This added functionality would make it possible to load a file containing specific parameter configurations for a specific application.

6 Conclusion

Single cell analysis of mitochondrial mutation load detected with padlock probes and RCA is an important aspect of the ongoing ENLIGHT project, but brings some complexities. As of now, no cytoplasmic stain compatible with this application is available and still the cytoplasm needs to be delineated. By using a fixed radius approach of delineating the cytoplasm an accuracy as high as 87% can be achieved when compared to a manual delineation. This is sufficient enough for performing an analysis that in the end will give satisfying results. By using the windows based user friendly software VIS together with the newly added functionality the analysis can be done without any extensive knowledge in image analysis. Furthermore, only few input parameters that often need not to be changed makes the analysis even easier to carry out. By using the masks and label functions in VIS, the segmentation results from the analysis can easily be edited before the final count of the signals. The data export to excel sheets makes it possible to easily do further statistical analysis of the result. The analysis is not perfect and many aspects can be improved and optimized. In addition, the Add-In has been developed and tested for this specific application but in the future it could probably be extended to work for wider spectrum of applications.

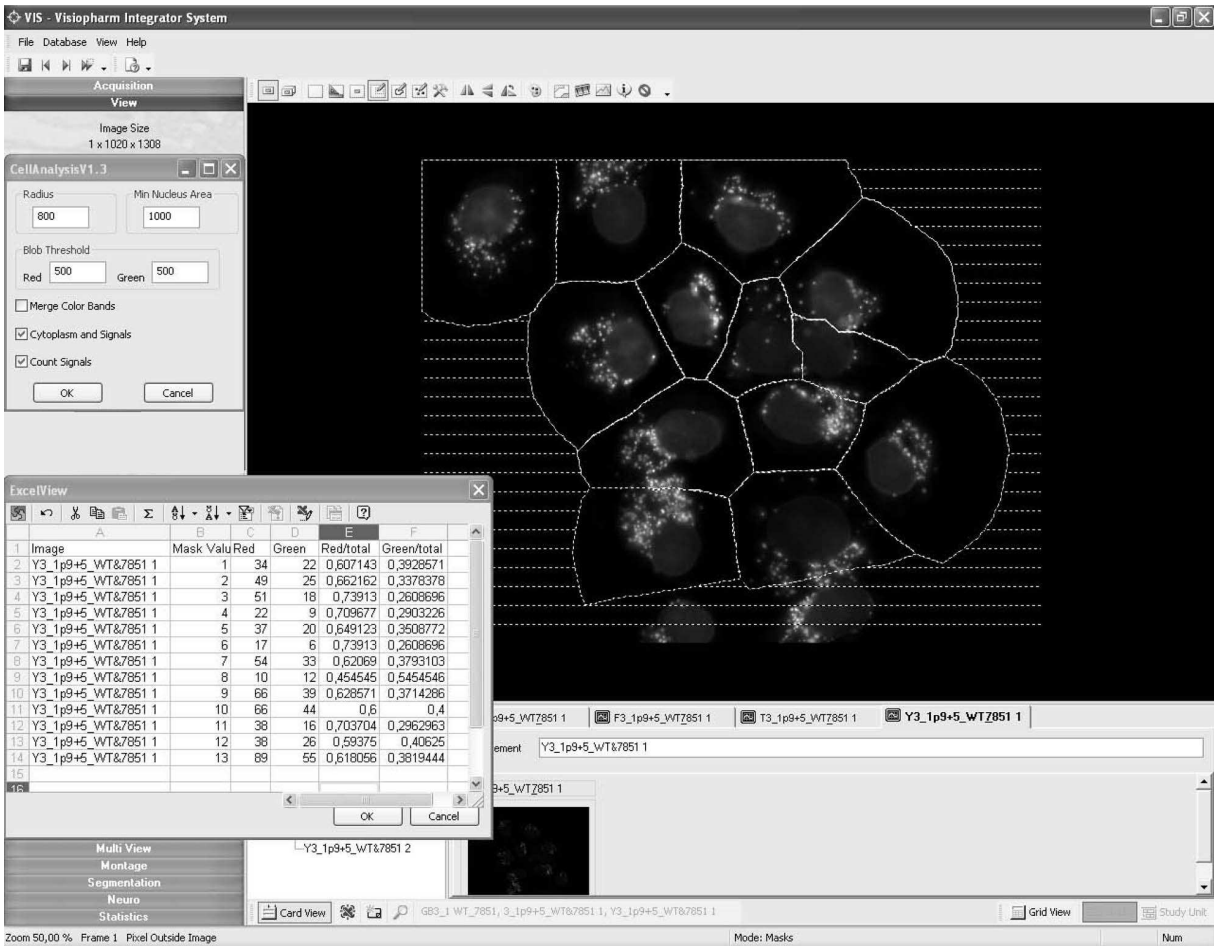


Figure 21: VIS workspace after analysis with the Add-In.

7 Acknowledgment

First of all, I would like to thank my supervisor Carolina Wählby for all her help and support during the work of this thesis. Also, I would like to thank A.K Raap and F.M van de Rijke at Leiden University Medical Center for providing me with images and making my visit to the Netherlands a pleasant one. Moreover, I want to thank all the staff at Visiopharm, Hörsholm, Denmark, for help with integration of new functionality in VIS. I would also like to thank the EU-Strep project ENLIGHT for providing me with this opportunity. Finally, thank you to everyone at Center for Image Analysis for taking their time to help me with various problems during this period.

8 Abbreviations

CCD	Charge-Coupled Device
CS	Cytoplasmic Stain
ENLIGHT	ENhanced LIGase based Histochemical Techniques
EU	European Union
FISH	Flourescence In Situ Hybridization
FNAF	False Negative Area Fraction
FPAF	False Positive Area Fraction
I/O	Input/Output
IDE	Integrated Development Environment
IMP	IMage Processing
IU-SDK	Image Utility-Software Development Kit
MSVC	Microsoft Visual C++
mtDNA	mitochondrial DNA
NCS	No Cytoplasmic Stain
PCR-RFLP	Polymerase Chain Reaction-Restriction Fragment Length Polymorphism
RCA	Rolling Circle Amplification
ROI	Region Of Interest
SNP	Single Nucleotide Polymorphism
TPAF	True Positive Area Fraction
tRNA	transfer-RNA
VIS	Visiopharm Integrator System

9 References

References

[1] A. Eames. Bonus for cancer tissue project. *BiotechSweden*, (8), 2006-11-07.

[2] C. Larsson, J. Koch, A. Nygren, G. Janssen, A. K. Raap, U. Landegren, and M. Nilsson. In situ genotyping individual DNA molecules by target- primed rolling- circle amplification of padlock probes. *Nature Methods*, 1:227–232, 2004.

[3] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular biology of the cell*. Garland Publishing, Inc, New York, 1994.

[4] T. Acharya and A. K. Ray. *Image processing principles and applications*. A Wiley-Interscience Publication, New Jersey, 2005.

[5] R. C. Gonzales, R. E. Woods, and Steven L. Eddins. *Digital Image Processing Using Matlab*. PEARSON Prentice Hall, New Jersey, 2004.

[6] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. on System Man and Cybernetics*, 9(1):62–69, 1979.

[7] P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, Berlin, 1999.

[8] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics and Image Processing*, 34:344–371, 1986.

[9] C. Lantuéjoul and S. Beucher. On the use of geodesic metric in image analysis. *Journal of Microscopy*, 121:39–49, 1981.

[10] J. Little and C. Moler. The Mathworks, Inc. www.mathworks.com, Accessed 2006 Dec 30.

[11] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–597, 1991.

[12] J. Lindblad, C. Wählby, E. Bengtsson, and A. Zaltsman. Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation. *Cytometry*, 57(1):22–33, January 2004.

[13] C. Wählby, J. Lindblad, M. Vondrus, E. Bengtsson, and L. Björkesten. Algorithms for cytoplasm segmentation of fluorescence labelled cells. *Analytical Cellular Pathology*, 24:101–111, 2002.

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

[15] J. K. Udupa, V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn. A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 30:75–87, 2006.