Cytometry
PART A
Journal of the International Society for Advancement of Cytometry

Single-Cell-Based Image Analysis of High-Throughput Cell Array Screens for Quantification of Viral Infection

Petr Matula,[1,2,3*] Anil Kumar,[4] Ilka Wörz,[4] Holger Erfle,[5] Ralf Bartenschlager,[4] Roland Eils,[1,2] Karl Rohr[1,2]

[1]University of Heidelberg, Department of Bioinformatics and Functional Genomics, BIOQUANT, IPMB, Heidelberg, Germany

[2]German Cancer Research Center (DKFZ), Department of Theoretical Bioinformatics, Heidelberg, Germany

[3]Masaryk University, Faculty of Informatics, Center for Biomedical Image Analysis, Brno, Czech Republic

[4]University of Heidelberg, Department of Molecular Virology, Heidelberg, Germany

[5]University of Heidelberg, BIOQUANT Center, Heidelberg, Germany

Received 15 July 2008; Accepted 22 September 2008

A preliminary report of this work was presented at the SPIE Medical Imaging conference (San Diego, February 2008) and the Thirteenth Leipziger workshop (Leipzig, April 2008).

Grant sponsor: BMBF (FORSYS) project VIROQUANT; Grant sponsor: The Ministry of Education of the Czech Republic; Grant number: 2B06052.

*Correspondence to: Dr. Petr Matula, University of Heidelberg, BIOQUANT, Department of Bioinformatics and Functional Genomics, Biomedical Computer Vision Group, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany.

Email: p.matula@dkfz.de

Published online 11 November 2008 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20662

ISAC
International Society for Advancement of Cytometry

• Abstract

The identification of eukaryotic genes involved in virus entry and replication is important for understanding viral infection. Our goal is to develop a siRNA-based screening system using cell arrays and high-throughput (HT) fluorescence microscopy. A central issue is efficient, robust, and automated single-cell-based analysis of massive image datasets. We have developed an image analysis approach that comprises (i) a novel, gradient-based thresholding scheme for cell nuclei segmentation which does not require subsequent postprocessing steps for separation of clustered nuclei, (ii) quantification of the virus signal in the neighborhood of cell nuclei, (iii) localization of regions with transfected cells by combining model-based circle fitting and grid fitting, (iv) cell classification as infected or noninfected, and (v) image quality control (e.g., identification of out-of-focus images). We compared the results of our nucleus segmentation approach with a previously developed scheme of adaptive thresholding with subsequent separation of nuclear clusters. Our approach, which does not require a postprocessing step for the separation of nuclear clusters, correctly segmented 97.1% of the nuclei, whereas the previous scheme achieved 95.8%. Using our algorithm for the detection of out-of-focus images, we obtained a high discrimination power of 99.4%. Our overall approach has been applied to more than 55,000 images of cells infected by either hepatitis C or dengue virus. Reduced infection rates were correctly detected in positive siRNA controls, as well as for siRNAs targeting, for example, cellular genes involved in viral infection. Our image analysis approach allows for the automatic and accurate determination of changes in viral infection based on high-throughput single-cell-based siRNA cell array imaging experiments. © 2008 International Society for Advancement of Cytometry

• Key terms

image analysis; cell nucleus segmentation; quantification of viral infection; siRNA screening; cell-based arrays; immunofluorescence microscopy; image quality control

INTEREST in understanding virus-host cell interactions has increased in recent years (1). It is believed that targeting specific proteins within the host cell instead of viral components may lead to significant improvements in antiviral treatments. Host functional genetic profiling using RNA interference (RNAi) screens provide a systematic approach to obtain a comprehensive overview of cellular pathways that are exploited by viruses. RNAi screening allows for the systematic knock down of each gene of the host cell's genome (2) to determine the effect of cellular gene silencing on virus infection (3).

Our overall aim is the genome-wide identification of cellular genes involved in virus entry and replication. To screen for changes in virus infection when knocking down a certain cellular gene we use small interfering RNA (siRNA) cell arrays in combination with high-throughput fluorescence microscopy (4). With this approach, siRNA and transfection reagents are robotically spotted on a chamber plate at known locations in a grid pattern. As cells are cultured and treated on the printed plates, only those cells located within a printed spot area take up siRNA to undergo gene

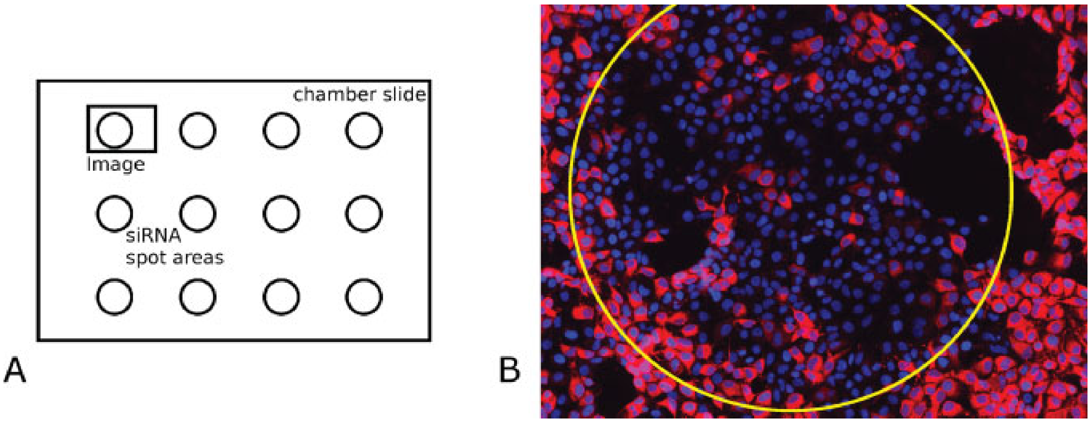

Figure 1. Layout of the experiment and image data example. **A**: Cell array containing M × N siRNA spot areas, **B**: Example of a two-channel image corresponding to one spot area with overlaid marked spot area (diameter of 400 μm). Typically there are 150–350 cells within one spot area. Cell nuclei are shown in blue (Channel 1) and corresponding viral protein expression in red (Channel 2).

silencing. For each spot area one two-channel image is acquired using a fully automated fluorescence microscope. The advantages of cell arrays over multi-well plates, which are often used for RNAi screening (5,6), are identical treatment of all samples with respect to cell seeding, infection, and staining, as well as significantly reduced costs due to lower reagent usage.

Genome-wide screens with more than 20,000 genes can generate more than 200,000 fluorescence images and therefore fully automatic, efficient, and robust image analysis methods are needed. In recent years, a number of approaches for cell nuclei or whole cell segmentation of fluorescence microscopy images has been reported (e.g., 7–11). In high-throughput applications, approaches based on adaptive thresholding (e.g., 12,13) yielded good results for cell nuclei segmentation, in particular, with unclustered nuclei. To separate clusters of nuclei, watershed-based techniques (e.g., 12,14,15) and approaches employing geometric properties (16) have been proposed. Recently, an approach based on multiscale entropy-based thresholding and region merging has been described (17). An approach based on the zero-crossings of the Laplacian was used in (18). For the segmentation of whole cells or the cytoplasm approaches based on deformable models (19), Voronoi diagrams (20), or a combination of both (21) have been proposed. Nevertheless, none of these approaches can cope with all image types and one has to carefully select and adapt algorithms for a particular application.

The key tasks for image analysis in our application are cell nucleus segmentation, detection of regions with transfected cells (siRNA spots), quantification of the virus infection level, and quality control of single images as well as whole plates. In this article, we describe the workflow and image analysis approaches of a system addressing these key tasks. In particular, we propose novel approaches for (i) the segmentation of cell nuclei using a gradient-based thresholding scheme which does not require subsequent postprocessing steps for separating clustered nuclei, (ii) the segmentation of viral protein expression in the cytoplasm based on a combination of a watershed algorithm and region growing, (iii) the localization of regions with transfected cells within cell array images, (iv) cell classification as infected or noninfected, and (v) the detection of out-of-focus images for the purpose of automatic quality control which increases the discrimination power by taking into account only pixels near segmented boundaries. We have evaluated the main algorithms of our system and compared their performance with that of previous approaches. Our whole approach was applied to a large number of images from siRNA high-throughput screens of cells infected by either dengue virus or hepatitis C virus. To our best knowledge, this is the first image-based approach for automatic single-cell-based quantification of viral replication from large datasets, which measures the level of virus replication in transfected cells only, and allows for image quality control.

Materials and methods

Image Analysis Workflow

In our application, the input for image analysis are two-channel images acquired from a chamber plate with printed siRNA spots. Each image corresponds to one siRNA spot (Fig. 1). The first channel displays DAPI stained cell nuclei. The second channel represents expressed fluorescently stained viral protein. To analyze these images we have developed a workflow and image analysis approaches as depicted in the diagram in Fig. 2. First, the cell nuclei are segmented in Channel 1. Then, a binary mask of pixels defining the neighborhood of each nucleus is computed and the viral signal is measured in Channel 2. Next, siRNA spot areas with transfected cells are localized exploiting prior knowledge about the layout of the cell arrays. Then, the cells are classified as infected or noninfected. Finally, the quality of single images and of a whole plate is checked. Each step is described in detail below.

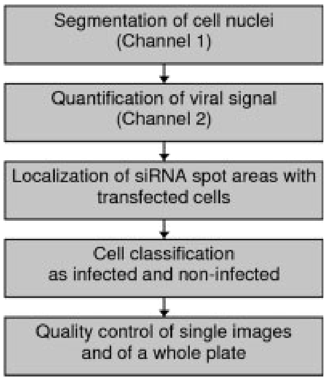
Segmentation of cell nuclei (Channel 1)
Quantification of viral signal (Channel 2)
Localization of siRNA spot areas with transfected cells
Cell classification as infected and non-infected
Quality control of single images and of a whole plate

Figure 2. Image analysis workflow.

Segmentation of Cell Nuclei

The cell nuclei in our images manifest themselves as bright objects on a dark background. The average intensity is related to the DNA density and varies for different nuclei as well as depends on the cell cycle. Also, the images generally suffer from uneven illumination and varying staining. Therefore, it is difficult to find a single global threshold suitable for all nuclei. Even if adaptive thresholding approaches are used and the thresholds are determined in image sub-regions, it is difficult to cope well with clustered cell nuclei. Here, we propose an edge-based approach which analyzes gradient magnitude images instead of thresholding the original image intensities.

In the 1D case, edges can be determined by computing the gradient (first derivative) of the signal and locating points that have locally maximal gradient magnitudes. In 2D case, edges correspond to ridge points of gradient magnitude images. Points close to edges can be determined by thresholding the gradient magnitude image. An alternative approach for edge finding is to calculate second derivatives of an image and to locate zero-crossings of the Laplacian operator. A nice property of the Laplacian operator is that zero-crossings always produce closed boundaries. Another property is that for bright objects on a dark background the Laplacian is positive on the dark side of an edge and negative on the bright side of an edge (22). In our approach we exploit this property to distinguish between interior and exterior cell nuclei parts and to improve the segmentation result for clustered cells.

The main idea of our approach is to detect pixels on the bright side of edges that have a large gradient magnitude and to subsequently morphologically process the result to obtain the final segmentation. We have observed that cell nuclei with different average intensities that are close to each other are often well separated by determining regions for which the Laplacian operator yields negative values (Fig. 3B). Relevant edge regions are detected by thresholding the gradient magnitude (Fig. 3C). Combining the results of these two operations improves the separation of clustered nuclei without requiring additional postprocessing steps.

The whole scheme for nuclei segmentation consists of six main steps (see Fig. 3): In the first step, a binary image f is obtained from the input image g by combining the results of the gradient magnitude and the Laplacian operator (Figs. 3B–D):

$$f(x,y) = \begin{cases} 1 & \text{if } |\nabla g| > T \text{ and } \nabla^2 g < 0, \\ 0 & \text{otherwise,} \end{cases}$$

where, $\bigtriangledown$ denotes the Nabla operator, $|\nabla g| = \sqrt{g_x^2 + g_y^2}$, $\nabla^2 g = g_{xx} + g_{yy}$, with g_x, g_y, g_{xx}, g_{yy} denoting first and second order partial derivatives of g. To reduce the noise influence the image is smoothed by a Gaussian before applying the Nabla and Laplacian operators. In both cases, we used $\sigma = 1$ for the standard deviation of the Gaussian (which was determined based on experimental experience). We automatically determine the threshold T by using the unimodal background symmetry method (23), which assumes that there is one dominant peak in the histogram corresponding to the background pixels (outside edges). The position p of the peak is found and the width w at half of the maximum is computed. The threshold T is set to $p + kw$, where k is a parameter (in our case we used $k = 1/2$).

In the second step of our approach, small objects are removed (Fig. 3E). An object is considered small if it has a less number of pixels than a given constant. This constant was determined by analyzing the histogram of the sizes of segmented objects in a number of images after the first step (in our application, we used 80 pixels for this constant). Third, connected components of pixels are labeled (Fig. 3F). In this step a unique identifier is assigned to each 8-connected component of pixels to identify the objects. In the fourth step, each object is morphologically closed with a small disk structuring element, i.e., background structures that cannot contain the structuring element are added to the object (Fig. 3G). We used a 3 × 3 structuring element (24). In this step each object is treated individually preventing merging of objects with different identifiers. Fifth, holes in each object are filled by a standard hole-filling algorithm (24) (Fig. 3H), and finally, cell nuclei are identified (Fig. 3I) based on size, intensity level, and circularity. The appropriate range for each feature was determined based on an analysis of real image data and experimental experience.

Quantification of Changes in Virus Infection

For quantifying changes in viral infection after knocking down a certain cellular gene, there exist two main approaches. Either (i) the level of expressed viral protein (i.e., the viral signal) can be measured for each cell and the changes are quantified for each siRNA spot, e.g., by comparing the mean viral signal over all cells in the siRNA spot or (ii) the percentage of infected cells (called infection rate) can be computed for each siRNA spot and the changes in infection rates are studied. In both cases the viral signal must be determined for each cell. In the second approach, the cells must also be classified as infected or noninfected.

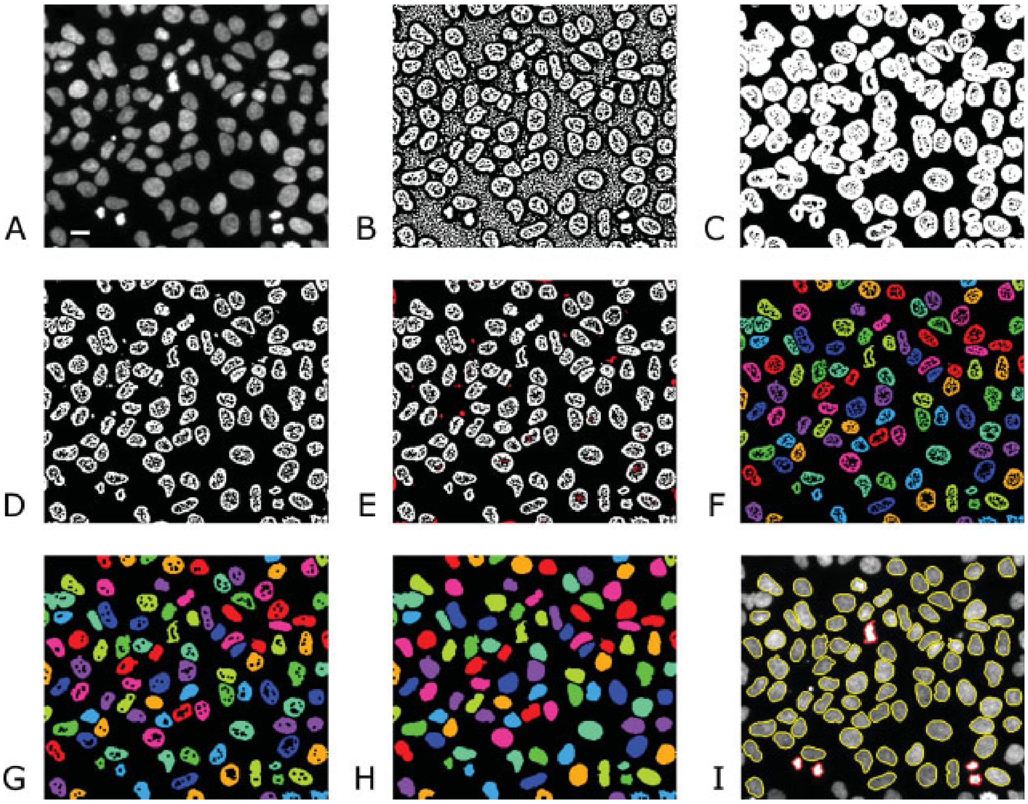

Figure 3. Segmentation of cell nuclei. **A**: Section of an input image. **B**: Regions with negative Laplacian of Gaussian (pixels colored white). **C**: Result of thresholding the gradient magnitude image. **D**: Intersection of binary images in B and C (''AND'' operation). **E**: Small objects to be removed (colored red). **F**: Labeled connected components. **G**: Objects after morphological closing. **H**: Objects after hole-filling. **I**: Overlay of segmentation results with the original image. Objects recognized as cells (yellow outline) and excluded objects (red outline). The scale bar is 10 μm (Fig. A, bottom left).

In our case, we compute the viral signal for each segmented cell nucleus in its neighborhood by computing the mean of the pixel values in the virus signal channel (Channel 2). To prevent overlaps between neighborhoods of different cells we partition an image into influence zones (IZs) of segmented nuclei (Figs. 4A and 4B). The IZs are computed using a seeded watershed transform of the Gaussian filtered ($\sigma = 1$) and inverted virus channel (Channel 2) with the segmented cell nuclei as initial seeds. We have implemented three different approaches for defining the neighborhood of a cell nucleus based on: (i) dilating the cell nucleus mask (Fig. 4C), (ii) dilating the cell nucleus mask inside its IZ (Fig. 4D), and (iii) region growing inside IZ (Fig. 4E). The region growing algorithm is started from the pixels at the cell nucleus boundary. For each IZ, the mean (μ_{IZ}) and standard deviation (σ_{IZ}) of the pixel values at the cell nucleus boundary are computed. All pixels within an IZ with intensities inside the range $[\mu_{\mathrm{IZ}} - k\sigma_{\mathrm{IZ},}\ \mu_{\mathrm{IZ}} + k\sigma_{\mathrm{IZ}}]$ that are connected to the cell nucleus boundary are included (we used $k = 1$). Connected pixels to the cell nucleus boundary are determined using morphological reconstruction by dilation (24).

The cells are classified as infected or noninfected according to the estimated viral signal. Cells with a viral signal less than a threshold are classified as noninfected, whereas the others are classified as infected. The threshold is determined automatically by maximizing the difference in infection rates between positive and negative controls. In positive controls the viral protein production is blocked and therefore the signal is reduced. In negative controls the virus replication is not altered. To detect changes in viral infection we use a measure denoted as infection rate ratio, which is defined by $\mathrm{IRR} = \frac{\mathrm{IR}_i}{\mathrm{IR}_\mathrm{N}}$ where IR_i is the infection rate within the siRNA spot i and IR_N is the normal infection rate of a virus, i.e. the percentage of infected cells without knocking down a certain gene.

Determination of Circular Regions with Transfected Cells

Only cells that are located within printed siRNA spots can be transfected and only those should be quantified (note that changes in the viral signal can be observed only in transfected cells). Cells outside printed circular regions (siRNA spots) cannot take up siRNA (neglecting rare migration of cells) and thus should exhibit a normal virus signal. Therefore, it is important to localize siRNA spot areas with transfected cells to select cells that are located within printed siRNA spots. The problem of finding siRNA circular areas is not easy because cells are present on the whole plate and if a particular siRNA has no effect on virus infection there is no change in viral signal, which could be exploited to find the spot. On the

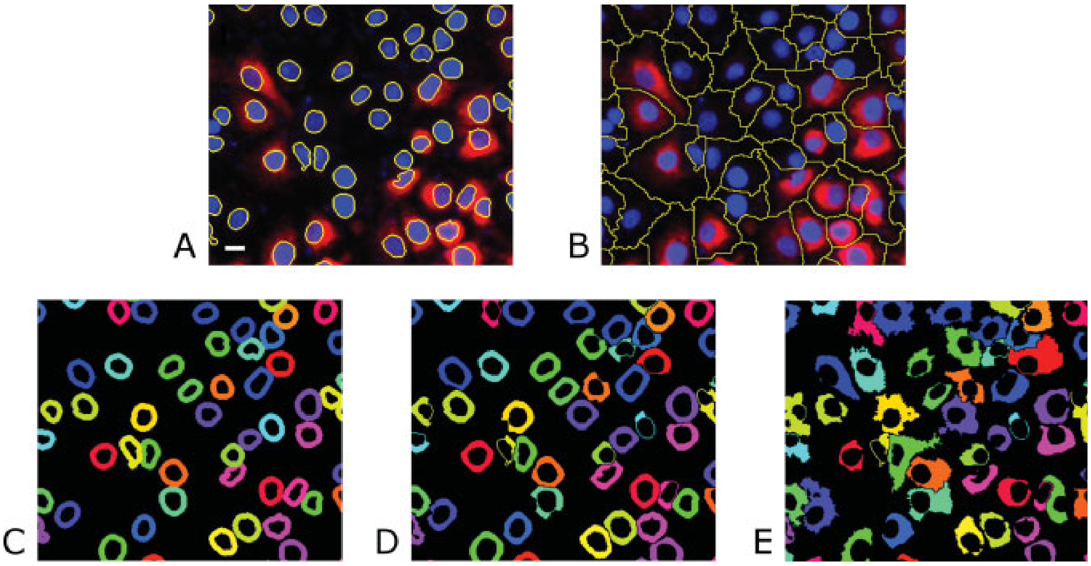

Figure 4. Quantification of the viral signal. **A**: Section of an input image with overlaid contours of segmented cell nuclei. **B**: Influence zones of cell nuclei (IZ). **C**: Cell nucleus neighborhoods for the quantification of the viral signal based on dilation (SD). **D**: Cell nucleus neighborhoods based on dilation restricted by IZ (RD). **E**: Cell nucleus neighborhoods based on region growing inside influence zones (RG). The scale bar is 10 μm. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

other hand, if the knockdown of a certain gene leads to a change in virus signal a clear difference between transfected and nontransfected cells is observable (e.g., Fig. 1B).

A usual approach for localizing siRNA spot areas in cell arrays is to manually determine a circle within the first image and to apply the same position for the other images. In this case, the position of the first spot is marked on the plate with a pen by a biologist in a laboratory. However, using the same spot position for all images does not take into account possible tilting of the whole plate and hence errors are generally introduced. To improve the accuracy, we have developed an automatic approach for localizing siRNA spots directly from the image data. This approach combines local information extracted from single images and global information about the dimension and layout of the printed spots.

The idea is to detect siRNA spot areas in images with altered viral protein expression and to extrapolate the locations of the areas to other images by using prior knowledge about the printed grid (spot diameter, spotting distance). Our approach for localizing siRNA spots consists of two steps:

1. For each image g of a plate, a position $[x_c^g, y_c^g]$ of a circle of known fixed diameter d is found, for which the difference d^g in the mean viral signal μ_{IN} of cells inside the circle and the mean viral signal μ_{OUT} of cells outside the circle is maximal, i.e.

$$[x_c^g, y_c^g] = \underset{[x,y]}{\operatorname{argmax}}(d^g(x,y)) = \underset{[x,y]}{\operatorname{argmax}}(|\mu_{\mathrm{IN}}(x,y) - \mu_{\mathrm{OUT}}(x,y)|)$$

2. All images with differences d^g in a certain range are selected and a grid of known parameters is fitted to the computed circle positions $[x_c^g, y_c^g]$ using a least-squares approach. The idea behind selecting only differences in a certain range is to choose only those images from Step 1 for which the siRNA spot was detectable.

The appropriate range of differences was determined based on simulations. We found that too large or too small differences were correlated with erroneous estimations of circle positions.

Quality Control

Since in high-throughput screening applications a large number of images need to be analyzed and some of them may be of poor quality (e.g., out-of-focus, no cells in certain areas, image artifacts), we need algorithms that can assess the quality of the data to exclude failures from statistics. Quality checks can be carried out on two levels: on the whole plate level and on the single image level.

On the whole plate level, the main goal is to sort out unsuccessful experiments. To this end, statistical parameters can be computed based on the results for positive and negative controls (25). In our approach, we use measures which are directly computed from the images. First, we calculate the percentage of saturated pixels in Channel 2 (viral protein). To limit the effect of image artifacts we determine saturated pixels only in the neighborhoods of identified cells. A high percentage of saturated pixels are related to overexposure. Overexposed plates can be excluded and the images can be reacquired with decreased exposure times.

To visualize whole plate related problems, e.g., due to improper staining or cell seeding, we have implemented a graphical user interface (GUI) which displays all images of a plate in one overview tiled image. A user can view the original

Table 1. Performance comparison of approaches for cell nucleus segmentation

SEGMENTATION APPROACH	SEGMENTATION ACCURACY (%)	COMPUTATION TIME (s)
ATO	78.5	5.3
ATO + SCC	95.8	15.3
GBT	97.1	7.6

For all approaches, a hole-filling step was included. The values for the computation time are mean values over 9 different images (Dual Core AMD Opteron Processor, 2.6 GHz, 64 bit Linux).

ATO, Adaptive thresholding by Otsu's method; ATO + SCC, ATO followed by a model-based strategy for separating cell clusters; GBT, Novel gradient-based thresholding scheme.

data, the segmentation results, as well as quality tags of single images. The images are tagged automatically as described below. The GUI plays a significant role especially during optimization of the sample preparation and validation of the automatic quality control of an experiment.

On the single image level, we automatically tag images as "low quality" if (i) the number of cells is outside a given range or if (ii) the images are classified as "out-of-focus". The first criterion enables excluding images with a too small or too large number of cells, which is typically related to uneven seeding. This criterion also excludes most out-of-focus images because no or only a small number of cell nuclei are usually recognized in blurred images. The images satisfying criterion (i) are classified as "out-of-focus" if the average gradient magnitude calculated from image regions near the boundaries of segmented cell nuclei is lower than a threshold. Our approach is motivated by the fact that the boundaries of segmented objects are related to image edges and the gradient magnitude of an edge point is related to image sharpness. The threshold is computed as the mean minus three standard deviations of the gradient magnitude calculated from all images of a plate. To increase the robustness of estimating the mean and standard deviation we exclude average gradient magnitude values larger than a certain threshold. This threshold was determined by the unimodal background symmetry method (23) as described earlier (we used $k = 1$).

Samples and Image Acquisition

We have applied our whole approach to real image data of hepatitis C virus (HCV) and dengue virus (DV) high-throughput screening experiments. The cell arrays were prepared on chambered cover glass tissue culture plates (Nalge Nunc, Wiesbaden, Germany). Transfection reagents together with siRNAs were printed typically in a 12 × 32 grid using a chip writer compact robot (Bio-Rad) with solid pins (Point Technologies) resulting in a spot diameter of ~400 μm. The siRNAs used in our experiments were taken from kinase or cytoskeleton libraries (Ambion). Seven different plates are needed for a kinase experiment and four different plates for a cytoskeleton experiment. Each experiment was repeated several times (between four and eight repetitions). Cells of a human hepatoma cell line (Huh7 (26) for DV and Huh7.5 (27) for HCV, respectively) were seeded on spotted plates and incubated for a certain time period (24–48 h). Subsequently, the cells were infected with a virus, namely a green fluorescent protein (GFP)-tagged HCV (28) or with DV 2 (strain New Guinea C). After 24–48 h the cells were fixed and fluorescently stained. Cell nuclei (cellular DNA) were labeled by DAPI and a viral protein was labeled by immunofluorescence. For DV we used primary antibody anti DENV E mouse monoclonal, HB46 [American Type Culture Collection (ATCC), Manassas, VA] and secondary antibody anti-mouse Alexafluor 546, (Invitrogen, Karlsruhe, Germany). For HCV we used a monoclonal anti-GFP-antibody (Roche Diagnostics GmbH, Mannheim, Germany) and secondary antibody anti-mouse Alexafluor 546, (Invitrogen, Karlsruhe, Germany). For each siRNA spot area and each channel one grayscale image was acquired using the high-content scanning system Scanˆ R (Olympus, Heidelberg, Germany) with magnification 10×, NA = 0.40, and CCD camera pixel size 6.45 μm × 6.45 μm. The typical image size is 1,344 × 1,024 pixels. As an output of the high-throughput scanning, we typically obtain a set of 384 two-channel images for each plate of HCV or DV experiments.

Results

Prior to applying the overall approach to real image data, we tested and evaluated single algorithms and compared them with previous approaches.

Segmentation of Cell Nuclei

The cell nucleus segmentation algorithm was evaluated using real image data, where ground truth was obtained from two experts who marked cell nuclei in randomly selected real images from different experiments (in total 1,914 nuclei were used). We compared the results of our approach with two previous approaches, namely adaptive thresholding by Otsu's method (29) (ATO) as well as ATO followed by a three-step model-based strategy for separating cell clusters (SCC) based on the watershed transform as implemented in CellProfiler (15,30). With ATO 78.5% of cell nuclei were correctly segmented. ATO followed by SCC yielded 95.8% and using our gradient-based thresholding (GBT) approach we achieved a segmentation accuracy of 97.1% (1,859 correctly segmented nuclei), see Table 1. Our approach was particularly superior to ATO in segmenting clustered cells (compare Figs. 5A and 5C). SCC significantly improved the results obtained by ATO (compare Figs. 5A and 5B), but failures in separating elongated cell nuclei were observed (Fig. 5B). The reason for this is that SCC assumes a circular shape of the nuclei. We also measured the computation time averaged over nine different images. ATO required 5.3 s and ATO followed by SCC needed 15.3 s. The computation time of our approach turned out to be 7.6 s (see Table 1).

Quantification of the Virus Signal

We studied the effect of using different cell nucleus neighborhoods for quantifying the virus signal, which is computed as the mean intensity in Channel 2 in the given neighborhood. For 112,067 cells in total from one plate, we measured the

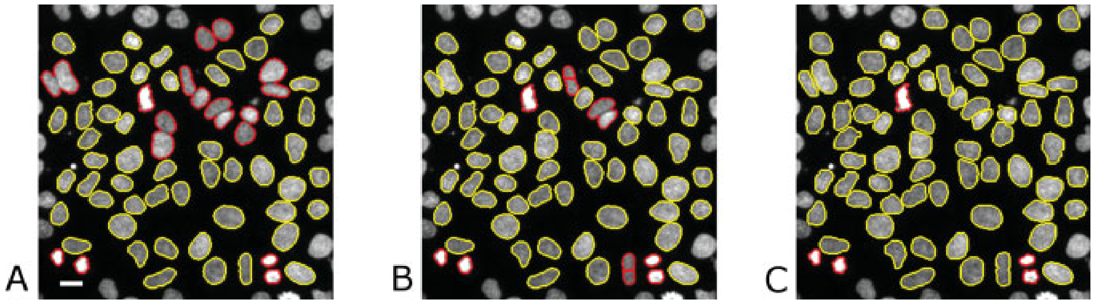

Figure 5. Comparison of approaches for cell nucleus segmentation. **A**: Adaptive thresholding by Otsu's method (ATO), **B**: Result after separation of cell clusters (ATO + SCC), **C**: Result of the novel gradient-based thresholding (GBT) scheme. Yellow contours: objects identified as cell nuclei. Red contours: object not identified as cell nuclei. The scale bar is 10 μm.

virus signal for each cell by using three different types of neighborhoods: neighborhoods based on simple dilation (SD), restricted dilation (RD), and region growing (RG) in IZ (see Material and Methods). To compare the results, we computed differences in the measured signal for each cell while using a different type of neighborhood. It turned out that we obtain almost the same results with SD and RD (mean difference: 0%, standard deviation: 2.5% of the dynamic intensity range). With RG we measured on average a slightly higher virus signal level than with SD as well as RD (mean difference: 1% of the dynamic intensity range, standard deviation: 3.8%). The reason for this difference is that with RG we segment the cytoplasm and compute the mean intensity only from pixels inside the cytoplasm, whereas with SD as well as RD we compute the mean intensity from pixels near a cell nucleus without segmenting the cytoplasm and therefore background pixels are generally included. We do not consider these differences to be significant and therefore we mostly use the approach based on SD in our application because of its simplicity and significantly lower computation time.

Detection of Out-of-Focus Images

We have also analyzed the performance of our approach for the detection of out-of-focus images. We have determined the discrimination power of our gradient-based out-of-focus measure for 384 images from one plate and have compared it with another algorithm using the normalized variance of pixel values in the image (31). An expert classified the images into three different categories: "in-focus", "out-of-focus", and "hard to decide". The latter class was mostly assigned to images where some cells were in-focus and some were out-of-focus. We computed the discrimination power as the maximal possible percentage of correctly distinguished "in-focus" and "out-of-focus" images by

$$\mathrm{DP}(m) = \max_{T_m} \frac{|\{g|\mathrm{class}(g) = \text{“in-focus”} \wedge m(g) \geq T_m\}| + |\{g|\mathrm{class}(g) = \text{“out-of-focus”} \wedge m(g) < T_m\}|}{|\{g|\mathrm{class}(g) = \text{“in-focus”} \vee \mathrm{class}(g) = \text{“out-of-focus”}\}|},$$

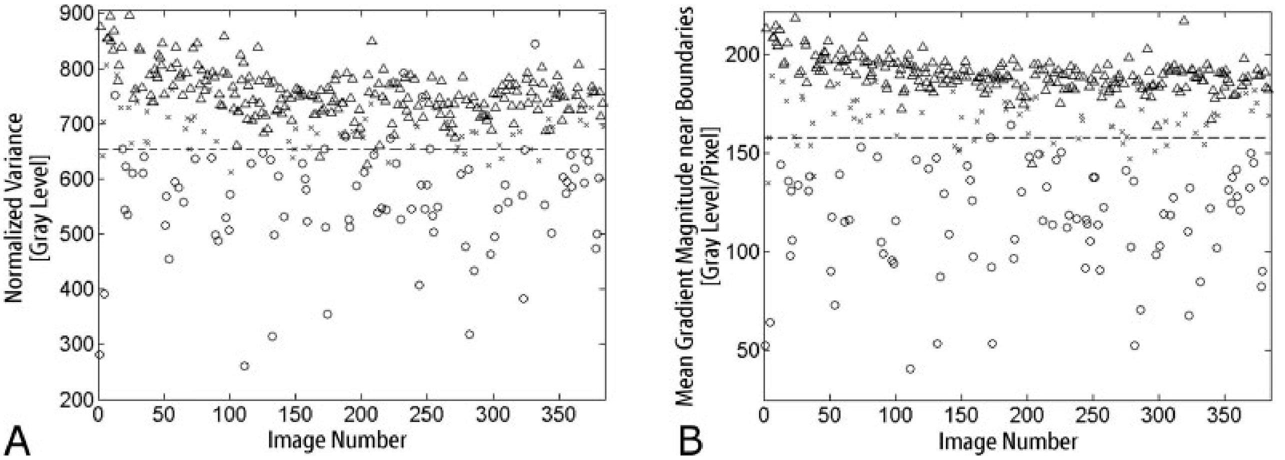

Figure 6. Calculated discrimination power (DP) for 384 images. **A**: Normalized variance of the image intensities (DP = 97.4%), and **B**: Our out-of-focus measure based on the mean gradient magnitude calculated from pixels near the boundaries of segmented objects (DP = 99.4%). To compare the measures, the images were classified by an expert into three different classes: ''in-focus'' (triangles), ''out-of-focus'' (circles), and ''hard to decide'' (crosses). The thresholds used to compute DP are visualized as dashed horizontal lines.

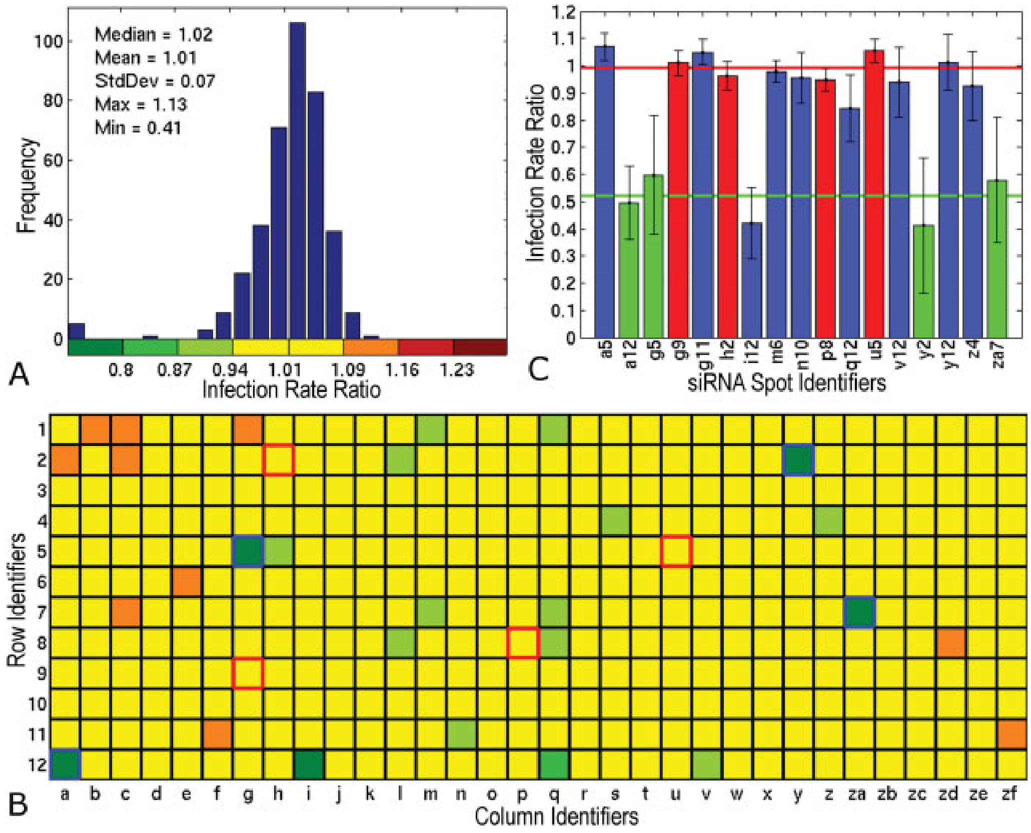

Figure 7. Infection rate ratios (IRRs) computed from a dengue virus (DV) screening experiment. The IRR values were averaged over seven repetitions of the same plate with siRNA spots printed on the 12 × 32 grid. The values are represented by colors (**B**). See the scale below the histogram (**A**). The bar plot (**C**) shows the result for a subset of spots demonstrating a decrease in IRR in positive controls (a12, g5, y2, za7; green bars), normal infection rate (IRR ≈ 1) in negative controls (g9, h2, p8, u5; red bars) together with the results for some other spots (blue bars). The red and green horizontal lines indicate the mean calculated from all positive and negative controls, respectively.

where, m is the out-of-focus measure for which discrimination power (DP) is computed and T_m is a threshold which is chosen such that DP is maximized. Using our approach we achieved DP = 99.4% whereas for the normalized variance we obtained DP = 97.4% (Fig. 6). Note that the latter approach was ranked best in the performance study in (31).

Application to High-Throughput Screens

Our overall approach has been applied to more than 55,000 images of kinase and cytoskeleton screens of cells infected either by HCV or DV using cell arrays with 384 spots. For both types of viruses we obtained a good agreement with decreased infection rate ratios (IRRs) in positive controls as compared to IRRs in negative controls. Besides positive controls, reduced IRRs were also observed in other siRNA spot areas targeting, for example, cellular genes involved in viral infection. This is illustrated for some siRNA spots in Fig. 7C which were selected from a kinase DV screening experiment presented in Fig. 7B. Seven different plates with the same siRNA spot layouts were prepared and imaged on different days. The mean values of these seven repetitions and their standard deviations are shown. Note that we obtained a clear decrease in IRR not only in positive controls but also for other spots (e.g., at position i12). This indicates the applicability of the whole approach.

Discussion

We have described an automatic approach for analyzing image-based high-throughput screens using cell arrays to identify genes involved in virus entry and replication. The overall image analysis approach allows for fully automatic and accurate quantification of a large number of images on single cell basis. Analyzing phenotypes at the level of single cells is critical to determine and study distributions of measured quantities in contrast to using only average values (32). The approach presented here was designed based on requirements of a specific application and the individual algorithms were carefully selected and adapted. Nevertheless, we believe that our approach as well as the presented ideas and experimental comparisons may be applicable in other cytometry high-throughput applications.

In particular, we have described a novel gradient-based thresholding scheme for cell nucleus segmentation which does not require postprocessing steps for cluster separation. This approach does not use the image intensities directly, but is based on thresholding the gradient magnitude while only taking into account pixels where the Laplacian of Gaussian operator yields negative values. In our approach we assume that after the first step of the algorithm (result shown in Fig. 3D) there is exactly one connected component of pixels based on which a nucleus is identified in the subsequent steps. In our experiments we did not observe that a nucleus is separated into two different connected components after the first step. In general, however, this may occur. For example, if the threshold T used to obtain the result in Fig. 3C would be chosen much higher. In this case, the approach could be extended by introducing an additional step comprising morphological reconstruction by dilation (24) of a binary image corresponding to the negative values of the Laplacian of Gaussian (Fig. 3B) from the intermediate result shown in Fig. 3D.

In our validation study, the proposed cell nucleus segmentation algorithm correctly segmented 1,859 nuclei from a total of 1,914 available nuclei, resulting in 97.1% accuracy. In comparison, a smaller number of correctly segmented cells were obtained by adaptive thresholding (78.5%) as well as by adaptive thresholding followed by model-based cluster separation (95.8%). The increase in performance using our approach was observed particularly for clustered cell nuclei which had different average intensities. Also, our algorithm is about two times faster than adaptive thresholding followed by model-based cluster separation. The segmentation accuracy we achieved is also better than the accuracy of 84% recently reported by Gudla et al. (17) using an algorithm that was designed to work in the presence of uneven illumination and clustered nuclei. Note, however, that the stated segmentation accuracy values are difficult to compare because different images were used.

To eliminate low quality images we described an approach for out-of-focus detection which measures the mean gradient magnitude from regions near segmented cell nucleus boundaries. In recent years, several autofocusing approaches have been proposed (31,33,34). In these approaches, a certain measure is typically calculated for several focal planes and the optimal plane is determined by maximizing the measure. The problem of out-of-focus detection addressed in this article is different in that we need to discriminate blurred and sharp images. Nevertheless, autofocus measures can also be used for the task considered here. We experimentally compared our approach with an approach based on the normalized variance of the image intensities, which was ranked best in (31). For our approach we obtained a higher discrimination power, namely of 99.4%, whereas using the normalized variance we obtained 97.4%. The main difference to other measures is that we exploit the result of image segmentation.

Our integrated approach has been applied to more than 55,000 images in different screening experiments with cells infected by either hepatitis C or dengue viruses. It turned out that the obtained results are in good agreement with the expected behavior and encourage the application to image datasets from other high-throughput experiments, in particular, genome-wide screens.

Acknowledgments

The authors thank Nathalie Harder for providing an implementation of adaptive thresholding by Otsu's method, Carolin Wohlfarth for her help with annotating the images, Nathan Brady for checking the English, and the Delft University, Netherlands, for providing the DIPImage toolbox.

Literature Cited

1. Damm E, Pelkmans L. Systems biology of virus entry in mammalian cells. Cell Microbiol 2006;8:1219–1227.
2. Carpenter AE, Sabatini DM. Systematic genome-wide screens of gene function. Nat Rev Genet 2004;5:11–22.
3. Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, Krausz E, Zerial M. Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 2005;436:78–86.
4. Erfle H, Simpson JC, Bastiaens PIH, Pepperkok R. siRNA cell arrays for high-content screening microscopy. Biotechniques 2004;37:454–462.
5. Perrimon N, Mathey-Prevot B. Applications of high-throughput RNA interference screens to problems in cell and developmental biology. Genetics 2007;175:7–16.
6. Echeverri CJ, Perrimon N. High-throughput RNAi screening in cultured cells: A user's guide. Nat Rev Genet 2006;7:373–384.
7. Elter M, Daum V, Wittenberg T. Maximum-intensity-linking for segmentation of fluorescence-stained cells. In: Metaxas D, Whitaker R, Rittscher J, Sebastian T, editors. Proceedings of MICCAI, Workshop MIAAB. Copenhagen: Fraunhofer Publica; 2006. pp 46–50.
8. Lin G, Chawla MK, Olson K, Guzowski JF, Barnes CA, Roysam B. Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei. Cytometry Part A 2005;63A:20–33.
9. Lindblad J, Wählby C, Bengtsson E, Zaltsman A. Image analysis for automatic segmentation of cytoplasm and classification of Rac1 activation. Cytometry Part A 2004;27A:22–33.
10. Wählby C, Lindblad J, Vondrus M, Bengtsson E, Björkesten L. Algorithms for cytoplasm segmentation of fluorescence labelled cells. Anal Cell Pathol 2002;24:101–111.
11. Adiga PSU, Chaudhuri BB. An efficient method based on watershed and rule-based merging for segmentation of 3-D histo-pathological images. Pattern Recognit 2001;34:1449–1458.
12. Li F, Zhou X, Ma J, Wong STC. An automated feedback system with the hybrid model of scoring and classification for solving over-segmentation problems in RNAi high-content screening. J Microsc 2007;226:121–132.
13. Harder N, Mora-Bermúdez F, Godinez WJ, Ellenberg J, Eils R, Rohr K. Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. In: Larsen R, Nielsen M, Sporring J, editors. Proceedings of MICCAI'06, LNCS 4190. Copenhagen: Springer-Verlag; 2006. pp 840–848.
14. Yang X, Li H, Zhou X. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy. IEEE Trans Circuits Syst I 2006;53:2405–2414.
15. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. Genome Biol 2006;7: R100.
16. Raman S, Maxwell CA, Barcellos-Hoff MH, Parvin B. Geometric approach to segmentation and protein localization in cell cultures assays. J Microsc 2007;225:22–30.
17. Gudla PR, Nandy K, Collins J, Meaburn KJ, Misteli T, Lockett SJ. A high-throughput system for segmenting nuclei using multiscale techniques. Cytometry Part A 2008;73A:451–466.
18. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science 2004;306:1194–1198.
19. Xiong GL, Zhou XB, Ji L. Automated segmentation of Drosophila RNAi fluorescence cellular images using deformable models. IEEE Trans Circuits Syst I 2006;53:2415–2424.
20. Jones TR, Carpenter A, Golland P. Voronoi-based segmentation of cells on image manifolds. In: Liu Y, Juany T, Zhang C, editors. Proceedings of Computer Vision for Biomedical Image Applications Conference. LNCS 3765. Berlin, Heidelberg: Springer-Verlag; 2005. pp 535–543.
21. Chang H, Yang Q, Parvin B. Segmentation of heterogenous blob objects through voting and level set formulation. Pattern Recognit Lett 2007;28:1781–1787.
22. Gonzalez RC, Woods RE. Digital Image Processing. Prentice Hall. New Jersey, USA; 2002.
23. DIPImage Toolbox. Available at: http://www.diplib.org/.
24. Soille P. Morphological Image Analysis: Principles and Applications, 2nd ed. Berlin, Germany: Springer; 2004.

25. Zhang JH, Chung TD, Oldenburg KR. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J Biomol Screen 1999; 4:67–73.
26. Nakabayashi H, Taketa K, Miyano K, Yamane T, Sato J. Growth of human hepatoma cell lines with differentiated functions in chemically defined medium. Cancer Res 1982;42:3858–3863.
27. Blight KJ, McKeating JA, Rice CM. Highly permissive cell lines for subgenomic and genomic hepatitis C virus RNA replication. J Virol 2002;76:13001–13014.
28. Schaller T, Appel N, Koutsoudakis G, Kallis S, Lohmann V, Pietschmann T, Bartenschlager R. Analysis of hepatitis C virus superinfection exclusion by using novel fluorochrome gene-tagged viral genomes. J Virol 2007;81:4591–4603.
29. Otsu N. A threshold selection method from gray level histograms. IEEE Trans Syst Man Cybern 1979;9:62–66.
30. CellProfiler. Available at: http://www.cellprofiler.org/.
31. Sun Y, Duthaler S, Nelson BJ. Autofocusing in computer microscopy: Selecting the optimal focus algorithm. Microsc Res Tech 2004;65:139–149.
32. Levsky JM, Singer RH. Gene expression and the myth of the average cell. Trends Cell Biol 2003;13:4–6.
33. Firestone L, Cook K, Culp K, Talsania N, Preston K Jr. Comparison of autofocus methods for automated microscopy. Cytometry 1991;12:195–206.
34. Groen FCA, Young IT, Ligthart G. A comparison of different focus functions for use in autofocus algorithms. Cytometry 1985;6:81–91.