

Evaluation of siRNA Screens of Cells Infected by Hepatitis C and Dengue Viruses based on Immunofluorescence Microscopy

Petr Matula¹, Anil Kumar², Ilka Wörz², Holger Erfle³, Ralf Bartenschlager²,
Roland Eils¹, and Karl Rohr¹

¹University of Heidelberg, BIOQUANT, IPMB, and German Cancer Research Center,
Dept. Bioinformatics and Functional Genomics, Biomedical Computer Vision Group,

Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

² University of Heidelberg, Department of Molecular Virology,

Im Neuenheimer Feld 345, 69120 Heidelberg, Germany

³ University of Heidelberg, BIOQUANT, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

Abstract—We present an image analysis approach as part of a high-throughput microscopy siRNA-based screening system using cell arrays for the identification of cellular genes involved in hepatitis C and dengue virus replication. Our approach comprises: cell nucleus segmentation, quantification of virus replication level in the neighborhood of segmented cell nuclei, localization of regions with transfected cells, cell classification by infection status, and quality assessment of a whole plate and single images. For the latter task, we have developed a graphical user interface (GUI) to ease quality control of a large number of images. We also describe an approach for the classification of cells as infected or non-infected which works well also in case of low infection rates. The whole approach is fully automatic and has been successfully applied to a large number of cell array images from screening experiments. The experimental results show a good agreement with the expected behavior of positive as well as negative controls and encourage the application to screens from further high-throughput experiments.

I. INTRODUCTION

For understanding viral infection it is important to identify cellular genes involved in virus entry and replication. Viruses need to enter cells and exploit their cellular machinery to replicate. Prospective discoveries of host cell components that are crucial for the replication of a certain class of viruses are expected to lead to significant improvements in antiviral treatments [1]. It is believed that by knocking down cellular genes indispensable for virus entry or replication, a cell can still perform its function due to redundant pathways whereas it would be hard to accommodate for a simple organism such as a virus.

Our ultimate goal is to develop a high-throughput screening system for genome-wide identification of cellular genes potentially involved in virus entry and replication. The general idea is to systematically knock down host cell genes, to quantify changes in the level of viral protein expression, and to detect genes whose knockdown leads to significant changes in virus replication. Such a systematic approach became possible by using RNA interference (RNAi), which allows to knock down

the expression of single genes [2]. RNAi screening has been applied to answer a wide range of biological questions (see Ref. [3] for a review). Genome-wide screens with more than 20,000 genes generate more than 200,000 fluorescence images and therefore fully automatic image analysis methods are needed.

In recent years, a number of approaches for cell nuclei or whole cell segmentation based on fluorescence microscopy have been reported [4], [5], [6], [7], [8], [9]. In *high-throughput* applications, approaches based on adaptive thresholding [10], [11] proved to give good results for cell nuclei segmentation, especially if the nuclei are not clustered. To separate clusters of nuclei, watershed based techniques [10], [12] or approaches employing geometric properties [13], [14] have been proposed. Recently, an approach using multi-scale entropy-based thresholding and region merging has been described [15]. For the segmentation of whole cells or the cytoplasm approaches based on deformable models [16], Voronoi diagrams [17], or a combination of both [18] have been introduced. Recently, a multi-scale approach based on region growing has been described [19].

In our application we rely on high-throughput fluorescence microscopy imaging of small interfering RNA (siRNA) cell arrays [20]. In cell arrays, cells are cultured on chamber plates with printed (ready to transfect) siRNA spots organized in a grid pattern. Only those cells that are located within a printed spot area can take up siRNA. This screening platform enables to increase throughput and to significantly reduce costs compared to multi-well plates. Key tasks for image analysis in this application are cell nucleus segmentation, detection of regions with transfected cells (siRNA spots), and quantification of the virus infection level. The approaches should be efficient, robust, and fully automatic.

In this contribution, we describe an image analysis approach as part of a high-throughput system for genome-wide identification of cellular genes that are important for hepatitis C (HCV) and dengue virus (DV) replication. Currently, to our best knowledge, there exists no image analysis system for quantifying viral signals given a large number of images

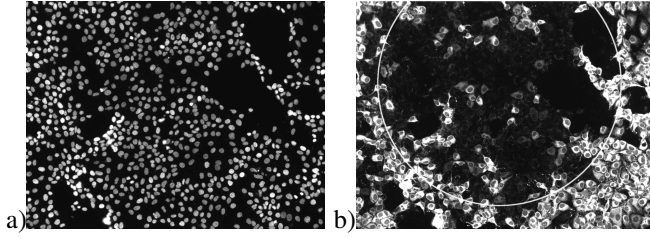


Fig. 1. Image data. (a) Image of cell nuclei corresponding to one spot area (channel 1), (b) Image of viral protein expression corresponding to the same spot (channel 2) with marked siRNA spot area

on per cell basis for cell arrays, which measures the status of virus replication in transfected cells only, and combines the results from many different and repeated experiments to produce reliable statistics. We propose approaches (1) for the segmentation and quantification of cells in two-channel images, (2) for the localization of regions with transfected cells within cell array images, and (3) for classification of cells as infected or non-infected. We have also developed a graphical user interface (GUI) to ease quality control of a large number of images.

II. IMAGE DATA AND OVERVIEW OF THE APPROACH

The cell array images in our application are acquired by high-throughput fluorescence microscopy. A cell array is a microscopic chamber plate overlaid with a grid of siRNA spots of $400\ \mu\text{m}$ diameter that are printed on the plate by an automatic robot. In this study, grids of 6×6 or 12×32 siRNA spots were used for hepatitis C virus or dengue virus. A biological experiment is performed using spotted plates. The cells are seeded on spotted plates and incubated for 24–48 hours. Cells within the spot areas take up siRNAs and the expression of cellular genes corresponding to the applied siRNAs is knocked down. Subsequently, the cells are infected with a virus and after 24–48 hours they are fixed and fluorescently stained. Cell nuclei (cellular DNA) are labeled by DAPI or Hoechst staining (channel 1, Fig. 1a) and a viral protein is labeled by immunofluorescence (channel 2, Fig. 1b). One grayscale image is acquired for each siRNA spot area of each channel using the scanning platform Scan^R (Olympus). In Fig. 1b an siRNA spot area has been marked in an acquired image. The typical image size is 1344×1024 pixels. As an output of high-throughput scanning we obtain a set of 36 or 384 two-channel images for each plate.

The image analysis workflow consists of the following main steps: 1. Segmentation of cell nuclei in the nucleus channel (channel 1), 2. Quantification of the viral protein production level in the neighborhood of each nucleus (channel 2), 3. Localization of siRNA spot areas with transfected cells, 4. Cell classification by infection status, 5. Quality assessment of single images and of a plate. In the first two steps of the workflow single images are processed. In the remaining three steps the information from the whole set of images is exploited.

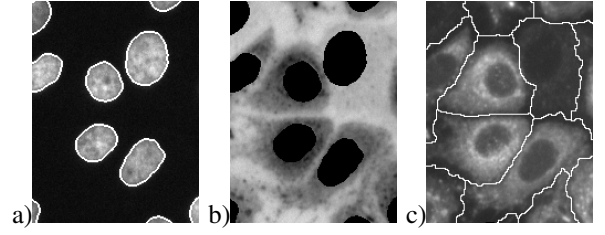


Fig. 2. Segmentation results for an image section. (a) Cell nuclei (channel 1) with overlaid contours, (b) inverted channel 2 with segmented nuclei as seeds, (c) viral protein expression (channel 2) with overlaid influence zones (IZ) of each nucleus

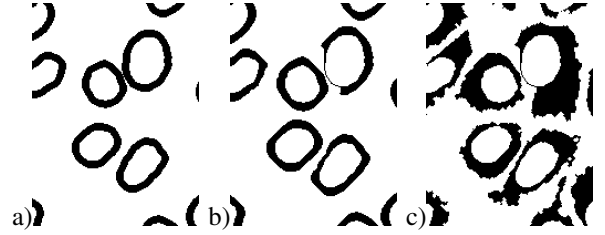


Fig. 3. Neighborhoods of cell nuclei for the quantification of the level of expressed viral protein: (d) Simple dilation, (e) Constrained dilation by IZ, (f) Region-growing inside IZ.

III. SEGMENTATION AND QUANTIFICATION

There are two main approaches for the quantification of cells infected by viruses. Either (1) the percentage of virus infected cells (called infection rate) is determined for each siRNA spot and changes in infection rates are studied or (2) the level of expressed viral protein (called viral signal) is measured for each cell and distributions of viral signal within different siRNA spots are compared (e.g., by comparing their means). In both cases, the required tasks are cell nucleus segmentation and determination of the neighborhood of each nucleus, where the virus signal is measured. In the first case, an additional task is needed to classify each cell as infected or non-infected based on the measured virus signal.

To segment cell nuclei we use a gradient-based thresholding approach and mathematical morphology operations. We determine cell nucleus boundary regions based on the gradient magnitude and the response of the Laplacian of Gaussian. Then, we apply connected component labeling and remove small and large objects. Next, the remaining objects are closed while preventing merging objects with different identifier. Afterwards, holes in objects are filled. Finally, cell nuclei are identified among the segmented objects by applying size, intensity, and circularity criteria.

To quantify the level of expressed viral protein for each cell we compute the mean intensity in channel 2 inside its nucleus neighborhood. We have implemented and compared three different approaches for defining the neighborhood of a cell nucleus: simple dilation (Fig. 3a), restricted dilation by influence zones (IZ) (Fig. 3b), and region growing in IZ (Fig. 3c). The first approach is the simplest one and the neighborhood is computed by dilating the cell nucleus masks and subtracting the original masks. The second and third approaches rely on a partition of an image into IZ of each nucleus (Fig. 2c). Influence zones were computed using

a seeded watershed transform of the Gaussian filtered and inverted virus channel (channel 2) with the segmented cell nuclei as initial seeds (Fig. 2b). The region growing algorithm was applied only inside corresponding influence zones and was started from the pixels on the cell nucleus boundary. Only pixels within an IZ with intensity values similar to the average intensity of pixels at the boundary of a cell nucleus were included during the growing process.

In order to compute infection rates, the cells must first be classified as infected or non-infected. A cell is classified as infected if the measured virus signal is higher than a certain threshold and as non-infected if the measured virus signal is lower than this threshold. The threshold is set so that the difference in infection rates in positive and negative controls is maximized. In positive controls (PC) the viral protein production is blocked and therefore the signal is reduced. In negative controls (NC) the virus replication is not altered (see Figs. 4a,c). An advantage of the approach based on the maximization of the difference between PC and NC is that it works well also for viruses with lower infection rates as presented in Figs. 4c,d.

IV. LOCALIZATION OF siRNA SPOTS

Only cells that are located within printed siRNA spot areas can generally be transfected and only those should be quantified. Gene expression of these cells was knocked down and changes in viral protein expression can potentially be observed. The cells outside printed siRNA spots cannot take up siRNA (rare migration of cells is neglected) and therefore should exhibit a normal infection rate. An example of an image with clear differences in viral protein expression is shown in Fig. 1c. The goal of localizing siRNA spot areas with transfected cells is to find cells that are located within printed siRNA spots, regardless of the exhibited changes in the viral protein expression.

The idea of our approach is to detect siRNA spot areas in images with altered viral protein expression and to extrapolate siRNA spot areas to other images by using information about the grid printed on the plate. Our approach comprises the following two steps.

- 1) For each image I of an experiment, a position $[x_c^I, y_c^I]$ of a circle of known fixed diameter is found, for which the difference d^I in mean protein production level in cells inside and outside the circle is maximal.
- 2) All images with differences d^I in a certain range are selected and a grid of known parameters is fitted to the computed circle positions $[x_c^I, y_c^I]$ using a least-squares approach.

The used range of differences d^I in step 2 helps to select only those images from step 1, for which significant differences in viral protein expression inside and outside the circle were obtained, i.e. where the siRNA spot was detectable.

The appropriate range of differences was determined based on simulations. We randomly generated N points in a plane to simulate nucleus centers. Then, we randomly generated a position of a circle of a given diameter and all points inside this circle were marked as "inside spot" and all points outside

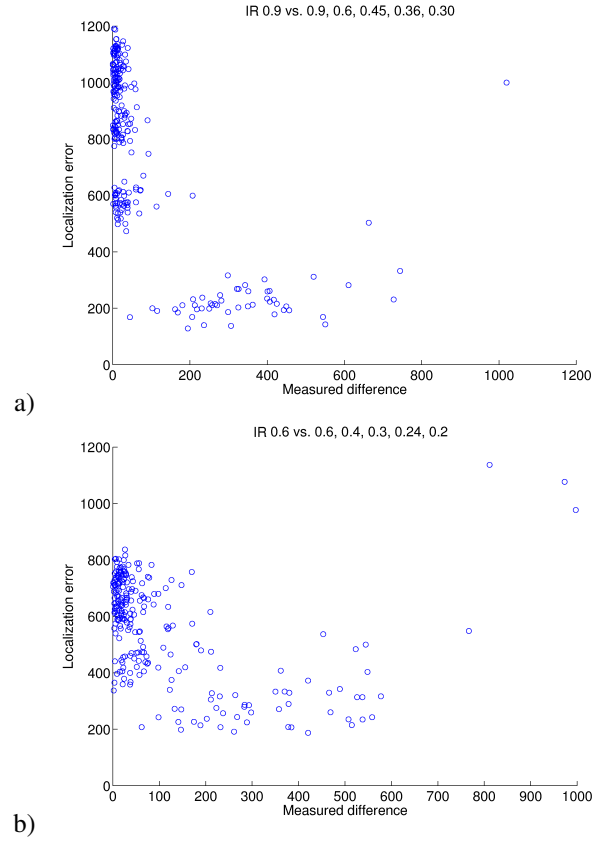


Fig. 5. Correlation between the output of the first step of the spot localization algorithm (measured difference d^I , see text) and the correctness of the computed spot position (represented by the localization error) based on simulations. (a) Result for high reference infection rate (0.9), (b) Result for low reference infection rate (0.6).

the circle were marked as "outside spot". Then, we assigned a value to each point representing the signal level of viral expression either based on the Gaussian distribution G_I for infected cells or the Gaussian distribution G_N for non-infected cells. The probability of selecting the distribution G_I (called infection rate) was different for cells marked as "inside spot" and "outside spot". The infection rate (IR) for cells marked as "outside spot" was fixed in each simulation (so called reference IR). The IR for cells marked as "inside spot" was varied using five different scaling factors of the reference IR, namely 1, 2/3, 1/2, 2/5, and 1/3. The generated data was used as input for our spot localization algorithm and the Euclidean distance between the found circle position and the generated circle position was calculated yielding the localization error. Simulation results for two different reference infection rates, namely 0.9 (typical for dengue virus) and 0.6 (typical for hepatitis C virus) are presented in Fig. 5a and Fig. 5b, respectively. The figures show the calculated localization error as a function of the measured difference d^I . The lowest errors were obtained for differences in the range of $d^I = 300 \dots 500$. Too large or too small differences are correlated with erroneous estimations of circle positions. A similar behavior was observed for real data.

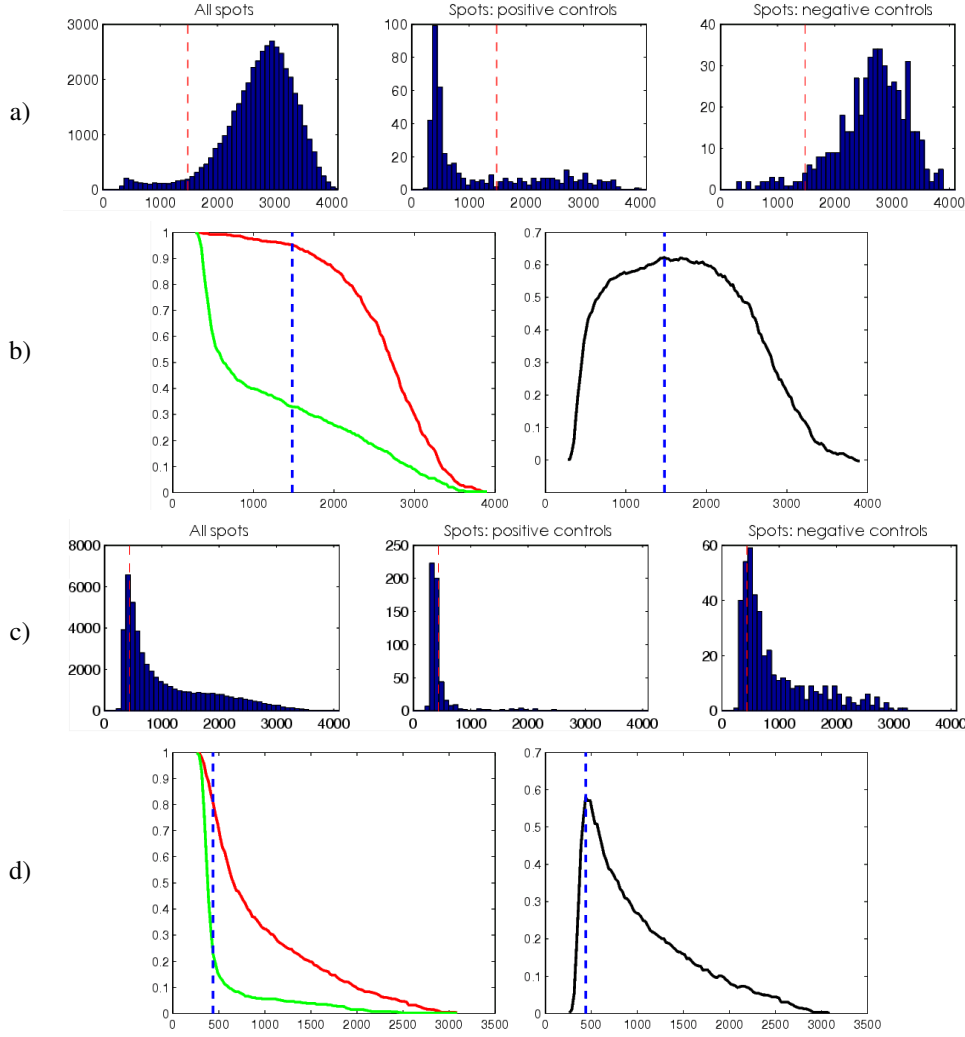


Fig. 4. Selection of infection rate threshold. **(a), (c)** Distribution of the measured viral signal level within all spots (left), all spots corresponding to positive controls (PC) (middle), and all spots corresponding to negative controls (NC) (right); **(b), (d)** Infection rate for PC (light gray or green in color version) and for NC (dark gray or red in color version) as a function of the threshold (left) and difference in infection rates between NC and PC as a function of the threshold (right). The maximal difference defines the optimal threshold value (dashed line). The results for two different viruses with different infection rates are shown: **(a), (b)** High infection rate: dengue virus; **(c), (d)** Low infection rate: hepatitis C virus.

V. QUALITY ASSESSMENT

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. We perform quality checks on two levels: on the whole plate level and on single images level.

On the single image level, images are automatically classified as low quality if (1) they contain a too large or too small number of cells or if (2) they are out-of-focus.

Most out-of-focus images are excluded already by applying criterion (1) because in these images no or a small number of cell nuclei are usually segmented. In addition, we exclude images for which the average gradient magnitude calculated from pixels near the boundaries of segmented objects is lower than a threshold. The threshold is computed as the mean minus three standard deviations of the gradient magnitudes calculated from all images of a plate. We use robust estimators

for the mean and standard deviations to reduce the effect of outliers. The main difference of this approach as compared to autofocusing approaches [21], [22] is that we exploit the segmentation results.

On the whole plate level, we compute the percentage of overexposed pixels in a mask given by the union of the neighborhoods of all cell nuclei. A high percentage of overexposed pixels indicates a too long exposure time and such plates need to be reacquired.

To visualize whole plate related problems, e.g., due to improper staining or cell seeding, we have developed a graphical user interface (GUI) which displays all images of a plate in one overview tiled image (Fig. 6). All images classified as low quality are tagged with red numbers. A user can also view the original data, the segmentation results, and can alter the assigned flags. In case of serious problems the whole plate is excluded. In case of less serious problems only good quality images are included in the statistical evaluation. The implemented GUI plays an important role especially during

optimization of the sample preparation and validation of the automatic quality control of single images as well as whole plates.

VI. EXPERIMENTAL RESULTS

Prior to applying the overall approach, we tested and evaluated single algorithms based on real data.

The cell nucleus segmentation algorithm was evaluated using real data, where ground truth was obtained from two experts who marked cell nuclei in 6 randomly selected real images from 3 different experiments (1914 nuclei in total). We compared the results of our approach with that of adaptive thresholding by Otsu's method [23]. With our approach we were able to segment more than 97% of all nuclei whereas Otsu's method yielded about 79% only. The proposed approach was particularly superior in segmenting clustered cells. Whereas Otsu's method often segmented these cell clusters as one object, which was subsequently rejected by the classifier, our approach correctly separated clustered cells more frequently. The obtained percentage of correctly segmented cells by our gradient-based approach is considered to be high enough for our application.

We also studied the performance of the three approaches for defining the cell nucleus neighborhood. For each approach, we computed the mean intensity in channel 2 for more than 100,000 cells. It turned out that the results for all three approaches were similar. However, for the region growing approach we generally measured slightly higher values than for the other two approaches (simple and restricted dilation). The reason for this is that with the latter approaches we include, in general, also pixels outside of the cytoplasm. In our application, we use the approach based on simple dilation because of its simplicity and significantly lower computation time.

Our overall approach has been applied to more than 55,000 images of HCV and DV screens. A grid of 12×32 was used in most screens. The overall results of 10 repeated experiments of an HCV screen (with a smaller grid of 6×6) are presented in Fig. 7. It turned out that we obtain a good agreement with reduced infection rate ratios in positive controls as compared to the infection rate ratios in negative controls. Besides positive controls, reduced infection rate ratios were also observed in other siRNA spot areas targeting, e.g., cellular genes involved in viral infection. A similar behavior was observed also in other screens. This indicates the applicability of the whole approach.

VII. CONCLUSION

We have described an image analysis approach for the identification of genes involved in HCV and DV replication based on high-throughput screening experiments. The whole approach relies on (1) a gradient-based approach for segmentation of cell nuclei in fluorescence microscopy images, (2) combination of model-based circular region fitting and grid fitting for the localization of siRNA spot regions, which allows to exclude non-transfected cells, and (3) classification of cells as infected and non-infected, which works well also

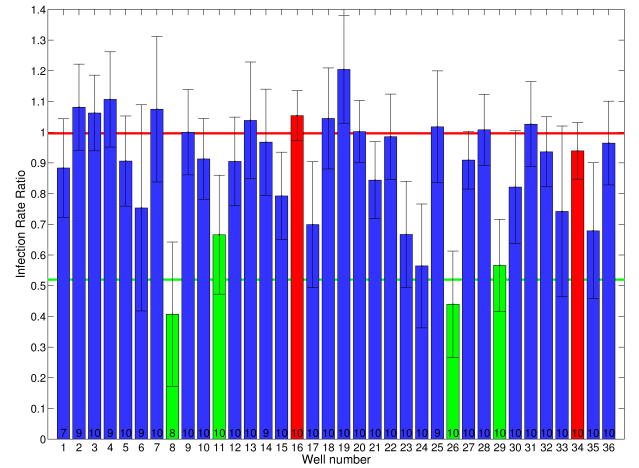


Fig. 7. Infection rate ratios computed from screening experiments. Results averaged (a) from 10 repeated experiments of an HCV screen (positive controls: columns 8, 11, 26, 29; negative controls: columns 16, 34). The red and green horizontal lines indicate the mean calculated from all positive and negative controls, respectively.

for low infection rates. We have developed a graphical user interface (GUI) to ease quality control of a large number of images. The overall approach enables to fully automatically quantify a large number of images on single cell basis. The obtained results are in good agreement with the expected behavior and encourage the application to images from other high-throughput experiments, in particular, from genome-wide screens.

ACKNOWLEDGEMENTS

This work has been funded by the BMBF (FORSYS) project VIROQUANT. We thank Nathalie Harder for providing an implementation of adaptive thresholding by Otsu's method and Carolin Wohlfarth for her help with annotating the images, and the Delft University, Netherlands, for providing the DIPimage toolbox.

REFERENCES

- [1] E.-M. Dam and L. Pelkmans, "Systems biology of virus entry in mammalian cells," *Cellular Microbiology*, vol. 8, no. 8, pp. 1219–1227, 2006.
- [2] A. E. Carpenter and D. M. Sabatini, "Systematic genome-wide screens of gene function," *Nature Reviews Genetics*, vol. 5, no. 1, pp. 11–22, 2004.
- [3] N. Perrimon and B. Mathey-Prevot, "Applications of high-throughput RNA interference screens to problems in cell and developmental biology," *Genetics*, vol. 175, pp. 7–16, 2007.
- [4] M. Elter, V. Daum, and T. Wittenberg, "Maximum-intensity-linking for segmentation of fluorescence-stained cells," in *Proc. Workshop Microscopic Image Analysis with Applications in Biology (MIAAB'06)*, D.N. Metaxas, R.T. Whitaker, J. Rittscher, and T.B. Sebastian, Eds., Copenhagen, 2006, pp. 46–50.
- [5] G. Lin, M. K. Chawla, K. Olson, J. F. Guzowski, C. A. Barnes, and B. Roysam, "Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei," *Cytometry*, vol. 63A, pp. 20–33, 2005.
- [6] J. Lindblad, C. Wahlby, E. Bengtsson, and A. Zaltsman, "Image analysis for automatic segmentation of cytoplasm and classification of Rac1 activation," *Cytometry*, vol. 27A, pp. 22–33, 2004.

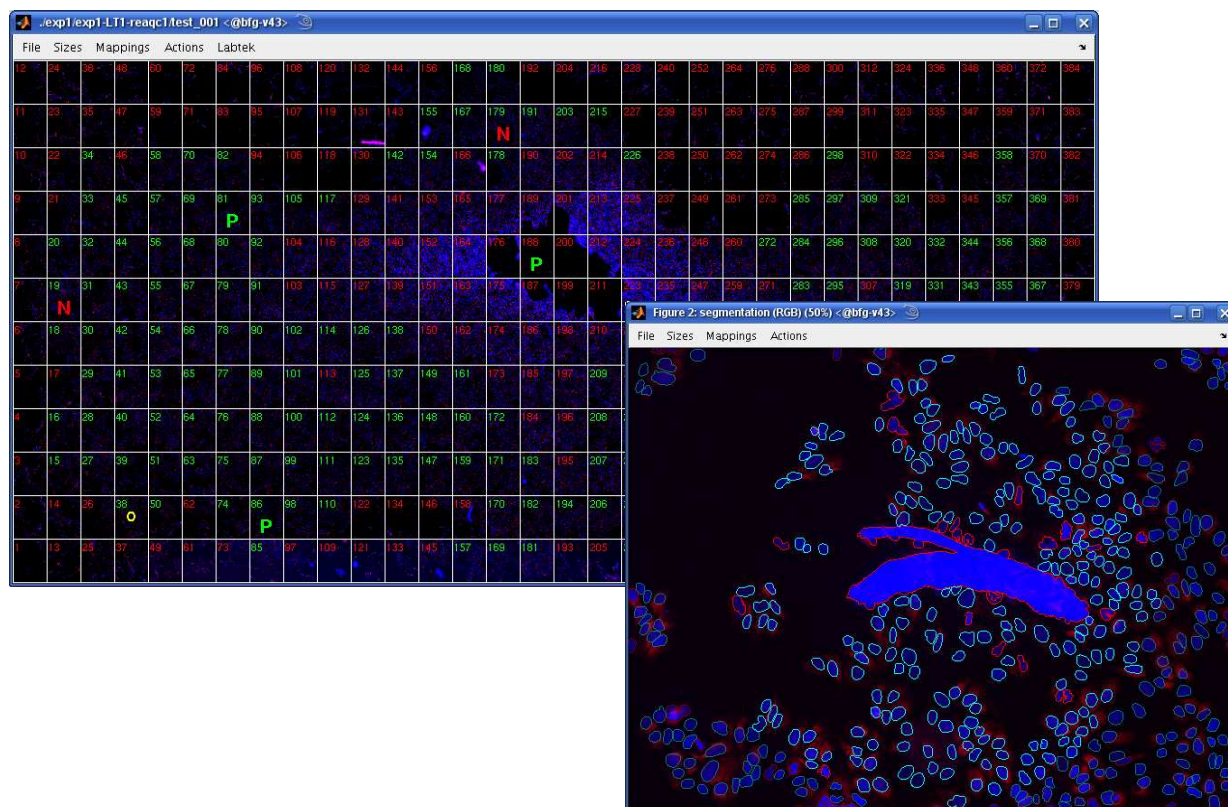


Fig. 6. Graphical user interface (GUI) for quality control. Images of all 384 spots of one plate are tiled into one overview image to visualize whole slide problems. An example of a bad quality plate during the cell seeding optimization process for an experiment with problems in cell seeding is shown in the window in the background. Single images classified as low quality are tagged by red numbers. Good quality images are tagged by green numbers. A user can view each single image with overlaid segmentation and classification results, and can change the quality tags. Positions of positive controls are marked by the letter "P" and positions of negative controls by the letter "N". The segmentation results for one siRNA spot are shown in the window in the foreground. Light cyan contours: objects identified as cell nuclei within siRNA spot area. Red contours: excluded objects.

- [7] G. Lin, U. Adiga, K. Olson, J. F. Guzowski, C. A. Barnes, and B. Roysam, "A hybrid 3D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks," *Cytometry*, vol. 56A, pp. 23–36, 2003.
- [8] C. Wählby, J. Lindblad, M. Vondrus, E. Bengtsson, and L. Björkstén, "Algorithms for cytoplasm segmentation of fluorescence labelled cells," *Analytical Cellular Pathology*, vol. 24, pp. 101–111, 2002.
- [9] P. S. Umesh Adiga and B. B. Chaudhuri, "An efficient method based on watershed and rule-based merging for segmentation of 3-D histopathological images," *Pattern Recognition*, vol. 34, no. 7, pp. 1449–1458, 2001.
- [10] F. Li, X. Zhou, J. Ma, and Stephen T. C. Wong, "An automated feedback system with the hybrid model of scoring and classification for solving over-segmentation problems in RNAi high-content screening," *Journal of Microscopy*, vol. 226, no. 2, pp. 121–132, 2007.
- [11] N. Harder, F. Mora-Bermúdez, W. J. Godinez, J. Ellenberg, R. Eils, and K. Rohr, "Automated analysis of the mitotic phases of human cells in 3d fluorescence microscopy image sequences," in *Proc. International Conference Medical Image Computing and Computer-Assisted Intervention (MICCAI'06)*, R. Larsen, M. Nielsen, and J. Sporring, Eds., Berlin, 2006, vol. 4190 of *LNCS*, pp. 840–848, Springer-Verlag.
- [12] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy," *IEEE Transactions on Circuits and Systems I*, vol. 53, no. 11, pp. 2405–2414, 2006.
- [13] S. Raman, C. A. Maxwell, M. H. Barcellos-Hoff, and B. Parvin, "Geometric approach to segmentation and protein localization in cell culture assays," *Journal of Microscopy*, vol. 225, no. 1, pp. 22–30, 2007.
- [14] X. Zhou, K.-Y. Liu, P. Bradley, N. Perrimon, and S. T. C. Wong, "Towards automated cellular image segmentation for RNAi genome-wide screening," in *Proc. International Conference Medical Image Computing and Computer-Assisted Intervention (MICCAI'05)*, Berlin, 2005, vol. 3749 of *LNCS*, pp. 885–892, Springer-Verlag.
- [15] Prabhakar R. Gudla, K. Nandy, J. Collins, K. J. Meaburn, T. Misteli, and S. J. Lockett, "A high-throughput system for segmenting nuclei using multiscale techniques," *Cytometry*, vol. 73A, pp. 451–466, 2008.
- [16] G. L. Xiong, X. B. Zhou, and L. Ji, "Automated segmentation of Drosophila RNAi fluorescence cellular images using deformable models," *IEEE Transactions on Circuits and Systems I*, vol. 53, pp. 2415–2424, 2006.
- [17] T. R. Jones, A. Carpenter, and P. Golland, "Voronoi-based segmentation of cells on image manifolds," in *Proc. Conference Computer Vision for Biomedical Image Applications (CVBIA'05)*, Y. Liu, T. Jiang, and C. Zhang, Eds., Berlin, 2005, vol. 3765 of *LNCS*, pp. 535–543, Springer-Verlag.
- [18] H. Chang, Q. Yang, and B. Parvin, "Segmentation of heterogeneous blob objects through voting and level set formulation," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1781–1787, 2007.
- [19] A. Verma and A. Kriete, "Robust object-oriented segmentation for high-content screening," in *Proc. Workshop Microscopic Image Analysis with Applications in Biology (MIAAB'07)*, D.N. Metaxas, J. Rittscher, S. Lockett, and T.B. Sebastian, Eds., New York, 2007.
- [20] H. Erfle, J. C. Simpson, P. I. H. Bastiaens, and R. Pepperkok, "siRNA cell arrays for high-content screening microscopy," *BioTechniques*, vol. 37, no. 3, pp. 454–462, 2004.
- [21] F. C. A. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms," *Cytometry*, vol. 6, pp. 81–91, 1985.
- [22] L. Firestone, K. Cook, K. Culp, N. Talsania, and K. Preston Jr, "Comparison of autofocus methods for automated microscopy," *Cytometry*, vol. 12, pp. 195–206, 1991.
- [23] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, 1979.