Cellular Phenotype Recognition for High-Content RNA Interference Genome-Wide Screening

JUN WANG,[1,4] XIAOBO ZHOU,[1,2] PAMELA L. BRADLEY,[3,5] SHIH-FU CHANG,[4] NORBERT PERRIMON,[3] and STEPHEN T. C. WONG[1,2]

Genome-wide, cell-based screens using high-content screening (HCS) techniques and automated fluorescence microscopy generate thousands of high-content images that contain an enormous wealth of cell biological information. Such screens are key to the analysis of basic cell biological principles, such as control of cell cycle and cell morphology. However, these screens will ultimately only shed light on human disease mechanisms and potential cures if the analysis can keep up with the generation of data. A fundamental step toward automated analysis of high-content screening is to construct a robust platform for automatic cellular phenotype identification. The authors present a framework, consisting of microscopic image segmentation and analysis components, for automatic recognition of cellular phenotypes in the context of the Rho family of small GTPases. To implicate genes involved in Rac signaling, RNA interference (RNAi) was used to perturb gene functions, and the corresponding cellular phenotypes were analyzed for changes. The data used in the experiments are high-content, 3-channel, fluorescence microscopy images of *Drosophila* Kc167 cultured cells stained with markers that allow visualization of DNA, polymerized actin filaments, and the constitutively activated Rho protein Rac^{V12}. The performance of this approach was tested using a cellular database that contained more than 1000 samples of 3 predefined cellular phenotypes, and the generalization error was estimated using a cross-validation technique. Moreover, the authors applied this approach to analyze the whole high-content fluorescence images of *Drosophila* cells for further HCS-based gene function analysis. (*Journal of Biomolecular Screening* 2008:29-39)

Key words: high-content screening, RNA interference, microscopic image segmentation, phenotype feature extraction and selection, phenotype classification

INTRODUCTION

FLUORESCENCE MICROSCOPY IMAGES OF CELLS stained to reveal complex cellular structures, such as cytoarchitecture, are considered to be "high-content" images due to the large amount of information they contain. These images reveal numerous biological readouts, including cell size, cell viability, DNA content, cell cycle, and cell morphology. With such a wealth of biological information in high-content images, scientists now perform high-content screens (HCS) in disease diagnosis and prognosis, as well as drug target validation.[1,2] The application of high-throughput screening (HTS) procedures coupled with automated microscopy to generate large quantities of high-content image data can be used to identify genes or small molecules involved in a particular complex biological process.

A gene's function can be assessed by analyzing alterations in a biological process caused by absence or disruption of that gene. In *Drosophila* cells, specific and reproducible loss-of-function phenotypes can be generated by the addition of gene-specific double-stranded RNA (dsRNA), which causes reduction or elimination of target gene expression by a process known as RNA interference (RNAi).[3] Automated fluorescence microscopy then allows unattended acquisition of unbiased cellular images in large quantity. A rate-limiting factor in realizing the full potential of cellular and molecular imaging studies is the ability to automate the analysis of the large number of images generated in such screens because currently, scientists have to resort to the slow manual analysis of complex phenotypes. Thus, a major advance in the field of HCS would be to automate the extraction of patterns and convert them into cellular signatures that can be used to answer in-depth biological questions.

[1]Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair, Harvard Medical School, Boston, Massachusetts.
[2]Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts.
[3]Department of Genetics and Howard Hughes Medical Institute Harvard Medical School, Boston, Massachusetts.
[4]Department of Electrical Engineering, Columbia University, New York.
[5]Present address: National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland.

Received Mar 25, 2007, and in revised form Sep 5, 2007. Accepted for publication Sep 24, 2007.

Journal of Biomolecular Screening 13(1); 2008
DOI: 10.1177/1087057107311223

Automating HCS image analysis could allow functional analysis of complex cellular processes that require information about individual cells.[4] In a small-scale screen using manual analysis of fluorescence microscopy images, we observed a wide range of phenotypes with affected cytoskeletal organization and cell shape.[5,6] Genome-wide screens, however, produce huge volumes of image data not amenable to manual analysis. Thus, without the aid of proper automated image analysis techniques, it becomes intractable to characterize morphological phenotypes quantitatively and to identify genes and their dynamic relationships on a genome-wide scale. Therefore, it is critical and urgent to develop automated, reliable, and fast methods to identify and quantify cellular phenotypes as the basis for a computerized image scoring system.

A robust high-content, image-based genome analysis system using image processing and pattern analysis techniques requires an automated informatics framework that includes the following components: cellular segmentation, cellular morphology and texture feature extraction, cellular phenotype classification, and clustering analysis (**Fig. 1**). In this article, we investigate image-based cellular phenotype recognition for large-scale, high-content RNAi screening. First, individual cells are identified by segmentation using the deformable model[7] and the watershed method using CellProfiler™.[8] Then, we scrutinized the geometric properties and appearance of 3 distinct cellular phenotypes by extracting 5 types of features—namely, wavelet features, moment features, Haralick co-occurrence features, region property features, and problem-specific shape descriptors. To improve the classification performance and reduce the computational cost, we used a genetic algorithm to select a subset of the most discriminative features. Finally, 4 different classification models were trained to predict the phenotypes of test cells. We validated the classifier with HCS image data of Rho GTPase activity in *Drosophila* cells. In the phenotype identification experiments, the classifier correctly recognized 87.94% *normal* cells, 67.41% *spiky* cells, and 70.10% *ruffling* cells (detailed description of the predefined phenotypes in the "High-content images of Rho GTPase activity" subsection) and achieved an overall accuracy of 76% on all 3 phenotypes. Considering the complexity and diversity of cellular phenotypes in the fluorescence images, this is acceptable for genome-wide analysis. Moreover, during the study on gene expression function analysis using RNAi HCS, the gene effects are evaluated based on the statistical properties on a population of cells in the whole screen, which is much more robust than phenotype representation of individual cells.

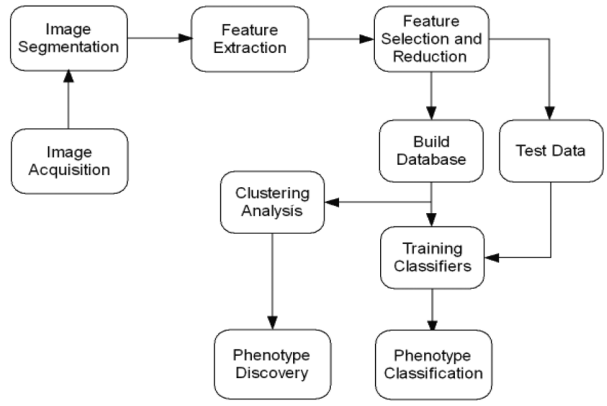


FIG. 1. The information-processing pipeline of the automatic cellular phenotype analysis system for high-content screening (HCS) mainly consists of image segmentation, cellular feature extraction, feature selection and reduction, and phenotype classification.

MATERIALS AND METHODS

High-content images of Rho GTPase activity

Rho proteins are required for the cytoskeletal reorganization critical to cell shape change and migration, as well as many other cellular processes in multiple cell types.[9] Extensive work in mammalian cell culture systems using dominant negative and constitutively active mutant forms of the Rho protein Rac demonstrated the contribution of Rac to specific cytoskeletal structures, such as lamella.[9] Moreover, Rac is required for the invasive behavior of breast cancer cells and is thus an important player in cancer metastasis.[10] To identify novel downstream effectors that mediate Rac cellular responses, we have developed a *Drosophila* cell-based assay for GTPase function that can be used in combination with high-throughput RNAi technology to screen for dsRNAs that block GTPase activity. Specifically, we use the *Drosophila* Kc167 embryonic cell line, which consists of small (10 μm) and uniformly round cells. Visualizing the actin cytoskeleton with phalloidin staining reveals little filamentous actin (F-actin) cytoskeletal structure in Kc167 cells: only small puncta and nonuniform cortical expression are seen.

To facilitate HCS, we generated a construct containing sequences encoding a GFP-Rac^{V12} fusion protein under the transcriptional control of a copper sulfate ($CuSO_4$)–inducible promoter on the same plasmid with a hygromycin resistance gene. We used dsRNA specific to predicted *Drosophila* genes to elicit the RNAi response, which mimics loss-of-function mutations in the targeted gene.[11] To perform the screen, dsRNAs were robotically arrayed individually in 384-well plates. *Drosophila* cells were plated in each well, where they were taken up the dsRNA from the culture media. After 3 days of incubation at 25 °C, Rac^{V12} expression was induced with 0.75 mM $CuSO_4$ for 24 h. The cells were then fixed in 4% formaldehyde and stained to visualize DNA and actin using 4′,6-diamidino-2-phenylindole (DAPI) and fluorescently labeled phalloidin, respectively.

Two or 3 images per well in each of 3 channels (ultraviolet [UV], fluorescein isothiocyanate [FITC], and tetramethyl rhodamine isothiocyanate [TRITC]) were acquired by automated microscopy with a Universal Imaging AutoScope, a Nikon TE300

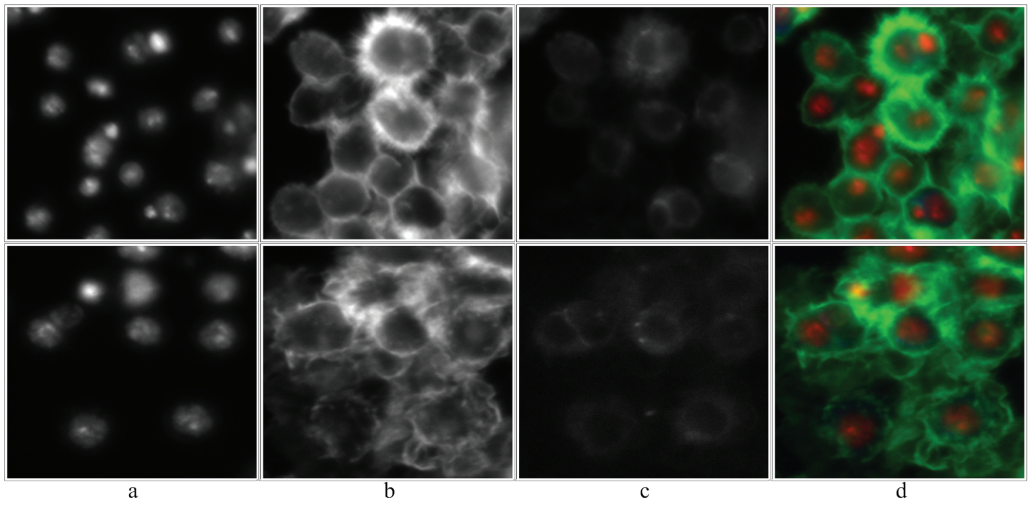


FIG. 2. An example of the fluorescence microscopy image. Images are taken in 3 channels of *Drosophila* Kc167 cells stained with 4′,6-diamidino-2-phenylindole (DAPI) to visualize DNA (**a**) and tetramethyl rhodamine isothiocyanate (TRITC)–labeled phalloidin to visualize F-actin (**b**); these cells also express a GFP-Rac^{V12} fusion protein (**c**). The 3-channel merged color image is shown in **d**.

inverted fluorescence microscope, using a 40× air objective. A single focal plane of the highest intensity was chosen automatically by the focusing software within the Metamorph program. An example of the 3-channel RNAi fluorescence images used for cytological profiling is shown in **Figure 2**. The DNA channel shows the nuclei of cells (**Fig. 2a**). The actin channel reveals the cytoskeletal structure, used to determine the morphology of cell bodies (**Fig. 2b**). Because relatively little visual information is available from the GFP-Rac^{V12} channel (**Fig. 2c**), cytological profiling was based on analyzing cell shapes in the actin channel. Each image was visually examined to determine if the dsRNA altered the cell morphology induced by Rac^{V12}.

The prominent 3 cellular phenotypes observed in this assay have specific morphological characteristics that can be identified and quantified automatically. Cells with the normal phenotype have a smooth contour, and the actin intensity in the cell body is even (**Fig. 3a**). Cells with the spiky phenotype have actin-rich, spike-like protrusions at the periphery (**Fig. 3b**). For the ruffling cellular phenotype, cell body size is increased and actin accumulates with variable intensity and patterns (**Fig. 3c**).

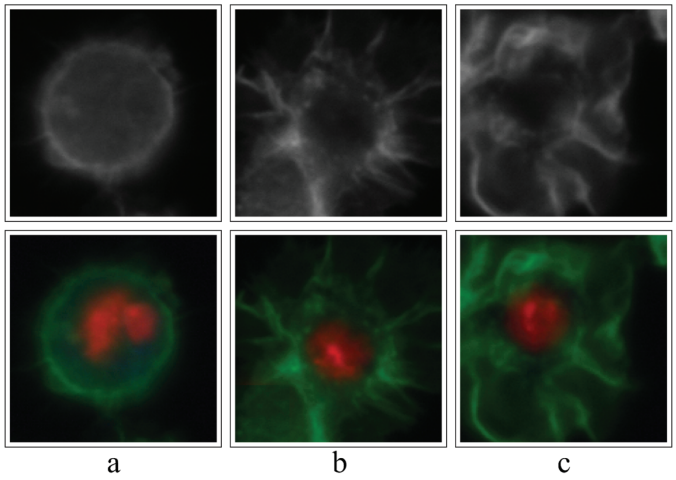


FIG. 3. *Drosophila* Kc167 cells have different cellular phenotypes: (**a**) normal, (**b**) spiky, and (**c**) ruffling. The upper row is the image of the F-actin channel, and the bottom row is the merged color images; green is F-actin and red is DNA.

Fluorescence image segmentation

The first step to automating the identification of cellular phenotypes in high-content images is to partition the images into regions that correspond to individual cells. Segmentation of these fluorescence microscopy images consists of 2 steps. The first step is to extract individual nuclei from the DNA channel, and the second step is to extract the individual cell bodies, using the F-actin channel. The entire segmentation procedure should provide detailed shape and boundary information of individual cells. The DNA signal is fairly strong, protruding from a relatively uniform dark background; thus, nuclei are easily segmented by a histogram thresholding technique. This approach correctly segments most isolated nuclei, but it is unable to segment touching-together or overlapping nuclei. Therefore, we applied a watershed algorithm to separate attached objects, as presented by Chen et al.[12] Although nuclear segmentation is straightforward, cytoplasmic

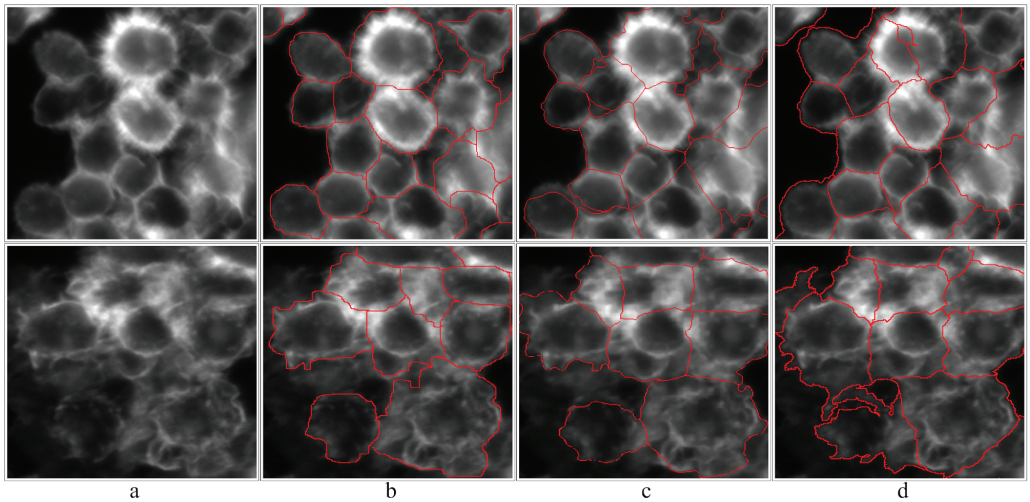


FIG. 4. Two examples of the segmentation of images of *Drosophila* Kc167 cells: (**a**) F-actin staining with no segmentation, (**b**) manual segmentation, (**c**) deformable model segmentation, and (**d**) CellProfiler™ segmentation.

segmentation remains a challenging task due to intensity variation and cellular phenotype diversity.

A large array of segmentation algorithms have been developed in the past 30 years; however, there is no state-of-the-art technique that segments fluorescence microscopy images with robust performance and tolerable computation cost.[13-15] In our work, we used 2 recently developed segmentation techniques, a deformable model by Xiong et al.[7] and the watershed method implemented in CellProfiler™,[16] both of which were specially designed for segmenting fluorescence cellular images.

The validation of image segmentation results is difficult as it is expensive and time-costly to create a ground truth segmentation result. In our work, we validated the segmentation results in 2 aspects: 1) comparison of the 2 automated computer techniques and 2) comparison with manual segmentation results. Qualitatively, we can draw some general conclusions about the 2 automated methods from this study. First, both methods work well to segment the normal cellular phenotypes. The reason for this lies in the smooth shape and even intensity distribution of fluorescence in cells with the normal phenotype. Second, the performance of both methods is degraded when processing cells with complicated textures and boundaries. The spiky phenotype, with its jaggy contour, is especially difficult to segment accurately. Third, when cells are close to or touching each other, the CellProfiler™ method generates fragmental boundaries, whereas the deformable model method captures accurately the complex boundaries. Fourth, the CellProfiler™ method is much faster than the deformable model method, which is computationally costly. An example comparing the segmentation results is shown in **Figure 4**. Although both automated methods are capable of segmenting many cells appropriately, neither automated method is capable of discerning all cell boundaries. Although more accurate segmentation of such complex images is always desirable, we argue that reasonable performance in phenotype classification may be attainable by combining existing segmentation results with innovative techniques of feature extraction, selection, and machine learning methods. We validate such approaches with experimental results described in the following sections.

Phenotype feature extraction

As stated earlier, HCS images contain a variety of phenotypes. In our study of *Drosophila* Kc167 cells, the most prominent cellular phenotypes were categorized as normal, spiky, and ruffling. Automated phenotype identification relies on feature extraction, the most critical step for pattern recognition problems. Even for a single cellular phenotype, the overall shape and appearance can be quite different because the cells could be in different stages of a certain phenotype. To capture the geometric and appearance properties, we extracted 5 types of features: wavelet features, Zernike moments features, Haralick features, region property features, and phenotype shape descriptor features.

Wavelet features. The discrete wavelet transformation (DWT) has been adopted to investigate image characteristics in both scale and frequency domains. In our work, we applied 2 important wavelets techniques, the Gabor wavelet[17] and the Cohen-Daubechies-Feauveau wavelet (CDF9/7),[18] to extract phenotype texture.

The Gabor wavelet features were developed by Manjunath and Ma[17] and are formed by a set of multiscale and multiorientation coefficients to describe texture variations in an image. The Gabor wavelet features have been used as the texture signature for numerous image analysis applications, such as image retrieval, segmentation, and recognition.[19,20] As defined by

Daugman,[20] the 2-dimensional complex-value Gabor function is a plane wave restricted by a Gaussian envelope. After conducting the Gabor wavelet transformation on the cell image, the real and imaginary parts of the transformation coefficients C_R, C_I can be obtained. The magnitude of the transformed coefficients $C = \sqrt{C_R^2 + C_I^2}$ is used as the Gabor vector. Because the transformed coefficients are computed based on pixels, this procedure derives the magnitudes for each pixel in the image. In the texture feature extraction method,[17] the statistics, such as mean μ and standard deviation η of these magnitudes on the entire image, are calculated as the feature representation. Through changing the scales and orientations, a set of feature representations can be calculated, which provide rich texture signatures in the frequency domain. In our experiments, 4 scales and 6 orientations are used to compute a 70-dimensional feature ($\mu_{0,0}$, $\eta_{0,0}$, $\mu_{0,1}$, . . . $\mu_{4,6}$, $\eta_{4,6}$) for each segmented cell.

Furthermore, we performed the 3-level CDF97 wavelet transformation[18] on images to extract additional texture signatures. The minimum value, maximum value, mean value, the median value of maximum distribution, and the standard derivation are calculated for each transformed image. In total, we obtained 15 wavelet features of each segmented cell.

Zernike moments features. Simply speaking, the Zernike moments features of an image are calculated based on the particular weighted averages of the intensity values. They are generated with the basis functions of Zernike polynomials. As classical image features, Zernike moments have wide applications.[21] Here, we give a brief description for calculating Zernike moments features for each cell: 1) calculate the center of mass for each cell polygon image and redefine the cell pixels based on this center; 2) compute the radius for each cell, and define the average of the radii as R; and 3) map the pixel (x, y) of the cell image to a unit circle and obtain the projected pixel as (x', y'). Because the Zernike moments polynomials are defined over a circle of radius 1, only the pixels (x', y') within the unit circle will be used to calculate Zernike moments. Finally, 49 Zernike moments features are computed.

Haralick co-occurrence features. As a traditional image feature extraction technique, the Haralick co-occurrence features use co-occurrence distribution of the gray image to generate the texture signature. Roughly speaking, given an offset on the image, the co-occurrence distribution of the image is referred to as the distribution of co-occurring values of pixel intensities. In our method, a total of 14 attributes were extracted from each of the gray-level spatial-dependence matrices.[22] The extracted co-occurrence features were as follows: angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measures of correlation, and maximal correlation coefficient.[23]

Region property. We also used a set of common region properties to describe the shape and texture characteristics of the cells. For general texture description, the maximum, minimum mean value, and standard deviation of the intensity in the segmented cell area were used. Moreover, we used some weak shape descriptions, such as the lengths of the longest axis l_{max} and the shortest axis l_{min}, the ratio l_{max}/l_{min}, the area s of the cell, the perimeter p of the cell, and the compactness of the cell, which is calculated as compactness $= p^2/(4\pi \cdot s)$. If the perimeter of the minimum convex shape is p_c, then the roughness is roughness $= p/p_c$. In all, we extracted 12 general texture and shape features for each segmented cell region.

Phenotype shape descriptor. Ideally, if we precisely define the boundary of each cell, the 3 different phenotypes of *Drosophila* Kc167 cells have more obvious differences in shape rather than texture. Because the shape information provided in the region property features is inexact, we developed 2 additional kinds of shape descriptors as our problem-specific features. One feature is based on the ratio length of the central axis projection, and the other feature is the area distribution over each equal sector.

From the original cellular patch $I(x, y)$, the binary image $f(x, y)$ can be derived. The value of $f(x, y)$ equals 1 when the pixel (x, y) is located in the cell area; otherwise, $I(x, y) = 0$. The centroid of the cellular area (m_x, m_y) is obtained. Centered at the centroid, we get a series of the central axis as the line L_α, shown in **Figure 5**. The central projection along L_α denotes the length of the axis. The equation of L_α is based on the angle α of the axis and the centroid coordinate (m_x, m_y). The ratio length of the central projection is defined as the length of L_α divided by the perimeter of the cellular contour.

$$r_{L_\alpha} = \frac{1}{p} \int_{L_\alpha} f(x, y), \tag{1}$$

where p is the same with the perimeter calculated in region property. For each different angle α, the ratio length for the central axis is calculated. The angles are evenly sampled with 36 different values to derive a 36-dimensional ratio length feature that represents the shape of the cellular boundary.

The second shape descriptor is based on the distribution of sector areas. As shown in **Figure 5**, a fan bin S_β centered at the centroid is denoted. The ratio area is defined as the area of the bin to the area of the entire cellular region.

$$r_{S_\beta} = \frac{\iint_{(x, y) \in S_\beta} f(x, y)}{\iint f(x, y)} = \frac{\iint_{(x, y) \in S_\beta} f(x, y)}{s}. \tag{2}$$

The value of s is the same as the area calculated for the region property feature. The entire cellular region is angle-evenly

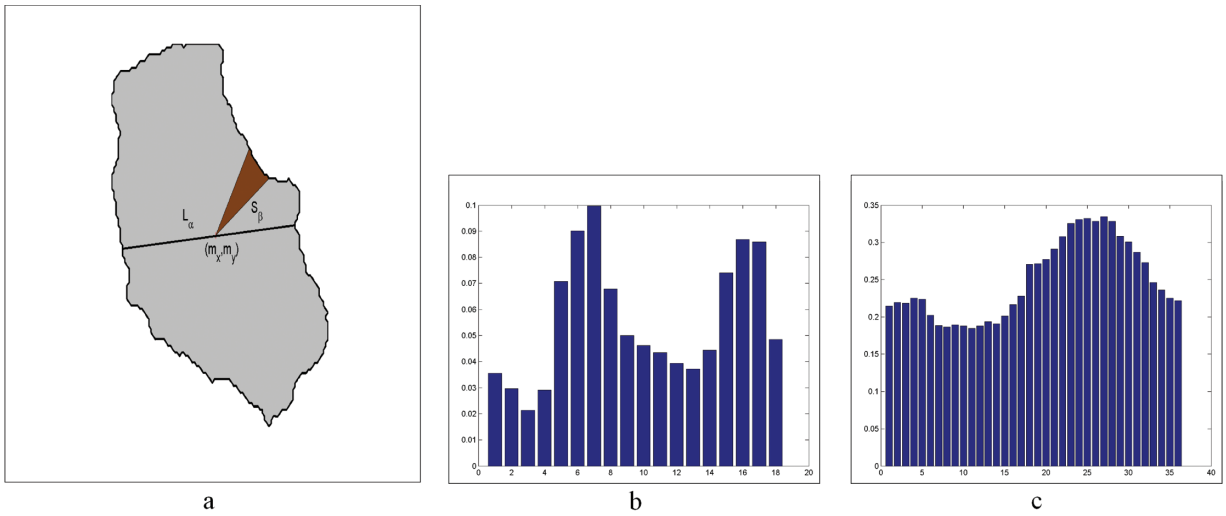


FIG. 5. The computation of the shape descriptor using ratio length and ratio area: (**a**) cellular region, central axis, and fan bin; (**b**) the 18-dimensional ratio area feature; and (**c**) the 36-dimensional ratio length feature.

partitioned into 18 sectors. Hence, the ratio area feature is constructed by the ratios of each sector. **Figure 5b,c** gives the computed shape features represented by the histograms, where the *x*-axis denotes the index of the sector and the value of the *y*-axis represents the percentage of area of the cell region or the length of the cell boundary falling into this sector. These 2 shape descriptors are scale and translation invariant but rotation variant. To achieve independence of rotation, the calculated ratio length and ratio area are sorted by value.

Phenotype feature subset selection

Using the various feature extraction procedures, we obtained 214 texture and shape features for each segmented cell image, covering diverse shape and texture properties. Our proposed framework for automated HCS analysis will be able to identify varied phenotypes in a wide range of cellular fluorescence images through the abundance of features. However, in each specific biomedical application, such as the fluorescence image of *Drosophila* Kc167 cells with 3 predominant phenotypes in our experiments, a concise subset of features must be obtained using different techniques for more efficient analysis. Thus, we added the automatic feature selection as one of the kernel components in the system, which makes the framework highly adaptable to different kinds HCS fluorescence images.

The objective of feature selection is to remove irrelevant and redundant features from the original feature space. Intrinsically, feature selection is a kind of optimization procedure with a predefined evaluation function. Neither global optimization nor full-space searching is feasible for feature selection. Thus, we applied a random search technique to derive an optimal feature subset. A genetic algorithm (GA) is a classical random optimization method that mimics the evolutionary process of survival of the fittest.[24] In brief, some individual feature subsets are initially created as the candidate sets, which are so called *population*. In successive iterations, the well-fitted individual subsets are selected from the population based on the evaluation of the fitness function. This selected portion of population breeds a new generation. There are 2 operators during the breeding, *crossover* and *mutation*. The crossover operator recombines portions of good selected individuals and is likely to create even better individuals. The mutation operation is executed with some low probability. A portion of the new individuals will have some of their bits flipped. Unlike the traditional hill-climbing optimization method, which has the drawback of easily falling into a local optimization, the mutation optimization can maintain the diversity during the search for optimal solutions because it induces a random walk through the search space. The evolution procedure of the GA can be terminated based on conditions, such as the maximum generations, running time, or fitness value threshold, which can be chosen based on the specific application. In our phenotype feature selection using GA, we selected the following parameters: population size of 200, maximum generation size of 50, and mutation rate of 0.3. In practice, we selected 12, 15, and 18 features from the original feature set and compared the performances (**Fig. 6**); the 15 features selected by the GA achieved better performance. As a comparison study, we also used principal component analysis (PCA) to reduce the dimensionality of the extracted features. Both GA-selected features and PCA-reduced features are validated in the phenotype identification experiments (see below). For the GA-selected features, in all the 3 cases (12, 15, 18 features), the region property and shape

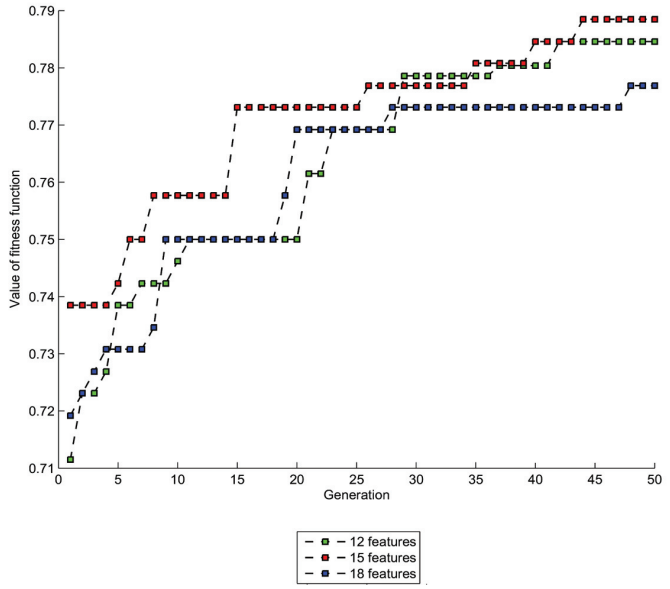


FIG. 6. The evolving procedure of feature selection by the genetic algorithm: the 3 lines represent selection of a subset with 12, 15, and 18 features. The *x*-coordinate is the generation and the *y*-coordinate is the value of the fitness function.

Table 1. Some of the Extracted Region Features for the 3 Cell Phenotypes Shown in Figure 3

Phenotype	*Normal*	*Spiky*	*Ruffling*
Long axis	117.08	154.07	163.09
Short axis	109.28	128.21	137.74
Perimeter	401.46	916.14	595.87
Area	9992	13367	16424
Compactness	1.28	4.98	1.72
Roundness	0.95	0.58	0.89

This table confirms that the phenotype differences may be best captured by compactness and roundness.

descriptors show superior performance to other features. It is reasonable because for the current set of phenotypes of *Drosophila* cells, the most salient difference among the predefined cellular phenotypes involves the geometric attributes of the cells, such as cell shape. For example, normal cells usually have a smooth shape; the actin-rich, spike-like protrusions of spiky cells generate complex boundaries; and ruffling is in between. As a simple illustration, **Table 1** shows some region features extracted for the 3 example cell phenotypes in **Figure 3**. It is obvious that the normal cell has the most roundness (more like a circle) and compactness (close to 1), and the spiky cell has the most irregular shape (large compactness value and small roundness).

RESULTS

Fluorescence images of Drosophila *cells*

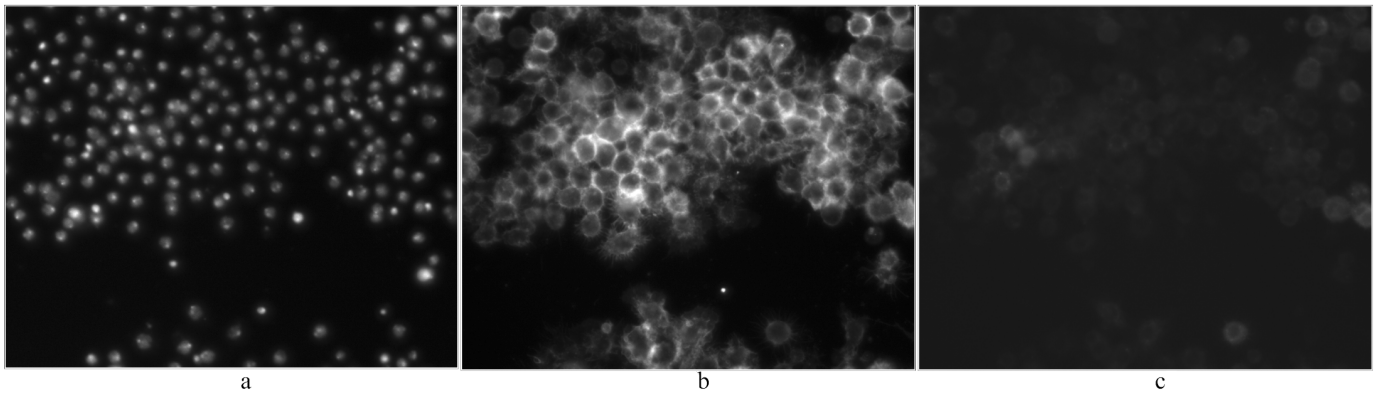
The proposed framework for automated HCS fluorescence image analysis was validated by its application in the phenotype classification of images of *Drosophila* cells. The original fluorescence images were 1280 × 1024 pixels and stored in a 12-bit format. Images contain 100 to 400 cells. One of the image examples used in our experiments is shown in **Figure 7**, where the DNA channel, F-actin channel, and Rac^{V12} channel are displayed separately. To build a cellular phenotype database, we first performed the segmentation process on the image and then manually classified the segmented cells into 3 categories of phenotypes (normal, spiky, and ruffling), ignoring inaccurately segmented cells. We constructed a cellular phenotype database, including around 600 normal phenotype cells, 200 spiky phenotype cells, and 200 ruffling phenotype cells. The database was used for training the classifiers and validating the proposed method.

Classification models

To evaluate the features and our system, we constructed 4 well-known classifiers—naive Bayesian classifier (NBC), linear discriminant analysis (LDA), K nearest neighbors (KNN), and support vector machine classifier (SVC)—and used them in the standard model based on the handbook by Duda et al.[25] 1) NBC is based on the Bayes theorem, with the strong constraint of feature independence. Although the assumption of feature independence is not strict for real cases, NBC remains a popular classifier because of its robustness and easy implementation. 2) LDA finds an optimal linear transformation, which retains class separability while reducing variation within each class. 3) KNN is an incremental and lazy classification algorithm. This classifier is simple and easily implemented. 4) SVC, originally explored by Vapnik,[26] is an important kernel technique used for pattern analysis. SVC can estimate a maximum-margin hyperplane to separate the data. Usually, the selection of the appropriate kernel functions and the corresponding parameters is vital for a real application.

Phenotype classification results

We first conducted the classification experiments on the 1000-cell-patch database. To estimate the generalization error, we applied the 10-folder cross-validation technique. The leave-one-out strategy, as one of the simplest and widely used versions of cross-validation, efficiently estimated the results. The procedure was repeated 100 times, and the performance is calculated as the mean value of the 100-time tests. Both feature selection by GA and feature reduction by PCA were tested. The GA was used to select a subset with 12, 15, and 18 features for tests.



a b c

FIG. 7. An example of fluorescence images used in our experiments: (**a**) DNA channel, (**b**) F-actin channel, and (**c**) Rac^{V12} channel.

Table 2. Experimental Results of GA Features and PCA Features

Feature/Classifier	*NBC, %*	*LDA, %*	*KNN, %*	*SVC, %*	
GA features	12	62.68	74.26	65.90	67.10
	15	62.29	76.08	65.75	69.73
	18	61.21	74.67	64.83	65.17
PCA features	12	59.84	64.90	60.73	61.16
	18	60.26	65.33	60.96	61.18
	48	53.37	68.15	59.80	59.89

All 5 types of features with 214 attributes are used. The principal component analysis (PCA) features are obtained using PCA to reduce the dimensionality to 12, 18, and 48, which conserves respectively 85%, 90%, and 98% energy. GA, genetic algorithm; NBC, naive Bayesian classifier; LDA, linear discriminant analysis; KNN, K nearest neighbors; SVC, support vector machine classifier.

Table 3. Confusion Matrix of LDA Classifier with 15 GA-Selected Features (in Percentages)

Input/Output	*Normal*	*Spiky*	*Ruffling*
Normal	86.94	3.21	9.85
Spiky	12.74	67.41	19.85
Ruffling	14.08	15.81	70.10

GA, genetic algorithm; LDA, linear discriminant analysis.

Table 4. Sensitivity and Specificity of Different Phenotypes (jn Percentages)

Input/Output	*Normal*	*Spiky*	*Ruffling*
Sensitivity	86.94	67.41	70.10
Specificity	86.44	90.65	86.40

Meanwhile, the PCA was applied to reduce the feature dimensionality to 12, 15, and 39, which conserved respectively 85%, 90%, and 98% energy (**Table 2**). The reliability of the LDA was also evaluated by the Cohen kappa coefficient, which was around 0.58.[27] Two conclusions were derived from the experiments. First, the GA-selected features are superior to PCA-reduced features. Second, the LDA classifier achieved the highest performance on both GA features and PCA features. When the 15 GA-selected features were used, the LDA classifier achieved the best performance of more than 76% accuracy. The confusion matrix of the LDA classifier using the 15 GA-selected features is calculated (**Table 3**). Altogether, the normal phenotype was the most easily identified, and spiky was the most difficult to identify and easily misclassified into the other 2 phenotypes (**Table 4**). Compared with the convenience of a LDA classifier without any parameter optimization requirements, the performance of SVC highly depends on the kernel selection and parameters estimation. There are no efficient guidelines for the determination of kernels. The usual empirical solution is based on an exhaustive search of both kernel space and parameter space. Therefore, more training data are required to avoid the overfitting problem. In our current study, the kernel and parameter selection issue of SVC has not been investigated because the specific classifier optimization is not the main focus.

Sensitivity and *specificity* are 2 widely used criteria for evaluating binary testing. To investigate these 2 criteria for a certain phenotype, we first transformed the multiple-category problem to a 2-category problem. For example, the cellular phenotypes were categorized as normal or abnormal, which included both spiky and ruffling. The calculated evaluation by sensitivity and specificity is shown in **Table 4**, where the result demonstrates that the spiky cells are most likely to be misclassified as the other 2 phenotypes.

Finally, we applied the proposed approach on whole fluorescence images of *Drosophila* cells, which were captured for a HCS-based gene functional analysis. An example of the process is shown in **Figure 8**, indicating the output of nuclei segmentation (**Fig. 8a**), cytoplasm segmentation (**Fig. 8b**), and

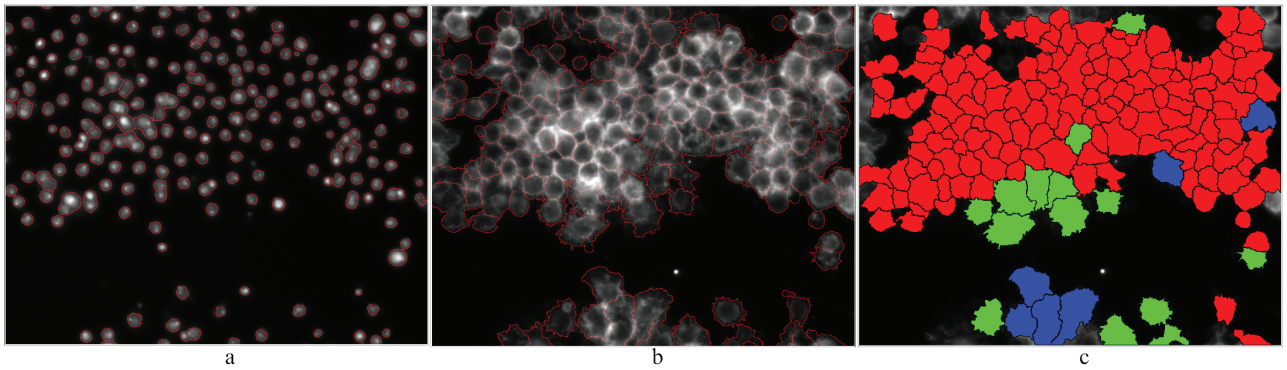


FIG. 8. The segmentation results for images in **Figure 7** are shown: (**a**) nuclei segmentation and (**b**) cell bodies segmentation. The phenotype classification results are displayed in **c**. Red represents normal cells, green represents spiky cells and blue represents ruffling cells.

cellular phenotype classification (**Fig. 8c**). As shown in this figure, the extracted information from the raw images, such as cellular segmentation and phenotype identification results, can provide abundant quantitative information for investigating the function of the related genes in the biological process. Hence, one of the future directions is to apply the proposed framework to investigate the gene function using RNAi HCS image data. In our ongoing study on the gene function evaluation using RNAi HCS images, we have applied the proposed image analysis technique on an image data set of 255 scanned screens, which are captured to test individual gene knock-down effects of 85 genes (3 sites scanned for each gene knock-down condition). Given the phenotype classification results on each image, we compute statistical measures such as the ratio of different cell phenotypes present in the image. Such statistics may be used to assess the impact of knocking down each individual gene and thus quantitatively predict the relevance of the gene with respect to the biological problem (e.g., drug design, protein interaction) under study. In other words, we can predict a scalar score to estimate the importance of each knocked-down gene. Our preliminary study has shown promising results of gene function estimation using the automatic image analysis tool proposed in this article, comparable to the results based on manual scoring by experts.

DISCUSSION AND CONCLUSION

Rho GTPases are essential for the cytoskeletal rearrangements that drive morphogenesis and cell migration during normal development, as well as in cancer metastasis. Thus, genome-wide screens designed to identify novel components of Rho signaling are critical to further our understanding of developmental and cancer biology. The most fundamental and vital step for the success of high-content RNAi genome-wide screening-based research is the development of automated cellular phenotype identification of high-content images.

In this report, we designed a framework for identifying a cellular phenotype for automated HCS analysis. Cell segmentation is executed using 2 recently developed fluorescence image segmentation techniques, the deformable model and CellProfiler™ method. Both of these approaches start from the nuclear segmentation in the high-contrast DNA channel and derive the cell body contour in the F-actin channel.

As one of the most important steps toward achieving automated phenotype classification, several different image feature extraction methods are used to generate an extensive phenotype descriptor of the segmented cells. Wavelet features, moment features, and co-occurrence features are extracted to define the texture signature of the cellular phenotype. The region property feature, as a kind of simple statistical property, is also used to describe the phenotype feature in both intensity and region shape. Furthermore, we designed ratio length and area-based shape features to describe additional details of the cellular shape. A total of 214 image features are combined to create the candidate attributes for the general purpose of phenotype classification. For specific biomedical applications, an optimal panel of efficient phenotype features can be selected to remove the redundant and irrelevant information. A genetic algorithm was used to select a discriminant subset of features for the input of the classification module. Four widely used classification methods—NBC, LDA, KNN, and SVC—were used to identify the cellular phenotype.

Although our studies were conducted on fluorescence images of *Drosophila* Kc167 cells, the proposed system is intrinsically designed for generalized phenotype identification of HCS images. The extracted diverse features, the feature selection procedure, and the training of classifiers guarantee that this framework can be easily adapted to classify other predefined cellular phenotypes in fluorescence microscopy images of other species.

In summary, we have built a prototype of an automatic cellular phenotype identification system for RNAi genome screening. Clearly, the low quality of images, inaccurate segmentation techniques, and imprecise biological definition of complex cellular

phenotypes present extreme challenges to automated HCS analysis.[28] In our studies, the accuracy of classification was between 67% and 87% for different phenotypes, and the overall accuracy was around 76%. Yi and Coppolino[29] recently reported their automatic ruffling classification method in the literature, which applied ImageJ as the image analysis tool and SPSS for statistical analysis. They used the moment features, measured metrics, and geometric descriptor to identify ruffling and nonruffling in confocal micrographs (no clear statistical performance compared with manual ground truth reported). Ideally, the state-of-the-art automatic analysis system for microscopic images is expected to recognize all the possible phenotypes with high efficiency and robustness under various imaging and biological conditions. However, note that the HCS screening data used in our experiments have more challenges in terms of image complexity and quality. Thus, even though this performance is not yet state of the art, this accuracy should be useful for the purpose of automated screening. Importantly, several issues need to be addressed to improve the performance of this system. First, the current cellular segmentation cannot achieve satisfactory results with both fast computation and high precision. Second, some cells exhibit ambiguous phenotypes that are difficult to classify. A conventional unsupervised learning technique, clustering, may help discover new phenotypes. Regardless of these issues, our current phenotype classification method could be applied to building a scoring system to quantitatively evaluate fluorescence microscopy images by establishing certain regression models based on the ratios of each phenotype, thus providing a score that allows biologists to easily find novel genes of interest.

ACKNOWLEDGMENTS

The authors thank Baillie Yip for labeling training data sets. They also thank research members of the Life Science Imaging Group of the Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair (HCNR) and Brigham and Women's Hospital, Harvard Medical School, particularly Ashley Tarokh and Kuang Yu Liu, for their technical comments. The research is funded by the HCNR, Harvard Medical School (S. T. C. Wong).

REFERENCES

1. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschule SJ: Multidimensional drug profiling by automated microscopy. *Science* 2004; 306:1194-1198.
2. Yarrow JC, Feng Y, Perlman ZE, Kirchhausen T, Mitchison TJ: Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb Chem High Throughput Screen* 2003;6:279-286.
3. Hannon GJ: RNA interference. *Nature* 2002;418:244-251.
4. Boland MV, Murphy RF: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* 2001;17:1213-1223.
5. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, et al: Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* 2004;303:832-835.
6. Kiger AA, Baum B, Jones S, Jones MR, Coulson A, Echeverri C, et al: A functional genomic analysis of cell morphology using RNA interference. *J Biol* 2003;2:27.
7. Xiong G, Zhou X, Ji L, Bradley P, Perrimon N: Automated segmentation of *Drosophila* RNAi fluorescence cellular images using deformable models. *IEEE Trans Circuit Systems* 2006;53:2415-2424.
8. Lamprecht MR, Sabatini DM, Carpenter AE: CellProfiler™: free, versatile software for automated biological image analysis. *BioTechniques* 2007;42:71-75.
9. Aelst LV, Symons M: Role of Rho family GTPases in epithelial morphogenesis. *Genes Dev* 2002;16:1032-1054.
10. Bouzahzah B, Albanese C, Ahmed F, Pixley F, Lisanti MP, Segall JD, et al: Rho family GTPases regulate mammary epithelium cell growth and metastasis through distinguishable pathways. *Mol Med* 2001;7:816-830.
11. Echeverri CJ, Perrimon N: High-throughput RNAi screening in cultured cells: a user's guide. *Nat Rev Genet* 2006;7:373-384.
12. Chen X, Zhou X, Wong STC: An automated method for cell phase identification in high throughput time-lapse screens. *IEEE Trans Biomed Eng* 2006;53:762-726.
13. Duncan JS, Ayache N: Medical image analysis: progress over two decades and the challenges ahead. *IEEE Trans Pattern Anal Machine Intell* 2000;22:85-106.
14. Pham TD, Crane DI, Tran TH, Nguyen TH: Extraction of fluorescent cell puncta by adative fuzzy segmentation. *Bioinformatics* 2004;20: 2189-2196.
15. Zhou X, Liu KY, Bradley P, Perrimon N, Wong STC: Towards automated cellular image segmentation for RNAi genome-wide screening. *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv* 2005;8(pt 1):885-892.
16. Jones TR, Carpenter AE, Golland P: Voronoi-based segmentation of cells on image manifolds. Paper presented at the ICCV Workshop on Computer Vision for Biomedical Image Applications, Beijing, China, October 2005.
17. Manjunath BS, Ma WY: Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Machine Intell* 1996;18:837-842.
18. Cohen A, Daubechies I, Feauveau JC: Bi-orthogonal bases of compactly supported wavelets. *Pure Appl Math* 1992;45:485-560.
19. Bovic AC, Clark M, Geisler WS: Multichannel texture analysis using localized spatial filters. *IEEE Trans Pattern Anal Machine Intell* 1990;12:55-73.
20. Daugman JG: Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Trans Acoustics Speech Signal Processing* 1988;36:1169-1179.
21. Zernike F: Beugungstheorie des Schneidencerfarhens undseiner Verbesserten Form, der Phasenkontrastmethode. *Physica* 1934;1:689-704.
22. Haralick RM, Shanmugam K, Dinstein I: Textural features for image classification. *IEEE Trans Systems Man Cybernet* 1973;3:610-620.
23. Haralick RM: Statistical and structural approaches to texture. *Proc IEEE* 1979;67:786-804.
24. Holland JH: *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. 2nd ed. Cambridge, MA: MIT, 1996.
25. Duda RO, Hart PE, Stork DH: *Pattern Classification*. 2nd ed. New York: Wiley Interscience, 2000.

26. Vapnik V: *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
27. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
28. Zhou X, Wong STC: Informatics challenges of high-throughput microscopy. *IEEE Signal Processing Magazine* 2006;23:63-72.
29. Yi Q, Coppolino MG: Automated classification and quantification of F-actin-containing ruffles in confocal micrographs. *BioTechniques* 2006;40:745-756.

Address correspondence to:
Xiaobo Zhou
Department of Radiology
Brigham and Women's Hospital
Boston, MA 02115

E-mail: zhou@crystal.harvard.edu