Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An image score inference system for RNAi genome-wide screening based on fuzzy mixture regression modeling

Jun Wang [a,*], Xiaobo Zhou [b,*], Fuhai Li [b], Pamela L. Bradley [c], Shih-Fu Chang [a], Norbert Perrimon [d], Stephen T.C. Wong [b]

[a] *Department of Electrical Engineering, Columbia University, 1300 S.W. Mudd, 500 West 120th Street, New York, NY 10027, USA*
[b] *Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair, Harvard Medical School, 3rd floor, 1249 Boylston, Boston, MA 02215, USA*
[c] *National Institute of Neurological Disorders and Stroke, National Institutes of Health, 37 Convent Drive, Bethesda, MD 20892, USA*
[d] *Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA*

ARTICLE INFO

Article history:
Received 10 November 2007
Available online 29 April 2008

Keywords:
High-content screening
Image score inference

ABSTRACT

With recent advances in fluorescence microscopy imaging techniques and methods of gene knock down by RNA interference (RNAi), genome-scale high-content screening (HCS) has emerged as a powerful approach to systematically identify all parts of complex biological processes. However, a critical barrier preventing fulfillment of the success is the lack of efficient and robust methods for automating RNAi image analysis and quantitative evaluation of the gene knock down effects on huge volume of HCS data. Facing such opportunities and challenges, we have started investigation of automatic methods towards the development of a fully automatic RNAi–HCS system. Particularly important are reliable approaches to cellular phenotype classification and image-based gene function estimation.

We have developed a HCS analysis platform that consists of two main components: fluorescence image analysis and image scoring. For image analysis, we used a two-step enhanced watershed method to extract cellular boundaries from HCS images. Segmented cells were classified into several predefined phenotypes based on morphological and appearance features. Using statistical characteristics of the identified phenotypes as a quantitative description of the image, a score is generated that reflects gene function. Our scoring model integrates fuzzy gene class estimation and single regression models. The final functional score of an image was derived using the weighted combination of the inference from several support vector-based regression models. We validated our phenotype classification method and scoring system on our cellular phenotype and gene database with expert ground truth labeling.

We built a database of high-content, 3-channel, fluorescence microscopy images of *Drosophila* Kc_{167} cultured cells that were treated with RNAi to perturb gene function. The proposed informatics system for microscopy image analysis is tested on this database. Both of the two main components, automated phenotype classification and image scoring system, were evaluated. The robustness and efficiency of our system were validated in quantitatively predicting the biological relevance of genes.

1. Introduction

High-content screening (HCS) utilizes automated microscopy techniques and fluorescence probes to visualize details of complex cellular activities, such as mitosis and cell migration. As a result of recent technological advances, such as fast and reliable digital scanning, high-performance computing, and high-precision bioimaging techniques, HCS has increasingly been applied in genomic research and medicine. For example, changes in cellular phenotypes resulting from gene perturbation can be captured in high-throughput, allowing rapid identification of genes involved in a cellular process of interest. Furthermore, HCS technology is being actively applied in research on disease diagnosis and prognosis, drug target validation, and lead compound selection [1–3]. Although our manual annotation of images from fluorescence-based screens has provided encouraging results in small-scale screens [4], biologists face the enormous challenge of analyzing the immense volumes of images generated by genome-scale studies. The goal of informatics for genome-wide HCS is to convert, or translate, the information displayed in fluorescence images into quantitative descriptors, which then can be linked to statistical analysis that scores the image's overall phenotype. Existing imaging analysis tools are extremely limited in their scope and capacity to analyze individual difference and spatial information in high-content, fixed-cell imaging. Major informatics challenges of high-

[*] Corresponding authors.
E-mail addresses: jwang@columbia.edu (J. Wang), zhou@crystal.harvard.edu (X. Zhou).
URL: http://authors.elsevier.com/locate/latex (J. Wang).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.04.007

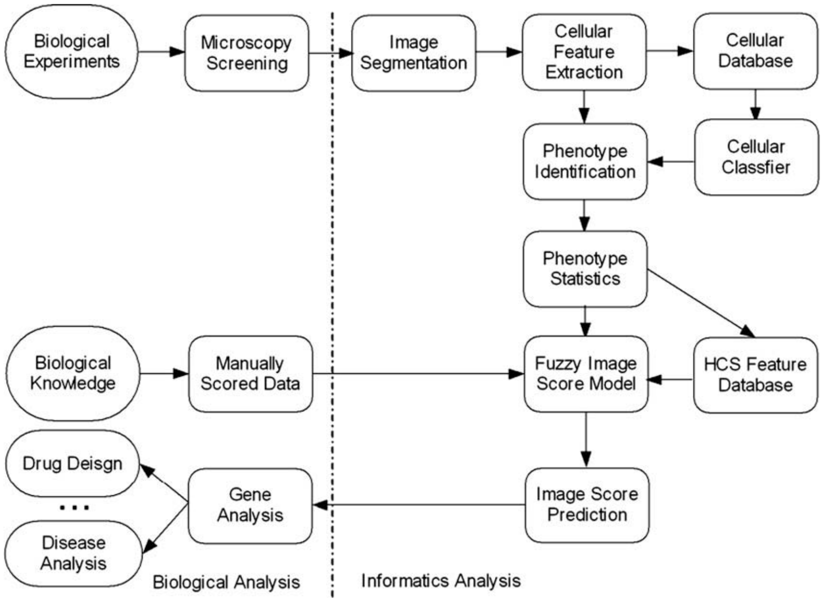

Fig. 1. Information processing pipeline for RNAi genome-wide high-content screening-based image score inference system.

content images include fluorescence image processing, cellular feature quantification and identification, statistical inference modeling and validation of gene scoring (reviewed by [5]). Image-based functional analysis and target discovery using HCS requires the cross-disciplinary collaboration of biology, computer science and mathematics. Recently, there are some emerging approaches proposed for automatic HCS image analysis, such as CellProfiler [6] and the cellular phenotype classification system proposed in [7]. However, these frameworks mainly focus analysis of cells instead of quantitatively analysis of HCS screening.

Here, we have built a pipeline for automatically cellular phenotype identification and RNAi HCS image scoring, summarized in the flowchart in Fig. 1 The key components of the proposed informatics pipeline are: microscopic image processing, automatic phenotype classification, and image scoring for gene function evaluation. With the well developed fluorescence image segmentation approach, the cellular phenotypes are identified using the extracted cell level morphological and appearance features. Hence, the visual characteristics of an image are obtained statistically based on the distribution of these phenotypes in the fluorescence screening corresponding to a particular treatment. Finally, a statistical modeling was built to map the image level description to the quantitative identification of gene function regard to the distribution and changes of the phenotypes observed in HCS. Our experimental results show the proposed system is a robust and efficient tool for genome-wide functional analysis using HCS.

The remainder of this paper is organized as follows. In Section 2, the generation procedure of RNAi HCS images of *Drosophila* Kc_{167} cells of is introduced, followed by a brief summary of segmentation approach for microscopic image. In Section 3, we present the phenotype classification model. Section 4 gives the methodology of image score inference. The experiments on phenotype recognition and image scoring are conducted in Section 5. Finally, concluding remarks and our future work will be given in Section 6.

2. Fluorescence image acquiring and preprocessing

2.1. Fluorescence image-based screening of Drosophila Kc_{167} cells

The morphological diversity of cells results in large part from the dynamic control of the cytoskeleton. Some of the major cytoskeletal regulators are members of the Rho family of small GTPase proteins, which are essential for morphological changes during normal development, as well as during disease states such as cancer [8,9]. Rho proteins also regulate many other facets of cell behavior, such as endocytosis, vesicle trafficking, cell polarity, and cell cycle. Rho GTPases cycle between an active, GTP-bound state, and an inactive, GDP-bound state. In the active state, Rho proteins interact with effector molecules and modulate their activities to relay upstream signals and implement downstream responses [10]. Identification of novel Rho effectors will elucidate the mechanisms by which Rho proteins orchestrate their varied cellular outcomes. Thus, we have designed a cell-based assay for Rho GTPase activity that is amenable to HCS with the intent of identifying novel effectors.

Expression of the constitutively active forms of Rho proteins causes distinct morphological changes to a multitude of cell types [11], including the *Drosophila* Kc_{167} embryonic cell line. Kc_{167} cells are small (10 μm) and uniformly round with little filamentous actin (F-actin) cytoskeletal structure. Expression of the constitutively active form of Drac1 (Rac^{V12}) induces an increase in the levels of F-actin, as well as the formation of large flat protrusions called lamella, which are dynamic and ruffle, and spike-like protrusions.

To facilitate HCS, we generated a construct containing sequences encoding a $GFP\text{–}Rac^{V12}$ 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 double stranded RNA (dsRNA) specific to predicted *Drosophila* genes to elicit the RNA interference (RNAi) response, which mimics loss-of-function mutations in the targeted gene [12]. To perform the screen, dsRNAs were robotically arrayed individually in 384-well plates. *Drosophila* cells were plated in each well, where they take up the dsRNA from the culture media. Two or three images per well in each of three channels were acquired by automated microscopy with a Universal Imaging AutoScope, a Nikon TE300 inverted fluorescence microscope, using a 40× air objective. An example of the 3-channel RNAi fluorescence images used for cytological profiling is shown in Fig. 2. The signal in DNA channel indicates the locations and shape of nuclei of cells (Fig. 2a). The actin channel reveals cytoskeletal structure, used to determine the morphology of cell bodies (Fig. 2b). Since relatively little visual information is available from the $GFP\text{–}Rac^{V12}$ channel (Fig. 2c), cytological profiling

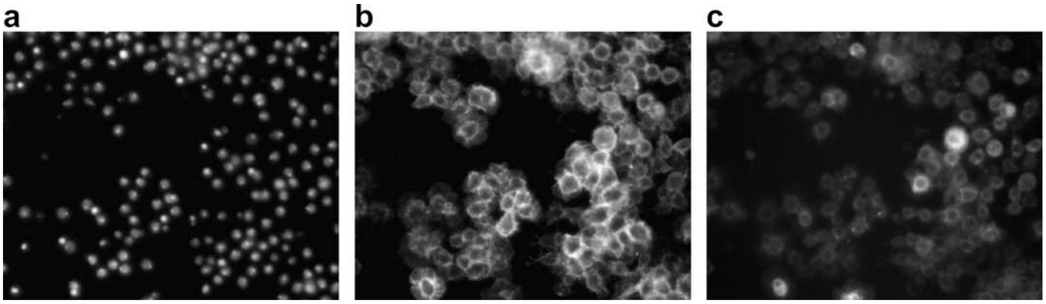

Fig. 2. *Drosophila Kc*$_{167}$ cells fluorescence images with three channels: (a) DAPI-stained DNA, (b) TRITC-phalloidin-stained F-actin, and (c) *GFP–Rac*^{V12}.

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}.

2.2. Fluorescence microscopy image segmentation

In order to build the automated phenotype classification system, the first step was to perform segmentation of the fluorescence microscopy images. The objective of segmentation is to partition the images into regions that correspond to the cells, thus allowing subsequent quantification of cellular phenotypic features. Although numerous algorithms have been developed for image segmentation in the past 30 years, there is no state-of-the-art technique that can be used to segment fluorescence microscopy images with robust performance and tolerable computation cost [13,14].

The details of our segmentation method are presented elsewhere [15–17]. In the proposed framework, we used a two-step enhanced watershed-based method obtain rough segments, followed by segmentation revision by a deformable model approach. As described in [15], the general seeded watershed method for cell image segmentation consists of two stages: nuclear segmentation by the ISODATA thresholding method [18] on DNA channel images, and cytoplasm segmentation on Actin channel. Basically, this algorithm correctly segmented most isolated nuclei. However, it has the drawback of over-segmentation problems caused by dividing cells and isolated nuclei. Therefore, we enhanced the seeded watershed method using an automatic feedback scheme to interactively validate the segmentation results. Specially, the approach proposed in [17] was applied in our experiments, in which an automated feedback system was built to reduce the over-segmentation in both nuclei and cytoplasm segmentation. Starting from the initial cell segmentation, the deformable model-based approach was applied to refine the cell boundaries [16], which especially improved segmentation accuracy for non-regular cells.

3. Cellular phenotype identification

3.1. Phenotype feature extraction and selection

As stated above, HCS images contain a variety of phenotypes. In our study of *Drosophila Kc*$_{167}$ 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 a total of 214 cellular attributes, which belong to five types of features: wavelet features, Zernike moments features, Haralick features, region property features, and phenotype shape descriptor features.

3.1.1. 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 two important wavelets techniques, the Gabor wavelet [19] and the Cohen–Daubechies–Feauveau wavelet (CDF9/7) [20], to extract phenotype texture. The Gabor wavelet features were developed by Manjunath et al. and is formed by a set of multi-scale and multi-orientation coefficients to describe texture variations in an image [19]. The Gabor wavelet features have been used as the texture signature for numerous image analysis applications, such as image retrieval, segmentation and recognition [21,22]. As defined in [22], the two-dimensional complex-value Gabor function is a plane wave restricted by a Gaussian envelope:

$$g(\omega,k) = \frac{\omega^2}{\sigma^2}\mathrm{e}^{-\frac{\omega^2k^2}{2\sigma^2}}\left(\mathrm{e}^{\mathrm{i}\omega k} - \mathrm{e}^{-\frac{\omega^2}{2}}\right) \tag{1}$$

where $\mathrm{e}^{\mathrm{i}\omega k}$ is the complex-value plane wave, and $\mathrm{e}^{-\frac{\omega^2k^2}{2\sigma^2}}$ is the Gaussian envelope function, which is applied to restrict the complex-valued plane wave. Assume that the cell image is represented as $I(x,y)$, the Gabor wavelet transformation can be computed as the spatial convolution with the Gabor wavelet function given certain parameters of scale and orientation. Corresponding to the real and imaginary parts of the Gabor wavelet function, the wavelet transformation outputs real and imaginary components C_R,C_I, respectively. The magnitude of the transformed coefficients $\|C(x,y)\| = \sqrt{C_R^2 + C_I^2}$ is used as the Gabor vector. In the texture feature extraction method of [19], the mean μ and standard deviation η of these magnitudes are calculated as the feature representation. Considering M scales and N orientations, we obtained a $2(M+1)\cdot(N+1)$ dimensional features ($\mu_{0,0}$, $\eta_{0,0}$, $\mu_{0,1}$, $\eta_{0,1}$, ..., $\mu_{M,N}$, $\eta_{M,N}$) for each segmented cell. In our experiments, we set the values as: $M = 6$, $N = 4$. Hence, we finally get a Gabor wavelet feature with 70 dimensionality.

Furthermore, we performed the 3-level CDF97 wavelet transformation [20] 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. For both of these wavelet transformations, the feature extraction is conducted on a rectangle region with the cell segment sitting in the center. The region outside the cell segments is filled with zero intensity pixels. In total, we obtained 15 CDF97 wavelet features of each segmented cell.

3.1.2. Zernike moments features

Zernike moments are classical image features that have wide applications [23]. 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. (3) Map the pixel (x,y) of the cell image to a unit circle and obtain the projected pixel as (x',y'). Since the Zernike moments polynomials are defined over a circle of radius 1, only the pixels within the unit circle will be used to calculate Zernike moments. (4) Calculate the zernike moments based on the projected image $I(x',y')$. We select the order as 12 and use the magnitude of Zernike moments as the feature to obtain 49 moments features in total ([24]).

3.1.3. Haralick co-occurrence features

As a traditional texture signature, the Haralick Co-occurrence features, with a total of 14 attributes, were extracted from each of the gray-level spatial-dependence matrices [25]. The extracted co-occurrence features were: 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 [26].

3.1.4. 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 of the cell s, the perimeter p of the cell, and the compactness of the cell compactness = $p^2/(4\pi s)$. If the perimeter of 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.

3.1.5. Phenotype shape descriptor

Since the shape information provided in the region property features is inexact, we developed two additional kinds of shape descriptors, ratio length of the central axis projection and the area distribution over each equal sector, as our problem-specific features. From the original cellular patch $I(x,y)$, the binary image $f(x,y)$ can be derived. The centroid of the cellular area (m_x, m_y) is the first order moments of the binary cell patch. Centered at the centroid, we get a series of central radial axis as the line L_a. The central projection along L_a denotes the length of the axis. The equation of is based on the angle of the axis and the centroid coordinate. The ratio length of the central projection r_{L_a} is defined as the value of the axial length divided by the perimeter of the cellular contour $r_{L_a} = \frac{1}{p}\int_{L_a} f(r)\mathrm{d}r$, where p is the perimeter of the cell. For each different angle, the ratio length of the central axis is calculated. The angles are evenly sampled with different values to derive a 36-dimensional ratio length feature that represents the shape of the cellular boundary. The other shape descriptor is based on the distribution of sector areas. The ratio area is defined as the area of the fan bin S_β center at the cellular centroid with even angle to the area of entire cellular region: $r_{S_\beta} = \frac{\int_{(x,y)\in S_\beta} f(x,y)\mathrm{d}x\mathrm{d}y}{\int f(x,y)\mathrm{d}x\mathrm{d}y}$. The entire cellular region is angle-evenly partitioned into 18 sectors. Hence, the ratio area feature is constructed by the ratios of each sectors. These two 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.

With the above feature extraction procedures, we obtained a abundant feature pool for cell segments, covering diverse shape and texture properties. The abundance of features used for phenotype description will be applicable for identifying varied phenotypes in a wide range of cellular fluorescence images. However, in each specific biomedical study, such as the fluorescence image of *Drosophila* Kc_{167} cells with three predominant phenotypes in our experiments, a concise subset of features will make the system more computationally efficient and well-fit this certain study. Hence, as shown in the system diagram of Fig. 1, a kernel component of automatic feature subset selection is incorporated in the system to make the framework highly scalable and adaptable to various HCS study. Simply speaking, the procedure of feature selection is to remove irrelevant and redundant features from the original feature space. In the research community of machine learning and pattern analysis, there are some techniques proposed for feature selection [27]. However, most approaches are specifically developed to certain applications, which may not fit to the HCS analysis. Thus, we applied a very general random search technique, Genetic algorithm (GA), to derive an optimal feature subset. GA is a classical random optimization method, which mimics the evolutionary process of survival of the fittest [28]. In brief, some individual feature subsets are initially created as the candidates 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. 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 practice, we selected 12, 15 and 18 features from the original feature set and compared the performances; the 15 features selected by the GA achieved better performance [7].

3.2. Phenotype classification

Traditionally, gene function has been assessed by analyzing alterations in a biological process caused by the absence or disruption of a gene. Combining high-throughput methodologies, such as automated fluorescence microscopy, with techniques to interfere with gene function, such as RNAi, has become an efficient way to conduct large-scale functional analysis. Quantification of cellular phenotypes in fluorescence images of RNAi-treated cells allows the identification of genes that have a role in the process of interest. In the present work, we sought to identify genes in the Rho signaling pathway by asking which dsRNAs alter the distribution of the cellular phenotypes caused by expression of Rac^{V12} (Spiky and Ruffling). The 'Normal' cellular phenotype, present in wild-type Rac^{V12} cells (not expressing Rac^{V12}), had a smooth contour, and the intensity or energy distributed in cell body region is relatively even (Fig. 4a). The 'Spiky' cellular phenotype was characterized by spike-like extensions of the cell body (Fig. 4b). For the 'Ruffling' cellular phenotype, the cytoplasm was increased and large protrusions (lamella) were seen at the periphery (Fig. 4c).

To classify a segmented cell as one of these three cellular phenotypes by automated analysis, we needed to identify features of the phenotypes that distinguish them from each other. The geometric properties and appearance of the different phenotypes were represented by the congenital texture features, as described in Section 3.1. In order to achieve the computational simplicity and classification efficiency, Genetic algorithm was applied to select discriminate subsets of the extracted features that would facilitate classification. Then, to identify the phenotype of a segmented cell, we applied different classification methods, including linear discriminant analysis (LDA) and support vector classifier (SVC) [29,30]. The phenotype classification results will be reported in Section 5.

3.3. Phenotype statistical property of HCS screening

Thus far, we have captured images of cells treated with specific dsRNAs, segmented the images to identify individual cells, and classified and labeled the segmented cells as N (Normal), S (Spiky) and R (Ruffling) (Fig. 4). The question remains: did the addition of a

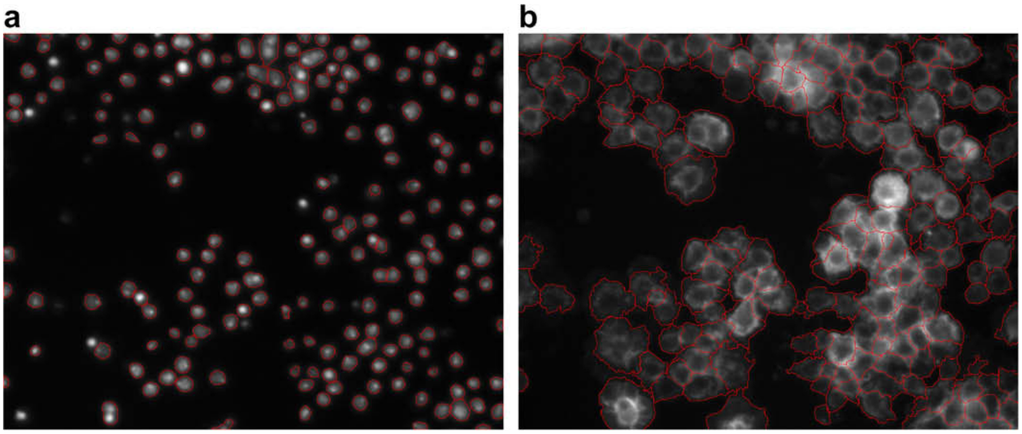

Fig. 3. Fluorescence image segmentation results by seeded two-step enhanced watershed-based method. (a) the extracted nuclei from DNA channel; (b) cell body segmentation on Actin channel.

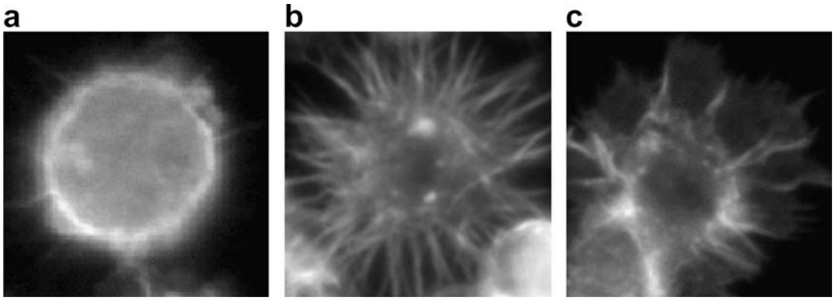

Fig. 4. Three cellular phenotypes of *Drosophila* Kc_{167} cells: (a) Normal; (b) Spiky; (c) Ruffling.

particular dsRNA alter the phenotype? In other words, is it a Screen Hit? To answer this question, we extracted various parameters from the entire NSR-labeled fluorescence image to generate statistical properties that describe the overall phenotype. For each cellular phenotype we extracted three kinds of statistical properties related to cell number, cell area, and cell perimeter. The ratios of each cellular phenotype, especially the Rac^{V12}-induced phenotypes, S and R, were key factors in evaluating the image's phenotype. To achieve a stable description for a given treatment, images from three independent sites for each treatment were analyzed; we refer to these as 'screened sites'.

The first statistical property was the ratio of different phenotypes. We obtained the number of segmented cells with each phenotype as $\mathbf{n} = \{n_i^N, n_i^S, n_i^R\}, i = 1, 2, \ldots, K$, where i is the index of the screened sites and K is the number of screened sites. We used the average ratios of the three cellular phenotypes to represent image characteristics. Hence, we transformed the features from $\mathbf{n} = \{n_N^i, n_S^i, n_R^i\}$ to $\mathbf{r}_{\text{num}} = \{r_{\text{num}}^N, r_{\text{num}}^S, r_{\text{num}}^R\}$, where the ratios were calculated as: $r_{\text{num}}^N = \frac{1}{k}\sum_i^K \frac{n_i^N}{n_i}$, $r_{\text{num}}^S = \frac{1}{k}\sum_i^K \frac{n_i^S}{n_i}$, $r_{\text{num}}^R = \frac{1}{k}\sum_i^K \frac{n_i^R}{n_i}$, where $n_i = n_i^N + n_i^S + n_i^R$.

The second statistical property was based on the area of each segmented cell. The ratios of areas of each cellular phenotype in the image were obtained and represented as $\mathbf{r}_{\text{area}} = \{r_{\text{area}}^N, r_{\text{area}}^S, r_{\text{area}}^R\}$.

The third statistical screening property was the ratio of the perimeters of each cellular phenotype, which is based on the sum of the perimeters of all cells of a particular phenotype within an image. The ratio of cell perimeters was $\mathbf{r}_{\text{pre}} = \{r_{\text{pre}}^N, r_{\text{pre}}^S, r_{\text{pre}}^R\}$.

Finally, these three statistical properties were also calculated using the average value of the three screened sites corresponding to a single treatment to achieve reliable and stable results. Thus, we obtained the overall statistical description of an HCS image's phenotype as: $\mathbf{x} = \{\mathbf{r}_{\text{num}}, \mathbf{r}_{\text{area}}, \mathbf{r}_{\text{pre}}\}$.

4. Fuzzy image scoring regression model

After computing the statistical properties for the image were computed, the task was to model the relationship between the phenotype property and the image score. Once the model was estimated, we can predict the score of test images. We derived the statistical properties-based scoring system as following:

$$F : \mathbf{x} = \{\mathbf{r}_{\text{num}}, \mathbf{r}_{\text{area}}, \mathbf{r}_{\text{pre}}\} \to y \tag{2}$$

where F is the prediction function and y is the image score. Although there are many models to mathematically describe the relationship between variables and their response values, the fuzzy theory was a better way to describe such an image scoring problem because of its intrinsic flexibility. The fuzzy system handles problems with imprecise and incomplete data and models non-linear functions of arbitrary complexity. The ground truth score for the training data was determined manually; therefore, the score was only an approximation rather than an exact, true value. Moreover, the manual scoring rules varied depending on complex parameters, such as cell density, and thus the prediction functions of the model had to be flexible.

Assume the image has the phenotype distribution descriptor as $\mathbf{x} = \{\mathbf{r}_{\text{num}}, \mathbf{r}_{\text{area}}, \mathbf{r}_{\text{pre}}\}$ and there exists C image classes, which are corresponding to C types of genes $\{g_1, g_2, \ldots, g_C\}$. The fuzzy model-based scoring system consists of two steps. The first step was to estimate the fuzzy membership function of the test image $\mathbf{x}$ belonging to different image class $g_i, i = 1,2, \ldots ,C$. We had two strategies to estimate the fuzzy membership value $p(g_i$—$\mathbf{x})$. One was applying fuzzy logic rules to determine the value. The other was representing the value by the posterior probability using standard Bayesian theorem $p(g_i|\mathbf{x}) = \frac{p(\mathbf{x}|g_i)}{\sum_i^C} p(\mathbf{x}|g_i)$. The second step of the scoring model was to build a single regression model for each image class as $y^i = f^i(\mathbf{x})$. Therefore, the scoring model has the formula:

$$y = F(\mathbf{x}) = \sum_{i=1}^{C} p(g_i|\mathbf{x})y^i = \sum_{i=1}^{C} p(g_i|\mathbf{x})f^i(\mathbf{x}) \tag{3}$$

4.1. Fuzzy membership estimation

The images used in our experiments were manually classified into three categories under the experimental hypothesis: negative controls (NC), positive controls (PC), and screen hits (SH). As described in the former sections, a gene's function can be assessed by analyzing alterations in a biological process caused by the absence of that gene. If disruption of a specific gene results in differences in the statistical description of the corresponding images, it is considered a screen hit. For example, PC images were distinctive because there existed few or no spiky or ruffling cells in the images, and the manual scores were identical. The phenotype distribution descriptor clearly distinguished this group of images, and it was fairly straightforward to apply fuzzy logic rules to derive the score and membership value. The linguistic terms characterized by the ratios of spiky and ruffling-phenotypes as:

$$\text{if } r^N_{\text{num}} \approx 0 \text{ and } r^R_{\text{num}} \approx 0 \textit{ then}:$$

$$P(g_i \mid \mathbf{x}) = \begin{cases} 1 & g_i = \text{positive-controls} \\ 0 & g_i \neq \text{positive-controls} \end{cases} \tag{4}$$

$$y = \frac{1}{l_{\text{pc}}} \sum_{k=1}^{l_{\text{pc}}} e^{-\|\mathbf{x}-\mathbf{x}_k\|} y_k$$

where l_{pc} is the number of PC images in the training set.

On the contrary, the distinction between NC and SH images was less clear. In some cases there existed a big discrepancy between the manual scoring and the phenotype statistical representation. This was especially concerning in cases where images of dsRNAs targeting genes known to cause phenotypes were identified by manual scoring but not by the automated scoring models. To address this discrepancy and assign scores that more accurately reflect the manually identified phenotypes, we modeled the variation and similarity of the measured statistical properties using Gaussian mixture models (GMMs) that treat each gene class as a Gaussian distribution with parameters in the feature space. The image training data set was modeled as a sampling from a mixture probabilistic model. For the training data set, images with enough confidence in both the biological and informatics domains were manually fixed with a certain membership value while others held floating fuzzy membership values. The floating fuzzy membership values were interactively approximated using the Expectation–Maximization (EM) algorithm [31,32]. Assume the cellular features are represented as $\mathbf{x}$ (Eq. (3)), which is sampled from a mixture of Gaussian processes $\mathcal{N} = \{\mathcal{N}_1, \cdots, \mathcal{N}_C\}$, where C corresponds to the number of phenotypes ($C = 3$ in this paper). Each Gaussian process can be described by the prior probability p, mean vector μ and covariance matrix Σ as: $\mathcal{N}_i = \{p_i; \mu_i, \Sigma_i\}$. The parameter optimization approach consisted of two steps:

E-step:

$$p(g_i|\mathbf{x}, \Theta_i^t) = \frac{p(\mathbf{x}|g_i, \Theta_i^t)p(g_i|\Theta_i^t)}{p(\mathbf{x}|\Theta_i^t)} = \frac{p(\mathbf{x}|p_i^t, \mu_i^t, \Sigma_i^t)p_i^t(g_i)}{\sum_{j=1}^{C} p(\mathbf{x}|p_i^t, \mu_i^t, \Sigma_i^t)p_i^t(g_i)} \tag{5}$$

M-step:

$$\begin{aligned} p_i^{t+1} &= \frac{\sum_k p(\mathbf{x}_k|g_i, \Theta^t)}{K} \\ \mu_i^{t+1} &= \frac{\sum_k p(\mathbf{x}_k|g_i, \Theta^t)\mathbf{x}_k}{\sum_k p(\mathbf{x}_k|g_i, \Theta^t)} \\ \Sigma_i^{t+1} &= \frac{p(\mathbf{x}_k|g_i, \Theta^t)(\mathbf{x}_k - \mu_i^{t+1})(\mathbf{x}_k - \mu_i^{t+1})^T}{p(\mathbf{x}_k|g_i, \Theta^t)} \end{aligned} \tag{6}$$

where t denotes the iteration step. After the iteration convergence, the final fuzzy membership value was estimated with the optimal parameters as $p^{\star}(g_i\text{—}\mathbf{x},\Theta^{\star})$.

4.2. Single scoring model by support vector regression

For those images that were labeled with both biological and informatics confidence, we used support vector regression to build a single score prediction model for each image class. The conventional empirical risk minimization (ERM) based regression models have the drawback of the "over-fitting problem", such as overlearning in neural network design. Support vector machine technique embodies the structure risk minimization (SRM) principle to ERM to formulate the optimization objective function, which has better generalization ability [33]. Moreover, support vector regression (SVR) was appropriate to the image scoring model-based on a small sample set.

Suppose the manually scored training set has the samples as $(\mathbf{x}_1^i, y_1^i), (\mathbf{x}_2^i, y_2^i), \ldots, (\mathbf{x}_l^i, y_l^i)$, where $\mathbf{x}^i \in \mathscr{R}_X^i$ and $y^i \in \mathscr{R}_Y^i$ are the input variables of image class and the target image score, respectively. The regression function is then:

$$y^i = f^i(\mathbf{x}) = \sum_{k=1}^{l_i} \left(a_k^{i^{\star}} - a_k^i\right) \cdot k(\mathbf{x}_k^i, \mathbf{x}^i) + b^i \tag{7}$$

where $a^{i^{\star}}, a_k^i$ are Lagrange multipliers, which can be obtained by solving the optimization problem:

$$\max_{a_i^{i*}, a_i^i} \left[\begin{aligned} & -\tfrac{1}{2} \sum_{s,t=1}^{l_i} (a_s^i - a_s^{i*})(a_t^i - a_t^{i*}) < \mathbf{x}_s^i - \mathbf{x}_t^i > \\ & -\varepsilon \sum_{i=1}^{n} (a_s^i + a_s^{i*}) + \varepsilon \sum_{t=1}^{l} y_t^i (a_t^i - a_t^{i*}) \end{aligned} \right] \tag{8}$$

$$s.t. \sum_{s=1}^{l_i} (a_s^i - a_s^{i*}) \text{ and } a_s^i, a_s^{i*} \in [0, \xi]$$

ζ is a constant and the value of ε is defined for the so-called insensitive loss function. The computing of b^i in Eq. (7) can be exploited by applying the Karush–Kuhn–Tucker conditions [34]. We used the Gaussian radial basis function as the kernel function, which is defined as:

$$k(\mathbf{x}, \mathbf{y}) = \mathrm{e}^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{2\delta^2}} \tag{9}$$

with the parameters set as: $\varepsilon = 0.01, \sigma = 0.8$.

4.3. Mixture fuzzy scoring model

Finally, we calculated an approximate image score by combining the fuzzy membership estimation and the SVR single scoring models as:

$$\begin{aligned} y = F(\mathbf{x}) &= \sum_{i=1}^{C} p(g_i|\mathbf{x})y^i = \sum_{i=1}^{C} p(g_i|\mathbf{x}, \Theta)f^i(\mathbf{x}) \\ &= \sum_{i=1}^{C} \left(p(g_i|\mathbf{x}, \Theta) \cdot \sum_{k=1}^{l_i} \left(a_k^{i^*} - a_k^i\right) k(\mathbf{x}_k^i, \mathbf{x}^i) + b^i \right) \end{aligned} \tag{10}$$

There are 2-fold merits of the mixture scoring model. First, it is flexibly extended, which allows the model to adapt easily to new image classes. Second, the model can be incrementally updated. If the new image data are fed into the system, the estimated fuzzy membership will be refined to match the new training data.

5. Experiments

5.1. Materials

As described in Section 2, the original fluorescence cellular images are in the scale of 1280 × 1024 and stored in 12-bit format. For each image, usually there are over 100 cells in the scope of the screen. We first built a cell database based on the segmentation results on the fluorescence images and manually classify the selected cell patches into the three predefined categories of phenotypes. Because of the diversity and complexity of the genome-wide RNAi screening image data, the current cellular segmentation method cannot generate state-of-art results, especially in the case that many cells touch together or even overlap each other. Therefore, when we collect the segmented data for building training set, those ill-segmented cellular patches are ignored. Finally, we have a cellular phenotype image database, including around 600 normal-phenotype cells, 200 spiky-phenotype cells and 200 ruffling-phenotype cells. Fig. 5 illustrates some examples of the cellular patches in our cellular phenotype database. This cell database is used to train and evaluate the phenotype classification model.

Moreover, we built a database to test our high-content image scoring model. The database contains three images each for 89 different treatments/dsRNA, including 54 NC, 6 PC and 35 SH, which were manually scored, ranging from 0.3 to 5.0. The fluorescence screening image database consists of 2 gigabytes, with more than 800 images from 3 channels. Cells (40,221) were identified from all the images. Following the cellular phenotype classification, the statistical description of the cellular phenotypes was obtained for each image of every treatment. There were distinctive statistical variations for these fluorescence screenings with different treatments. A simple statistical result for the different image types is summarized in Table 1.

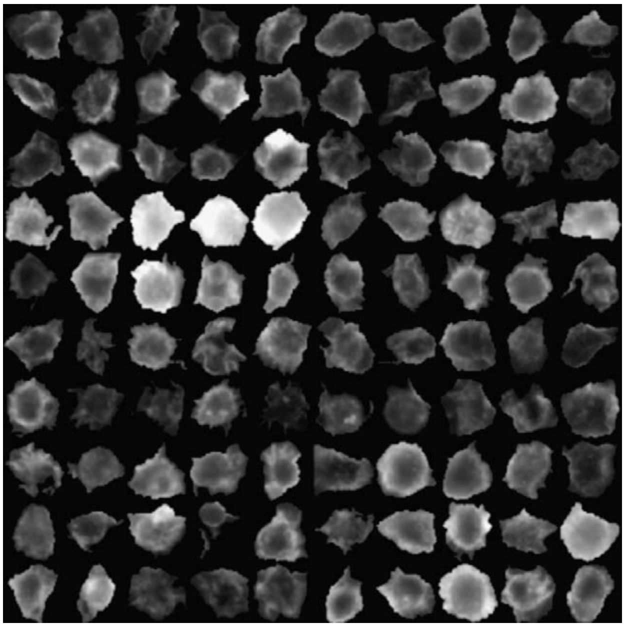

Fig. 5. Hundred cellular phenotype samples from the database.

Table 1
Statistical variations for the three image types: negative control (NC), positive control (PC), and screen hits (SH)

Image type	$\bar{n}_N$	$\bar{n}_S$	$\bar{n}_R$	$\bar{y}$	σ_y
NC	153.3	15.9	5.8	1.37	0.40
SH	98.0	11.9	2.2	3.57	0.65
PC	172.4	5.3	0.1	5.00	0.00

$\bar{n}_N, \bar{n}_S, \bar{n}_R$ represent the average number of Normal, Spiky and Ruffling-phenotypes, respectively. The value $\bar{y}$ and σ_y is the mean and standard deviation of manual score for each image class.

5.2. Results

First, in order to train and test our phenotype classification procedure, we conduct the 10-folder leave-one-out cross validation strategy on the cell database to derive the generalization classification error [35]. The experiments were repeated 100 times with random folder partition. Using 12, 15, and 18 features selected by GA, LDA classifier achieved the average performance 74.26%, 76.08% and 74.67%, respectively. Moreover, with the same experimental setting, SVC obtained the average performance as 67.10%, 69.73%, and 65.17%. Considering the complexity and diversity of cellular phenotypes in the high-content fluorescence screening, most cells contain multiple-phenotype characteristics. Strictly speaking, it is not reasonable to set hard and exclusive labels to cells (only belongs to a certain phenotype and not belong to any other phenotypes). However, to be realistic, it is not feasible to assign a ground truth training set in cell level with soft labels (showing the likelihoods of the cells belongs to different phenotypes). Therefore, we impose the fuzzy information on the screening level to reduce the disadvantage of hard labeling. The image level statistics provide more flexible description than the cell level hard labels in terms of image score prediction.

Moreover, we applied the trained phenotype classifiers on whole fluorescence images of *Drosophila* cells, which were captured for a HCS-based gene functional analysis. After cell level processing and analysis, the cellular segmentation (Fig. 3) and phenotype identification results can provide abundant quantitative information for investigating the function of the related genes. An example of the process is shown in Fig. 6, indicating the results of cellular phenotype classification. Notice that some cells are located near the boundary of the screen without whole cell body extracted. We may ignore those cells during the phenotype identification stage because the broken shape will generate inconvincible prediction.

Second, we use the image database described in Section 5.1 to evaluate the image score prediction model. In the experiments, 80% of the images were used to train the regression model. The remaining 20% was used to test the model. Two different evaluation criteria were used to validate the results. The first criterion was the mean square loss on test data, which was calculated as: $E = \sum_{i=1}^{N}(y_i - F(\mathbf{x}_i))^2$, where $F(\mathbf{x}_i)$ is the predicted score and y_i is target score. The other criterion was the coefficient of determination (COD), which can be computed as:

$$COD = \frac{\sum_{i=1}^{m}(F(\mathbf{x}_i) - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \qquad (11)$$

where m is the number of samples in test set and $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ is the ground mean value. Tables 2 and 3 show the experimental results of the score prediction on the NC, PC, and SH images. Generally speaking, the automatically predicted scores achieve a certain accuracy and consistence comparing to the manually expert scores. Specially, the negative control image samples has better prediction performance in terms of mean square loss (around 0.021), while positive control and screen hits samples have a slightly higher COD value (around 0.713).

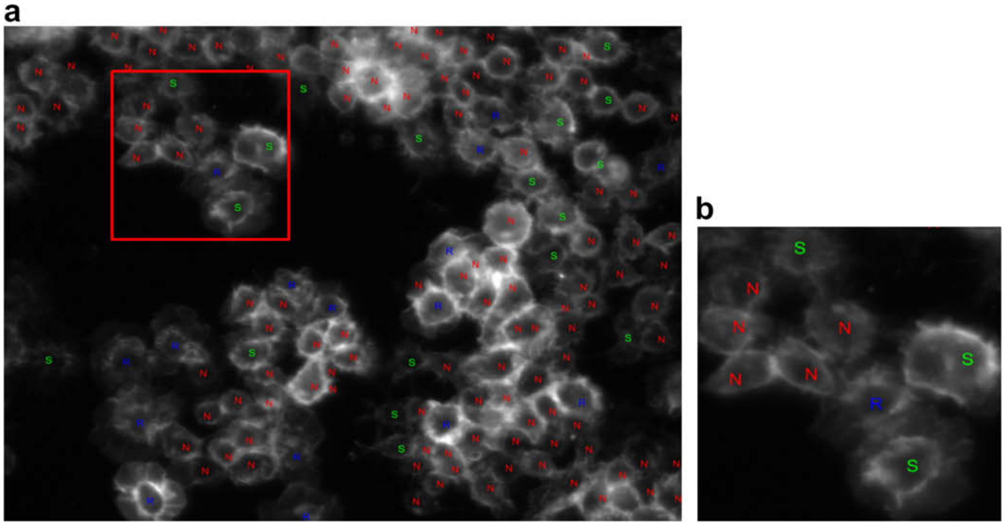

Fig. 6. Fluorescence image analysis of *Drosophila Kc*$_{167}$ cells: (a) Cellular phenotype classification results on the high-content fluorescence screening. The markers N, S and R represent Normal, Spiky and Ruffling-phenotype, respectively. (b) Zoom-in view of the rectangular area of (a).

Table 2
Scores of NC images: The first row is the location of the well within the 384-well plate

Well	L05			L01			G15			N22			O16			L03			K16			M20			P17			A17		
# Normal	145	251	117	182	239	204	111	97	120	135	200	102	134	198	206	192	172	149	134	123	163	174	113	155	193	212	128	67	89	107
# Spiky	16	19	18	9	6	10	12	16	21	18	15	6	15	13	9	21	34	25	6	10	22	26	21	9	15	13	26	23	22	36
# Ruffling	10	0	5	8	4	3	4	6	4	7	3	2	3	3	2	5	5	6	3	7	3	4	10	6	6	6	6	12	8	4
Manual score	1.3			1.2			1.3			1.3			1.3			1.3			1.0			1.3			1.5			1.7		
Auto-score	1.41			1.18			1.37			1.40			1.32			1.43			1.39			1.29			1.39			1.70		

The second to fourth rows give the number of phenotypes in each of three scanned sites for that well.
Images were scored 0.0–5.0; the higher the score, the more likely that dsRNA is a Hit.
Comparison of the manual score (fifth row) and the automatically predicted score (sixth row) reveals a mean square error is 0.021 and a coefficient of determination is 0.6910.

Table 3
Scores of PC and SH images: The first row is the location of the well within the 384-well plate

Well	I18			P09			N01			C02			I02			C21			H04			O22		
# Normal	26	26	38	69	97	118	42	55	33	18	65	37	180	173	203	101	104	151	73	118	109	280	205	162
# Spiky	0	1	2	13	8	11	10	7	7	6	4	3	29	19	9	28	23	28	15	13	13	1	1	0
# Ruffling	0	0	0	1	2	1	1	0	2	1	0	0	2	0	4	6	3	1	3	2	5	0	0	0
Manual score	3.3			3.7			3.7			3.3			3.2			3.0			3.7			5.0		
Auto-score	3.20			3.52			3.43			3.69			3.65			3.80			3.25			4.85		

The second to fourth rows give the number of phenotypes in each of three scanned sites for that well.
Images were scored 0.0–5.0; the higher the score, the more likely that dsRNA is a Hit.
Comparison of the manual score (fifth row) and the automatically predicted score (sixth row) reveals a mean square error is 0.1669 and a coefficient of determination is 0.7131.

6. Discussion and conclusion

Genome-wide RNAi HCS can provide critical visual information to elucidate the underlying mechanisms of complex cellular processes, such as cell division and cell migration. However, manual analysis of high-content image data sets is a tremendous hurdle. Here, we report an informatics processing system, containing phenotype identification and image scoring models, to quickly process a huge volume of high-content images and conduct quantitative analysis of the visual data.

Our automated image scoring system has two key components. The first part is fluorescence image analysis, including cellular segmentation, feature extraction and cellular phenotype recognition, and statistical description of the image. The second part of the system is the image score prediction model, which incorporates the fuzzy logical concepts and EM algorithm to conduct gene fuzzy membership approximation. Support vector machine technique is applied to derive the single scoring model for each image class with confident samples in both biological and informatics domains. Using the fuzzy membership values, a fuzzy mixture model automatically predicted the image score. In other word, a series of relatively simple regression models are combined with certain weights to achieve the structural simplicity and generalization of prediction.

We built the database of high-content, 3-channel, fluorescence microscopy images of *Drosophila Kc*$_{167}$ cultured cells that were treated with RNAi to perturb gene function. Images were analyzed for alterations in cellular phenotypes, suggestive of a role in Rac signaling. The performance of the automated phenotype classification was evaluated on the constructed cell database. Forevermore, the proposed image scoring system generated scores that were similar to manual annotation. The robustness and efficiency of our system were validated in quantitatively predicting the biological relevance of genes.

Finally, we remain concerned about how to generalize the proposed approach to process RNAi HCS images from other biological studies. For the biological study presented in this paper, we kept the biologist incorporating domain knowledge, such as predefining phenotypes. However, in order to achieve the general utility, there are two key problems need to be attacked. The first issue lies in cell level, which aims to build a semi-supervised learning model to efficiently obtain the phenotype samples for training to replace the current manually labeling procedure, which is costly prohibited in large-scale study. Moreover, a more ambitious target is to develop a new machine learning technique to automatically discover novel phenotypes in different HCS image dataset without exhausted expert interaction. Second, we need to refine an adaptive image-level feature extraction and selection technique, which can choose the distinctive image descriptor for evaluating the gene effects in diverse biological studies. In the current approach, the feature selection is directly linked to the phenotype classification instead of high level image scoring. Both of these two open problems are the interest of our future work.

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, for their technical comments. The research is funded by the HCNR, Harvard Medical School (Wong).

References

[1] 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;6:279–86.

[2] 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;17:1213–23.

[3] Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschule SJ. Multidimensional drug profiling by automated microscopy. Science 2004;306:1194–8.

[4] Kiger AA, Baum B, Jones J, Jones MR, Coulson A, Echeverri C, Perrimon N. A functional genomic analysis of cell morphology using rna interference. J Biol 2:27.

[5] Zhou X, Wong STC. Informatics challenges of high-throughput microscopy. IEEE Signal Process Mag 2006;5:63–72.

[6] Carpenter A, Jones T, Lamprecht M, Clarke C, Kang I, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 2006;7(10):R100.

[7] Wang J, Zhou X, Bradley PL, Chang S-F, Perrimon N, Wong STC. Cellular phenotype recognition for high-content RNAi genome-wide screening. J Biomol Screen 2008;13(1):29–39.

[8] Settleman J. Rac'n rho: the music that shapes a developing embryo. Dev Cell 2001;1:321–31.

[9] Yamazaki D, Kurisu S, Takenawa T. Regulation of cancer cell motility through actin reorganization. Cancer Sci 2005;96:379–86.

[10] Bishop AL, Hall A. Rho gtpases and their effector proteins. Biochem J 2000;2:241–55.

[11] Burridge K, Wennerberg K. Rho and rac take center stage. Cell 2004;116:167–79.

[12] Echeverri CJ, Perrimon N. High-throughput rnai screening in cultured cells: a user's guide. Nat Rev Genet 2006;7:373–84.

[13] Pham TD, Crane DI, Tran TH, Nguyen TH. Extraction of fluorescent cell puncta by adaptive fuzzy segmentation. Bioinformatics 2004;20: 2189–96.

[14] Duncan JS, Ayache N. Medical image analysis: progress over two decades and the challenges ahead. IEEE Trans Pattern Anal Mach Intell 2000;22: 85–106.

[15] Zhou X, Liu KY, Bradley P, Perrimon N, Wong STC. Towards automated cellular image segmentation for rnai genome-wide screening. Lecture Notes in Comp Sci (MICCAI 2005) 2005;3749:885–92.

[16] Xiong G, Zhou X, Ji L, Bradley P, Perrimon N, Wong STC. Automated segmentation of *Drosophila* rnai fluorescence cellular images using deformable models. IEEE Trans Circuit Syst 2006;53:2415–24.

[17] Li FH, Zhou X, 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;226:121–32.

[18] Ridler TW, Calvard S. Picture thresholding using an interactive selection method. IEEE Trans Syst, Man, Cybernet 1978;8:1264–91.

[19] Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 1996;18:837–42.

[20] Cohen A, Daubechies I, Feauveau JC. Bi-orthogonal bases of compactly supported wavelets. Pure Appl Math 1992;45:485–560.

[21] Bovic AC, Clark M, Geisler WS. Multichannel texture analysis using localized spatial filters. IEEE Trans Pattern Anal Mach Intell 1990;12:55–73.

[22] Daugman JG. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. IEEE Trans Acoustics, Speech, Signal Process 1988;36:1169–79.

[23] Zernike F. Beugungstheorie des schneidencerfarhens undseiner verbesserten form, der phasenkontrastmethode. Physica 1934;1:689–704.

[24] Teague MR. Image analysis via the general theory of moments. Opt Soc Am, J 1980;70:920–30.

[25] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Sys, Man Cybernet 1973;6:610–20.

[26] Haralick RM. Statistical and structural approaches to texture. Proc IEEE 1979;67:786–804.

[27] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1(3):131–56.

[28] Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology. Control and artificial intelligence. Cambridge, MA: MIT Press; 1996.

[29] Duda RO, Hart PE, Stork DH. Pattern classification. 2nd ed. New Haven: Wiley Interscience; 2000.

[30] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discov 1998;2:121–67.

[31] Dempster AP, Laird NM, Rubin DB. Maximum-likelihood from incomplete data via the em algorithm. J Royal Statist Soc Ser B 1977;39:1–38.

[32] Bilmes JA. A gentle tutorial of the em algorithm and its applications to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, International Computer Science Institute, Berkeley, California, 1998.

[33] Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.

[34] Kuhn HW, Tucker AW. Nonlinear programming. Proc 2nd Berkeley Symp Math Stat Probabilist 1951:481–92.

[35] Bishop C. Neural networks for pattern recognition. Oxford: Clarendon Press; 1995.