Journal of Microscopy, Vol. 238, Pt 2 2010, pp. 145–161 doi: 10.1111/j.1365-2818.2009.03337.x
Received 26 January 2009; accepted 9 September 2009

Image analysis benchmarking methods for high-content screen design

C.J. FULLER & A.F. STRAIGHT
Department of Biochemistry, Stanford Medical School, 279 Campus Drive, Beckman 409, Stanford, CA

Key words. Benchmarking, high-content screening, image analysis.

Summary

The recent development of complex chemical and small interfering RNA (siRNA) collections has enabled large-scale cell-based phenotypic screening. High-content and high-throughput imaging are widely used methods to record phenotypic data after chemical and small interfering RNA treatment, and numerous image processing and analysis methods have been used to quantify these phenotypes. Currently, there are no standardized methods for evaluating the effectiveness of new and existing image processing and analysis tools for an arbitrary screening problem. We generated a series of benchmarking images that represent commonly encountered variation in high-throughput screening data and used these image standards to evaluate the robustness of five different image analysis methods to changes in signal-to-noise ratio, focal plane, cell density and phenotype strength. The analysis methods that were most reliable, in the presence of experimental variation, required few cells to accurately distinguish phenotypic changes between control and experimental data sets. We conclude that by applying these simple benchmarking principles an *a priori* estimate of the image acquisition requirements for phenotypic analysis can be made before initiating an image-based screen. Application of this benchmarking methodology provides a mechanism to significantly reduce data acquisition and analysis burdens and to improve data quality and information content.

Introduction

The sequencing of multiple eukaryotic genomes coupled with the discovery of RNA-mediated inhibition (RNAi) has enabled genome-wide screens for gene depletion phenotypes in metazoan cells and organisms. Libraries of small molecules, traditionally restricted to pharmaceutical drug screening, are now available to a wide range of researchers for phenotype-based screening of chemical compounds. These inhibitor collections have led to numerous successful high-throughput screens for protein and pathway inhibition in metazoans, and high-throughput imaging has become a central method for recording and assessing biological phenotypes in these screens (Perlman *et al.*, 2004; Agaisse *et al.*, 2005; Erhardt *et al.*, 2008). Microscopic imaging can quantitatively capture depletion, localization and morphological changes in cells, but manual analysis of large high-throughput image data sets is impractical and thus requires the development of automated tools to extract, classify and measure the relevant information in the images.

Automated image analysis involves segmenting the image to identify the relevant objects for analysis followed by quantification or classification of the segmented objects. Numerous image segmentation methods have been developed and applied to problems in computer vision, medical imaging and biological imaging, but no one method is optimal for all image segmentation problems. Image data sets for development and benchmarking of segmentation techniques have been developed for many different image types, but there exists no generally applicable data set for all object identification tasks (Martin *et al.*, 2001; Li *et al.*, 2007; Shamir *et al.*, 2008) reviewed in (Peng, 2008).

Analysis methods for biological image data sets from high-throughput screens have ranged from the most laborious manual inspection of the images to automated computational methods to rapidly screen through large image data sets (Mayer *et al.*, 1999; Gonczy *et al.*, 2000; Kiger *et al.*, 2002; Yarrow *et al.*, 2003; Neumann *et al.*, 2006). Computational tools and software suites have been developed to facilitate identification of cells and subcellular features in high-throughput image data sets (Goldberg *et al.*, 2005; Baatz *et al.*, 2006; Carpenter *et al.*, 2006; Glory & Murphy, 2007; Orlov *et al.*, 2008). To date, the workflow for most image-based high-throughput screens involves collecting the screening data and empirically sampling and adjusting different image analysis methods until the desired information set is extracted

Correspondence to: Aaron F. Straight. Tel: +650-723-2941; fax: +650-723-6783; e-mail: astraigh@stanford.edu

from the screen data. This process is usually undertaken after the experimental phase of the screen is complete, and therefore little information about the analysis method feeds back into the data collection.

Here we propose and implement a new screening methodology that uses statistical analysis of benchmarking experiments preceding a screen to quantitatively assess the performance of image analysis. We then propose the use of those analysis method assessments to guide the creation of the entire screen. Specifically, we generate benchmark image data sets that are tailored to our desired screens and that recapitulate common variation encountered in high-throughput screening (e.g. variation in signal-to-noise, phenotype strength, focal plane and cell density). We quantitatively compare different image segmentation and analysis routines on our benchmark data sets to refine the image analysis methods. We show that the best analysis method of a group can be quantitatively determined, and the amount of data required to detect a particular phenotype to a desired significance level can be accurately predicted before performing the entire screen. The methods we present should allow significant savings in acquisition and analysis costs and facilitate the collection of higher quality data with improved information content.

Materials and methods

Experimental

Cell culture. Drosophila melanogaster Kc167 cells were grown in Schneider's media (Gibco), supplemented with 100 U/mL penicillin, 100 μg/mL streptomycin (P/S) and 10% foetal bovine serum (FBS) and passaged by trypsin/EDTA treatment and manual scraping. For RNAi experiments, cells were washed in unsupplemented Schneider's media and resuspended to a concentration of 2.5×10^6 cells/mL. Two hundred microlitres of the cell suspension was added to a sterile plastic tube, and dsRNA was added from a stock in sterile water to a final concentration of 100 nM for standard experiments or lower to generate intermediate phenotypes. Control cells were treated with an equivalent volume of sterile water. Cells were incubated for 45 min at room temperature and then plated in one well per treatment of a 24-well culture dish. Schneider's media, supplemented with P/S and FBS, was added to each well to give a final volume of 1 mL. Cells were incubated in the dark at room temperature for 16–72 h as indicated.

Human HeLa cells were grown in Dulbecco's modified Eagle's medium (Invitrogen), supplemented with 100 U/mL penicillin, 100 μg/mL streptomycin (P/S) and 10% FBS and passaged by trypsin/EDTA treatment.

For drug treatment experiments, HeLa cells were treated for 8 h with nocodazole (Sigma) to a final concentration of 0.5 μg/mL in Dulbecco's modified Eagle's medium with P/S and FBS, after which mitotic cells were shaken off by striking the plate of cells, and replated onto poly(L-lysine)-coated cover slips in serum-free Dulbecco's modified Eagle's medium with P/S. For control populations, cells were passaged directly onto cover slips.

Immunofluorescence and microscopy (Kc167 cells). Cells growing in 24-well dishes were harvested, after aspiration of the growth media, by treatment with 0.5 mL trypsin/EDTA for 5 min, addition of 1 mL Schneider's with P/S and FBS, and vigorous pipetting with a 2 mL plastic pipet. Wells were washed with an additional 1 mL Schneider's with P/S and FBS, and this wash was pooled with the original. Cells were collected by centrifugation and resuspended in 100 μL Schneider's medium with P/S and FBS then pipetted over a 16-fold range of dilutions onto acid-washed, poly-L-lysine-coated glass 12-mm cover slips and allowed to adhere for 1 h in a dark, humidified chamber. Cells on cover slips were fixed for 3 min in room-temperature methanol. Cover slips were rehydrated in 20 mM Tris–HCl pH 7.4, 150 mM NaCl with 0.1% Triton X-100 (TBST) and blocked for 30 min in antibody dilution buffer (AbDil, TBST with 2% bovine serum albumin and 0.1% sodium azide). Mouse anti-heterochromatin protein 1 (HP1) antibodies (Developmental Studies Hybridoma Bank, clone C1A9) were diluted 1:500 in AbDil and incubated on each cover slip for 1 h at room temperature. Cover slips were washed in AbDil and incubated for 1 h at room temperature with a 1:500 dilution of Alexa488-conjugated goat anti-mouse IgG (Invitrogen) in AbDil. Cells were stained with 10 μg/mL Hoechst in AbDil, washed in TBST and mounted on slides in 20 mM Tris pH 8.8, 0.5% *p*-phenylenediamine and 90% glycerol.

Cells were imaged using a 40× 0.9 NA objective using a Nikon Eclipse 80i microscope, and images were acquired with a Princeton Instruments Coolsnap HQ 12-bit CCD camera. Images were axially adjusted during acquisition to correct for chromatic aberration according to predetermined calibration standards created using 0.5-μm Tetraspeck fluorescent microspheres (Molecular Probes). Exposure time was set so that control cells brought the camera to three-fourths of its saturation value. The benchmarking set of images was taken by acquiring 25 Z-sections of each cover slip at 0.4 μm offset (corresponding to −4.8 to +4.8 μm) with 1/512, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4 or 0 neutral density filters in the excitation light path, or with double the exposure time and no neutral density filters in place.

Immunofluorescence and microscopy (HeLa cells). Rabbit polyclonal antibodies against human HP1α were generated using an N-terminal GST fusion to amino acids 66–119 of human HP1α. This fragment was chosen as the region of the protein least conserved among the three human HP1 variants. The GST-tagged protein fragment was purified using glutathione agarose for antibody generation. Rabbit serum was purified over an affinity column of the same HP1α peptide (untagged) used to generate the antibody.

Cells on 12-mm glass cover slips were fixed for 5 min in 150 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4, 0.1% Triton X-100 (PBST) with formaldehyde added to 3.7%. Cover slips were washed in TBST, blocked in AbDil for 30 min and stained with a solution in AbDil of mouse anti-αTubulin (1:500) (Sigma, mouse clone DM1α), rabbit anti-HP1α (1:500) and Atto488-conjugated phalloidin (13.3 U/mL) (Sigma), for 1 h at room temperature. Cover slips were washed in AbDil and incubated for 1 h at room temperature in a solution of Alexa647-conjugated goat anti-mouse IgG (1:500) (Invitrogen) and Alexa568-conjugated goat anti-rabbit IgG (1:500) (Invitrogen) in AbDil. Cells were stained with 10 μg/mL Hoechst in AbDil, washed in TBST and mounted on slides in 20 mM Tris pH 8.8, 0.5% *p*-phenylenediamine and 90% glycerol.

Cells were imaged using the same equipment and calibration as for the *Drosophila* cells. Exposure time was set so that control cells brought the camera to three-fourths of its saturation value, except in the Alexa647 channel, where this resulted in a prohibitively long exposure time, and the exposure was set to bring the camera to one-fourth of its saturation value. The benchmarking set of images was taken by acquiring 21 Z-sections of each cover slip at 0.8 μm offset (corresponding to −8.0 to +8.0 μm) with 1/512, 1/64, 1/16, 1/4 or 0 neutral density filters in the excitation light path, or with double the exposure time and no neutral density filters in place.

Preparation of RNAi constructs. A fragment of the HP1 gene was PCR amplified from plasmid ASP508 using primers forward: tggcgcccctagatgCCCTCTGGCAATAAATCAAAA, and reverse: cgacgcccgctgataTTAATCTTCATTATCAGAGTACCA, and purified by agarose gel electrophoresis and extraction (Qiagen). T7 promoters were added to each end by a second PCR step using primers forward: GCGTAATACGACTCACTATAGGGtggcgcccctagatg, and reverse: GCGTAATACGACTCACTATAGGGcgacgcccgctgata, and purified as before. The purified template was *in vitro* transcribed using T7 polymerase at 0.4 μg/mL and yeast inorganic phosphatase at 0.015 U/μL (Sigma) in 40 mM Tris–HCl pH 8.0, 10 mM DTT, 20 mM spermidine, 20 mM $MgCl_2$ and 7.5 mM (each) NTPs, and incubated for 5 h at 37°C. Subsequently, the reaction was treated with 3 μL RNAse-free DNAse at 2 U/μL (Ambion) and incubated for 30 min at 37°C. The RNA was then purified using an RNeasy kit (Qiagen) and eluted into RNAse-free sterile water at a final concentration of 1.1 mg/mL.

Computational

Algorithms for object detection. For the *Drosophila* data set, five methods for object detection based upon the literature were implemented using C++. Source code for all programs used is available online at http://straightlab.stanford.edu/software under the Mozilla Public License. Method 1, referred to as the recursive thresholding method, was based upon the recursive Otsu method described by Xiong *et al.* (2006). Briefly, the original image was thresholded according to Otsu's method. Subsequently, objects larger than 1000 pixels in area were isolated, and Otsu's method was applied recursively within each area until regions smaller than 1000 pixels were obtained. During this process, if regions smaller than 25 pixels were produced they were discarded.

Method 2, referred to as the graph/thresholding method, was based upon the algorithm of Felzenszwalb and Huttenlocher (2004). After the graph-based approach was applied to the image, the original image was thresholded according to the method described by Otsu (1979). Objects determined by the graph method were then excluded if their mean intensity value fell below the determined threshold. Objects that were still in contact after the thresholding were joined into single objects, and then regions that were more than 1.4 standard deviations (SDs) below or 1 SD above the mean area of the remaining regions were excluded.

Method 3, referred to as the graph/size method, was based on the same graph-based method as the method 2. The original image was blurred with a Gaussian filter of SD three pixels, and then the graph method was applied as described to segment the entire image into regions. Regions were then separated into foreground and background by applying an absolute size-exclusion filter excluding objects smaller than 500 pixels or larger than 4000 pixels, and then applying a second size-exclusion filter in which objects were excluded if their area was more than 1.4 SDs below or 1 SD above the mean area of the remaining regions.

Method 4, referred to as the thresholding/size method, was assembled based upon various textbook pieces for object detection in images [as described, for instance, in (Gonzalez & Woods, 2002)]. Briefly, a 16-bit TIFF image was convolved with a Gaussian kernel of 21-pixel SD, and then the Sobel gradient was calculated by convolution with a 3 × 3 pixel kernel. The Sobel gradient was then thresholded at a very low level (an intensity value of 9 of 4095) to loosely determine the areas in which the cells were located. The original (smoothed) image was also thresholded based on intensity at a low level (10 of 4095 intensity), and within the intersection of the low gradient-thresholded and low intensity-thresholded areas, the image was then intensity thresholded again as follows. A histogram of pixel intensities (within the intersection of the two prior thresholded areas), and the mode value (i.e. the maximum of the histogram) was calculated. This value represented the background noise in the image. The approximate half-width at half-maximum value of the background noise was then calculated from the histogram, and the image intensity was thresholded at the mode value plus 45 times the half-width at half-maximum of the noise distribution. In practice, setting the threshold at a fixed number of half-width at half-maximum values only worked due to the low-level thresholding of the gradient, which restricted the background noise to a roughly similar area surrounding the objects in each image. Finally, the foreground of the image was labelled as individual regions, and each region was excluded from the final mask if it was smaller than 500 pixels in area or larger than 4000 pixels in area.

Method 5, called the watershed method, is a commonly used algorithm for segmentation of biological images, and was implemented as described by Gonzalez & Woods (2002). Briefly, the original image is successively thresholded starting near the maximum image intensity and decreasing in steps equal to 10% of the difference between the minimum and maximum intensity. At each step, the connected components above the threshold are identified. If at any step after the first two formerly disjoint components are joined, these are divided as follows. The two joined components are restored to their last disjoint state and gradually grown by morphological dilation until they fill the entire space of the joined component. If at any point a pixel is added to the dilation that would cause joining of the components, this pixel is permanently set as a 'dam' dividing the two components. If two pixels are simultaneously added that would cause the joining of two components only when considered together, a dam of zero thickness is added between them, permanently separating the two components. After all the thresholding iterations are complete, a mask resulting from the thresholding of the original image at 10% of background-corrected maximum intensity is applied to the mask resulting from the formation of dams so that only foreground pixels are contained in the final mask.

Seeded algorithms for object detection. For processing the human cell data set, a seeded version of each of the five algorithms was implemented using C++; source code is available in the same location and under the same license as for the non-seeded versions. In general, a seeded segmentation algorithm incorporates pre-existing information on the general location and number of regions that should result from segmentation to be able to solve more difficult segmentation problems. Specifically, the seeded algorithms developed use cell nuclei (as found by applying the recursive thresholding method to the DNA-stained images) to seed the segmentation of cells based on actin-stained images. Except as described later, the seeded version of each algorithm was the same as the non-seeded version. For all seeded methods, a final step was added where regions were filled such that any pixels not in a region but completely surrounded by a single region were added to that region. Each method can produce boundaries of zero thickness between objects, so what appears as a single object when viewing the masks may be multiple objects.

Method 1, the recursive thresholding method, was modified so that the decision to recursively apply the thresholding algorithm to a region was based not on size, but whether that region had pixels in common with more than one seed region. If a region had more than one seed, it was recursively segmented until only one seed per region was present. Finally, regions not containing any pixels in common with a seed region were discarded from the final output.

In method 2, the graph/thresholding method, the size-exclusion step was removed, and regions not containing any pixels in common with a seed region were discarded from the final output.

For method 3, the graph/size method, the step in which regions were excluded based on the mean/SD of region size was removed. Regions not containing any pixels in common with a seed region were discarded from the final output.

Method 4, the thresholding/size method, was modified most extensively. The Gaussian filtering and gradient-based thresholding were eliminated completely. Instead, the method began directly with the thresholding based upon the noise histogram step. Then, the threshold was iteratively adjusted (up or down) to make the number of connected components in the thresholded image match as closely as possible to the number of seed regions. Regions not containing any pixels in common with a seed region were discarded. Finally, the same size-exclusion step as for the non-seeded version was applied, but with a size cutoff of 500–100 000 pixels to reflect the larger size of HeLa cells.

Method 5, the watershed method, was modified according to standard practice (Gonzalez & Woods, 2002) for using this method in a seeded context. Briefly, the seed regions were used as a starting point for the watershed segmentation before beginning the successively lower thresholding. At each step, any components not containing pixels in common with a seed region were discarded, and 'dams' were only inserted between components that each contained a seed region.

Quantitative phenotype assessment. For the Kc167 cells, once regions had been created by one of the segmentation methods, the phenotype was assessed as a scalar value for each region by dividing the integrated intensity of the FITC channel (corresponding to HP1 protein) by the integrated intensity of the Hoechst channel (DNA) within each region.

For the HeLa cells, once regions had been created using the segmentation methods, the integrated intensity of the Alexa568 channel (HP1) in each region but not in a seed, and the integrated intensity of the HP1 in each region's seed were both calculated. The phenotype was assessed as the area-normalized ratio of the HP1 in the seed (DNA-associated) to the HP1 in the segmented region but not in the seed (cytoplasmic).

Benchmarking comparisons. We made two types of assessments with regard to benchmarking: phenotype- and segmentation-based. For phenotype-based comparisons, we simply applied each method to the full set of benchmark images and compared the average phenotype assessment for each image over the range of benchmarks.

To assess the quality of image segmentation by each method, we manually segmented each benchmarking image using Metamorph software (Molecular Devices). For the Kc167 cells, manual segmentation consisted of placing circles, 31 pixels in diameter, around each nucleus. This method was chosen because placing circles of fixed diameter is significantly faster than placing variable-sized circles, and in practice most nuclei were approximately this size. As assessed by eye, the largest nuclei were typically no more than 1.5 times as large as this. For HeLa cells, manual segmentation consisted of drawing a polygon approximation to the cell outline in the actin channel. The number of sides varied from cell to cell and was chosen to be near the minimum number that still produced a close approximation to the cell outline. The manually segmented regions were then converted to a binary mask to identify the regions of interest.

For a given image, the manual mask and the masks produced by each of the five automated methods were compared using a Python script or Java program that classified each cell as belonging to (possibly multiple, but at least one of) five categories: correct, missed, false positive, undersegmented or oversegmented. Correct regions were defined as those where the mapping between manual and automatic regions was one-to-one. Missed regions were those regions found in the manual mask but with no region overlapping them in the automated mask. False positives were those regions in the automated masks with no overlapping region in the manual mask. Undersegmented were those regions in the automated masks that overlapped with multiple regions in the manual mask. Oversegmented were those regions in the manual mask with multiple regions overlapping in the automated mask. For the Kc167 cells, 'overlap' was defined as having at least one pixel in common. For the HeLa cells, this definition was loosened to having 10% of the number of pixels in the manually segmented region in common (otherwise all regions were in general labelled as overlapping and no useful information resulted from this analysis).

The signal-to-noise ratio was calculated for each image using the manually created mask using the formula $S/N = [(\text{average foreground intensity} - \text{average background intensity}) \times \text{photons per grey level}]/\sqrt{(\text{average foreground intensity} \times \text{photons per grey level})}$, where the number of photons per grey level is 3.1 for the camera and settings used. This is the signal-to-noise resulting from photon counting noise in the foreground [as discussed in (Rasnik *et al.*, 2007), e.g.], which is a reasonable estimate when concerned with how the S/N varies with differing exposure or filtering for the same field of cells. Another more common way of calculating signal-to-noise is to use S/N = (average foreground intensity)/(SD of foreground intensity). However, this method was not appropriate for our images because both the DNA staining for the Kc167 cells and the actin staining for the HeLa cells were highly variable within single regions, due to variations in DNA density (heterochromatin) or density of the actin network, which would cause the S/N resulting from the properties of the imaging itself to be obscured by cell-to-cell and image-to-image differences in the variance of this density.

Determination of data acquisition requirements. We used the two-sided rank sum test to assess the statistical significance of any observed phenotypes. As a nonparametric test, the two-sided rank sum test was ideally suited for our observed phenotype distributions, which in many cases were far from Gaussian or any other recognizable distribution. As a threshold for significance, we chose $P < 10^{-9}$ because we desired significance at the 10^{-3} level, and screens often contain on the order of 10^5–10^6 images.

To determine how many cells were required to attain significance at this level, we repeated the rank sum test using the entire population of control cells but smaller random samples of varying sizes from the phenotype population (ranging from five cells to the entire population). We then took the number of cells required as the smallest sample that produced an average P-value less than 10^{-9} over at least 100 random samples of that size.

Results

Image analysis and benchmarking

To comparatively assess image segmentation methods for phenotypic screening, we implemented five different segmentation methods (described in detail in Section 'Materials and methods') (Fig. 1A,B). We chose one method (recursive thresholding) based on recursive application of intensity thresholding that had been used previously for high-throughput image analysis on the Kc167 cells used in this study (Xiong *et al.*, 2006; Chen *et al.*, 2008), a second method (graph/thresholding) that combined the recursive thresholding with a graph-joining segmentation algorithm from the computer vision field (Felzenszwalb & Huttenlocher, 2004), a third method (graph/size) combining the graph-joining segmentation with object size filtering, a fourth method (thresholding/size) assembled from simple textbook image analysis pieces (Gonzalez & Woods, 2002) and a fifth method (watershed) based on the widely used watershed algorithm (Gonzalez & Woods, 2002). As a model data set for our benchmarking analysis, we tested our segmentation methods on a genome-wide image-based RNAi depletion screen in *D. melanogaster* Kc167 cells designed to identify defects in HP1 nuclear localization.

Figure 1A shows the results of applying each of the five image segmentation algorithms to the epifluorescence images of Hoechst-stained Kc167 cells. These methods produced distinct segmentation patterns ranging from the efficient identification of nuclei in the recursive thresholding method to the graph/size method that produced several improperly

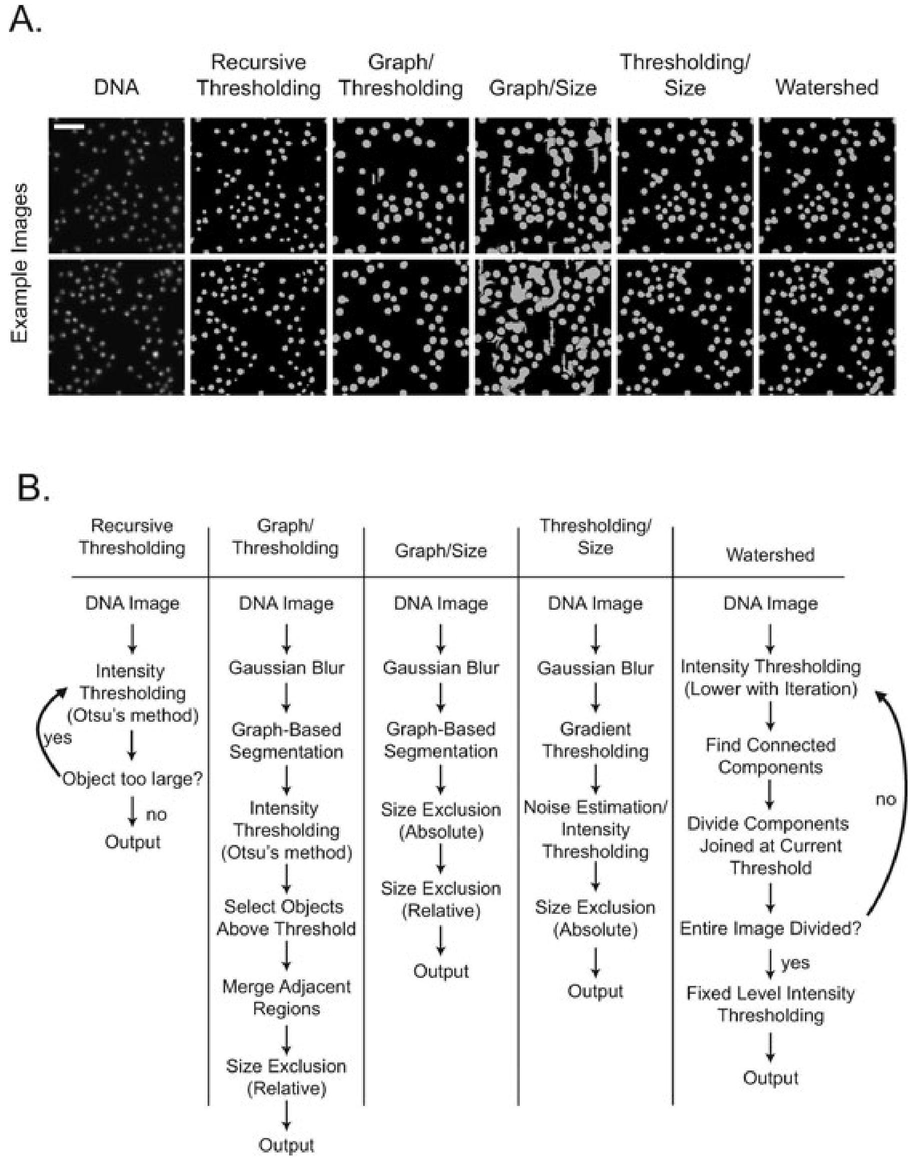

Fig. 1. Image segmentation methods used for benchmarking. (A) Image segmentation methods applied to DNA-stained Kc167 cells. The leftmost panels (DNA) show two example images (bar = 25 μm). The series of 10 panels on the right depict the binary mask created after each image segmentation method was applied to the two sample images. Grey pixels represent objects and black pixels represent background. (B) Flowchart of each of the five segmentation methods used for benchmarking.

segmented areas. Based on visual inspection, these methods provide a range of segmentation performance for testing our benchmarking approach.

We previously performed an image-based high-throughput screen and found that image data from automated microscopy can vary widely in quality with regard to focus, signal-to-noise and cell density (Erhardt *et al.*, 2008). We expected that these deviations from optimal image quality would compromise the efficiency of image segmentation; thus, we evaluated the performance of each image segmentation method with image data sets designed to mimic these variations. We created a set of benchmarking images representing variation in focus (Fig. 2A), signal-to-noise (Fig. 2C) and cell density (Fig. 2E) by systematically defocusing images, putting neutral density filters into the light path and plating cells on cover slips at different densities.

To compare the performance of different automated segmentation methods in analyzing our benchmark data set, we manually segmented the benchmarking images by drawing circles around nuclei in the Hoechst stained images and used this as our reference segmentation. We then quantitatively evaluated the robustness of each segmentation

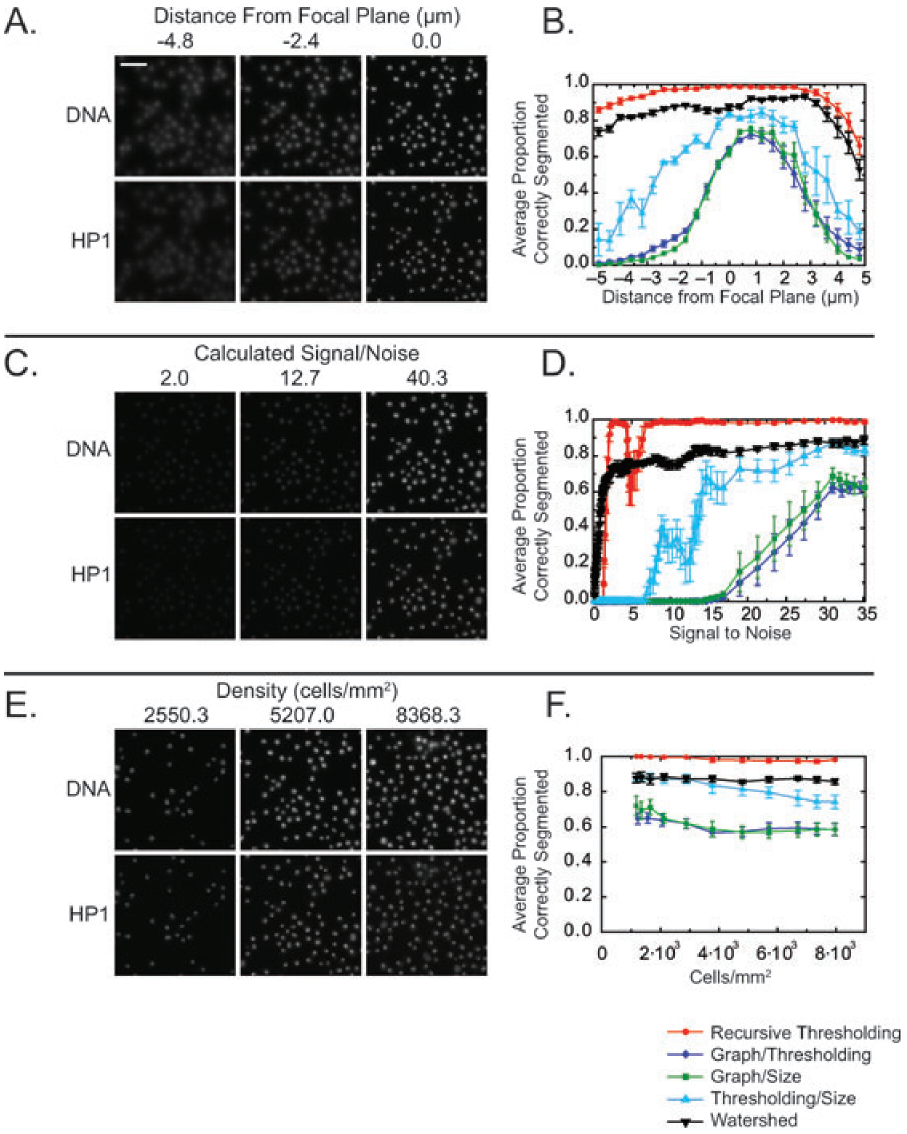

Fig. 2. Quantitative benchmarking of image segmentation. Sample images of Kc167 cells fixed and stained for DNA and HP1 are shown in panels (A), (C) and (E) (bar = 25 µm), and quantification of the proportion of nuclei correctly segmented versus the quantitative benchmark is shown in panels (B), (D) and (F). (A, B) Benchmarking based on focal plane. Cells were imaged at 0.4 µm axial intervals from −4.8 to +4.8 µm around the focal plane. The quantification is the mean ± SEM of the proportion of objects correctly segmented in three images. (C, D) Benchmarking based on signal-to-noise. The signal-to-noise was varied both by inserting neutral density filters into the light path and by changing the exposure time. (C) Representative images acquired at different signal-to-noise ratios. (D) The quantification represents a seven-point moving average ± SEM (of these seven points) of the proportion of objects correctly segmented in each image. (E, F) Benchmarking based on cell density. (E) Images of cells plated over a 16-fold density range. The cell density was calculated by counting the number of nuclei per field and dividing by the field size. (F) The quantification is a five-point moving average ± SEM (of these five points) of the proportion of objects correctly segmented in each image.

method in two ways: how closely the segmentation obtained from a given method matched that obtained manually, and how closely the phenotypic scoring matched the score obtained from the manually segmented image.

To quantify the segmentation efficiency of each algorithm, we classified each segmented object as belonging to one of five categories: correct, indicating a one-to-one matchup with the manual segmentation; missed, indicating an object in the manual segmentation with no overlapping object in the automated segmentation; false positive, indicating an object in the automated segmentation with no overlapping object in the manual segmentation; oversegmented, indicating an object in the manual segmentation overlapping with multiple objects in the automated segmentation and undersegmented, indicating an object in the automated segmentation overlapping with multiple objects in the manual segmentation. The proportion of correctly segmented objects with respect to variation in focal plane, signal-to-noise and cell density is plotted in Fig. 2B, D and F, respectively. The proportion of objects classified as missed, false positive, undersegmented and oversegmented with respect to each type of variation are shown in Supporting Figure S1.

The proportion of objects classified as correct varied strongly with focus and signal-to-noise, but only weakly with cell density (Fig. 2B, D and F). Variation with focal plane was continuous (Fig. 2B), but the variation with respect to signal-to-noise ratio was more steplike: below a certain characteristic signal-to-noise value, each method was ineffectual at correctly identifying any objects (Fig. 2D). Above that value, the fraction of correctly identified nuclei increased until it reached a final value that varied little with further increase in signal-to-noise. The recursive thresholding method performed significantly better than the other three methods for all varied parameters while the graph/size method performed the least well. Surprisingly, even 4.8 μm out of focus, and at signal-to-noise as low as 3.5, the recursive thresholding method was able to correctly identify more than half of the objects.

We observed that some methods produced characteristic errors when they did not correctly segment images. The graph/size method and the watershed method (at low signal-to-noise and out of focus) tended to produce many false positives, whereas the graph/thresholding and thresholding/size methods produced few false positives but tended to undersegment images (Supporting Figure S1). Stereotyped segmentation errors may have a more detrimental effect in certain applications; thus, a segmentation method that minimizes these errors can be chosen accordingly.

The ultimate goal of image segmentation is to identify objects that can be analyzed and phenotypically scored. We quantified the success of each segmentation method by calculating the ratio of the HP1 fluorescence to the DNA fluorescence in the regions identified by automated segmentation and comparing that ratio to the same ratio produced by manual segmentation. The phenotypic value produced by the segmentation (which included all objects, not just ones classified as correct in the previous analysis) showed little variation over the full range of benchmarking conditions (Supporting Figure S2). The graph/size method produced phenotype values consistently significantly different than the other methods, indicating that it is possible to distinguish this method from the others based on phenotype value as well as segmentation quality.

Determination of data acquisition requirements for statistical significance

A clear advantage of performing a segmentation benchmarking experiment independent of a high-throughput screen is that the information gained from the benchmarking can feed back into the design of the screen. In particular, the screen can be designed to avoid problems in the image segmentation methods (e.g. to focus images more carefully if the method fails on out-of-focus images).

Benchmarking experiments also provide the potential to inform the amount of data collection required during the screen to ensure statistically significant phenotype detection. To test this idea, we determined the data acquisition requirements for a high-throughput screen given a particular phenotype and a particular image analysis method. In practice, many single-cell distributions of phenotypes are decidedly non-Gaussian, so to avoid any statistical test that makes assumptions about the distribution of the data we used the rank sum test as a nonparametric method to determine statistical significance. We depleted HP1 from Kc167 cells by dsRNA treatment and performed a rank sum test to compare this population of cells with a control population treated with water. The *P*-value for the hypothesis that the two populations were drawn from the same distribution was less than 10^{-290}. We compared two separately treated control populations against each other and found that the *P*-value was greater than 0.7, indicating that the 10^{-290} *P*-value between the control and dsRNA-treated populations was not simply caused by variation between two uniformly treated populations.

We estimated the amount of data required to detect the HP1 depletion phenotype at a desired significance level by performing a data-removal test. We randomly sampled a chosen number of cells from the experimental HP1 depleted population, applied the rank sum test to this subset and the entire control population, and calculated the average *P*-value over many such subsets. We plotted the *P*-value versus the sample size to determine the number of cells required to achieve a particular significance level (Fig. 3A). Surprisingly, we found that we only needed a sample size of about 15 cells to achieve significance at the 10^{-9} level for the recursive thresholding, graph/thresholding and thresholding/size methods. It is possible to achieve high significance in this experiment with such a small sample size because we compared each experimental sample against the entire population of control cells (500–1000 cells was typical). Additional control data is inexpensive and easy to collect, and because the same population of cells can serve as controls for many experimental treatments increasing the amount of control data can partially compensate for the lower significance values associated with a smaller experimental population.

The dependence of the *P*-value on a benchmarking parameter such as focal plane change should closely mirror the dependence of the phenotypic value on the same parameter. To test this, we performed the data removal test for each segmentation method over the full range of focal planes previously analyzed. We compared the entire population of control cell images taken at a specific focal plane with the random images from the population of RNAi-treated cells at the same focal plane (Fig. 3B). The graph/thresholding,

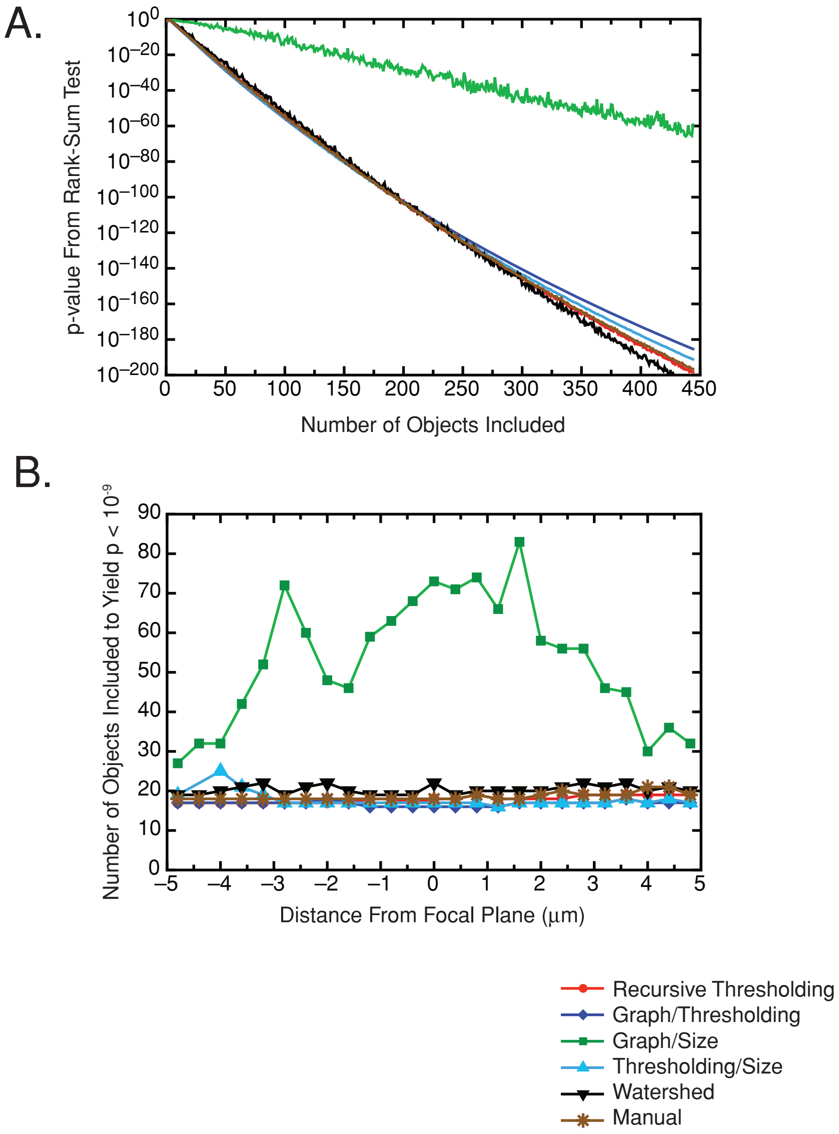

Fig. 3. Data removal test to determine data acquisition requirements. (A) Determination of the number of cells required for given statistical significance level for each analysis method. Control or HP1-depleted cell populations were compared by dividing the integrated HP1 intensity by the integrated DNA intensity for each cell then using a data removal test to calculate the number of cells necessary to distinguish between the control and experimental populations at different significance levels. The mean *P*-value from the rank sum test over 100 random samples is plotted versus the size of the sample for each of the five segmentation methods. (B) Dependence of the sample size needed for statistical significance at $P < 10^{-9}$ on focal plane. Cells treated as in (A) were imaged at 0.4 μm intervals from −4.8 to +4.8 μm around the focal plane. The data removal test was repeated for each set of control and RNAi images at each focus step. The sample size required to achieve significance at the 10^{-9} level is plotted versus distance from optimal focus.

thresholding/size, recursive thresholding and watershed methods showed very little dependence of the *P*-value on focal plane, consistent with the result that the phenotype value shows little dependence on focal plane (Supporting Figure S2A).

In an actual screen, many phenotypes of interest will not have the full strength of the positive control phenotype. To test the ability of each algorithm to detect intermediate phenotypes, we depleted Kc167 cells with HP1 dsRNA for 16, 24, 40 or 72 h to generate cells with partial HP1 depletion. Figure 4A shows images of the intermediate phenotypes, and Fig. 4B shows a quantification of the phenotype value for each of the five automated methods as well for manual segmentation. All of the automated methods except the graph/size method were able to accurately quantify the intermediate phenotypes as well as the manual segmentation. The graph/size method underestimated the phenotype at intermediate times.

We determined the amount of data required to significantly detect intermediate phenotypes by repeating the data removal test for each intermediate phenotype using each analysis method. The ability to significantly detect intermediate phenotypes at significance levels ranging from 0.05 to 10^{-9} showed a steep dependence on the sample size (Fig. 5A–C). After 16 h of dsRNA treatment resulting in 4% knockdown (quantified as 100% – HP1-to-DNA ratio normalized to a no RNA control), no method produced significance to the 10^{-9} level even when using the full population of ∼350 cells (Fig. 5A), only the recursive thresholding method was significant at the 10^{-4} level (Fig. 5B) and all but the graph size method were significant at the $P < 0.05$ level (Fig. 5C). By 24 h treatment resulting in 22% knockdown, all but the graph/size

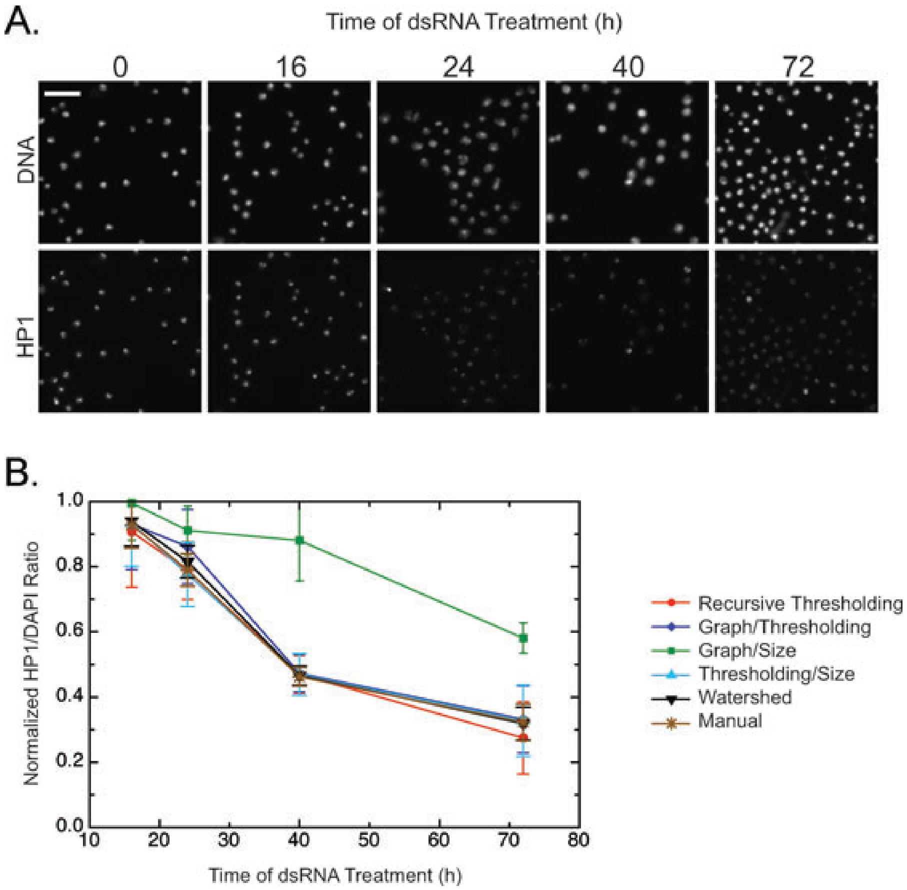

Fig. 4. Quantification of intermediate HP1 depletion phenotypes. A) Images of Kc167 cells treated with HP1 dsRNA for 0, 16, 24, 40 or 72 h, and fixed and stained for DNA and HP1 (bar = 25 μm). (B) A phenotype value for each object in an image was determined by calculating the ratio of HP1 to DNA integrated intensity within each object. (B) The mean HP1/DNA ratio ± SD on a per cell basis normalized to a no RNA treatment control versus time of RNAi treatment for all five segmentation methods and the manual segmentation.

method produced significance with less than 100 cells, and by 40 h treatment causing 55% knockdown, graph/size was significant with around 600 cells (variability in DNA staining caused a decrease in graph/size method segmentation efficiency at this timepoint), and the other methods needed fewer than 20 cells for significance.

Application to different phenotypes

To demonstrate that our methodology can be used for a variety of phenotypes, and to examine its ability to discern among analysis methods on a more demanding segmentation problem, we created a second benchmarking data set to analyze the distribution of HP1 between chromatin and the cytoplasm at different cell cycle stages. This is particularly relevant to the biology of HP1, which is known to dissociate from chromatin during mitosis and then relocalize to heterochromatin in interphase (Kellum *et al.*, 1995; Sugimoto *et al.*, 2001). We stained HeLa cells with phalloidin (to visualize actin), Hoechst (to visualize DNA) and anti-HP1 antibody, and used the DNA and actin information to segment not only nuclei but also the cell boundary. We expected this to be a much more difficult task for automated segmentation as the actin staining of adjacent cells often appears continuous (Fig. 6).

We developed modified (seeded) versions of all five segmentation algorithms used for the Kc167 cell data set that perform the segmentation on the actin channel, but incorporate the location of the nuclei in the image to refine estimates of cell boundaries and numbers; these modifications are detailed in Section ‘Materials and methods’ and outlined in Fig. 6B. Figure 6A shows the results of applying these seeded methods to actin images of HeLa cells. As we observed with the Kc167 cell data set, the five methods display widely variable segmentation efficiencies.

To determine the segmentation efficiency of each modified method, we generated benchmarking image data sets to test the methods by systematically varying the focus (Supporting Figure S3A), signal-to-noise (Supporting Figure S3C) and cell density (Supporting Figure S3E). These benchmarking images were manually segmented by drawing many-sided polygons to approximate the cell boundaries, and this manual

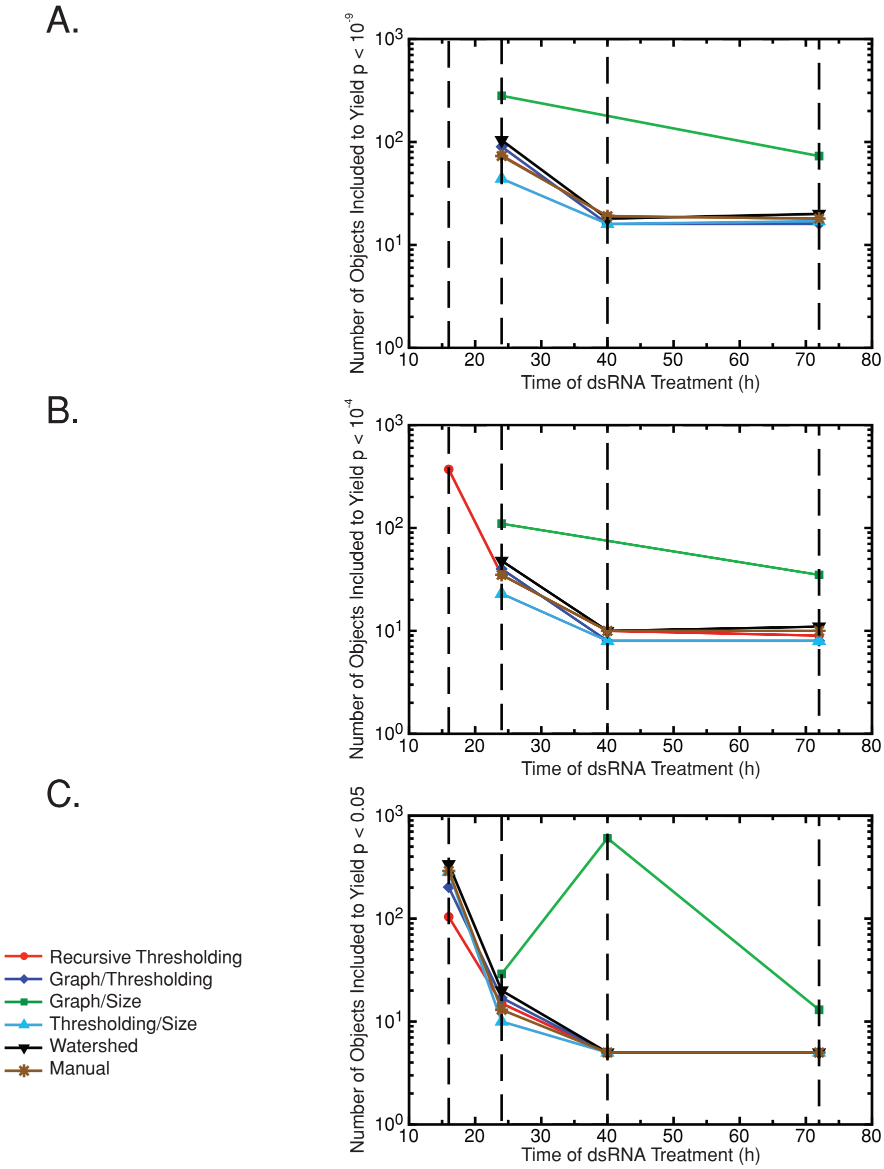

Fig. 5. Determination of the data acquisition requirements by data removal test for intermediate phenotypes. The populations of cells shown in Fig. 4 were used for a data removal test. The absence of a data point for a given time, method and significance level denotes that not even the full population of cells tested as significant. (A) Sample size required for significance at $P < 10^{-9}$ versus time of RNAi treatment. (B) Sample size required for significance at $P < 10^{-4}$ versus time of RNAi treatment. (C) Sample size required for significance at $P < 0.05$ versus time of RNAi treatment.

segmentation was used to classify each object produced by an automated method as correct, missed, false positive, undersegmented or oversegmented. Supporting Figures 3B, 3D and 3F show the proportion of objects classified as correctly segmented as a function of focus, signal-to-noise and cell density, respectively. The graph/size method showed a strong dependence on focus and cell density but all other methods were robust to focus and cell density variation. All methods showed a strong dependence on signal-to-noise ratio, failing below a value of about 5. The seeded version of the graph/size method consistently produced the best segmentation over a wide range of conditions, by contrast to its performance on the previous Kc167 cell data set.

Different segmentation methods showed different characteristic errors on our human cell data set (Supporting Figure S4). The recursive thresholding and thresholding/size methods tended to miss cells; the watershed method tended to over- or undersegment cells. False positives were close to zero for all methods, as the information incorporated about the location of nuclei led the methods to discard cells that clearly had no nucleus. Quantification of each classification for each benchmark for all five methods is shown in Supporting

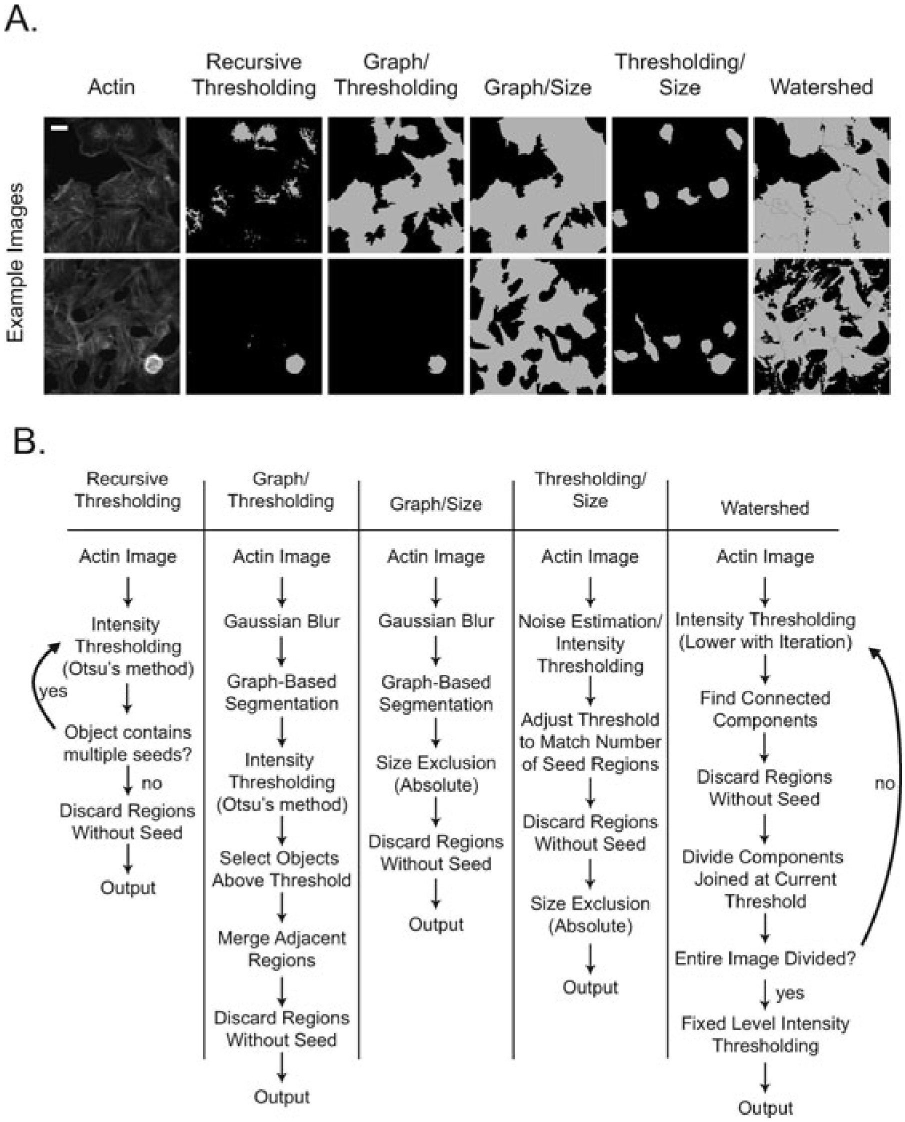
A.
Actin
Recursive Thresholding
Graph/ Thresholding
Graph/Size
Thresholding/ Size
Watershed
Example Images
B.
Recursive Thresholding
Graph/ Thresholding
Graph/Size
Thresholding/ Size
Watershed
Actin Image
Intensity Thresholding (Otsu's method)
yes
Object contains multiple seeds?
no
Discard Regions Without Seed
Output
Actin Image
Gaussian Blur
Graph-Based Segmentation
Intensity Thresholding (Otsu's method)
Select Objects Above Threshold
Merge Adjacent Regions
Discard Regions Without Seed
Output
Actin Image
Gaussian Blur
Graph-Based Segmentation
Size Exclusion (Absolute)
Discard Regions Without Seed
Output
Actin Image
Noise Estimation/ Intensity Thresholding
Adjust Threshold to Match Number of Seed Regions
Discard Regions Without Seed
Size Exclusion (Absolute)
Output
Actin Image
Intensity Thresholding (Lower with Iteration)
Find Connected Components
Discard Regions Without Seed
Divide Components Joined at Current Threshold
Entire Image Divided?
no
yes
Fixed Level Intensity Thresholding
Output

Fig. 6. Seeded image segmentation methods used for benchmarking. (A) Image segmentation methods applied to actin-stained HeLa cells. The leftmost panels (actin) show two example images (bar = 20 μm). The series of 10 panels on the right depict the binary mask created after applying each seeded image segmentation method. Grey pixels represent objects and black pixels represent background. (B) Flowchart of each of the five seeded segmentation methods used for benchmarking.

Figure S4. The systematic errors we observed on the human cell data set were distinct from those observed during the segmentation of our *Drosophila* cell data set. This demonstrates the importance of generating segmentation problem specific benchmarking data sets when comparing different image analysis methods.

To extract a phenotype value (chromatin vs. cytoplasmic localization of HP1) from the images, we quantified the ratio of the average intensity of HP1 staining in the nucleus to the average intensity of HP1 staining in the cytoplasm on a per cell basis. The results of this quantification over the full range of benchmarking conditions are shown in Supporting Figure S5. By contrast to our results with the *Drosophila* cell data set, the phenotype value showed a strong dependence on both focal plane and signal-to-noise. This dependence was also present for manually segmented images; thus, this strong dependence on signal-to-noise does not result from poor segmentation at lower signal-to-noise, but rather from an actual change in the phenotype value. This can be explained by the fact that the cytoplasmic staining was already at the noise level in most cells over the full range of benchmark conditions as would be expected for a nuclear protein in predominantly interphase cells. However, the phenotype values obtained from the different methods varied widely with respect to the value obtained from the manual segmentation (even for those methods that performed very well), indicating that the phenotype value for this data set is very sensitive to the quality of image segmentation.

To determine whether we could detect a statistically significant change in phenotype when the variation in phenotype values was high, we generated image data sets under conditions that forced HP1 to localize predominantly to the DNA or throughout the cytoplasm. Phosphorylation of histone H3 by aurora kinase causes HP1 dissociation from chromatin during mitosis in human cells (Fischle *et al.*, 2005; Hirota *et al.*, 2005). We treated cells with nocodazole to depolymerize microtubules and arrest cells in mitosis, and then we stained these cells with phalloidin (for actin), with Hoechst (for DNA), and with anti-tubulin and anti-HP1 antibodies. Example images from untreated and nocodazole-treated cells are shown in Fig. 7A. The phenotype was quantified for each of the drug treatments and normalized to the value for the control-treated population analyzed by manual segmentation. Although the values obtained from each segmentation method varied, all methods except the thresholding/size method could detect the decrease in the ratio of chromatin-associated to cytoplasmic HP1 (Fig. 7B).

To assess the statistical power of each method to distinguish the phenotype, we again turned to the data removal test, where a rank sum test was applied to the full population of control-treated cells and decreasing numbers of nocodazole-treated cells. Supporting Figure S6A shows the dependence of the average *P*-value over 100 repeats of random data removal on the number of cells included in the test. The graph/size method had statistical power comparable to the manual segmentation; the watershed and graph/thresholding methods had somewhat better power than manual segmentation (these segmented mitotic cells accurately, but non-mitotic cells less accurately, which selected for cells with the phenotype); the thresholding/size and recursive thresholding methods had significantly lower statistical power than the manual segmentation. We confirmed that the statistical power varied with benchmarking conditions by examining the dependence of the number of cells required

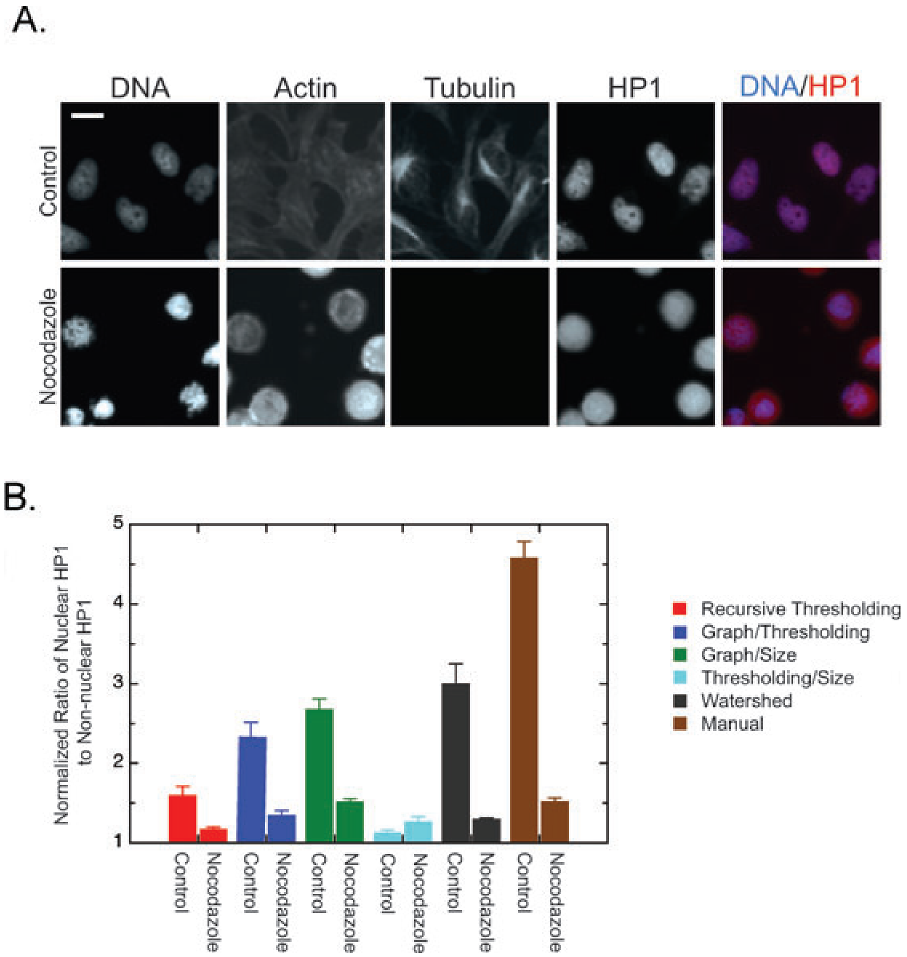

Fig. 7. Quantification of HP1 redistribution. (A) Images of control HeLa cells and cells synchronized by nocodazole treatment stained for DNA, actin, tubulin and HP1 (bar = 20 μm). (B) The ratio of nuclear to cytoplasmic HP1 staining in nocodazole-arrested and untreated control cells. The mean ratio ± SEM normalized to the value of manual segmentation of the control treated cells from 16 images is shown for all segmentation methods.

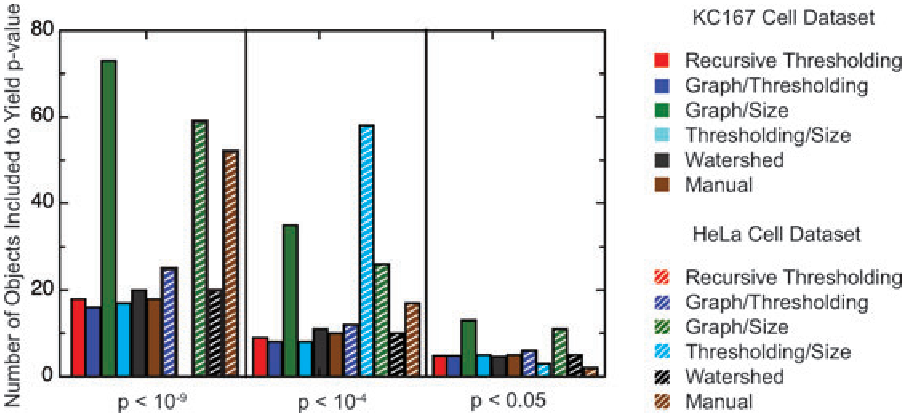

Fig. 8. Comparison of the benchmarking methodology applied to two different analysis problems. The Kc167 cell and HeLa cell data sets were subjected to the data removal test (Fig. 3 and Supporting Figure S4, respectively) but with *P*-value cutoffs at $P < 10^{-9}$, $P < 10^{-4}$ or $P < 0.05$. The number of objects required to attain these *P*-value cutoffs are plotted versus *P*-value for each segmentation method and each data set. The seeded recursive thresholding method did not produce statistical significance at any of the three *P*-value cutoffs, and the seeded thresholding/size method did not produce statistical significance at the $P < 10^{-9}$ level.

for significance at $P < 10^{-9}$ on focal plane (Supporting Figure S6B). In agreement with the benchmarking results, the number of cells required for significance depended strongly on focus when the segmentation efficiency or phenotype value depended strongly on focus.

To compare the results of the benchmarking methodology between the *Drosophila* and human cell data sets, we determined the number of cells required for significance at $P < 10^{-9}$, $P < 10^{-4}$ and $P < 0.05$ for each method in the Kc167 cell data set and in the HeLa cell data set (Fig. 8). This comparison demonstrates that a single segmentation and analysis method is unlikely to be appropriate for different image analysis problems, but that the best method among a group can be determined a priori. For instance, although the thresholding/size method performs very well on the Kc167 cell data set, its seeded counterpart is less able to discern the phenotype in the HeLa cell data set. This provides confidence in the ability to distinguish a phenotype for a given task as well as an estimation of the data acquisition requirements that will be needed to discern a given phenotype in a large-scale screen.

Discussion

The use of RNAi and small molecule inhibitor collections in cell-based phenotypic screening has enabled the collection of an unprecedented amount of image-based phenotypic data, yet the efficient computational analysis of this data remains a significant challenge. Many image segmentation and analysis algorithms, as well as many image data sets for benchmarking them, exist, but no method or benchmarking data set is optimal for all screening problems. Currently, most image-based screening efforts involve performing the experimental component of a screen then creating new analysis methods or modifying existing methods until the desired information is extracted from the images.

We have demonstrated that performing small benchmarking experiments combined with statistical analysis can be used (1) to quantitatively assess the quality of different image segmentation methods for a given screening problem and (2) to determine the data acquisition requirements during screening to ensure statistically significant detection of a given phenotype.

Because benchmarking and analysis can be carried out before performing a screen, the information gleaned from benchmarking can be used to determine the parameters (e.g., focus or signal-to-noise) that need to be most tightly controlled and the amount of data that must be collected for each treatment in the full screen. In the absence of universally applicable segmentation algorithms or benchmarking data sets, this methodology enables image-based screens to proceed even with imperfect methods through *a priori* determination of the experimental parameters and data collection requirements for significant phenotype detection.

The steps that we followed for setting up our benchmarking experiment should be generally applicable to any high-throughput screen. First, we chose the benchmarking parameters for our screen. In our case, intermediate phenotype generation, commonly encountered image variation and changes in cell density are reasonable parameters to vary for image-based RNAi or chemical screens. We treated cells with RNAs or chemicals for varying amounts of time to generate intermediate phenotypes, and varied signal-to-noise, cell density and focal plane because these variations commonly occur in high-throughput imaging screens. Second, we selected image analysis methods for comparison. We have chosen five widely applicable methods for our benchmarking that were likely to be appropriate for segmenting our data set.

The specific image analysis methods will vary significantly depending on the nature of the screen but the software we have developed is modular in design to enable relative ease in new method addition. Third, once the methods were chosen, we generated segmentation masks and phenotypic values and if desired a manually segmented data set as a 'gold standard'. We have automated the classification and comparison of segmentation methods so that the efficiency of segmentation can be quickly assessed and the results plotted to determine which analysis method to use. Finally, we ran a data removal test on the phenotype values to estimate the number of cells required for significance at a desired level.

In practice, high-throughput screening often aims to find multiple phenotypes. Our methodology is equally applicable to such screens, as long as a positive and negative control data set can be generated for each phenotype of interest. Because this methodology identifies the optimal method among a group for the detection of a given phenotype, this benchmarking can be performed separately for each phenotype of interest to identify effective image segmentation methods for any phenotype. Our approach is likely to produce more reliable results than choosing a one-size-fits-all method that may not be optimal for any phenotype.

Application of this benchmarking and analysis methodology to feed back into experimental screen design should prove generally useful for high-throughput high-content imaging screens. Furthermore, this method allows quantitative statements to be made about detectable phenotypes in a screen, and quantitative assessment of which image analysis methods are optimal for any given data set. The methods we present should allow significant savings in acquisition and analysis costs and facilitate the collection of higher quality data with improved information content.

References

Agaisse, H., Burrack, L.S., Philips, J.A., Rubin, E.J., Perrimon, N. & Higgins, D.E. (2005) Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science* **309**, 1248–1251.

Baatz, M., Arini, N., Schape, A., Binnig, G. & Linssen, B. (2006) Object-oriented image analysis for high content screening: detailed quantification of cells and sub cellular structures with the cellenger software. *Cytom. Part A* **69A**, 652–658.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R. *et al.* (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100.1–R100.11.

Chen, C., Li, H., Zhou, X. & Wong, S.T. (2008) Constraint factor graph cut-based active contour method for automated cellular image segmentation in RNAi screening. *J. Microsc.* **230**, 177–191.

Erhardt, S.E., Mellone, B.G., Betts, C.M., Zhang, W., Karpen, G.H. & Straight, A.F. (2008) Genome-wide analysis reveals a cell cycle dependent mechanism controlling centromere propagation. *J. Cell Biol.* **183**, 805–818.

Felzenszwalb, P.F. & Huttenlocher, D.P. (2004) Efficient graph-based image segmentation. *Int. J. Comput. Vision* **59**, 167–181.

Fischle, W., Tseng, B.S., Dormann, H.L. *et al.* (2005) Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* **438**, 1116–1122.

Glory, E. & Murphy, R.F. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev. Cell* **12**, 7–16.

Goldberg, I.G., Allan, C., Burel, J.M. *et al.* (2005) The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol.* **6**, R47.41–R47.13.

Gonczy, P., Echeverri, C., Oegema, K. *et al.* (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408**, 331–336.

Gonzalez, R.C. & Woods, R.E. (2002) *Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ.

Hirota, T., Lipp, J.J., Toh, B.H. & Peters, J.M. (2005) Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. *Nature* **438**, 1176–1180.

Kellum, R., Raff, J.W. & Alberts, B.M. (1995) Heterochromatin protein 1 distribution during development and during the cell cycle in Drosophila embryos. *J. Cell Sci.* **108**, 1407–1418.

Kiger, A., Baum, B., Armknecht, S. *et al.* (2002) Functional genomic analysis of cellular morphology using high-throughput RNAi screens. *Dev. Biol.* **247**, 480–480.

Li, F.F., Fergus, R. & Perona, P. (2007) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Und.* **106**, 59–70.

Martin, D., Fowlkes, C., Tal, D. & Malik, J. (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proc. 8th Int. Conf. Comput. Vision* **2**, 416–423.

Mayer, T.U., Kapoor, T.M., Haggarty, S.J., King, R.W., Schreiber, S.L. & Mitchison, T.J. (1999) Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. *Science* **286**, 971–974.

Neumann, B., Held, M., Liebel, U., Erfle, H., Rogers, P., Pepperkok, R. & Ellenberg, J. (2006) High-throughput RNAi screening by time-lapse imaging of live human cells. *Nat. Methods* **3**, 385–390.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M. & Goldberg, I.G. (2008) WND-CHARM: multi-purpose image classification using compound image transforms. *Pattern Recogn. Lett.* **29**, 1684–1693.

Otsu, N. (1979) Threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cyb.* **9**, 62–66.

Peng, H.C. (2008) Bioimage informatics: a new area of engineering biology. *Bioinformatics* **24**, 1827–1836.

Perlman, Z.E., Slack, M.D., Feng, Y., Mitchison, T.J., Wu, L.F. & Altschuler, S.J. (2004) Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198.

Rasnik, I., French, T., Jacobson, K. & Berland, K. (2007) Electronic cameras for low-light microscopy. *Methods Cell Biol.* **81**, 219–249.

Shamir, L., Orlov, N., Eckley, D.M., Macura, T.J. & Goldberg, I.G. (2008) IICBU 2008: a proposed benchmark suite for biological image analysis. *Med. Biol. Eng. Comput.* **46**, 943–947.

Sugimoto, K., Tasaka, H. & Dotsu, M. (2001) Molecular behavior in living mitotic cells of human centromere heterochromatin protein HP1alpha ectopically expressed as a fusion to red fluorescent protein. *Cell Struct. Funct.* **26**, 705–718.

Xiong, G., Zhou, X., Ji, L., Bradley, P., Perrimon, N. & Wong, S. (2006) Segmentation of drosophila RNAi fluorescence images using level sets. *IEEE International Conference on Image Processing*, pp. 73–76. Proceedings of the IEEE International Conference on Image Processing.

Yarrow, J.C., Feng, Y., Perlman, Z.E., Kirchhausen, T. & Mitchison, T.J. (2003) Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb. Chem. High Thr. Scr.* **6**, 279–286.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Quantitative benchmarking of segmentation efficiency. For each panel, objects were classified into one or more of five categories: correct, missed, false positives, undersegmented or oversegmented. Correct segmentation is shown in Fig. 2. The proportion of objects classified as (from top to bottom) missed, false positive, undersegmented and oversegmented is plotted versus the benchmark parameter. (A) Benchmarking based on focal plane. The mean proportion of cells segmented in each category ± SEM from three images. (B) Benchmarking based on signal-to-noise. The quantification is a seven-point moving average ± SEM (of these seven points) of the object classification for each image. (C) Benchmarking based on cell density. The quantification is a five-point moving average ± SEM (of these five points) of the object classification for each image.

Fig. S2. Quantitative benchmarking of phenotype detection. Kc167 cells were fixed and stained for HP1 and DNA, and segmented by each of the five methods. A phenotype value for each object in an image was determined by dividing the integrated HP1 intensity within that object by the integrated DNA intensity within the same object. (A) Benchmarking based on focal plane. Cells were imaged at 0.4 μm axial intervals from −4.8 to +4.8 μm around the focal plane. The quantification is the mean ± SEM from three images. (B) Benchmarking based on signal-to-noise. Signal-to-noise was varied by inserting neutral density filters into the light path and adjusting the exposure time. The quantification is a seven-point moving average ± SEM (of these seven points) on a per image basis of the set of images where at least one object was detected. (C) Benchmarking based on cell density. The cell density was calculated by counting the number of nuclei per field and dividing by the field size. The quantification is a five-point moving average ± SEM (of these five points) of the phenotype value for each image.

Fig. S3. Quantitative benchmarking of seeded image segmentation. Sample images of HeLa cells fixed and stained for DNA and actin are shown in panels (A), (C) and E (bar = 20 μm), and quantification of the proportion of nuclei correctly segmented versus the quantitative benchmark is shown in panels (B), (D) and (F). (A, B) Benchmarking based on focal plane. Cells were imaged at 0.8 μm axial intervals from −8.0 to +8.0 μm around the focal plane. The quantification indicates the mean ± SEM of the proportion of objects correctly segmented in 16 images. (C, D) Benchmarking based on signal-to-noise. Signal-to-noise was varied by inserting neutral density filters into the light path and changing the exposure time. The quantification represents a seven-point moving average ± SEM (of these seven points) of the proportion of objects correctly segmented in each image. (E, F) Benchmarking based on cell density. The cell density was calculated by counting the number of nuclei per field and dividing by the field size. The quantification is a five-point moving average ± SEM (of these five points) of the proportion of objects correctly segmented in each image.

Fig. S4. Quantitative benchmarking of seeded segmentation efficiency. The proportion of objects classified as (from top to bottom) missed, false positive, undersegmented and oversegmented is plotted versus the quantitative benchmark. (A) Benchmarking based on focal plane. The mean proportion of cells segmented in each category ± SEM from 16 images is shown. (B) Benchmarking based on signal-to-noise. The quantification is a seven-point moving average ± SEM (of these seven points) of the object classification for each image. (C) Benchmarking based on cell density. The quantification is a five-point moving average ± SEM (of these five points) of the object classification for each image.

Fig. S5. Quantitative benchmarking of phenotype detection with seeded segmentation methods. The variation in phenotype values, representing the ratio of nuclear to cytoplasmic HP1 staining, is shown for each benchmark parameter. (A) Benchmarking based on focal plane. Cells were imaged at 0.8 μm axial intervals from −8.0 to +8.0 μm around the focal plane. The quantification is the mean ± SEM on a per cell basis with cells pooled from three images. (B) Benchmarking based on signal-to-noise. Signal-to-noise was varied by inserting neutral density filters into the light path and changing the exposure time. The quantification is a seven-point moving average ± SEM on a per image basis of the set of images where at least one object was detected. (C) Benchmarking based on cell density. The cell density was calculated by counting the number of nuclei per field and dividing by the field size. The quantification is a five-point moving average ± SEM of the phenotype value for each image.

Fig. S6. Data removal test to determine data acquisition requirements. (A) Determination of the number of cells required for given statistical significance level for each analysis method. Control or nocodazole-treated cell populations were compared by dividing the average HP1 intensity within the seed DNA region by the average HP1 intensity within the same cell but not in the seed region and then using a data removal test to calculate the number of cells necessary to

distinguish between the control and experimental populations at different significance levels. The mean *P*-value from the rank sum test over 100 random samples is plotted versus the size of the sample for each of the five segmentation methods. (B) Dependence of the sample size needed for statistical significance at $P < 10^{-9}$ on focal plane. Cells treated as in (A) were imaged at 0.8 μm axial intervals from −8.0 to +8.0 μm around the focal plane. The data removal test was repeated for each set of control and nocodazole images at the same focal plane. The sample size required to achieve significance at the 10^{-9} level is plotted versus distance from optimal focus.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.