AD________________

Award Number: W81XWH-05-1-0138

TITLE: Identifying Novel Drug Targets for the Treatment of Tuberous Sclerosis Complex Using High Throughput Technologies

PRINCIPAL INVESTIGATOR: David Sabatini, M.D., Ph.D.

CONTRACTING ORGANIZATION: Whitehead Institute for Biomedical Research
Cambridge, MA 02142-1479

REPORT DATE: January 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE 01-01-2006	**2. REPORT TYPE** Final	**3. DATES COVERED** 8 Dec 2004 – 7 Dec 2005

4. TITLE AND SUBTITLE

Identifying Novel Drug Targets for the Treatment of Tuberous Sclerosis Complex Using High Throughput Technologies

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0138

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

David Sabatini, M.D., Ph.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Whitehead Institute for Biomedical Research
Cambridge, MA 02142-1479

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT

In a patient with Tuberous Sclerosis Complex (TSC), the problematic cells that initiate and constitute tumors have lost TSC1 or TSC2 function. A promising approach for treatment would be to target members of the pathway with which TSC1/2 proteins interact. In cultured drosophila cells, we proposed to rapidly identify genes whose RNAi-mediated reduction in expression (1) Prevents growth/proliferation of TSC1 or TSC2-deficient cells without affecting normal cells. (2) Induces apoptosis/cell death in TSC1 or TSC2-deficient cells without killing normal cells. (3) Reverts TSC1 or TSC2-deficient cells to a normal phenotype, as determined by measuring a reporter of cell growth pathway activation and cell morphology. We have (1) advanced genome-wide RNA interference living cell microarrays from proof-of-principle to a robust technology. (2) developed software to analyze these screens, a previously formidable challenge, and (3) completed genome-wide experiments on the scale required to complete the goals of this proposal. We will repeat these experiments under several experimental conditions in order t identify genes involved in the TSC pathway.

15. SUBJECT TERMS
Genome-wide, RNA interference, TSC

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	28	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction

In a patient with Tuberous Sclerosis Complex (TSC), the problematic cells that initiate and constitute tumors have lost TSC1 or TSC2 function. A promising approach to develop an effective small-molecule drug for TSC would be to target members of the pathway with which TSC1/2 proteins interact. Such a drug could either bypass the requirement for TSC1/2 or specifically kill or arrest cells that have lost TSC function. Unfortunately, very little is known about the proteins in this pathway or how they interact, so we cannot make an educated guess about which proteins should be targeted to kill, arrest, or revert TSC mutant cells. We are aware of significant gaps in our knowledge of the proteins in this pathway and how they interact[4]. Even for those proteins which are already known to be in the pathway[5], predicting the effects on living cells of targeting those proteins is unreliable. Clearly, we need to better understand this pathway in order to more rationally design drugs to treat TSC. These patients could likely be successfully treated with small molecule drugs if drug targets could be identified which cause only cells deficient in TSC1 or TSC2 function to arrest, die, or revert to normal without disrupting the patient's normal remaining cells. In cultured drosophila cells, we proposed to rapidly identify genes whose RNAi-mediated reduction in expression (1) Prevents growth/proliferation of TSC1 or TSC2-deficient cells without affecting normal cells. (2) Induces apoptosis/cell death in TSC1 or TSC2-deficient cells without killing normal cells. (3) Reverts TSC1 or TSC2-deficient cells to a normal phenotype, as determined by measuring a reporter of cell growth pathway activation and cell morphology.

Body

We have made good progress towards the goals outlined in the original research proposal:

Task 1: Prepare for genome-wide screens:

Print the genome-wide RNAi microarrays using our existing libraries of dsRNA

We have developed the dsRNA printing technology and scaled it up to a higher-throughput, more robust format to allow genome-wide screens. Previously, we had printed a maximum of 384 genes on a single slide[6]. We have now successfully printed ~20,000 dsRNAs, covering the vast majority of Drosophila genes, onto four glass microscope slides. This printing has been developed at a feature density which allows 5600 genes to be spotted per slide.

We overcame technical difficulties relating to the unwanted spreading of dsRNA in the spots and the imprecision of the DNA microarraying robot. We have adapted the technology to work with a new cell type, Drosophila S2 cells expressing an adapted S6 reporter protein. This cell type is required for the aims of this proposal. We also were able to train new personnel to successfully carry out the experiments from printing, to cell seeding to image acquisition and analysis. The transferability of the techniques to new personnel indicates that this technology is robust enough to also be transferred to other laboratories.

Optimize software for these cells and these assays and write code to improve data extraction

Image cytometry (automated cell image analysis) simultaneously measures many valuable features of cells: the intensity, texture and localization of each fluorescently labeled cellular component (e.g. DNA or protein) within each subcellular compartment, as well as the number, size, and shape of each subcellular compartment. This type of full phenotypic analysis is necessary for our aim of identifying reversion to a normal phenotype. Drosophila cells were notoriously difficult to identify in images[7] using existing software. In addition, our project required the accurate measurement of a large number of cellular features, many of which were not measurable using commercial software.

Our laboratory therefore initiated an open-source software project, CellProfiler, to address these substantial challenges (Figure 1 and 2). CellProfiler allows accurate quantitative measurement of many cellular properties, including cell count, cell size, cell cycle distribution, organelle size, and the levels and localization of proteins and phospho-proteins. The software is user-friendly, flexible, modular, open-source, and free, making it a useful tool to share, compare, test, adapt, and further develop image analysis methods in the scientific community.

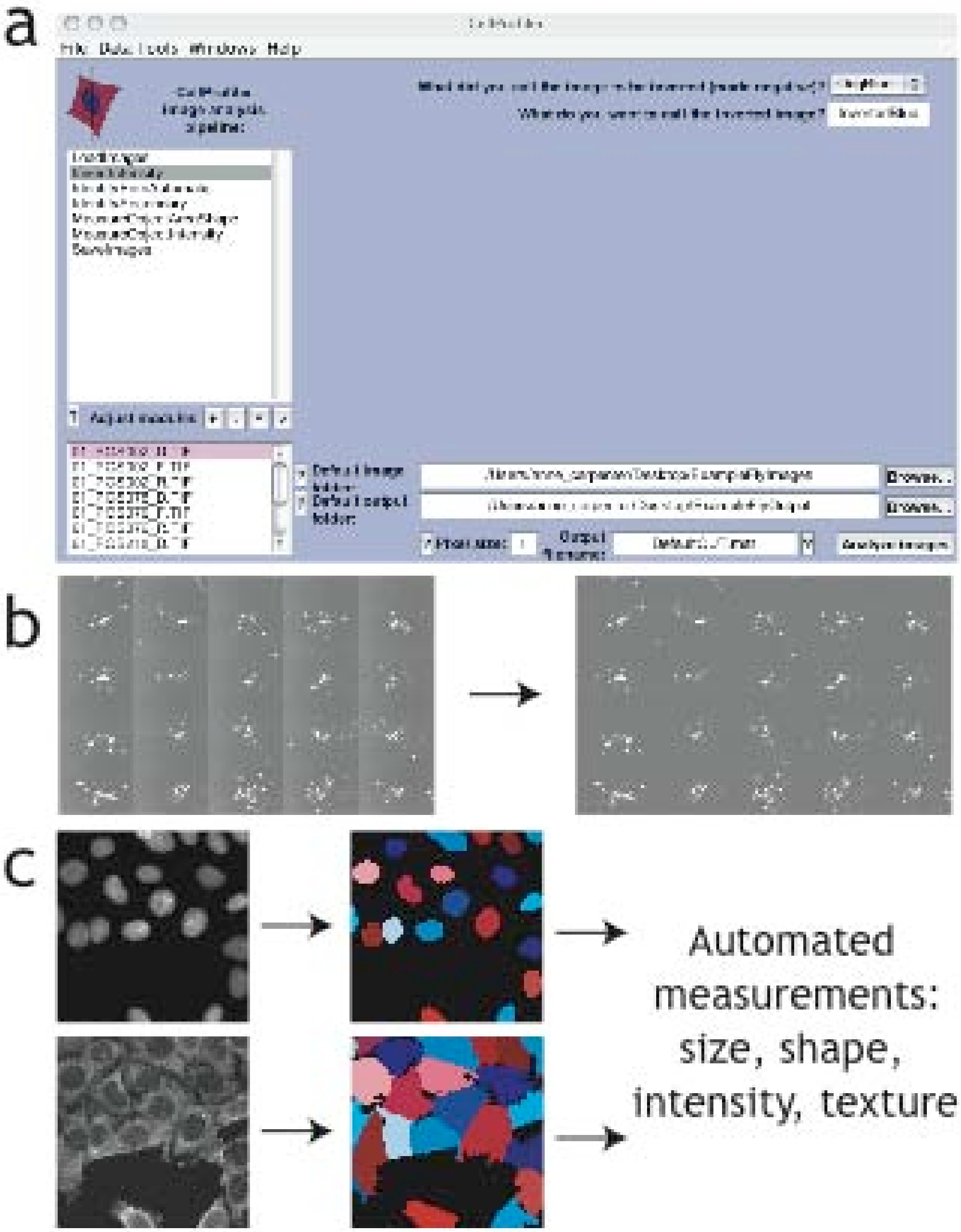


Figure 1: a. Main CellProfiler interface, with a simple analysis pipeline displayed. b. Image processing example: Uneven illumination of the field of view is noticeable when nearby images are placed adjacent to each other and can significantly affect intensity measurements from uncorrected images (left). The left side of each image in this 3 row, 5 column tiled image is brighter than the right side. CellProfiler's CorrectIllumination_Calculate and CorrectIllumination_Apply modules correct these anomalies (right). Images were brightness- and contrast-enhanced to display this effect. c. In typical usage, CellProfiler identifies nuclei first and cell edges are identified surrounding each nucleus using a cell stain image. Measurements are then made.

During this project period, CellProfiler was adapted and optimized for the research projects proposed in several ways:

1. Algorithms for the successful identification of Drosophila nuclei and cells in images were developed and validated[8].
2. We added the ability to measure a large number of sophisticated measurements to CellProfiler, including many measures of the size, shape, intensity, and texture of cells.
3. CellProfiler was validated for many key phenotypes (examples, Figure 2). We tested its ability to measure cell number, cell size, cell cycle distribution (based on DNA content), and amount of fluorescence per cell using cells with known variations in these features.
4. The user interface of CellProfiler was improved dramatically to eliminate some of the tedium of high-throughput image analysis and to allow non-experts the ability to conduct image analysis experiments, including documentation and a manual.
5. CellProfiler was adapted to make use of a cluster of computers, so that it can analyze images at a pace faster than image acquisition.
6. CellProfiler was adapted to export measurements to Excel and also to a database, a necessary feature for large-scale genome-wide experiments.
7. CellProfiler was beta-tested by several academic and industry research groups.
8. CellProfiler was released for free to the public, allowing further development by the open-source community.

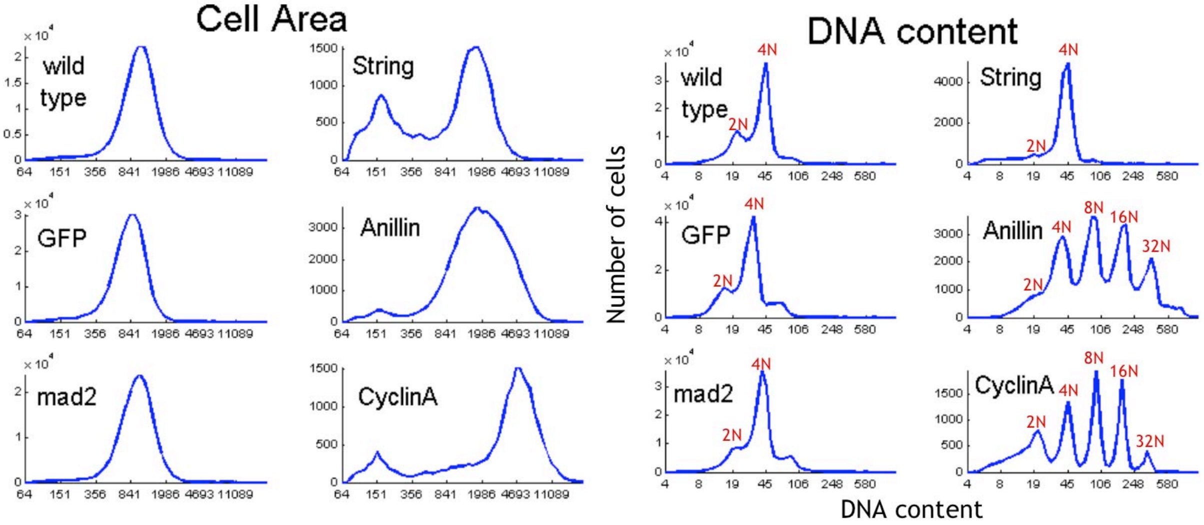
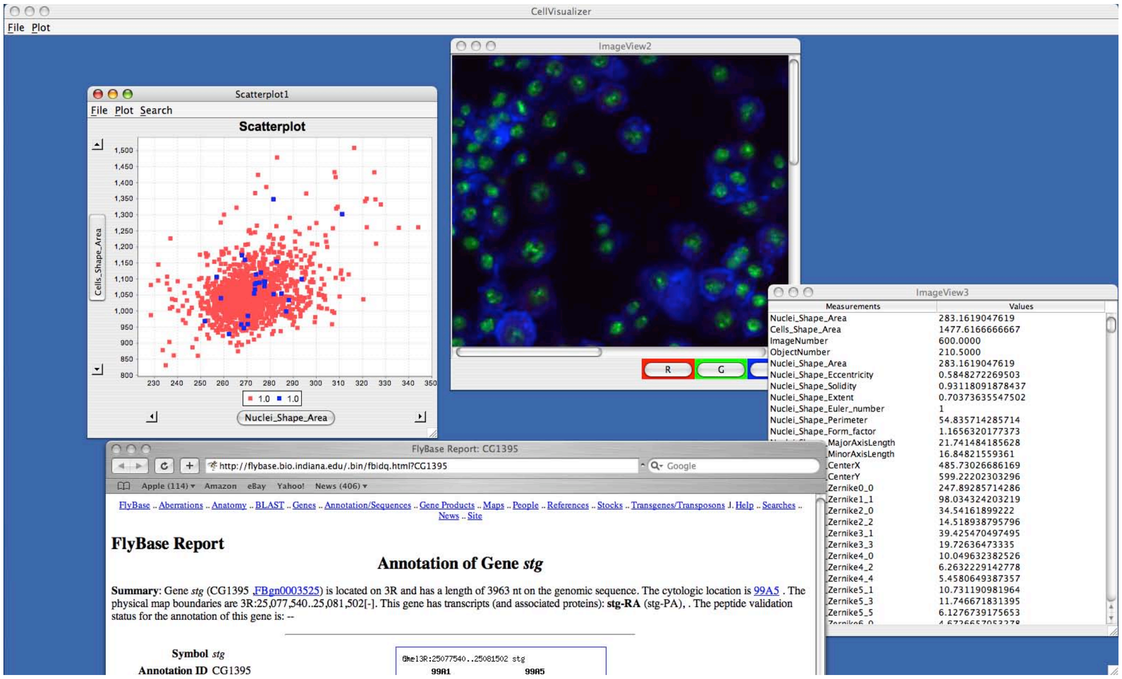


Figure 2: Images from slides with the indicated gene knocked down by RNA interference were analyzed using CellProfiler software to create histograms of cell area (left) and DNA content (right). The results are consistent with existing literature, validating the software.

In addition, we initiated another software project, CellVisualizer, to allow visualization and extraction of the measurements made by CellProfiler (Figure 3). Because the number and richness of measurements from image-based genome-wide assays are unprecedented, other fields must advance to accommodate this influx of data. A full phenotypic analysis of 20,000 images, each containing about 400 cells, includes ~8 million cells and produces roughly three billion measurements (400 measures per cell, including size, shape, and the intensity and texture of three fluorescent stains). CellVisualizer allows biologists without database experience to analyze data from genome-wide screens, including identifying unusual samples based on their quantitative measurements and viewing the original images corresponding to those samples.

We have discovered that these developments in image analysis and data visualization have been critical for our ability to accurately identify genes of interest in the large-scale assays relating to the aims of this proposal.



Measurements	Values
Nuclei_Shape_Area	283.1619047619
Cells_Shape_Area	1477.6166666667
ImageNumber	600.0000
ObjectNumber	210.5000
Nuclei_Shape_Area	283.1619047619
Nuclei_Shape_Eccentricity	0.5848272269503
Nuclei_Shape_Solidity	0.93118091878437
Nuclei_Shape_Extent	0.70373635547502
Nuclei_Shape_Euler_number	1
Nuclei_Shape_Perimeter	54.835714285714
Nuclei_Shape_Form_factor	1.1656320177373
_MajorAxisLength	21.741484185628
_MinorAxisLength	16.84821559361
_CenterX	485.73026686169
_CenterY	599.22202303296
_Zernike0_0	247.89285714286
_Zernike1_1	98.034324203219
_Zernike2_0	34.54161899222
_Zernike2_2	14.518938795796
_Zernike3_1	39.425470497495
_Zernike3_3	19.72636473335
_Zernike4_0	10.049632382526
_Zernike4_2	6.2632229142778
_Zernike4_4	5.4580649387357
_Zernike5_1	10.731190981964
_Zernike5_3	11.746671831395
_Zernike5_5	6.1276739175653

http://flybase.bio.indiana.edu/.bin/fbidq.html?CG1395

FlyBase Report

Annotation of Gene *stg*

Summary: Gene *stg* (CG1395 ,FBgn0003525) is located on 3R and has a length of 3963 nt on the genomic sequence. The cytologic location is 99A5 . The physical map boundaries are 3R:25,077,540..25,081,502[-]. This gene has transcripts (and associated proteins): **stg-RA** (stg-PA), . The peptide validation status for the annotation of this gene is: --

Symbol *stg*
Annotation ID CG1395

Dmel3R:25077540..25081502 stg
99A1 99A5

Figure 3: We developed CellVisualizer, shown here, to allow visualization of data from genome-wide screens. Genes are shown in a scatterplot based on measurements of cells from images with that gene knocked down (top left). Clicking a gene in the scatterplot allows viewing the raw, original image (top middle), the numerical measurements from that image (right) and links to the external gene database FlyBase (bottom).

Task 2: Conduct genome-wide screens:

<u>Prepare RNAi/cell microarrays in duplicate (Total = 24 microscope slides)</u>

During this project period, we have successfully printed, seeded, fixed, stained, imaged, and analyzed genome-wide cell microarrays. So far, we have conducted the following genome-wide experiments:

1. wild type drosophila Kc167 cells treated with dsRNA against nearly every gene in the Drosophila genome (quadruplicate).
2. drosophila S2 cells expressing a modified S6 reporter protein (a reporter of a branch of the cell growth pathway), stained for phospho-S6 (duplicate + rapamycin, duplicate - rapamycin). Note: RNA interference was not very effective on this set of slides. We have since improved the efficacy of RNAi for this cell type and therefore plan to repeat this experiment to obtain higher quality data.

To complete the work described in the research proposal, we must repeat these experiments under two more conditions: in the presence of dTSC1 and dTSC2 RNA interference reagent. Given the substantial progress made in the past year on technology development, these experiments should be feasible in the upcoming months. We have been granted an extension through July 7th, 2007 to accomplish this work.

ADDITIONAL UPDATE, revised report:

As of August 2006, we have taken a step back to smaller-scale experiments in an attempt to work out these techniques more robustly. We have now completed some of the original project goals:

1. We have completed screens covering all the kinases and phosphatases (288 genes) in Kc167 cells that are:
   - Wild type
   - TSC1 knockdown
   - TSC2 knockdown

This data has been mined and hits are being tested in follow-up experiments to confirm the original phenotypes, other signaling pathway outcomes, and interactions with TSC1 and 2.

2. The first genome-wide experiments like (1) have been imaged for wild type and will soon be replicated (with TSC1 and TSC2 conditions), mined and followed up.

3. The first kinase and phosphatase screens in S6 wild-type cells have been prepared and imaged. When robust, these will be performed at the genome-wide scale (with TSC1 and TSC2 conditions), imaged, mined, and followed up. Some preliminary hits from the wild-type condition alone are currently being followed up.

Persons paid from this grant:

David M. Sabatini
Michael Lamprecht
Ola Friman

Key Research Accomplishments

- The dsRNA printing technology was developed, improved, and scaled up to high-throughput such that we were able to print ~20,000 dsRNAs, covering the vast majority of Drosophila genes, onto four glass microscope slides. This was the proof of principle needed to demonstrate that genome-wide screens are feasible using this method.
  - In particular, technical difficulties related to the unwanted spreading of dsRNA in the spots and the imprecision of the DNA microarraying robot were resolved.
- We determined that software we wrote, CellProfiler, produces accurate quantitative cell measurements and is another key enabling technology now in place to analyze genome-wide screens.
  - Algorithms for the successful identification of Drosophila nuclei and cells in images were developed and validated.
  - We added the ability to measure a large number of sophisticated measurements to CellProfiler, including many measures of the size, shape, intensity, and texture of cells.
  - CellProfiler was validated for many key phenotypes. We tested its ability to measure cell number, cell size, cell cycle distribution (based on DNA content), and amount of fluorescence per cell using cells with known variations in these features.
  - The user interface of CellProfiler was improved dramatically to eliminate some of the tedium of high-throughput image analysis and to allow non-experts the ability to conduct image analysis experiments, including documentation and a manual.
  - CellProfiler was adapted to make use of a cluster of computers, so that it can analyze images at a pace faster than image acquisition.
  - CellProfiler was adapted to export measurements to Excel and also to a database, a necessary feature for large-scale genome-wide experiments.
  - CellProfiler was beta-tested by several academic and industry research groups.
  - CellProfiler was released for free to the public, allowing further development by the open-source community.
- We determined that software we wrote, CellVisualizer, allows visualization and extraction of the measurements made by CellProfiler so that genome-wide screens can be rapidly analyzed and conclusions drawn. This is another necessary component now in place to analyze genome-wide screens.
- We were able to complete genome-wide screens in wild type cells and determined that, barring technological/reproducibility challenges, we will be able to complete the genome-wide screens in the proposal.

Reportable Outcomes

Completed Manuscripts/Abstracts/Publications:

- Lamprecht MR, Sabatini DM, Carpenter AE (in press) CellProfiler: free, versatile software for automated biological image analysis. BioTechniques. Attached as an appendix.
- Jones TR, Carpenter AE, Golland P, Sabatini DM (in press) Methods for high-content, high-throughput image-based cell screening. MIAAB 2006 Workshop Proceedings. Preprint available at www.cellprofiler.org/papers.htm and attached as an appendix.

Manuscripts/Abstracts/Publications in progress:

- Anne E. Carpenter*, Thouis Ray Jones, Douglas B. Wheeler, Michael Lamprecht, Colin Clarke, Ola Friman, David A. Guertin, In Han Kang, Robert Lindquist, Joo Han Chang, Jason Moffat, Polina Golland, and David M. Sabatini, CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes, submitted.

Open-source software released to the public, funded in part by this grant:

- CellProfiler cell image analysis software (www.cellprofiler.org)

Presentations discussing this work:

David M. Sabatini (principal investigator):

- (talk) Center for Cancer Research, Mass General Hospital, Feb. 2005, Boston, MA
- (talk) Neuroscience Monday Seminars, March 2005, Children's Hospital, Boston, MA
- (talk) TargetTalk Meeting, March 2005, San Diego, CA
- (talk) Department of Pharmacology, University of Virginia, April 2005, Charlottesville, VA
- (talk) Pfizer Research Technology Center, June 2005, Cambridge, MA
- (talk) American Diabetes Association, June 2005, San Diego, CA
- (talk) GlaxoSmithKline, July 2005, Philadelphia, PA
- (talk) Cancer Models & Mechanisms Gordon Research Conf., July 2005, Smithfield, RI
- (talk) Protein Kinases and Protein P'n FASEB Meeting, July 2005, Snowmass, CO
- (talk) Glucose Transporter Biology FASEB Meeting, August 2005, Snowmass, CO
- (talk) Ariad Pharmaceuticals, August 2005, Cambridge, MA

Anne E. Carpenter (postdoctoral fellow):

- (talk) Harvard Department of Systems Biology, November 2005, Boston, MA
- (talk) Cytometry Development Workshop, Asilomar, October 2005, Pacific Grove, CA
- (talk) Merck Automated Biotechnology group, October 2005, North Wales, PA
- (talk) MipTec Enabling Technologies for Drug Discovery, May 2005, Basel, Switzerland
- (talk) Roche, May 2005, Nutley, NJ
- (poster) Life Sciences Research Foundation Annual Meeting, Oct. 2005, Wash., DC
- (poster) Discovery on Target, October 2005, Boston, MA
- (poster) Whitehead Institute Annual Retreat, September 2005, Waterville Valley, NH

- (poster) Society Biomolecular Screening Annual Mtg, Sept. 2005, Geneva, Switzerland

Colin Clarke (undergraduate student):

- (talk) American Society for Cell Biologists Annual Mtg, Dec., 2005, San Francisco, CA

Funding applied for based on this work:

- Culpeper Biomedical Pilot Initiative grant

Conclusions

In the past year, Drosophila genome-wide RNA interference living cell microarrays have gone from a proof-of-principle concept to a robust technology. We have also developed and adapted a number of software tools to allow the analysis of these genome-wide image-based screens, a previously formidable challenge. We have completed genome-wide experiments on the scale required to complete the goals of this proposal. Future work remaining to complete these goals are to repeat these genome-wide experiments under several more experimental conditions in order to identify genes involved in the TSC pathway.

References

1. Krymskaya, V. P. Tumour suppressors hamartin and tuberin: intracellular signalling. Cell Signal 15, 729-39 (2003).
2. Tucker, T. & Friedman, J. M. Pathogenesis of hereditary tumors: beyond the "two-hit" hypothesis. Clin Genet 62, 345-57 (2002).
3. Narayanan, V. Tuberous sclerosis complex: genetics to pathogenesis. Pediatr Neurol 29, 404-9 (2003).
4. Hall, M. N., Raff, M. & Thomas, G. Cell Growth: Control of Cell Size (Cold Spring Harbor Laboratory, 2004).
5. Li, Y., Corradetti, M. N., Inoki, K. & Guan, K. L. TSC2: filling the GAP in the mTOR signaling pathway. Trends Biochem Sci 29, 32-8 (2004).
6. Wheeler, D. B. et al. RNAi living-cell microarrays for loss-of-function screens in Drosophila melanogaster cells. Nat Methods 1, 127-32 (2004).
7. Armknecht, S. et al. High-throughput RNA interference screens in Drosophila tissue culture cells. Methods Enzymol 392, 55-73 (2005).
8. Jones, T. R., Carpenter, A. E. & Golland, P. Voronoi-based segmentation of cells on image manifolds. ICCV Workshop on Computer Vision for Biomedical Image Applications, 535-543 (2005).

CellProfiler™: free, versatile software for automated biological image analysis

Michael R. Lamprecht[1], David M. Sabatini[1,2], and Anne E. Carpenter[1]

[1]*Whitehead Institute for Biomedical Research, Cambridge, and* [2]*Massachusetts Institute of Technology, Cambridge, MA, USA*

BioTechniques 41:XXX-XXX (October 2006)
doi 10.2144/000112257

Careful visual examination of biological samples is quite powerful, but many visual analysis tasks done in the laboratory are repetitive, tedious, and subjective. Here we describe the use of the open-source software, CellProfiler™, to automatically identify and measure a variety of biological objects in images. The applications demonstrated here include yeast colony counting and classifying, cell microarray annotation, yeast patch assays, mouse tumor quantification, wound healing assays, and tissue topology measurement. The software automatically identifies objects in digital images, counts them, and records a full spectrum of measurements for each object, including location within the image, size, shape, color intensity, degree of correlation between colors, texture (smoothness), and number of neighbors. Small numbers of images can be processed automatically on a personal computer, and hundreds of thousands can be analyzed using a computing cluster. This free, easy-to-use software enables biologists to comprehensively and quantitatively address many questions that previously would have required custom programming, thereby facilitating discovery in a variety of biological fields of study.

INTRODUCTION

One of the most powerful methods in biology is the visual analysis of a sample. While nothing can fully replace the expertise of a trained biologist, observing many samples by eye is time-consuming, subjective, and nonquantitative. Certain repetitive tasks in visual analysis are suitable for automation by collecting digital images and processing them with image analysis software. This liberates biologists for more interesting work and has several advantages over visual observations including speed, quantitative and reproducible results, and simultaneous measurement of many features in the image. Efforts to automate visual analysis in biology began several decades ago, but many aspects still need improvement (1).

While numerous commercial and free software packages exist for image analysis, many of these packages are designed for a very specific purpose, such as cell counting (2). Other packages are sold with accompanying hardware for image acquisition (e.g., yeast colony counters), but these are expensive and do not allow measurement of features beyond those that are already built-in. Most commercial software is proprietary, meaning that the underlying methods of analysis are hidden from the researcher. At the other end of the continuum, some software packages are very flexible, especially for interactive analysis of individual images [e.g., Image-Pro Plus, MetaMorph®, and the open-source ImageJ/National Institutes of Health (NIH) Image (3)]. While users can program custom algorithms or record macros, these customized routines are challenging to adapt without knowing a programming language or interacting directly with the macro code.

The CellProfiler™ project was developed to address these software challenges by providing the scientific community with an easy-to-use open-source platform for automated image analysis. The compiled software is freely available for Macintosh®, PC, and Unix platforms at www.cellprofiler.org. It can accommodate adaptation to many biological objects and assays without requiring programming, due to its modular design and graphical user interface. There are many existing software packages available for specific applications in biology, but CellProfiler accomplishes many of the same goals

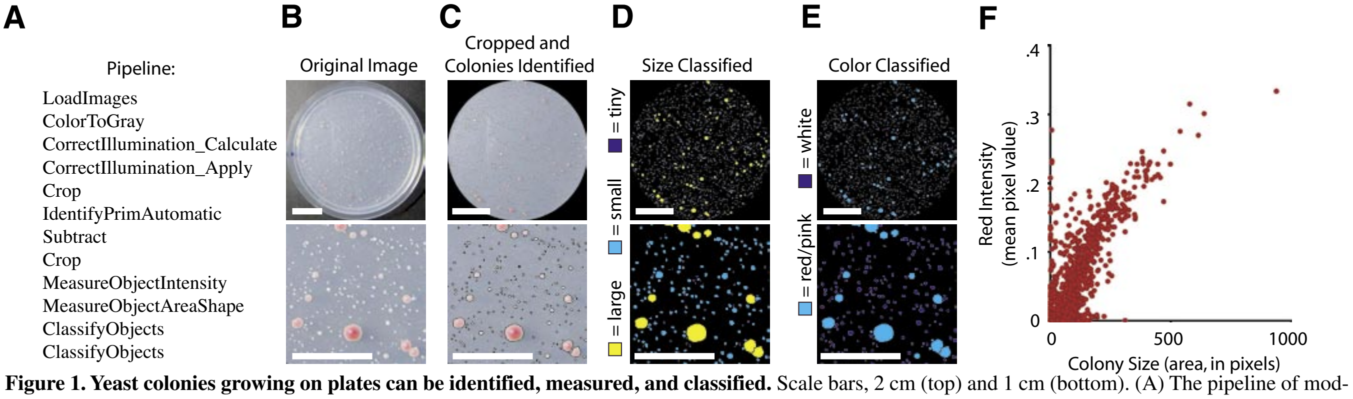


Figure 1. Yeast colonies growing on plates can be identified, measured, and classified. Scale bars, 2 cm (top) and 1 cm (bottom). (A) The pipeline of modules used for this analysis. (B) Original image of yeast colonies growing on an agar plate. (C) Image after automatic cropping of the plate and colony identification by CellProfiler, with individual colonies outlined in black. (D) Image with the identified colonies classified by size (see legend for color-coding). (E) Same as panel D, but classified by red intensity rather than size. (F) The red intensity (vertical axis) and size (horizontal axis) of each colony on the plate is plotted, revealing the relationship between these measurements.

in one open-source program. We recently described CellProfiler's use for cell identification, cell size, intensity and texture of fluorescent stains, cell cycle distributions, and other features of individual cells in images (4). Here we describe its use for a variety of other applications such as yeast colony counting, grid analysis, wound healing, and other visually quantifiable assays.

MATERIALS AND METHODS

All of the image analysis in this paper used the freely available CellProfiler cell image analysis software. The pipelines and images for these examples, as well as others, are available for download (www.cellprofiler.org/examples.htm). The image of yeast colonies (Figure 1) is a plate of Hi90-strain cells plated on synthetic defined medium with 128 ⊠g/mL fluconazole as previously described (5). Images of *Drosophila* Kc167 cells on cell microarrays (Figure 2, A–C) were prepared as described previously (4,6). Briefly, spots of double-stranded RNA (dsRNA) were printed onto plain slides, and cells were grown on these slides for 3 days before being fixed, stained with Hoechst 33342, and imaged. Images of yeast patches (Figure 2, D–G) were prepared by manually spotting cells (with a 96-well pinning device) onto agar plates containing galactose to induce the expression of α-synuclein and a gene of interest. The cells were grown for 2 days at 30°C prior to imaging (Aaron Gitler, personal communication). Images of green fluorescent protein (GFP)-labeled mouse tumors (Figure 3, A–C) are faces of a mouse lung lobe, dissected out at 8 weeks post-tail vein injection of an established metastatic human cancer cell line overexpressing a gene of interest as described (7). Images of wound healing (Figure 3, E–G) were prepared using MDA-MB-435 cells and imaged at the time points indicated (Lynne Waldman, personal communication). A *Drosophila* wing imaginal disc from a third larval instar (Figure 3, H–J) was stained with rhodamine-phalloidin to label F-actin, which is concentrated at points of cell-cell contact at the level of the adherens junction lattice (Matt Gibson, personal communication).

RESULTS AND DISCUSSION

CellProfiler's main window allows the user to point and click to do most tasks, including the design of a new assay. The software uses the concept of a pipeline, which is a series of modules. Each module performs a specific task on the image or on identified objects (Figure 1A). A typical pipeline consists of loading the images, correcting for uneven illumination, identifying the objects, and then taking measurements on those objects. These modules can easily be added, removed, or rearranged within a pipeline. The resulting measurements can be viewed by (*i*) using CellProfiler's built-in viewing and plotting data tools; (*ii*) exporting in a tab-delimited spreadsheet format that can be opened in programs like Microsoft® Excel® and OpenOffice.org Calc; (*iii*) exporting in a format that can be imported into a database like Oracle® or MySQL® (MySQL, Cupertino, CA, USA); or (*iv*) opening in MATLAB® (Mathworks, Natick, MA, USA). An analysis can be done on one specific image, a group of images, or thousands of images by using a computing cluster.

CellProfiler bridges the gap between powerful computational methods and their practical application in the biological laboratory. Computer scientists can prototype new computational methods and contribute them to the project, and then biologists can easily use these new additions in their work. Further, the functionality of existing modules can be enhanced by researchers with some programming experience, because the code is open-source, well-documented, and in a language that is relatively easy to understand. While most users will download the completely free, compiled version of CellProfiler, the CellProfiler Developer's version requires the software package MATLAB and its image processing toolbox.

As described in the manual, available at www.cellprofiler.org/linked_files/CellProfilerManual.pdf, CellProfiler already contains advanced object identification algorithms from

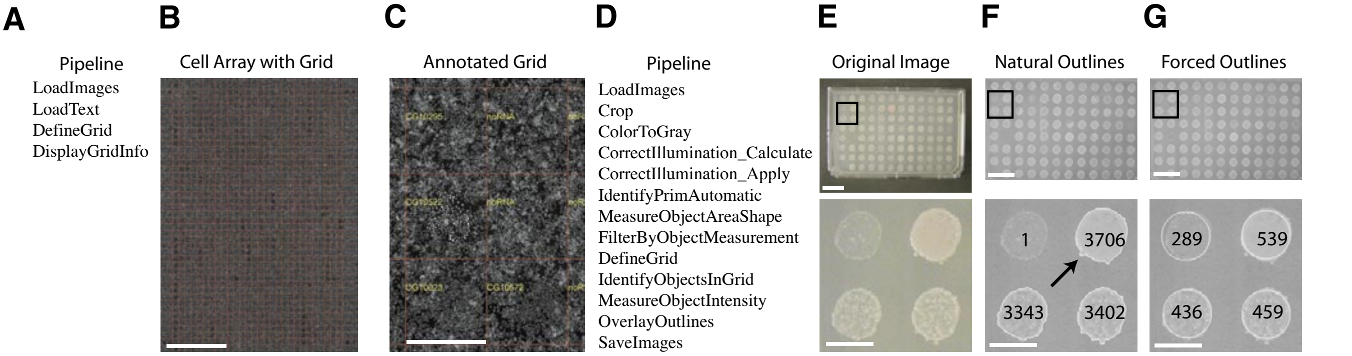


Figure 2. Grids of samples can be annotated and analyzed. (A) The pipeline of modules used for the analysis shown in panels B and C. (B) A living cell RNA interference microarray with 40 rows and 28 columns of spots, stained for DNA. Each spot contains a double-stranded RNA (dsRNA) that knocks down a particular gene. The grid placed on the image by CellProfiler is shown as red lines. Scale bar, 4.5 mm. (C) Enlarged portion of panel B, with the annotations placed by CellProfiler shown in yellow. noRNA is the control. Scale bar, 450 ⊠m. (D) The pipeline of modules used for the analysis shown in panels E–G. (E) Original image of yeast patches growing in a grid with 8 rows and 12 columns. Box indicates the region shown enlarged below. (F) Image showing the patches' natural outlines determined by CellProfiler, including wiggly protrusions (arrow). The measured area of each patch is shown numerically on top of each patch, in pixels. (G) Image showing outlines of patches that were forced into a standard circular shape to measure the amount of growth in each patch, using intensity units. Scale bars for panel E–G: top, 20 mm; bottom, 5 mm.

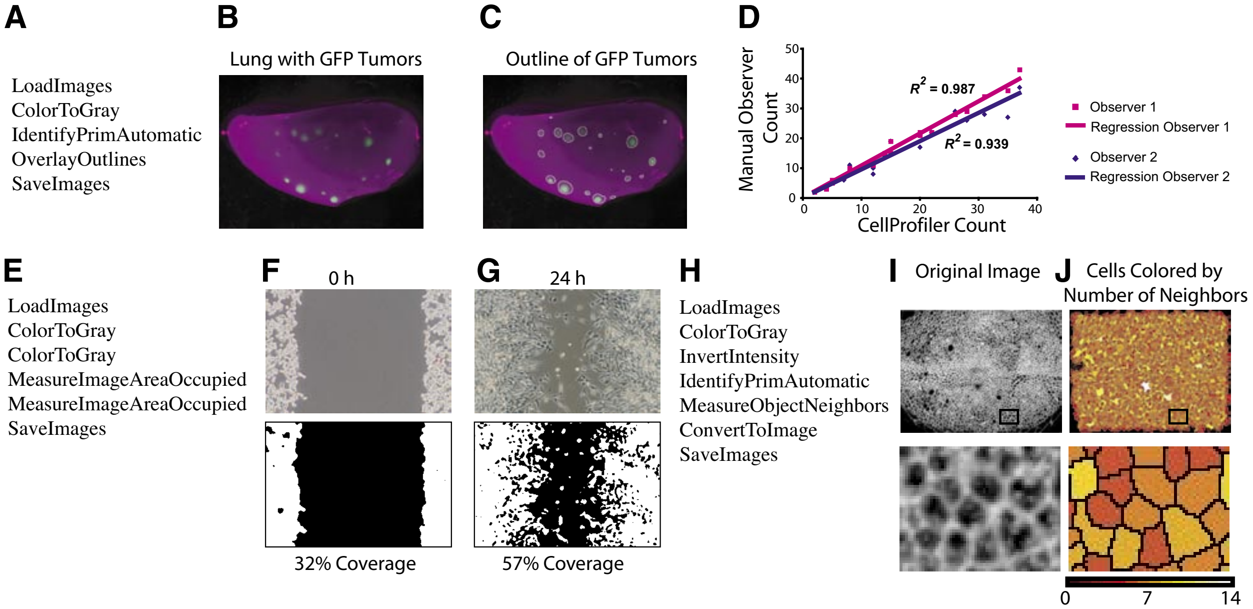


Figure 3. Identification and measurement of green fluorescent protein (GFP)-labeled tumors in mouse lungs, the wound healing assay, and *Drosophila* tissue topology. (A) The pipeline of modules used for the analysis shown in panels B–D. (B) Original image. (C) Image with tumors outlined by CellProfiler. (D) Tumors in a set of 20 images were counted by CellProfiler and by two researchers. The manual tumor count for each image (vertical axis) is plotted versus the CellProfiler count (horizontal axis), revealing strong concordance (R^2 value is shown). (E) The pipeline of modules used for the analysis shown in panels F and G. (F) At time point zero, the wound visible in the original image (top) is large and the cells present at the edges of the image cover a small percentage of the area of the image, as quantified by CellProfiler (bottom). (G) After 24 h, the wound has recovered due to cells migrating from the edges (top) and now is much smaller in measured size (bottom). (H) The pipeline of modules used for the analysis shown in panels I and J. (I) Original image of *Drosophila* epithelial cells growing in a sheet. Box indicates the region shown enlarged below. (J) Image showing cells identified and color-coded by CellProfiler based on how many neighbors they have. Box indicates the region shown enlarged below.

the literature (4,8–15) and is open to adding new algorithms as described above. In object identification modules, users can rapidly select the best solution for their application using a Test Mode to see the results of various methods. In the following examples, we show the identification of objects by CellProfiler and select measurements for each. Note that the full spectrum of measurements, including many not often measured by biologists (16–18), can be recorded for each identified object, including location within the image, size, shape, color intensity, texture (smoothness), correlation between colors, and number of neighbors. Moreover, each broad category contains many different specific measurements. For example, size includes area, perimeter, and major/minor axis length, and shape includes eccentricity (elongation), solidity, form factor, and 32 other shape-related measures.

Yeast Colonies

Counting colonies on agar plates and classifying them by size, color, or shape is tedious, time-consuming, and subjective. While complete systems for automated colony counting exist, they are more expensive and less flexible than using a digital camera or off-the-shelf flatbed scanner to acquire images for analysis by CellProfiler. The cost of this solution can be less than $100. Furthermore, the algorithms in CellProfiler are accurate and adaptable, and unusual features of colonies, which commercial software and even the human eye cannot detect, can be measured (e.g., certain measures of texture and shape). After the initial analysis strategy has been established, plates can be analyzed automatically in large batches.

Here we show an example of yeast colonies (*Saccharomyces cerevisiae*) that were analyzed by CellProfiler (Figure 1B). In this analysis, the plates are automatically cropped to remove the edges, and individual colonies are identified, even when clumped (Figure 1C). Measurement modules then calculate measurements of interest for each individual colony. Any of the available measurements can then be used to classify the colonies, for example, by size (Figure 1D), by color (Figure 1E), or by a combination of measurements, such as size and color. In this example, the apparent correlation between size-classified (Figure 1D) and color-classified (Figure 1E) yeast colonies is verified by a scatter plot of these two measurements (Figure 1F). Each class of colonies can be analyzed separately to allow the researcher to focus on classes of interest. This allows for addressing questions like “Are the red colonies larger than white colonies?” or “Do the larger colonies have more irregularly-shaped borders?” In this example, the colonies all display a smooth round phenotype, but the colony shape and texture of yeast strains with unusual morphology could be quantified using these methods.

Grid Analysis

Many experiments are designed in a grid format, such as cell microarrays, agar plates of yeast patches, and multiwell plates. In experiments with large or numerous grids, it is difficult to identify which reagent corresponds to a spot on this grid. CellProfiler can manually or automatically place a grid on an image and associate each grid location with an annotation such as a sample number or gene name. Each grid location can also be identified as an object, and all available measurements can be made on those objects.

For example, on a cell microarray, more than 5000 individual clusters of cells are grown on a single microscope slide (4,6,19–21). Each cluster has been treated with a perturbant, which could be a small molecule, an overexpression plasmid, or an RNA interference reagent. The goal is to determine which perturbants alter cells. When attempting to quantify even a simple phenotype such as the extent of cell death in a spot, it is difficult to determine which spot correlates to which reagent. In this example, CellProfiler places a grid on top of an image (Figure 2B) in a position defined by the user, who specifies the location of known control spots on the array and the number and size of rows and columns. The LoadText module allows the user to load a text file with corresponding sample information for each of the spots on the grid, and the DisplayGridInfo module allows this imported data to be assigned to each of the grid locations (Figure 2C). This quickly allows the user to associate a location in the grid with its reagent. For example, knockdown of the cytokinesis-related gene CG10522 (*sticky*) shows an unusual large, bright-nuclei phenotype that is visible at low resolution (Figure 2C).

Plates of yeast patches also make use of grids (Figure 2E). Although large screens of yeast patches have been analyzed by eye (22), the number of these screens in progress is rapidly increasing, and visual analysis cannot keep up with the rapid pace at which samples are being generated. Furthermore, quantitative analysis is much preferred, because subtle changes in growth can be identified and the screen can be analyzed statistically. Some software for this application has been developed, but none to our knowledge is fully automated, open-source, and flexible to new assays/unusual measurements.

Because thousands of plates are typically analyzed, the entire process of finding the grid and making measurements is performed automatically by CellProfiler. In the pipeline for this analysis, the images are cropped to remove the plate edges, and any yeast patches that are present are identified. Unlike the cell microarray example, a grid is then defined automatically based on the yeast patches that are identified. This allows for nonuniformity in the precise placement of the grid on the plate, to allow for experimental variation. For this function to work correctly, none of the outside rows or columns can be completely blank. This condition can be satisfied if most patches tend to grow well in the experiment or if control patches exist in two or more opposite corners. These yeast patches can be analyzed in their naturally identified shapes if the patch size or shape is of interest (Figure 2F). Alternately, a circle can be forced into the location of the identified objects to measure, for example, the intensity of each patch, which is a measure of growth (Figure 2G). Once the grid and objects are identified, all available measurements can be calculated for each patch, including measures of growth (Figure 2, F–G, bottom).

Tumor Counting

When tumorigenic cells are labeled with GFP and injected into mice, the resulting tumors in the lungs are readily visible by fluorescence microscopy of the dissected lungs. Accurate, objective quantification of the number and size of the resulting tumors is necessary to understand the process of tumor metastasis (7). The GFP signal from each tumor can be identified by CellProfiler (Figure 3, A–C). CellProfiler counts of identified tumors were comparable to counting by eye (Figure 3D). If the GFP brightness or the shape or texture/smoothness of the tumors is of interest, these measurements can also be recorded.

Wound Healing

The wound healing assay is a standard technique to determine the migration of different cell types in different conditions. In this assay, a confluent monolayer of cells is wounded by scratching it with a pipet tip (23). The monolayer is then imaged at time points to record the size of the wound. In this example, the area of the images covered by cells is calculated by CellProfiler (Figure 3, E–G). While this is not a particularly challenging application, the structure of CellProfiler makes it simple to carry out this quantitative analysis for hundreds of thousands of images, enabling high-throughput screens. In addition, the shape characteristics of the wound border can be measured; for example, to distinguish between samples where all cells have steadily grown toward the middle versus samples where a few individual cells extend into the wound space.

Tissue Topology

In a developing tissue or at other sites of cell-cell contact (e.g., tumors and surrounding stromal cells), it is useful to determine the number of neighbors each cell has to better understand the processes underlying the topology (24). CellProfiler can identify cells in tissues (Figure 3, H–J). In addition to typical measurements, the MeasureObjectNeighbors module can determine the number of cells neighboring each cell and record this measurement. The cells can then be color-coded by how many neighbors it has (Figure 3J), or the data can be exported to further analyze the topology of the tissue.

CellProfiler is a flexible platform that can automate the analysis of images to address a wide variety of biological questions. For many assays, described here and previously (4), it eliminates the tedium of repetitive visual analysis and produces rapid, quantitative, and accurate results. The modular design of the software provides an infrastructure for image analysis that is applicable to diverse assays. Its open-source code allows programmers to design and contribute new algorithms to the project. It is our hope that CellProfiler will become a widely used platform, through which advanced algorithms are made conveniently available for automatic biological image analysis.

ACKNOWLEDGMENTS

We gratefully acknowledge contributions of images from researchers at the Harvard Medical School: Matt Gibson (Figure 3, H–J) and at the Whitehead Institute for Biomedical Research and Massachusetts Institute of Technology: Leah Cowen (Figure 1), Douglas B. Wheeler (Figure 2, A–C), Aaron Gitler (Figure 2, D–G), Kimberly A. Hartwell (Figure 3, A–D), and Lynne K. Waldman (Figure 3, E–G). We also thank the other members of the CellProfiler project software team for software development and lab members for helpful comments: Robert A. Lindquist, Shomit Sengupta, and David A. Guertin. This work was supported by a Merck/CSBi postdoctoral fellowship (A.E.C.), a Novartis fellowship from the Life Sciences Research Foundation (A.E.C.), a Society for Biomolecular Screening Academic grant (A.E.C.), Department of Defense (DOD) Tuberous Sclerosis Complex (TSC) research program grant no. W81XWH-05-1-0318-DS (D.M.S.), NIH grant no. R01 GM072555-01 (D.M.S.), and the Keck Foundation (D.M.S.).

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. **Murphy, R.F., E. Meijering, and G. Danuser.** 2005. Special issue on molecular and cellular bioimaging. IEEE Trans. Image Process. *14*:1233-1236.
2. **Selinummi, J., J. Seppala, O. Yli-Harja, and J.A. Puhakka.** 2005. Software for quantification of labeled bacteria from digital microscope images by automated image analysis. BioTechniques *39*:859-863.
3. **Abramoff, M.D., P.J. Magalhaes, and S.J. Ram.** 2004. Image processing with ImageJ. Biophotonics Int. *11*:36-42.
4. **Carpenter, A.E., T.R. Jones, D.B. Wheeler, O. Friman, C. Clarke, D.A. Guertin, I.H. Kang, J.H. Chang, et al.** CellProfiler: image analysis for high-throughput microscopy (In preparation).
5. **Cowen, L.E. and S. Lindquist.** 2005. Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. Science *309*:2185-2189.
6. **Wheeler, D.B., S.N. Bailey, D.A. Guertin, A.E. Carpenter, C.O. Higgins, and D.M. Sabatini.** 2004. RNAi living-cell microarrays for loss-of-function screens in *Drosophila melanogaster* cells. Nat. Methods *1*:127-132.
7. **Hartwell, K.H., B. Muir, F. Reinhardt, A.E. Carpenter, D.C. Sgroi, and R.A. Weinberg.** (In preparation).
8. **Jones, T.R., A.E. Carpenter, and P. Golland.** 2005. Voronoi-based segmentation of cells on image manifolds. ICCV Workshop on Computer Vision for Biomedical Image Applications, p. 535-543.
9. **Jones, T.R., A.E. Carpenter, D.M. Sabatini, and P. Golland.** Methods for high-content, high-throughput image-based cell screening (In preparation).
10. **Malpica, N., C.O. de Solorzano, J.J. Vaquero, A. Santos, I. Vallcorba, J.M. Garcia-Sagredo, and F. del Pozo.** 1997. Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry *28*:289-297.
11. **Meyer, F. and S. Beucher.** 1990. Morphological segmentation. J. Vis. Commun. Image Rep. *1*:21-46.
12. **Ortiz de Solorzano, C., E.G. Rodriguez, A. Jones, D. Pinkel, J.W. Gray, D. Sudar, and S.J. Lockett.** 1999. Segmentation of confocal microscope images of cell nuclei in thick tissue sections. J. Microsc. (Oxford) *193*:212-226.
13. **Wahlby, C.** 2003. Algorithms for applied digital image cytometry, p. 75. *In* Center for Image Analysis. Uppsala University, Uppsala.
14. **Wahlby, C., I.M. Sintorn, F. Erlandsson, G. Borgefors, and E. Bengtsson.** 2004. Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. J. Microsc. *215*:67-76.
15. **Vincent, L. and P. Soille.** 1991. Watersheds in digital spaces—an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. *13*:583-598.
16. **Boland, M.V., M.K. Markey, and R.F. Murphy.** 1998. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry *33*:366-375.
17. **Gabor, D.** 1946. Theory of communication. J. Instit. Electr. Engineer. *93*:429-441.
18. **Haralick, R.M., K. Shanmuga, and I. Dinstein.** 1973. Textural features for image classification. IEEE Trans. Syst. Man. Cybern. *SMC3*:610-621.
19. **Bailey, S.N., D.M. Sabatini, and B.R. Stockwell.** 2004. Microarrays of small molecules embedded in biodegradable polymers for use in mammalian cell-based screens. Proc. Natl. Acad. Sci. USA *101*:16144-16149.
20. **Wheeler, D.B., A.E. Carpenter, and D.M. Sabatini.** 2005. Cell microarrays and RNA interference chip away at gene function. Nat. Genet. *37(Suppl)*:S25-S30.
21. **Bailey, S.N., S.M. Ali, A.E. Carpenter, C.O. Higgins, and D.M. Sabatini.** 2006. Microarrays of lentiviruses for gene function screens in immortalized and primary cells. Nat. Methods *3*:117-122.
22. **Measday, V., K. Baetz, J. Guzzo, K. Yuen, T. Kwok, B. Sheikh, H. Ding, R. Ueta, et al.** 2005. Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. Proc. Natl. Acad. Sci. USA *102*:13956-13961.
23. **Yu, A.C., Y.L. Lee, and L.F. Eng.** 1993. Astrogliosis in culture: I. The model and the effect of antisense oligonucleotides on glial fibrillary acidic protein synthesis. J. Neurosci. Res. *34*:295-303.
24. **Classen, A.K., K.I. Anderson, E. Marois, and S. Eaton.** 2005. Hexagonal packing of *Drosophila* wing epithelial cells by the planar cell polarity pathway. Dev. Cell *9*:805-817.

Received 15 May 2006; accepted 7 July 2006.

Address correspondence to Anne E. Carpenter, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA. e-mail: carpenter@wi.mit.edu

To purchase reprints of this article, contact: Reprints@BioTechniques.com

Methods for High-Content, High-Throughput Image-Based Cell Screening

Thouis R. Jones[1,2], Anne E. Carpenter[2], David M. Sabatini[2], Polina Golland[1]

[1]Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, Cambridge, MA 02139, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA

***Abstract*— Visual inspection of cells is a fundamental tool for discovery in biological science. Modern robotic microscopes are able to capture thousands of images from massively parallel experiments such as RNA interference (RNAi) or small-molecule screens. Such screens also benefit from lab automation, making large screens, e.g., genome-scale knockdown experiments, more feasible and common. As such, the bottleneck in large, image-based screens has shifted to visual inspection and scoring by experts.**

In this paper, we describe the methods we have developed for automatic image cytometry. The paper demonstrates illumination normalization, foreground/background separation, cell segmentation, and shows the benefits of using a large number of individual cell measurements when exploring data from high-throughput screens.

I. Introduction

One of the most basic tools of modern biology is visual inspection of cells using a microscope. Modern techniques, such as immunofluorescent staining and robotic microscopes, have only magnified its importance for the elucidation of biological mechanisms. However, visual analysis has also become a major bottleneck in large, image-based screens, where tens to hundreds of thousands of individual cell populations are perturbed (genetically or chemically) and examined to find those populations yielding an interesting phenotype. Several genome-scale screens have relied on visual scoring by experts [1], [2]. There are benefits to manual scoring, such as the ability of a trained biologist to quickly intuit meaning from appearance, the robustness of the human visual system to irrelevant variations in illumination and contrast, as well as humans' ability to deal with the wide variety of phenotypes that cells can present.

However, automatic image cytometry has several advantages over manual scoring: simultaneous capture of a wide variety of measurements for each cell in each image (versus scoring a few features per image), quantitative rather than qualitative scoring, ease of reproducibility, detection of more subtle changes than is possible by eye, and the main benefits, elimination of tedious manual labor and much faster analysis of images.

Several groups have made use of automated cell image analysis [3], [4], [5], [6], [7], demonstrating the efficacy of such an approach. These groups have either made use of expensive and inflexible commercial systems, often bundled with a particular imaging platform, or they have developed their own software, seldom used outside of the original lab because of its specificity to a particular screen. In order to reduce the duplication of effort in this area, and to make tools for automated cell-image analysis more widely available, we have created CellProfiler, an open-source, modular system for cell-image analysis [8], [9].

This paper describes the key algorithms in CellProfiler, and our overall strategies for accomplishing high-throughput image analysis. These include illumination correction to normalize for biases in the illumination and optical path of the microscope, identification of cells versus background, segmentation of individual cells, and capture of a wide variety of per-cell measurements (the "high-content" aspect of our work). We discuss methods and techniques for exploration and analysis of the resulting data and illustrate their application to real-world biological experiments.

II. Challenges in Image-Based High-Content Screening

We have analyzed several large screens with our system. This paper presents some of the challenges inherent to image-based screens, and the methods we use to address those difficulties. We will use two screens in particular as examples. The first is a set of cell microarrays, single glass slides with cells growing on an array of "spots" printed with gene-knockdown reagents [10]. The second is from an experiment screening $\sim$5000 RNA-interference lentiviral vectors targeted to silence $\sim$ 1000 human genes, run in a set of 384-well plates [11]. Both experiments produced thousands of high-resolution (512×512 pixels or larger) images, each containing hundreds of cells. Each image contains cells with a single gene's expression knocked down (decreased).

These experiments suffered from a variety of biases and sources of noise. Both showed illumination variation of around a factor of 1.5 within the field of view, swamping many measurements with noise if not corrected. The cell microarray experiment was performed with *Drosophila melanogaster* Kc167 cells, which are notoriously difficult to segment accurately [12]. Also in this experiment, significant post-measurement biases were detected based on spot position on the slide, due to variations in cell seeding density, concentrations of nutrients or stain, or other factors.

For both screens, discovery of unknown "interesting" phenotypes was and is an open-ended goal. We take a wide variety of per-cell measurements, because we do not know which measurements will be most useful or interesting *a priori,* both in the particular screen and for future explorations. Capturing a wide variety of measurements provides the most freedom in post-cytometry analysis, but also leads to difficulties in finding which subset of hundreds of measurements can most effectively discriminate a particular phenotype.

Moreover, even in the more goal-directed screens, we are often focused on identifying cells that are different from the "usual" cell in ways that may not be completely specified. Algorithms and methods that work well on normal cells can fail completely when faced with cells that vary significantly in appearance. Robustness to wide variation in cell appearance is therefore an overarching concern in all of our work.

In the following section, we discuss how each of the issues discussed above arose during screens, and the methods we used to overcome these challenges.

III. Methods

A. Illumination Normalization

Any image- and cell-based screen involves several devices whose physical limitations lead to biased measurements. One of the most pervasive of these is non-uniformity in the optical path of the microscope and the imager. It is typical for the overall illumination to vary by almost a factor of two across the field of view, making segmentation of individual cells more difficult, and seriously compromising intensity-based measurements. Since many such measurements vary less than two-fold in a group of cells, they will be useless unless the illumination is normalized. Fortuitously, such variation is consistent from image to image within a single screen, provided as many elements as possible do not change within the screen, i.e., the microscope and optical components are kept the same, the same type of slide or plate is used consistently throughout the experiment, and the images are taken in as short a span as is feasible. We include uneven incoming illumination, sensor biases, and illumination variation due to lens and slide imperfections under the single term "illumination variation."

We need to estimate the illumination variation in order to correct for it within each image. We model the image-forming process at pixel (x, y) in a particular image I as,

$$I_{x,y} = L_{x,y}(C_{x,y} + b), \tag{1}$$

where I is the image, L is the illumination function, C is an indicator function which is 1 if a cell overlaps pixel (x, y), and 0 otherwise, and b is a term to account for background staining. Note that this model conflates the magnitude of $C + b$ and L, but since we lack any data that give L physical units, we only need to estimate it up to a scale factor. Lindblad and Bengtsson use a similar model for single image normalization (after log-transformation of the pixels), but without the background term [13]. In our experience, non-specific (background) staining is not always low enough to disregard during normalization.

In the cell microarray experiments, we found that the cell distribution was uniform in the field of view (figure 1) and, in this case, background staining b negligible (as judged from a histogram of pixel intensities). In this case, we estimate L as a smoothed per-channel average of I across all the images in the screen. The average intensities for the three channels (i.e., stains) and the (uniform) cell distribution are shown in figure 1. Note that we smooth the intensity images to reduce sampling noise prior to using them for illumination correction.

In the well-based experiments, each well was imaged in four different locations. Each location had a significantly different cell distribution, but the background staining level b and the illumination function L were the same across locations (as judged by eye). We use a smoothed regression via equation (1) to estimate L for a range of values of b, taking the pair that best fit the position-wise averages. Cell distribution was estimated by smoothing DNA-stained images and adaptively thresholding to approximately locate nuclei (more accurate identification of nuclei is described in the next section).

Illumination correction is necessary for accurate segmentation and measurement of cells. We note that the optimal solution would combine the estimation of illumination variation and the segmentation steps into a single procedure, similarly to well-known EM-segmentation methods that simultaneously fit a smooth bias field and discrete segmentation labels to image data [14], [15]. In our initial implementation, we have focused on each step separately, with the goal of understanding the nature of the signal. In addition, the high-throughput nature of the experiments places substantial run-time limitations on the algorithms used for the analysis of individual images. We are currently working on a fast implementation for simultaneous illumination correction and segmentation of cell images.

B. Segmentation

The primary benefits of image-based assays are the capture of per-cell data, with a large number of per-cell measurements. This prevents the conflation of multimodal populations, as in expression profiling with gene-chips, and provides a much richer data source than other methods, such as flow cytometry. To exploit the full potential of this data, however, it is necessary to accurately segment individual cells within each image.

Unfortunately, the appearance of cells is highly variable from assay to assay. Experiments use different types of cells, different staining protocols, different growth substrates, and of course, different conditions within each assay. All of these prevent a single approach from being optimal for all cases. We have implemented several methods in our system in a modular fashion so we can easily adapt to new screens.

We have developed a successful, general approach for cell segmentation. Nuclei are more uniform in shape and more easily separated from one another than cells, so we first segment nuclei, then use segmented nuclei to seed the segmentation of individual cells. We threshold the nuclear image using a regularized version of Otsu's method [16] or our own implementation that fits a Gaussian mixture to pixel intensities. After thresholding the nuclear channel, we separate nuclei that appear to abut or overlap by locating well-separated peaks in the intensity image, and use either a watershed transformation [17] or Voronoi regions of the peaks to place nuclear boundaries, as in related work [18], [19], [20]. Our thresholding and segmentation system are modular, so the user can experiment with different approaches on a small set

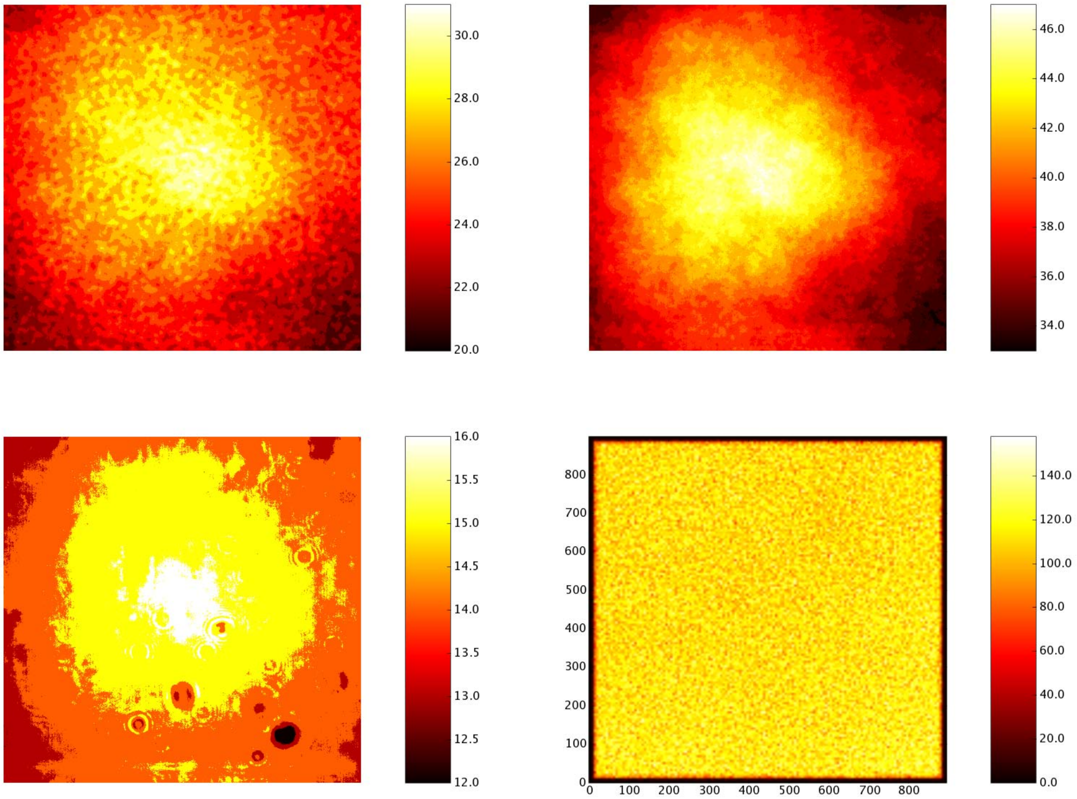


Fig. 1. Top: Mean intensity for DNA (nuclear) and Actin (cytoskeletal) stained channels in the cell microarray experiment. Bottom left: Mean intensity for phospho-Akt stained channel (a protein of interest in this screen). Bottom right: The flat distribution of nuclear centers. Nuclei that overlap the image boundary are eliminated before measurement. All images are false color.

of images to determine the best option, or modify existing modules for a particular experiment.

Given segmented nuclei, segmentation of individual cells is a matter of locating the borders between adjacent cells. The wide variety of cellular phenotypes discussed above prevents us from knowing the particular appearance of cell borders, and in fact, in many screens the borders may change significantly in response to a particular condition, such as a gene's knockdown. For this reason, we use a very general method for placing cell borders.

A priori, we assume that a pixel we have classified as being "within some cell" is more likely to be associated with the closest nucleus in the image. This naturally leads to using the Voronoi regions of the nuclei to place borders between cells. Another approach is to assume the borders of the cells are brighter or darker, and use a watershed transformation to place boundaries. Both of these approaches are commonly used in image cytometry [21], [22]. However, the first approach makes no reference to the cytoskeletal stain (i.e., information on where the border of the cell is actually located), and the second relies on the borders of the cells being brighter or darker and is overly sensitive to noise in pixels at cell boundaries. In our experience, both of these methods provide poor results in practice. We combine and extend these approaches by defining a distance between pixels that makes dissimilar pixels farther apart, and use this metric to compute nearest-neighbor regions.

We define similarity in terms of pixel neighborhoods. The distance between adjacent pixels at positions i and j is computed as

$$((i-j)^T \nabla g(I))^2 + \lambda ||i-j||^2 \tag{2}$$

where $g(I)$ is a smoothed version of the image, $||i-j||$ is the Euclidean distance between pixels i and j, and λ is a regularization term that balances between image-based and Euclidean distances. Distances between non-adjacent pixels are computed as the shortest path stepping between adjacent pixels, and cells are segmented via Voronoi regions of nuclei under this metric. More details of this approach are given in our earlier work [23].

C. Measurements

After segmentation, it is possible to make per-cell measurements for each image. Even if the screen is very targeted and the staining protocol has been tuned to give a simple binary answer, we capture a wide variety of measurements in order to maximize our ability to make inferences from the data.

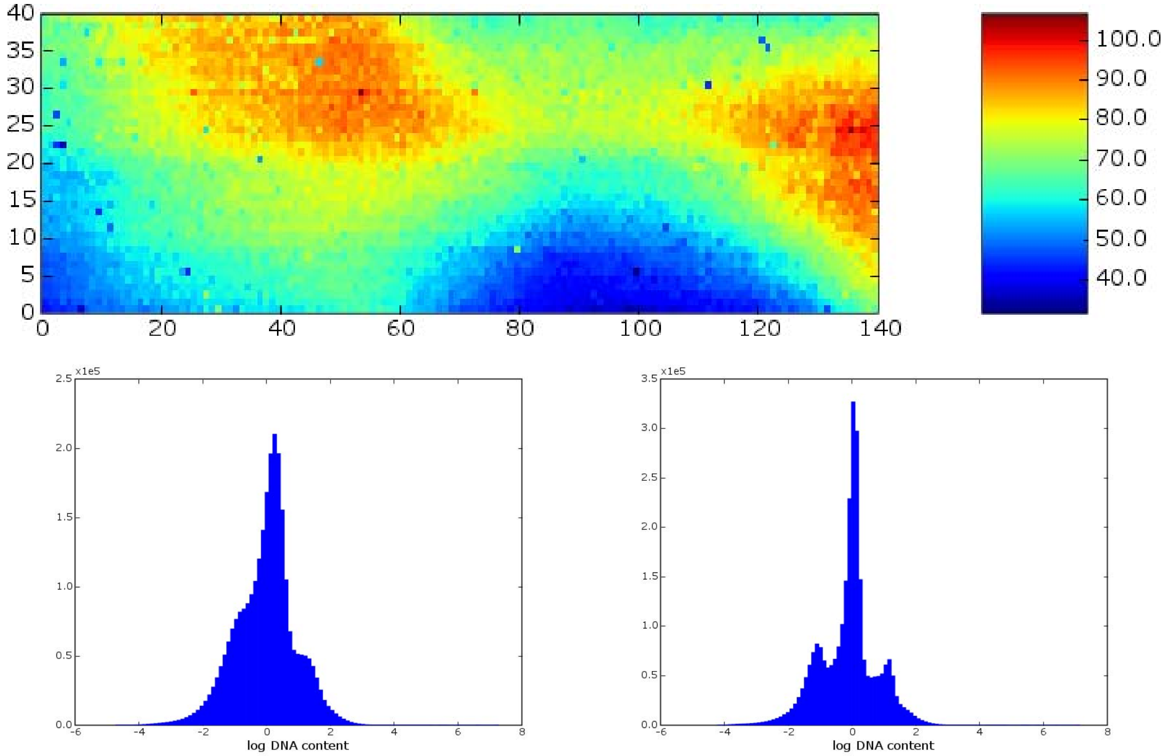


Fig. 2. Top: Median cell DNA content plotted on the physical layout of the slide in the cell microarray experiment. Bottom left: DNA content histogram for all cells on the slide, prior to spatial bias correction. Bottom right: DNA content histogram after bias correction. DNA content is measured by total intensity of the DNA stain within the nucleus, with unknown units. For this cell line, the dominant peak is made up of 4N cells, in which the DNA has duplicated, but the cells have not yet divided. The horizontal axis labels show relative values only.

For each cell, we make measurements of its morphology (e.g., area, perimeter, extent, convexity, and several Zernike moments), and intensity and texture of the various stains (e.g., mean and standard deviation of intensity, correlation of stains, and Gabor filter response at various scales). Measurements are also broken down by cellular compartment (nucleus, cytoplasm, and entire cell). A full discussion of which measurements to use in a given screen is not germane to this paper, but our guiding principle has been that, although it can make inference more difficult, taking too many measurements is better than taking too few. Adding new measurements to our system is simple because of its modular design.

Many of the measurements we capture have a clear biological meaning, such as cell size, or total DNA staining intensity in the nucleus (proportional to the amount of DNA present). Others have a less obvious connection to the biology of the cell, such as the eccentricity of the nucleus, or amount of variation in the cytoskeletal stain. Although we may not be able to assign meaning to every measurement, we can still make use of them when performing analyses or when classifying cells, as discussed in section III-E.

D. Spatial Bias Correction

Before measurements can be used to make biologically useful statements, we must control for systematic biases as much as possible. Biases in the data often come from variation across the physical layout of the slide or multi-well plate in which the experiments were performed (a.k.a, “plate effects” and “edge effects” [24]).

Some measurements can be corrected by fitting a smooth function to the data on the physical layout, and dividing the corresponding per-cell measurements at each position by the smooth function. For example, if we plot median per-cell DNA intensity on the slide layout for a 5600-spot (140×40) slide (figure 2, top), we observe a spatially varying bias, most likely due to inhomogeneity in the stain for DNA. We correct for this bias by applying a 2D median filter to the 140×40 values and dividing each cell’s measurement by the smoothed value. The improvement in the per-cell DNA content histogram is obvious (figure 2, bottom left vs. right).

In some cases, it is difficult to determine how to correct a particular measurement or combination of measurements. Nonlinear interactions of cells with their environment makes it nearly impossible to remove all biases before making inferences from the data. Therefore, we make maximum use of nearby control spots or wells and check each measurement we use against the physical layout (as in figure 2). Bias correction is an active area of research [24], [25].

E. Exploration and Inference

We take several approaches to exploring data from high-throughput, high-content screens. For example, (1) per-cell measurements can be combined to give per-gene values by taking means, medians, etc., or by using other data reduction techniques. (2) Pairs of populations produced by different gene knockdowns can be compared directly using distribution-based metrics. Or, (3) individual cells can be classified by their measurements, and gene knockdowns compared by how they change the balance of different classes of cells. We discuss each of these approaches below.

1) Per-Gene Measurements: Each per-cell measurement can be converted to a per-gene measurement by taking the mean, median, or otherwise reducing each measurement to a small set of parameters. This approach works particularly well when the screen focuses on a simple single parameter readout (e.g., presence of a given protein), or if the goal is to find gene knockdowns that have an easily measured effect (e.g., cause cells to grow larger). For example, figure 3 shows a scatterplot of per-gene mean cell size vs. mean nuclear size. Three replicates knocking down the gene *ial* are highlighted, in which cells and nuclei have grown larger than controls.

This approach is also effective for early, open-ended exploration, where identification of outliers is the primary task, especially since it can be applied to any measurement without prior knowledge about that measurement's biological implications.

Reducing the data in this way makes it weakly analogous to the data from gene-chips, in which mean expression level is measured for a large number of proteins. Like gene-chips, this approach can suffer from an over-reduction of measurements. For example, knocking down a gene may cause some cells to double in size, and an equivalent fraction to halve in size, but this would not affect the mean cell size [26]. In contrast, if we work with measurements' distributions directly, such differences are easily detected.

2) Population comparisons: To compare two populations' measurements directly without first reducing to a single per-gene value, we can apply distribution comparisons such as the Kolmogorov-Smirnov [3] or Kuiper [27] tests, or compute sample-based information-theoretic estimates, such as the KL-divergence between the two distributions [28]. These can be used to compare each sample against a set of positive or negative controls, or against the full slide-wide cell population, yielding a more experiment-specific per-gene measurement as discussed above.

Comparing gene knockdowns' populations via a single or small set of per-cell measurements, as in figure 3, top right, is similar to exploring data from flow cytometry, in which a few measurements are taken for a large number of cells. Flow cytometry is generally lower-throughput than image-cytometry. The number of measurements is also much more limited compared to automatic image-based cytometry.

3) Per-Cell Classification: To take full advantage of the large number of per-cell measurements, our primary method of exploration is via per-cell classifiers. We build or train classifiers that identify a phenotype of interest, and apply them to the full screen in order to determine which conditions or gene knockdowns cause enrichment or depletion in those phenotypes. Our goal is to understand the function of genes, with the underlying assumption that gene knockdowns that cause similar changes in phenotype have similar functions in the cell.

In particular, we advocate the per-cell classifier approach because it detects very small changes in the percentage of cells falling into a particular class. Some phenotypes, such as mitotic (replicating) cells, are $<1\%$ of cells at the background level and increase only three-fold above this level in outliers and positive controls [8]. These changes are so small relative to the full population that they are swamped if measurements are blindly combined into per-gene values, or when comparing two otherwise similar distributions.

Given a classifier for cells showing a known phenotype, the list of gene knockdowns that enrich or deplete that phenotype can be used to impute function for those genes. For example, if we build a classifier for cells in metaphase, knockdowns that cause enrichment of that phenotype probably have a regulatory function in the metaphase to anaphase transition. Simplified examples of per-cell classifiers are shown in figure 4, in which classifiers were constructed to identify different phases of the cell cycle based on a pair of measurements, total nuclear DNA content (as measured by the DNA stain), and mean nuclear phospho-H3 content (a marker for mitosis). If a gene knockdown significantly changes the fraction of cells landing in one (or more) of these classifiers, it is likely to be a regulator for those phases of the cell cycle. Most classifiers are more complicated than this, involving a larger number of per-cell measurements [8].

To compute enrichments and p-values, we treat the output of classifiers as Bernoulli random variables. If negative controls are available in the screen, then enrichments are computed relative to those controls. Otherwise, we use the full screen-wide cell population as the control, the operative assumption being that for each phenotype, knockdown of most genes will not affect that phenotype. There are two justifications for this assumption: many genes are not expressed under experimental conditions, so they cannot be depleted by knockdown, and most genes' knockdown will have no effect on a particular phenotype.

The phenotypes targeted by the classifier can be biologically well-characterized, such as cells in a particular phase of the cell cycle (as above), or simply cells that have a novel appearance, without a well-defined biological interpretation attached. For an uncharacterized phenotype, the group of gene knockdowns causing enrichment or depletion in that phenotype can be informative depending on the group of gene knockdowns causing similar effects. For example, the genes in the group might share a physical or biochemical property, suggesting a mechanism for the phenotypic change. Or, if the group contains genes with a similar, known function, the uncharacterized genes in the group can be hypothesized to also share that function. This also allows for the identification of new, hypothetical cellular processes, rather than simply identifying genes involved in known processes.

The per-cell classifier approach can also be applied to a

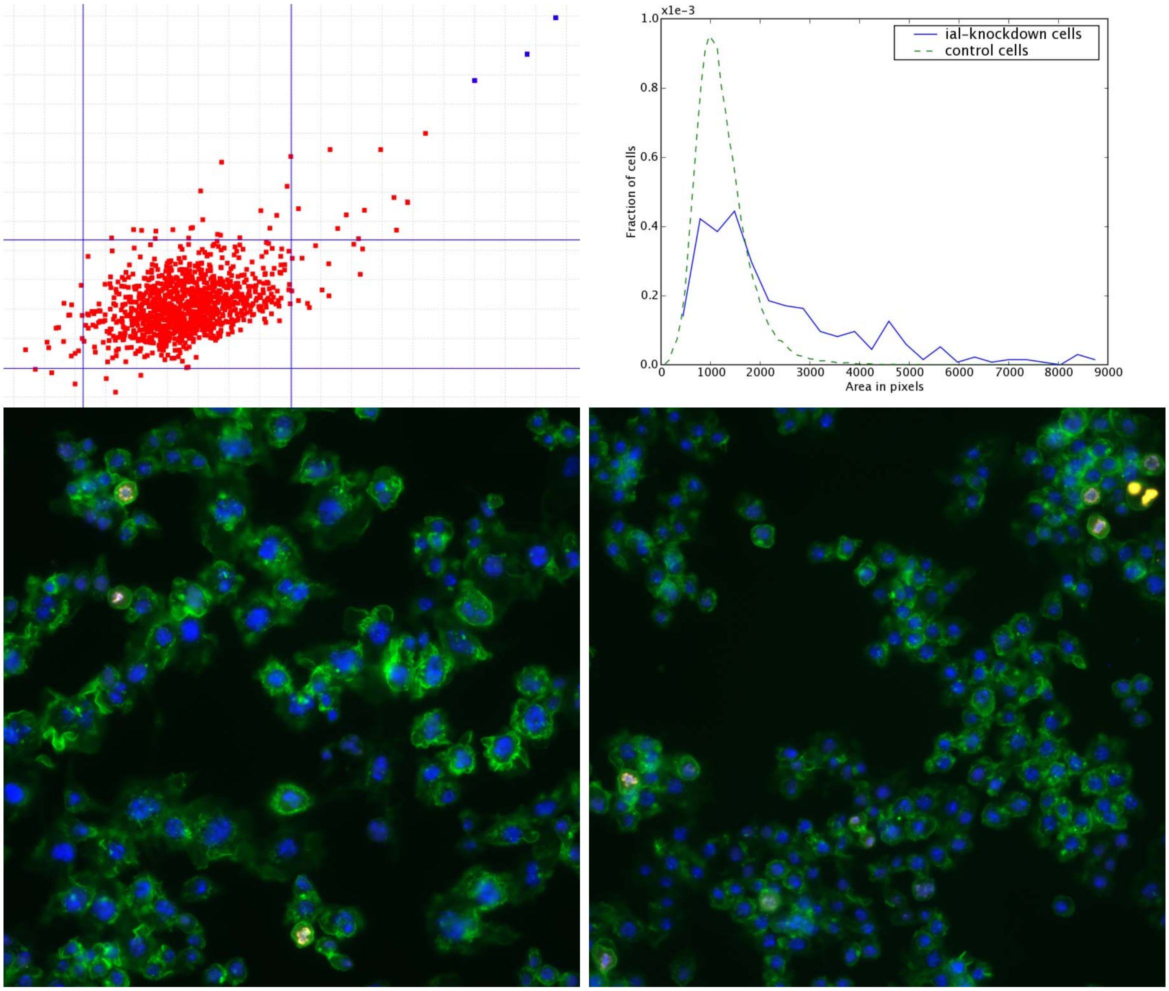


Fig. 3. Top left: Scatterplot of per-image mean cell area vs mean nuclear area in a cell microarray experiment. Three replicates knocking down the gene *ial* are highlighted in the upper right corner. The lines show two standard deviations around the mean. Top right: Per-cell histograms of cell area for the *ial* replicates compared to controls. Lower left: Cells with *ial* knocked down. Lower right: Control cells (Blue: DNA, Green: Actin, Red: phospho-Histone H3).

particular gene's knockdown which might not show a human-discernable phenotype, but for which we can still build a classifier. If the classifier is effective at separating the cells with the target gene knocked down from the cell population at large, the implication is that there is a measurable phenotype caused by the target gene's knockdown, and other knockdowns that cause the same phenotype have a similar function.

One of the benefits of the classifier-based approach is that it is less susceptible to spatial biases when the classifier is trained by a human, compared to data-reduction or full-population comparison methods, because of the robustness of the human visual system to these biases. Note, however, that the nonlinear effect of environment on cells can cause biases in the fraction of cells of a particular phenotype, so the results of applying the classifier should be checked for spatial bias, similar to figure 2.

The classifier approach is reminiscent of example-based image retrieval [29], [30]. However, rather than searching for images as the primary goal, we are using similar techniques to quickly categorize subimages of cells, with the intent of determining the number and distribution of cells matching our query specification.

After finding hits in a particular screen, followup work may be necessary to validate the results. If there are a sufficient number of replicates, or data from other screens are available, it may be possible to make a categorical statement about a gene knockdown's effect without further experimental work. However, in almost all cases, biologists will investigate the mechanism of the effect in traditional followup experiments.

IV. Discussion

This paper has presented several methods for high-content, high-throughput image-based screens of cells. Such screens are particularly valuable in biological and pharmaceutical research. We have developed CellProfiler [9], a modular, open-source system incorporating these methods.

High-throughput, high-content screening is a powerful technique for making discoveries about cellular processes, genetic

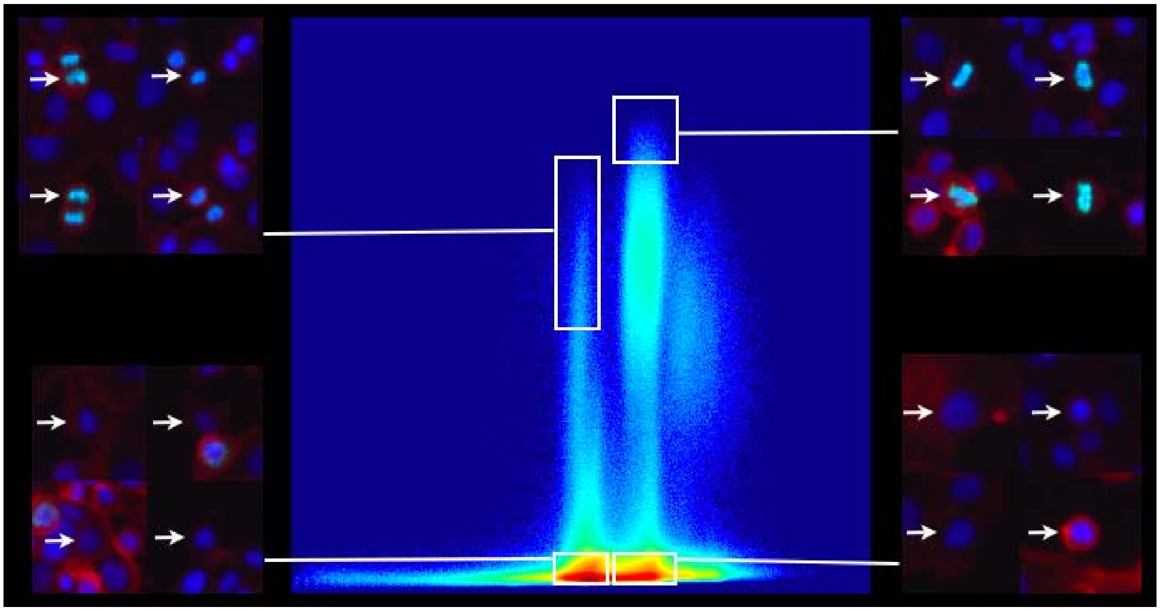


Fig. 4. Simplified examples of per-cell classifiers (from [8]). The central scatterplot shows total DNA content (horizontal axis) vs mean phospho-histone H3 staining intensity for *all* cells in the screen of human genes. Phospho-histone H3 is present in cells undergoing cell division (mitosis). Selecting different regions in the scatterplot selects different subpopulations of cells, as shown in the insets. Each inset shows 4 subimages; each subimage shows a random cell from the corresponding subpopulation (marked) and its surrounding image neighborhood. Counter-clockwise from lower left: 2N cells (normal complement of chromosomes), 4N cells (DNA duplicated), Metaphase (condensed DNA, preparing to separate), Anaphase/Telophase (daughter cells separating). This is also the progression of the cell cycle. A gene knockdown causing enrichment in any of these subpopulations relative to controls is likely a regulator of that phase of the cell cycle. Most classifiers involve many more measurements.

pathways, and drug candidates. It also poses new challenges and requires novel techniques to realize its full potential as a discovery tool. Algorithms for identifying, segmenting, and measuring individual cells must deal with noise and biases, be robust to a wide variety of cell appearances, and must be accurate enough to allow very small ($<1\%$) subpopulations to be identified accurately. However, the payoff for the increased effort is a dramatically more powerful method for detecting changes in cells under different experimental conditions, through the use of per-cell data and classifiers, compared to more traditional techniques. These methods have been proven in some of the first large-scale automatic screens to appear in the biological literature [8], [11].

For reasons of scope, this paper does not include a discussion of the architecture and design of our system implementing the techniques presented here [9]. We aimed to make the system as modular and extensible as possible, while maintaining a user-friendly interface. We believe it has been successful in these respects, particularly given its use in a variety of non-cell screening tasks (e.g., counting and classifying yeast colonies on Petri dishes, counting nuclear subcompartments/speckles, tumor measurement, etc. [31]).

In the future, we will incorporate methods for simultaneous illumination normalization and segmentation, and automatic methods for spatial bias correction. We have also started to explore general clustering based on per-cell measurements.

Acknowledgements

This work was supported by an academic grant from the Society for Biomolecular Sciences (AEC), a fellowship from the Merck/MIT computational and systems biology initiative (AEC), the MIT EECS/Whitehead/Broad Training Program in Computational Biology (NIH grant DK070069-01; TRJ), NIH grant R01 GM072555-01 (DMS), and a grant awarded by the U.S. Army Medical Research Acquisition Activity: W81XWH-05-1-0318-DS (DMS), although no official endorsement by the government should be inferred. AEC is a Novartis fellow of the Life Sciences Research Foundation.

References

[1] A. A. Kiger, B. Baum, S. Jones, M. R. Jones, A. Coulson, C. Echeverri, and N. Perrimon, “A functional genomic analysis of cell morphology using RNA interference,” *J Biol*, vol. 2, no. 1475-4924 (Electronic), p. 27, 2003.

[2] J. K. Kim, H. W. Gabel, R. S. Kamath, M. Tewari, A. Pasquinelli, J.-F. Rual, S. Kennedy, M. Dybbs, N. Bertin, J. M. Kaplan, M. Vidal, and G. Ruvkun, “Functional genomic analysis of RNA interference in C. elegans,” *Science*, vol. 308, no. 1095-9203 (Electronic), pp. 1164–7, 2005.

[3] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler, “Multidimensional drug profiling by automated microscopy,” *Science*, vol. 306, no. 1095-9203 (Electronic), pp. 1194–8, 2004.

[4] J. A. Philips, E. J. Rubin, and N. Perrimon, “Drosophila rnai screen reveals cd36 family member required for mycobacterial infection,” *Science*, vol. 309, no. 1095-9203 (Electronic), pp. 1251–3, 2005.

[5] J. N. Harada, K. E. Bower, A. P. Orth, S. Callaway, C. G. Nelson, C. Laris, J. B. Hogenesch, P. K. Vogt, and S. K. Chanda, “Identification of novel mammalian growth regulatory factors by genome-scale quantitative image analysis,” *Genome Res*, vol. 15, no. 1088-9051 (Print), pp. 1136–44, 2005.

[6] L. Pelkmans, E. Fava, H. Grabner, M. Hannus, B. Habermann, E. Krausz, and M. Zerial, "Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis," *Nature*, vol. 436, no. 1476-4687 (Electronic), pp. 78–86, 2005.

[7] N. H. Pipalia, A. Huang, H. Ralph, M. Rujoi, and F. R. Maxfield, "Automated microscopy screening for compounds that partially revert cholesterol accumulation in niemann-pick c cells," *J Lipid Res*, vol. 47, no. 0022-2275 (Print), pp. 284–301, 2006.

[8] A. E. Carpenter, T. R. Jones, M. Lamprecht, D. B. Wheeler, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "CellProfiler: image analysis for high throughput microscopy," 2006, in preparation.

[9] [Online]. Available: http://cellprofiler.org

[10] D. B. Wheeler, S. N. Bailey, D. A. Guertin, A. E. Carpenter, C. O. Higgins, and D. M. Sabatini, "RNAi living-cell microarrays for loss-of-function screens in Drosophila melanogaster cells." pp. 127–32, 2004.

[11] J. Moffat, D. A. Grueneberg, X. Yang, S. Y. Kim, A. M. Kloepfer, G. Hinkle, B. Piqani, T. M. Eisenhaure, B. Luo, J. K. Grenier, A. E. Carpenter, S. Y. Foo, S. A. Stewart, B. R. Stockwell, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini, and D. E. Root, "A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen," *Cell*, vol. 124, no. 0092-8674 (Print), pp. 1283–98, 2006.

[12] S. Armknecht, M. Boutros, A. Kiger, K. Nybakken, B. Mathey-Prevot, and N. Perrimon, "High-throughput RNA interference screens in Drosophila tissue culture cells," *Methods Enzymol*, vol. 392, no. 0076-6879 (Print), pp. 55–73, 2005.

[13] J. Lindblad and E. Bengtsson, "A comparison of methods for estimation of intensity nonuniformities in 2D and 3D microscope images of fluorescence stained cells." in *Proceedings of the 12th Scandinavian Conference on Image Analysis (SCIA)*, 2001, pp. 264–271, a.

[14] W. Wells, W. Grimson, R. Kikinis, and F. Jolesz, "Adaptive segmentation of MRI data," in *IEEE TMI*, vol. 15, 1996, pp. 429–442.

[15] K. V. Leemput, F. Maes, D. Vanermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," in *IEEE TMI*, vol. 18, no. 10, 1999, pp. 885–895.

[16] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Systems, Man and Cybernetics*, vol. 9, pp. 62–66, Mar. 1979.

[17] S. Beucher, "The watershed transformation applied to image segmentation," in *Scanning Microscopy International*, vol. 6, 1992, pp. 299–314.

[18] N. Malpica, C. O. de Solorzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M. Garcia-Sagredo, and F. del Pozo, "Applying watershed algorithms to the segmentation of clustered nuclei," *Cytometry*, vol. 28, no. 0196-4763 (Print), pp. 289–97, 1997.

[19] F. Meyer and S. Beucher, "Morphological segmentation," in *Journal of Visual Communication on Image Representation*, vol. 1, no. 1, 1990, pp. 21–46.

[20] C. Wahlby, I.-M. Sintorn, F. Erlandsson, G. Borgefors, and E. Bengtsson, "Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections," *J Microsc*, vol. 215, no. 0022-2720 (Print), pp. 67–76, 2004.

[21] C. Wählby, "Algorithms for applied digital image cytometry," Ph.D. dissertation, 2003.

[22] S. Beucher, "The watershed transformation applied to image segmentation," in *Scanning Microscopy International*, vol. 6, 1992, pp. 299—314.

[23] T. R. Jones, A. E. Carpenter, and P. Golland, "Voronoi-based segmentation of cells on image manifolds," in *CVBIA*, ser. Lecture Notes in Computer Science, Y. Liu, T. Jiang, and C. Zhang, Eds., vol. 3765. Springer, 2005, pp. 535–543.

[24] D. E. Root, B. P. Kelley, and B. R. Stockwell, "Detecting spatial patterns in biological array experiments," *J Biomol Screen*, vol. 8, no. 1087-0571 (Print), pp. 393–8, 2003.

[25] D. Kevorkov and V. Makarenkov, "Statistical analysis of systematic errors in high-throughput screening," *J Biomol Screen*, vol. 10, no. 1087-0571 (Print), pp. 557–67, 2005.

[26] J. M. Levsky and R. H. Singer, "Gene expression and the myth of the average cell," *Trends Cell Biol*, vol. 13, no. 0962-8924 (Print), pp. 4–6, 2003.

[27] N. H. Kuiper, "Tests concerning random points on a circle," in *Proc. K. Ned. Akad. Wet., Ser. A*, vol. 63, 1962, pp. 38—47.

[28] S. Kullback and R. A. Leibler, "On information and sufficiency," in *Annals of Mathematical Statistics*, vol. 22, no. 1, 1951, pp. 79—86.

[29] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions," *Int J Med Inform*, vol. 73, no. 1386-5056 (Print), pp. 1–23, 2004.

[30] K. Tieu and P. Viola, "Boosting image retrieval," *Int. J. Comput. Vision*, vol. 56, no. 1-2, pp. 17–36, 2004.

[31] A. Carpenter, M. Lamprecht, and D. Sabatini, in preparation.

End of report