CHAPTER 9

Data Analysis Pipeline for High Content Screening in Drug Discovery

Daniel W. Young, Charles Y. Tao, and Yan Feng

9.1 Introduction

Automated fluorescent microscopy based high content screening (HCS) is making a major impact on many aspects of the drug discovery process by enabling simultaneous measurement of multiple features of cellular phenotype that are relevant to both therapeutic and toxic activities of compounds. HCS employs highly automated processes and typically generates immense datasets reflecting cell phenotypic information. Besides substantial research articles published in the past 10 years, three recent books were devoted entirely to HCS, encompassing the scientific and technological aspects of assay development, automation, image acquisition, and image analysis [1–3]. However, mining the large HCS dataset remains an ongoing challenge in the field. In this chapter, we will briefly summarize current HCS technology and aim to focus our discussion mainly on the integration of necessary computation and statistical methods for data handling and mining. Examples of data normalization, dose response estimation, cytometry analysis, and data driven factor analysis will be discussed in detail.

9.2 Background

Drug discovery is a complex process that entails many aspects of technology. A recent estimate put the cost of developing a new drug at more than $1 billion, and the time for developing such a drug at 12 to 14 years [4]. How to reduce the cost and increase the efficiency of the drug discovery process has become increasingly important.

Typical drug discovery efforts employ a ''targeted'' process, which starts with the identification of specific disease related target, usually gene products that are either mutated or misregulated in patients. Then functional assays, such as enzyme or binding based in vitro assays, are designed to identify small molecule or biological ''hits'' that can be used to perturb or restore the function of the target. Such ''hit'' molecules are further tested for its efficacy and toxicity in cellular disease models to become a ''lead'' molecule. Drug leads with a desired activity profile are then tested in animals and humans before they are finally submitted to Food and Drug Administration (FDA) for approval. The targeted approach has been increasingly criticized recently due to the high failure rate of identifying active

but nontoxic drug leads. Hence, there is a need for refined approaches to predict success earlier in the discovery process [5]. HCS, which can simultaneously measure multiple cellular phenotypes which are relevant to both efficacy and toxicity, became increasingly popular because it holds the promise of expediting the drug discovery process [6].

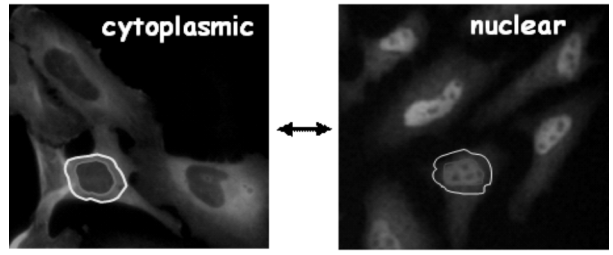
Early HCS applications were mainly focused on secondary assays for the hit-to-lead process. Evaluating the effects of hit compounds in cellular models is crucial for making decisions on the fate of the compound, as well as the direction of the drug discovery program. Hit compounds are usually tested in a variety of assays reflecting their on-target activities, off-target activities, toxicities, and physicochemical properties. Cellular imaging based HCS assays have been developed in all these areas. Successful cases include nuclear and plasma membrane translocation of transcription factors or signaling molecules, endocytosis of G-protein coupled receptor (GPCR) after activation, cell cycle, and neurite morphology assays [7–10]. With the increasing realization of the pitfalls of targeted based drug discovery, HCS has now been adopted earlier in the hit discovery stage at many major pharmaceutical companies [11]. Interestingly, phenotypic profiling of drug activities using high content analysis has been shown to be able to predict drug mechanism of action (MOA) with remarkable accuracy, comparable to the more expensive and low throughput transcription profiling technology [12, 13]. The large quantity of single cell descriptors from high content analysis might hold the key to such a result. High-throughput prediction of drug MOA and toxicity is one of the most important challenges facing drug discovery. High content analysis is predicted to have an increasingly important role in this area as the technology develops.

9.3 Types of HCS Assay

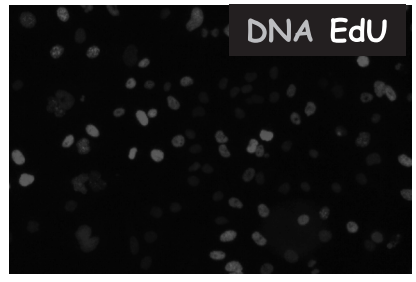
Thanks to the rapid development of knowledge on molecular and cellular functions in the past two decades, many cellular processes can be adopted into HCS assays format in rather straightforward ways and usually in a short timeframe. Current HCS assays can be roughly divided into three main formats: (1) translocation assays, (2) cytometry assays, and (3) morphology assays. Figure 9.1 illustrates examples of the three assay formats.

Translocation assays score for concentration shift of a particular protein from one cellular compartment to another. Immunocytochemistry using specific antibodies and expression of green fluorescence protein (GFP) fusion protein are among the most commonly used methods for detection. The ratio or difference of the concentration in different cellular compartments was used as a quantitative score for such events [14]. Nuclear-cytoplasmic translocation of transcription factors or signal transduction molecules, such as NFκB, NFAT, FOXO, p38MK2, and HDAC4/5 are among the most common ones of interest. Because of the simple format and robust quantitation method, nuclear-cytoplasmic translocation was also adopted in engineering biosensors to detect binding or cleavage events in cells, such as p53/MDM2 interaction or caspase3 activation [15, 16].

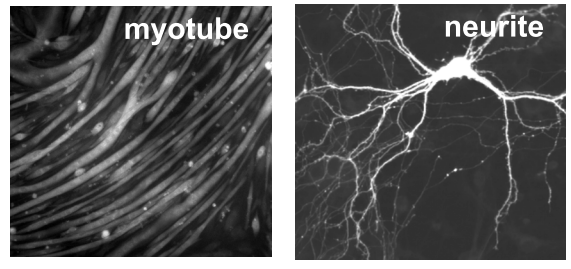
Cytometry assay treats each cell as a single entity and scores for population change of cells with certain phenotypic characteristics. Fluorescence activated cell



(a) Translocation



(b) Cytometry



(c) Morphology

Figure 9.1 HCS assay formats. (a) Cytoplasmic-nuclear translocation of GFP-Foxo3a. The ratio of GFP-Foxo3a intensity in the nuclear mask region (circle) and in the cytoplasmic mask region (ring) was used as the quantitative measurement. (b) Cytometry assay measuring percentage of S-phase cells. S-phase cells were labeled with Ethynyl-dU incorporation and azido-rhodamine label and all cells were labeled with Hoechest 33342 dye. (c) Morphology assays on myotube formation and neurite formation. Myotube length and width, neurite length, and branches, were among the parameters of interest.

sorting (FACS) is the most common format for cytometry based assays. FACS requires a large number ($>10^5$) of cells and does not provide enough resolution to score intracellular or morphological changes in cells, whereas HCS based cytometry requires many fewer ($\sim 10^2$) cells and can be performed in high spatial resolution, thus providing higher throughput and information not accessible to conventional FACS. Cell cycle assay is a cytometry based assay widely adopted in oncology research where interfering with cell cycle and cell growth is among the main goals of therapeutic intervention. Cell cycle stages are determined by DNA content, as well as elevation of other specific cell cycle markers such as phospho-Histone H3 for mitotic cells and cleaved PARP for apoptotic cells [17].

Morphological based assays are generally more challenging for image analysis, and thus more difficult to develop into a fully automated high throughput format. But a few exceptions, such as neurite outgrowth and myotube differentiation, where no other surrogate assay formats are available, have been studied more extensively [18, 19].

9.4 HCS Sample Preparation

Sample preparation is the first (and crucial) step for obtaining high-quality image data. Here we provided a typical protocol used in our lab as the starting point for HCS assay development. In practice, we found more than 90% of the assays can be covered by this simple procedure.

9.4.1 Cell Culture

Cells are plated at $\sim 10^5$/ml, 30 μl for 384 well and 100 μl for 96 well plate, respectively. We use a Multidrop (ThermoFisher), Microfill (BioTek), or PlateMate (Matrix) to assist automation of cell plating. Cells are then grown at 37°C in 5% CO_2 from overnight to 2 days, when treatment can be performed in between.

9.4.2 Staining

We use an ELX405 plate washer (BioTek) running at the lowest dispensing/aspiration speed to assist all fixation and staining processes. In brief, an equal volume of 2× concentrated fixatives, 7.5% formaldehyde in phosphate buffered saline (PBS) or Mirsky's (National Diagnostics), were added directly to the wells. After 15 minutes, the wells were washed 1× with PBS. For cells expressing GFP fusion protein, these are ready for image acquisition. For immunocytochemistry staining, the cells were then incubated 1× with PBS-TB (PBS with 0.2% Triton X-100, 0.1% BSA) for 10 minutes to permeabilize the cell membrane. Primary antibody was then added at ~0.5 to 5 μg/ml in PBS-TB and incubated at room temperature for 1 hour. The cells were then washed 2× with PBS-TB. Fluorescently labeled secondary antibody was added at 2 μg/ml together with 1 μg/ml nuclear stain Hoechst33342 in PBS-TB and incubated at room temperature for 30 minutes. The cells were then washed 2× with PBS-TB and 1× PBS before image acquisition.

9.5 Image Acquisition

Currently more than 10 commercial vendors provide automated fluorescence imaging hardware, including confocal and nonconfocal imagers. In principle, a confocal imager can provide images with reduced out-of-focus light. But in practice most of the applications use low magnification objectives on a flat monolayer of cells; a wide field microscope can already provide images of sufficient quality.

All the HCS imagers can focus on each field and acquire fluorescence images at various magnifications and wavelengths in fully automated mode. A robotic

Table 9.1 A Partial List of High Content Imaging Instruments

Instrument	*Manufacture*	*Technology*	*Live Cell*	*Liquid Handling*	*Data Management*	*Information*
ArrayScan VTi	ThermoFisher	Widefield with Apotome	Available	Available	STORE	http://www.cellomics.com
ImageXpress Micro	Molecular Devices	Widefield	Available	Available	MDCStore	http://www.moleculardevices.com
ImageXpress Ultra	Molecular Devices	Laser Scanning Confocal	No	No	MDCStore	http://www.moleculardevices.com
IN Cell Analyzer 1000	GE Healthcare	Widefield with structured light	Yes	Yes	In Cell Miner	http://www.gelifesciences.com
IN Cell Analyzer 3000	GE Healthcare	Laser Scanning Confocal	Yes	Yes	In Cell Miner	http://www.gelifesciences.com
Opera	PerkinElmer	Laser confocal with Nipkow disk	Available	Available	File directory	http://las.perkinelmer.com
CellWoRx	Applied Precision	Widefield with oblique illumination	No	No	STORE	http://www.api.com
Pathway 435	BD Biosciences	Widefield with Nipkow disk	No	No	File directory	http://www.atto.com
iCyte	Compucyte	Widefield Laser scanning	No	No	File directory	http://www.compucyte.com
MIAS-2	Maia Scientific	Widefield	No	No	File directory	http://www.maia-scientific.com
Explorer ^{e}X3	Acumen	Widefield Laser scanning	No	No	File directory	http://www.ttplabtech.com

plate handler can usually be integrated with the imager for a large screening effort. Several instruments also provide live imaging and liquid handling capabilities. In Table 9.1 we provide a partial list summarizing their main features. Technical details can easily be obtained by contacting the specific vendors. One has to consider many factors before choosing a high content imager that fits a specific need. Obviously, price, capability, IP status, and maintenance are all issues that need to be considered carefully. The most important thing is to test several instruments with real samples and questions in mind. For example, if counting responsive cells is the only desired application, one can choose a high speed instrument such as Explorer with low magnification and totally avoid image analysis. But if detailed intracellular structures such as endosomes are the primary interest, one has to go for an instrument with a higher magnification and a more sensitive camera, perhaps with confocal capability. High content imagers generate a large volume of raw data, especially when fitted with an automated plate loader. If large scale screening is the main application, one has to consider data archiving solutions. A database built in with the instrument is good for handling a large, multiuser environment but sufficient IT support is often required. All data generated for this chapter was from an upgraded version of Cellomics ArrayScan VI or MetaXpress Ultra Confocal imagers.

9.6 Image Analysis

Image analysis is a complex topic, and comprehensive coverage of the topic is beyond the scope of this chapter. At this moment most of the HCS instrument vendors provide some level of image analysis capabilities which are developed for specific biological applications, such as nuclear translocation, object content and texture measurements, vesicle counting, and neurite length/branch scoring. Vendor software usually gets the most use, largely because the proprietary format used by different vendors makes it hard to export and analyze in other software packages. But most vendor software is sufficient in generating wellular summary of image data for simple analysis. Generally speaking, most of the HCS image analysis starts with nuclei recognition, because nuclei are spatially well separated with sharp boundaries. A nuclear counter stain such as Hoechst33342 is most often used to outline the nuclear boundary. Thus, they can be easily recognized and accurately separated with almost any boundary detection algorithm. For the same reason, bright spots such as endosomes and Golgi can be detected and segmented easily [21].

Accurate cell boundary detection is a complex task. Because cells are irregularly shaped and frequently overlapped, efforts to accurately identify cell boundaries often suffer from serious under or over-segmentation problems. In practice, a proportional ring expansion or watershed expansion from the nuclear area is often used to approximate the cellular area. More complex iterative decision making processes have been shown to be able to detect cell boundary accurately. First, over-segmentation of cell bodies and then joining at a later stage with complex, environment-sensitive criteria were employed in these exercises [22, 23]. Once boundary detection is achieved, it is rather straightforward to score a multitude of

parameters for each object, such as dimension, intensity, and texture features. Cellomics and Metamorph software, for example, provide more than 30 parameters for each channel. A more comprehensive list was described by Haralick [24].

In addition to vendor software there are several attempts being made to develop generic image and data analysis software that can be used with all acquisition platforms. These include the free and open source image and data analysis software Cell Profiler, from the Broad/MIT and Whitehead Institutes (http://www.cellprofiler.org/index.htm) and Zebrafish image analysis software (ZFIQ) from the Methodist Hospital in Houston, Texas (http://www.cbi-platform.net/download.htm). More advanced image analysis software can be found in the Definiens software package (http://www.definiens.com/), which uses a novel over-segmentation and reconstruction approach to identify cellular regions of interest. Finally, another example of a free and robust (but rudimentary) software package is Image J from the NIH (http://rsb.info.nih.gov/ij). This package is extremely useful for viewing and quickly manipulating images for display and analysis. This software has a recordable macro language for analysis of images in batch and has a large user group base to freely exchange macros.

9.7 Data Analysis

In contrast to all the assays, hardware, and image analysis software developed for HCS in the past 10 years, little has been done on the data analysis aspects, one reason being that most application developers have neither experience nor the desire to handle a large amount of data, and without proper integration of data analysis pipeline, HCS has little use for high throughput screening and remains on par with a manually operated fluorescence microscope. In an effort to make HCS practical in a drug screen scenario, we have developed a series of tools for data extraction, normalization, dose response estimation, and multiparameter cell classification. We will use most of the rest of the chapter to discuss these tools, and introduce examples when necessary.

9.7.1 Data Process Pipeline

HCS generates a large amount of primary image data. For example, a typical medium scale screen of 25,000 compounds in duplicate can easily be achieved in one week with one HCS instrument and limited liquid handling automation access. Assuming four individual channels and four image fields were collected in each treatment condition, the experiment generates ~600 GB of raw image data, and a full HTS screen at 1M compounds will generate over 25 TB of raw image data (Figure 9.2). For small scale exploration such as HCS assay development, a local hard drive is sufficient. A corporate storage and archive system is necessary if screening effort runs at full scale. Some image analysis routines also generate a large amount of data, especially when every single cell analysis data is stored. We estimated that 2.5 billion parameters will be generated in the typical medium-sized experiment, as mentioned above, and a full million compound screen will produce over 100 billion data points. The dataset is big enough that it cannot be handled by

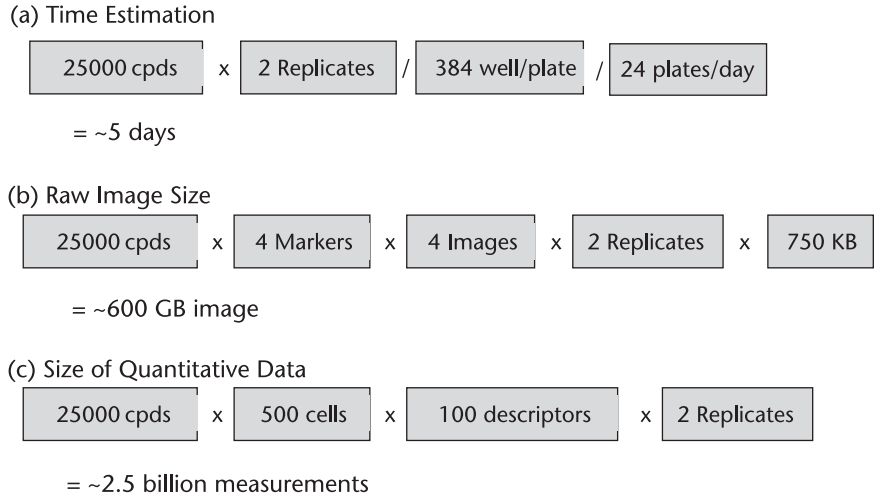


Figure 9.2 HCS data size. (a) A typical medium sized HCS screen with 25,000 compound treatments in duplicates took one week to complete, assuming imaging takes one hour per 384 well plates. (b) 600 GB of raw image data was generated, assuming four image fields for each treatment and four marker channels were taken with a 2 × 2 binned 1M pixel CCD camera and 12-bit digitizing format. (c) 2.5 billion measurements were generated, assuming 500 cells were quantified with 100 descriptors each in every treatment condition.

an Excel spreadsheet used by most bench biologists unless dramatic data reduction was achieved.

In our lab, a relational database was provided by Cellomics for storage of all raw image and image analysis data. Cellomics, like many other vendors, also provides simple data visualization software at the plate, well, and cell levels, but not at the screen level. Data treatment is limited to simple statistics such as averaging and Z function. None of the current commercial vendors provides tools for large scale data normalization and data mining, which are absolutely necessary for a successful high throughput screening campaign.

Here we developed/adopted a series of data analysis tools for a large numerical dataset resulting from image analysis algorithms. These data analysis tools are grouped into three modules, preprocessing, data mining, and data integration modules (Figure 9.3). The preprocessing module handles retrieval of cell level data from the image analysis software, data normalization, and generation of quality control plots. Most of the data reduction was achieved by a variety of data mining modules generating treatment level data. A series of data mining modules has been built in our lab. We will describe the dose response curve/confidence estimation module and the automated cytometry classification module in detail. A data driven factor model was also recently published, and we will briefly introduce the concept.

9.7.2 Preprocessing Normalization Module

Systematic errors, such as plate-to-plate and well-to-well variations, often exist in a high content screen. Data normalization is a necessary procedure to minimize the

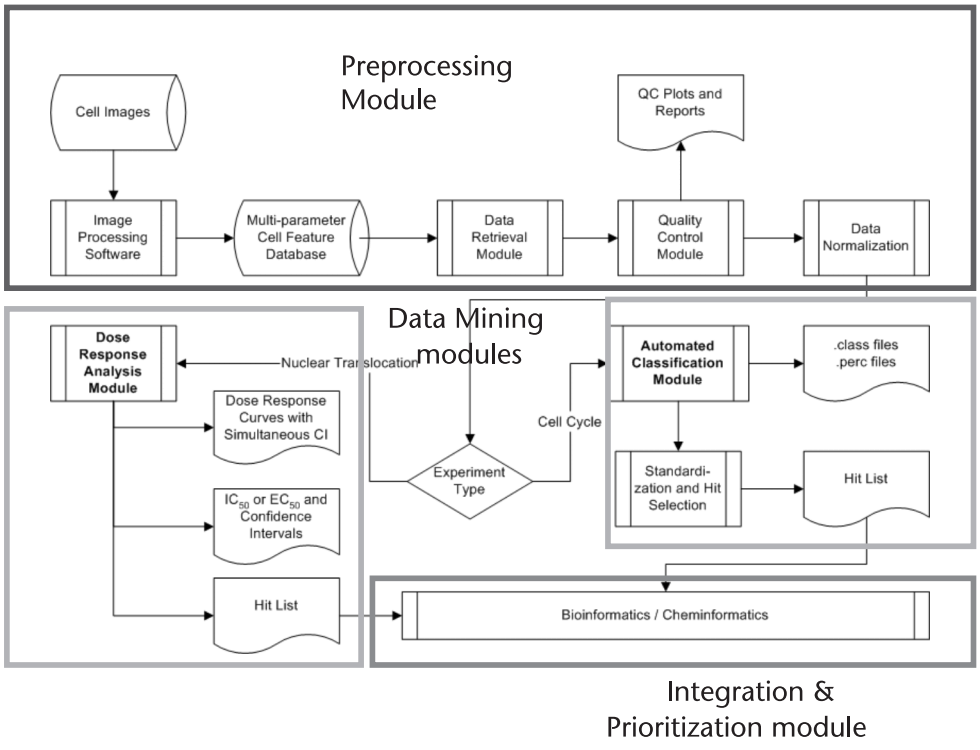


Figure 9.3 HCS data processing pipeline. Flowchart of HCS data processing pipeline consisting of preprocessing, data mining, and data integration modules. The preprocessing module extracts image analysis data and performs normalization. Data mining modules reduce the cell centric data to treatment centric data. The multiple data mining module can be tailored to specific needs such as dose response curve estimation, cytometry analysis, and morphometry analysis. The integration module links treatment centric data back to treatment annotation files.

impact of such errors. For each parameter, we first calculated the median for each well, x_{ij}, which is the median for the jth well (for a 384-well plate, $j = 1,2,\ldots,384$) on the ith plate. The following multiplicative model was then used:

$$x_{ij} = p_i \cdot w_j \cdot \varepsilon_{ij}$$

where p_i is the plate effect, w_j is the well effect, and the residue, ε_{ij} is the value after normalization.

Applying logarithmic transformation to both sides of the above equation converts it into an additive model:

$$\log x_{ij} = \log p_i + \log w_j + \log \varepsilon_{ij}$$

ε_{ij}, which was subsequently used to correct the value of each cell in well j on plate i, was then determined using the median polish method by alternatively removing row and column medians [25]. The effects of normalization on plate-to-plate variation and plate edge effect—two typical systematic errors in high throughput screening—are shown in Figure 9.4.

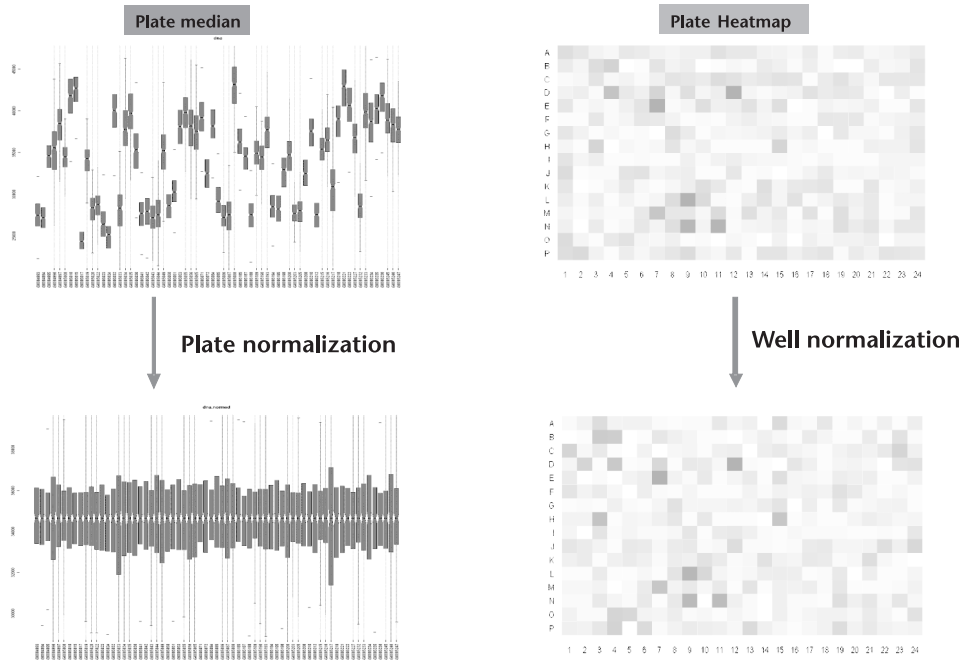


Figure 9.4 Preprocessing data normalization. Normalization removes plate-to-plate and well-to-well variations, thus allowing uniform analysis of the entire HCS dataset in other modules.

9.7.3 Dose Response and Confidence Estimation Module

EC50 estimation is the most commonly used method for determining the compound effect on cells. Here we used a four-parameter method to fit the dose response curve using all single cell measurements:

$$f(x) = \alpha + (\beta - \alpha)/(1 + exp[(x - \lambda)/\theta]$$

where α and β are the low and high responses, respectively. λ is the EC50 value, and θ is the slope of the dose response curve reflecting cooperativity.

Quantitative image analysis often results in significant data variation between cells and a rather small window of response to treatment (Figure 9.5). Fortunately we were still able to get an accurate measurement of the treatment effect because it is relatively easy to collect information for hundreds of cells in each HCS experiment, thus greatly increasing the confidence level of our estimation. But that comes with a price: acquiring more data requires more storage space and slows down the data analysis process. One important question here is what minimal cell number is necessary for a sufficiently confident measurement. The rule of thumb that was used was to measure 500 cells per treatment. To have an accurate estimation, we performed both theoretical analysis and numerical simulation. We found that the confidence interval (CI) for the cumulative distribution function is inversely proportional to the square root of the cell numbers. As shown in Figure 9.5, we can reach CI < 0.1 at about 250 cells.

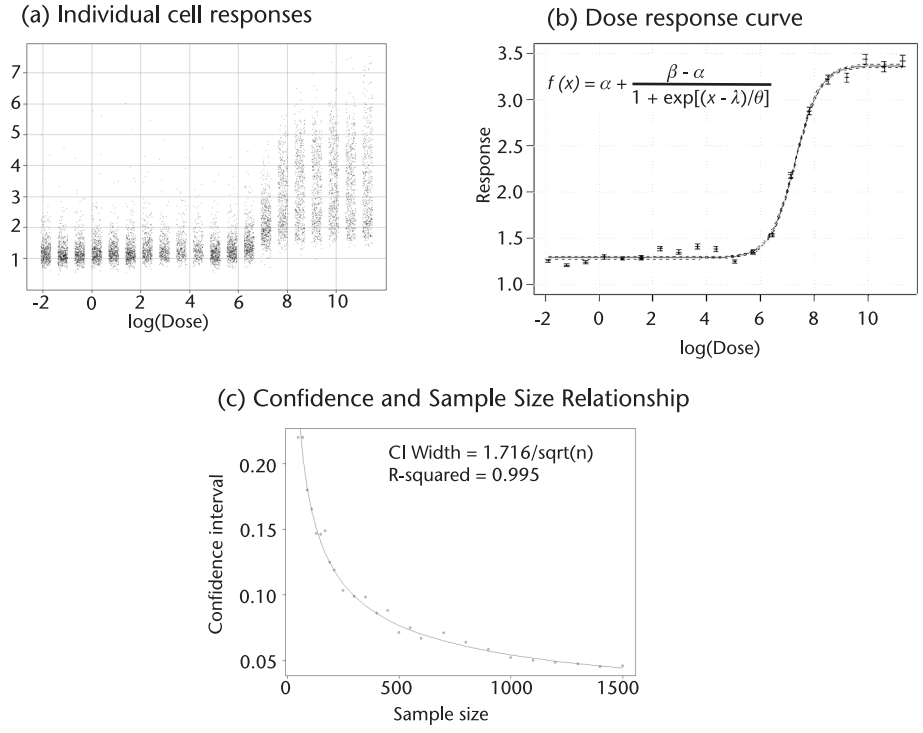


Figure 9.5 Dose response and confidence interval estimation. (b) Dose response curve was derived from (a) single cell responses using a four-parameter model. 95% confidence bands of each treatment point were shown by error bars, and curve fitting was shown by dotted lines. (c) The confidence interval of the cumulative distribution function is inversely proportional to the square root of sample size.

9.7.4 Automated Cytometry Classification Module

In a cell cycle assay, the percentage of cells in each cell cycle stage is the most crucial measurement. Cells were automatically classified into different phases of

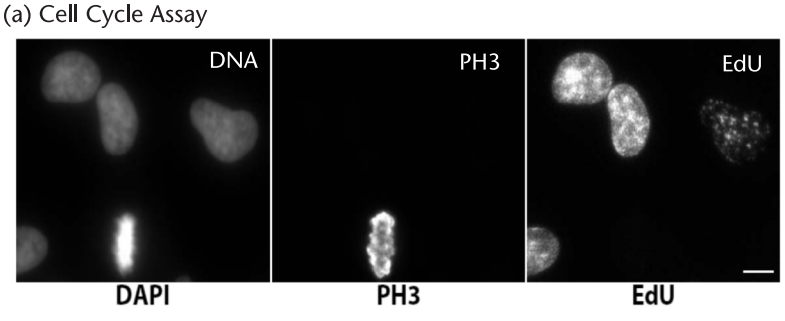


Figure 9.6 Decision tree cell cycle phase classification. (a) Cells were stained to show their levels of DNA, PH3, and EdU in each nuclei. (b) Every cell was classified as either G1 (2N DNA), G2 (4N DNA), M (PH3 positive), or S (EdU positive) phases using the automated four-parameter decision tree model. (c) The results of the classification were shown in two scatter plots where cells in different phases were shown.

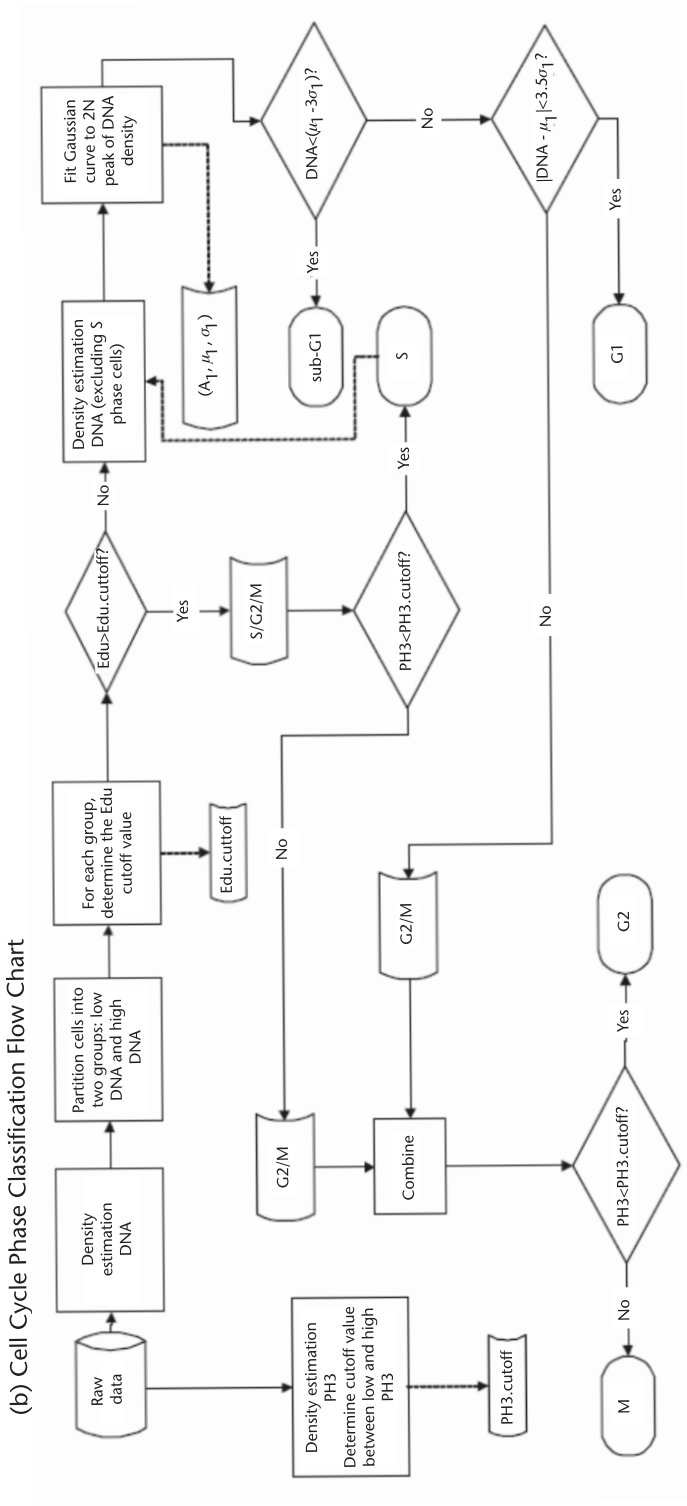


Figure 9.6 *(continued)*

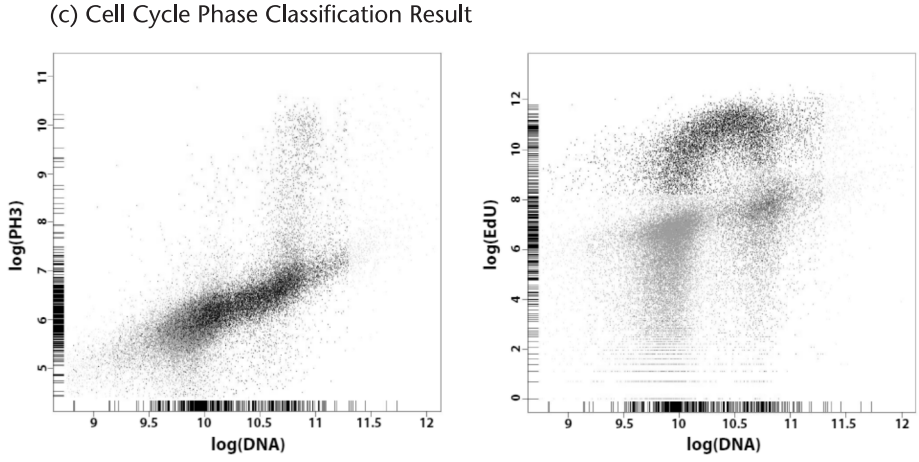


Figure 9.6 (*continued*)

the cell cycle on a plate-by-plate basis using a decision-tree based algorithm. An outline of the algorithm has the following steps (Figure 9.6):

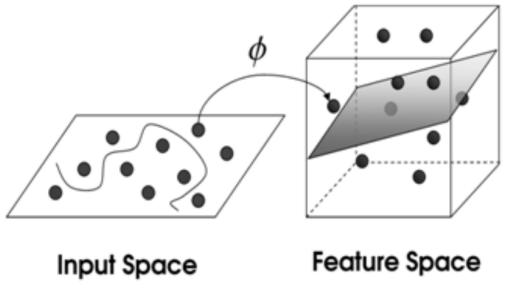
1. For each plate, estimate the distribution of DNA content and determine a cutoff value to roughly divide the cells into two groups: low and high DNA content. This step is necessary because the EdU distribution for the low DNA group usually differs from that of the high DNA group.
2. For each group, estimate the distribution of EdU content to partition the cells into low EdU and high EdU intensity groups.
3. Cells in the high EdU group are candidates for S, G2, or M phases. Those cells with low PH3 level are classified as S-phase; otherwise, G2/M.
4. Estimate the distribution of the DNA content again; this time excluding those cells in S-phase.
5. Fit a Gaussian curve, determined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right]$$

   to the 2*N* peak of the DNA distribution.
6. Classify cells with DNA content less than $\mu_1 - 3\sigma_1$ as *sub-G1*.
7. Cells with DNA content less than $\mu_1 + 3.5\sigma_1$ and greater than $\mu_1 - 3\sigma_1$ are classified as in G1 phase.
8. Combine cells with DNA content greater than $\mu_1 + 3.5\sigma_1$ with the G2/M cells from Step 3.
9. Classify those cells from step 8 and with low PH3 content as G2, otherwise, M.

Alternatively, a supervised method, such as neural network or support vector machine, can be used (Figure 9.7). This approach requires a small, often manually curated training set [26].

A. Support Vector Machine (SVM) Model



B. SVM Classification Accuracy

		Expert Annotated				
		Pro	Prometa	Meta	Ana/Telo	Lobed
SVM Predicted	Pro	**36**	1	2	0	4
	Prometa	10	**94**	7	0	1
	Meta	4	6	**95**	0	0
	Ana/Telo	2	0	0	**78**	0
	Lobed	2	0	0	1	**26**
Sensitivity (%)		66.7	93.1	91.3	98.7	83.9
Specificity (%)		97.8	93.3	96.2	99.3	99.1

Figure 9.7 Support vector machine mitotic subphase classification. (a) A supervised support vector machine (SVM) model was implemented to generate a multidimensional classification filter based on expert annotated images of cells in mitotic subphases. (b) Overall sensitivity and specificity of the method were above 90%.

The percentage of cells in each phase was then calculated for each well. Typically, most cells are in interphase (G1, S, or G2), with only a small number of cells in M-phase. Therefore, the percentages of cells in each cell cycle phase vary significantly among the phases and need to be standardized before the comparison of cell cycle profiles of different treatments can be made.

We introduced the *NZ* score for this purpose. If the percentage for well i is x_i, where $i = 1 \ldots N$ and N is the total number of wells in the reagent set, then the *NZ* score was calculated as follows:

$$NZ_i = \frac{x_i - \mu'}{1.4826 \cdot \sigma'}$$

where μ' is the median of $\{x_i; i = 1 \ldots N\}$ and σ' is the median absolute deviation, defined as the median of $\{|x_i - \mu'|; i = 1 \ldots N\}$. It can be shown that if x_i is normally distributed, *NZ* score is the same as the commonly used *Z* score, defined as $Z_i = (x_i - \mu)/\sigma$, where μ is the mean and σ is the standard deviation. *NZ* score was used as an alternative to the *Z* score as it is resistant to outliers which can occur frequently in high throughput screening; screen hits are by definition such outliers.

As an example, a positive S-phase *NZ* score indicates that the percentage of S-phase cells is higher than usual for the particular well and that treatment has likely resulted in a delay in S-phase progression. A negative S-phase *NZ* score indicates that the percentage of S-phase cells is low, and that treatment has likely resulted in a block at S-phase entry. Together, the four *NZ*-score numbers (one for each of G1, S, G2, M) give the cell cycle profile of a particular treatment.

9.8 Factor Analysis

A typical HCS experiment might generate gigabytes of numbers extracted from the images describing the amount and location of biomolecules on a cell-to-cell basis. Most of these numbers have no obvious biological meaning; for example, while the amount of DNA per nucleus has obvious significance, that of other nuclear measures, such as DNA texture, or nuclear ellipticity, are much less clear. This leads biologists to ignore the nonobvious measurements, even though they may report usefully on compound activities. A standard method in other fields for analyzing large, multidimensional datasets is factor analysis. It allows a large data-reduction but retains most of the information content, and quantifies phenotype using data-derived factors that are biologically interpretable in many cases. For this reason, factor analysis is highly appropriate to high content imaging, as it seeks to identify these underlying processes [27].

HCS data are contained in an $n \times m$ matrix, $\mathbf{X}$ consisting of a set of n image-based features measured on m cells. In mathematical terms, the so-called Common Factor Model posits that a set of measured random variables $\mathbf{X}$ is a linear function of common factors, $\mathbf{F}$ and unique factors, $\boldsymbol{\varepsilon}$:

$$\mathbf{X} = \mathbf{LF} + \boldsymbol{\varepsilon}$$

In HCS the common factors in $\mathbf{F}$ reflect the set of major phenotypic attributes measured in the assay. The loading matrix $\mathbf{L}$ relates the measured variables in $\mathbf{X}$ to $\mathbf{F}$. $\boldsymbol{\varepsilon}$ is a matrix of unique factors and is comprised of the reliable effects and the random error that is specific to a given variable. Rooted in this model is the concept that the total variance of $\mathbf{X}$ is partitioned into common and specific components. Therefore, after fitting the factor model and performing the rotations, we estimate the common attribute $\mathbf{F}$ on each of the k factors for each observation (i.e., cell) using a regression equation derived from the factor model (Figure 9.8) This is accomplished using the score procedure in SAS [28]. The complete factor structure and underlying phenotypic traits are outlined in Figure 9.8(d). As we can see in this case, the top six common factors have significant value and each contains a specific interpretable attribute.

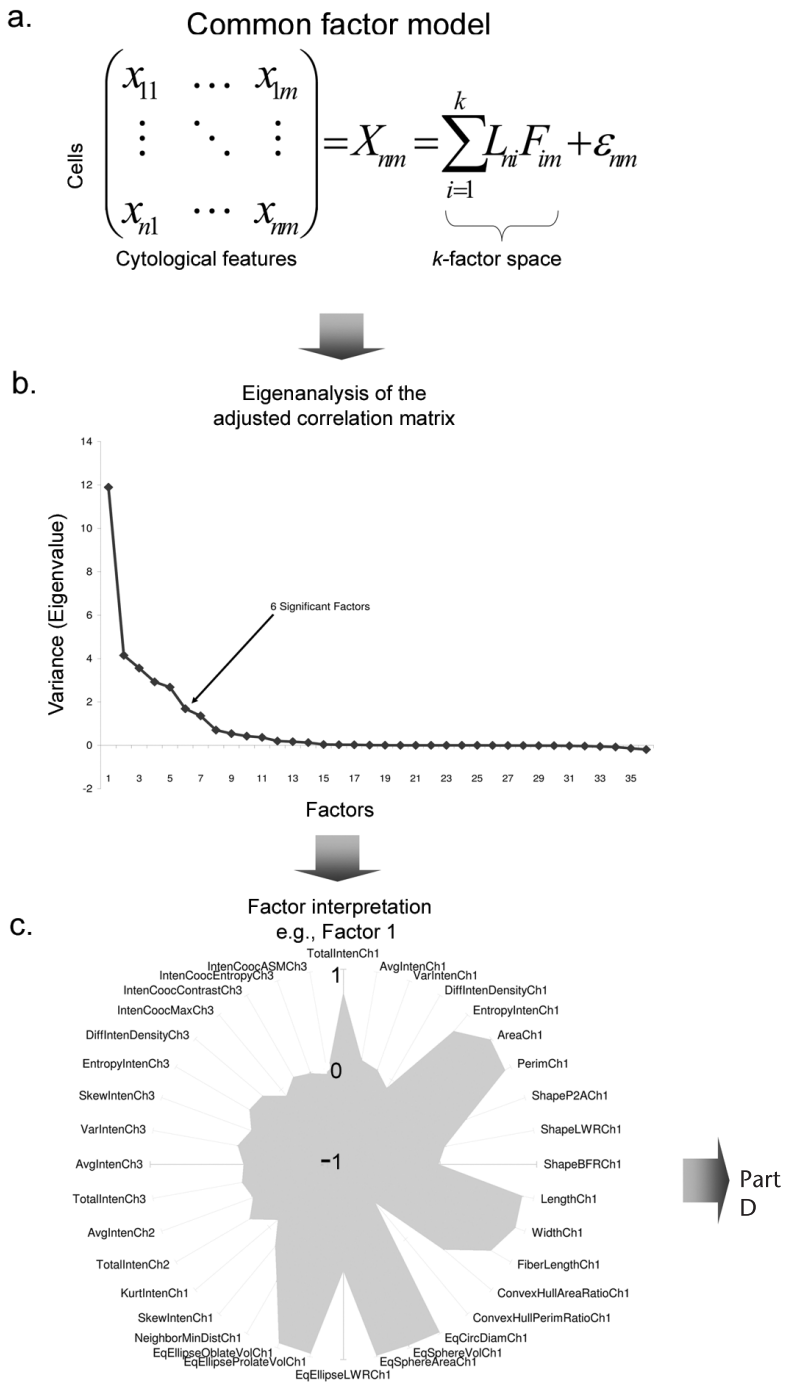


Figure 9.8 Data driven factor model of cell cycle. (a) HCS data are contained in an $n \times m$ matrix, X consisting of a set of n image-based cytological features measured on m cells. The common factor model maps the $n-$cytological features to a reduced $k-$dimensional space described by a set of factors, F, that reflect the major underlying phenotypic attributes measured in the assay. (b) The dimensionality of the factor space is determined by an eigen-analysis of the correlation matrix of the data matrix, X. We determine that there are six significant factors. (c) The loading matrix for factor 1, as an example. (d) The complete factor structure of the cell cycle assay is shown in this schematic. Each of the six factors is drawn with lines connected to the cytological features with which they are most significantly correlated. Our interpretation of the phenotypic attributes characterized by each factor is shown on the right.

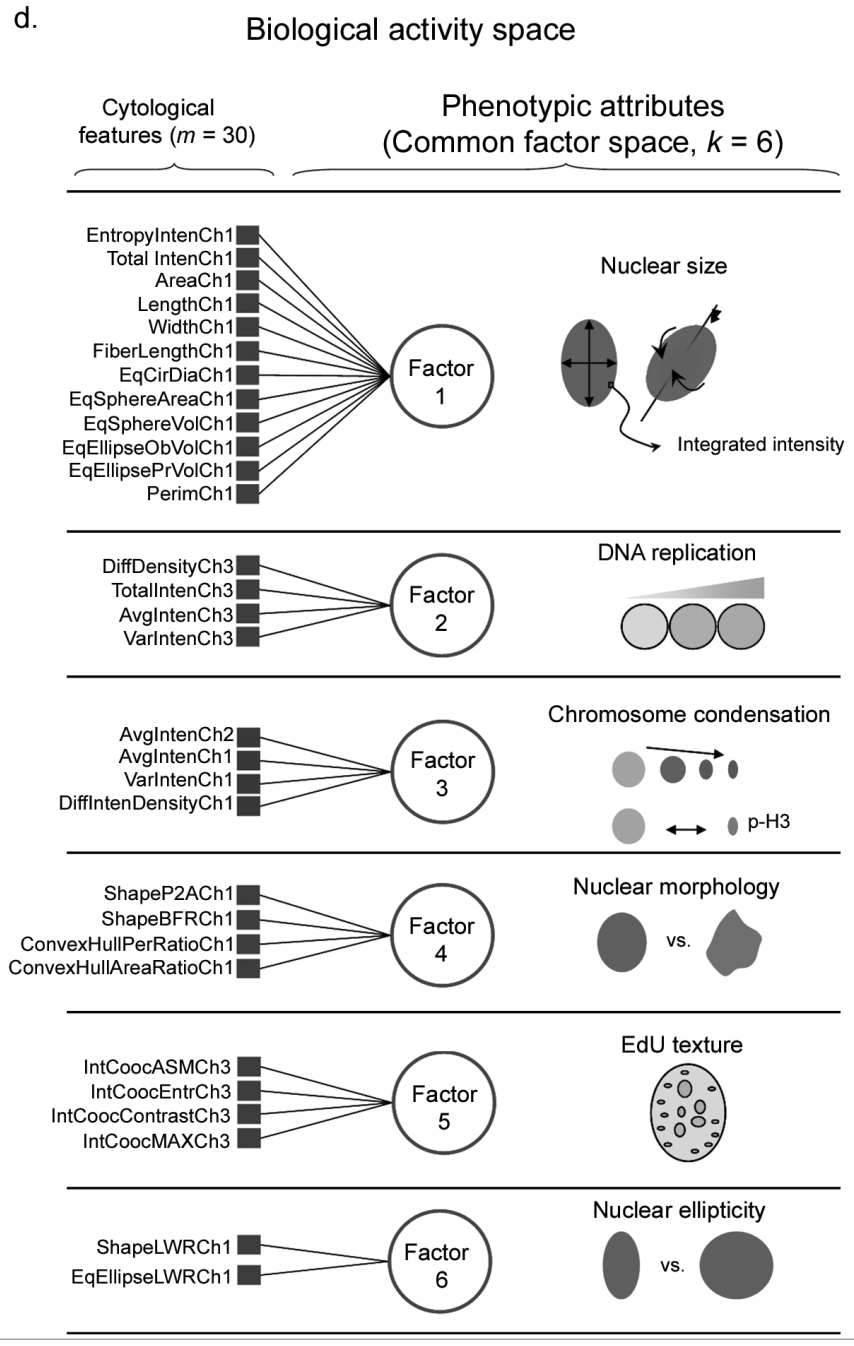


Figure 9.8 (*continued*)

9.9 Conclusion and Future Perspectives

HCS has increasingly been adopted in many aspects of drug discovery as well as system cell biology research. It has the potential to make major impacts on compound mechanistic and toxicology studies. During the past 10 years of development, many technical hurdles have been surpassed, including automated image acquisition, basic forms of image analysis, and data storage and archiving. Here we provide a practical basic framework on large scale HCS data analysis. We envision that more advanced statistical methods will be adopted in the near future to better mine the rich information contained in the HCS dataset. There are also significant efforts spent on improving image analysis, especially for morphology based HCS, and analysis of time lapsed datasets, as well as making basic image analysis tools more accessible for academic research. Statistical methods to mine and compare these rich datasets will become increasingly necessary.

Acknowledgments

We thank Jonathan Hoyt and Elizabeth McWhinnie for providing technical support.

References

[1] Taylor, D. L., J. R. Haskins, and K. A. Giuliano, (Eds.), "High Content Screening: A powerful approach to systems cell biology and drug discovery," *Methods Mol. Biol.*, 2007.

[2] Inglese, J., (Ed.), "Measuring biological responses with automated microscopy," *Methods Enzymol.*, 2006.

[3] Haney, S. A., (Ed.), *High Content Screening: Science, Techniques, and Applications*, Wiley Interscience Series, 2008.

[4] General Accounting Office report, "New drug development," 2006.

[5] Butcher, E. C., "Can cell systems biology rescue drug discovery?" *Nat. Rev. Drug Disc.*, 2005, Vol. 4, No. 6, pp. 461–467.

[6] Taylor, D. L., "Past, present, and future of high content screening and the field of cellomics," *Methods Mol. Biol.*, 2007, Vol. 356, pp. 3–18.

[7] Venkatesh, N., et al., "Chemical genetics to identify NFAT inhibitors: potential of targeting calcium mobilization in immunosuppression," *Proc. Natl. Acad. Sci. USA*, 2004, Vol. 101, No. 24, pp. 8969–74.

[8] Oakley, R. H., et al., "The cellular distribution of fluorescently labeled arrestins provides a robust, sensitive, and universal assay for screening G protein-coupled receptors," *Assay Drug Dev. Technol.*, 2002, Vol. 1, pp. 21–30.

[9] Wilson, C. J., et al., "Identification of a small molecule that induces mitotic arrest using a simplified high-content screening assay and data analysis method," *J. Screen.*, 2006, Vol. 11, No. 1, pp. 21–28.

[10] Liu, D., et al., "Screening of immunophilin ligands by quantitative analysis of neurofilament expression and neurite outgrowth in cultured neurons and cells," *J. Neurosci. Methods.*, 2007, Vol. 163, No. 2, pp. 310–320.

[11] Haney, S. A., et al., "High-content screening moves to the front of the line," *Drug Discov. Today*, 2006, Vol. 11, No. 19-20, pp. 889–894.

[12] Mitchison, T. J., "Small-molecule screening and profiling by using automated microscopy," *Chembiochem.*, 2005, Vol. 6, No. 1, pp. 33–39.

[13] Young, D. W., et al., "Integrating high-content screening and ligand-target prediction to identify mechanism of action," *Nat. Chem. Biol.*, 2008, Vol. 4, No. 1, pp. 59–68.

[14] Ding, G. J., et al., "Characterization and quantitation of NF-kappaB nuclear translocation induced by interleukin-1 and tumor necrosis factor-alpha. Development and use of a high capacity fluorescence cytometric system," *J. Biol. Chem.*, 1998, Vol. 273, No. 44, pp. 28897–28905.

[15] Knauer, S. K., "Translocation biosensors to study signal-specific nucleo-cytoplasmic transport, protease activity and protein-protein interactions," *Traffic*, 2005, Vol. 6, No. 7, pp. 594–606.

[16] Giuliano, K. A., "Fluorescent proteins in drug development," *Expert Rev. Mol. Diagn.*, 2007, Vol. 7, No. 1, pp. 9–10.

[17] Moffat, J., "A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen," *Cell*, 2006, Vol. 124, No. 6, pp. 1283–1298.

[18] Srinivasan, R., et al., "Automated axon tracking of 3D confocal laser scanning microscopy images using guided probabilistic region merging," *Neuroinformatics*, 2007, Vol. 5, No. 3, pp. 189–203.

[19] Wagner, B. K., "Small-molecule fluorophores to detect cell-state switching in the context of high-throughput screening," *J. Am. Chem. Soc.*, 2008, Vol. 130, No. 13, pp. 4208–4209.

[20] Carpenter, A. E., "CellProfiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biol.*, 2006, Vol. 7, No. 10, p. R100.

[21] Barak, L. S., et al., "A beta-arrestin/green fluorescent protein biosensor for detecting G protein-coupled receptor activation," *J. Biol. Chem.*, 1997, Vol. 272, No. 44, pp. 27497–27500.

[22] Li, F., "An automated feedback system with the hybrid model of scoring and classification for solving over-segmentation problems in RNAi high content screening," *J. Microsc.*, 2007, Vol. 226, No. 2, pp. 121–132.

[23] Baatz, M., "Object-oriented image analysis for high content screening: detailed quantification of cells and sub cellular structures with the Cellenger software," *Cytometry A*, 2006, Vol. 69, No. 7, pp. 652–658.

[24] Haralick, R. M., "Statistical and structural approaches to texture," *Proceedings of the IEEE*, 1979, Vol. 67, No. 5, pp. 786–804.

[25] Tukey, J. W., *Exploratory Data Analysis*, Reading, MA: Addison-Wesley, 1977.

[26] Tao, C. Y., J. Hoyt, and Y. Feng, "A support vector machine classifier for recognizing mitotic subphases using high-content screening data," *J. Biomol. Screen.*, 2007, Vol. 12, No. 4, pp. 490–496.

[27] Spearman, C., "General Intelligence, Objectively Determined and Measured," *American Journal of Psychology*, 1904, Vol. 15, pp. 201–293.

[28] Hatcher, L. *A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling,* Cary, NC: SAS Institute, Inc., 1994.